

# Analysis of the projected Coupled Cluster Method in Electronic Structure Calculation

Reinhold Schneider

8.11.2008

## Abstract

The electronic Schrödinger equation plays a fundamental role in molecular physics. It describes the stationary nonrelativistic behaviour of an quantum mechanical N electron system in the electric field generated by the nuclei. The (Projected) Coupled Cluster Method has been developed for the numerical computation of the ground state energy and wave function. It provides a powerful tool for high accuracy electronic structure calculations. The present paper aims to provide a rigorous analytical treatment and convergence analysis of this method. If the discrete Hartree Fock solution is sufficiently good, the quasi-optimal convergence of the projected coupled cluster solution to the full CI solution is shown. Under reasonable assumptions also the convergence to the exact wave function can be shown in the Sobolev  $H^1$ -norm. The error of the ground state energy computation is estimated by an Aubin Nitsche type approach. Although the Projected Coupled Cluster method is nonvariational it shares advantages with the Galerkin or CI method. In addition it provides size consistency, which is considered as a fundamental property in many particle quantum mechanics.

**Key words:** electronic Schrödinger equation, quantum chemistry, coupled cluster method, Galerkin method, error analysis

**Mathematics Subject Classification:** 65N25,65N30,65Z05,81V70,81V55

This research was supported by DFG priority program SSP 1145 *Modern and universal first-principles methods for many-electron systems in chemistry and physics*

## 1 Introduction

Reliable computations of molecular behavior are requested e.g. in chemistry, molecular biology, material sciences or semi-conductor devices etc. with increasing importance recent in science

and technology. On an atomic or molecular length scale, the physics is completely governed by the laws of quantum mechanics. Consequently, numerical modeling of such processes should be based on *first principles* of quantum mechanics. The basic equation for a quantitative description of most phenomena is the *electronic Schrödinger equation*. It describes the stationary and non-relativistic behavior of an ensemble of  $N$  electron in an electric field resulting from fixed nuclei. A quantity of major interest is the ground state energy  $E_0$ , which is the lowest eigenvalue of the Schrödinger-Hamilton operator. Despite the relatively simple form of the electronic Schrödinger equation, the numerical approximation of the solution is an extremely difficult and challenging problem. This is mainly due to the fact that the corresponding eigenfunction  $\Psi$  depends on at least  $3N$  spatial variables. The complexity of traditional approximation methods like Finite Elements, Fourier series, splines etc. is scaling exponentially with respect to the spatial dimensions, an effect often called the *curse of dimensions*. Furthermore, the wave function  $\Psi$  is not smooth, and approximating these cusps, in particular the electron-electron cusp, causes additional severe problems. After 70 years of progress, recent electronic structure calculation algorithm are powerful and demonstrate impressive accuracy. The Nobel Prize in chemistry has been awarded to Kohn, Pople for their contributions to this success. The curse of dimensions might be circumvented by replacing the original many particle Coulomb system by an artificial (nonlinear) system of non-interacting particles, as done in Density Functional Theory. Even this approach is extremely successful, there remains an unavoidable a modeling error and no constructive method is known, which can provide the convergence to the exact results, in principle. Whenever highly accurate results are required one needs a reasonable approximations of the wave function. In particular, if one is asks for derivatives of the ground state energy w.r.t. to some parameters, like e.g. forces,

Among those methods based on computing an approximate wave function, the *Coupled Cluster Method* (CC-Method) provides an extremely powerful tool. Nowadays it is considered as the most efficient and preferred approach whenever highly accurate calculations are required, although the CC method fails in many situations. The CC method, more precisely the Projected Coupled Cluster Method, was invented by Coester and Kümmel for many body quantum theory, based on an exponential ansatz introduced by Hubbard. It has has been applied to problems in quantum chemistry first by Paldus & Czischek in the sixties, and from that time on it has been developed further, see e.g. [18] for further historical remarks. It was motivated by the general observation, that for most many-body quantum system the energy is an extensive quantity being proportional to the particle number. For independent systems this behavior is formulated as *size consistency*, a property on which scientists in physics and chemistry pay severe attention and emphasize. Size consistency can be easily guaranteed using an exponential parametrization of the wave function as done in the Coupled Cluster method. When it comes to numerical computations, the approximate Coupled Cluster parametrization is found by solving a system of nonlinear equations, namely the amplitude equations, which are derived by the Projected Coupled Cluster approach. Since this ansatz is not variational, it was disputed in literature for several times, although it works extremely well, see e.g. [20] for a very detailed and extensive study of the convergence of the projected Coupled Cluster method and comparison with truncated CI methods as well as electron pair approximations (CEPA) and alternative Coupled Cluster ap-

proaches. The present paper, is to the authors knowledge, a first mathematical attempt to analyse the Projected Coupled Cluster, or shortly the Coupled Cluster approach, from the perspective of numerical analysis. The outcome of the paper is, roughly speaking, that the Projected Coupled Cluster method almost preserves some advantages of variational methods, like quasi-optimal convergence of the wave function and super-convergence of the energy, provided that the reference determinant is sufficiently closed to the exact wave function. In this situation, the Coupled Cluster method becomes a favorable tool, due to its size consistency. This confirms the experience that it is extremely powerful for describing what is called *dynamical correlation*, and is in accordance to conclusions of [20] and others. But it fails in difficult situations, where a single Hartree-Fock determinant is not sufficient for approximating the wave functions appropriately. Examples for this difficult situations are e.g. open shell systems or quantitatively correct descriptions of the dissociation of a closed shell system into open shell systems.

For the present investigations, we will consider of the CC method in the same framework way it is traditionally applied in practise. Therefore, the (approximate) Hartree-Fock (HF) wave function, as a result of a previous Hartree-Fock computation, is taken as the reference Slater determinant, which provides a canonical choice. We consider canonical orbitals, which are eigenfunctions of the Fock operator, as the underlying basis set for approximating the wave function. For sake of simplicity, we always use the spin orbital formulation. In the present framework, there arises a principle difficulty, since the HF Slater determinant as well as the Fock operator depend on the actual discretisation, i.e. the underlying (finite) basis set. Consequently, the approximation of the wave function by a Projected Coupled Cluster solution will be considered here in two steps. The first part is the convergence of the Full CI solution towards the exact wave function. Since the Full CI method is a Galerkin method on the Full CI ansatz space, and we refer only to known results about convergence of the Galerkin method for solving symmetric eigenvalue problems. Although the full CI ansatz space is finite dimensional, its dimension increases exponentially with the numbers of electrons  $N$ . Therefore the computation of the full CI solution becomes untractable, or at least prohibitively expensive, except for very small systems. Even this space is finite-dimensional, we always try to keep our estimates independent of the size of the underlying basis set and the dimensions of this space. We show rigorously that any wave function, which is not orthogonal to the reference determinant can be parameterized by an exponential ansatz. Probably this results has been proved earlier in physical literature, where nowdays it is taken for granted without any further reference. This implies that the solution of the full set of amplitude equation exists. To make the method computational accessible, only a small subset of amplitude equations are considered for a restricted CC ansatz. The standard approach is the Projected Coupled Cluster method. Since the Projected CC Method is not variational, it is not clear whether amplitude equations of the Projected Coupled Method are uniquely solvable or not, and how accurate this approximation performs. In the present paper, the Projected Coupled Cluster Method is considered as a Galerkin scheme for the nonlinear amplitude equations. To guarantee solvability and convergence we consider strict monotonicity as a sufficient condition. For instance this condition can be satisfied if the fluctuation potential is sufficiently small compared to the Fock operator. To prove local existence and error estimates rigorously, we consider an equivalent but different set of nonlinear amplitude equations.

In order to show the convergence of the projected CC-solution to the exact wave function, the convergence estimates for the CC solution to the CI solution should be uniform with respect to actual discretisations, e.g. the size of basis sets. This means that the constants involved in these estimates do not depend explicitly on the size of the basis set. In particular, we will introduce a weighted  $l_2$ -norm for the coupled cluster amplitudes, which is shown to be equivalent to the  $H^1$  Sobolev norm of the approximate wave functions to a certain extend. The CC method is trying to compute an excitation operator  $T$ . The norm of the amplitudes appear in two different circumstances now. One role is in a linear and an exponential parametrization applied to the reference. By this way, we will measure the (full CI) wave function and the approximation error. And on the other hand, the amplitudes defining the operator  $T$  and the norm of the amplitude vector is related to the operator norms of the excitation operator  $T$  acting on  $L_2$  as well as on  $H^1$ . A major result of this paper states that, in the present situation, all these norms are more or less equivalent, with constants depending on the actual Schrödinger equation, e.g. the particle numbers, but not on its discretisation. With this framework and results at hand, the Projected CC Method can be interpreted as a Galerkin scheme for a non-linear operator acting on the space of amplitude vectors. Under the assumption that the reference determinant is sufficiently closed to the wave function, strict monotonicity of the modified amplitude function, solvability and quasi-optimal convergence of the projected CC method can be concluded. Finally, to estimate the convergence of the computed eigenvalue to the exact ground state energy, we use the *dual weighted residual* approach, e.g. [3], to apply some Aubin Nitsche like argument for the derivation of a priori estimates. Since the present analysis is based on canonical orbitals and discrete Fock operators, the present investigations will be completed by a detailed analysis of a uniform behavior of such functions and operators, e.g. uniform boundedness, uniform gap conditions etc. The present analysis highlights the role of the HOMO-LUMO gap and the underlying reference determinant on the performance of the Projected Coupled Cluster method. We would like to mention that the dual weighted residual method has been developed for the design of local error estimators for adaptive finite element methods in order to optimize the finite element meshes to compute certain functionals of the solution. However adaptive finite element method (FEM) techniques cannot be transferred immediately to the present setting, but we hope that the present investigations could be a starting point for developing adaptive techniques for the solution of the electronic Schrödinger equation. Procedures for optimizing the basis set together with an automatic selection of amplitudes giving the most significant contribution to the ground energy could improve recent ab initio computations of the electronic structure.

The presentation is kept fairly self-contained to make it accessible to numerical analysts which are not particular familiar with wave function methods in quantum chemistry. For instance, we have avoided the notion of second quantization completely. The paper is organized as follows. After the formulation of the electronic Schrödinger equation, in Chapter 3, we provide a collection of required analytical results, notations, definitions, which are basic for this paper, but cannot be easily in the present form in existing literature. A best reference is [5]. The Coupled Cluster method is introduced in Chapter 4. The exponential ansatz is verified rigorously, by a global existence result Theorem 4.3. The Projected Coupled Cluster Method for its numerical is introduced as Galerkin solution of the nonlinear amplitude equations on a certain normed vector

space of amplitude vectors. The excitation operator is analysed in detail, showing that its norm is equivalent to the norm of the wave functions, Theorem 4.11 , as well as an appropriate operator norm, see e.g. Theorem 4.15. With these crucial results at hand, a local existence result together with quasi-optimal convergence w.r.t the wave function could be proved in Section 5. The error in the energy is considered in Chapter 6, based on duality techniques. The performance of the CC method has frequently been reported about in the quantum chemistry literature, we illustrate this by a simple comparison with CISD computations and refer to further results in literature. A particular reference for a comparison of CCSD (Coupled Cluster Single Double) with CISD (Configuration Interaction Single Double) and further numerical examples are presented in the monograph [14].

There exists several monographs concerning the numerical solution of the electronic Schrödinger equation from the perspective of quantum chemistry and physics, in particular [31] contains classical introduction. For a recent comprehensive and relatively complete and modern treatment we would like to refer to [14] and the literature cited therein. In contrast to highly developed numerical analysis in fluid dynamics, continuum mechanics or electro dynamics, there exist only very few contributions and mathematical investigations to numerical problems in quantum mechanics, in particular to those who are related to the convergence of the numerical methods including a priori and/or a posteriori error estimates. A priori estimates for the CI method are following directly from known results about the Galerkin method, see e.g. [34] for a short summary. The more complex optimization problem tackled with MCSFC (Multi Configuration Self Consistent Field) methods have been investigated in [24] and [13]. For a survey of state of art in numerical analysis, we refer to [7, 23] and the literature cited therein. Due to the authors knowledge, the convergence of the Coupled Cluster (CC) Method, or more precisely of the *Projected Coupled Cluster Method* has not been investigated from a numerical analysis perspective so far, and we would like to refer to [20] for a different but highly interesting treatment of error analysis of the CC method in quantum chemistry.

## 2 The electronic Schrödinger equation

A quantum mechanical system of  $N$  identical non-relativistic electrons is completely described by a *state-function*  $\Psi$  depending on  $3N$  spatial variables  $\mathbf{x}_i \in \mathbb{R}^3$ ,  $i = 1, \dots, N$ , together with  $N$  discrete spin variables  $s_i \in \{\pm \frac{1}{2}\}$ ,  $i = 1, \dots, N$ ,

$$\Psi : (\mathbb{R}^3 \otimes \{\pm \frac{1}{2}\})^N \rightarrow \mathbb{C} \quad , \quad \Psi(\mathbf{x}_1, s_1; \dots; \mathbf{x}_N, s_N) .$$

The function  $\Psi$  belongs to the Hilbert space  $L_2((\mathbb{R}^3 \times \{\pm \frac{1}{2}\})^N)$  with an inner product defined by

$$\langle u, v \rangle = \sum_{s_i=\pm \frac{1}{2}} \int_{\mathbb{R}^{3N}} u(\mathbf{x}_1, s_1; \dots; \mathbf{x}_N, s_N) \overline{v(\mathbf{x}_1, s_1; \dots; \mathbf{x}_N, s_N)} d\mathbf{x}_1 \dots d\mathbf{x}_N .$$

The *Pauli antisymmetry principle* states that the wave function of fermions, in particular electrons, must be antisymmetric with respect of permutation of variables, i.e.

$$\Psi(\mathbf{x}_1, s_1; \dots; \mathbf{x}_i, s_i; \dots; \mathbf{x}_j, s_j; \dots; \mathbf{x}_N, s_N) = -\Psi(\mathbf{x}_1, s_1; \dots; \mathbf{x}_j, s_j; \dots; \mathbf{x}_i, s_i; \dots; \mathbf{x}_N, s_N).$$

The well known *Born-Oppenheimer-approximation* considers a nonrelativistic quantum mechanical system of  $N$  electrons in an exterior field generated by the nuclei. Therein the wave function  $\Psi$  is supposed to satisfy the *stationary Schrödinger equation*, i.e. the wave function is an eigenfunction of a differential operator, namely the *Hamilton operator*  $H$ ,

$$H\Psi := \sum_{i=1}^N \left[ -\frac{1}{2} \Delta_i - \sum_{j=1}^M \frac{Z_j}{|\mathbf{x}_i - \mathbf{R}_j|} + \frac{1}{2} \sum_{j \neq i}^N \frac{1}{|\mathbf{x}_i - \mathbf{x}_j|} \right] \Psi = \left[ -\frac{1}{2} \Delta + V \right] \Psi = E\Psi,$$

where the eigenvalue  $E \in \mathbb{R}$  is the energy of the corresponding state  $\Psi$ . The most interesting quantity is the *ground state energy*  $E_0$ , which is the lowest eigenvalue.

Since the Hamilton operator is a linear second order differential operator, the analysis for the electronic Schrödinger equation has been already established to certain extent. We would like to briefly summarize some basic results and refer to e.g. [29], the survey article [30] and the extensive literature cited therein. The Sobolev spaces  $H^m := H^m((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N)$ ,  $m \in \mathbb{N}_0$  are defined as the function spaces for which all derivatives up to order  $m$  are in  $H^0 := L_2((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N)$ . Consequently, the operator  $H$  maps the Sobolev space  $H^1$  continuously into its dual space  $H^{-1}$ , i.e.  $H : H^1 \rightarrow H^{-1}$  boundedly [16, 29, 34]. The potential operator  $V$  maps the Sobolev spaces  $H^1$  continuously into  $H^0$  i.e.  $V : H^1 \rightarrow H^0 = L_2$  boundedly [29, 34]. The electronic Schrödinger operator admits a rather complicated spectrum. We are interested mainly in the ground state energy  $E_0$ . If  $\sum_{j=1}^M Z_j \geq N$ , in particular for electrical neutral systems, it is known that  $E_0$  is an eigenvalue of finite multiplicity of the operator  $H : H^2 \rightarrow H^0$  below the essential spectrum  $\sigma_{ess}(H)$  of  $H$ , i.e.  $-\infty < E_0 < \inf \sigma_{ess}(H)$ . see e.g. [29, 30] for further details. This situation will be considered in the sequel. For sake of simplicity, we will also always assume in the sequel that  $E_0$  is a simple eigenvalue. Without loss of generality, we assume further that the wave function is real valued. According to the well known mini-max principle, see e.g. [29], the ground state energy and the corresponding wave function satisfies the Rayleigh Ritz variational principal, see [34, 29], i.e. the lowest eigenvalue is the minimum of the Rayleigh quotient

$$E_0 = \min_{\Psi \neq 0} \frac{\langle H\Psi, \Psi \rangle}{\langle \Psi, \Psi \rangle}, \quad \Psi = \operatorname{argmin}_{\Psi \neq 0} \frac{\langle H\Psi, \Psi \rangle}{\langle \Psi, \Psi \rangle}.$$

In this variational framework, the energy space for the electronic Schrödinger equation is

$$\mathcal{V} = H^1((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N) \cap \bigwedge_{i=1}^N L_2(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}) \tag{1}$$

supplied with the  $H^1$  norm  $\|\Psi\|_{\mathcal{V}} := \|\Psi\|_{H^1}$ . Here  $\wedge$  denotes the outer or anti-symmetric product. Since the Hamilton operator maps  $H : \mathcal{V} \rightarrow \mathcal{V}'$  boundedly, we will put the eigenvalue

problem into the following weak formulation, see e.g. [34], to find  $\Psi \in \mathcal{V}$ , normalized by  $\langle \Psi, \Psi \rangle = 1$  satisfying

$$\langle \Phi, (H - E_0)\Psi \rangle = 0 , \forall \Phi \in \mathcal{V} . \quad (2)$$

The above framework will be considered throughout the present paper.

