

# Direct minimization for calculating invariant subspaces in density functional computations of the electronic structure \*

R. Schneider, T. Rohwedder, A. Neelov, J. Blauert

## Abstract

In this article, we analyse three related preconditioned steepest descent algorithms, which are partially popular in Hartree-Fock and Kohn-Sham theory as well as invariant subspace computations, from the viewpoint of minimization of the corresponding functionals, constrained by orthogonality conditions. We exploit the geometry of the of the admissible manifold, i.e. the invariance with respect to unitary transformations, to reformulate the problem on the Grassmann manifold as the admissible set. We then prove asymptotical linear convergence of the algorithms under the condition that the Hessian of the corresponding Lagrangian is elliptic on the tangent space of the Grassmann manifold at the minimizer.

## 1 Introduction

On the length-scale of atomistic or molecular systems, physics is governed by the laws of quantum mechanics. A reliable computation required in various fields in modern sciences and technology should therefore be based on the first principles of quantum mechanics, so that *ab initio* computation of the electronic wave function from the stationary electronic Schrödinger equation is a major working horse for many applications in this area. To reduce computational demands, the high dimensional problem of computing the wave function for  $N$  electrons is often, for example in Hartree-Fock and Kohn-Sham theory, replaced by a nonlinear system of equations for a set  $\Phi = (\varphi_1, \dots, \varphi_N)$  of single particle wave functions  $\varphi_i(\mathbf{x}) \in V = H^1(\mathbb{R}^3)$ . This ansatz corresponds to the following abstract formulation for the minimization of a suitable energy functional  $\mathcal{J}$ :

---

\*This work was supported by the DFG SPP 1445: “Modern and universal first-principles methods for many-electron systems in chemistry and physics” and the EU NEST project BigDFT.

**Problem 1:** Minimize

$$\mathcal{J} : V^N \rightarrow \mathbb{R}, \quad \mathcal{J}(\Phi) = \mathcal{J}(\varphi_1, \dots, \varphi_N) \longrightarrow \min, \quad (1.1)$$

where  $\mathcal{J}$  is a sufficiently often differentiable functional which is

(i) invariant with respect to unitary transformations, i.e.

$$\mathcal{J}(\Phi) = \mathcal{J}(\Phi \mathbf{U}) = \mathcal{J}\left(\left(\sum_{j=1}^N u_{i,j} \phi_j\right)_{i=1}^N\right), \quad (1.2)$$

for any orthogonal matrix  $\mathbf{U} \in \mathbb{R}^{n \times n}$ , and

(ii) subordinated to the orthogonality constraints

$$\langle \varphi_i, \varphi_j \rangle := \int_{\mathbb{R}^3} \varphi_i(x) \varphi_j(x) dx = \delta_{i,j}. \quad (1.3)$$

In the present article, we shall be concerned with minimization techniques for  $\mathcal{J}$  along the admissible manifold characterized by (1.3). The first step towards this will be to set up the theoretical framework of the *Grassmann manifold* to be introduced in section 2, reflecting the constraints (i) and (ii) imposed on the functional  $\mathcal{J}$  and the minimizer  $\Phi$ , respectively. In applications in electronic structure theory, formulation of the first order optimality (necessary) condition for the problem (1.1) results in a nonlinear eigenvalue problem of the kind:

$$A_\Phi \varphi_i = \lambda_i \varphi_i, \quad \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N \quad (1.4)$$

for  $N$  eigenvalues  $\lambda_i$  and the corresponding solution functions assembled in  $\Phi$ . In these equations, the operator  $A_\Phi$ , is a symmetric bounded linear mapping  $A_\Phi : V = H^1(\mathbb{R}^3) \rightarrow V' = H^{-1}(\mathbb{R}^3)$  depending on  $\Phi$ , so that we are in fact faced with a nonlinear eigenvalue problem.  $A_\Phi$  is called the *Fock operator* in Hartree-Fock theory, and *Kohn-Sham Hamiltonian* in *density functional theory* (DFT) respectively. We will illustrate the relation between (1.4) and the minimization task above in further detail in section 3. In this work, our emphasis will rather be on the algorithmic approximation of the minimizer of  $\mathcal{J}$ , i.e. an invariant subspace  $\text{span}[\Phi] := \text{span}\{\varphi_1, \dots, \varphi_N\}$ , of (1.4), in the corresponding energy space  $V^N$  than on computation of the eigenvalues  $\lambda_1, \dots, \lambda_N$ .

One possible procedure for computing the minimum of  $\mathcal{J}$  is the so-called *direct minimization*, utilized e.g. in DFT calculation, which performs a steepest descent algorithm by updating the gradient of  $\mathcal{J}$ , i.e. the Kohn-Sham Hamiltonian or Fock operator, in each iteration step. Direct minimization, as proposed in [2], is prominent in DFT calculations if good preconditioners are available and the systems under consideration are large, e.g.

for the computation of electronic structure in bulk crystals using plane waves, finite differences [7] and the recent wavelet code developed in the BigDFT project (see [45]). In contrast to the direct minimization procedure is the *self consistent field iteration (SCF)*, which keeps the Fock operator fixed until convergence of the corresponding eigenfunctions and updates the Fock operator thereafter, see section 3.

In the rest of this article, we will pursue different variants of projected gradient algorithms to be compiled in section 4. In addition, we will (for the case where the gradient  $\mathcal{J}'(\Phi)$  can be written as an operator  $A_\Phi$  applied to  $\Phi$ , as it is the case in electronic structure calculation) investigate an algorithm based on [4] following a preconditioned steepest descent along geodesics on the manifold. so that no re-projections onto the admissible manifold are required. It turns out that all these algorithms to be proposed perform in a similar way. For matters of rigorous mathematical analysis, let us note at this point that the mathematical theory about Hartree-Fock is still too incomplete to prove the assumptions required in the present paper; even less is known for Kohn-Sham equations, due to the fact that there are so many different models used in practice. If the assumptions are not met for a particular problem, it is not clear whether it is a deficiency of the problem or a real pathological situation. Along with (1.1), we will therefore consider the following simplified prototype problem for a *fixed* operator  $A$ :

**Simplified problem 2:** Minimize

$$\mathcal{J}_A(\varphi_1, \dots, \varphi_N) := \sum_{i=1}^N \langle \varphi_i, A\varphi_i \rangle \longrightarrow \min, \quad \langle \varphi_i, \varphi_j \rangle = \delta_{i,j}. \quad (1.5)$$

Analogous treatment with Lagrange techniques shows that this special case of problem 1 is the problem of computing the first  $N$  eigenfunctions, resp. the lowest  $N$  eigenvalues of  $A$  (see Lemma 3). While this is an interesting problem by itself, e.g. if  $\lambda$  is an eigenvalue of multiplicity  $N$ , it is also of interest as a sort of prototype: Properties that can be proven for this problem may hold in the more general case for Hartree-Fock or Kohn-Sham. In particular, we will show that for  $A$  symmetric and bounded from below, the Hessian of the Lagrangian, taken at the solution  $\Psi$ , is elliptic on a specific tangent manifold at  $\Psi$ , an essential ingredient to prove linear convergence of all of the proposed algorithms in section 5. The same convergence results will be shown to hold for (1.1) if we impose this ellipticity condition on the Lagrangian of  $\mathcal{J}$  of the nonlinear problem. Note that the problem type (1.5) also arises in many other circumstances, which we will not consider here in detail. Let us just note that the algorithms presented in section 4 also provide reasonable routines for the inner cycles of the SCF procedure.

In the context of eigenvalue computations, variants of our basic algorithm 1, applied to problem 2, have been considered by several authors (see e.g. [8, 28, 34]) reporting excellent performance, in particular if subspace acceleration techniques are applied and the preconditioner is chosen appropriately; in [40, 11], an adaptive variant was recently proposed

and analysed for the simpler case  $N = 1$ . In contrast to all these papers, we will view the algorithms as steepest descent algorithms for optimization of  $\mathcal{J}$  under the orthogonality constraints given above, as such a systematic treatment does not only simplify the proofs but also provides the insight necessary to understand the direct minimization techniques for the more complicated nonlinear problems of the kind (1.1) in DFT and HF.

Our analysis will cover closed (usually finite dimensional) subspaces of  $V_h \subset V$  as well as the energy space  $V$  itself, so that finite dimensional approximations by Ritz-Galerkin methods and also finite difference approximations are included in our analysis. In particular, our results are also valid if Gaussian type basis functions are used. The convergence rates will be independent of the discretization parameters like mesh size. However, the choice of an appropriate preconditioning mapping to be used in our algorithms is crucial. Fortunately, such preconditioners can often easily be constructed, e.g. by the use of multi-grid methods for finite elements, finite differences or wavelets, polynomials [25, 2, 7]. Our analysis will show that for the gradient algorithms under consideration, it suffices to use a *fixed* preconditioner respectively relaxation parameter. In particular, no expensive line search is required.

All results proven will be local in nature meaning that the initial guess is supposed to be already sufficiently close to the exact one. At the present stage, we will for the sake of simplicity consider only real valued solutions for the minimization problem. Nevertheless, complex valued functions can be treated by minor modifications. Note that since the present approach is completely based on a variational framework, i.e. considering a constrained optimization problem, it does not include unsymmetric eigenvalue problems or the computation of other eigenvalues than the lowest ones.

## 2 Optimization on Grassmann manifolds

The invariance of the functional  $\mathcal{J}$  with respect to uniform transformations among the eigenfunctions shows a certain redundancy inherent in the formulation of the minimization task (1.1). Therefore, it will be more advantageous to factor out the unitary invariance of the functional  $\mathcal{J}$ , resulting in the usage of the Stiefel and Grassmann manifolds, originally defined in finite dimensional Euclidean Hilbert spaces in [4], see also [1] for an extensive exposition. In this section, we will generalize this concept for the present infinite dimensional space  $V^N$  equipped with the  $L_2$  inner product. In the next section, we will then apply this framework to the minimization problems for the HF and KS functionals. First of all, we shall briefly introduce the spaces under consideration and some notations.

## 2.1 Basic notations

Letting  $H = L_2 := L_2(\mathbb{R}^3)$  or a closed subspace of  $L_2$ , we will work with a Gelfand triple  $V \subset H \subset V'$  with the usual  $L_2$  inner product  $\langle \cdot, \cdot \rangle$ , as dual pairing on  $V' \times V$ , where either  $V := H^1 = H^1(\mathbb{R}^3)$  or an appropriate subspace corresponding to a Galerkin discretization. Because the ground state is determined by a set  $\Phi$  of  $N$  one-particle functions  $\varphi_i \in V$ , we will formulate the optimization problem on an admissible subset of  $V^N$ . To this end, we extend inner products and operators from  $V$  to  $V^N$  by the following

**Definitions 1.** For  $\Psi = (\psi_1, \dots, \psi_N) \in V^N$ ,  $\Phi = (\varphi_1, \dots, \varphi_N) \in (V^N)' = (V')^N$ , and the  $L_2$  inner product  $\langle \cdot, \cdot \rangle$  given on  $H = L_2$ , we denote

$$\langle \Phi^T \Psi \rangle := (\langle \varphi_i, \psi_j \rangle)_{i,j=1}^N \in \mathbb{R}^{N \times N},$$

and introduce the dual pairing

$$\langle \langle \Phi, \Psi \rangle \rangle := \text{tr} \langle \Phi^T \Psi \rangle = \sum_{i=1}^N \langle \varphi_i, \psi_i \rangle$$

on  $(V')^N \times V^N$ .

Because there holds  $V^N = V \otimes \mathbb{R}^N$ , we can canonically expand any operator  $R : V \rightarrow V'$  to an operator

$$\mathcal{R} := R \otimes I : V^N = V \otimes \mathbb{R}^N \rightarrow V'^N, \Phi \mapsto \mathcal{R}\Phi = (R\varphi_1, \dots, R\varphi_N). \quad (2.1)$$

Throughout this paper, for an operator  $V \rightarrow V'$  denoted by a capital letter as  $A, B, D, \dots$ , the same calligraphic letter  $\mathcal{A}, \mathcal{B}, \mathcal{D}, \dots$ , will denote this expansion to  $V^N$ .

Further, we will make use of the following operations:

**Definitions 2.** For  $\Phi \in V^N$  and  $\mathbf{M} \in \mathbb{R}^{N \times N}$ , we define the set  $\Phi \mathbf{M} = (I \otimes \mathbf{M})\Phi$  by  $(\Phi \mathbf{M})_j := \sum_{i=1}^N m_{i,j} \varphi_i$ , cf. also the notation in (1.2), and for  $\phi \in V$  and  $v = (v_1, \dots, v_N) \in \mathbb{R}^N$  the element  $\phi \otimes v \in V^N$  by  $(v_1 \phi, \dots, v_N \phi)$ . Finally, we denote by  $O(N)$  the orthogonal group of  $\mathbb{R}^{N \times N}$ .

## 2.2 Geometry of Stiefel and Grassmann manifolds

Let us now introduce the admissible manifold and prove some of its basic properties. Note in this context that well established results of [4] for the case in the finite dimensional Euclidean spaces cannot be applied to our setting without further difficulties, because the norm induced by the  $L_2$  inner product is weaker than the present  $V$ -norm.

Our aim is to minimize the functionals  $\mathcal{J}(\Phi)$ , where  $\mathcal{J}$  is either  $\mathcal{J}_{HF}$ ,  $\mathcal{J}_{KS}$  or  $\mathcal{J}_A$ , under the orthogonality constraint  $\langle \varphi_i, \varphi_j \rangle = \delta_{i,j}$ , i.e.

$$\langle \Phi^T \Phi \rangle = \mathbf{I} \in \mathbb{R}^{N \times N}. \quad (2.2)$$

The subset of  $V^N$  satisfying the property (2.2) is called the *Stiefel manifold* (cf. [4])

$$\mathcal{V}_{V,N} := \{ \Phi = (\varphi_i)_{i=1}^N \mid \varphi_i \in V, \langle \Phi^T \Phi \rangle - \mathbf{I} = \mathbf{0} \in \mathbb{R}^{N \times N} \},$$

i.e. the set of all orthonormal bases of  $N$ -dimensional subspaces of  $V$ .

