Coupled-Cluster Theory Revisited

Mihaly Andras Csirik, Andre Laestadius

COUPLED-CLUSTER THEORY REVISITED*

MIHÁLY A. CSIRIK† AND ANDRE LAESTADIUS†

Abstract. We propose a comprehensive mathematical framework for Coupled-Cluster-type methods. These aim at accurately solving the many-body Schrödinger equation. The present work has two main aspects. First, we rigorously describe the discretization scheme involved in Coupled-Cluster methods using graph-based concepts. This allows us the discuss different methods in a unified and more transparent manner, including multireference methods. Second, we analyze the nonlinear equations of the single-reference Coupled-Cluster method using topological degree theory. We establish existence results and qualitative information about the solutions of these equations that also sheds light on some of the numerically observed behavior. For the truncated Coupled-Cluster method, we derive an energy error bound for approximate eigenstates of the Schrödinger equation.

1. Introduction. The Coupled-Cluster (CC) method is one of the most popular methods in computational quantum chemistry among Hartree–Fock (HF) and Density-Functional Theory (DFT). In its full generality, the quantum many-body problem is intractable, and it is one of the greatest challenges of quantum mechanics to devise practically useful methods to approximate the solutions of the many-body Schrödinger equation. Although the stationary Schrödinger equation itself is a linear eigenvalue problem, it is extremely high-dimensional even for a few particles and a low-dimensional one-particle space.[1] Here, we focus on those fermionic systems which are described by the so-called *molecular Hamilton operator*—on which most electronic-structure models are based in quantum chemistry. The Galerkin method applied to the Schrödinger equation (sometimes combined with an initial HF "guess") is branded Configuration Interaction (CI) in computational quantum chemistry; unfortunately, its applicability is limited due to the aforementioned high-dimensionality issue. The HF method is perhaps conceptually the simplest, whereby the ground state is approximated by minimizing the energy of the system over determinantal wavefunctions; the resulting Euler–Lagrange equations constitute a nonlinear eigenvalue problem that yields the HF ground state. HF theory has attracted much interest in the mathematical physics community, see e.g. [39, 40, 3, 4, 11, 54, 18, 33]. The spiritual successor to the statistical mechanics-motivated Thomas–Fermi theory—DFT—is the single most used method in quantum chemistry, and some of its mathematical aspects are also highly non-trivial [36, 15, 34, 35].

CC theory is a vast and highly active subfield of quantum chemistry, consisting of many variants and refinements. However, among the aforementioned methods, the CC approach has arguably received the least attention in the mathematics community.

1.1. Previous work. It is beyond the scope of this paper to give a historical review of the CC method and its vast number of variants. The interested reader is pointed to [21, 30, 7, 6, 53]. The survey article [56, pp. 99–184] is somewhat more mathematically-oriented and also proposes a rather general framework.

Our approach is based on the analysis of the single-reference CC method by R. Schneider [51]. In that work, a thorough description of the basic building blocks of the

*This work has received funding from the Norwegian Research Council through Grant Nos. 287906 (CCerror) and 262695 (CoE Hylleraas Center for Quantum Molecular Sciences).

†Hylleraas Centre for Quantum Molecular Sciences, Department of Chemistry, University of Oslo, P.O. Box 1033 Blindern, N-0315 Oslo, Norway (m.a.csirik@kjemi.uio.no, andre.laestadius@kjemi.uio.no)

[1] The dimension is $\binom{K}{N}$, where N is the number of particles, and K is dimension of the one-particle Hilbert space.

method, namely excitation- and cluster operators, and their algebraic and functional-analytic properties are given. Using the standard tools of nonlinear analysis, the CC equation (a nonlinear system of equations consisting of quartic polynomials) is formulated in terms of a locally strongly monotone and locally Lipschitz operator defined on an appropriate space. Under certain assumptions, this establishes local existence and uniqueness of a (Galerkin projected) solution of the equation and moreover *quasi-optimality* of the projected CC solution (Theorem 5.8 ibid.). Perhaps the most important contribution of [51] is a quadratic energy error estimate (Theorem 6.3 ibid.). It is worth emphasizing that Schneider's analysis is a local one: *"[...] 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 at best."* (ibid. p. 30)

Schneider's original analysis was carried out in the finite-dimensional case only. This was remedied in two subsequent articles by T. Rohwedder [50] and then by both of them [49]. The article [50] establishes important technical tools and rigorously proves that the untruncated CC problem is equivalent to the Full CI problem, i.e. to essentially the Schrödinger equation. Using the said tools, the subsequent paper [49] also establishes local uniqueness and existence of a solution to the truncated CC equations in a neighborhood of the untruncated CC solution. Further, [49] also extends the energy error estimates of [51] to the infinite-dimensional case.

This line of investigation was continued by S. Kvaal and A. L. in [32] for the Extended CC (ECC) method based on the "bivariational principle" [2]. In this case, strong monotonicity for the ECC mapping can be established so that quasi-optimality follows along similar lines as previously done by Schneider and Rohwedder. In the ECC theory, the traditional CC theory is recovered as a special case.

Furthermore, the local strong monotonicity-based analysis was applied to a variant of the traditional CC method by F. M. Faulstich et al. in 2019 [16], namely the Tailored Coupled-Cluster (TCC) method. The TCC approach splits the computational task into two parts: solving for the statically correlated wave function on a complete active space, and then on top of that accounting for the dynamical correlation using the CC method. Numerical investigations based on [16] were conducted in [17].

Finally, we mention the survey article [31] for more details on the use of local strong monotonicity-based methods in the analysis of CC methods.

1.2. Outline. It is our intention to present both known and new results in a self-contained manner and primarily with a mathematical audience in mind. In section 2, we describe the setting of the quantum-mechanical problems the CC theory is aimed at. The CC method typically takes a HF solution as an "input", so we give a brief discussion of the HF method in subsection 2.3. Next, subsection 2.4 gives a rough sketch of the most basic CI and CC methods.

We begin our discussion in subsection 3.1 with the definition of a partial order relation which will be used to encode the relevant transitions of the system, called *excitations*. This partial order, and the induced lattice operations will be used in subsection 3.2 to define the *excitation graph*, which fully describes the CC discretization scheme. We give a few examples of the generality of our concepts and also extend the definition of the excitation graph to the multireference (MR) case. After this, the corresponding *excitation operators* (subsection 3.3), *cluster operators* (subsection 3.4) and *cluster amplitude spaces* (subsection 3.5) are constructed, which are the essential building blocks for the formulation of any CC-like method.

In section 4, we give short derivations of the SRCC and JM-MRCC methods. We do so by generalizing the known procedure which is based on perturbation theory.

The analysis of the single-reference CC (SRCC) method begins in section 5. Basic properties of the SRCC mapping are discussed in subsection 5.1. After this, the local properties of the SRCC mapping are considered in subsection 5.2, such as strong monotonicity and topological index in both the non-degenerate-, and in the degenerate case. We also look at the complex SRCC mapping in subsection 5.3.

In subsection 5.4, an important class of homotopies is defined that can be used for proving the existence of a solution for the truncated SRCC mappings. In subsection 5.5 a homotopy is considered that was invented specifically to connect CC methods of different truncation levels. We prove an existence result and calculate the topological index of the homotopy. Finally, we derive an energy error estimate in subsection 5.6 using the results of Appendix E.

In Appendix A and Appendix B we briefly summarize the results that we use from finite-dimensional topological degree theory. In Appendix C we calculate various graph-theoretic properties of the excitation graph. In Appendix D we propose a method based on linear programming to select reference determinants for the multi-reference setting in an optimal way. In Appendix E we re-prove certain results related to the method used in subsection 5.5 and subsection 5.6.

2. Background. In this section we collect the concepts and results that are necessary for the forthcoming discussion. For proofs and more about the mathematical foundations of quantum mechanics, see e.g. [46, 47, 48, 20, 57, 22, 38].

We use the convention that complex inner products are conjugate-linear in their *second* arguments (as opposed to the convention used in physics). Complex conjugation is denoted by $\overline{z}$. The usual notation $B(a,r)$ is used for the open ball of radius r and center a, also $B^*(a,r) = B(a,r) \setminus \{a\}$ denotes the punctured ball. The spectrum of a linear operator A is written $\sigma(A)$, the elements of its discrete spectrum as $\mathcal{E}_n(A)$, where $n = 0, 1, 2, \ldots$, if A is bounded from below. We use the usual notation $[A, B] = AB - BA$ for the commutator. For normed spaces V and W, the symbol $\mathcal{L}(V, W)$ denotes normed space of *bounded* linear mappings $V \to W$ endowed with the operator norm $\|\cdot\|_{\mathcal{L}(V,W)}$. Furthermore, V^* denotes the (continuous) dual space.