## 3 Full Configuration Interaction Approximation

### 3.1 Approximation by Slater-determinants

A convenient way to approximate the wave function is to use a tensor product approximation, i.e. to apply separation of variables w.r.t. to individual particle variables  $(\mathbf{x}_i, s_i)$ . Due to the antisymmetry constraint the usual tensor product has to be replaced by an outer (resp. antisymmetric) product, which can be realized by determinants. To this end, let us consider functions  $\varphi_i : (\mathbf{x}, s) \mapsto \varphi_i(\mathbf{x}, s) \in \mathcal{X} := H^1(\mathbb{R}^3 \times \{\pm \frac{1}{2}\})$ , which will be called spin orbitals, because they depend on one spin variable  $s = \pm \frac{1}{2}$  and the spatial variable  $\mathbf{x} \in \mathbb{R}^3$ . In the sequel we will confine ourselves to spin orbital formulations. We build Slater determinants  $\Psi_{SL}$  in the energy space of an  $N$ -electron system by selecting  $N$  nonidentical indices, say  $i = 1, \dots, N$ , together with  $N$  ortho-normal spin orbitals  $\varphi_i$ , i.e.

$$\langle \varphi_i, \varphi_j \rangle := \sum_{s=\pm \frac{1}{2}} \int_{\mathbb{R}^3} \varphi_i(\mathbf{x}, s) \varphi_j(\mathbf{x}, s) d\mathbf{x} = \delta_{i,j} \quad (3)$$

to define the *Slater determinant*

$$\Psi_{SL}[\nu_1, \dots, \nu_N](\mathbf{x}_1, s_1; \dots; \mathbf{x}_N, s_N) = \frac{1}{\sqrt{N!}} \det(\varphi_{\nu_i}(\mathbf{x}_j, s_j))_{i,j=1}^N. \quad (4)$$

For simplicity of notation, we will use the same brackets  $\langle \cdot, \cdot \rangle$  for designating inner products in Hilbert spaces, independent of the underlying Hilbert space. To avoid confusion, the concrete definition of the inner product will be highlighted if it is not obviously identified from the context. Due to the antisymmetry, the sign of the above Slater determinant depends on the order of the indices. We will say, that the indices  $\nu_1, \dots, \nu_N$  are in *canonical order* if  $\nu_i < \nu_{i+1}$ ,  $i = 1, \dots, N$ ; i.e., the indices are ordered monotonically.

### 3.2 The Full CI Ansatz and the Galerkin-Ritz Approximation

For numerical computations, we consider finite dimensional subspaces. Natural candidates to built finite dimensional subspaces of  $\mathcal{V}$  are those spaces spanned by Slater determinants of functions from a basis set of  $\mathcal{X}_K \subset \mathcal{X} := H^1(\mathbb{R}^3 \times \{\pm \frac{1}{2}\})$ .

Taking  $K$  with  $N \leq K$  and  $\varphi_i \in H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$ ,  $i = 1, \dots, K$  as an ( $L_2$ -) ortho-normal system of spin orbitals, we define the  $K$ -dimensional subspace of  $\mathcal{X}$  by

$$\mathcal{X}_K := \text{span}\{\varphi_i : i = 1, \dots, K\} \subset \mathcal{X} = H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}). \quad (5)$$

To ensure convergence, we suppose the family  $\{\mathcal{X}_K\}_{K \in \mathbb{N}}$  to be dense in  $X = L_2(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$  as well as in  $X = \mathcal{X} = H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$ , i.e. for each  $\epsilon > 0$  and  $\phi \in \mathcal{X}$  there is  $K_0$  such that for all  $K > K_0$  there exists  $\phi_K \in \mathcal{X}_K$  with  $\|\phi - \phi_K\|_X < \epsilon$ . We define the *full configuration interaction space* or *FCI space* as the finite dimensional space

$$\begin{aligned} \mathcal{V}_K &:= \text{span}\{\Psi_{SL}[\nu_1, \dots, \nu_N] : \nu : \{1, \dots, N\} \rightarrow \{1, \dots, K\}, i \mapsto \nu(i) := \nu_i \text{ is monotone}\} \\ &= \text{span}\{\Psi_\mu : \mu \in \mathcal{I} = \mathcal{I}_K\} = \bigwedge_{i=1}^N \mathcal{X}_K. \end{aligned} \quad (6)$$

Let  $\nu_1 < \nu_2 < \dots < \nu_N$  be a monotonously increasing selection of indices  $1 \leq \nu_i \leq K$ ,  $1 \leq i \leq N$ . The Slater determinants built by spin orbitals  $\varphi_i \in \mathcal{X}_K$  of the type  $\Psi[\nu_1 \dots \nu_N]$  form a complete ortho-normal basis of  $\mathcal{V}_K$ . The dimension of  $\mathcal{V}_K$  is

$$\dim \mathcal{V}_K = \binom{K}{N} = \frac{K!}{N!(K-N)!} \sim \mathcal{O}(K^N), \quad K \rightarrow \infty.$$

Since  $\{\mathcal{X}_K\}_{K \in \mathbb{N}}$  is a dense family in  $H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$ ,  $\{\mathcal{V}_K\}_{K \in \mathbb{N}}$  is a dense family in  $\mathcal{V}$ .

To obtain an approximate solution of the electronic Schrödinger equation, one may apply the Ritz-Galerkin method using subspaces  $\mathcal{V}_K \subset \mathcal{V}$ . I.e.

$$E_{0,K} := \min\{\langle \Psi, H\Psi \rangle : \langle \Psi, \Psi \rangle = 1, \Psi \in \mathcal{V}_K\}. \quad (7)$$

The CI wave -function  $\Psi_K \in \mathcal{V}_K$  is the solution of the Galerkin scheme

$$\Psi_K := \text{argmin}\{\langle \Psi, H\Psi \rangle : \langle \Psi, \Psi \rangle = 1, \Psi \in \mathcal{V}_K\}. \quad (8)$$

From the above definitions it becomes obvious, that the approximate ground state energy obtained by the Galerkin method provides an upper bound for the exact energy value  $E_0 \leq E_{0,K}$ . The convergence theory of the Galerkin scheme for the numerical solution of eigenvalue problems is already established. Here we recall the following known result for a dense family of finite dimensional subspaces,  $\{\mathcal{V}_K\}_{K \in \mathbb{N}}$ , see e.g. [34].

**Theorem 3.1.** *Let and  $E_0$  be the lowest eigenvalue of  $H$ , having multiplicity 1, together with a corresponding eigenfunction  $\Psi$ , normalized by  $\langle \Psi, \Psi \rangle = 1$  and  $K_0$  sufficiently large. Then, for all  $K > K_0$ , there exist functions  $\Psi_K \in \mathcal{V}_K$  corresponding to a single lowest eigenvalue  $E_{0,K}$  solving the discrete Galerkin scheme given by (7) and normalized as in (8) which satisfy the following error bounds:*

$$\|\Psi - \Psi_K\|_{\mathcal{V}} \leq C \inf_{\phi_K \in \mathcal{V}_K} \|\Psi - \phi_K\|_{\mathcal{V}} \quad (9)$$

$$0 \leq E_{0,K} - E_0 \leq C \|\Psi - \Psi_K\|_{\mathcal{V}}^2. \quad (10)$$

This result states that the Galerkin method provides an approximate eigenfunction  $\Psi_K$  which approximates the exact eigenfunction quasi-optimally. Moreover, the eigenvalue converges quadratically compared with the convergence of the eigenfunction, which is sometimes referred as superconvergence of the eigenvalues. The eigenvalue  $E_{0,K}$  is also non-degenerate under the above assumption that the approximation space is sufficiently good, i.e.  $K_0 < K$ . The Galerkin scheme using all Slater determinants built by orbitals from  $\mathcal{X}_K$ , i.e.  $\mathcal{V}_K$ , is called the *Full Configuration Interaction Method* or shortly *Full CI Method*. Since  $\dim \mathcal{V}_K \sim K^N \sim \mathcal{O}(K^N)$ ,  $K \rightarrow \infty$ , this approach scales exponentially with respect to  $K$  as well as to the number of electrons  $N$ . Let us remark, that in order to resolve the electron-electron cusp with accuracy  $\epsilon$ , the number  $K$  has to be sufficiently large, one expects  $K \sim \epsilon^{-\alpha}$ , for some  $\alpha > 1$ . Therefore, for larger systems, e.g molecules etc., this approach is practically not feasible, except for a very small number of electrons. In the sequel, we will consider the *Projected Coupled Cluster Method* as an approximation of the full CI wave function  $\Psi_K$  along a low dimensional nonlinear manifold in the space  $\mathcal{V}_K$ . However, we will pay attention that the constants involved in convergence estimates etc. should be uniform with respect to the parameter  $K_0 \leq K$ .

### 3.3 Excitation Operators

To built Slater determinants  $\Psi_{SL}$  of an  $N$ -electron system, one has to select  $N$  nonidentical indices  $\nu_i$ ,  $i = 1, \dots, N$  and basis functions  $\varphi_{\nu_i} \in \mathcal{X}_K$ ,

$$\Psi_{SL}[\nu_1, \dots, \nu_N](\mathbf{x}_1, s_1, \dots, \mathbf{x}_N, s_N) = \frac{1}{\sqrt{N!}} \det(\varphi_{\nu_i}(\mathbf{x}_j, s_j))_{i,j=1}^N.$$

Starting with a single *reference Slater determinant*, e.g.

$$\Psi_0 = \Psi[1, \dots, N] = \frac{1}{\sqrt{N!}} \det(\varphi_i(\mathbf{x}_j, s_j))_{i,j=1}^N \in \mathcal{V}_K,$$

we will call the first  $N$  orbitals  $\varphi_i \in \mathcal{X}_K$ ,  $i = 1, \dots, N$ , *occupied orbitals* and the remaining  $K - N$  orbitals will be called *unoccupied orbitals*.

We can build further Slater determinants

$$\Psi_j^a := \Psi[1, \dots, j-1, a, j+1, \dots, N] = (-1)^{N-j} \Psi[1, \dots, j-1, j+1, \dots, N, a]$$

by replacing, for instance, the orbital  $\varphi_j$ ,  $j \leq N$ , by another orbital  $\varphi_a$ ,  $a > N$ .  $\Psi_j^a$  will be called a *single excited Slater determinant*. The expressions *unoccupied orbital* and *excitations* are standard terms in quantum chemistry literature and have been motivated by Hartree-Fock theory.

Analogously, if we replace two orbitals  $\varphi_j$  and  $\varphi_k$  by  $\varphi_a$  and  $\varphi_b$ , we obtain the Slater determinant

$$\Psi_{j,k}^{a,b} = \Psi[1, \dots, a, \dots, b, \dots, N].$$

If  $\Psi_{SL} = \Psi_0$  then  $\Psi_{j,k}^{a,b}$  will be called *doubly excited*. Higher excited determinants are defined analogously, e.g.  $\Psi_{j,k,l}^{a,b,c}$ (triple) or  $\Psi_{j,k,l,m}^{a,b,c,d}$  etc.. By this procedure all Slater determinants  $\Psi_\mu = \Psi[\nu_1, \dots, \nu_N] \in \mathcal{V}_K$  can be expressed by excitations of a single reference Slater determinant  $\Psi_0 \in \mathcal{V}_K$ . In the sequel we will use the convention that the sign of the Slater determinant  $\Psi_\nu$  is determined by assuming that for  $\nu \in \mathcal{I}_K$  the indices are in canonical order.

Alternatively, if  $\Psi_\mu$  is  $k$ -times excited each index  $\mu \in \mathcal{I}_K$  can be exactly characterized by the set indices  $\{l_1, \dots, l_k\}$  of *occupied orbital functions* which are removed from the reference determinant  $\Psi_0$ , and therefore no longer contained in  $\Psi_\mu$ , together with the set of indices  $\{a_1, \dots, a_k\}$  of *unoccupied orbital functions* which are replacing them in the Slater determinant  $\Psi_\mu$ . We will use the following notation

$$\mu \simeq \begin{pmatrix} a_{\mu_1}, & \dots, & a_{\mu_k} \\ l_{\mu_1}, & \dots & l_{\mu_k} \end{pmatrix} \in \mathcal{J}. \quad (11)$$

Here  $\mathcal{J}$  denotes the index set of a Slater determinant basis of  $\mathcal{V}_K$ , except the one for the reference determinant  $\Psi_0$ . For sake of brevity of notation, we have dropped the subscript  $K$  for  $\mathcal{J} = \mathcal{J}_K$ .

**Definition 3.2.** For  $\mu \in \mathcal{J}$ , we can define an excitation operator  $X_\mu = X_{l_1, \dots, l_k}^{a_1, \dots, a_k} : \mathcal{V}_K \rightarrow \mathcal{V}_K$  such that  $X_\mu \Psi_0 = \Psi_\mu$  by its action on Slater determinants. Let  $\mu \in \mathcal{J}$  as defined above, then according to the above introduced notation

$$\Psi_\mu = X_\mu \Psi_0 = X_{l_1, \dots, l_k}^{a_1, \dots, a_k} \Psi_0 = (-1)^{p(\mu)} \Psi_{l_1, \dots, l_k}^{a_1, \dots, a_k},$$

with an appropriate phase factor  $(-1)^{p(\mu)} = \pm 1$ . By this, we have established an obvious one-to-one correspondence between Slater determinants  $\Psi_\mu$  and excitation operators  $X_\mu$ . However, the action of the excitation operators  $X_\mu$  can be extended such that it is well defined for arbitrary Slater determinant  $\Psi[\nu_1, \dots, \nu_N] \in \mathcal{V}_K$ . If  $\Psi \in \mathcal{V}_K$  is a Slater determinant containing all spin orbitals  $\varphi_{l_i}$ ,  $i = 1, \dots, k$ , then  $X_\mu \Psi = X_{l_1, \dots, l_k}^{a_1, \dots, a_k} \Psi$  is defined by replacing the orbitals  $\varphi_{l_i}$ ,  $i = 1, \dots, l_k$  by the orbitals  $\varphi_{a_1}, \dots, \varphi_{a_k}$ . If an occupied spin orbital  $\varphi_{l_i}$ ,  $i = 1, \dots, k$  is not contained in the Slater determinant  $\Psi$ , then we simply define  $X_\mu \Psi := 0$ . We also note that if  $\Psi$  already contains an unoccupied orbital  $\varphi_{a_i}$ ,  $i = 1, \dots, k$ , then  $X_\mu \Psi = 0$  due to the fact that a Slater determinant which contains two identical orbitals vanishes.

From this definiton, we easily conclude the following proposition.

**Proposition 3.3.** For  $\mu, \nu \in \mathcal{J}$  there holds  $[X_\mu, X_\nu] = X_\mu X_\nu - X_\nu X_\mu = 0$ . The adjoint operator  $X_\mu^\dagger$  defined by  $\langle X_\mu^\dagger \Phi, \Psi \rangle = \langle \Phi, X_\mu \Psi \rangle$ ,  $\forall \Phi, \Psi$ , is a de-excitation operator  $X_\mu^\dagger = X_{a_1, \dots, a_k}^{l_1, \dots, l_k}$ .

**Proposition 3.4.** Let  $\Psi_0 = \Psi[1, \dots, N] \in \mathcal{V}_K$  be a single reference Slater-determinant and  $\Psi_1$  be an arbitrary function in the full CI space  $\mathcal{V}_K$  satisfying the intermediate normalization  $\langle \Psi, \Psi_0 \rangle = 1$ . Then there exists exactly one excitation operator of the form

$$T = I + \sum_{\mu \in \mathcal{J}} c_\mu X_\mu \quad \text{such that } \Psi = T \Psi_0.$$

The coefficients  $c_\mu = \langle \Psi, \Psi_\mu \rangle$  are called the *excitation amplitudes*. Let  $\mu, \nu \in \mathcal{J}$ , then the *excitation rank*  $\#\mu$  will be defined by the numbers of unoccupied orbitals contained in the Slater determinant  $\Psi_\mu$ . We can decompose the operator  $T = I + \sum_{k=1}^N T_k$  according to its excitation rank into a part containing single excitation operators  $T_1 = \sum_{l,a} t_l^a X_l^a$ , another part with double excitation operators  $T_2 = \sum_{l,k;a,b} t_{l,k}^{a,b} X_{l,k}^{a,b}$ , the triple excitation operators  $T_3$  and so on.

### 3.4 Slater-Condon Rules

If Slater determinants are used as basis functions in  $\mathcal{V}_K$ , the matrix coefficients of the Galerkin matrix  $\langle H\Psi_\mu, \Psi_\nu \rangle$  can be computed explicitly. The resulting formulas are known as Slater-Condon rules and can be found in textbooks [31, 14]. Due to their fundamental importance, we summarize these results briefly. To this end, it is convenient to decompose the Hamilton operator of the electronic Schrödinger equation into *single particle* and *two particle operators*  $H =: \sum_{i=1}^N (\mathbf{h}_i + \sum_{j < i} \mathbf{G}_{i,j}) = \mathbf{h} + G$ , where  $\mathbf{h}_i = -\frac{1}{2}\Delta_i + V_{core}(\mathbf{x}_i) = -\frac{1}{2}\Delta_i + \sum_{j=1}^M \frac{-Z_j}{|\mathbf{R}_j - \mathbf{x}_i|}$  and  $\mathbf{h} = \sum_{i=1}^N \mathbf{h}_i$  are *single particle operators* (notice that all individual one particle operators are identical, which is consistent with the Pauli principle), and  $G = \sum_{i=1}^N \sum_{j < i} \mathbf{G}_{i,j}$  represents the potential of the repulsion between two electrons, which consists of *two particle operators*  $\mathbf{G}_{i,j} = \frac{1}{|\mathbf{x}_i - \mathbf{x}_j|}$ . For brevity of notation, we define the following expressions for single particle operators

$$h_{i,j} := \langle i | \mathbf{h} | j \rangle := \langle \varphi_i, \mathbf{h}_j \varphi_j \rangle$$

and for the two particle repulsion term

$$\langle a, b | i, j \rangle = \sum_{s,s'=\pm\frac{1}{2}} \int \int \frac{\varphi_a^*(\mathbf{x}, s) \varphi_b^*(\mathbf{x}', s') \varphi_i(\mathbf{x}, s) \varphi_j(\mathbf{x}', s')}{|\mathbf{x} - \mathbf{x}'|} d\mathbf{x} d\mathbf{x}'$$

together with

$$\langle a, b || i, j \rangle := \langle a, b | i, j \rangle - \langle a, b | j, i \rangle. \quad (12)$$

**Theorem 3.5.** Let  $\Psi_{SL}^1[1, \dots, N]$  and  $\Psi_{SL}^2[\mu_{i_1}, \dots, \mu_{i_N}]$  be two Slater determinants built by the orbitals  $\varphi_i$  and  $\varphi_{\mu_i}$ , where  $i = 1, \dots, N$ .