All functionals  $\mathcal{J}$  under consideration are unitarily invariant, i.e. there holds (1.2). To get rid of this nonuniqueness, we will identify all orthonormal bases  $\Phi \in \mathcal{V}_{V,N}$  spanning the same subspace  $V_\Phi := \text{span} \{ \varphi_i : i = 1, \dots, N \}$ . To this end we consider the *Grassmann manifold*, defined as the quotient

$$\mathcal{G}_{V,N} := \mathcal{V}_{V,N} / \sim$$

of the Stiefel manifold with respect to the equivalence relation  $\Phi \sim \tilde{\Phi}$  if  $\tilde{\Phi} = \Phi \mathbf{U}$  for any  $\mathbf{U} \in O(N)$ . We usually omit the indices and write  $\mathcal{V}$  for  $\mathcal{V}_{V,N}$ ,  $\mathcal{G}$  for  $\mathcal{G}_{V,N}$  respectively. To simplify notations we will often also work with representatives instead of equivalence classes  $[\Phi] \in \mathcal{G}$ .

The interpretation of the Grassmann manifold as equivalence classes of orthonormal bases spanning the same  $N$ -dimensional subspace is just one way to define the Grassmann manifold. We can as well identify the subspaces with orthogonal projectors onto these spaces. To this end, let us for  $\Phi = (\varphi_1, \dots, \varphi_N) \in \mathcal{V}^N$  denote by  $D_\Phi$  the  $L_2$ -orthogonal projector onto  $\text{span}\{\varphi_1, \dots, \varphi_N\}$ . It is straightforward to verify

**Lemma 1.** *There is a one to one relation identifying  $\mathcal{G}$  with the set of rank  $N$   $L_2$ -orthogonal projection operators  $D_\Phi$ .*

In the following, we will compute the tangent spaces of the manifolds defined above for later usage.

**Proposition 1.** *The tangent space of the Stiefel manifold at  $\Phi \in \mathcal{V}$  is given by*

$$\mathcal{T}_\Phi \mathcal{V} = \{ X \in V^N \mid \langle X^T \Phi \rangle = - \langle \Phi^T X \rangle \in \mathbb{R}^{N \times N} \}.$$

*The tangent space of the Grassmann manifold is*

$$\begin{aligned} \mathcal{T}_{[\Phi]} \mathcal{G} &= \{ W \in V^N \mid \langle W^T \Phi \rangle = \mathbf{0} \in \mathbb{R}^{N \times N} \} \\ &= (\text{span}\{\varphi_1, \dots, \varphi_N\}^\perp)^N. \end{aligned}$$

*Thus, the operator  $(\mathcal{I} - D_\Phi)$ , where  $D_\Phi$  is the  $L_2$ -projector onto the space spanned by  $\Phi$ , is an  $L_2$ -orthogonal projection from  $V^N$  onto the tangent space  $\mathcal{T}_{[\Phi]} \mathcal{G}$ .*

*Proof.* If we compute the Fréchet derivative of the constraining condition

$$g(\Phi) := \langle \Phi^T \Phi \rangle - \mathbf{I} = \mathbf{0}$$

for the Stiefel manifold, the first result follows immediately. To prove the second result, we consider the quotient structure of the Grassmann manifold and decompose the tangent space  $\mathcal{T}_\Phi \mathcal{V}$  of the Stiefel manifold at the representative  $\Phi$  into a component tangent to the set  $[\Phi]$ , which we call the *vertical space*, and a component containing the elements of  $\mathcal{T}_\Phi \mathcal{V}$  that are orthogonal to the vertical space, the so-called *horizontal space*. If we move on a curve in the Stiefel manifold with direction in the vertical space, we do not leave the equivalence class  $[\Phi]$ . Thus only the horizontal space defines the tangent space of the quotient  $\mathcal{G} = \mathcal{V}/O(N)$ . The horizontal space is computed in the following lemma, from which the claim follows.  $\square$

**Lemma 2.** *The vertical space at a point  $\Phi \in \mathcal{V}$  (introduced in the proof of proposition 1) is the set*

$$\{\Phi \mathbf{M} \mid \mathbf{M} = -\mathbf{M}^T \in \mathbb{R}^{N \times N}\}.$$

*The horizontal space is given by*

$$\{W \in V^N \mid \langle W^T \Phi \rangle = \mathbf{0} \in \mathbb{R}^{N \times N}\}.$$

*Proof.* To compute the tangent vectors of the set  $[\Phi]$ , we consider a curve  $c(t)$  in  $[\Phi]$  emanating from  $\Phi$ . Then  $c$  is of the form  $c(t) = \Phi \mathbf{U}(t)$  for a curve  $\mathbf{U}(t) \in O(N)$  with  $\mathbf{U}(0) = \mathbf{I}_{N \times N}$ . Differentiating  $\mathbf{I}_{N \times N} = \mathbf{U}(t)\mathbf{U}(t)^T$  at  $t = 0$  yields  $\mathbf{U}'(0) = -\mathbf{U}'(0)^T$  and we get that every vector of the vertical space is of the form  $\Phi \mathbf{M}$  where  $\mathbf{M}$  is skew symmetric. Reversely, for any skew symmetric matrix  $\mathbf{M}$  we find a curve  $\mathbf{U}(t)$  in  $O(N)$  emanating from  $\Phi$  with direction  $\mathbf{M}$ , and  $c(t) := \Phi \mathbf{U}(t)$  is a curve with direction  $\dot{c}(0) = \Phi \mathbf{M}$ , and thus the first assertion follows.

To compute the horizontal space, we decompose  $W \in \mathcal{T}_\Phi \mathcal{V}$  into  $W = \Phi \mathbf{M} + W_\perp$ , where  $W_\perp := W - \Phi \langle \Phi^T W \rangle \in \Phi^\perp$ ,  $\mathbf{M} := \langle \Phi^T W \rangle$ . Then  $\mathbf{M}$  is an antisymmetric matrix, which implies that  $\Phi \mathbf{M}$  is in the vertical space, and that the horizontal space is given by all  $\{W_\perp = W - \Phi \langle \Phi^T W \rangle \mid W \in \mathcal{T}_\Phi \mathcal{V}\}$ . Let us note that this set is the range of the operator  $(I - \mathcal{D}_\Phi)$ . This operator is continuous and of finite codimension. If  $W_\perp = W - \Phi \langle \Phi^T W \rangle$  is in the horizontal space, then

$$\langle W_\perp^T \Phi \rangle = \langle W^T \Phi \rangle - \langle \Phi^T \Phi \rangle \langle W^T \Phi \rangle = \mathbf{0}.$$

Reversely, if  $W \in V^N$  with  $\langle W^T \Phi \rangle = \mathbf{0}$ , then  $W$  is in  $\mathcal{T}_\Phi \mathcal{V}$  and from  $(I - \mathcal{D}_\Phi)W = W - \Phi \langle \Phi^T W \rangle = W$  we get that  $W$  is in the range of  $I - \mathcal{D}_\Phi$ , being the  $L_2$ -orthogonal projection from  $V^N$  onto the tangent space  $T_{[\Phi]} \mathcal{G}$ .  $\square$

To end this section, let us prove a geometric result needed later.

**Lemma 3.** Let  $[\Psi] \in \mathcal{G}$ ,  $D^*$  the  $L_2$ -projector on  $\text{span}[\Psi]$ ,  $\mathcal{D}^*$  is its expansion as above and  $\|\cdot\|$  is the norm induced by the  $L_2$  inner product. For any  $\Phi = (\varphi_1, \dots, \varphi_N) \in \mathcal{V}$  sufficiently close to  $[\Psi] \in \mathcal{G}$  in the sense that for all  $i \in \{1, \dots, N\}$ ,  $\|(I - D^*)\varphi_i\| < \delta$ , there exists an orthonormal basis  $\bar{\Psi} \in \mathcal{V}$  of  $\text{span}[\Psi]$  for which

$$\Phi - \bar{\Psi} = (I - \mathcal{D}^*)\Phi + \mathcal{O}(\|(I - \mathcal{D}^*)\Phi\|^2).$$

*Proof.* For  $i = 1, \dots, N$ , let

$$\tilde{\psi}_i = \arg \min\{\|\psi - \varphi_i\|, \psi \in \text{span}\{\psi_i | i = 1, \dots, N\}, \|\psi\| = 1\} = D^*\varphi_i / \|D^*\varphi_i\|,$$

and set  $\tilde{\Psi} := (\tilde{\psi}_1, \dots, \tilde{\psi}_N)$ . If we denote by  $\tilde{P}_i$  the  $L_2$  projector on the space spanned by  $\tilde{\psi}_i$ , it is straightforward to see from the series expansion of the cosine that

$$(I - D^*)\varphi_i = (I - \tilde{P}_i)\varphi_i = \varphi_i - \tilde{\psi}_i + \mathcal{O}(\|(I - D^*)\varphi_i\|^2) \quad (2.3)$$

The fact that  $\tilde{\Psi} \notin \mathcal{V}$  is remedied by orthonormalization of  $\tilde{\Psi}$  by the Gram-Schmidt procedure. For the inner products occurring in the orthogonalization process (for which  $i \neq j$ ), there holds

$$\begin{aligned} \langle \tilde{\psi}_i, \tilde{\psi}_j \rangle &= \langle \tilde{\psi}_i - \varphi_i, \tilde{\psi}_j \rangle + \langle \varphi_i, \tilde{\psi}_j - \varphi_j \rangle + \langle \varphi_i, \varphi_j \rangle \\ &= -\langle (I - D^*)\varphi_i, \tilde{\psi}_j \rangle - \langle (I - D^*)\varphi_i, (I - D^*)\varphi_j \rangle + \mathcal{O}(\|(I - D^*)\varphi_i\|^2). \\ &= \mathcal{O}(\|(I - \mathcal{D}^*)\Phi\|^2) \end{aligned}$$

where we have twice replaced  $\varphi_i - \tilde{\psi}_i$  by  $(I - D^*)\varphi_i$  according to (2.3) and made use of the orthogonality of  $D^*$ . In particular, for  $\Phi$  sufficiently close to  $[\Psi]$ , the Gramian matrix is non-singular because the diagonal elements converge quadratically to one while the off-diagonal elements converge quadratically to zero. By an easy induction for the orthogonalization process and a Taylor expansion for the normalization process, we obtain that  $\tilde{\Psi}$  differs from the orthonormalized set  $\bar{\Psi} := (\bar{\psi}_1, \dots, \bar{\psi}_N)$  only by a error term depending on  $\|(I - \mathcal{D}^*)\Phi\|^2$ . Therefore,

$$\varphi_i - \bar{\psi}_i = \varphi_i - \tilde{\psi}_i + \mathcal{O}(\|(I - \mathcal{D}^*)\Phi\|^2) = (I - D^*)\varphi_i + \mathcal{O}(\|(I - \mathcal{D}^*)\Phi\|^2),$$

so that

$$\Phi - \bar{\Psi} = (I - \mathcal{D}^*)\Phi + \mathcal{O}(\|(I - \mathcal{D}^*)\Phi\|^2),$$

and the result is proven.  $\square$

## 2.3 Optimality conditions on the Stiefel manifold

By the first order optimality condition for minimization tasks, a minimizer  $[\Psi] \in \mathcal{G}$  of the functional  $\mathcal{J} : \mathcal{G} \rightarrow \mathbb{R}, \Phi \mapsto \mathcal{J}(\Phi)$  over the Grassmann manifold  $\mathcal{G}$  satisfies

$$\langle \langle \mathcal{J}'(\Psi), \delta\Phi \rangle \rangle = 0 \quad \text{for all } \delta\Phi \in \mathcal{T}_{[\Psi]}\mathcal{G}, \quad (2.4)$$

i.e. the gradient  $\mathcal{J}'(\Psi) \in (V')^N = (V^N)'$  vanishes on the tangent space  $\mathcal{T}_{\Psi}\mathcal{G}$  of the Grassmann manifold. This property can also be formulated by

$$\langle \langle \delta\Phi \rangle^T \mathcal{J}'(\Psi) \rangle = \mathbf{0} \quad \text{for all } \delta\Phi \in \mathcal{T}_{[\Psi]}\mathcal{G},$$

or equivalently, by Lemma 1,

$$\langle \langle (\mathcal{I} - \mathcal{D}_{\Psi})\mathcal{J}'(\Psi), \Phi \rangle \rangle = 0 \quad \text{for all } \Phi \in V^N, \quad (2.5)$$

that is, in strong formulation,

$$(\mathcal{I} - \mathcal{D}_{\Psi})\mathcal{J}'(\Psi) = \mathcal{J}'(\Psi) - \Psi\Lambda = 0 \in (V')^N, \quad (2.6)$$

where  $\Lambda = (\langle \langle \mathcal{J}'(\Psi) \rangle_j, \psi_i \rangle)_{i,j=1}^N$  and  $(\mathcal{J}'(\Psi))_i \in V'$  is the  $i$ -th component of  $\mathcal{J}'(\Psi)$ . Note that this corresponds to one of the optimality conditions for the Lagrangian yielded from the common approach of the Euler-Lagrange minimization formalism: Introducing the Lagrangian

$$\mathcal{L}(\Phi, \Lambda) := \frac{1}{2} \left( \mathcal{J}(\Phi) + \sum \lambda_{i,j} (\langle \varphi_i, \varphi_j \rangle_{L_2} - \delta_{i,j}) \right), \quad (2.7)$$

the condition for the derivative restricted to  $V^N$ , here denoted by  $\mathcal{L}^{(1,\Psi)}(\Psi, \Lambda)$  for convenience, is given by

$$\mathcal{L}^{(1,\Psi)}(\Psi, \Lambda) = \mathcal{J}'(\Psi) - \left( \sum_{k=1}^N \lambda_{i,k} \psi_k \right)_{i=1}^N = 0 \in (V')^N. \quad (2.8)$$