2.1. Function spaces. In the context of many-body quantum mechanics, the complex Lebesgue-, and Sobolev spaces $L^2(\mathbb{R}^3)$ and $H^1(\mathbb{R}^3)$ are viewed as "one-particle spaces" which are used to define the N-particle *fermionic* spaces (see e.g. [37])

$$\mathfrak{L}^2 = \bigwedge^N L^2(\mathbb{R}^3), \quad \text{and} \quad \mathfrak{H}^1 = \mathfrak{L}^2 \cap H^1(\mathbb{R}^{3N}),$$

endowed with the inner products

$$\langle \Psi, \Phi \rangle = \int_{\mathbb{R}^{3N}} \Psi(\mathbf{X}) \Phi(\mathbf{X}) \, \mathrm{d}\mathbf{X}$$

and

$$\langle \Psi, \Phi \rangle_{\mathfrak{H}^1} = \langle \Psi, \Phi \rangle + \sum_{k=1}^{N} \int_{\mathbb{R}^{3N}} \boldsymbol{\nabla}_{\mathbf{x}_k} \Psi(\mathbf{X}) \cdot \boldsymbol{\nabla}_{\mathbf{x}_k} \Phi(\mathbf{X}) \, \mathrm{d}\mathbf{X},$$

respectively. Here, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbb{R}^{3N}$ and $\mathbf{z} \cdot \mathbf{w}$ denotes the usual inner product. Also, $\boldsymbol{\nabla}_{\mathbf{x}_k} = (\partial_{x_k^1}, \partial_{x_k^2}, \partial_{x_k^3})$ is the distributional gradient operator acting on the kth triple of the arguments. We define the second order Sobolev space as $\mathfrak{H}^2 = \mathfrak{L}^2 \cap H^2(\mathbb{R}^{3N})$.

Let $K \geq N$ or $K = \infty$ and assume that an L^2-orthonormal *(spin-)orbital set* $\mathcal{B} = \{\varphi_p\}_{p=1}^K \subset H^1(\mathbb{R}^3)$ is given. We define the subspace $H^1_K(\mathbb{R}^3) = \operatorname{Span} \mathcal{B} \subset H^1(\mathbb{R}^3)$. Corresponding to $\mathcal{B}$ we can construct the *determinantal wavefunctions* (a.k.a. Slater determinants)

$$\mathfrak{B} = \{\Phi_\alpha \in \mathfrak{H}^1 : 1 \leq \alpha_1 < \ldots < \alpha_N \leq K,\ \Phi_\alpha(\mathbf{X}) = N!^{-1/2} \det(\varphi_{\alpha_i}(\mathbf{x}_j))_{1\leq i,j\leq N}\}.$$

Then $\mathfrak{B}$ is $\mathfrak{L}^2$-orthonormal. Set

$$\mathfrak{H}^1_K = \operatorname{Span} \mathfrak{B} \subset \mathfrak{H}^1,$$

and we will use the convention that the subscript K is dropped if $\mathcal{B} \subset H^1(\mathbb{R}^3)$ forms a *basis*.

The negative exponent Sobolev space $\mathfrak{H}^{-1}$ will also be used in the sequel, which is given by the continuous dual space $(\mathfrak{H}^1)^*$ (see e.g. [1]). We will exploit that the dense continuous embeddings $\mathfrak{H}^1 \hookrightarrow \mathfrak{L}^2 \hookrightarrow \mathfrak{H}^{-1}$ hold true, i.e. they form a Gelfand triple. In this case, instead of the dual pairing $\langle \cdot, \cdot \rangle_{\mathfrak{H}^1 \times \mathfrak{H}^{-1}}$ it suffices to use $\langle \cdot, \cdot \rangle$ on $\mathfrak{H}^1 \times \mathfrak{L}^2$ and then extend to the whole space by density.

2.2. Hamilton operator. In this section, we introduce the model Hamilton operator for concreteness. Let $V, w : \mathbb{R}^3 \to \mathbb{R}$ be Kato class[2] potentials: $V, w \in L^{3/2}(\mathbb{R}^3) + L^\infty_\varepsilon(\mathbb{R}^3)$ with w even and define the quadratic form $\mathcal{E}$ on $\mathfrak{H}^1$ as

$$\mathcal{E}(\Psi) = \frac{1}{2}\|\boldsymbol{\nabla}\Psi\|^2 + \int_{\mathbb{R}^{3N}} \left(\sum_{i=1}^N V(\mathbf{x}_i) + \sum_{\substack{i,j=1 \\ i<j}}^N w(\mathbf{x}_i - \mathbf{x}_j) \right) |\Psi(\mathbf{X})|^2 \, \mathrm{d}\mathbf{X}$$

for any $\Psi \in \mathfrak{H}^1$. For every $\varepsilon > 0$, there is a $C_\varepsilon > 0$ so that Kato's inequality (see e.g. [18] for a detailed proof),

$$\frac{1-\varepsilon}{2}\|\boldsymbol{\nabla}\Psi\|^2 - C_\varepsilon\|\Psi\|^2 \leq \mathcal{E}(\Psi) \leq \frac{1+\varepsilon}{2}\|\boldsymbol{\nabla}\Psi\|^2 + C_\varepsilon\|\Psi\|^2 \quad \text{for all} \quad \Psi \in \mathfrak{H}^1,$$

holds true. This implies that the quadratic form induced by V and w is infinitesimally form bounded with respect to $-\triangle$ (and $\mathcal{E}$ is continuous and closed on $\mathfrak{H}^1$). The KLMN theorem implies that there exists a unique self-adjoint operator $\mathcal{H} : D(\mathcal{H}) \to \mathfrak{L}^2$ associated to $\mathcal{E}$, having form domain $Q(\mathcal{H}) = Q(\mathcal{E}) = \mathfrak{H}^1$ and being lower semibounded. This $\mathcal{H}$ is given by

$$(\mathcal{H}\Psi)(\mathbf{X}) = -\frac{1}{2}\sum_{i=1}^N \triangle_{\mathbf{x}_i}\Psi(\mathbf{X}) + \left(\sum_{i=1}^N V(\mathbf{x}_i) + \sum_{\substack{i,j=1 \\ i<j}}^N w(\mathbf{x}_i - \mathbf{x}_j) \right) \Psi(\mathbf{X}),$$

for all $\Psi \in D(\mathcal{H})$ and $\mathbf{X} \in \mathbb{R}^{3N}$. Kato's inequality implies that $\mathcal{H}$ is $\mathfrak{H}^1$-bounded: there is a constant $M > 0$, such that

$$\langle \mathcal{H}\Psi, \Phi \rangle \leq M\|\Psi\|_{\mathfrak{H}^1}\|\Phi\|_{\mathfrak{H}^1} \tag{2.1}$$

for all $\Psi, \Phi \in \mathfrak{H}^1$. Thus, $\mathcal{H}$ can be extended to a bounded mapping $\mathfrak{H}^1 \to \mathfrak{H}^{-1}$, which we denote with the same symbol. We say that $\Psi \in \mathfrak{H}^1$ and $\mathcal{E} \in \mathbb{R}$ satisfy the *weak Schrödinger equation* if $\langle \mathcal{H}\Psi, \Phi \rangle = \mathcal{E}\langle \Psi, \Phi \rangle$ for all $\Phi \in \mathfrak{H}^1$.

[2]By definition $f \in L^{3/2}(\mathbb{R}^3) + L^\infty_\varepsilon(\mathbb{R}^3)$, if for every $\varepsilon > 0$ there is an $f_1 \in L^{3/2}(\mathbb{R}^3)$ and $f_2 \in L^\infty(\mathbb{R}^3)$ with $\|f_2\|_\infty < \varepsilon$ so that $f = f_1 + f_2$.

As far as the finite-dimensional case $K < \infty$ is concerned, we simply consider the Galerkin projection of the weak Schrödinger equation. More precisely, let $\mathfrak{H}^1_K \subset \mathfrak{H}^1$ be as defined in subsection 2.1. Then $\Psi \in \mathfrak{H}^1_K$ and $\mathcal{E} \in \mathbb{R}$ are said to satisfy the *projected Schrödinger equation* if $\langle \mathcal{H}\Psi, \Phi\rangle = \mathcal{E}\langle \Psi, \Phi\rangle$ for all $\Phi \in \mathfrak{H}^1_K$.