For single particle operators, e.g.  $A := \mathbf{h} = \sum_{i=1}^N \mathbf{h}_i$  there holds

$$\langle \Psi^1, A \Psi^1 \rangle = \sum_{l=1}^N \langle l | \mathbf{h} | l \rangle \text{ if } \Psi^1 = \Psi^2, \quad \langle \Psi^1, A \Psi^2 \rangle = \langle l | \mathbf{h} | a \rangle \text{ if } \Psi^1 = X_j^a \Psi^2$$

and  $\langle \Psi^1, A \Psi^2 \rangle = 0$  otherwise, i.e. for higher excitations.

The two particle operator  $G = \sum_{i=1}^N \sum_{j>i}^N \frac{1}{|\mathbf{x}_i - \mathbf{x}_j|}$  satisfies

$$\begin{aligned}\langle \Psi^1, G\Psi^1 \rangle &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \langle i, j | i, j \rangle , \\ \langle \Psi^1, G\Psi^2 \rangle &= \sum_{l=1}^N \langle j, l | a, l \rangle \text{ if } \Psi^2 = X_j^a \Psi^1 , \\ \langle \Psi^1, G\Psi^2 \rangle &= \langle i, j | a, b \rangle \text{ if } \Psi^2 = X_{i,j}^{a,b} \Psi^1 ,\end{aligned}$$

and  $\langle \Psi^1, G\Psi^2 \rangle = 0$  for higher excitations, e.g.  $\Psi^2 = X_{i,j,l}^{a,b,c} \Psi^1$ .

### 3.5 The Hartree-Fock approximation and the Fock operator

The discrete Hartree-Fock approximation is defined by an approximation of  $\Psi \in \mathcal{V}$ , resp.  $\Psi_K \in \mathcal{V}_K$  by a single Slater determinant, i.e. by optimizing the energy functional

$$E_{HF} = \min \{ \langle H\Psi_{SL} \Psi_{SL} \rangle : \Psi_{SL} \in \mathcal{V}_K \text{ is a single Slater determinant} \} \quad (13)$$

as is a functional of  $N$  unknown orbital function  $\varphi_i \in \mathcal{X}_K$ ,  $i = 1, \dots, N$ , subordinated to orthogonality constraints  $\langle \varphi_i, \varphi_j \rangle = \delta_{i,j}$ . Let us introduce the functions

$$\rho(\mathbf{x}, s, \mathbf{y}, s') := \sum_{j=1}^N \varphi_j(\mathbf{x}, s) \varphi_j(\mathbf{y}, s') , \quad n(\mathbf{y}) := \sum_{s'=\pm\frac{1}{2}} \rho(\mathbf{y}, s', \mathbf{y}, s') , \quad V_H(\mathbf{x}) = \int_{\mathbb{R}^3} \frac{n(\mathbf{y})}{|\mathbf{x} - \mathbf{y}|} d\mathbf{y} . \quad (14)$$

With this definition at hand one can define the corresponding Fock operator  $F : \mathcal{X} \rightarrow \mathcal{X}'$  defined by

$$Fu(\mathbf{x}, s) := \left( \frac{1}{2} \Delta - \sum_{k=1}^M \frac{Z_k}{|\mathbf{x} - \mathbf{R}_k|} - V_H(\mathbf{x}) \right) u(\mathbf{x}, s) + \sum_{s'=\pm\frac{1}{2}} \int_{\mathbb{R}^3} \frac{\rho(\mathbf{x}, s, \mathbf{y}, s')}{|\mathbf{x} - \mathbf{y}|} u(\mathbf{y}, s') d\mathbf{y} .$$

Note that the Fock operator  $F = F_\rho$  depends on the minimizer  $\{\varphi_i : 1 \leq i \leq N\}$ . If  $\sum_{j=1}^M Z_j \geq N$ , it is known [25] that there exists a minimizer of (13) containing orbitals  $\varphi_i$ ,  $i = 1, \dots, N$ , which obey the Hartree-Fock equations (in the spin orbital formulation) as a nonlinear eigenvalue problem in weak formulation

$$\langle F\varphi_i - \lambda_i \varphi_i, \phi \rangle = 0 , \quad \forall \phi \in \mathcal{X}_K , \quad (15)$$

where the *Fock* operator (or Hamilton-Fock operator)  $F$  depends on the solution  $\varphi_i$ ,  $i = 1, \dots, K$  itself. Here the orbitals  $\varphi_i$ ,  $i = 1, \dots, N$ , are chosen according to the *aufbau principle*, i.e. they are eigenfunction with respect to the  $N$  lowest eigenvalues  $\lambda_1 \leq \dots \leq \lambda_N < \lambda_{N+1} \leq \dots \leq \lambda_K$ .

Since the eigenfunctions of the discrete Fock operator  $\varphi_i$ ,  $i = 1, \dots, K$  are forming an orthonormal basis in  $\mathcal{X}_K$ , they can be used as a basis of  $\mathcal{X}_K$  for further computations like the *CI Method* or the *Møller Plesset Perturbation Methods* [31, 14] or the *Coupled Cluster Method*. These eigenfunctions of the Fock operator are called *canonical orbitals*. Usually wave function methods are based on the above mentioned Hartree-Fock computation, i.e. we will take the Hartree-Fock determinant  $\Psi_{HF} = \frac{1}{\sqrt{N!}} \det(\varphi_i(\mathbf{x}_j, s_j))_{i,j=1}^N$  as the reference determinant  $\Psi_0$ , and occupied resp. unoccupied orbitals are defined by this choice. Actually, in most Coupled Cluster computations the closed shell restricted Hartree-Fock approximation is used for the reference determinant. However, we will not go into further details about specific treatments of the spin variables. Since the Fock operator  $F_i = F : H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}) \rightarrow H^{-1}(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$  acts only on functions depending on one particle variable  $\mathbf{x}_i, s_i$ , i.e. three spatial variables and one spin variable, it defines a *single-particle operator*. Therefore, we can define also a multi-dimensional operator acting in the configuration space  $\mathcal{V}$  or  $\mathcal{V}_K$  by  $F\Psi := \sum_{i=1}^N F_i \Psi$ . Here the Fock operators  $F_i = F$ ,  $i = 1, \dots, N$ , act only on single particle variable  $(\mathbf{x}_i, s_i) \in \mathbb{R}^3 \times \{\pm\frac{1}{2}\}$  as defined by (15).

**Definition 3.6.** *We decompose the (discretized) Hamiltonian*

$$H = F + U,$$

where  $F$  is the Fock operator and  $U := H - F$  denotes the fluctuation potential.

We remind that the Fock operator  $F$  is a one particle operator and the fluctuation potential  $U$  contains the electron-electron repulsion terms. Computations are rather simplified if we use *canonical orbitals* as basis functions in  $\mathcal{X}_K$ , since they are eigenfunctions of the corresponding (discretized) Fock operator  $F : \mathcal{X}_K \rightarrow \mathcal{X}'_K$  in the variational (weak) sense, i.e. a solution of the Galerkin scheme. Applying Theorem 3.5, one immediately obtains the following result.

**Proposition 3.7.** *The Hartree-Fock Slater determinant  $\Psi_0 = \Psi_{HF}$  given by the solutions of the (discrete) Hartree-Fock equations satisfying the aufbau principle, is an eigenfunction of the operator  $F : \mathcal{V}_K \rightarrow \mathcal{V}'_K$  corresponding to the lowest eigenvalue  $\Lambda_0 = \sum_{i=1}^N \lambda_i$  in the variational sense, i.e.*

$$\langle (F - \Lambda_0) \Psi_0, \Phi \rangle = 0 \quad \forall \Phi \in \mathcal{V}_K.$$

Moreover, for  $\mu \in \mathcal{J}$ , an excited Slater determinant  $\Psi_\mu = X_\mu \Psi_0 = \Psi[\mu_1, \dots, \mu_N]$  is also an eigenfunction of  $F : \mathcal{V}_K \rightarrow \mathcal{V}'_K$  with eigenvalue

$$\Lambda_\mu = \sum_{j=1}^N \lambda_{\mu_j} = \Lambda_0 + \epsilon_\mu, \quad \Lambda_0 = \sum_{j=1}^N \lambda_j, \quad \epsilon_\mu := \sum_{j=1}^k (\lambda_{a_j} - \lambda_{l_j}).$$

The following result, known as *Brillouin's Theorem* is presented without proof.

**Proposition 3.8.** *The interaction of the Hamiltonian  $H$  between the Hartree-Fock determinant  $\Psi_{HF} = \Psi_0$  and an arbitrary singly excited orbital  $\Psi^{Single} = \Psi_j^a$ , where  $j = 1, \dots, N$  refers to an occupied orbital and  $a > N$  to an unoccupied orbital, is zero,*

$$\langle \Psi_j^a, H \Psi_{HF} \rangle = \langle \Psi_j^a, U \Psi_{HF} \rangle = 0. \tag{16}$$

It has been shown that there is a positive HOMO (highest occupied molecular orbital) -LUMO (lowest un-occupied molecular orbital) gap  $\lambda_{N+1} - \lambda_N > 0$ , due to a result in [2]. The Hartree-Fock equations together with the aufbau principle are known to be necessary conditions for a minimizer. It is not clear whether they are sufficient. It is also known by [26] that the Fock operator has countably many negative eigenvalues  $\lambda_i$ , it admits also a continuous spectrum and the eigenfunctions are not forming a complete basis in  $\mathcal{V}$ . In view of these principal difficulties, we confine ourselves to finite dimensional subspaces  $\mathcal{X}_K \subset \mathcal{X}, \mathcal{V}_K \subset \mathcal{V}$  and define the excitation operators only in discrete spaces  $\mathcal{V}_K$ . This means, in the sequel, we will consider only the discrete Hartree-Fock approximation, i.e.  $\varphi_i \in \mathcal{X}_K$ , the above occupied orbitals are solutions of the Hartree Fock equation satisfying the aufbau principle, which are called aufbau solutions. Therefore they satisfy the necessary conditions to be a minimizer of the corresponding energy functional in  $\mathcal{X}_K$ . It is not clear yet, whether the discrete aufbau solutions are converging to a minimizer of the continuous Hartree-Fock model or not. In order to obtain convergence results of the Projected Coupled Cluster Method, we will see later that we do not need the convergence of the Hartree-Fock determinant if  $K \rightarrow \infty$  explicitly. Instead of convergence, we require some uniformity conditions, e.g.  $\|\Psi_0\|_{H^1} \leq C$  uniformly w.r.t.  $\Psi_0 \in \mathcal{V}_K, K \rightarrow \infty$ . This is a reasonable condition and satisfied in case of convergence.

It is remarkable, that in physics literature one does not distinguish between the Hamilton operator  $H : \mathcal{V} \rightarrow \mathcal{V}'$  and its finite dimensional counterpart  $\mathcal{H} : \mathcal{V}_K \rightarrow \mathcal{V}_K$ . Even in the sense of the canonical embedding  $\mathcal{V}_K \subset \mathcal{V}$  they are not the same operators nor converging to each other in the sense of operator norms. They are the same only in the weak formulation on  $\mathcal{V}_K$ , i.e. in the Galerkin scheme. This is sometimes misleading or erranous, and requires a careful treatment the concrete setting.

## 4 Coupled Cluster Approximation

Since the dimension of the full CI space is too large, it is common practise to apply the Galerkin method to subspaces, which are defined via their excitation ranks, eg.  $\mathcal{V}_{SD}, \mathcal{V}_{SDT}$ , for single, double, triples excitations and so on. For example, the CI method using only singly and doubly excited Slater determinants is called *CISD*. Besides the superior qualities of the variational Ritz-Galerkin approach (see Theorem 3.1), all kind of truncated CI methods defined above are violating the *size consistency*, meaning the following. If one consider  $M$  independent electronic systems, then the total energy is the sum over all energies of the individual systems. Moreover, the wave function of the total system is the outer (antisymmetric) product of the wave functions of the individual system. For instance, let  $A, B$  be to independent electronic systems with  $N_A, N_B$  particles and ground state energy  $E_A, E_B$ . Further let  $\mathcal{V}_{K,A}, \mathcal{V}_{K,B}$  be corresponding full CI spaces. Then, for the combined system  $AB$ , the ansatz space  $\mathcal{V}_{K,AB} = \mathcal{V}_{K,A} \wedge \mathcal{V}_{K,B}$  yields the Galerkin solution  $\Psi_{AB} = \Psi_A \wedge \Psi_B$  with approximate ground state energy  $E_{AB,K} = E_{A,K} + E_{B,K}$ . However, for the doubly truncated CI method with trial spaces  $\mathcal{V}_{SD,A}, \mathcal{V}_{SD,B}$ , it is obvious that in general  $\mathcal{V}_{SD,AB} \neq \mathcal{V}_{SD,A} \wedge \mathcal{V}_{SD,B}$ . Approximating each individual system e.g. by *CISD*,

i.e. at most doubly excited Slater determinants, the corresponding approximate wave function of the total system is then fourfold excited. An approximation of this function by only at most doubly excited Slater determinants can be quite poor. This effect will become even worse for large systems. All these methods are underestimating the correlation energy significantly. In fact, *size consistency* is a major concern in electronic structure calculations supported by a rather long experience and multiple experimental data.

In the sequel, let us fix  $K \in \mathbb{N}$  and define the Coupled Cluster Method for the space  $\mathcal{V}_K \subset \mathcal{V}$  as an approximation to the full CI solution  $\Psi_K$ . Although the analysis of the Projected Coupled Cluster method will be performed on each particular  $\mathcal{V}_K$ , we will often drop the index  $K$  during these sections and write instead  $\mathcal{V}_{FCI}$ , for simplicity of notation. However we have to take in mind that in order to consider the convergence to the exact wave function, the constants involved in the convergence estimates should be uniform w.r.t.  $K \rightarrow \infty$ . For this reason, we will introduce discrete norms which represent the error between the full CI solution  $\Psi_K$  to the projected coupled cluster solution  $\Psi_{CC,K} \in \mathcal{V}_K$  in such a way that one obtains norm equivalences to the  $H^1$  Sobolev norm with constants uniform with respect to  $K \rightarrow \infty$ . We note that these constants may depend on the particular atomic configuration, namely  $N$ ,  $Z_j$  and  $\mathbf{R}_j$ .

## 4.1 The Exponential Ansatz

The idea of the Coupled Cluster method is to replace the linear parametrization in the Galerkin ansatz, respectively in the CI-method, by a nonlinear parametrization. Therefore, instead of searching an approximate solution in linear subspaces, we are looking for an approximation on a nonlinear manifold. Let  $\Psi_0 \in \mathcal{V}_K$  be a reference Slater determinant and, for  $\mu \in \mathcal{J} = \mathcal{J}_K$ ,  $X_\mu$  denote the corresponding excitation operator. In the present paper, we will consider the reference Slater determinant  $\Psi_0 = \Psi_{HF}$  to be the discrete Hartree-Fock determinant in the sequel, which is the canonical choice, mainly used in electronic structure codes.

**Lemma 4.1.** *1. Let  $\mu, \nu \in \mathcal{J}$  be given with  $\#\mu \geq \#\nu$ . Then there holds*

$$\langle \Psi_\nu, X_\alpha \Psi_\mu \rangle = 0, \quad \forall \alpha \in \mathcal{J},$$

*2. Let  $\mu, \nu \in \mathcal{J}$ ,  $\#\nu > \#\mu$ , be given, then there exists at most one  $\alpha \in \mathcal{J}$  such that there holds*

$$\langle \Psi_\nu, X_\alpha \Psi_\mu \rangle \neq 0,$$

*3. Let  $\nu \in \mathcal{J}$  be given, then*

$$\#\{(\alpha, \mu) \in \mathcal{J} \times \mathcal{J} : \langle \Psi_\nu, X_\alpha \Psi_\mu \rangle \neq 0\} = \sum_{\alpha, \mu \in \mathcal{J}} |\langle \Psi_\nu, X_\alpha \Psi_\mu \rangle| \leq C_N \leq 2^{2N}.$$

*Proof.* Since  $X_\alpha$  increases the excitation rank at least by one, the first assertion follows immediately from the orthogonality of the Slater determinants. Referring to the notation (11), the Slater

determinant is nonzero only if

$$\{a_{\mu_1}, \dots, a_{\mu_{\sharp\mu}}\} \subset \{a_{\nu_1}, \dots, a_{\nu_{\sharp\nu}}\} \text{ and } \{l_{\mu_1}, \dots, l_{\mu_{\sharp\mu}}\} \subset \{l_{\nu_1}, \dots, l_{\nu_{\sharp\nu}}\}.$$

In this case we set

$$\{a_{\alpha_1}, \dots, a_{\alpha_{\sharp\alpha}}\} = \{a_{\nu_1}, \dots, a_{\nu_{\sharp\nu}}\} \setminus \{a_{\mu_1}, \dots, a_{\mu_{\sharp\mu}}\}$$

and

$$\{l_{\alpha_1}, \dots, l_{\alpha_{\sharp\alpha}}\} = \{l_{\nu_1}, \dots, l_{\nu_{\sharp\nu}}\} \setminus \{l_{\mu_1}, \dots, l_{\mu_{\sharp\mu}}\}.$$

For  $A := \{a_{\alpha_1}, \dots, a_{\alpha_{\sharp\alpha}}\}$ ,  $B := \{a_{\mu_1}, \dots, a_{\mu_{\sharp\mu}}\} \subset C := \{a_{\nu_1}, \dots, a_{\nu_{\sharp\nu}}\}$ , there hold  $A \cap B = \emptyset$  and  $A \cup B = C$ . This suggests the notations  $\mu \leq \nu$  and  $\alpha =: \nu \ominus \mu$ , in case  $X_\nu = X_\alpha X_\mu$ . A very rough estimate for the number of all possibilities building a set  $C$  as a union of disjoint subsets  $A, B$  is provided by square of the number of subsets  $2^{\sharp\nu} \leq 2^N$ , which gives the required result.  $\square$

According to Proposition 3.4, for any state  $\Psi$ , with  $\langle \Psi, \Psi_0 \rangle = 1$ , there exists an operator  $T = \sum_{i=1}^N T_i = \sum_{\mu \in \mathcal{J}} t_\mu X_\mu$  with  $\Psi = (I + T)\Psi_0$ . By definition there holds  $t_\mu = \langle \Psi_\mu, T\Psi_0 \rangle$ . Applying  $T$  to any  $\Psi_\nu$ , if  $T\Psi_\nu \neq 0$ , there is some  $\mu \in \mathcal{J}$  with  $t_\mu \neq 0$ , and  $T$  increases the excitation order of  $\Psi_\nu$  at least by one. In the corresponding matrix representation

$$\mathbf{T} = (T_{\mu,\nu})_{\mu,\nu \in \mathcal{J}} \in \mathbb{R}^{\#\mathcal{J} \times \#\mathcal{J}}, \quad T_{\mu,\nu} = \langle \Psi_\mu, T\Psi_\nu \rangle, \quad \mu, \nu \in \mathcal{J},$$

we observe that

$$T_{\mu,\nu} = 0 \text{ if } \sharp\nu \geq \sharp\mu,$$

that is  $\mathbf{I} + \mathbf{T}$  is a lower triangular matrix with diagonal  $\mathbf{I}$ .