Testing this equation with  $\psi_j, j = 1, \dots, N$ , verifies the Lagrange multipliers indeed agree with the  $\Lambda$  defined above, so that (2.5) and (2.8) are equivalent. Note also that the remaining optimality conditions,

$$\frac{\partial \mathcal{L}}{\partial \lambda_{i,j}} = \frac{1}{2} (\langle \langle \psi_i, \psi_j \rangle_{L_2} - \delta_{i,j} \rangle) = 0,$$

of the Lagrange formalism are now incorporated in the framework of the Stiefel manifold. From the representation (2.6), it follows that the Hessian  $\mathcal{L}^{(2,\Psi)}(\Psi, \Lambda)$  of the Lagrangian (2.8), taken at the minimum  $\Psi$  and with the derivatives taken with respect to  $\Psi$ , is given by

$$\mathcal{L}^{(2,\Psi)}(\Psi, \Lambda)\Phi = \mathcal{J}''(\Psi)\Phi - \Phi\Lambda.$$

As a necessary second order condition for a minimum,  $\mathcal{L}(\Psi, \Lambda)^{(2, \Psi)}$  has to be positive semidefinite on  $\mathcal{T}_{[\Psi]}\mathcal{G}$ . For our convergence analysis, we will have to impose the stronger condition on  $\mathcal{L}^{(2, \Psi)}(\Psi, \Lambda)$  being elliptic on the tangent space, i.e.

$$\langle \langle \mathcal{L}^{(2, \Psi)}(\Psi, \Lambda) \delta\Phi, \delta\Phi \rangle \rangle \geq \gamma \|\delta\Phi\|_{V^N}^2, \quad \text{for all } \delta\Phi \in \mathcal{T}_{[\Psi]}\mathcal{G}. \quad (2.9)$$

It is an unsolved problem if this condition holds in general for the minimization problems of the kind (1.1) or if it depends on the functional under consideration; in particular, it is not clear whether it holds for the functionals of Hartree-Fock and density functional theory. In the case of Hartree-Fock, it suffices to demand that  $\mathcal{L}^{(2, \Psi)}(\Psi, \Lambda) > 0$  on  $\mathcal{T}_{[\Psi]}\mathcal{G}$  because this already implies  $\mathcal{L}^{(2, \Psi)}(\Psi, \Lambda)$  is bounded away from zero, cf [35]. For the simplified problem, we will show in Lemma 4 that the assumption holds for symmetric operators  $A$  fulfilling a certain gap condition.

### 3 Minimization tasks in electronic structure calculations

We will now particularize the results of the last section to the functionals common in electronic structure calculation. As the following section will show, the applications of interest in electronic structure calculations deal with the minimization of functionals  $\mathcal{J}$  for which the gradient can be written as  $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi \Phi$ , where  $\mathcal{A}_\Phi : V \rightarrow V'$  (and  $\mathcal{A}_\Phi$  its extension to  $V^N$  by (2.1)). We conjecture that if the functional  $\mathcal{J}$  only depends on the electronic density, that is, if condition (1.2) holds, this form of  $\mathcal{J}(\Phi)$  is valid in general, i.e. for each  $\Phi \in \mathcal{G}$ , there is an operator  $\mathcal{A}_\Phi$  so that  $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi \Phi$ . Nevertheless, we decided to formulate the algorithms (except algorithm 3) for  $\mathcal{J}'(\Phi)$  rather than for  $\mathcal{A}_\Phi$  to emphasize the minimization viewpoint we pursue in this work and to display that the concrete structure of the Fock or Kohn-Sham operators does not enter anywhere in the proof of convergence given in section 5.

In this section, we will remind the reader of some basic facts about Hartree-Fock and Kohn-Sham theory, where our emphasis will be on the ansatzes leading to the problem of minimizing a nonlinear functional (1.1). Also, we will review the concrete form the operator  $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi \Phi$  in (1.4) has in these applications. For a more detailed introduction to electronic structure calculations, we refer the reader to the standard literature [9, 12, 26, 43]. At the end of this section, we will investigate the simplified problem (1.5) and its connection to eigenvalue computations.

### 3.1 Hartree-Fock and Kohn-Sham energy functionals in quantum chemistry

The commonly accepted model to describe atoms and molecules is by means of the Schrödinger equation, which is in good agreement with experiments as long as the energies remain on a level at which relativistic effects can be neglected. We are mainly interested in the stationary ground state of quantum mechanical systems, given by the eigenfunction belonging to the lowest eigenvalue of the Hamiltonian  $H$  of the system. In the Born-Oppenheimer approximation the Hamiltonian of the (time-independent) *electronic Schrödinger equation*  $H\Psi = E\Psi$  is given by

$$H := -\frac{1}{2} \sum_{i=1}^{N^*} \Delta_i - \sum_{i=1}^{N^*} \sum_{\nu=1}^M \frac{Z_\nu}{\|x_i - R_\nu\|} + \frac{1}{2} \sum_{i,j=1, i \neq j}^{N^*} \frac{1}{\|x_i - x_j\|}.$$

Here,  $N^*$  denotes the number of electrons,  $M$  the number of the nuclei, and  $Z_\nu$ ,  $R_\nu$  the charge respectively the coordinates of the nuclei, which are the only fixed input parameters of the system. Note that we use atomic units, so that no physical constants appear in the Schrödinger equation. We also neglect the interaction energy between the nuclei, since for a given constellation  $(R_1, \dots, R_M)$  of the  $M$  nuclei this only adds a constant to the energy eigenvalues. Due to the Pauli principle for fermions, the wave function is required to be antisymmetric with respect to permutation of particle coordinates. It is easy to see that every such antisymmetric solution can be represented by a convergent sum of Slater determinants of the form

$$\psi_{SL}^\Phi(x_1, s_1, \dots, x_{N^*}, s_{N^*}) := \frac{1}{\sqrt{N^*!}} \det(\varphi_i(x_j, s_j)), \quad x_i \in \mathbb{R}^3, \quad s_i = \pm \frac{1}{2}$$

where  $\Phi = (\varphi_i)_{i=1}^{N^*} \in H^1(\mathbb{R}^3 \times \{\pm \frac{1}{2}\})^{N^*}$  and  $\langle \varphi_i, \varphi_j \rangle = \delta_{i,j}$ . In *Hartree-Fock (HF) theory*, one approximates the ground state of the system by minimizing the Hartree-Fock energy functional  $\Phi \mapsto \mathcal{J}_{HF}(\Phi) := \langle H\psi_{SL}^\Phi, \psi_{SL}^\Phi \rangle$  over the set of all wave functions consisting of *one single* Slater determinant  $\psi_{SL}^\Phi(x_1, s_1, \dots, x_{N^*}, s_{N^*})$ . Additional simplification is made by the *Closed Shell Restricted Hartree-Fock model* (RHF), given in a spin-free formulation for  $N = N^*/2$  pairs of electrons, so that  $\Phi = (\varphi_i)_{i=1}^N \in H^1(\mathbb{R}^3)^N =: V^N$ . Abbreviating  $V(x) := -\sum_{\nu=1}^M \frac{Z_\nu}{\|x - R_\nu\|}$ , the corresponding functional reads

$$\begin{aligned} \mathcal{J}_{HF}(\Phi) := & \sum_{i=1}^N \int_{\mathbb{R}^3} \left( \frac{1}{2} |\nabla \varphi_i(x)|^2 + V(x) |\varphi_i(x)|^2 + \frac{1}{2} \sum_{j=1}^N \int_{\mathbb{R}^3} \frac{|\varphi_j(y)|^2}{\|x - y\|} dy |\varphi_i(x)|^2 \right. \\ & \left. - \frac{1}{2} \sum_{j=1}^N \int_{\mathbb{R}^3} \frac{\varphi_i(x) \varphi_j(x) \varphi_j(y) \varphi_i(y)}{\|x - y\|} dy \right) dx. \end{aligned} \quad (3.1)$$

A minimizer of  $\mathcal{J}_{HF}$  is named Hartree-Fock ground state. Its existence has been proven in the case that  $\sum_{\mu=1}^K Z_\mu > N - 1$  ([29], [30]).

The energy functional of the *Kohn-Sham (KS) model* can be derived from the Hartree-Fock energy functional by two modifications: First of all, as a consequence of the Hohenberg-Kohn theorem (cf. [27]), it is formulated in terms of the electron density  $n(x) = \sum_{i=1}^N |\varphi_i(x)|^2$  rather than in terms of the single particle functions; secondly, it replaces the nonlocal and therefore computationally costly exchange term in the Hartree-Fock functional (i.e. the fourth term in (3.1)) by an additional (a priori unknown) exchange correlation energy term  $E_{xc}(n)$  also depending only on the electron density. The resulting energy functional reads

$$\mathcal{J}_{KS}(\Phi) = \frac{1}{2} \sum_{i=1}^N \int_{\mathbb{R}^3} |\nabla \varphi_i(x)|^2 dx + \int_{\mathbb{R}^3} n(x)V(x) + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{n(x)n(y)}{\|x-y\|} dx dy + E_{xc}(n).$$

Determining the ground state energy of Kohn-Sham theory then consists in a minimization of  $\mathcal{J}_{KS}$  over all  $\Phi = (\varphi_1, \dots, \varphi_N) \in V^N$  with  $\langle \varphi_i, \varphi_j \rangle = \delta_{i,j}$ . Since the exchange correlation energy  $E_{xc}$  is not known explicitly, further approximations are necessary. The most simple approximation for  $E_{xc}$  is the local density approximation (LDA, cf. [13]) defined as  $E_{xc}^{LDA}(n) = \int_{\mathbb{R}^3} n(x)\epsilon_{xc}^{LDA}(n(x)) dx$ , where  $\epsilon_{xc}^{LDA}$  denotes the exchange-correlation energy of a particle in an electron gas with density  $n$ . If we split this expression in an exchange and a correlation part, we get

$$E_{xc}^{LDA}(n) = E_x^{LDA}(n) + E_c^{LDA}(n) = \int_{\mathbb{R}^3} n(x)\epsilon_x^{LDA}(n(x)) dx + \int_{\mathbb{R}^3} n(x)\epsilon_c^{LDA}(n(x)) dx,$$

where in the exchange part,  $\epsilon_x^{LDA}(n) = -C_D n^{\frac{1}{3}}$  and  $C_D := \frac{3}{4}(\frac{3}{\pi})^{1/3}$  is the Dirac constant. For the correlation part  $E_c^{LDA}(n)$ , the expression  $\epsilon_c^{LDA}$  is analytically unknown, but can be calibrated e.g. by Monte-Carlo methods. We note that a combination of both HF and density functional models, namely the hybrid B3LYP, is experienced to provide the best results in benchmark computations.

### 3.2 Canonical Hartree-Fock and Kohn-Sham equations

For the HF and KS functionals, we can compute the derivative of  $\mathcal{J}$  and the Lagrange multipliers at a minimizer explicitly.

**Proposition 2.** *For the functional  $\mathcal{J}_{HF}$  of Hartree-Fock,  $\mathcal{J}'_{HF}(\Phi) = \mathcal{A}_\Phi \Phi \in (V')^N$ , where  $\mathcal{A}_\Phi = F_\Phi^{HF} : H^1(\mathbb{R}^3) \rightarrow H^{-1}(\mathbb{R}^3)$  is the so-called Fock operator and  $\mathcal{A}_\Phi$  is defined by  $\mathcal{A}_\Phi$  through (2.1); using the notation of the density matrix*

$$\begin{aligned} \rho_\Phi(x, y) &:= N \int_{\mathbb{R}^{3(N-1)}} \psi_{SL}^\Phi(x, x_2, \dots, x_N) \psi_{SL}^\Phi(y, x_2, \dots, x_N) dx_2 \cdots dx_N \\ &= \sum_{i=1}^N \varphi_i(x) \varphi_i(y) \end{aligned}$$

and the electron density  $n_\Phi(x) := \rho_\Phi(x, x)$  already introduced above. It is given by

$$F_\Phi^{HF} \varphi(x) := -\frac{1}{2} \Delta \varphi(x) + V(x) \varphi(x) + \int_{\mathbb{R}^3} \frac{n_\Phi(y)}{\|x-y\|} dy \varphi(x) - \int_{\mathbb{R}^3} \frac{\rho_\Phi(x, y) \varphi(y)}{\|x-y\|} dy.$$

For the gradient of the Kohn-Sham functional  $\mathcal{J}_{KS}$ , there holds the following: Assuming that  $E_{xc}$  in  $\mathcal{J}_{KS}$  is differentiable and denoting by  $v_{xc}$  the derivation of  $E_{xc}$  with respect to the density  $n$ , we have  $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi \Phi \in (V')^N$ , with  $\mathcal{A}_\Phi = F_n^{KS}$  the Kohn-Sham Hamiltonian, given by

$$F_n^{KS} \varphi_i := -\frac{1}{2} \Delta \varphi_i + V(x) \varphi_i + \left( n \star \frac{1}{\|\cdot\|} \right) \varphi_i + v_{xc}(n) \varphi_i.$$

In both cases, the Lagrange multiplier  $\Lambda$  of (2.8) at a minimizer  $\Psi = (\psi_1, \dots, \psi_N)$  is given by

$$\lambda_{i,j} = \langle A_\Psi \psi_i, \psi_j \rangle. \quad (3.2)$$

There exists a unitary transformation  $\mathbf{U} \in O(N)$  amongst the functions  $\psi_i$ ,  $i = 1, \dots, N$  such that the Lagrange multiplier is diagonal for  $\Psi \mathbf{U} = (\tilde{\psi}_1, \dots, \tilde{\psi}_N)$ ,

$$\lambda_{i,j} := \langle A \tilde{\psi}_i, \tilde{\psi}_j \rangle = \lambda_i \delta_{i,j}.$$

so that the ground state of the HS resp. KS functional (i.e. minimizer of  $\mathcal{J}$ ) satisfies the nonlinear Hartree-Fock resp. Kohn-Sham eigenvalue equations

$$F_\Psi^{HF} \psi_i = \lambda_i \psi_i, \quad \text{resp.} \quad F_n^{KS} \psi_i = \lambda_i \psi_i, \quad \lambda_i \in \mathbb{R}, \quad i = 1, \dots, N, \quad (3.3)$$

for some  $\lambda_1, \dots, \lambda_N \in \mathbb{R}$  and a corresponding set of orthonormalized functions  $\Psi = (\psi_i)_{i=1}^N$  up to a unitary transformation  $\mathbf{U}$ .