The so-called (electronic) molecular Hamilton operator $\mathcal{H}$ corresponds to the special case

$$V(\mathbf{x}) = -\sum_{j=1}^{M} \frac{Z_j}{|\mathbf{x} - \mathbf{r}_j|} \quad \text{and} \quad w(\mathbf{x}) = \frac{1}{|\mathbf{x}|},$$

where $Z_j \in \mathbb{N}$ $(j = 1, \dots, M)$ and $\mathbf{r}_1, \dots, \mathbf{r}_M \in \mathbb{R}^3$ denote the charges and the positions of the $M \in \mathbb{N}$ nuclei.

2.3. The Hartree–Fock method. In practice, the CC method usually takes the HF orbitals as an input and therefore the performance of the method hinges on this preliminary HF calculation. Here, we collect the basic facts about the HF method that will be used later on.

The HF method[3] is based on the minimiziation of the energy over the determinantal wavefunctions. It is fairly easy to see by direct calculation that the energy $\mathcal{E}(\Psi)$ of a determinantal wavefunction $\Psi = N!^{-1/2} \det(\varphi_i(\mathbf{x}_j))_{1\leq i,j\leq N}$, with $\{\psi_j\}_{j=1}^N$ being L^2-orthonormal is given by

$$\mathcal{E}(\psi_1, \dots, \psi_N) := \mathcal{E}(\Psi) = \frac{1}{2}\sum_{j=1}^{N} \int_{\mathbb{R}^3} |\boldsymbol{\nabla}\psi_j|^2 + \sum_{i=1}^{N} \int_{\mathbb{R}^3} V|\psi_i|^2$$
$$+ \frac{1}{2} \iint_{\mathbb{R}^3\times\mathbb{R}^3} \left[\sum_{i,j=1}^{N} |\psi_i(\mathbf{x})|^2 |\psi_i(\mathbf{y})|^2 - \left| \sum_{i=1}^{N} \psi_i(\mathbf{x})\overline{\psi_i(\mathbf{y})} \right|^2 \right] w(\mathbf{x}-\mathbf{y})\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y},$$

see e.g. [18, 33]. Hence the task is to determine the *HF energy*

$$\mathcal{E}_{\mathrm{HF}} = \mathcal{E}(\Phi_{\mathrm{HF}}) = \min_{\substack{\psi_1,\dots,\psi_N \\ \text{orthonormal}}} \mathcal{E}(\psi_1, \dots, \psi_N), \tag{2.2}$$

along with a *HF minimizer* (or *HF determinant*) Φ_{HF}. The existence of a HF minimizer Φ_{HF} is guaranteed for the case of positive ions and neutral atoms (corresponding to the electronic molecular Hamilton operator).

THEOREM 2.1. *[39] If $N < Z + 1$, then there exists a minimizer Φ_{HF} to (2.2).*

Recently, much more has been discovered about the mathematical structure of the HF energy functional [18, 33]. As usual, a minimizer satisfies the corresponding Euler–Lagrange equations (which are called *Hartree-Fock equations* in this context). In practice, it is this system of nonlinear integro-differential equations which is discretized and solved. To describe the HF equations, we make a definition that will be convenient for later purposes. Fix $\Phi(\mathbf{X}) = N!^{-1/2} \det(\varphi_i(\mathbf{x}_j))_{1\leq i,j\leq N}$, where $\{\varphi_p\}_{p=1}^N \subset H^1(\mathbb{R}^3)$ is L^2-orthonormal. Define a self-adjoint operator $F_\Phi : L^2(\mathbb{R}^3) \to L^2(\mathbb{R}^3)$ with domain $D(F_\Phi) = H^2(\mathbb{R}^3)$ via the instruction

$$(F_\Phi\psi)(\mathbf{x}) = -\frac{1}{2}\triangle\psi(\mathbf{x}) + V(\mathbf{x})\psi(\mathbf{x})$$
$$+ \sum_{i=1}^{N} \int_{\mathbb{R}^3} w(\mathbf{x}-\mathbf{y})(|\varphi_i(\mathbf{y})|^2 - \varphi_i(\mathbf{x})\overline{\varphi_i(\mathbf{y})})\psi(\mathbf{x})\, \mathrm{d}\mathbf{y}$$

[3] a.k.a. Self-Consistent Field (SCF) method

for all $\psi \in D(F_\Phi)$ and all $\mathbf{x} \in \mathbb{R}^3$. The operator F_Φ is called the *mean-field operator*.[4] The form domain of F_Φ is $H^1(\mathbb{R}^3)$. The essential spectrum of F_Φ is $[0, +\infty)$. We summarize the basic properties of the mean-field operator in the next theorem. Let

$$\mu_n(F_\Phi) = \min_{\psi_1,\dots,\psi_n \in H^1(\mathbb{R}^3)} \max_{\substack{\psi \in \mathrm{Span}\{\psi_1,\dots,\psi_n\} \\ \|\psi\|_{L^2(\mathbb{R}^3)}=1}} \langle F_\Phi \psi, \psi \rangle$$

denote the min-max values of F_Φ.

THEOREM 2.2. *Assume that there exists a HF minimizer*

$$\Phi_{\mathrm{HF}} = N!^{-1/2} \det(\varphi_i(\mathbf{x}_j))_{1 \le i,j \le N}.$$

(i) *(Hartree-Fock equations) There exists a unitary matrix* $\mathbf{U} \in \mathbb{C}^{N \times N}$ *so that with*

$$(\widetilde{\varphi}_1, \dots, \widetilde{\varphi}_N) = \mathbf{U}(\varphi_1, \dots, \varphi_N),$$

$\widetilde{\varphi}_i$ are eigenfunctions of F_Φ corresponding to its N lowest eigenvalues $\lambda_1 \le \dots \le \lambda_N$,

$$F_\Phi \widetilde{\varphi}_i = \lambda_i \widetilde{\varphi}_i, \quad \text{for all } i = 1, \dots, N. \tag{2.3}$$

(ii) *(Aufbau principle) If $\mu_{N+1}(F_\Phi)$ is an eigenvalue of F_Φ, then*

$$\lambda_N = \mu_N(F_\Phi) < \mu_{N+1}(F_\Phi) \le 0.$$

The eigenvalue λ_N is called the *highest occupied molecular orbital* (HOMO) and λ_{N+1} the *lowest unoccupied molecular orbital* (LUMO). Their difference, $\varepsilon_{\min} := \lambda_{N+1} - \lambda_N$ is called the *HOMO-LUMO gap*, which is an important quantity in quantum chemistry [4].

The N-particle "lifted" version of the mean-field operator is called the *Fock operator* and is defined as the self-adjoint operator $\mathcal{F}_\Phi : \mathfrak{L}^2 \to \mathfrak{L}^2$ with domain $D(\mathcal{F}_\Phi) = \mathfrak{H}^2$ via

$$\mathcal{F}_\Phi = F_\Phi \otimes I \otimes \dots \otimes I + \dots + I \otimes \dots \otimes I \otimes F_\Phi.$$

Henceforth we omit Φ from the notation, and let $\mathcal{F} := \mathcal{F}_\Phi$ whenever Φ is clear from the context. It is immediate that the HF determinant $\Phi_0 := \Phi_{\mathrm{HF}}$ is an eigenfunction of $\mathcal{F}$,

$$\mathcal{F}\Phi_0 = \Lambda_0 \Phi_0, \quad \text{with} \quad \Lambda_0 = \sum_{i=1}^{N} \lambda_i.$$

The Fock operator gives rise to a splitting of the molecular Hamilton operator

$$\mathcal{H} = \mathcal{F} + \mathcal{W}, \quad \text{where} \quad \mathcal{W} = \mathcal{H} - \mathcal{F} \tag{2.4}$$

is called the *fluctuation operator*.

For the rest of the section, we consider the finite-dimensional case. In practice, the Galerkin projection of the Hartree–Fock equations (2.3) are solved to obtain the orbitals $\{\varphi_i\}_{i=1}^N$. Since the mean-field operator F_Φ is self-adjoint, its eigenfunctions can be used to extend these orbitals to an orthonormal basis $\{\varphi_i\}_{i=1}^K \subset H^1_K(\mathbb{R}^3)$. In this orbital basis, the Fock operator takes the diagonal form $\mathcal{F} = \sum_{i=1}^K \lambda_i a_i^\dagger a_i$, where

[4] It is sometimes called the Fock operator, but we reserve that name for its N-particle version.

$a_i^\dagger$ and a_i are the fermionic creation-, and annihilation operators corresponding to the orbital φ_i.