We introduce the following class of excitation operators mapping  $\mathcal{V}_K \rightarrow \mathcal{V}_K$

$$\mathcal{L} := \{t_0 I + T : T = \sum_{\mu \in \mathcal{J}} t_\mu X_\mu, t_0, t_\mu \in \mathbb{C}\}.$$

The following lemma shows that this is an involutory algebra.

**Lemma 4.2.** *The set of operators  $\mathcal{L}$  forms a closed commutative subalgebra in the algebra of linear operators  $\mathcal{V}_K \rightarrow \mathcal{V}_K$ . This algebra is closed under inversion, and the spectrum of  $L = I - T \in \mathcal{L}$  is  $\sigma(L) = \{1\}$ . Furthermore, the excitation operators  $T$  are nilpotent, there holds  $T^k = 0$  for all  $k > N$ .*

*Proof.* We first note that for  $\mu = \begin{pmatrix} a \\ l \end{pmatrix}$ ,  $\alpha = \begin{pmatrix} b \\ k \end{pmatrix}$ , there holds

$$X_\mu X_\alpha = X_\alpha X_\mu = X_\nu \neq 0$$

only if  $k \neq l$  and  $a \neq b$ , and then the excitation rank is increased at least by one since  $\#\nu = 2$ . Otherwise one obtains  $X_\mu X_\alpha = 0$ . Therefore, in general there exists either  $\nu \in \mathcal{J}$  such that

$$X_\mu X_\alpha = X_\alpha X_\mu = X_\nu$$

holds or the product of both operators is zero. This proves that the composition of two excitation operators  $T, T'$  is again an excitation operator. In other words, the class of excitation operators  $\mathcal{L}$  form an algebra.

It is also clear that the operators of form  $T = \sum_{\mu \in \mathcal{J}} t_\mu X_\mu$  cannot be invertible in  $\mathcal{L}$  since  $\mathcal{L}$  does not contain operators reducing the excitation rank. Additionally, since, for each excitation operator  $X_\mu$ , we have  $X_\mu^k = 0$  for  $k > N$ , there holds  $T^k = 0$ , for all  $k > N$ , i.e. the operators  $T$  are nilpotent. Let  $L = I - T \in \mathcal{L}$  and  $L^{-1} := \sum_{k=0}^{\infty} T^k = \sum_{k=0}^N T^k$ , we compute

$$L^{-1}L = LL^{-1} = I - T^{N+1} = I.$$

I.e. each operator of the form  $I - T$  is invertible and the inverse is contained in the algebra, which implies the remaining assertions.  $\square$

The following theorem establishes the exponential parametrization which is fundamental for the Coupled Cluster ansatz.

**Theorem 4.3.** *Let  $\Psi_0 \in \mathcal{V}_K$  be a (single) Slater determinant, e.g. the Hartree-Fock solution  $\Psi_0 = \Psi_{HF}$ , and  $\Psi \in \mathcal{V}_K$  satisfying  $\langle \Psi, \Psi_0 \rangle \neq 1$  there exists an excitation operator of the form  $T = \sum_{i=1}^N T_i = \sum_\mu t_\mu X_\mu \in \mathcal{L}$  such that*

$$\Psi = e^T \Psi_0. \quad (17)$$

*Proof.* Let  $\Psi \in \mathcal{V}_K$  normalized by the intermediate normalization  $\langle \Psi, \Psi_0 \rangle = 1$ . Then, according to Proposition 3.4, there exists an excitation operator  $C = \sum_{\mu \in \mathcal{J}} c_\mu X_\mu$  such that  $\Psi = \Psi_0 + \sum_\mu c_\mu \Psi_\mu = (I + C)\Psi_0$ .

The spectrum of the operator  $A := I + C \in \mathcal{L}$  is  $\sigma(A) = \{1\}$  by the previous lemma. We define the following operator  $T$  in  $\mathcal{L}$  by the Cauchy-Dunford calculus

$$T := \frac{1}{2\pi i} \oint_{\Gamma} (zI - A)^{-1} \log z \, dz,$$

where  $\Gamma$  denotes, for instance, a circle in the complex plain surrounding the spectrum  $\sigma(A)$ , which consist of a single point  $z_0 = 1$ . The logarithm is defined in a usual way as a complex analytic function on  $\mathbb{C} \setminus \{x \in \mathbb{C} : x \in \mathbb{R}, x \leq 0\}$  with imaginary part  $-\pi < \operatorname{Im} \log z \leq \pi$ . Moreover this operator satisfies  $e^T = A$ . Since due to Lemma 4.2, the algebra  $\mathcal{L}$  is closed under inversion and  $(zI - A)^{-1} = (z - 1)^{-1}(I - \frac{1}{z-1}C)^{-1}$ , the operator  $T \in \mathcal{L}$  is in the same algebra  $\mathcal{L}$  and satisfies  $e^T = A$ .

To show  $t_\mu \in \mathbb{R}$ , we expand  $T$  in a formal power series  $T = \log(I + C) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k-1} C^k$ . Observe that the series terminates after finitely many steps,  $\log(I + C) = \sum_{k=1}^N \frac{(-1)^{k-1}}{k-1} C^k \in \mathcal{L}$ , since  $A = I + C \in \mathcal{L}$  and  $C$  is nilpotent. From the calculus above it follows that  $e^T = A$ .  $\square$

The following basic results follow from the Slater Condon rules.

**Theorem 4.4.** *Let  $\Psi \in \mathcal{V}_K$  be a solution of the discretized Schrödinger equation*

$$\langle H\Psi - E\Psi, \Phi \rangle = 0, \quad \forall \Phi \in \mathcal{V}_K$$

and the reference Slater determinant  $\Psi_0 = \Psi_{HF} \in \mathcal{X}_K$  be the corresponding Hartree-Fock determinant. If the wave function  $\Psi$  is normalized by  $\langle \Psi, \Psi_0 \rangle = 1$ , for the operator  $T$  with  $\Psi = e^T \Psi_0$ , according to (17), there holds

$$E = \langle \Psi_0, H\Psi \rangle = \langle \Psi_0, H(I + T_2 + \frac{1}{2}T_1^2)\Psi_0 \rangle. \quad (18)$$

*Proof.* Since  $\Psi = e^T \Psi_0 = \sum_{k=0}^{\infty} \frac{1}{k!} T^k \Psi_0$  and  $T = T_1 + T_2 + \dots$ , this is an immediate consequence of the Slater Condon rule (3.5)  $\langle \Psi_0, H\Psi_\mu \rangle = 0$ ,  $\#\nu \geq 3$ , and Proposition 3.8.  $\square$

For a general reference determinant  $\Psi_0$ , there will be an additional term  $\langle \Psi_0, HT_1\Psi_0 \rangle$  which vanishes in case the Hartree-Fock determinant is taken to be the reference, due to Brillouin's theorem, i.e. Proposition 3.8.

Let us recall the well known *Baker-Campbell-Hausdorff formula*

$$e^{-T} A e^T = A + [A, T] + \frac{1}{2!} [[A, T], T] + \frac{1}{3!} [[[A, T], T], T] + \dots = A + \sum_{k=1}^{\infty} \frac{1}{k!} [A, T]_{(k)}, \quad (19)$$

where  $[A, T]_{(k)}$  denotes the  $k$ -th nested commutator. Since the electronic Hamilton operator is a two particle operator, it is known that the Baker Campell Hausdorff expansion of  $e^{-T} H e^T$  terminates. This observation, stated in the following lemma, is crucial for the practical use of the Coupled Cluster method.

**Lemma 4.5.** *For a two particle operator  $H$ , in particular the Hamiltonian  $H$ , the Campell-Baker Hausdorff expansion terminates*

$$\langle \Phi, e^{-T} H e^T \Psi_0 \rangle = \langle \Phi, \sum_{k=0}^4 \frac{1}{k!} [H, T]_{(k)} \Psi_0 \rangle \quad \forall \Phi \in \mathcal{V}_k. \quad (20)$$

For the Fock operator  $F$ , using canonical orbitals, there holds,

$$\langle \Phi, e^{-T} F e^T \Psi_0 \rangle = \langle \Phi, (F + [F, T]) \Psi_0 \rangle \quad \forall \Phi \in \mathcal{V}_k. \quad (21)$$

For a proof of this known, but nontrivial result, we refer to [14] pp. 660ff and 663ff.

**Proposition 4.6.** *Let  $F$  be the Fock operator and  $X_\mu = X_{l_1, \dots, l_k}^{a_1, \dots, a_k}$ , then the following commutator can be computed explicitly*

$$[F, X_\mu] = [F, X_{l_1, \dots, l_k}^{a_1, \dots, a_k}] = \left( \sum_{j=1}^k (\lambda_{a_j} - \lambda_{l_j}) \right) X_\mu =: \epsilon_\mu X_\mu.$$

Consequently, there holds  $[F, T] = \sum_{\mu \in \mathcal{J}} \epsilon_\mu t_\mu X_\mu$ .

**Proposition 4.7.** Let  $\mathcal{V}_K = \text{span}\{\Psi_0, \Psi_\mu : \mu \in \mathcal{J}\}$  and  $\Psi = e^T \Psi_0 \in \mathcal{V}_k$ , according to (17) satisfying  $\langle H\Psi, \Phi \rangle = \langle E\Psi, \Phi \rangle$  for all  $\Phi \in \mathcal{V}_K$  with the intermediate normalization  $\langle \Psi_0, \Psi \rangle = 1$ , then there holds for  $T = T_1 + T_2 + \dots$ ,

$$E = \langle \Psi_0, e^{-T} H e^T \Psi_0 \rangle = \langle \Psi_0, H(I + T_2 + \frac{1}{2} T_1^2) \Psi_0 \rangle \quad (22)$$

$$0 = \langle \Psi_\mu, e^{-T} H e^T \Psi_0 \rangle, \mu \in \mathcal{J}, \quad (23)$$

or alternatively the following set of equations

$$\begin{aligned} E &= \langle \Psi_0, H e^T \Psi_0 \rangle = \langle \Psi_0, H(I + T_2 + \frac{1}{2} T_1^2) \Psi_0 \rangle \\ E \langle \Psi_\mu, e^T \Psi_0 \rangle &= \langle \Psi_\mu, H e^T \Psi_0 \rangle, \mu \in \mathcal{J}. \end{aligned} \quad (24)$$

*Proof.* Since  $E = E_K$  is an (discrete) eigenvalue of  $H$  in the weak (Galerkin) sense and  $\Psi = \Psi_K$  the corresponding full CI solution, there holds

$$0 = \langle \Phi, H\Psi \rangle - \langle \Phi, E\Psi \rangle = \langle \Phi, H e^T \Psi_0 \rangle - E \langle \Phi, e^T \Psi_0 \rangle, \forall \Phi \in \mathcal{V}_K.$$

Using  $e^{-T} e^T = I$ , this implies

$$\langle \Phi, e^{-T} H e^T \Psi_0 \rangle - \langle \Phi, E\Psi_0 \rangle = 0, \forall \Phi \in \mathcal{V}_K.$$

Formula (24) and hence formula (22) have been shown in the proof of Theorem 4.4. The last equation follows directly from  $E \langle \Phi, \Psi \rangle = \langle \Phi, H\Psi \rangle = \langle \Phi, H e^T \Psi_0 \rangle$ , for all  $\Phi \in \mathcal{V}_K$ .  $\square$

**Remark 4.8.** The above results hold also in the continuous formulation  $\mathcal{V}_K = \mathcal{V}$ . This is because the results stated above are formulated in the weak sense for all space  $\mathcal{V}_K$ , which in the limit  $K \rightarrow \infty$  are defining a dense subset of  $\mathcal{V}$ . Let us remark, that the presentation in the quantum chemistry literature is sometimes misleading or erranous, and does not apply directly to the present setting. In the discrete case, as considered here,  $(E, \Psi) = (E_K, \Psi_K)$  is the full CI ground state solution, which is not an eigen-pair of the Hamiltonian, but a Galerkin approximation of it. The results here mostly are formulated for the finite dimensional Galerkin approximation. Many of them hold also in the infinite dimensional case. Only if we consider canonical orbitals as eigenfunctions of the Fock operator we must be careful. In the infinite dimensional case, the Fock operator is likely to admit also a continuous spectrum. Its eigenfunctions are not forming a complete ortho-normal basis. Therefore the arguments based on spectral representation of the Fock operator cannot be easily transferred to the infinite dimensional setting. An extensive treatment of the continuous formulation of the Coupled Cluster ansatz will be deferred to a forthcoming paper.

## 4.2 The Projected Coupled Cluster Method

For the vector of unknown *Coupled Cluster amplitudes*  $\mathbf{t} = (t_\nu)_{\nu \in \mathcal{J}} \in \mathbb{R}^{\#\mathcal{J}}$ , we obtain a function  $\mathbf{t} \mapsto \mathbf{f}(\mathbf{t}) = (f_\mu(\mathbf{t}))_{\mu \in \mathcal{J}} \in \mathbb{R}^{\#\mathcal{J}}$  defined by

$$f_\mu(\mathbf{t}) = \langle \Psi_\mu, e^{-T} H e^T \Psi_0 \rangle = \langle \Psi_\mu, e^{-\sum t_\nu X_\nu} H e^{\sum t_\nu X_\nu} \Psi_0 \rangle, \mu \in \mathcal{J}.$$

The *amplitude equations* (23), which are necessary conditions can be rewritten in the form

$$\mathbf{f}(\mathbf{t}) = \mathbf{0}, \text{ or } \langle \mathbf{v}, \mathbf{f}(\mathbf{t}) \rangle = 0, \forall \mathbf{v} \in \mathbb{R}^{\#\mathcal{J}}.$$

This formulation is preferred to (24), because the Baker Campell Haussdorff expansion terminates and leads to computational terms. By definition, the solution of the amplitude equations will reproduce the CI solution  $\Psi \in \mathcal{V}_K$  exactly. We note there are  $\#\mathcal{J}$  amplitude equations for  $\#\mathcal{J}$  unknown amplitude parameters. Since  $\#\mathcal{J} = \mathcal{O}(K^N)$  becomes prohibitively large for increasing  $N$ , we will approximate the Coupled Cluster solution by an ansatz which contains much less parameters  $L \ll \#\mathcal{J}$ , choosing an appropriate subset  $\mathcal{J}_h \subset \mathcal{J}$  of cardinality  $\#\mathcal{J}_h = L$ . Usually, but not necessarily, this truncation, i.e. the choice of  $\mathcal{J}_h$  is made according to the excitation rank, analogously to the truncated CI approximation. For example the CCSD (Coupled Cluster Single Double) ansatz consists in the choice  $\mathcal{J}_h = \{\mu \in \mathcal{J} : \#\mu \leq 2\}$ .

Let  $\mathcal{N} = \dim \mathcal{V}_K - 1 = \#\mathcal{J}$  the total number of amplitudes. We introduce the  $\mathcal{N}$ -dimensional parameter space  $V := \mathbb{R}^{\mathcal{N}}$  equipped with the norm

$$\|\mathbf{t}\|_V^2 := \sum_{\mu \in J} (\epsilon_\mu^{1/2} |t_\mu|)^2, \quad \mathbf{t} \in V, \quad (25)$$

together with the dual norm  $\|\mathbf{v}\|_{V'}^2 := \sum_{\mu \in J} (\epsilon_\mu^{-1/2} |v_\mu|)^2$ ,  $\mathbf{v} \in V'$ . Furthermore, for a subset  $\mathcal{J}_h \subset \mathcal{J}$ , we consider a subspace  $V_h \subset V$  consisting of vectors  $\mathbf{t}_h = (t_\mu)_{\mu \in \mathcal{J}}$  with  $t_\mu = 0$  if  $\mu \notin \mathcal{J}_h$ . Then  $\dim V_h = \#\mathcal{J}_h$ , which is usually chosen much smaller than  $\#\mathcal{J}$ .

The *Projected Coupled Cluster Method* consists in the ansatz

$$T = T(\mathbf{t}_h) = \sum_{k=1}^l T_k = \sum_{\mu \in \mathcal{J}_h} t_\mu X_\mu,$$

e.g. the CCSD method is given by  $T = T_1 + T_2 = T(\mathbf{t}_h)$ , where the unknown cluster amplitudes  $t_\mu$ ,  $\mu \in \mathcal{J}_h$  must satisfy the following (nonlinear) equations

$$0 = \langle \Psi_\mu, e^{-T} H e^T \Psi_0 \rangle =: f_\mu(\mathbf{t}), \mu \in \mathcal{J}_h. \quad (26)$$

In particular, these conditions can be cast into a weak formulation

$$\langle \mathbf{f}(\mathbf{t}_h), \mathbf{v}_h \rangle = 0, \forall \mathbf{v}_h \in V_h. \quad (27)$$

These are  $L = \#\mathcal{J}_h$  nonlinear equations for  $L$  unknown excitation amplitudes  $t_\mu$ ,  $\mu \in \mathcal{J}_h$ . The corresponding approximate Coupled Cluster energy is computed by  $E_h = \langle \Psi_0, H e^{T_h} \Psi_0 \rangle$  with  $T_h = T_h(\mathbf{t}_h) = \sum_{\mu \in \mathcal{J}_h} t_\mu X_\mu$  according to formula (18).