The converse result, i.e. if for a collection  $\Phi = (\varphi_1, \dots, \varphi_N)$  belonging to the  $N$  lowest eigenvalues of the Fock operator in (3.3), the corresponding Slater determinant actually gives the Hartree-Fock energy by  $\mathcal{J}(\Phi) = \langle H \psi_{SL}^\Phi, \psi_{SL}^\Phi \rangle$ , is not known yet.

### 3.3 Simplified problem

The practical significance of the simplified problem (1.5) is given by the following result, which shows that for symmetric  $A$ , the minimization of  $\mathcal{J}_A$  is indeed equivalent to finding an orthonormal basis  $\{\psi_i : 1 \leq i \leq N\}$  spanning the invariant subspace of  $A$  given by the first eigenfunctions of  $A$ .

**Proposition 3.** *Let  $A$  in the simplified problem (1.5) a bounded symmetric operator. The gradient of the functional  $\mathcal{J}_A$  is then given by  $\mathcal{J}'(\Phi) = \mathcal{A}\Phi \in (V')^N$ . Therefore,  $\Psi$  is a stationary point of  $\mathcal{L}$  if and only if there exists an orthogonal transformation  $\mathbf{U}$  such that  $\Psi\mathbf{U} = (\tilde{\psi}_1, \dots, \tilde{\psi}_N) \in V^N$  consists of  $N$  pairwise orthonormal eigenfunctions of  $A$ , i.e.  $A\psi_k = \lambda_k\psi_k$  for  $k = 1, \dots, N$ ; in this case, there holds  $\mathcal{J}(\Psi) = \sum_{k=1}^N \lambda_k$ . The minimum of  $\mathcal{J}$  is attained if and only if the corresponding eigenvalues  $\lambda_k$ ,  $k = 1, \dots, N$  are the  $N$  lowest eigenvalues. This minimum is unique up to orthogonal transformations if there is a gap  $\lambda_{N+1} - \lambda_N > 0$ , so that in this case, the minimizers  $\Psi = \operatorname{argmin} \mathcal{J}$  are exactly the bases of the unique invariant subspace spanned by the eigenvectors according to the  $N$  lowest eigenvalues.*

### 3.4 Comparison of direct minimization and self consistent iteration

Self consistent iteration consists of fixing the Fock operator  $F^{(n)} = F_{\Phi^{(n)}}$  for each iterate  $\Phi^{(n)}$ ; the simplified problem is then solved in an inner iteration loop for  $A = F^{(n)}$ ; the solution  $\Phi$  defines the next iterate  $\Phi^{(n+1)}$  of the outer iteration, by which the Fock operator is then updated to form  $F^{(n+1)}$ , defining the simplified problem for the next iteration step. For the solution of the inner problems with a fixed Fock operator, Proposition 3 from the last section applies and the algorithms presented in the next section can be used. Self consistent iteration is faced with convergence problems though, which can be remedied by advanced techniques: With an appropriate choice of the update, the ODA-optimal damping algorithm [10], convergence can be guaranteed.

Direct minimization corresponds to the treatment of the nonlinear problem (1.1) for the Hartree-Fock or Kohn-Sham functional with the gradient algorithm 1 from the next section. Direct minimization thus differs from the self consistent iteration only in that the Fock operator is updated after each inner iteration step. Therefore, direct minimization is preferable if the update of the Fock operator is sufficiently cheap. This is mostly the case for Gaussians but not for the plane wave or wavelet basis or finite differences.

## 4 Algorithms for minimization

In this section we will introduce three related algorithms to tackle the minimization problem (1.1) in a rather general form. Their convergence properties will be analysed in the next section.

## 4.1 Gradient and projected gradient algorithm

We will consider a gradient algorithm for the constrained minimization problem; the motivation for this is given by the following related formulation (cf. [32] for this concept): With an initial guess  $\Phi^{(0)} \in \mathcal{V}$ ,  $[\Phi^{(0)}] \in \mathcal{G}$ , the gradient flow on  $\mathcal{V}$ , resp.  $\mathcal{G}$  is given by the differential

$$\left\langle \left\langle \frac{d\Phi(t)}{dt} - \mathcal{J}'(\Phi(t)), \delta\Phi \right\rangle \right\rangle = 0 \quad \forall \delta\Phi \in \mathcal{T}_{[\Phi(t)]}\mathcal{G}. \quad (4.1)$$

Using the fact that  $\mathcal{I} - \mathcal{D}_{[\Phi]}$  is projecting onto the tangent space  $\mathcal{T}_{[\Phi]}\mathcal{G}$ , this algebraic differential initial value problem can be rewritten by an ordinary initial value problem for the gradient flow on  $\mathcal{V}$ ,

$$\frac{d}{dt}\Phi(t) = (\mathcal{I} - \mathcal{D}_{[\Phi(t)]})\mathcal{J}'([\Phi(t)]), \quad \Phi(0) = \Phi^{(0)}, \quad (4.2)$$

or, equivalently,

$$\frac{d}{dt}\Phi(t) = [\mathcal{J}', \mathcal{D}_{[\Phi(t)]}](\Phi(t)), \quad \Phi(0) = \Phi^{(0)}, \quad (4.3)$$

where the bracket  $[\cdot, \cdot]$  denotes the usual commutator. Denoting by  $(\mathcal{J}'(\Phi(t)))_i$  the  $i$ -th component of the gradient  $\mathcal{J}'(\Phi(t))$  and letting  $\Lambda(t) := \left( \langle (\mathcal{J}'(\Phi(t)))_i, \varphi_j(t) \rangle \right)_{i,j=1}^N$ , we obtain the identification

$$\mathcal{J}'(\Phi(t)) - \Phi(t)\Lambda(t) = [\mathcal{J}', \mathcal{D}_{[\Phi(t)]}](\Phi(t)), \quad (4.4)$$

which we will make use of later.

There holds  $\frac{d\Phi(t)}{dt} \rightarrow 0$  for  $t \rightarrow \infty$ , so we are looking for the fixed point of this flow  $\Psi = \lim_{t \rightarrow \infty} \Phi(t)$  rather than its trajectory. Equation (4.3) suggests the projected gradient type algorithms presented below. In algorithm 1, corresponding to an Euler procedure for the differential equation (4.1), the gradient at a certain point  $\Phi(t)$  is kept fixed (and being preconditioned) for non-differential stepsize, so that the manifold is left in each iteration step. Therefore, a projection on the admitted set is performed in each iteration step.

Note also that the role of the preconditioners  $\mathcal{B}_n^{-1}$  is crucial, see the remarks following algorithm 1.

---

### **Algorithm 1: Projected Gradient Descent**

---

**Require:** Initial iterate  $\Phi^{(0)} \in \mathcal{V}$ ;

evaluation of  $\mathcal{J}'(\Phi^{(n)})$  and of preconditioner(s)  $\mathcal{B}_n^{-1}$  (see comments below)

**Iteration:**

for  $n = 0, 1, \dots$  do

(1) Update  $\Lambda^{(n)} := \langle \mathcal{J}'(\Phi^{(n)}), \Phi^{(n)} \rangle \in \mathbb{R}^{N \times N}$ ,

(2) Let  $\hat{\Phi}^{(n+1)} := \Phi^{(n)} - \mathcal{B}_n^{-1}(\mathcal{J}'(\Phi^{(n)}) - \Phi^{(n)}\Lambda^{(n)})$ ,

( =  $\Phi^{(n)} - \mathcal{B}_n^{-1}(\mathcal{A}_{\Phi^{(n)}}\Phi^{(n)} - \Phi^{(n)}\Lambda^{(n)})$  for the case that  $\mathcal{J}'(\Phi) = \mathcal{A}_{\Phi}\Phi$ .)

(3) Let  $\Phi^{(n+1)} = P\hat{\Phi}^{(n+1)}$  by projection  $P$  onto  $\mathcal{V}$  resp.  $\mathcal{G}$

endfor

---

Some remarks about this algorithm are in order. First of all, note that if Algorithm 1 is applied to the ansatzes in electronic structure calculation as portrayed in section 3, the gradient  $\mathcal{J}'(\Phi)$  is given by  $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi \Phi$  with  $\mathcal{A}_\Phi$  the Fock- or Kohn-Sham operator or a fixed operator  $\mathcal{A}_\Phi = A$  for the simplified problem. Therefore,  $(\mathcal{J}'(\Phi^{(n)} - \Phi^{(n)}\Lambda^{(n)})_i = A_{\Phi^{(n)}}\phi_i^{(n)} - \sum_{j=1}^N \langle A_{\Phi^{(n)}}\phi_i^{(n)}, \phi_j^{(n)} \rangle \phi_j^{(n)}$  is the usual “subspace residual” of the iterate  $\Phi^{(n)}$ , which is a crucial fact for capping the complexity of the algorithm in section 6.

Next, let us specify the role of the preconditioner  $\mathcal{B}_n^{-1}$  used in each step. This preconditioner is induced (according to (2.1)) by an elliptic symmetric operator  $B_n : V \rightarrow V'$ , which we require to be equivalent to the norm on  $H^1$  in the sense that

$$\langle B_n \varphi, \varphi \rangle_{L_2} \sim \|\varphi\|_{H^1}^2 \quad \forall \varphi \in V = H^1(\mathbb{R}^3). \quad (4.5)$$

For example, one can use approximations of the shifted Laplacian,  $B \approx \alpha(-\frac{1}{2}\Delta + C)$ , as is done in the BigDFT project. This is also a suitable choice when dealing with plane wave ansatz functions using advantages of FFT, or a multi-level preconditioner if one has finite differences, finite elements or multi-scale functions like wavelets [25, 7, 16, 3].

For the simplified problem, the choice  $B^{-1} = \alpha A^{-1}$  corresponds to a variant of simultaneous inverse iteration. The choice

$$B|_{V_0^\perp := \{v | \langle v, \varphi_i^{(n)} \rangle = 0 \forall i=1, \dots, N\}} = \alpha(A - \lambda_j^{(n)} I)|_{V_0^\perp := \{v | \langle v, \varphi_i^{(n)} \rangle = 0 \forall i=1, \dots, N\}}$$

corresponds to a simultaneous Jacobi-Davidson iteration.

To guarantee convergence of the algorithm, the preconditioner  $B$  chosen according to the guidelines above also has to be properly scaled by a factor  $\alpha > 0$ , cf. Lemma 6. The optimal choice of  $\alpha$  is provided by minimizing the corresponding functional over span  $\{\Phi^{(n)}, \widehat{\Phi}^{(n+1)}\}$  (a line search over this space), which can be done for the simplified problem without much additional effort. For the Kohn-Sham energy functional, it will become prohibitively expensive. However, line search and subspace acceleration like DIIS [37] will improve the convergence speed. Note that in this context, one might as well use different step sizes for every entry, i.e.  $\mathcal{B}\Phi = (\alpha_1 B\varphi_1, \dots, \alpha_N B\varphi_N)$ .

Next, let us make a remark concerning the projection onto  $\mathcal{G}$ . It only has to satisfy span  $\{\varphi_i^{(n+1)} : 1 \leq i \leq N\} = \text{span} \{\widehat{\varphi}_i^{(n+1)} : 1 \leq i \leq N\}$ . For this purpose any orthogonalization of  $\{\widehat{\varphi}_i^{(n+1)} : 1 \leq i \leq N\}$  is admissible. For example, three favorable possibilities which up to unitary transformations yield the same result are

- Gram-Schmidt orthogonalization,
- Diagonalization of the Gram matrix  $\mathbf{G} = (\langle \widehat{\varphi}_i^{(n+1)}, \widehat{\varphi}_j^{(n+1)} \rangle)_{i,j=1}^N$  by Cholesky factorization,

- (For the problems of section 3, i.e. where  $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi \Phi$ .)  
 Diagonalisation of the matrix  $\mathbf{A}_{\Phi^{(n+1)}} := (\langle \mathcal{A}_{\Phi^{(n)}} \hat{\varphi}_i^{(n+1)}, \hat{\varphi}_j^{(n+1)} \rangle)_{i,j=1}^N$  by solving an  $N \times N$  eigenvalue problem.

Parallel to the above algorithm, we consider the following variant in which the descent direction is projected onto the tangent space  $\mathcal{T}_{[\Phi^{(n)}]} \mathcal{G}$  in every iteration step. It will play an important theoretical role considering convergence of the local exponential parametrization, i.e. algorithm 3.

---

**Algorithm 2: Modified Projected Gradient Descent**

---

*Require:* see Algorithm 1

**Iteration:**

for  $n = 0, 1, \dots$  do

- (1) Update  $\Lambda^{(n)} := \langle \mathcal{J}'(\Phi^{(n)}), \Phi^{(n)} \rangle \in \mathbb{R}^{N \times N}$ ,
- (2) Let  $\hat{\Phi}^{(n+1)} := \Phi^{(n)} - (\mathcal{I} - \mathcal{D}_{\Phi^{(n)}}) \mathcal{B}_n^{-1} (\mathcal{J}'(\Phi^{(n)}) - \Phi^{(n)} \Lambda^{(n)})$ ,  
 ( $= \Phi^{(n)} - \mathcal{B}_n^{-1} (\mathcal{A}_{\Phi^{(n)}} \Phi^{(n)} - \Phi^{(n)} \Lambda^{(n)})$  for the case that  $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi \Phi$ .)
- (3) Let  $\Phi^{(n+1)} = P \hat{\Phi}^{(n+1)}$  by projection  $P$  onto  $\mathcal{V}$  resp.  $\mathcal{G}$ ,

endfor

---

Note again that the algorithms are given in a general form, where the preconditioner (or the corresponding parameter  $\alpha$ , e.g. obtained by a kind of line search) may be chosen in each iteration step. In our analysis, we will consider a fixed preconditioner  $\mathcal{B}_n = \mathcal{B}$  in every iteration step, for which we will show linear convergence without further line search invoked. Thus, our analysis is in a way a worst case analysis for the algorithms under consideration. See also section 6 for improvements on the speed of convergence.