Furthermore, if $\Phi_\alpha = N!^{-1/2}\det(\varphi_{\alpha_i}(\mathbf{x}_j))_{1\le i,j\le N}$ with $\alpha_1 < \ldots < \alpha_N$ that is obtained from $\Phi_0 = N!^{-1/2}\det(\varphi_i(\mathbf{x}_j))_{1\le i,j\le N}$ by swapping r orbitals φ_{I_j} with φ_{A_j} where $I_j \in \{1,\ldots,N\}$, $A_j \notin \{1,\ldots,N\}$ $(j = 1,\ldots,r \le N)$, then

$$\mathcal{F}\Phi_\alpha = \Lambda_\alpha\Phi_\alpha, \quad \text{with} \quad \Lambda_\alpha = \sum_{i=1}^{N}\lambda_{\alpha_i} = \Lambda_0 + \varepsilon_\alpha, \quad \varepsilon_\alpha = \sum_{j=1}^{r}\lambda_{A_j} - \lambda_{I_j}. \tag{2.5}$$

Note that in the infinite-dimensional case, it might not be possible to construct a complete eigenbasis $\{\varphi_i\}_{i=1}^\infty \subset H^1(\mathbb{R}^3)$ for the mean-field operator F_Φ due to the presence of the essential spectrum.

2.4. The CI and the CC method. We now give a very rough description of the single-reference CI and CC methods. For the rigorous derivations, we refer to section 4.

In a preliminary step—typically using the HF method—the *reference determinant*

$$\Phi_0 = N!^{-1/2}\det(\varphi_i(\mathbf{x}_j))_{1\le i,j\le N}$$

is constructed and normalized so that $\|\Phi_0\| = 1$. We restrict our discussion here to the case when relevant function spaces are real. The *occupied orbitals* $\mathcal{B}_{\rm occ} = \{\varphi_p\}_{p=1}^N \subset H^1(\mathbb{R}^3)$ are extended to a basis $\mathcal{B} = \{\varphi_p\}_{p=1}^K \subset H^1_K(\mathbb{R}^3)$ by adding $K-N$ *virtual orbitals* $\mathcal{B}_{\rm virt} = \{\varphi_p\}_{p=N+1}^K$, so that $\mathcal{B} = \mathcal{B}_{\rm occ} \cup \mathcal{B}_{\rm virt}$. Here, $K = \infty$ is allowed. The orthonormal set $\mathcal{B}$ generates the determinantal wavefunctions $\mathfrak{B}$ and the subspace $\mathfrak{H}^1_K \subset \mathfrak{H}^1$ (see subsection 2.1). For later convenience, we introduce the space $\mathfrak{H}^{1,\perp}$ as the $\mathfrak{L}^2$-orthogonal complement of $\mathrm{Span}\{\Phi_0\}$ in $\mathfrak{H}^1$. Further, we also set $\mathfrak{H}^{1,\perp}_K = \mathfrak{H}^{1,\perp} \cap \mathfrak{H}^1_K$.

In both the CI and the CC method, the Schrödinger equation $\mathcal{H}\Psi = \mathcal{E}\Psi$ is solved based on the reference wavefunction Φ_0. For simplicity,[5] we consider the case when Ψ is sought after in the form $\Psi = \Phi_0 + \underline{\Psi}$, where $\langle\underline{\Psi},\Phi_0\rangle = 0$. In other words, Ψ is calculated via a *correction* $\underline{\Psi}$ to Φ_0. Note that $\langle\Psi,\Phi_0\rangle = 1$, which is called the *intermediate normalization* condition. If the "targeted" wavefunction Ψ happens to be orthogonal to the reference determinant Φ_0, then the Ansatz $\Psi = \Phi_0 + \underline{\Psi}$ cannot yield a solution (see, however, Lemma 5.1).

The *Full Configuration Interaction* (FCI) method can be summarized as follows: find $\underline{\Psi} \in \mathfrak{H}^{1,\perp}_K$ such that

$$\langle\mathcal{H}(\Phi_0 + \underline{\Psi}),\Phi\rangle = \mathcal{E}_{\rm CI}\langle\Phi_0 + \underline{\Psi},\Phi\rangle \quad \text{for all } \Phi \in \mathfrak{H}^{1,\perp}_K. \tag{2.6}$$

Here, $\mathcal{E}_{\rm CI} = \|\Psi\|^{-2}\langle\mathcal{H}\Psi,\Psi\rangle$ is called the *CI-*, or *variational energy*. The *projected CI* method is simply the Galerkin projection of the previous problem to some finite dimensional subspace $\mathfrak{V}_d \subset \mathfrak{H}^{1,\perp}_K$, i.e. to find $\underline{\Psi}_d \in \mathfrak{V}_d$ such that

$$\langle\mathcal{H}(\Phi_0 + \underline{\Psi}_d),\Phi_d\rangle = \mathcal{E}_{d,\rm CI}\langle\Phi_0 + \underline{\Psi}_d,\Phi_d\rangle \quad \text{for all } \Phi_d \in \mathfrak{V}_d. \tag{2.7}$$

The choice of the Galerkin subspace $\mathfrak{V}_d$ is typically based on so-called truncation rank, for instance $\mathfrak{V}_d = \mathfrak{V}_{\rm SD}$, is the span of singly-, and doubly excited determinants in $\mathfrak{B}$. The corresponding (projected) CI method in this case is designated as "CISD".

[5] Although the CI method is more general.

The CI equations are more commonly expressed using *cluster operators*. A cluster operator $C : \mathfrak{L}^2_K \to \mathfrak{L}^2_K$ is a bounded linear operator that is a linear combination of special products of fermionic creation and annihilation operators $a_i^\dagger$ and a_i, so that the action of each such product is to replace some occupied orbitals $\mathcal{B}_{\mathrm{occ}}$ with the same number of virtual orbitals $\mathcal{B}_{\mathrm{virt}}$ (see Remark 3.18). A cluster operator C can therefore be parametrized with the said linear-combination coefficients, denoted by the lower case c and are called *cluster amplitudes*. The vector space of all cluster amplitudes are denoted by $\mathbb{V}$. There is a one-to-one correspondence between functions in $\mathfrak{L}^{2,\perp}$ (resp. $\mathfrak{H}^{1,\perp}_K$) and functions of the form $C\Phi_0$, where $C : \mathfrak{L}^2_K \to \mathfrak{L}^2_K$ (resp. $C : \mathfrak{H}^1_K \to \mathfrak{H}^1_K$) is a cluster operator. Therefore, (2.6) can be expressed as a follows: find a cluster operator C (or, equivalently cluster amplitudes c), such that

$$\langle \mathcal{H}(I+C)\Phi_0, S\Phi_0\rangle = \mathcal{E}_{\mathrm{CI}}\langle (I+C)\Phi_0, S\Phi_0\rangle \quad \text{for all cluster operators } S.$$

Although this might seem an unnecessary complication at first, cluster operators are essential for the formulation of the CC method.

In the CC method, the "exponential Ansatz" is assumed for the intermediately normalized wavefunction Ψ. Substituting $\Psi = e^T\Phi_0$ into the Schrödinger equation, where T is a cluster operator, we get

$$\mathcal{H}e^T\Phi_0 = \mathcal{E}_{\mathrm{CC}}e^T\Phi_0, \tag{2.8}$$

for some $\mathcal{E}_{\mathrm{CC}} \in \mathbb{R}$. First, to determine $\mathcal{E}_{\mathrm{CC}}$ we premultiply (2.8) by e^{-T} (e^T is always invertible), and take the inner product with Φ_0 to obtain the *CC energy*

$$\mathcal{E}_{\mathrm{CC}} := \mathcal{E}_{\mathrm{CC}}(t) = \langle e^{-T}\mathcal{H}e^T\Phi_0, \Phi_0\rangle, \tag{2.9}$$

where we used the normalization $\|\Phi_0\| = 1$. Second, by premultiplying (2.8) by e^{-T} again, but now testing against functions in $\mathfrak{H}^{1,\perp}_K$, we get the *Full CC* (FCC) method: find cluster amplitudes $t_* \in \mathbb{V}$ such that

$$\langle e^{-T_*}\mathcal{H}e^{T_*}\Phi_0, S\Phi_0\rangle = 0, \quad \text{for all } s \in \mathbb{V}. \tag{2.10}$$

The *projected CC* method is the Galerkin projection of the FCC problem with respect to some subspace $\mathbb{V}_d \subset \mathbb{V}$. More precisely, the task is to find $t_{d,*} \in \mathbb{V}_d$ such that

$$\langle e^{-T_{d,*}}\mathcal{H}e^{T_{d,*}}\Phi_0, S_d\Phi_0\rangle = 0 \quad \text{for all } s_d \in \mathbb{V}_d. \tag{2.11}$$

For the moment, we denote the corresponding CC energy by $\mathcal{E}_{d,\mathrm{CC}}$. Again, $\mathbb{V}_d$ is based on some truncation, such as SD, in which case the corresponding method is called "CCSD".