**Remark 4.9.** *In contrast to the CI method, which minimizes the energy expectation value over a linear subspace  $\mathcal{V}_h$ , the Projected Coupled Cluster Method in general does not provide the optimal energy expectation value which can be achieved on the nonlinear manifold*

$$E_{var} = \inf \left\{ \langle \Psi, H \Psi \rangle : \Psi = \frac{1}{\|e^T \Psi_0\|_{L_2}} e^T \Psi_0, T = T(\mathbf{t}_h), \mathbf{t}_h \in V_h \right\},$$

i.e.  $E_{var} \neq E_h$ , because  $e^{-T}He^T$  is no longer symmetric. Therefore it is not clear whether  $E_h \geq E$  holds in general or not.

**Size Consistency:** The Projected Coupled Cluster method is automatically size consistent. Here we follow the argumentation in physics literature.

**Theorem 4.10.** Let  $H_A, H_B$  be two independent Hamilton operators, (i.e.  $[H_A, H_B] = 0$ ), and  $\Psi_{0,A}, \Psi_{0,B}$  be the according reference Slater determinants and  $T_A, T_B$  be the excitation operators computed by individual Projected Coupled Cluster calculations with approximate energies  $E_A, E_B$ . Then  $\Psi_{AB} = e^T \Psi_0 = e^{T_A+T_B} (\Psi_{0,A} \wedge \Psi_{0,B}) = e^{T_A} \Psi_{0,A} \wedge e^{T_B} \Psi_{0,B}$  will be the corresponding (projected) CC solution with an approximate energy  $E_{AB} = E_A + E_B$ .

*Proof.* The full CI space of the total system is  $\mathcal{V}_{AB} = \mathcal{V}_A \wedge \mathcal{V}_B$ , and the projected Coupled Cluster solution  $T = \sum_{\mu \in \mathcal{J}_{h,A} \cup \mathcal{J}_{h,B}} t_\mu X_\mu$  solves

$$\langle \Psi_\nu, e^{-T}(H_A + H_B)e^T \Psi_0 \rangle = 0 \quad \forall \nu \in \mathcal{J}_{h,A} \cup \mathcal{J}_{h,B}.$$

Setting  $T = T_A + T_B$ , since the following operators are commuting  $[H_B, T_A] = 0, [H_A, T_B] = 0$  and consequently  $[H_B, e^{T_A}] = 0, [H_A, e^{T_B}] = 0$ , there holds

$$e^{-(T_A+T_B)}(H_A + H_B)e^{T_A+T_B} = e^{-T_A}H_Ae^{T_A} + e^{-T_B}H_Be^{T_B}.$$

From this, we infer that for all Slater determinants  $\Psi_\nu = \Psi_{\nu_A} \wedge \Psi_{\nu_B} \in \mathcal{V}_{AB}$  we have

$$\begin{aligned} 0 &= \langle \Psi_\nu, e^{-T}(H_A + H_B)\Psi_{AB} \rangle \\ &= \langle \Psi_\nu, e^{-T_A-T_B}(H_A + H_B)e^{T_A+T_B} \Psi_{0,A} \wedge \Psi_{0,B} \rangle \\ &= \langle \Psi_{\nu_A}, e^{-T_A}H_Ae^{T_A}\Psi_{0,A} \rangle + \langle \Psi_{\nu_B}, e^{-T_B}H_Be^{T_B}\Psi_{0,B} \rangle = 0 \quad \forall \nu_A \in \mathcal{J}_{h,A}, \nu_B \in \mathcal{J}_{h,B}. \end{aligned}$$

This implies that  $\Psi_{AB} = e^{T_A+T_B}(\Psi_{0,A} \wedge \Psi_{0,B}) = e^{T_A} \Psi_{0,A} \wedge e^{T_B} \Psi_{0,B}$  will be the corresponding projected CC solution. We compute the approximate energy by

$$\begin{aligned} E_{AB} &= \langle \Psi_0, (H_A + H_B)\Psi_{AB} \rangle \\ &= \langle \Psi_0, (H_A + H_B)e^{T_A}\Psi_{0,A} \wedge e^{T_B}\Psi_{0,B} \rangle \\ &= \langle \Psi_{0,A}, H_Ae^{T_A}\Psi_{0,A} \rangle + \langle \Psi_{0,B}, H_Be^{T_B}\Psi_{0,B} \rangle = E_{AB} = E_A + E_B, \end{aligned}$$

as predicted for exact results. □

In the sequel, we will consider the Projected Coupled Cluster Method as a Galerkin method applied to the nonlinear equation  $\mathbf{f}(\mathbf{t}) = \mathbf{0}$ , where  $\mathbf{f} : V \rightarrow V'$ . The following theorem establishes the relationship between the norm of the amplitude vector and the  $H^1$ -norm of a wave function.

**Theorem 4.11.** There holds for  $\mathbf{t} \in V$  and  $T := \sum_{\mu \in \mathcal{J}} t_\mu X_\mu$ ,

$$\|\mathbf{t}\|_V^2 = \langle T\Psi_0, e^{-T}Fe^T\Psi_0 \rangle \sim \|T\Psi_0\|_{H^1((\mathbb{R}^3 \times \{\pm \frac{1}{2}\})^N)}^2.$$

*Proof.* Since  $e^{-T}Fe^T = F + [F, T]$  and  $\Psi_0$  is an eigenfunction of the Fock operator  $F : \mathcal{V}_K \rightarrow \mathcal{V}'_K$ , we get

$$\langle \Phi, e^{-T}Fe^T\Psi_0 \rangle == \langle \Phi, (F + [F, T])\Psi_0 \rangle = \langle \Phi, \Lambda_0\Psi_0 + \sum_{\mu \in \mathcal{J}} \epsilon_\mu t_\mu \Psi_\mu \rangle .$$

for all  $\Phi \in \mathcal{V}_K$  where  $\Lambda_0 = \sum_{i=1}^N \lambda_i$  is the lowest (discrete) eigenvalue of  $F$ . Therefore

$$\langle T\Psi_0, (F + [F, T])\Psi_0 \rangle = \sum_{\mu \in \mathcal{J}} \epsilon_\mu |t_\mu|^2 = \|\mathbf{t}\|_V^2.$$

For  $\mu = \begin{pmatrix} a_1, & \dots, & a_k \\ l_1, & \dots, & l_k \end{pmatrix} \in \mathcal{J}$ , the  $\epsilon_\mu = \sum_{i=1}^k \lambda_{a_i} - \lambda_{l_i}$  are bounded from below by the HOMO-LUMO gap,

$$0 < \epsilon_0 := \lambda_{N+1} - \lambda_N \leq \epsilon_\mu .$$

The lowest discrete eigenvalue of the Fock operator is denoted by  $\Lambda_0$ , i.e.  $\langle \Phi, (F - \Lambda_0)\Psi_0 \rangle = 0$ , for all  $\Phi \in \mathcal{V}_K$ . From this we conclude that a lower estimate of the spectrum of the matrix  $(\langle \Psi_\mu, (F - \Lambda_0)\Psi_\nu \rangle)_{\mu, \nu \in \mathcal{J}}$  is provided by  $\epsilon_0$ . In view of the continuity of  $F : H^1 \rightarrow H^{-1}$  this implies further that  $\|\Phi\|_{H^1}^2 \sim \langle \Phi, (F - \Lambda_0)\Phi \rangle$ , holds for all  $\Phi \perp \Psi_0$ ,  $\Phi \in \mathcal{V}_K$ . We estimate the norm of  $\Phi := T\Psi_0$  by

$$\|\Phi\|_{H^1}^2 = \|T\Psi_0\|_{H^1}^2 \sim \langle T\Psi_0, (F - \Lambda_0)T\Psi_0 \rangle = \sum_{\mu \in \mathcal{J}} \epsilon_\mu |t_\mu|^2 = \|\mathbf{t}\|_V^2.$$

□

The above theorem states that for  $\Phi = \sum_{\mu \in \mathcal{J}} t_\mu \Psi_\mu$  the  $H^1$  norm of  $\Phi = T\Psi_0$  can be expressed by  $\|\Phi\|_{H^1} \sim \|\mathbf{t}\|_V$ . It will be important to investigate how the constants involved depend on the actual choice of the basis sets for approximation in the orbital space, i.e. these constants may depend on  $\mathcal{X}_K$  resp.  $K$ . Since  $\Psi_0 = \Psi_{HF, K} \in \mathcal{V}_K$ , the (discrete) Fock operator  $F = F_K$  and also  $\epsilon_0 = \epsilon_{0, K}$  actually depend on  $K \in \mathbb{N}$ , resp. on  $\mathcal{X}_K$ . In order to obtain uniform bounds we will introduce the following reasonable assumptions:

- (U1) There exists  $C_1 > 0$  such that  $\|\Psi_{HF, K}\|_{H^1} \leq C_1$  holds for all  $K \in \mathbb{N}$ .
- (U2) There exists a constant  $C_2 > 0$  such that  $\|F_K \Phi\|_{H^{-1}} \leq C_2 \|\Phi\|_{H^1}$  holds for all  $\Phi \in \mathcal{V}_K$  and  $K \in \mathbb{N}$ .
- (U3) (Uniform HOMO-LUMO gap) There exists a constant  $C_3 > 0$  such that

$$\lambda_{N+1, K} - \lambda_{N, K} > C_3 .$$

The proof of Lemma 4.11 shows that under these assumptions the norms are uniformly equivalent with respect to  $K \in \mathbb{N}$ .

**Theorem 4.12.** *Under the above assumptions (U1) - (U3), for  $\mathbf{t} \in V$  and  $T := \sum_{\mu \in \mathcal{J}} t_\mu X_\mu$ , the following norm equivalence holds uniformly with respect to  $K \in \mathbb{N}$*

$$\|\mathbf{t}\|_V \sim \|T\Psi_0\|_{H^1((\mathbb{R}^3 \times \{\pm \frac{1}{2}\})^N)}^2.$$

We would like to mention that we did not explicitly assume that the aufbau solutions of the (discrete) HF equations, are actually global minimizers of the Hartree-Fock energy functional in  $(\mathcal{X}_K)^N$ , nor that the discrete Hartree-Fock determinants  $\Psi_{HF,K}$  converge in  $H^1$  if  $K \rightarrow \infty$ . In this respect, the present conditions are slightly weaker than what is usually assumed in praxis. Theoretical investigations whether the assumptions (U1)–(U3) may hold in general are left for future research. For example, (U1) together with (U3) will be guaranteed if  $\|\Psi_{HF,K} - \Psi_{HF}\|_{H^1} \rightarrow 0$ ,  $K \rightarrow \infty$ , where  $\Psi_{HF}$  minimizes the Hartree-Fock energy in  $\mathcal{V}$ . Under these perspectives these assumptions are reasonable. Let us also remark that the HOMO-LUMO gap can be relatively small. Our present analysis exhibits also that small gaps may cause principal difficulties in the CC approximation process.

**Lemma 4.13.** *For  $\mathbf{t} \in l_2(\mathcal{J})$ , the operator  $T := \sum_{\nu \in \mathcal{J}} t_\nu X_\nu$  satisfies the following continuity estimate*

$$\|T\Psi\|_{L_2} \leq C_N \|\mathbf{t}\|_{l_2} \|\Psi\|_{L_2}$$

for all  $\Psi \in \mathcal{V}_K \subset \bigwedge_{i=1}^N L_2(\mathbb{R}^3 \times \{\pm \frac{1}{2}\})$ , uniformly w.r.t.  $K \rightarrow \infty$ .

*Proof.* Let us consider the matrix  $\mathbf{A} = (a_{\nu,\mu})_{\mu,\nu \in \mathcal{J}}$  with coefficients  $a_{\nu,\mu} := \langle \Psi_\nu, T\Psi_\mu \rangle$ . In case  $\#\mu \geq \#\nu$  we have  $a_{\nu,\mu} := \langle \Psi_\nu, T\Psi_\mu \rangle = 0$ . If  $\#\mu < \#\nu$ , then, according to Lemma 4.1, there exists at most one  $\alpha = \alpha(\nu, \mu) = \nu \ominus \mu$  such that  $X_\alpha X_\mu = X_\nu$ , and  $a_{\nu,\mu} := \langle \Psi_\nu, T\Psi_\mu \rangle = t_\alpha \langle \Psi_\nu, X_\alpha \Psi_\mu \rangle$ . This implies that for each  $\nu$  there are only finitely many nonzero coefficients  $a_{\nu,\mu} \neq 0$ . The number of nonzero entries is bounded by the constant  $C_N$  introduced in Lemma 4.1. Let  $\Psi = \sum_\mu c_\mu X_\mu \Psi_0$ , we estimate

$$\begin{aligned} \|T\Psi\|_{L_2}^2 &= \sum_\nu \|\Psi_\nu \sum_\mu a_{\nu,\mu} c_\mu\|_{L_2}^2 \leq \sum_\nu |\sum_\mu a_{\nu,\mu} c_\mu|^2 \leq C_N \sum_\nu \sum_\mu |a_{\mu,\nu} c_\mu|^2 \\ &= C_N \sum_{\alpha=\nu \ominus \mu} \sum_\mu |a_{\nu,\mu} c_\mu|^2 = C_N \sum_\alpha |t_\alpha|^2 \sum_\mu |c_\mu|^2 = C_N \|\mathbf{t}\| \|\Psi\|, \end{aligned}$$

since  $|\langle \Psi_\nu, X_\alpha \Psi_\mu \rangle| = 1$  holds only if  $\alpha = \alpha(\nu, \mu) = \nu \ominus \mu \in \mathcal{J}$ . □

Let us remark, that the estimate in Lemma 4.1 for the constant  $C_N$  is a worst case estimate and usually is far too pessimistic. For instance it is easy to see, that  $C_N$  does not exceed the actual number  $\#\mathcal{J}_h$  of excitation operators.

**Lemma 4.14.** *Let  $\nu \in \mathcal{J}$  and  $\#\alpha, \#\mu < \#\nu$  and  $\alpha = \nu \ominus \mu$  such that  $\langle \Psi_\nu, X_\alpha \Psi_\mu \rangle \neq 0$ , then there exists a constant  $C = C_N$  (not depending on  $\#\mathcal{J}$ ) such that*

$$\frac{\epsilon_\nu}{\epsilon_\mu} \leq C \epsilon_\alpha.$$

*Proof.* It holds  $0 < \epsilon_0 \leq \epsilon_\nu$  for all  $\nu \in \mathcal{J}$ , hence  $\epsilon_0^{-1} \geq \epsilon_\nu^{-1}$ . Let

$$0 < \bar{\lambda}_\nu := \max\{\lambda_{a_i} : i = 1, \dots, \#\nu\} - \delta$$

where  $\delta := \frac{1}{2}(\lambda_N + \lambda_{N+1})$  be a chemical potential. Obviously, there holds  $\epsilon_\nu \leq N(\bar{\lambda}_\nu - (\lambda_1 - \delta))$  and  $\bar{\lambda}_\nu \geq \frac{1}{2}\epsilon_0$ . From these estimates, we conclude  $|\lambda_1 - \delta| \leq 2\bar{\lambda}_\nu |\lambda_1 - \delta| \epsilon_0^{-1}$  and obtain  $\epsilon_\nu \leq N(\bar{\lambda}_\nu + |\lambda_1 - \delta|) \leq N\bar{\lambda}_\nu(1 + 2|\lambda_1 - \delta| \epsilon_0^{-1})$ . Therefore there exists a constant  $C > 0$  such that for all  $\nu \in \mathcal{J}$ , there holds  $C^{-1}\epsilon_\nu \leq \bar{\lambda}_\nu \leq \epsilon_\nu$ . This constant can be chosen such that the above estimates hold uniformly w.r.t.  $K \rightarrow \infty$ .

Since  $\Psi_\nu = X_\alpha \Psi_\mu$  there holds  $\bar{\lambda}_\nu = \bar{\lambda}_\alpha$  if  $\bar{\lambda}_\alpha \geq \bar{\lambda}_\mu$  and  $\bar{\lambda}_\nu = \bar{\lambda}_\mu$  if  $\bar{\lambda}_\alpha \leq \bar{\lambda}_\mu$ . In the first case it follows that

$$\frac{\epsilon_\nu}{\epsilon_\mu} \leq C \frac{\bar{\lambda}_\nu}{\epsilon_0} \leq \frac{C}{\epsilon_0} \bar{\lambda}_\alpha \leq \frac{C}{\epsilon_0} \epsilon_\alpha .$$

In the second case we estimate

$$\frac{\epsilon_\nu}{\epsilon_\mu} \leq C \frac{\bar{\lambda}_\nu}{\bar{\lambda}_\mu} = C \leq \frac{2C}{\epsilon_0} \bar{\lambda}_\alpha \leq \frac{2C}{\epsilon_0} \epsilon_\alpha .$$

□

The following theorem establishes an equivalence between the  $\|\mathbf{t}\|_V$  norm of the amplitude vector and the operator norm of the corresponding excitation operator  $T$  in  $H^1$ .

**Theorem 4.15.** *For  $\mathbf{t} \in V$ , the operator  $T := \sum_{\nu \in \mathcal{J}} t_\nu X_\nu$  satisfies the estimate*

$$\|T\Psi\|_{H^1} \leq C_N \|\mathbf{t}\|_V \|\Psi\|_{H^1}$$

for all  $\Psi \in \mathcal{V}$ . Under the assumptions (U1) - (U3), this estimates hold uniformly with respect to  $K \rightarrow \infty$ .