## 4.2 Exponential parametrization

Instead of projecting the iterate  $\Phi^{(n)}$  onto the Grassmann manifold  $\mathcal{G}$  in every iteration step, we will now develop an algorithm in which the iterates remain on the manifold without further projection. This will be achieved by following geodesic paths on the manifold instead of straight lines in Euclidean space, which has the advantage that during our calculations we do not leave the constraining set at any time so that no orthonormalization process is required. To apply the result of proposition 4, we will for this algorithm limit our treatment to the case where  $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi \Phi$  is given by a linear operator (see the discussion after algorithm 1).

Recall that a geodesic is a curve  $c$  on a manifold with vanishing second covariant derivative, i.e.

$$\frac{\nabla}{dt} \dot{c}(t) := \pi_{c(t)} \ddot{c}(t) = 0 \quad \text{for all } t, \quad (4.6)$$

where  $\pi_{c(t)}$  denotes the projection onto the tangent space at the point  $c(t)$ .

**Proposition 4.** For any operator  $X : V \rightarrow V$  for which  $\mathcal{X}\Phi \in \mathcal{T}_{[\Phi]}\mathcal{G}$  (where as always,  $\mathcal{X}$  is defined by  $X$  by (2.1)), the antisymmetric operator

$$\hat{X} = (I - D_{\Phi})XD_{\Phi} - D_{\Phi}X^{\dagger}(I - D_{\Phi}), \quad (4.7)$$

satisfies  $\hat{\mathcal{X}}\Phi = \mathcal{X}\Phi$ , and  $c(t) := \exp(t\hat{\mathcal{X}})\Phi$  is a geodesic in  $\mathcal{G}$  emanating from point  $\Phi$  with direction  $\dot{c}(0) = \mathcal{X}\Phi$ .

*Proof.* The proof is straightforward; application of the projection equation yields  $(\frac{\nabla}{dt}\dot{c}(t)) = (\mathcal{I} - \mathcal{D}_{c(t)})\ddot{c}(t) = \mathbf{0}$ .  $\square$

If we now let, for any iterate  $\Phi^{(n)}$ ,

$$X^{(n)} = (I - D_{\Phi^{(n)}})B^{-1}(I - D_{\Phi^{(n)}})A_{\Phi^{(n)}}, \quad (4.8)$$

the curve

$$c(t) := \exp(-t\hat{\mathcal{X}}^{(n)})\Phi^{(n)}$$

with  $\hat{\mathcal{X}}^{(n)}$  from (4.7) is by the previous Lemma a geodesic in  $\mathcal{G}$  with direction

$$\dot{c}(0) = -(I - D_{\Phi^{(n)}})B^{-1}(I - D_{\Phi^{(n)}})A_{\Phi^{(n)}}\Phi^{(n)}$$

which equals the (preconditioned) descent direction of the projected gradient descent algorithm of the preceding section. If we now choose the next iterate as a point on this geodesic, we get the following algorithm:

---

**Algorithm 3: Preconditioned exponential parametrization**

---

**Require:** see Algorithm 1

**Iteration:**

for  $n = 0, 1, \dots$  do

▷ Follow a geodesic path on the Grassmann manifold with stepsize  $\alpha$ ,

$$\Phi^{(n+1)} := \exp(-\alpha\hat{\mathcal{X}}^{(n)})\Phi^{(n)} \text{ (with } \mathcal{X} \text{ from (4.8) and } \hat{\mathcal{X}} \text{ defined by } \mathcal{X} \text{ via (4.7))}$$

endfor

---

Note that a similar algorithm, Conjugate Gradient on the Grassman Manifold, has already been introduced in [4], page 327. That paper also included numerical tests for a model system. The algorithm was also tested for electronic structure applications very different from those of the BigDFT program in [38]. A similar approach using the density matrix representation for electronic structure problems was also proposed in [42], where the authors move along the geodesics in a gradient resp. Newton method direction without preconditioning.

Like in this work, the stepsize  $\alpha$  may be calculated in each iteration step using line search algorithms like backtracking linesearch or quadratic approximations to the energy term [36]. These often time consuming line searches may be omitted though if we choose a suitable preconditioner  $B = B_n$  and set the stepsize  $\alpha = 1$  once and for all.

The efficiency of this algorithm strongly depends on the computation of matrix exponentials needed to follow geodesic paths on the Grassmann manifold. A variety of methods can be found in [33], see also [41] for an analysis of selected methods. For some of these algorithms, there exist powerful implementations like the software package Expokit [44], which contain both Matlab and Fortran code thus supplying a convenient tool for numerical experiments.

## 5 Convergence results

### 5.1 Assumptions, error measures and main result

In this section, we will show linear convergence of the algorithms of the last section under the ellipticity assumption 1 given below. Additional results we give include the equivalence of the error of  $\Phi$ , measured in a norm on  $V$ , and the error of the gradient residual  $(\mathcal{I} - \mathcal{D})\mathcal{J}'(\Phi)$ , and quadratic reduction of the energy error  $\mathcal{J}(\Phi^{(n)}) - \mathcal{J}(\Psi)$ .

Recall that in our framework introduced in section 2, we kept the freedom of choice to either use  $V := H^1 = H^1(\mathbb{R}^3)$ , equipped with an inner product equivalent to the  $H^1$  inner product  $\langle \cdot, \cdot \rangle_{H^1}$ , for analysing the original equations, or to use  $V = V_h \subset H^1$  as a finite dimensional subspace for a corresponding Galerkin discretization of these equations. In practice, our iteration scheme is only applied to the discretized equations. However, the convergence estimates obtained will be uniform with respect to the discretization parameters. The main ingredient our analysis is based on is the following condition imposed on the functional  $\mathcal{J}$ , cf. section 2.3:

**Assumption 1.** *Let  $\Psi$  a minimizer of (1.1). The Hessian  $\mathcal{L}^{(2,\Psi)}(\Psi, \Lambda) : V^N \rightarrow (V^N)^N$  of the Lagrangian  $\mathcal{L}(\Psi, \Lambda)$  (given by (2.8)), where the derivatives are taken with respect to  $\Psi$ , is assumed to be  $V^N$ -elliptic on the tangent space, i.e. there is  $\gamma > 0$  so that*

$$\langle \langle \mathcal{L}^{(2,\Psi)}(\Psi, \Lambda)\delta\Phi, \delta\Phi \rangle \rangle \geq \gamma \|\delta\Phi\|_{V^N}^2, \quad \text{for all } \delta\Phi \in \mathcal{T}_{[\Psi]}\mathcal{G}. \quad (5.1)$$

Note again that for Hartree-Fock calculations, verification of  $\mathcal{L}^{(2,\Psi)}(\Psi, \Lambda) > 0$  on  $\mathcal{T}_{[\Psi]}\mathcal{G}$  already implies  $\mathcal{L}^{(2,\Psi)}(\Psi, \Lambda)$ , cf [35].

From section 2.2, we recall that  $\mathcal{L}^{(2,\Psi)}(\Psi, \Lambda)\Phi = \mathcal{J}''(\Psi)\Phi - \Phi\Lambda$ , so that (5.1) is verified if and only if

$$\langle \langle \mathcal{J}''(\Psi)\delta\Phi - \delta\Phi\Lambda, \delta\Phi \rangle \rangle \geq \gamma \|\delta\Phi\|_{V^N}^2, \quad \text{for all } \delta\Phi \in \mathcal{T}_{[\Psi]}\mathcal{G} \quad (5.2)$$

holds, where  $\Lambda = (\langle (\mathcal{J}'(\Psi))_j, \psi_i \rangle)_{i,j=1}^N$  as above. From the present state of Hartree-Fock theory, it is not possible to decide whether this condition is true in general; the same applies to DFT theory. For the simplified problem, the condition holds if the operator  $A$  fulfils the conditions of the following lemma.

**Lemma 4.** *Let  $A : V \rightarrow V'$ ,  $\psi \mapsto A\psi$  a bounded symmetric operator, such that  $A$  has  $N$  lowest eigenvalues  $\lambda_1 \leq \dots \leq \lambda_N$  satisfying the gap condition*

$$\lambda_N < \inf\{\lambda \mid \lambda \in \sigma(A) \setminus \{\lambda_1, \dots, \lambda_N\}\}. \quad (5.3)$$

*Then assumption 1 holds for the simplified problem (1.5).*

*Proof.* We estimate the two terms of (5.2) separately. Let us denote  $\lambda = \inf\{\lambda \mid \lambda \in \sigma(A) \setminus \{\lambda_1, \dots, \lambda_N\}\}$ . To the first term, the Courant-Fisher theorem ([39]) applies componentwise to give the estimate  $\langle \langle \mathcal{A}\delta\Phi, \delta\Phi \rangle \rangle \geq \lambda \|\delta\Phi\|_{V^N}^2$ . For the second, choosing  $\mathbf{U} = (u_{i,j})_{i,j=1}^N \in O(N)$  so that  $\mathbf{U}^T \Lambda \mathbf{U} = \text{diag}(\lambda_i)_{i=1}^N$ , where  $\lambda_i$  are the lowest  $N$  eigenvalues of  $A$ , gives

$$\begin{aligned} \langle \langle \delta\Phi \Lambda, \delta\Phi \rangle \rangle &= \langle \langle \delta\Phi (\mathbf{U} \mathbf{U}^T \Lambda \mathbf{U} \mathbf{U}^T), \delta\Phi \rangle \rangle := \sum_{i=1}^N \langle \sum_{j=1}^N u_{j,i} \lambda_j \delta\varphi_j, \sum_{k=1}^N u_{k,i} \delta\varphi_k \rangle \\ &= \sum_{j,k=1}^N \lambda_j \delta_{j,k} \langle \delta\varphi_j, \delta\varphi_k \rangle \leq \lambda_N \|\delta\Phi\|_{V^N}^2. \end{aligned}$$

so that  $\mathcal{L}^{(2)}(\Psi, \Lambda)$  is elliptic on  $\mathcal{T}_{[\Psi]}\mathcal{G}$  by the gap condition (5.3).  $\square$

To formulate our main convergence result, we now introduce a norm  $\|\cdot\|_{V^N}$  on the space  $V^N$ , which will be equivalent to the  $(H^1)^N$ -norm but more convenient for our proof of convergence. We will then state our convergence result in terms of these error measures.

**Lemma 5.** *Let  $B : V \rightarrow V'$  the preconditioning mapping introduced in section 4, so that in particular,  $B$  is symmetric and the spectral equivalence*

$$\vartheta \|x\|_{H^1}^2 \leq \langle Bx, x \rangle \leq \Theta \|x\|_{H^1}^2$$

*holds for some  $0 < \vartheta \leq \Theta$  and all  $x \in V$ . Let us consider the mapping*

$$\hat{B}^{-1} : V' \rightarrow V, \quad \hat{B}^{-1} := (I - D)B^{-1}(I - D) + D, \quad (5.4)$$

*where  $D = D_\Psi$  projects onto the sought subspace. Then the inverse  $\hat{B}$  satisfies  $\langle \hat{B}\varphi, \psi \rangle = \langle \varphi, \hat{B}\psi \rangle$  for all  $\varphi, \psi \in V$ , and for the induced  $\hat{B}$ -norm  $\|\cdot\|_{\hat{B}}$  on  $V$  there holds*

$$\langle \hat{B}\varphi, \varphi \rangle \sim \|\varphi\|_{H^1}^2.$$

Using the notation (2.1), a norm on  $V^N$  is now induced by the  $\|\cdot\|_{\widehat{\mathcal{B}}}$ -norm by

$$\|\Phi\|_{V^N}^2 := \langle \widehat{\mathcal{B}}\Phi, \Phi \rangle. \quad (5.5)$$

Note that this norm, as any norm defined on  $V^N$  in the above fashion, is invariant under the orthogonal group of  $\mathbb{R}^{N \times N}$  in the sense that

$$\|\Phi \mathbf{U}\|_{V^N} = \|\Phi\|_{V^N} \quad (5.6)$$

for all  $\mathbf{U} \in O(N)$ . In the Grassmann manifold, we measure the error between  $[\Phi_{(1)}], [\Phi_{(2)}] \in \mathcal{G}$  by a related metric  $d$  given by

$$d([\Phi_{(1)}], [\Phi_{(2)}]) := \inf_{\mathbf{U} \in O(N)} \|\Phi_{(1)} - \Phi_{(2)} \mathbf{U}\|_{V^N}.$$

If  $[\Phi_{(2)}]$  is sufficiently close to  $[\Phi_{(1)}] \in \mathcal{G}$  it follows from Lemma 3 that this measure given by  $d$  is equivalent to the expression

$$\|(\mathcal{I} - \mathcal{D}_{\Phi_{(1)}})\Phi_{(2)}\|_{V^N}, \quad (5.7)$$

in which we used the  $L_2$ -orthogonal projector  $\mathcal{D}_{\Phi_{(1)}}$  onto the subspace spanned by  $\Phi_{(1)}$ . In the following, let us use the abbreviation  $D = D_{\Psi}$  for the projector on the sought subspace, wherever no confusion can arise. An equivalent error measure for the deviation of  $\Phi \in \mathcal{V}$  from the sought element  $\Psi \in \mathcal{V}$  is then given by the expression

$$\|(\mathcal{I} - \mathcal{D})\Phi\|_{V^N}, \quad (5.8)$$

which will be used in the sequel. In terms of this notation, our main convergence result is the following.