We now discuss the relation between CI and CC. It was shown that the FCI (2.6) and the FCC (2.10) methods are equivalent, see [49, Theorem 5.3].

THEOREM 2.3 (Equivalence of FCI and FCC). *The problems (2.6) and (2.10) are equivalent, and the full CC solution $\Psi = e^{T_*}\Phi_0$ satisfies $\mathcal{E}_{\mathrm{CC}}(t_*) = \mathcal{E}_{\mathrm{CI}}$.*

However, the corresponding Galerkin-projected problems are *not* equivalent. Further, while $\mathcal{E}_{\mathrm{CI}} \leq \mathcal{E}_{d,\mathrm{CI}}$ due to the Rayleigh–Ritz variational principle, the same is not true for the CC method and numerical experience undoubtedly shows that *there is no obvious relation* in general between $\mathcal{E}_{\mathrm{CC}} = \mathcal{E}_{\mathrm{CI}}$ and $\mathcal{E}_{d,\mathrm{CC}}$; this last phenomenon is called the *nonvariational property* of CC theory. Note that according to Theorem 2.3, FCC *is* variational.

Despite this, the gains of CC over CI are significant. First, by construction, the CC method is size-consistent, even when truncated [51, Theorem 4.10]. This property is crucial for the precise determination of various chemical properties. Second, the evaluation of expressions involving the *similarity-transformed Hamilton operator* $e^{-T}\mathcal{H}e^{T}$ is greatly eased by the formula

$$e^{-T}\mathcal{H}e^{T} = \sum_{j=0}^{4} \frac{1}{j!}[\mathcal{H}, T]_{(j)}, \tag{2.12}$$

see [51, Theorem A.1],[6] where the iterated commutators are given by $[\mathcal{H}, T]_{(0)} = \mathcal{H}$ and $[\mathcal{H}, T]_{(j)} = [[\mathcal{H}, T]_{(j-1)}, T]$ for $j \geq 1$. Equation (2.12) may be referred to as the termination of the Baker–Campbell–Hausdorff series (2.12), and it makes the computer implementation of CC methods feasible even for moderately sized systems. In particular, this, and the Slater–Condon rules imply that the CC energy can be computed as[7]

$$\mathcal{E}_{\mathrm{CC}}(t) = \langle \mathcal{H}(I + T_1 + T_2 + \tfrac{1}{2}T_1^2)\Phi_0, \Phi_0\rangle. \tag{2.13}$$

Furthermore, (2.12) also implies that the polynomial system (2.10) (and hence its Galerkin projection (2.11)) is quartic in terms of the cluster amplitudes t. Despite their apparent simplicity, the CC equations usually involve many complicated terms and even their assembly is a nontrivial task. In summary, the CC method approximates an extremely high-dimensional linear problem (2.6) by a low-dimensional nonlinear problem (2.11).

Since most literature on CC theory is somewhat vague on how the actual discretization scheme is set up, we begin our discussion by introducing a framework that allows us to describe the CC discretization rigorously and also without the use of second-quantized formalism.

3. Coupled-Cluster discretization. Using an appropriate string of creation and annihilation operators, any fermionic state can be changed to any other one (see e.g. [21, 55]). In our context, a set of N occupied orbitals is given; its complement is called the set of virtual orbitals. The action of an excitation operator consists of annihilating a few occupied orbitals and creating the same number of virtual orbitals (hence the particle number N is conserved). A de-excitation operator amounts to the reverse action: annihilating some virtual orbitals and creating the same number of occupied ones. Obviously, any N-particle state can be achieved by acting with an appropriate excitation operator on the "reference state", which is the N-particle state composed of all the occupied orbitals. However, it might also be possible to arrive at the same state from another state through successive excitations. The concrete relationships are nontrivial and this section is devoted to their description.

3.1. Excitation order. Let Λ be a countable set called the *orbital set* and let 2^{Λ} denote the power set of Λ. In concrete examples, we will often use the numbers $\Lambda = \{1, 2, 3, \ldots\}$ to label the elements of Λ for the sake of simplicity, and set $K = |\Lambda|$. Let $N \geq 1$ denote the number of particles, and set $S = \{\alpha \in 2^{\Lambda} : |\alpha| = N\}$, the elements of which are called *states*. The particle number N is assumed to be fixed

[6] Their proof is straightforward to adapt to the more general Hamilton operator defined in subsection 2.2.

[7] Actually, the term $\langle \mathcal{H}T_1\Phi_0, \Phi_0\rangle$ vanishes if Φ_0 is the Hartree–Fock solution (Brillouin theorem).

throughout. Fix $M \geq 1$ *reference states*

$$\Omega = \{0_1, \dots, 0_M\} \subset S.$$

For every $m = 1, \dots, M$ define

$$L_m = S \setminus (\Omega \setminus \{0_m\})$$

and on it, the partial order relation

$$\alpha \preceq_m \beta \quad \iff \quad \underline{\beta}_m \subset \underline{\alpha}_m \quad \text{and} \quad \overline{\alpha}^m \subset \overline{\beta}^m$$

for any $\alpha, \beta \in L_m$, where

$$\underline{\alpha}_m = \alpha \cap 0_m \quad \text{and} \quad \overline{\alpha}^m = \alpha \cap (0_m)^c,$$

and the complement is to be understood relative to Λ. According to commonly used nomenclature, we call $\underline{\alpha}_m$ the *occupied part of* α *w.r.t.* 0_m and $\overline{\alpha}^m$ the *virtual part of* α *w.r.t.* 0_m. This partial order relation is a generalization of [50, Definition 4.2]. By definition, $L_m = \{\alpha \in S : 0_m \preceq_m \alpha\}$ and for the sake of convenience, we introduce the notations $\overline{S} = S \setminus \Omega$ and $\overline{L}_m = L_m \setminus \{0_m\}$. Note that the reference states are defined *not* to be comparable with respect to $\preceq_m$ with each other.

The partial order $\preceq_m$ generates the *join* and *meet* lattice operations

$$\alpha \vee_m \beta = \left(\underline{\alpha}_m \cap \underline{\beta}_m\right) \cup \left(\overline{\alpha}^m \cup \overline{\beta}^m\right),$$
$$\alpha \wedge_m \beta = \left(\underline{\alpha}_m \cup \underline{\beta}_m\right) \cup \left(\overline{\alpha}^m \cap \overline{\beta}^m\right),$$

for all $\alpha, \beta \in L_m$. Furthermore, we introduce the orthocomplementation $\alpha^\perp = \Lambda \setminus \alpha$.

For the so-called *single-reference* (SR) case, $M = 1$ and we will make the convention that all the m indices are dropped from the notation. For the next result, we extend $\preceq$, $\vee$ and $\wedge$ to the whole 2^Λ.

PROPOSITION 3.1. *The structure* $B = (2^\Lambda, \vee, \wedge, 0, 1, \perp)$ *is a Boolean algebra, that is, a distributive, bounded lattice in which the de Morgan laws hold true. Here, we set* $1 := \Lambda$*, the identity for* $\wedge$*.*

A similar statement holds true in the *multi-reference* (MR) case, for the individual structures $B_m = (2^\Lambda, \vee_m, \wedge_m, 0_m, 1, \perp)$. Even though the algebraic structure on B is nice, the subset S loses this structure. In fact, S is *not* a sublattice of B, since for example $\alpha \vee_m \beta, \alpha \wedge_m \beta \notin S$ for distinct α and β with $\underline{\alpha} = \underline{\beta} = \emptyset$. The reason why we stated Proposition 3.1, however, is because we will exploit the operational rules for $\vee$, $\wedge$ and $\perp$ in a few occasions; for instance, in the following trivial result.