Moreover, the operator norm of the operator  $T$  from  $H^1 \rightarrow H^1$  is equivalent to  $\|\mathbf{t}\|_V$

$$\|T\|_{H^1 \rightarrow H^1} \sim \|\mathbf{t}\|_V . \quad (28)$$

*Proof.* Let  $\Psi = \sum_{\mu \in \mathcal{J}} c_\mu \Psi_\mu \in \mathcal{V}$ ,  $\mathbf{c} = (c_\mu)_{\mu \in \mathcal{J}}$  and  $T = \sum_{\alpha \in \mathcal{J}} t_\alpha X_\alpha$ . Then according to Theorem 4.11

$$\|T\Psi\|_{H^1}^2 \sim \|(\langle \Psi_\nu, T\Psi \rangle)_{\nu \in \mathcal{J}}\|_V^2 = \sum_{\nu \in \mathcal{J}} (\epsilon_\nu^{1/2} \left| \sum_{\alpha, \mu \in \mathcal{J}} t_\alpha c_\mu \langle \Psi_\nu, X_\alpha \Psi_\mu \rangle \right|^2)^2 = \|\mathbf{DAD}^{-1} \mathbf{Dc}\|_V^2 ,$$

where  $\mathbf{D} := \text{diag}(\epsilon_\mu^{1/2})$  and  $\mathbf{A}$  as defined in the proof of Lemma 4.13. We estimate the spectral norm of the matrix  $\mathbf{A} := \mathbf{DAD}^{-1}$ ,

$$\tilde{a}_{\nu, \mu} = \left( \frac{\epsilon_\nu}{\epsilon_\mu} \right)^{1/2} \sum_{\alpha \in \mathcal{J}} t_\alpha \langle \Psi_\nu, X_\alpha \Psi_\mu \rangle \lesssim \sum_{\alpha \in \mathcal{J}} t_\alpha \epsilon_\alpha^{1/2} \langle \Psi_\nu, X_\alpha \Psi_\mu \rangle .$$

Invoking the analysis of Lemma 4.13 it turns out that the operator norm of  $\tilde{\mathbf{A}}$  can be bounded by

$$\|\tilde{\mathbf{A}}\|_{l_2 \rightarrow l_2} \leq C_N \left( \sum_{\alpha \in \mathcal{J}} |t_\alpha|^2 \epsilon_\alpha \right)^{1/2} = C_N \|\mathbf{t}\|_V .$$

Therefore  $C_N \|\mathbf{t}\|_V$  presents an upper bound for the operator norm of  $T$ . The norm equivalence established in Theorem 4.11 provides a lower bound  $\|\mathbf{t}\|_V \sim \|T\Psi_0\|_{H^1} \leq \|T\|_{H \rightarrow H^1}$ .  $\square$

Since the nonlinear function  $\mathbf{f}$  is at most quartic in the unknown parameters  $\mathbf{t} = (t_\mu)$  we obtain the following result.

**Theorem 4.16.** *The function  $\mathbf{f} : V \rightarrow V'$  is differentiable at  $\mathbf{t} \in V$  with the Frechet derivative  $\mathbf{f}'[\mathbf{t}] : V \rightarrow V'$  given by*

$$(\mathbf{f}'[\mathbf{t}])_{\nu,\mu} = \langle \Psi_\nu, e^{-T}[H, X_\mu]e^T \Psi_0 \rangle = \langle \Psi_\nu, [F, X_\mu]\Psi_0 \rangle \delta_{\mu,\nu} + \langle \Psi_\nu, e^{-T}[U, X_\mu]e^T \Psi_0 \rangle . \quad (29)$$

where  $T = \sum_{\nu \in \mathcal{J}} t_\nu X_\nu$ . Furthermore, the Frechet derivative  $\mathbf{t} \mapsto \mathbf{f}'[\mathbf{t}]$ ,  $\mathbf{f}'[\mathbf{t}] : V \rightarrow V'$  is Lipschitz continuous as well as all higher derivatives. In particular  $\mathbf{f}^{(5)} \equiv 0$ . Therefore the function  $\mathbf{t} \mapsto \mathbf{f}(\mathbf{t})$  is infinitely many times differentiable. In particular, for any neighborhood  $U_R(\mathbf{0})$  and  $\mathbf{f} : U_R(\mathbf{0}) \subset V \rightarrow V'$ , there exists a Lipschitz constant  $L(R)$  such that

$$\|\mathbf{f}(\mathbf{t}) - \mathbf{f}(\mathbf{t}')\|_{V'} \leq L(R) \|\mathbf{t} - \mathbf{t}'\|_V , \quad \|\mathbf{t}\|_V, \|\mathbf{t}'\|_V < R . \quad (30)$$

Analogous results hold for higher derivatives of  $\mathbf{f}$ .

*Proof.* The first derivative of  $\mathbf{f}$  at  $\mathbf{t} \in V$  is given by

$$\begin{aligned} (\mathbf{f}'[\mathbf{t}])_{\nu,\mu} &= \langle \Psi_\nu, e^{-T}[H, X_\mu]e^T \Psi_0 \rangle \\ &= \langle \Psi_\nu, [F, X_\mu]\Psi_0 \rangle + \langle \Psi_\nu, e^{-T}[U, X_\mu]e^T \Psi_0 \rangle \\ &= \epsilon_\mu \delta_{\mu,\nu} + \langle \Psi_\nu, e^{-T}[U, X_\mu]e^T \Psi_0 \rangle , \end{aligned} \quad (31)$$

which can be easily derived from the Baker-Campbell-Hausdorff expansion. The boundedness of the first term in (31) follows immediately by definition, whereas the second term is bounded according to the following observations.

According to Theorem 4.15, resp. Lemma 4.13, the operator  $T : \mathcal{V} \rightarrow \mathcal{V}$  is bounded with respect to the Sobolev norm  $\|\cdot\| := \|\cdot\|_{H^1} = \|\cdot\|_{\mathcal{V}}$  and also to the  $L_2$ -norm  $\|\cdot\| := \|\cdot\|_{L_2}$ , with the corresponding operator norm  $\|T\| \lesssim C_N \|\mathbf{t}\|_V$ . For  $e^T \Psi_0 = \sum_{k=0}^{\infty} \frac{1}{k!} T^k \Psi_0$ , we obtain

$$\|e^T \Psi_0\| \leq \sum_{k=0}^{\infty} \frac{1}{k!} \|T^k \Psi_0\| \leq \sum_{k=0}^{\infty} \frac{1}{k!} \|T\|^k \|\Psi_0\| < \infty .$$

For  $H = -\frac{1}{2}\Delta + V$ , it is known (see e.g. [29, 34]) that the potential operator  $V$  maps  $V : H^1 \rightarrow L_2$  continuously

$$\|V(\cdot)\Psi\|_{L_2} \lesssim \|\Psi\|_{H^1} .$$

The same result is valid for the fluctuation potential  $U$  ([25]),

$$\|U\Psi\|_{L_2} \lesssim \|\Psi\|_{H^1}.$$

as well as

$$\|[U, T]_{(k)}\Psi\|_{L_2} \lesssim \|\Psi\|_{H^1}, \quad \Psi \in \mathcal{V}, \quad k = 1, \dots, 4.$$

From the above results, we conclude the continuity of  $\mathbf{f}'[\mathbf{t}] : V \rightarrow V'$ ,  $\mathbf{t} \in V$ , including the Lipschitz continuity of  $\mathbf{f}$ . Higher derivatives exist due to the same conclusion and can be estimated in a similar way.

We note that differential operators are contained only in the Fock operator, which appears linearly in the above expression, and  $f$  is quartic in  $t$  where the nonlinear terms arise only in conjunction with the fluctuation potential  $U$ . We remark that higher derivatives do not contain the Fock operator and, consequently, do not contain differential operators. These results can also be derived from the fact that  $f$  is polynomial in  $t$ . In particular, there holds  $\mathbf{f}^{(5)} \equiv 0$ , due to Lemma 4.5.  $\square$

**Remark 4.17. Quasi-Newton method to solve CC amplitude equations:**

The amplitude equation given by 4.7 are  $\mathbf{f}(\mathbf{t}_h)|_{V_h} := ((\mathbf{f}(\mathbf{t})_\nu)_{\nu \in \mathcal{J}_h} = \mathbf{0} \in \mathbb{R}^{\#\mathcal{J}_h}$  for an unknown vector  $\mathbf{t}_h \in \mathbb{R}^{\#\mathcal{J}_h}$ . This equation can be solved by the following quasi-Newton scheme

1. Choose  $\mathbf{t}^0$ , e.g.  $\mathbf{t}^0 = \mathbf{0}$ .
2. While  $\|(\mathbf{f}(\mathbf{t}_h)|_{V_h}\|_V \geq tol$ , compute

$$\mathbf{t}^{n+1} := \mathbf{t}^n - \mathbf{A}^{-1}\mathbf{f}(\mathbf{t}^n), \quad (32)$$

where  $\mathbf{A} = diag(\epsilon_\mu)_{\mu \in \mathcal{I}}$ .

Since  $\epsilon_\mu \geq \epsilon_0 > 0$  the matrix  $\mathbf{A}$  is nonsingular. For example, convergence of the above quasi-Newton method can be guaranteed if the initial vector  $\mathbf{t}^{(0)}$  is sufficiently close to the solution  $\mathbf{t}$  and e.g.  $\frac{1}{\epsilon_0} \|\mathbf{f}'[\mathbf{t}] - \mathbf{A}\|_{l_2 \rightarrow l_2} < \frac{1}{2}$ . That means, if  $\Psi_0$  is a good reference solution and the HOMO LUMO gap is sufficiently large, the above quasi Newton iteration is converging and the CC methods is expected to work pretty well. This performance of the CC methods is experienced in practice. However if the above conditions are not satisfied, there are cases where the convergence is quite slow, or there is no convergence and the CC method practically fails.

## 5 Convergence Analysis

**Definition 5.1.** A function  $\mathbf{f} : V \rightarrow V'$  will be called strictly monotone in some neighborhood  $U_\delta(\mathbf{v})$  of  $\mathbf{t}$  if there exists  $\gamma > 0$  such that

$$\langle \mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \geq \gamma \|\mathbf{v} - \mathbf{w}\|_V^2 \quad (33)$$

holds for all  $\|\mathbf{w} - \mathbf{v}\|_V < \delta$ .

Under this assumption, local existence and convergence follow by standard arguments [6].

**Theorem 5.2.** *Let  $\mathbf{f} : V \rightarrow V$  be strictly monotone in some  $U_\delta(\mathbf{t})$  where  $\mathbf{t}$  solves  $\mathbf{f}(\mathbf{t}) = \mathbf{0}$ . If  $\inf\{\|\mathbf{v}_h - \mathbf{t}\| : \mathbf{v}_h \in V_h\} \leq \delta'$  is sufficiently small, then there exists a unique solution  $\mathbf{t}_h \in V_h$  of the Projected Coupled Cluster method, defined by the Galerkin method*

$$\langle \mathbf{f}(\mathbf{t}_h), \mathbf{v}_h \rangle = 0, \quad \forall \mathbf{v}_h \in V_h,$$

which satisfies a quasi optimal error estimate

$$\|\mathbf{t} - \mathbf{t}_h\|_V \leq \frac{L}{\gamma} \inf_{\mathbf{v} \in V_h} \|\mathbf{t} - \mathbf{v}_h\|_V, \quad (34)$$

where  $L = L(\delta)$  is the Lipschitz constant of  $\mathbf{f}$  in  $U_\delta(\mathbf{t}) = \{\mathbf{v} \in V : \|\mathbf{v} - \mathbf{t}\|_V \leq \delta\}$ .

*Proof.* Existence and the error estimate follow by standard arguments [6]. See also the proof of Theorem 5.8.  $\square$

The assumption on strict monotonicity has been introduced above to guarantee optimal convergence. If the fluctuation potential  $U$  is sufficiently small compared to the Fock operator, this property is satisfied. Indeed, this follows from the  $V$  ellipticity of the shifted Fock operator

$$\sum_{\mu, \nu \in \mathcal{J}} t_\mu t_\nu \langle \Psi_\nu, [\mathcal{F}, X_\mu] \Psi_0 \rangle = \sum_\mu \epsilon_\mu |t_\mu|^2 \geq \epsilon_0 \|\mathbf{t}\|_V^2.$$

Provided that  $\|\Psi - \Psi_0\|_{H^1} \leq \delta$  is small enough, but so far this is not proved. In order to prove local existence and quasi-optimal convergence, we proceed in alternative way and consider the Galerkin solution of a different but somehow equivalent problem. To this end, let us abbreviate  $T(\mathbf{t}) = \sum_{\mu \in \mathcal{J}} t_\mu X_\mu$ , with  $\mathbf{t} \in V$  solves the following nonlinear equations

$$g_\nu(\mathbf{t}) := \langle \Psi_\nu, (H - E(\mathbf{t})) e^{T(\mathbf{t})} \Psi_0 \rangle = 0, \quad \forall \nu \in \mathcal{J}, \quad (35)$$

where

$$E(\mathbf{t}) = \langle \Psi_0, H e^{T(\mathbf{t})} \Psi_0 \rangle. \quad (36)$$

Clearly the nonlinear function  $\mathbf{g}$  maps  $\mathbf{g} : V \rightarrow V'$ . The discrete solutions  $\mathbf{t}_h$  and  $T_h := T(\mathbf{t}_h) = \sum_{\mu \in \mathcal{J}_h} t_\mu X_\mu$  are defined by

$$\langle \mathbf{v}_h, \mathbf{g}(\mathbf{t}_h) \rangle = 0 \quad \forall \mathbf{v}_h \in V_h.$$

Note that according to Proposition 4.7 the exact CC amplitude  $\mathbf{t}$  solves (35) as well as (26). Under an additional assumption on the discrete space  $V_h$  we can prove that the Galerkin solutions  $\mathbf{t}_h \in V_h$  are also identical.

**Definition 5.3.** *An index set  $\mathcal{J}_h \subset \mathcal{J}$ , respectively a space  $\mathcal{V}_h = \text{span } \{\Psi_\mu : \mu \in \mathcal{J}_h\}$ , is called excitation complete iff for every  $\mu, \nu \in \mathcal{J}$  satisfying  $\mu < \nu$ , see Lemma 4.1, there holds  $\alpha = \nu \ominus \mu \in \mathcal{J}_h$ .*

For example, this property holds for all kind of projections based on the excitation rank, e.g. for CCSD, CCSDT etc.. Let us denote the adjoint operator of  $X_\mu$  by  $X_\mu^\dagger$ , which is in essence a de-excitation operator.

**Lemma 5.4.**  $\mathcal{J}_h$  is excitation complete iff, for all  $\mu, \nu \in \mathcal{J}_h$  there holds

$$X_\mu^\dagger \Psi_\nu \in \mathcal{V}_{h,0} := \mathcal{V}_h \oplus \text{span}\{\Psi_0\} .$$

*Proof.* In case  $\mu = \nu$ , we obtain  $X_\mu^\dagger \Psi_\nu = \Psi_0$ . If  $\mu < \nu$ , then there hold  $\alpha = \nu \ominus \mu \in \mathcal{J}_h$  and  $X_\mu^\dagger \Psi_\nu = \Psi_\alpha \in \mathcal{V}_h$  and vice versa. In all other cases, i.e. if  $\mu \cap \nu \neq \mu$ , we get  $X_\mu^\dagger \Psi_\nu = 0$ .  $\square$

Let  $P_0 : \mathcal{V} \rightarrow \text{span}\{\Psi_0\}$  the orthogonal projection onto  $\text{span}\{\Psi_0\}$ .

**Lemma 5.5.** Let  $\mathcal{J}_h$  be excitation complete and let  $T_h^\dagger$  be the adjoint operator to  $T_h = T(\mathbf{t}_h) = \sum_{\mu \in \mathcal{J}_h} t_\mu X_\mu$ , then  $e^{T_h^\dagger} : \mathcal{V} \rightarrow \mathcal{V}$  maps  $\mathcal{V}_{h,0}$  one-to-one onto  $\mathcal{V}_{h,0}$ . Furthermore,  $(I - P_0)e^{\pm T_h^\dagger} : \mathcal{V}_h \rightarrow \mathcal{V}_h$  are surjective linear mappings.

*Proof.* According to Lemma 5.4 the operator  $T_h^\dagger$  maps  $T_h^\dagger : \mathcal{V}_{h,0} \rightarrow \mathcal{V}_{h,0}$ , which implies, by expanding the exponential  $e^{\pm T_h^\dagger}$  into a power series, that  $e^{\pm T_h^\dagger} : \mathcal{V}_{h,0} \rightarrow \mathcal{V}_{h,0}$ . Note that there holds  $e^{\pm T_h^\dagger} \Psi_0 = \Psi_0$ . The present excitation operators  $X_\mu, X_\nu$  are commuting  $[X_\mu, X_\nu] = 0$ , as well as their adjoint operators  $[X_\mu^\dagger, X_\nu^\dagger] = 0$ . The operator  $e^{T_h^\dagger}$  can be easily inverted, namely  $(e^{T_h^\dagger})^{-1} = e^{-T_h^\dagger} : \mathcal{V}_{h,0} \leftrightarrow \mathcal{V}_{h,0}$ .

Let us consider an arbitrary function  $\Psi \in \mathcal{V}_h$  and set  $\Phi := e^{T_h^\dagger} \Psi = c_0 \Psi_0 + \Phi'$ , with  $\Phi' := (I - P_0)\Phi \in \mathcal{V}_h$ . Then  $e^{-T_h^\dagger} \Phi' = \Psi - c_0 e^{-T_h^\dagger} \Psi_0 = \Psi - c_0 \Psi_0$ ,  $c_0 = \langle \Phi, \Psi_0 \rangle$ , and therefore we obtain  $(I - P_0)e^{-T_h^\dagger} \Phi' = \Psi$ , which finally proves the desired result.  $\square$

The next result states that both nonlinear equations admit an identical solution.

**Theorem 5.6.** Let  $\mathcal{J}_h$  be excitation complete, then  $\mathbf{t}_h \in V_h$  solves

$$\langle \mathbf{v}_h, \mathbf{g}(\mathbf{t}_h) \rangle = 0 , \forall \mathbf{v}_h \in V_h , \quad (37)$$

iff it is a solution of (27)

$$\langle \mathbf{v}_h, \mathbf{f}(\mathbf{t}_h) \rangle = 0 , \forall \mathbf{v}_h \in V_h .$$

*Proof.* Abbreviating  $T_h := T(\mathbf{t}_h)$ , we compute

$$\begin{aligned} & \langle \Phi, (H - E(\mathbf{t}_h))e^{T_h} \Psi_0 \rangle \\ &= \langle e^{T_h^\dagger} \Phi, e^{-T_h} (H - E(\mathbf{t}_h))e^{T_h} \Psi_0 \rangle \\ &= \langle (I - P_0)e^{T_h^\dagger} \Phi, e^{-T_h} (H - E(\mathbf{t}_h))e^{T_h} \Psi_0 \rangle + \langle P_0 e^{T_h^\dagger} \Phi, e^{-T_h} (H - E(\mathbf{t}_h))e^{T_h} \Psi_0 \rangle \\ &= \langle (I - P_0)e^{T_h^\dagger} \Phi, e^{-T_h} H e^{T_h} \Psi_0 \rangle + c_0 \langle \Psi_0, e^{-T_h} (H - E(\mathbf{t}_h))e^{T_h} \Psi_0 \rangle \\ &= \langle (I - P_0)e^{T_h^\dagger} \Phi, e^{-T_h} H e^{T_h} \Psi_0 \rangle , \forall \Phi \in \mathcal{V}_h , \end{aligned}$$

where we have used the energy expression  $E(\mathbf{t}_h) = \langle \Psi_0, H e^{T(\mathbf{t}_h)} \Psi_0 \rangle$  in the last step. The first set of equations (37) can be cast into the following weak form,

$$\langle \Phi, (H - E(\mathbf{t}_h)) e^{T(\mathbf{t}_h)} \Psi_0 \rangle = 0, \forall \Phi \in \mathcal{V}_h. \quad (38)$$

We conclude by Lemma 5.5 the above conditions (38) are equivalent to

$$0 = \langle \Phi', e^{-T_h} H e^{T_h} \Psi_0 \rangle, \forall \Phi' \in \mathcal{V}_h,$$

which is nothing but the Projected Coupled Cluster conditions (26) resp. (27).  $\square$

Both formulation can be used to compute the approximate CC amplitudes  $\mathbf{t}_h$ . However, the original formulation (26) exploits the fact that the Baker Campell Haussdorff expansion terminates. Therefore, the alternative formulation (35) is for theoretical purpose mainly. Next we show the strict monotonicity of  $\mathbf{g}$ . For the convergence estimates  $K \rightarrow \infty$ , we will need that strict monotonicity has to be uniform with respect to  $K \rightarrow \infty$ .