**Theorem 1.** *Under the ellipticity assumption (5.1), the following holds for any of the three algorithms formulated in section 4: For  $\Phi^{(0)} \in U_{\delta}(\Psi)$  sufficiently close to  $\Psi$ , there is a constant  $\chi < 1$  such that for all  $n \in \mathbb{N}_0$ ,*

$$\|(\mathcal{I} - \mathcal{D})\Phi^{(n+1)}\|_{V^N} \leq \chi \cdot \|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}. \quad (5.9)$$

The rest of this section will be mainly dedicated to the proof of this theorem. For the sake of clarity, let us first sketch the proof to be performed: We will exploit the fact that the iteration mapping can be written in the form  $\Phi^{(n)} \mapsto \Phi^{(n)} - \mathcal{B}^{-1}(\mathcal{I} - \mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)})$  and is thus a perturbation of the mapping  $\Phi^{(n)} \mapsto \Phi^{(n)} - \mathcal{B}^{-1}(\mathcal{I} - \mathcal{D}_{\Psi})\mathcal{J}'(\Phi^{(n)})$ . The estimate then splits in two main parts: The first will be a linear part incorporating the Hessian of the Lagrangian and the task will be to show that application of this linear part to an iterate  $\Phi^{(n)} \in \mathcal{G}$  indeed reduces its error in the tangent space of  $\Psi$  (as defined by (5.8)); here, our ellipticity assumption enters as main ingredient. The second part consists of showing that the remaining perturbation terms (including those resulting from projection on the manifold) are of higher order and thus asymptotically neglectable; the main lemmas entering are Lemma 3 above and Lemma 8 to be proven below.

## 5.2 Ellipticity on the tangent space

In this section, we will first formulate a rather general result about how ellipticity on subspaces can be used to construct a contraction on these spaces and then specialize this to the tangent space at the solution  $\Psi$  and assumption 1 in the subsequent corollary. Finally, we will then prove that our assumption 5.1 entering here is indeed true for the simplified problem (1.5).

**Lemma 6.** *Let  $W \subset G \subset W'$  a Gelfand triple,  $U \subset W$  a closed subspace of  $W$  and  $S, T' : W \rightarrow W'$  two bounded elliptic operators, symmetric with respect to the  $G$ -inner product  $\langle \cdot, \cdot \rangle_G$ , satisfying*

$$\gamma \|x\|_W^2 \leq \langle Sx, x \rangle_G \leq \Gamma \|x\|_W^2, \quad (5.10)$$

$$\text{and} \quad \vartheta \|x\|_W^2 \leq \langle T'x, x \rangle_G \leq \Theta \|x\|_W^2 \quad (5.11)$$

for all  $x \in U$ . Moreover, let  $S, T'$  both map the subspace  $U$  to itself. Then there exists a scaled variant  $T = \alpha T'$ , where  $\alpha > 0$ , and a constant  $\beta < 1$  for which

$$\|(I - T^{-1}S)x\|_T \leq \beta \|x\|_{T'}, \quad (5.12)$$

for all  $x \in U$ , where  $\|x\|_T^2 := \langle Tx, x \rangle_G$  is the inner product induced by  $T$ .

*Proof.* It is easy to verify that for  $\beta := (\Gamma\Theta - \gamma\vartheta)/(\Gamma\Theta + \gamma\vartheta) < 1$  and  $\alpha := \frac{1}{2}(\Gamma/\vartheta + \gamma/\Theta)$  there holds

$$|\langle (I - T^{-1}S)x, x \rangle_T| \leq \beta \|x\|_T^2 \quad \text{for all } x \in U. \quad (5.13)$$

Due to the symmetry of  $T, S$  as mappings  $U \rightarrow U$ , the result (5.12) follows.  $\square$

Let  $\lambda_i, i = 1, \dots, N$  the lowest eigenvalues of  $A$ ,  $\psi_i, i = 1, \dots, N$ , the corresponding eigenfunctions, and

$$V_0 = \text{span} \{\psi_i : i = 1, \dots, N\} \quad (5.14)$$

By Lemma 1, there holds  $(V_0^\perp)^N = \mathcal{T}_{[\Psi]} \mathcal{G}$ , where  $\Psi = (\psi_1, \dots, \psi_N)$ . The following corollary is the main result needed for estimation of the linear part of the iteration scheme.

**Corollary 1.** *Let  $\mathcal{J}$  fulfil the ellipticity condition (5.1) and  $B' : V \rightarrow V'$  a symmetric operator that fulfils (5.11) with  $T' = B'$ . Then there exists a scaled variant  $B = \alpha B'$ , where  $\alpha > 0$ , for which for any  $\delta\Phi \in \mathcal{T}_{[\Psi]} \mathcal{G}$  there holds*

$$\|\delta\Phi - \hat{B}^{-1}(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2, \Psi)}(\Psi, \Lambda)\delta\Phi\|_{V^N} \leq \beta \|\delta\Phi\|_{V^N},$$

where  $\beta < 1$  and  $\hat{B}$  is defined by  $B$  via (5.4).

*Proof.* Note that the restriction of  $\hat{B}'$  is a symmetric operator  $V_0^\perp \rightarrow V_0^\perp$ , so that the same holds for the extension  $\hat{B}'$  as mapping  $\mathcal{T}_{[\Psi]}\mathcal{G} \rightarrow \mathcal{T}_{[\Psi]}\mathcal{G}$ .  $(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Psi)}$  also maps  $V_0^\perp \rightarrow V_0^\perp$  symmetricly, so Lemma 6 applies.  $\square$

### 5.3 Residuals and projection on the manifold

For the subsequent analysis, the following result will be useful. It also shows that the “residual”  $(\mathcal{I} - \mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)})$  may be utilized for practical purposes to estimate the norm of the error  $(\mathcal{I} - \mathcal{D})\Phi^{(n)}$ .

**Lemma 7.** *For  $\delta$  sufficiently small and  $\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{\hat{B}} < \delta$ , there are constants  $c, C > 0$  such that*

$$c\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N} \leq \|(\mathcal{I} - \mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)})\|_{(V^N)'} \leq C\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}. \quad (5.15)$$

*An analougeous result holds for gradient error  $\|(\mathcal{I} - \mathcal{D})\mathcal{J}'(\Phi^{(n)})\|_{(V^N)'}$ .*

*Proof.* Let us choose  $\bar{\Psi} \in [\Psi]$  according to Lemma 3 (applied to  $\Phi = \Phi^{(n)}$ ). Letting  $\Delta\Psi := \Phi^{(n)} - \bar{\Psi}$ , there holds by linearization and Lemma 3 (recall that we let  $D = D_\Psi$ )

$$\begin{aligned} (\mathcal{I} - \mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)}) &= (\mathcal{I} - \mathcal{D})\mathcal{J}'(\bar{\Psi}) + (\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Psi)}(\bar{\Psi}, \Lambda)\Delta\bar{\Psi} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}^2) \\ &= (\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Psi)}(\bar{\Psi}, \Lambda)(\mathcal{I} - \mathcal{D})\Phi^{(n)} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}^2) \end{aligned}$$

By assumption 1,  $\|(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Psi)}(\bar{\Psi}, \Lambda)(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{(V^N)'} \sim \|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}$ , from which the assertion follows. The assertion for  $\|(\mathcal{I} - \mathcal{D})\mathcal{J}'(\Phi^{(n)})\|_{(V^N)'}$  follows from the same reasoning by replacing  $\mathcal{L}^{(2,\Psi)}(\bar{\Psi}, \Lambda)$  by  $\mathcal{J}''(\bar{\Psi})$  in the above.  $\square$

The last ingredient for our proof of convergence is following lemma which will imply that the projection following each application of the iteration mapping does not destroy the asymptotic linear convergence.

**Lemma 8.** *Let  $\hat{\Phi}^{(n+1)} = (\hat{\phi}_1, \dots, \hat{\phi}_N)$  the intermediate iterates as resulting from iteration step (2) in algorithm 1 or 2, respectively. For any orthonormal set  $\Phi \in \mathcal{V}$  fulfilling  $\text{span}[\Phi] = \text{span}[\hat{\Phi}^{(n+1)}]$ , its error deviates from that of  $\hat{\Phi}^{(n+1)}$  only by quadratic error term:*

$$\|(\mathcal{I} - \mathcal{D})\Phi\|_{V^N} = \|(\mathcal{I} - \mathcal{D})\hat{\Phi}^{(n+1)}\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\hat{\Phi}^{(n)}\|_{V^N}^2) \quad (5.16)$$

*Proof.* First of all, note that if (5.16) holds for one orthonormal set  $\Phi$  with  $\text{span}[\Phi] = \text{span}[\hat{\Phi}^{(n+1)}]$ , it holds for any other orthonormal set  $\tilde{\Phi}$  with  $\text{span}[\tilde{\Phi}] = \text{span}[\hat{\Phi}^{(n+1)}]$  because  $\|(\mathcal{I} - \mathcal{D})\Phi\|_{V^N} = \|(\mathcal{I} - \mathcal{D})\tilde{\Phi}\|_{V^N}$  for all orthonormal  $\mathbf{U} \in O(N)$ . Therefore, we will show (5.16) for  $\Phi = (\varphi_1, \dots, \varphi_N)$  yielded from  $\hat{\Phi}^{(n+1)}$  by the Gram-Schmidt orthonormalization

procedure. Denote  $\hat{\varphi}_i = \varphi_i^{(n)} + r_i^{(n)}$ , where for  $s_i^{(n)} = B^{-1}(\mathcal{I} - \mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)})$ , we set  $r_i^{(n)} = s_i^{(n)}$  or  $r_i^{(n)} = (I - D_{\Phi^{(n)}})s_i^{(n)}$  for algorithm 1 or 2, respectively. From the previous lemma, we get in particular that  $\|r_i^{(n)}\|_V \lesssim \|(I - D)\phi_i^{(n)}\|_V$  for both cases (remember that  $D = D_{\Psi}$ ). With the Gram-Schmidt procedure given by  $\varphi'_k = \hat{\varphi}_k - \sum_{j < k} \langle \hat{\varphi}_k, \varphi_j \rangle \varphi_j$ ,  $\varphi_k = \varphi'_k / \|\varphi'_k\|$ , the lemma is now proven by verifying that in each of the inner products involved, there occurs at least one residual  $\|r_i^{(n)}\|$ ; and that, on top of this, for the correction directions  $\varphi_j$  there holds  $(I - D)\varphi'_j = \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}) + \mathcal{O}(\sum_{i < k} \|r_i^{(n)}\|_{V^N}) = \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N})$ . Therefore, the correction terms are of  $\mathcal{O}(\|(\mathcal{I} - \mathcal{D})\hat{\Phi}^{(n)}\|_{V^N}^2)$ , thus proving  $\varphi'_k - \hat{\varphi}_k = \mathcal{O}(\|(I - D)\Phi\|_{V^N}^2)$ . It is easy to verify that the normalization of  $\varphi'_k$  only adds another quadratic term, so the result follows.  $\square$

## 5.4 Proof of Convergence

To prove (5.9) for Algorithm 1, we define  $\mathcal{F}(\Phi) = \Phi - \mathcal{B}^{-1}(\mathcal{I} - \mathcal{D}_{\Phi})\mathcal{J}'(\Phi)$ , so that  $\Phi^{(n+1)} = P(\mathcal{F}(\Phi^{(n)}))$ , where  $P$  is a projection on the Grassmann manifold for which  $[P(\mathcal{F}(\Phi^{(n)}))] = [\mathcal{F}(\Phi^{(n)})]$ . For fixed  $n$ , let us choose  $\bar{\Psi} \in \text{span}[\Psi]$  according to Lemma 3, so that, using the abbreviation  $\mathcal{D} := \mathcal{D}_{\Psi}$ ,

$$\bar{\Psi} - \Phi^{(n)} = (\mathcal{I} - \mathcal{D})\Phi^{(n)} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{L_2^N}^2) \quad (5.17)$$

$$\leq (\mathcal{I} - \mathcal{D})\Phi^{(n)} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}^2) \quad (5.18)$$

Introducing  $\Delta\Psi := \Phi^{(n)} - \bar{\Psi}$ , there follows by linearization

$$\|(\mathcal{I} - \mathcal{D})\Phi^{(n+1)}\|_{V^N} \quad (5.19)$$

$$\stackrel{\text{Lemma 8}}{=} \|(\mathcal{I} - \mathcal{D})\mathcal{F}(\Phi^{(n)})\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}^2) \quad (5.20)$$

$$= \|(\mathcal{I} - \mathcal{D})\mathcal{F}(\bar{\Psi}) + (\mathcal{I} - \mathcal{D})\mathcal{F}'(\bar{\Psi})\Delta\Psi\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}^2) \quad (5.21)$$

$$= \|(\mathcal{I} - \mathcal{D})\mathcal{F}'(\bar{\Psi})(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}^2) \quad (5.22)$$

$$= \|(\mathcal{I} - \mathcal{D})(\mathcal{I} - \mathcal{B}^{-1}(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Psi)}(\bar{\Psi}, \Lambda))(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}^2) \quad (5.23)$$

where we have used (5.18) and the fact that  $(\mathcal{I} - \mathcal{D})\mathcal{F}(\bar{\Psi})$  is zero. The proof is now finished by noticing that

$$\begin{aligned} & (\mathcal{I} - \mathcal{D})\left(\mathcal{I} - \mathcal{B}^{-1}(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Psi)}(\bar{\Psi}, \Lambda)\right)(\mathcal{I} - \mathcal{D})\Psi \\ &= \left(\mathcal{I} - \hat{\mathcal{B}}^{-1}(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Psi)}(\bar{\Psi}, \Lambda)\right)(\mathcal{I} - \mathcal{D})\Psi, \end{aligned}$$

so that corollary 1 applies to give

$$\|(\mathcal{I} - \mathcal{D})\Phi^{(n+1)}\|_{V^N} \leq \vartheta\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}^2) \leq \chi\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N},$$

where  $\chi < 1$  for  $\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}$  small enough to neglect the quadratic term.  $\square$