LEMMA 3.2. *Let* $\gamma, \beta \in 2^\Lambda$ *be such that* $\beta \preceq_m \gamma$*. Then,* $\alpha \vee_m \beta = \gamma$ *if and only if* $\alpha = \beta^\perp \wedge_m \gamma$*.*

Proof. We have

$$\alpha \vee_m \beta = (\gamma \wedge_m \beta^\perp) \vee_m \beta = (\gamma \vee_m \beta) \wedge_m (\beta^\perp \vee_m \beta) = (\gamma \vee_m \beta) \wedge_m 1 = \gamma \vee_m \beta = \gamma,$$

where in the last step we used $\beta \preceq_m \gamma$. Further, if $\alpha' \vee_m \beta = \gamma$ as well, then $\alpha' \vee_m \beta = \alpha \vee_m \beta$. By joining $\beta^\perp$ to both sides, we get $\alpha' = \alpha$. ∎

The poset $(L_m, \preceq_m)$ also admits a rank function which makes it a *graded poset*. This means that the *rank function* $\mathrm{rk}_m : L_m \to \mathbb{N}$ satisfies $\mathrm{rk}_m(\alpha) < \mathrm{rk}_m(\beta)$ whenever $\alpha \prec_m \beta$, and $\mathrm{rk}_m(\beta) = \mathrm{rk}_m(\alpha) + 1$ if there is no element γ such that $\alpha \prec_m \gamma \prec_m \beta$. The choice $\mathrm{rk}_m(\alpha) = |\overline{\alpha}^m|$ is easily seen to satisfy the requirements. Obviously, the maximum value that $\mathrm{rk}_m(\alpha)$ can take is N. For a geometric description of the rank function, see Appendix D.

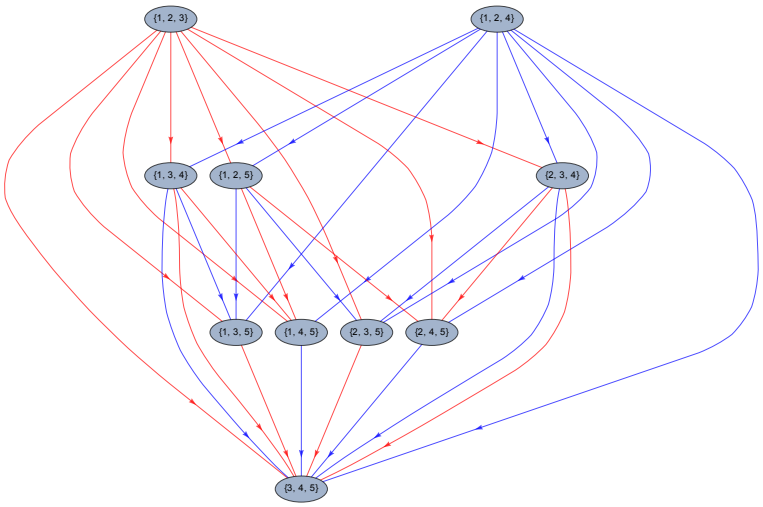

FIG. 1. *Full multi-reference excitation multigraph for* $\Lambda = \{1, \ldots, 5\}$ *and* $0_1 = \{1, 2, 3\}$, $0_2 = \{1, 2, 4\}$. *The edges corresponding to* 0_1 *and* 0_2 *are shown in red and blue, respectively.*

3.2. Excitation graphs. As we remarked in the previous section, L_m fails to be a sublattice of the Boolean algebra B_m. Therefore, let us consider pairs $(\alpha, \beta) \in L_m \times L_m$ for which $\alpha \vee_m \beta \in L_m$. In other words, pairs $(\alpha, \beta) \in L_m \times L_m$ for which $|\underline{\alpha}_m \cap \underline{\beta}_m| + |\overline{\alpha}^m \cup \overline{\beta}^m| = N$, or, equivalently,

$$|\underline{\alpha}_m \cup \underline{\beta}_m| + |\overline{\alpha}^m \cap \overline{\beta}^m| = N. \tag{3.1}$$

While still $\alpha \vee_m \beta \in L_m$ in the case $\overline{\alpha}^m \cap \overline{\beta}^m \neq \emptyset$, we wish to avoid that possibility on based physical grounds. Namely, on the account of the Pauli exclusion principle, since roughly speaking such an operation would introduce a repeated row in the determinantal wavefunction and that would render it identically zero. Therefore, we restrict our attention to the set

$$\mathcal{L}_m = \{(\alpha, \beta) \in L_m \times L_m : \ |\underline{\alpha}_m \cup \underline{\beta}_m| = N \ \text{ and } |\overline{\alpha}^m \cap \overline{\beta}^m| = 0\}. \tag{3.2}$$

Hence, if $(\alpha, \beta) \in \mathcal{L}_m$, then we have $\alpha \vee_m \beta \in L_m$. The set $\mathcal{L}_m$ is symmetric to the diagonal (which it does not contain), and $(0_m, \alpha), (\alpha, 0_m) \in \mathcal{L}_m$ for any $\alpha \in L_m$. Furthermore, the rank function rk_m is additive on $\mathcal{L}_m$ in the sense that

$$\mathrm{rk}_m(\alpha \vee_m \beta) = \mathrm{rk}_m(\alpha) + \mathrm{rk}_m(\beta),$$

for any $(\alpha, \beta) \in \mathcal{L}_m$. This property may also seen to be a reason why we want to exclude the case $\overline{\alpha}^m \cap \overline{\beta}^m \neq \emptyset$. Indeed, it could also be taken as the *definition* of $\mathcal{L}_m$.

PROPOSITION 3.3. *The set* $\mathcal{L}_m$ *can be written as*

$$\mathcal{L}_m = \{(\alpha, \beta) \in L_m \times L_m : \alpha \vee_m \beta \in L_m \text{ and } \mathrm{rk}_m(\alpha \vee_m \beta) = \mathrm{rk}_m(\alpha) + \mathrm{rk}_m(\beta)\}.$$

Proof. Let $\mathcal{L}'_m$ denote the set on the right hand side. Then, it is clear from the above that $\mathcal{L}_m \subset \mathcal{L}'_m$. Conversely, suppose that $(\alpha, \beta) \in \mathcal{L}'_m$. Then, $|\overline{\alpha}^m \cup \overline{\beta}^m| = |\overline{\alpha}^m| + |\overline{\beta}^m|$, from which $|\overline{\alpha}^m \cap \overline{\beta}^m| = 0$. Since $\alpha \vee_m \beta \in L_m$, (3.1) holds true, and we have that $|\underline{\alpha}_m \cup \underline{\beta}_m| = N$, so $(\alpha, \beta) \in \mathcal{L}_m$. Hence, $\mathcal{L}_m \supset \mathcal{L}'_m$. ∎

The set $\mathcal{L}_m$ is used for our main definition.

DEFINITION 3.4. *The digraph* $G_m^{\text{full}} = (L_m, E_m^{\text{full}})$ *is called the* full (SR) excitation graph w.r.t. 0_m, *where*

$$E_m^{\text{full}} = \{(\beta, \alpha \vee_m \beta) \in L_m \times L_m : (\alpha, \beta) \in \mathcal{L}_m,\ \alpha \neq 0_m\}.$$

A subgraph $G_m = (L_m, E_m)$, $E_m \subset E_m^{\text{full}}$ *is said to be an* (SR) excitation (sub)graph w.r.t. 0_m.

Notice that we excluded $\alpha = 0_m$ to omit loop edges. Lemma 3.2 has the following refinement on the excitation graph.

LEMMA 3.5. *Let* $(\beta, \gamma) \in E_m^{\text{full}}$. *Then* $\alpha = \beta^\perp \wedge_m \gamma \in L_m$ *is the unique* α *such that* $\alpha \vee_m \beta = \gamma$.

Proof. Using Lemma 3.2, we can uniquely solve the equation $\alpha \vee_m \beta = \gamma$ for α to obtain $\alpha = \beta^\perp \wedge_m \gamma \in 2^\Lambda$. Therefore, $(\beta, \alpha \vee_m \beta) \in E_m^{\text{full}}$, which implies that $\alpha \in L_m$ using the definition of E_m^{full}. ∎

COROLLARY 3.6. *The digraph* G_m^{full} *does not contain parallel edges.*

Various graph-theoretic quantities of the single-reference excitation graph are calculated in Appendix C.

A digraph $G = (V, E)$ is said to be *transitive* if $(u, v) \in E$ and $(v, w) \in E$ imply $(u, w) \in E$. It follows by induction that, if G is transitive, and whenever G contains a directed path $((v_0, v_1), (v_1, v_2), \ldots, (v_{n-1}, v_n))$, then $(v_0, v_n) \in E$.