**Theorem 5.7.** *Let us assume that the index set  $\mathcal{J}_h$  is excitation complete. If the reference determinant  $\Psi_0$  is sufficiently closed to the wave function  $\Psi$ , i.e.  $\|\Psi - \Psi_0\|_{H^1} \leq \delta$ . Then  $\mathbf{g} : V \rightarrow V'$  is strictly monotone in a neighborhood  $U_{\delta'}(\mathbf{0})$ , and there holds*

$$\langle \mathbf{g}(\mathbf{t}) - \mathbf{g}(\mathbf{t}'), \mathbf{t} - \mathbf{t}' \rangle \geq \gamma \|\mathbf{t} - \mathbf{t}'\|^2, \forall \|\mathbf{t}\|_V, \|\mathbf{t}'\|_V \leq \delta'. \quad (39)$$

There is a constant  $\gamma > 0$  such that the above estimate holds uniformly with respect the size of basis set  $K > K_0$ .

*Proof.* We compute the Frechét derivative, resp. the Jacobian  $g'_{\mu,\nu}$  of  $\mathbf{g}$  at point  $\mathbf{0}$ . Let  $E_0 := \langle \Psi_0, H \Psi_0 \rangle = E(\mathbf{0})$ , then

$$\begin{aligned} g'_{\mu,\nu}(\mathbf{0}) &= \lim_{t \rightarrow 0} \frac{1}{t} (\langle \Psi_\nu, (H - E(t\mathbf{e}_\mu)) e^{tX_\mu} \Psi_0 \rangle - \langle \Psi_\nu, (H - E_0) \Psi_0 \rangle) \\ &= \langle \Psi_\nu, (H - E_0) X_\mu \Psi_0 \rangle - \langle \Psi_\nu, \langle \Psi_\nu, H X_\mu \Psi_0 \rangle \Psi_0 \rangle = \langle \Psi_\nu, (H - E_0) \Psi_\mu \rangle. \end{aligned}$$

We observe that is nothing but the Galerkin matrix of the operator  $H - E_0$  with respect to the subspace  $\mathcal{V}_h \subset (I - P_0)\mathcal{V}$ . For  $E_0$  sufficiently closed to  $E$ , which follows from  $\|\Psi - \Psi_0\|_{H^1} \leq \delta$  sufficiently small, it is well known that we have coercivity of  $H - E_0$  on the orthogonal complement of  $\Psi_0$ ,

$$\langle \Phi, (H - E_0) \Phi \rangle \geq \gamma' \|\Phi\|_{H^1}, \forall \Phi \in \mathcal{V} \perp \Psi_0.$$

Let us remark, that the latter estimate holds for some  $\gamma' > 0$  uniformly with respect to  $K < K_0$  ( $\mathcal{V} = \mathcal{V}_K$ ).

By linearizing, we get for  $\|\mathbf{t}\| \leq \delta$  sufficiently small

$$\langle \mathbf{g}(\mathbf{t}) - \mathbf{g}(\mathbf{0}), \mathbf{t} \rangle = \langle \mathbf{g}'(\mathbf{0}) \mathbf{t}, \mathbf{t} \rangle + \mathcal{O}(\|\mathbf{t}\|_V^3) \geq \gamma \|\mathbf{t}\|_V^2.$$

From this we deduce, that there for  $\|\mathbf{t}'\|_V, \|\mathbf{t}\| \leq \delta'$ , there holds

$$\langle \mathbf{g}(\mathbf{t}) - \mathbf{g}(\mathbf{t}'), \mathbf{t} - \mathbf{t}' \rangle \geq \gamma \|\mathbf{t} - \mathbf{t}'\|_V^2.$$

$\square$

We obtain a local existence result for the Projected CC equations together with quasi-optimal convergence.

**Theorem 5.8.** *Let  $\|\Psi - \Psi_0\|_{H^1} < \delta$  sufficiently small and  $\Psi = e^{T(\mathbf{t})}\Psi_0$ . If  $\mathcal{J}_h$  is excitation complete and sufficiently large so that  $\inf_{\mathbf{v}_h \in V_h} \|\mathbf{v}_h - \mathbf{t}\|_V \leq \delta$ , then there exists a solution of the Projected Coupled Cluster amplitude equations  $\mathbf{t}_h \in V_h$  satisfying a quasi optimal error estimate*

$$\|\mathbf{t}_h - \mathbf{t}\|_V \lesssim \inf_{\mathbf{v}_h \in V_h} \|\mathbf{v}_h - \mathbf{t}\|_V$$

*The error of the corresponding wave functions can be estimated up to a constant factor by the best approximation error on the nonlinear manifold*

$$\|\Psi - e^{T(\mathbf{t}_h)}\Psi_0\| \lesssim \inf_{\mathbf{v}_h \in V_h} \|\Psi - e^{T(\mathbf{v}_h)}\Psi_0\|_{H^1} .$$

*Proof.* The existence is proved by standard techniques using Theorem 5.7 together with the implicit function theorem. The quasi-optimal convergence of the discrete solution  $\mathbf{t}_h$  can be concluded directly from (39). Using the Galerkin orthogonality  $\langle \mathbf{g}(\mathbf{t}) - \mathbf{g}(\mathbf{t}_h), \mathbf{v}_h \rangle = 0$ , there holds for any  $\mathbf{v}_h \in V_h$ ,

$$\begin{aligned} \|\mathbf{t} - \mathbf{t}_h\|^2 &\leq \gamma^{-1} \langle \mathbf{g}(\mathbf{t}) - \mathbf{g}(\mathbf{t}_h), \mathbf{t} - \mathbf{t}_h \rangle \\ &= \gamma^{-1} \langle \mathbf{g}(\mathbf{t}) - \mathbf{g}(\mathbf{t}_h), \mathbf{t} - \mathbf{v}_h \rangle \leq \frac{L}{\gamma} \|\mathbf{t} - \mathbf{t}_h\|_V \|\mathbf{t} - \mathbf{v}_h\|_V , \end{aligned}$$

where  $L = L(\delta)$  is a Lipschitz constant.

The last assertion is a consequence of the norm equivalence stated in Theorem 4.11, yielding that for all  $\mathbf{v}_h \in V_h$  with  $\|\mathbf{t} - \mathbf{v}_h\|_V \leq \delta$  there holds

$$\|\mathbf{t} - \mathbf{v}_h\|_V \sim \|(T(\mathbf{t}) - T(\mathbf{v}_h))\Psi_0\|_{H^1} \sim \|e^{T(\mathbf{t})}\Psi_0 - e^{T(\mathbf{v}_h)}\Psi_0\|_{H^1} .$$

□

**Remark 5.9.** *If the Hartree-Fock determinant, or any other single reference determinant, is a rather bad approximation, it is known that the Projected Coupled Cluster method fails. This situation is demanding for a multi-reference solution as a starting point (static correlation). This experience indicates that, in general, it cannot be expected that strong monotonicity always holds, or the constants might be extremely bad. Therefore we expect to get local existence results as best. We like to stress, that our analysis demonstrates that the HOMO-LUMO gap enters the constants involved crucially and will play an important role in the convergence of the CC method.*

## 6 Energy Error

From the results of Theorem 5.2 and the energy expression in Theorem 4.4, one can easily conclude that the error of the computed energy is bounded by a constant times the best approximation error of the amplitudes.

However, this result is less favorable than the corresponding result for the truncated CI methods. Since CI constitutes a variational scheme, the energy error decays quadratically with respect to the error of the wave function, compare Theorem 3.1. In the sequel, we will present a more detailed analysis based on a type of Aubin-Nitsche trick to get reasonable bounds for the error. These techniques have been developed by Rannacher and coworkers [3] for establishing goal oriented a posteriori error estimates in the context of finite element methods. Since the basis functions used in electronic structure calculation are not local and less systematic and moreover a rigorous convergence analysis is still in its infancy, a development of a posteriori error estimators for the Coupled Cluster methods, based on the results presented below will be not straightforward.

In the sequel, we assume that the approximate amplitude vector  $\mathbf{t}_h \in V_h$  is included in an appropriate neighborhood  $U_\delta(\mathbf{t})$ , meaning that the error  $\|\mathbf{t} - \mathbf{t}_h\|_V \leq \delta$  is sufficiently small. Let us consider the ground state energy as a functional depending on the cluster amplitudes by formula (18),

$$J(\mathbf{t}) := \langle \Psi_0, H(1 + T_2 + \frac{1}{2}T_1^2) \Psi_0 \rangle$$

where  $\mathbf{t}$  solves the amplitude equations

$$\mathbf{f}(\mathbf{t}) = (\langle \Psi_\nu, e^{-T} H e^T \Psi_0 \rangle)_{\nu \in \mathcal{J}} = \mathbf{0}.$$

Let us further consider the Lagrange functional

$$L(\mathbf{t}, \mathbf{a}) := J(\mathbf{t}) + \langle \mathbf{f}(\mathbf{t}), \mathbf{a} \rangle, \quad \mathbf{t} \in V, \quad \mathbf{a} \in V. \quad (40)$$

and its stationary points  $L'[\mathbf{t}, \mathbf{a}](\mathbf{r}, \mathbf{b}) = \mathbf{0}$ , i.e.

$$\langle J'[\mathbf{t}] + \mathbf{f}'[\mathbf{t}]^\top \mathbf{a}, \mathbf{r} \rangle = 0, \quad \forall \mathbf{r} \in V. \quad (41)$$

and

$$\langle \mathbf{f}[\mathbf{t}], \mathbf{b} \rangle = 0, \quad \forall \mathbf{b} \in V. \quad (42)$$

The last equation is exactly the cluster amplitude equation, whereas the first equation is an equation for a *dual solution*  $\mathbf{a}$  which has to satisfy

$$\mathbf{f}'[\mathbf{t}]^\top \mathbf{a} = -J'[\mathbf{t}] \text{ in } V'.$$

These equations can be solved approximately by the Galerkin method. The discrete solution  $\mathbf{t}_h = (t_\nu)_{\nu \in \mathcal{J}_h}$ ,  $\mathbf{t}_h \in V_h$  of the primal nonlinear system (42) is the solution of the projected Coupled Cluster amplitude equations

$$\langle \mathbf{f}(\mathbf{t}_h), \mathbf{b}_h \rangle = 0, \quad \forall \mathbf{b}_h \in V_h, \quad (43)$$

and the approximate Lagrange multipliers  $\mathbf{a}_h = (a_\mu)_{\mu \in \mathcal{J}_h}$  are the solution of the discrete dual problem

$$\langle \mathbf{f}'[\mathbf{t}_h]^\top \mathbf{a}_h, \mathbf{v}_h \rangle = \langle -J'[\mathbf{t}_h], \mathbf{v}_h \rangle, \quad \forall \mathbf{v}_h \in V_h. \quad (44)$$

Correspondingly, we define the *primal residual*  $\mathbf{r}(\mathbf{t}_h)$  by

$$\mathbf{r}(\mathbf{t}_h)_\mu = f_\mu(\mathbf{t}_h) = (\mathbf{f}(\mathbf{t}_h))_\mu \text{ if } \mu \notin \mathcal{J}_h . \quad (45)$$

and zero otherwise, together with the *dual residual*  $\mathbf{r}^*(\mathbf{t}_h, \mathbf{a}_h)$  by

$$\mathbf{r}^*(\mathbf{t}_h, \mathbf{a}_h)_\mu = (\mathbf{f}'[\mathbf{t}_h]^\top \mathbf{a}_h + J'[\mathbf{t}_h])_\mu \text{ if } \mu \notin \mathcal{J}_h , \quad (46)$$

and zero otherwise.

The derivatives  $J'[\mathbf{t}]$  and  $\mathbf{f}'[\mathbf{t}]$  can be explicitly computed using the notation  $T = \sum_{\nu \in \mathcal{J}} t_\nu X_\nu$ .

$$(J'[\mathbf{t}])_\mu = \begin{cases} \langle \Psi_0, UT_1\Psi_\mu \rangle & \sharp\mu = 1 \\ \langle \Psi_0, U\Psi_\mu \rangle , & \sharp\mu = 2 \\ 0 , & \text{otherwise} \end{cases} \quad (47)$$

and

$$(\mathbf{f}'[\mathbf{t}])_{\mu,\nu} = \epsilon_\mu \delta_{\mu,\nu} + \langle \Psi_\mu, e^T [U, X_\nu] e^T \Psi_0 \rangle, \mu, \nu \in \mathcal{J} . \quad (48)$$

By the same arguments as in the proof of Theorem 4.16, we conclude the analyticity of the map  $\mathbf{t} \rightarrow J(\mathbf{t})$ . In particular,  $J$  is only quadratic in  $\mathbf{t} \in V$ .

**Proposition 6.1.** *The function  $\mathbf{t} \rightarrow J(\mathbf{t})$ ,  $t \in V$ , is differentiable, with  $J'[\mathbf{t}] \in V'$ , and locally Lipschitz continuous in  $V$ . The same assertion holds for all derivatives of  $J$ .*

**Lemma 6.2.** *(Rannacher et al.) Let  $x := (\mathbf{t}, \mathbf{a}) \in V \times V$ ,  $x_h := (\mathbf{t}_h, \mathbf{a}_h) \in V_h \times V_h$  and  $e_h = x - x_h$ . If  $L'[x] = 0$  and  $L'[x_h](x_h) = 0$ , then*

$$L(x) - L(x_h) = L'[x_h](x - x_h) + R^3 ,$$

where the remainder term  $R^3 = \frac{1}{2} \int_0^1 L^{(3)}[x_h + se_h](e_h, e_h, e_h) s(s-1) ds$  depends cubically on the error  $e_h$ .

For a proof of this result see [3].

Since  $0 = \langle \mathbf{f}(\mathbf{t}), \mathbf{a} \rangle = \langle \mathbf{f}(\mathbf{t}_h), \mathbf{a}_h \rangle$  and  $L(\mathbf{t}, \mathbf{a}) = J(\mathbf{t}) + \langle \mathbf{f}(\mathbf{t}), \mathbf{a} \rangle$  there holds  $L(x) - L(x_h) = J(\mathbf{t}) + \langle \mathbf{f}(\mathbf{t}), \mathbf{a} \rangle - J(\mathbf{t}_h) - \langle \mathbf{f}(\mathbf{t}_h), \mathbf{a}_h \rangle = J(\mathbf{t}) - J(\mathbf{t}_h)$ . As an immediate result of this lemma we obtain the following result.

**Theorem 6.3.** *(Rannacher et al) For a solution  $\mathbf{t} \in V$  of the amplitude equations and a solution  $\mathbf{a} \in V$  of the dual problem (41) together with discrete solutions  $\mathbf{t}_h, \mathbf{a}_h \in V_h$ , the error of the energy functional can be represented by*

$$|J(\mathbf{t}) - J(\mathbf{t}_h)| = \frac{1}{2} \langle \mathbf{r}(\mathbf{t}_h), (\mathbf{a} - \mathbf{b}_h) \rangle + \frac{1}{2} \langle \mathbf{r}^*(\mathbf{t}_h, \mathbf{a}_h), (\mathbf{t} - \mathbf{u}_h) \rangle + R^3 , \quad \forall \mathbf{u}_h, \mathbf{b}_h \in V_h , \quad (49)$$

with a cubic remainder term  $R^3 = \mathcal{O}((\|\mathbf{t} - \mathbf{t}_h\|_V + \|\mathbf{a} - \mathbf{a}_h\|_V)^3)$ .

With this results at hand, we can state the following theorem.