The convergence estimate (5.9) for Algorithm 2 is easily derived from this: Consider

$$\mathcal{F}_2(\Phi) = \Phi - (\mathcal{I} - \mathcal{D}_\Phi)\mathcal{B}^{-1}(\mathcal{I} - \mathcal{D}_\Phi)\mathcal{J}'(\Phi), \quad (5.24)$$

for which  $\Phi^{(n+1)} = P(\mathcal{F}_2(\Phi^{(n)}))$  for the iterates of Algorithm 2. Differentiation of  $\mathcal{F}_2$  at  $\bar{\Psi}$  chosen as before gives

$$\mathcal{F}'_2(\bar{\Psi})\Delta\Psi = \mathcal{I} - (\mathcal{I} - \mathcal{D})\mathcal{B}^{-1}(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2)}(\bar{\Psi}, \Lambda)\Delta\Psi + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}^2),$$

(note that derivation of the projector  $D_{\bar{\Psi}}$  on the left hand side with respect to  $\bar{\Psi}$  results in a zero term), so that the same reasoning as above gives

$$\begin{aligned} & \|(\mathcal{I} - \mathcal{D})\Phi^{(n+1)}\|_{V^N} \\ & \leq \|(\mathcal{I} - \mathcal{D})(\mathcal{I} - \hat{\mathcal{B}}^{-1}(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2, \Psi)}(\bar{\Psi}, \Lambda))(\mathcal{I} - \mathcal{D})\Psi\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}^2) \\ & \leq \chi\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}, \end{aligned}$$

with  $\chi < 1$  for  $\Phi^{(n)}$  close enough to  $\Psi$ .  $\square$

To prove the convergence of the exponential parametrisation (Algorithm 3) defined by

$$\Phi^{(n+1)} := \exp\left(-\alpha\hat{\mathcal{X}}\right)(\Phi^{(n)}),$$

it is enough to notice, cf. the remarks after Lemma 4, that we follow a geodesic path in direction  $(\mathcal{I} - \mathcal{D}_{\Phi^{(n)}})\mathcal{B}^{-1}(A_{\Phi^{(n)}}\Phi^{(n)} - \Phi^{(n)}\Lambda^{(n)})$ , which is equal to the descent direction of Algorithm 2. Due to the definition of the tangent manifold,  $\Phi^{(n+1)}$  again differs from  $\mathcal{F}_2(\Phi^{(n)})$  (defined by (5.24)) only by an asymptotically neglectable quadratic error term.

$\square$

## 5.5 Quadratic convergence of the energy

For the Rayleigh quotient  $R(\phi^{(n)})$ , i.e. for the simplified problem and  $N = 1$ , it is known that  $R(\phi^{(n)}) - R(\psi) \lesssim \|\psi - \phi^{(n)}\|_V^2$ . To end this section, we will show that this property holds also for the computed energies, provided that the constraints are satisfied exactly and the functional is sufficiently often differentiable. The latter is only known for Hartree-Fock and the simplified problem. Since the exchange correlation potential is not known exactly, this question remains open in general for the density functional theory.

**Theorem 2.** *Provided that  $\mathcal{J}$  is two times differentiable on a neighborhood  $U_\delta(\Psi) \subseteq V^N$  of the minimizer  $\Psi$ , and that for fixed  $\Phi \in U_\delta(\Psi)$ ,  $\mathcal{J}''$  is continuous on  $\{t\Psi + (1-t)\Phi | t \in [0, 1]\}$ , the error in the energy depends quadratically on the approximation error of the minimizer  $\Psi$ , i.e.*

$$\mathcal{J}(\Phi) - \mathcal{J}(\Psi) \lesssim \|(I - \mathcal{D}_\Psi)\Phi^{(n)}\|_{V^N}^2. \quad (5.25)$$

*Proof.* Let us choose a representant of the solution  $\Psi$  according to Lemma 3. Abbreviating  $e = \Phi - \Psi$ , we can use  $\mathcal{J}'(\Psi)((\mathcal{I} - \mathcal{D})\Phi) = 0$  to find that

$$\mathcal{J}'(\Psi)(e) = \mathcal{J}'(\Psi)((\mathcal{I} - \mathcal{D})\Phi) + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi\|^2) = \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi\|^2)$$

so that

$$\begin{aligned} \mathcal{J}(\Phi) - \mathcal{J}(\Psi) &= \int_0^1 \mathcal{J}'(\Psi + se)(e) ds + \frac{1}{2} \mathcal{J}'(\Phi)(e) \\ &\quad - \frac{1}{2} (\mathcal{J}'(\Psi)(e) + \mathcal{J}'(\Phi)(e)) + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi\|^2). \end{aligned}$$

By integration by parts,

$$\frac{1}{2}(f(0) + f(1)) = \int_0^1 f(t) dt + \int_0^1 (s - \frac{1}{2}) f'(s) ds,$$

so that

$$\mathcal{J}(\Phi) - \mathcal{J}(\Psi) = \frac{1}{2} \langle \mathcal{J}'(\Phi), \Phi - \Psi \rangle - \int_0^1 (s - \frac{1}{2}) \mathcal{J}''(\Phi + se)(e, e) ds + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi\|^2).$$

For estimation of the first term on the right hand side, recall from (5.15) that

$$\|(\mathcal{I} - \mathcal{D})\mathcal{J}'(\Phi)\|_{V^N} \lesssim \|(I - \mathcal{D})\Phi\|_{V^N},$$

and therefore

$$\begin{aligned} \frac{1}{2} \langle \mathcal{J}'(\Phi), \Phi - \Psi \rangle &= \frac{1}{2} \langle (\mathcal{I} - \mathcal{D})\mathcal{J}'(\Phi), (\mathcal{I} - \mathcal{D})\Phi \rangle + \mathcal{O}(\|(I - \mathcal{D})\Phi\|^2) \\ &= \mathcal{O}(\|(I - \mathcal{D})\Phi\|^2), \end{aligned}$$

while for the second term,  $|\int_0^1 (s - \frac{1}{2}) \mathcal{J}''(\Phi + se)(e, e) ds| = \mathcal{O}(\|e\|^2) = \mathcal{O}(\|(I - \mathcal{D})\Phi\|^2)$  follows from the continuity of  $\mathcal{J}''$  and, again, the usage of Lemma 3.  $\square$

## 6 Further Comments and Conclusions

Before we conclude this article with numerical examples, we would like to make some comments about the complexity of the numerical schemes when applied to the problems of section 3, and about the potentialities for accelerating convergence of the iteration scheme.

**Complexity:** Concerning disk storage, the task is to compute  $N$  functions  $\psi \in V_h$ , so  $\mathcal{O}(N \dim V_h)$  memory is needed to store the orbital functions, while storage of the discretization of the Fock operator  $A$  requires at most  $\mathcal{O}((\dim V_h)^2)$  in the general and worst case, but only  $\mathcal{O}(\dim V_h)$  for sparse discretizations. Regarding computational demands, the non-zero entries of a sparse discretization of  $A$  are of  $\mathcal{O}(\dim V_h)$ , so that the complexity of the application of  $A$  depends linearly on  $\dim V_h$ . The computation of  $\langle A \hat{\phi}_i^{(n+1)}, \hat{\phi}_j^{(n+1)} \rangle$ , and  $\langle \hat{\phi}_i^{(n+1)}, \hat{\phi}_j^{(n+1)} \rangle$  needs  $\mathcal{O}(N^2(\dim V_h))$  operations in the case of sparse discretizations (and  $\mathcal{O}(N^2(\dim V_h)^2)$  in the worst case). The orthogonalization procedure, i.e. the projection onto the Stiefel manifold usually has a complexity  $\mathcal{O}(N^2 \dim V_h)$ . To relate the above complexities to the size  $N$  of the electronic system, it is also interesting to discuss how large  $\dim V_{h,min}$  has to be chosen for a given size  $N$ . To this end, we might fix a given maximal error  $e$  per atom or electron (usually requested to be smaller than the intrinsic modeling error of DFT or HF models) and determine the minimal ansatz space dimension  $\dim V_{h,min}(N)$  that keeps the numerical error under that error  $e$ . If we then consider the scaling of  $\dim V_{h,min}$  with respect to the size of the system  $N$ , it turns out that  $\dim V_{h,min}(N) = \mathcal{O}(N)$ , where the constant in front of  $N$  is extremely large for systematic basis functions and surprisingly small for Gaussian type basis functions. Therefore, the natural scaling of the orbital based DFT and/or HF computations with respect to the size  $N$  of the underlying system gives an overall complexity of  $\mathcal{O}(N^3)$  (or even  $\mathcal{O}(N^4)$  for non-sparse discretizations).

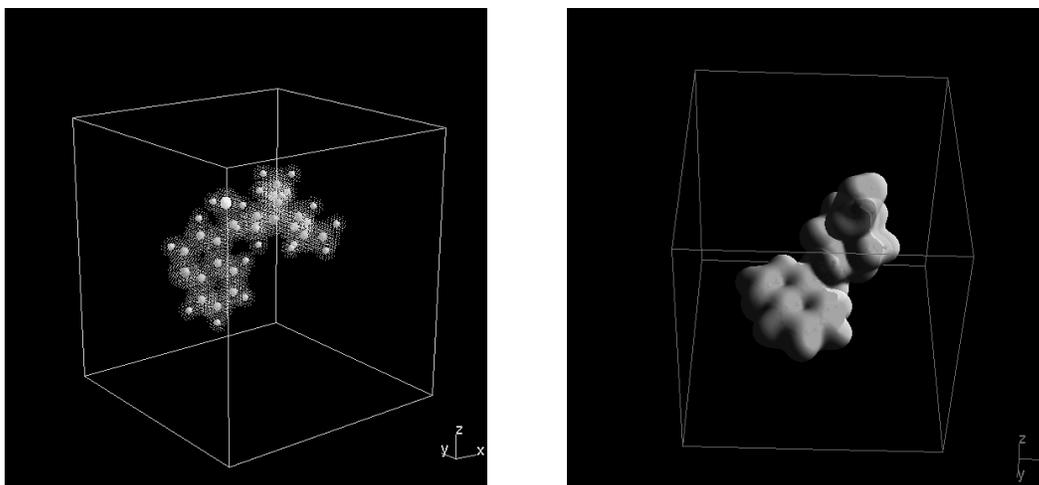
This can be improved if the discretization of the individual orbitals  $\phi_i^{(n)}$  requires substantially less than  $\dim V_h$  DOFs. In an optimal case, one may achieve  $\mathcal{O}(1)$  for a fixed accuracy per atom; this is for example the case if the diameter of support of  $\phi_i^{(n)}$  is of  $\mathcal{O}(1)$ , i.e. the support is local. In this case, the total complexity scales only linearly with respect to  $N$ . Usually, the eigenfunctions  $\check{\psi}_i$  have global support. For insulating materials, though, there exists a representation  $\Psi_{loc}$  such that  $[\Psi_{loc}] = [\check{\Psi}] \in \mathcal{G}$  and  $|\psi_{loc,i}(x)| \lesssim e^{-\alpha|x-x_i|}$ ,  $\alpha \gg 0$  sufficiently large. These representations are called maximally localized or Wannier orbitals. Linear scaling  $\mathcal{O}(N)$  can be achieved if, during the iteration, the representant  $\Phi_{loc}^{(n)}$  in the Grassmann manifold is selected and approximated in a way that the diameter of support is of  $\mathcal{O}(1)$ . This is the strategy pursued in Big DFT to achieve linear scaling, [17, 22]. We defer the further details to a forthcoming paper. A related approach, computing localized orbitals in an alternative way was proposed by [6] and exhibits extremely impressing results.

**Convergence and Acceleration:** In the present paper we have considered linear convergence of a preconditioned gradient algorithm. For the simplified model, this convergence is guaranteed by the spectral gap condition, in physics referred as the HOMO-LUMO gap (i.e. highest occupied molecular orbital-lowest unoccupied molecular orbital gap). For the Hartree-Fock model, this condition is replaced by the coercivity condition 5.1. The same condition applies to models in density functional theory, provided the Kohn-Sham energy functionals are sufficiently often differentiable. Let us mention that a verification of this conditions will answer important open problems in Hartree-Fock theory, like uniqueness etc. The performance of the algorithm may be improved by an optimal line search, replacing  $\mathcal{B}$  by an optimal  $\alpha_n \mathcal{B}$ . Except for the simplified problem, where an optimal line search performed like in the Jacobi-Davidson algorithm as a particular simple subspace acceleration, optimal line search is rather expensive though and not used in practice. Since the present preconditioned steepest decent algorithm is gradient directed, a line search based on the Armijo rule will guarantee convergence in principle, even without a coercivity condition [5, 14].

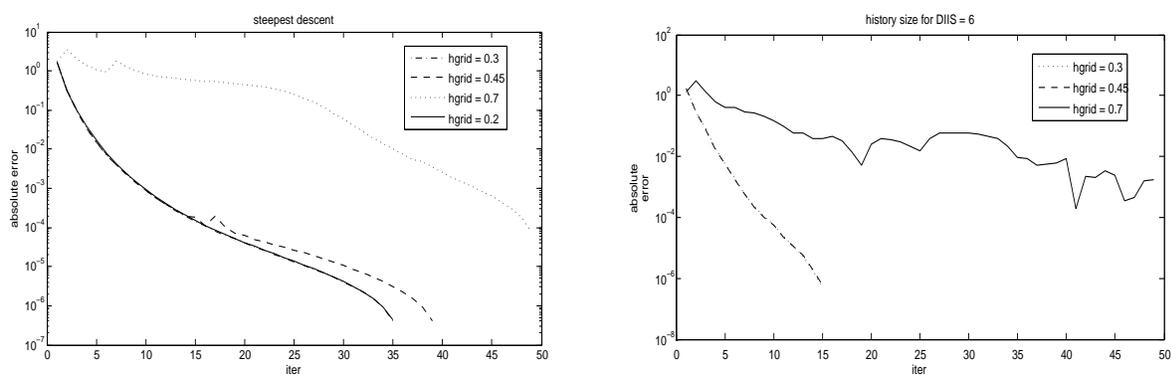
In practice, convergence is improved by subspace acceleration techniques, storing iterates  $\Phi^{(n-k)}, \dots, \Phi^{(n)}, \widehat{\Phi}^{(n+1)}$  and compute  $\Phi^{(n+1)}$  from an appropriately chosen linear combination of them. Most prominent examples are the DIIS [37] and conjugate gradient [2, 4] algorithm. The DIIS algorithm is implemented in the EU NEST project BigDFT, and frequently used in other quantum chemistry codes. Without going into detailed descriptions of those methods and further investigations, let us point out that the analysis in this paper provides the convergence of the worst case scenario. Second order methods, in particular Newton methods have been proposed in literature [35], but since these require the solution of a linear system of size  $N \dim V_h \times N \dim V_h$ , they are to be avoided.