PROPOSITION 3.7. *The digraph* G_m^{full} *is transitive.*

Proof. Suppose that $(\gamma_0, \gamma_1) \in E_m^{\text{full}}$ and $(\gamma_1, \gamma_2) \in E_m^{\text{full}}$. Then there exists α and β such that $\gamma_1 = \alpha \vee_m \gamma_0$ and $\gamma_2 = \beta \vee_m \gamma_1$. Since $\gamma_2 = (\alpha \vee_m \beta) \vee \gamma_0$ by the associativity of $\vee_m$, it follows easily from Proposition 3.3 that $(\gamma_0, \alpha \vee_m \beta) \in \mathcal{L}_m$. Therefore,

$$(\gamma_0, \gamma_2) = (\gamma_0, (\alpha \vee_m \beta) \vee_m \gamma_0) \in E_m^{\text{full}},$$

which is what we wanted to show. ∎

Transitivity of certain subgraphs, and of G_m^{full} itself will come up later, since vaguely speaking this property will imply the algebraic closedness of the set of excitation operators that we attach to the edges (see subsection 3.3 and subsection 3.4).

We label the edges of G_m^{full} with their corresponding α. Thus, to every directed edge $(\beta, \alpha \vee_m \beta) \in E_m^{\text{full}}$ there corresponds a map $x_{m,\alpha} : L_m \to L_m$ defined with the instruction $x_{m,\alpha}(\beta) = \alpha \vee_m \beta$. This way, the digraph G_m^{full} may be interpreted as a commutative diagram (cf. subsection 3.3). Note that a label $x_{m,\alpha}$ may appear on multiple edges.

Furthermore, for any subgraph $G_m = (L_m, E_m)$, we introduce the *set of excitations* $\Xi(G_m) \subset \overline{L}_m$ *of* G_m via

$$\Xi(G_m) = \{\alpha \in \overline{L}_m : (\beta, \alpha \vee_m \beta) \in E_m \text{ for some } \beta \in L_m\}. \tag{3.3}$$

Note that the excitations are indexed with the same set L_m as the states themselves, but in general $\Xi(G_m) \neq \overline{L}_m$. Nonetheless, for the full excitation graph G_m^{full}, we have in fact $\Xi(G_m^{\text{full}}) = \overline{L}_m$.

The reason why explicitly stated that we are considering the "full" excitation graphs is that, in practice, one is forced to ignore the "degree of freedom" (called

"cluster amplitudes", see subsection 3.5) corresponding to some edges.[8] This is done by considering certain subsets of the full edge set E_m^{full}.

DEFINITION 3.8. *An excitation subgraph $G_m = (L_m, E_m)$ is said to be a* consistent subgraph (of G_m^{full}) *if $E_m \subset E_m^{\text{full}}$, and whenever $(\beta, \alpha \vee_m \beta) \in E_m$ for some $\beta \in L_m$ and $\alpha \in \Xi(G_m)$, then $(\beta', \alpha \vee_m \beta') \in E_m$ for all $\beta' \in L_m$.*

The consistency criterion can be rephrased as follows: for a fixed $\alpha \in \Xi(G_m)$, *either* E_m contains the whole "orbit" $\{(\beta, \alpha \vee_m \beta) \in E_m : \beta \in L_m\}$ *or* it does not contain it at all. Note that the set $\Xi(G_m)$ can equally well be used to define a consistent subgraph.

DEFINITION 3.9. *For a given $r = 1, \dots, N$, define $G_m(r) = (L_m, E_m(r))$, where*

$$E_m(r) = \{(\beta, \alpha \vee_m \beta) \in E_m^{\text{full}} : \beta \in L_m,\ \alpha \in \overline{L}_m \text{ such that } \mathrm{rk}_m(\alpha) = r\}.$$

The subgraph $G_m(r_1, \dots, r_\rho) = (L_m, E_m(r_1, \dots, r_\rho))$ is called a rank-truncated excitation subgraph *if*

$$E_m(r_1, \dots, r_\rho) = E_m(r_1) \cup \dots \cup E_m(r_\rho) \quad \text{for} \quad r_1, \dots, r_\rho \in \{1, \dots, N\}.$$

We refer to $G_m(1)$, $G_m(1,2)$, $G_m(1,2,3)$, etc. more colloquially as $G_m(\mathrm{S})$, $G_m(\mathrm{SD})$, $G_m(\mathrm{SDT})$, etc.

Rank-truncation does not introduce isolated vertices in $G_m(r_1, \dots, r_\rho)$ as long as one of the r_j's is 1. However, in the doubles (D) case, $G(\mathrm{D})$ *does* in fact produce isolated vertices so that vertices of odd rank cannot be reached. Also, note that these truncated subgraphs like $G_m(\mathrm{S})$ and $G_m(\mathrm{SD})$ are *not* transitive in general.

We shall summarize these observations in the next theorem. Recall that a digraph is said to be *weakly connected* if every pair of vertices has an undirected path between them.

THEOREM 3.10. *Let $G_m = G_m(r_1, \dots, r_\rho)$ be a rank-truncated excitation subgraph. Then the following is true.*

(i) *G_m is a consistent subgraph.*

(ii) *G_m is weakly connected if one of the r_j's is 1.*

Proof. Obvious from the definition. ∎

Next, we briefly consider two rather "exotic" CC-like methods to demonstrate the generality of the excitation graph concept.

Example 3.11. The excitation graph corresponding to the Tailored CC method (see e.g. [16]) can be described as follows. In that SR method ($M = 1$), the orbital set Λ is partitioned according to $\Lambda_{\text{CAS}} = \{1, \dots, N, N+1, \dots, k\}$ and $\Lambda_{\text{ext}} = \Lambda \setminus \Lambda_{\text{CAS}}$ for some $k = N, \dots, |\Lambda|$. This induces a splitting $L = L(\text{CAS}) \dot\cup L(\text{ext})$, where

$$L(\text{CAS}) = \{\alpha \in L : \alpha \subset \Lambda_{\text{CAS}}\}, \quad L(\text{ext}) = L \setminus L(\text{CAS}).$$

Furthermore, the edge set E^{full} may also be split accordingly

$$E(\text{CAS}) = \{(\beta, \alpha \vee \beta) \in E^{\text{full}} : \alpha \subset \Lambda_{\text{CAS}}\}, \quad \text{and} \quad E(\text{ext}) = E^{\text{full}} \setminus E(\text{CAS}).$$

In other words, $E(\text{CAS})$ contains excitations which change CAS occupied orbitals to CAS virtual ones, and as such, no edge in $E(\text{CAS})$ leaves $L(\text{CAS})$ that starts from $L(\text{CAS})$. It is easy to see that both $G(\text{CAS}) = (L, E(\text{CAS}))$ and $G(\text{ext}) = (L, E(\text{ext}))$ are transitive and consistent subgraphs.

[8] Note that the vertex set is still the "full" vertex set L_m—some vertices might become isolated.

Example 3.12. A generalization of $E(\mathrm{CAS})$ is the "CAS-type subalgebra" (denoted as "$\mathfrak{g}^{(N)}(R,S)$" in [28]), which is constructed from two given subsets $\Lambda_R \subset \{1,\dots,N\}$ and $\Lambda_S \subset \{N+1,\dots\}$. Define $\Lambda_{\mathrm{int}} = \Lambda_R \dot\cup \Lambda_S$ and $\Lambda_{\mathrm{ext}} = \Lambda \setminus \Lambda_{\mathrm{int}}$. This induces a splitting $L = L(\mathrm{int}) \dot\cup L(\mathrm{ext})$, where

$$L(\mathrm{int}) = \{\alpha \in L : \alpha \subset \Lambda_{\mathrm{int}}\}, \quad \text{and} \quad L(\mathrm{ext}) = L \setminus L(\mathrm{int}).$$

The edge set E^{full} decomposes as

$$E(\mathrm{int}) = \{(\beta, \alpha \vee \beta) \in E^{\mathrm{full}} : \alpha \subset \Lambda_{\mathrm{int}}\}, \quad \text{and} \quad E(\mathrm{ext}) = E^{\mathrm{full}} \setminus E(\mathrm{int}).$$

In other words, $E(\mathrm{int})$ contains excitations that replace some orbitals in Λ_R with ones in Λ_S. Then $G(\mathrm{int}) = (L, E(\mathrm{int}))$ and $G(\mathrm{ext}) = (L, E(\mathrm{ext}))$ are transitive and consistent subgraphs. Clearly, Example 3.11 can be recovered with the choice $\Lambda_R = \{1,\dots,N\}$, $\Lambda_S = \{N+1,\dots,k\}$.

Finally, we define excitation graph in the multireference case, which is a natural extension of the above concepts.