**Theorem 6.4.** *Let  $\tilde{\mathbf{a}} \in V$ ,  $\tilde{\mathbf{a}}_h \in V_h$ , be solutions of the approximate adjoint problems*

$$\langle \mathbf{f}'[\mathbf{t}_h]^\top \tilde{\mathbf{a}} - J'[\mathbf{t}_h], \mathbf{v} \rangle = 0 \quad \forall \mathbf{v} \in V, \quad \langle \mathbf{f}'[\mathbf{t}_h]^\top \tilde{\mathbf{a}}_h - J'[\mathbf{t}_h], \mathbf{v}_h \rangle = 0 \quad \forall \mathbf{v}_h \in V_h. \quad (50)$$

*Then, under the assumption (33), the error in the energy  $E = J(\mathbf{t})$  and the discrete energy  $E_h = J(\mathbf{t}_h)$  can be estimated by*

$$|E - E_h| \lesssim \inf_{\mathbf{u}_h \in V_h} \|\mathbf{t} - \mathbf{u}_h\|_V (\inf_{\mathbf{b}_h \in V} \|\mathbf{a} - \mathbf{b}_h\|_V + \|\tilde{\mathbf{a}} - \mathbf{a}_h\|_V) \leq \|\mathbf{t} - \mathbf{t}_h\|_V (\|\mathbf{a} - \mathbf{a}_h\|_V + \|\tilde{\mathbf{a}} - \mathbf{a}_h\|_V).$$

*Proof.* Applying the Cauchy-Schwarz inequality

$$\langle \mathbf{r}(\mathbf{t}_h), \mathbf{a} - \mathbf{b}_h \rangle \leq \|\mathbf{r}(\mathbf{t}_h)\|_{V'} \|\mathbf{a} - \mathbf{b}_h\|_V$$

to the result of Theorem 6.3, we have to prove the estimate  $\|\mathbf{r}(\mathbf{t}_h)\|_{V'} = \|\mathbf{f}(\mathbf{t}_h)\|_{V'} \leq L \|\mathbf{t} - \mathbf{t}_h\|_V$ . Indeed, this estimate is a consequence of the Lipschitz continuity of  $\mathbf{f}$ ,

$$\|\mathbf{f}(\mathbf{t}_h)\|_{V'} = \|\mathbf{f}(\mathbf{t}_h) - \mathbf{f}(\mathbf{t})\|_{V'} \leq L \|\mathbf{t} - \mathbf{t}_h\|_V.$$

Theorem 5.2 implies

$$\|\mathbf{t} - \mathbf{t}_h\|_V \lesssim \inf_{\mathbf{u}_h \in V_h} \|\mathbf{t} - \mathbf{u}_h\|_V.$$

We apply the Lipschitz continuity of  $\mathbf{f}'$  and  $J'$ , together with

$$\langle \mathbf{r}^*(\mathbf{t}_h, \mathbf{a}_h), \mathbf{t} - \mathbf{u}_h \rangle = \langle \mathbf{f}'[\mathbf{t}_h]^\top (\mathbf{a}_h - \tilde{\mathbf{a}}), \mathbf{t} - \mathbf{u}_h \rangle$$

to estimate

$$\langle \mathbf{r}^*(\mathbf{t}_h, \mathbf{a}_h), \mathbf{t} - \mathbf{u}_h \rangle \lesssim \|\mathbf{t} - \mathbf{u}_h\|_V \|\tilde{\mathbf{a}} - \mathbf{a}_h\|_V,$$

for all  $\mathbf{u}_h \in V_h$ .

□

**Remark 6.5.** *The dual problem has been introduced in the context of the dual weighted residual approach [3]. Although this approach obviously seems to be new in quantum chemistry, the dual problem is already known in quantum chemistry. In fact, the introduced Lagrangian coincides with the variational Coupled Cluster Lagrangian (see [14]) and the dual solution appears explicitly in the equation of motion formulation [14].*

We have to take in mind that in the previous sections 4–6, we have dropped the index  $K$  at most places, for notational convenience. Let  $\mathbf{a}, \tilde{\mathbf{a}}, \mathbf{a}_h \in V$  be the solution of the dual problems (41) and (50), and  $\mathbf{t} \in V$ . We set  $\Phi_K := \sum_{\mu \in \mathcal{J}_K} a_\mu X_\mu \Psi_{0,K} \in \mathcal{V}_K \subset \mathcal{V}$ ,  $\tilde{\Phi}_K$  and  $\Phi_{K,h} := \sum_{\mu \in \mathcal{J}_{h,K}} a_{h,\mu} X_\mu \Psi_{0,K} \in \mathcal{V}_{h,K} \subset \mathcal{V}_K$  accordingly. Setting  $T = \sum_{\mu \in \mathcal{J}} t_\mu X_\mu$  the CI wave function is given by  $\Psi_K = e^T \Psi_{0,K}$ . We introduce the finite dimensional nonlinear manifold  $\mathcal{M}_{h,K} := \{\Phi = e^T \Psi_{0,K} \in \mathcal{V}_K : T = \sum_{\mu \in \mathcal{J}_{h,K}} t_\mu X_\mu, \mathbf{t} \in V_{h,K}\}$  together with the linear subspace  $\mathcal{V}_{h,K} := \{\Phi = T \Psi_{0,K} \in \mathcal{V}_K : T = \sum_{\mu \in \mathcal{J}_{h,K}} t_\mu X_\mu, \mathbf{t} \in V_{h,K}\} \subset \mathcal{V}_K$ .

**Theorem 6.6.** Under the assumptions (U1) - (U3), there exists  $K_0 \in \mathbb{N}$  such that for all  $K \geq K_0$  there is a  $\mathcal{J}_{h_0} \subset \mathcal{J}$  such that for all  $h < h_0$  with  $\mathcal{J}_{h_0} \subset \mathcal{J}_h \subset \mathcal{J}$ , there holds the following quasi-optimal a priori error estimate

$$\|\Psi - \Psi_h\|_{H^1} \lesssim \inf_{u_K \in \mathcal{V}_K} \|\Psi - u_K\|_{H^1} + \inf_{v_h \in \mathcal{M}_{h,K}} \|\Psi_K - v_h\|_{H^1} \quad (51)$$

together with

$$|E - E_h| \lesssim \inf_{u_K \in \mathcal{V}_K} \|\Psi - u_K\|_{H^1}^2 + \inf_{v_h \in \mathcal{M}_{h,K}} \|\Psi_K - v_h\|_{H^1} (\|\Phi_K - \Phi_{h,K}\|_{H^1} + \|\Phi_K - \Phi_{h,K}\|_{H^1}). \quad (52)$$

The constants involved in these estimates are uniform with respect to  $K \geq K_0$  and  $h < h_0$ .

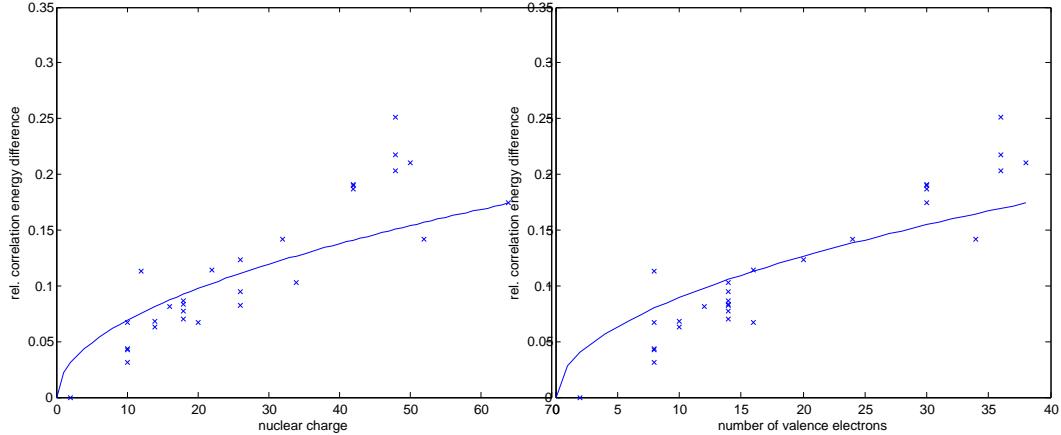
*Proof.* The convergence of the CI solution  $\Psi_K$  was established. We decompose  $\|\Psi - \Psi_h\|_{H^1} \leq \|\Psi_K - \Psi\|_{H^1} + \|\Psi_K - \Psi_h\|_{H^1}$ . It remains to show that there exists  $C > 0$  such that  $\|\Psi_K - \Psi_h\|_{H^1} \leq C \inf_{v_h \in \mathcal{M}_{h,K}} \|\Psi_K - v_h\|_{H^1}$ . Theorem 5.2 states that for fixed  $K$ ,  $\mathbf{t}_h$  converges quasi-optimally to  $\mathbf{t}$  in  $V$ . Due to assumptions (U1) - (U3), the constants in the a priori estimate do not depend on  $K$ . The uniform norm equivalence  $\|\mathbf{t} - \mathbf{t}_h\|_V \sim \|\Psi_K - \Psi\|_{H^1}$  is valid due to assumptions (U1) - (U3). This norm equivalence can also be applied to the result 6.4.  $\square$

## 7 Numerical Examples and Conclusions

By experience, it is well known that the (Projected) Coupled Cluster method outperforms the truncated CI methods, like CISD etc. see e.g. [14, 15]. Roughly speaking, considering a fixed basis set of sufficient quality, CCSD yield the same accuracy as CISDTQ or at least CISDT, but at a much lower cost. Of course CCSD is slightly more expensive than CISD. Additionally, CCSD is exact up to third order MP perturbation theory, again with a much lower cost and higher precision as MP3. A basic argument that CCSD is superior to CISD is the size consistency. There are some heuristic arguments that the error of truncated CI increases asymptotically like  $\mathcal{N}^{1/2}$  with the size of the system [17, 19]. This means that truncated CI methods are not appropriate for larger systems than only very few electrons. MCSCF methods (alone without improvements by perturbation methods or MRCI) cannot achieve those highly accurate results, due to their exponential scaling w.r.t to the particle number  $N$ . We like to present some numerical examples demonstrating that indeed the difference between the CISD ground state energy and the CCSD increases with the size of the system. For a numerical test, we considered several molecules and compared the CISD correlation energy  $E_{HF} - E_{CI}$  with the CCSD correlation energy  $E_{HF} - E_{CC}$  and also different basis sets.

The following diagrams depict the relative difference in the correlation energy versus the total nuclear charge  $Z = \sum Z_j$  and versus the number of valence electrons  $N'$  for a series of molecules. The line is plotting the curve  $\sim \sqrt{n}$ ,  $n = N, N'$ . All computations were performed in MOLPRO [15]. One clearly observes that the deviation between both method increases with

the size of the molecules with a rate which is roughly  $\mathcal{O}(N^{1/2})$ , which is due to the lack of size consistency in CISD. Let us remark this deficiency results only from the particular truncation w.r.t. to the excitation rank, which, by the structure of the codes, is a very canonical choice.



The relative error in the correlation CC energy depends on the underlying basis set. For most systems under consideration the correlation energy (in terms of full CI solutions) is expected to be below 1 – 5%. For some further comparisons we refer e.g. to [14, 15].

For highly accurate computations, recently higher excitations and also geminal and  $r_{1,2}$ -approaches could be included, see e.g. [17]. To compute relatively large molecular systems an original scaling  $\mathcal{O}(N^7)$  for the CCSD could be reduced to almost linear scaling by [28].

We have shown quasi-optimal a priori convergence estimates. It remains to investigate approximation properties of the underlying basis functions. The basis functions used in quantum chemistry have been particularly tailored for the present purpose and have shown to be by far the most efficient basis functions for all electron calculations. Some analysis has been performed for these special examples e.g. Hydrogen and Helium atoms [19, 21]. Furthermore there are several numerical studies about convergence behavior of Gaussian type basis functions [17], showing that the electron-nuclear cusp can be approximated by an exponential rate  $e^{-a\sqrt{K}}$  with respect to the size of basis sets  $\mathcal{X}_K$ . The experimental convergence rate for the energy of an electron-electron cusp is  $\mathcal{O}(K^{-2})$  for orbital methods, which could be improved by geminal methods to

$\mathcal{O}(K^{-4})$ . However the total number of degrees of freedom and computational complexity depends on the rank of excitations. For CCSD one has  $n = \mathcal{O}(K^2)$  degrees of freedom, which gives a convergence rate about  $\mathcal{O}(n^{-1})$ . For more precise calculation, higher excitations will be required. Despite the overwhelming success of these methods, by the traditional way to achieve convergence to the exact wave function, one has to increase the excitation rank, which would result finally in an exponential scaling of the complexity. A recent theoretical result shows that a sparse grid approximation could provide a way circumventing exponential scaling[32, 33]. From this result one may conclude a convergence rate about  $\mathcal{O}(n^{-1/3})$  for the ground state energy. The particular electron-electron cusp can be approximated by a rate  $\mathcal{O}(n^{-1})$  [10, 11]. This rate will be the best which can be obtained by orbital approximation in general, suggesting the following conjecture, that for an appropriate family of basis sets  $\{\mathcal{X}_K\}_{K \in \mathbb{N}}$ , there might exist  $n = \mathcal{O}(K)$  amplitudes, i.e.  $\#\mathcal{J}_h = n$ , providing a convergence behavior  $\mathcal{O}(n^{-1+\epsilon})$ , for any  $\epsilon > 0$ . These amplitudes cannot be selected only by their excitation rank. Adaptive algorithms and nonlinear approximation strategies should be developed for an automatic selection of these amplitudes, possibly together with an automatic selection of basis sets.

In a forthcoming paper we would like to consider a contionous coupled cluster method and a theoretical comaprison with other size consistent methods like CEPA [31, 19] etc..

**Acknowledgement:** I would like to thank Dr. H.-J. Flad for stimulating discussions. His expertise in chemistry and physics has guided me into the field of quantum chemistry. Moreover I thank Prof. H. Yserentant, T. Rohwedder and N. Pfeiffer for helpful comments and discussions.

## References

- [1] S. Agmon, *Lectures on Exponential Decay on Solutions of Second Order Elliptic Equations: Bounds on the Eigenfunctions of N-Body Schrödinger Operators*, Mathematical Notes **29**, Princeton University Press 1982.
- [2] V. Bach, E.H. Lieb, M. Loss, J.P. Solovej. There are no unfilled shells in unrestricted Hartree-Fock theory. *Phys. Rev. Lett.*, 72:2981–2983, 1994.
- [3] W. Bangert, R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag Basel 2003
- [4] R. J. Bartlett, M. Musial, Coupled-cluster theory in quantum chemistry, *Rev. of Modern Physics*, Vol. 79(1), pp. 291-352, 2007
- [5] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, Y. Maday. Computational Chemistry: a primer, in: *Handbook of Numerical Analysis*, Volume X. North-Holland 2003.
- [6] G.P. Ciarlet. *The Finite Element Method for Elliptic Problems*, Studies in Mathematics and its Application, North Holland, Amsterdam 1978.

- [7] G.P. Ciarlet. *Handbook of Numerical Analysis*, Volume X: Computational Chemistry. North-Holland, 2003.
- [8] T. D. Crawford, H. F. Schaefer III, An Introduction to Coupled Cluster Theory for Computational Chemists, *Reviews in Computational Chemistry*, Volume 14, pp. 33-136, 2000
- [9] H.-J. Flad, W. Hackbusch, D. Kolb, R. Schneider, Wavelet approximation of correlated wavefunctions. I. Basics. *J. Chem. Phys.* 116:9641–9857, 2002.
- [10] H.-J. Flad, W. Hackbusch, R. Schneider, Best N term approximation for electronic wavefunctions I One electron reduced density matrix, *M2AN Math. Model. Numer. Anal.* **40**, 2006
- [11] H.-J. Flad, W. Hackbusch, R. Schneider, Best N term approximation for electronic wavefunctions II, Jastrow factors, MIS, Leipzig Preprint 80/2005 to appear in *M2AN Math. Model. Numer. Anal.*
- [12] S. Fournais, M. Thomas-Ostenhof, T. Thomas-Ostenhof, T. Ostergaard Sorensen, Sharp regularity results for Coulombic many-electron wave functions, *Comm. Math. Phys.* **255**, 183–227, 2005.
- [13] G. Friesecke, The multiconfiguration equations for atoms and molecules: charge quantization and existence of solutions. *Arch. Rat. Mech. Anal.*, 169:35–71, 2003.
- [14] T. Helgaker, P. Jørgensen, J. Olsen. *Molecular electronic-structure theory*. John Wiley & Sons, New York, 2002.
- [15] C. Hampel, K. Peterson, H.-J. Werner. A comparison of efficiency and accuracy of the quadratic configuration interaction (QCISD), coupled cluster (CCSD) and Brückner coupled cluster (BCCD) methods, *Chem. Phys. Lett.* **190**, 1-12, 1992.
- [16] T. Kato, On the eigenfunctions of many-particle systems in quantum mechanics. *Comm. Pure Appl. Math.*, 10:151–177, 1957.
- [17] W. Klopper R12 methods, Gaussian geminals, in *Modern Methods and Algorithms of Quantum Chemistry*, Ed. Grothendorst, Proceedings Winterschool 2000, NIC Series, Vol. 3, Jülich 2000.
- [18] H. Kümmel, K.-H. Lührmann, J.G. Zaboglitzky, Many Fermion Theory in expS (or Coupled Cluster) Form, *Physics Reports* **36**, No. 1, 1978, 1–63
- [19] W. Kutzelnigg, Theory of the expansion of wave functions in a Gaussian basis. *Int. J. Quant. Chem.*, **51**, 1994, 447–463.
- [20] W. Kutzelnigg, Error analysis and improvement of coupled cluster theory. *Theoretica Chimica Acta*, **80**, 1991, 349–386.

- [21] W. Kutzelnigg, J.D. Morgan III, Rates of convergence of the partial wave expansions for atomic correlation energies
- [22] C. Le Bris, M. Defranceschi, *Mathematical models and methods for ab initio Quantum Chemistry*, Lecture Notes in Chemistry, volume 74. Springer, 2000.
- [23] C. Le Bris, P.-L. Lions, From Atoms to crystals: a mathematical journey, *Bulletin AMS*, **42**, 291–363, 2005
- [24] M. Levin, Solutions of the Multiconfiguration equations in quantum chemistry, *Arch. Rational Mech. Anal.* **171**, 83–114, 2004
- [25] E.H. Lieb, B. Simon, The Hartree-Fock theory for Coulomb systems. *Comm. Math. Phys.*, **53**, 1977, 185–194.
- [26] P.-L. Lions. Solutions of Hartree-Fock equations for Coulomb systems. *Comm. Math. Phys.*, **109**, 1987, 33–97.
- [27] Y. Maday, G. Turinici, Error bars and quadratically convergent methods for the numerical solution of the Hartree-Fock equations, *Numer. Math.* **94**, 739–770, 2003
- [28] M. Schütz, H.-J. Werner, Linear Scaling Local CCSD. *J.Chem.Phys.*, **114**, 661–681, 2001.
- [29] M. Reed, B. Simon, *Methods of Modern Mathematical Physics IV: Analysis of Operators*, San Diego, Academic Press, 1978.
- [30] B. Simon, Schrödinger operators in the 20th century, *Journal Math. Phys.* **41**, 3523–3555, 2000.
- [31] A. Szabo, N.S. Ostlund, *Modern quantum chemistry*. Dover publications, New York, 1996.
- [32] H. Yserentant. On the regularity of the electronic Schrödinger equation in Hilbert spaces of mixed derivatives. *Numer. Math.* **98**, 2004, 731–759.
- [33] H. Yserentant. Sparse grid approximation for the numerical solution of the electronic Schrödinger equation, *Numer. Math.* **105**, 2005, 381–389.
- [34] H. Yserentant. *On the electronic Schrödinger equation*. Lecture Notes, Preprint Universität Tübingen, 2003.