## 7 Numerical examples

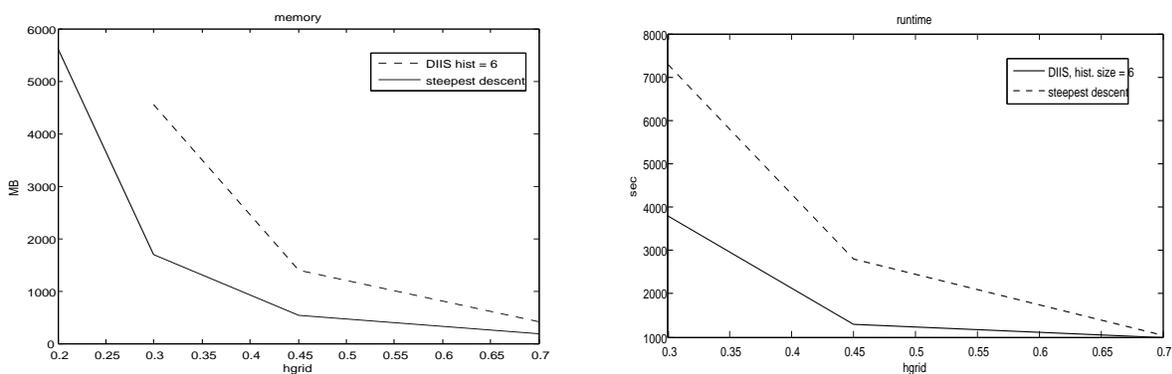
The proposed direct minimization algorithm 1 is realized in the recent density functional code bigDFT [45], which is implemented in the open source ABINIT package, a common project of the Université Catholique de Louvain, Corning Incorporated, and other contributors [46, 23, 24, 15]. It relies on an efficient Fast Fourier Transform algorithm [19] for the conversion of wavefunctions between real and reciprocal space, together with a DIIS subspace acceleration. We demonstrate the convergence for the simple molecule cinchonidine ( $C_{19}H_{22}N_2O$ ) of moderate size  $N = 55$  for a given geometry of the nuclei displayed in figure 7.1. Despite the fact that the underlying assumptions in the present paper cannot be verified rigorously, the proposed convergence behavior is observed by all benchmark computations. The algorithm is experienced to be quite robust also if the HOMO-LUMO gap is relatively small.



**Figure 7.1:** Atomic geometry and electronic structure of cinchonidine



**Figure 7.2:** Convergence history for the direct minimization scheme (left) and with DIIS acceleration (right) for different mesh sizes.



**Figure 7.3:** Memory requirements (left) and computing time (right) for direct minimization algorithm with and without DIIS acceleration.

For our computations, we have used a simple LDA (local density approximation) model proposed by [20] and norm-conserving non-local pseudopotentials [21]. The orbital functions  $\psi_i$  are approximated by Daubechies orthogonal wavelets with 8 vanishing moments based on an approximate Galerkin discretization [18]. For updating the nonlinear potential, the electron density is approximated by interpolating scaling functions (of order 16). The discretization error can be controlled by an underlying basic mesh size  $h_{grid}$ .

In figure 7.2, we demonstrate the convergence of the present algorithm for 4 different choices of mesh sizes, where the error is given in the energy norm of the discrete functions. The initial guess for the orbitals is given by the atomic solutions. Except in case of non-sufficient resolution ( $h_{grid} = 0.7$ ), where we obtain a completely wrong result, convergence is observed. If the discretisation is sufficiently good, we do not observe much difference in the convergence history for different mesh sizes. Since the convergence speed depends on the actual solution, it is only possible to observe that the convergence is bounded by a linear rate.

The number of iterations is relatively moderate bearing in mind that one iteration step only requires matrix-vector multiplications with the Fock operator and not a corresponding solution of linear equations. The DIIS implemented in BigDFT accelerates the iteration by almost halving the number of iterations and the total computing time at the expense of additional storage capacities, see also figure 7.2. Further benchmark computations have already been performed and will be reported in different publications by the groups involved in the implementation of BigDFT.

## References

- [1] P.-A. Absil, R. Mahony, R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2007
- [2] D. C. Allen, T. A. Arias, J. D. Joannopoulos, M. C. Payne, M. P. Teter, *Iterative minimization techniques for ab initio total energy calculation: molecular dynamics and conjugate gradients*, Rev. Modern Phys. 64 (4), 1045-1097, 1992.
- [3] T. A. Arias, *Multiresolution analysis of electronic structure: semicardinal and wavelet bases*, Rev. Mod. Phys. 71, 267 - 311, 1999.
- [4] T. A. Arias, A. Edelman, S. T. Smith, *The Geometry of Algorithms with Orthogonality Constraints*, SIAM J. Matrix Anal. and Appl., Vol. 20, No. 2, pp. 303-353, 1999.
- [5] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math, 1966.

- [6] M. Barrault, E. Cancès, W. W. Hager, C. Le Bris, *Multilevel domain decomposition for electronic structure calculations*, Journ. Comp. Phys., Volume 222 , 1, 86-109, 2007.
- [7] T. L. Beck, *Real-space mesh techniques in density-functional theory*, Rev. Mod. Phys. 72, 1041 - 1080, 2000.
- [8] J. H. Bramble, J. E. Pasciak, A.V. Knyazev, *A subspace preconditioning algorithm for eigenvector/eigenvalue computation*, Advances in Computational Mathematics 6 (1996) 159-189, 1999.
- [9] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, Y. Maday, *Computational Quantum Chemistry: A Primer*, Handbook of Numerical Analysis, Volume X, Elsevier Science, 2003.
- [10] E. Cancès, C. Le Bris, *On the convergence of SCF algorithms for the Hartree-Fock equations*, Mathematical Models and Methods in Applied Sciences, 1999.
- [11] W. Dahmen, T. Rohwedder, R. Schneider, A. Zeiser, *Adaptive Eigenvalue Computation - Complexity Estimates*, preprint, 2007 , obtainable at <http://arxiv.org/abs/0711.1070>
- [12] M. Defranceschi, C. Le Bris, *Mathematical Models and Methods for Ab Initio Quantum Chemistry*, Lecture Notes in Chemistry, Springer, 2000.
- [13] R. M. Dreizler, E. K. U. Gross, *Density functional theory*, Springer, 1990.
- [14] C. Geiger, C. Kanzow, *Theorie und Numerik restringierter Optimierungsaufgaben*, Springer, 2002.
- [15] L. Genovese, A. Neelov, S. Goedecker, T. Deutsch, S. A. Ghasemi, A. Willand, D. Caliste, O. Zilberberg, M. Rayson, A. Bergman, R. Schneider, *Daubechies wavelets as a basis set for density functional pseudopotential calculations*, preprint, 2008, obtainable at <http://arxiv.org/abs/0804.2583>
- [16] S. Goedecker, *Wavelets and their Application for the Solution of Partial Differential Equation*, Presses Polytechniques Universitaires et Romandes, Lausanne, 1998.
- [17] S. Goedecker, *Linear Scaling Methods for the Solution of Schrodinger's Equation*, in: *Handbook of Numerical Analysis Vol. X, Special volume on Computational Chemistry* , P.G. Ciarlet and C. Le Bris (editors), North-Holland, 2003.

- [18] A. Neelov, S. Goedecker, *An efficient numerical quadrature for the calculation of the potential energy of wavefunctions expressed in the Daubechies wavelet basis*, J. of. Comp. Phys. 217, 312-339, 2006.
- [19] S. Goedecker, *Fast radix 2, 3, 4 and 5 kernels for Fast Fourier Transformations on computers with overlapping multiply-add instructions*, SIAM J. on Scientific Computing 18, 1605, 1997.
- [20] S. Goedecker, C. J. Umrigar, *Critical assessment of the self-interaction-corrected local-density-functional method and its algorithmic implementation*, Phys. Rev. A 55, 1765 - 1771, 1997.
- [21] S. Goedecker, M. Teter, J. Hutter, *Separable dual-space Gaussian pseudopotentials*, Phys. Rev. B 54, 1703, 1996.
- [22] S. Goedecker, *Linear scaling electronic structure methods*, Rev. Mod. Phys. 71, 1085 - 1123, 1999.
- [23] X. Gonze, J.-M. Beuken, R. Caracas, F. Detraux, M. Fuchs, G.-M. Rignanese, L. Sindic, M. Verstraete, G. Zerah, F. Jollet, M. Torrent, A. Roy, M. Mikami, Ph. Ghosez, J.-Y. Raty, D.C. Allan, *First-principles computation of material properties : the ABINIT software project*, Computational Materials Science 25, 478-492, 2002.
- [24] X. Gonze, G.-M. Rignanese, M. Verstraete, J.-M. Beuken, Y. Pouillon, R. Caracas, F. Jollet, M. Torrent, G. Zerah, M. Mikami, Ph. Ghosez, M. Veithen, J.-Y. Raty, V. Olevano, F. Bruneval, L. Reining, R. Godby, G. Onida, D.R. Hamann, D.C. Allan, *A brief introduction to the ABINIT software package* Zeit. Kristallogr. 220, 558-562, 2005.
- [25] W. Hackbusch, *Iterative solution of large sparse systems of equations*, Springer, 1994.
- [26] T. Helgaker, P. Jorgensen, J. Olsen, *Molecular electronic-structure theory*, Wiley, New York, 2000.
- [27] P. Hohenberg, W. Kohn, *Inhomogeneous Electron Gas*, Phys. Rev. 136 p. 864-871, 1964.
- [28] A. V. Knyazev, K. Neymeyr, *A geometric theory for preconditioned inverse iteration III: A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl. 358, 95-114. 2003.
- [29] E. H. Lieb, B. Simon, *The Hartree-Fock Theory for Coulomb Systems*, Commun. Math. Phys. 53, 185-194, 1977.

- [30] P. L. Lions, *Solution of the Hartree Fock equation for Coulomb Systems*, Commum. Math. Phys. Volume 109, Number 1, 33-97, 1987.
- [31] D. Luenberger, *Optimization by Vector Space Methods*, Wiley, 1968.
- [32] C. Lubich, O. Koch, *Dynamical Low Rank Approximation*, preprint, Uni Tübingen, 2008.
- [33] C. Moler, C. van Loan, *Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later*, SIAM Vol. 45, No.1, pp. 3-49, 2003.
- [34] K. Neymeyr, *A geometric theory for preconditioned inverse iteration applied to a subspace*, Math. Comp. 71, 197-216, 2002.
- [35] Y. Maday, G. Turinici, *Error bars and quadratically convergent methods for the numerical simulation of the Hartree-Fock equations*, Numerische Mathematik, Springer, 2003.
- [36] J. Nocedal, S. J. Wright, *Numerical Optimization*, Springer, 1999.
- [37] P. Pulay, *Convergence Acceleration in Iterative Sequences: The Case of SCF Iteration*, Chem. Phys. Lett. 73, 393, 1980.
- [38] D. Raczkowski, C. Y. Fong, P.A. Schultz, R. A. Lippert, E. B. Stechel, *Unconstrained and constrained minimization, localization, and the Grassmann manifold: Theory and application to electronic structure* Phys. Rev. B 64, 2001.
- [39] M. Reed and B. Simon, *Methods of Modern Mathematical Physics IV: Analysis of Operators*, Academic Press, 1978.
- [40] T. Rohwedder, R. Schneider, A. Zeiser, *Perturbed preconditioned inverse iteration for operator eigenvalue problems with applications to adaptive wavelet discretization*, preprint, 2007, obtainable at <http://arxiv.org/abs/0708.0517>.
- [41] Y. Saad, *Analysis of some Krylov Subspace Approximations to the Matrix Exponential Operator*, SIAM Journal on Numerical Analysis, Vol. 29, NO. 1., pp. 209-228, 1992.
- [42] Y. Shao, C. Saravanan, M. Head-Gordon, C. A. White, *Curvy steps for density matrix-based energy minimization: Application to large-scale self-consistent-field calculations* J. Chem. Phys. 118, 6144 ,2003.
- [43] A. Szabo, N. S. Ostlund, *Modern Quantum Chemistry*, Dover Publications Inc., 1992.

[44] R. B. Sidje, *EXPOKIT: A Software Package for Computing Matrix Exponentials*, ACM. Trans. Math. Softw., 24(1):130-156, 1998.

[45] [http://www-drfmc.cea.fr/sp2m/L\\_Sim/BigDFT/index.en.html](http://www-drfmc.cea.fr/sp2m/L_Sim/BigDFT/index.en.html)

[46] <http://www.abinit.org>

**Authors:**

**Johannes Blauert**

**Reinhold Schneider**

**Thorsten Rohwedder**

---

**Institute for Mathematics  
Technical University Berlin  
Straße des 17. Juni 135  
10623 Berlin  
Germany**

**Alexej Neelov**

---

**Institute of Physics  
University of Basel  
Klingelbergstrasse 82  
4056 Basel  
Switzerland**