Remark 3.13. An important warning is in order. In general, $\alpha \vee_m \beta$ may or may not be equal to $\alpha \vee_\ell \beta$ for $m \neq \ell$. In fact, take $\Lambda = \{1,2,\dots,7\}$ and $0_1 = \{1,2,3\}$, $0_2 = \{1,2,4\}$. Then, with $\alpha = \{1,3,5\}$ and $\beta = \{2,6,7\}$, we have $\alpha \vee_1 \beta = \alpha \vee_2 \beta = \{5,6,7\}$. On the other hand, with $\alpha = \{2,3,4\}$ and $\beta = \{1,2,5\}$, we have $\alpha \vee_1 \beta = \{2,4,5\}$, but $\alpha \vee_2 \beta = \{2,3,5\}$. Note that in the first case, we actually have $(\alpha, \alpha \vee_1 \beta) \in E_1^{\mathrm{full}}$ and $(\alpha, \alpha \vee_2 \beta) \in E_2^{\mathrm{full}}$, i.e. a double edge.

Definition 3.14. *The* full MR excitation multigraph w.r.t. Ω, $G^{\mathrm{full}} = (L, E^{\mathrm{full}})$ *is defined as the union of the individual full SR excitation graphs* $G_m^{\mathrm{full}} = (L_m, E_m^{\mathrm{full}})$ *for all* $m = 1,\dots,M$, *i.e.*

$$L = \bigcup_{m=1}^{M} L_m, \quad E^{\mathrm{full}} = \biguplus_{m=1}^{M} E_m^{\mathrm{full}},$$

where $\uplus$ *denotes multiset union.*

Note that as opposed to the SR graph G_m^{full}, the MR graph G^{full} might have parallel edges (called "redundant" excitations), this justifies that G^{full} was introduced as a multigraph. Notice that other references cannot be "reached" from a given one (see Figure 1). An algorithm for choosing the set of reference states $\Omega = \{0_m\}_{m=1}^M$ in an optimal way, adhering to some given criteria is described in Appendix D.

3.3. Excitation operators. Recall that $\Omega = \{0_m\}_{m=1}^M$ denotes the set of references, and that L_m does *not* contain the other reference states $\Omega \setminus \{0_m\}$. The construction described below is to be repeated for every $m = 1,\dots,M$ separately.

First, we fix an ordering of the indices in $\alpha \in L$. Then, for every element $\alpha = \{\alpha_1,\dots,\alpha_N\} \in L$ we assign the lexicographically ordered N-tuple

$$\alpha^< = (\alpha_1^<,\dots,\alpha_N^<) \in \Lambda^N, \quad \alpha_1^< < \ldots < \alpha_N^<, \quad \text{where} \quad \alpha_j^< \in \alpha.$$

Without loss of generality, we can assume that the orbital indices contained in 0_m are strictly less than the virtual indicies $\Lambda \setminus 0_m$.

As in subsection 2.1, fix an orthonormal set $\mathcal{B} = \{\varphi_p\}_{p\in\Lambda} \subset H^1(\mathbb{R}^3)$ and the corresponding determinantal wavefunctions

$$\mathfrak{B} = \{\Phi_\alpha \in \mathfrak{H}^1 : \alpha \in L,\ \Phi_\alpha(\mathbf{X}) = N!^{-1/2} \det(\varphi_{\alpha_i^<}(\mathbf{x}_j))_{1\le i,j\le N}\}. \tag{3.4}$$

Recall the notation $\mathfrak{H}^1_K \subset \mathfrak{H}^1$ for the subspace spanned by $\mathfrak{B}$; which is allowed to be finite-, or infinite-dimensional depending on $K = |\Lambda|$.

DEFINITION 3.15. *Let $G_m = (L_m, E_m)$ be a subgraph of G_m^{full}. The family of linear operators $X_\alpha^{(m)} := X_\alpha(G_m) : \mathfrak{H}^1_K \to \mathfrak{H}^1_K$ given by*

$$X_\alpha(G_m)\Phi_\beta = \begin{cases} \sigma(\alpha,\beta)\Phi_{\alpha \vee_m \beta} & (\beta, \alpha \vee_m \beta) \in E_m \\ 0 & (\beta, \alpha \vee_m \beta) \notin E_m \end{cases}$$

for each $\alpha \in \Xi(G_m)$ and $\beta \in L$, and extended boundedly and linearly to the whole space $\mathfrak{H}^1_K$ is called the family of excitation operators on G_m. *Here, $\sigma(\alpha,\beta)$ is the sign of the permutation $\pi(\alpha,\beta)$ that puts the tuple $((\overline{\beta}^m)^<, (\overline{\alpha}^m)^<)$ in lexicographical order.*

Assuming $\Xi(G_m) \neq \emptyset$, by the definition of $\Xi(G_m)$ (see (3.3)) for every $\alpha \in \Xi(G_m)$ there is some $\beta \in L_m$ such that $(\beta, \alpha \vee_m \beta) \in E_m$ and therefore $X_\alpha(G_m) \not\equiv 0$. Recalling $\mathrm{rk}_m(\alpha \vee_m \beta) = \mathrm{rk}_m(\alpha) + \mathrm{rk}_m(\beta)$ (see Proposition 3.3), we can roughly say that an excitation operator $X_\alpha(G_m)$ increases the rank by $\mathrm{rk}_m(\alpha)$.

Since $G_m = (L_m, E_m)$ is a subgraph of $G_m^{\text{full}} = (L_m, E_m^{\text{full}})$, some excitations might be missing, i.e. $\Xi(G_m) \subset \Xi(G_m^{\text{full}})$. The next result shows that the excitation operators constructed for a *consistent subgraph* G_m (see Definition 3.8) are precisely the same as the ones constructed for G_m^{full}, with some of them possibly being absent. This explains the use of the word "consistent".

THEOREM 3.16. *Let G_m be a* consistent *subgraph of G_m^{full}. Then,*

$$X_\alpha(G_m) \equiv X_\alpha(G_m^{\text{full}}) \quad \text{for all } \alpha \in \Xi(G_m).$$

Proof. Fix $\alpha \in \Xi(G_m)$, then by (3.3) and Definition 3.8, $(\beta, \alpha \vee_m \beta) \in E_m$ for all $\beta \in L_m$. Consequently, $X_\alpha(G_m)\Phi_\beta = X_\alpha(G_m^{\text{full}})\Phi_\beta$ for all $\beta \in L$. ∎

Based on this result, if G_m is consistent, it is safe to drop the "G_m" from the notation $X_\alpha(G_m)$ and simply denote the excitation operators by $X_\alpha^{(m)}$, or by X_α in the SR case. However, it is important to note that for a given α, $X_\alpha^{(m)} \neq X_\alpha^{(\ell)}$ in general for differing reference states $m \neq \ell$, see Remark 3.13.

The excitation operators enjoy nice algebraic properties which we summarize in the next theorem (cf. [49, Lemma 2.5]).

THEOREM 3.17. *Let $G_m = (L_m, E_m)$ be a consistent subgraph of G_m^{full} and let $\{X_\alpha^{(m)}\}_{\alpha \in \Xi(G_m)}$ denote the set of excitation operators on G_m. Then the following properties hold true.*

(i) *(commutativity) For all $\alpha, \beta \in \Xi(G_m)$, there holds $X_\alpha^{(m)} X_\beta^{(m)} = X_\beta^{(m)} X_\alpha^{(m)}$. In detail, for any $\gamma \in L$,*

$$X_\alpha^{(m)} X_\beta^{(m)} \Phi_\gamma = \begin{cases} \sigma(\alpha, \beta \vee_m \gamma)\sigma(\beta,\gamma)\Phi_{\alpha \vee_m \beta \vee_m \gamma} & (\beta \vee_m \gamma, \alpha \vee_m \beta \vee_m \gamma), \\ & (\gamma, \beta \vee_m \gamma) \in E_m \\ 0 & \text{otherwise} \end{cases}$$

(ii) *If G_m is transitive, then $\{0\} \cup \{\pm X_\alpha^{(m)}\}_{\alpha \in \Xi(G_m)}$ is multiplicatively closed. In particular, $\{0\} \cup \{\pm X_\alpha^{(m)}\}_{\alpha \in \Xi(G_m^{\text{full}})}$ is multiplicatively closed.*

(iii) *(nilpotency) For all $\alpha \in \Xi(G_m)$, $(X_\alpha^{(m)})^2 = 0$.*

Proof. To see (i), first observe that if $(\beta \vee_m \gamma, \alpha \vee_m (\beta \vee_m \gamma))$, $(\gamma, \beta \vee_m \gamma) \in E_m$, then $(\alpha \vee_m \gamma, \beta \vee_m (\alpha \vee_m \gamma))$, $(\gamma, \alpha \vee_m \gamma) \in E_m$ due to the consistent subgraph property of G_m. It is obvious that $\Phi_{\alpha \vee_m \beta \vee_m \gamma} = \Phi_{\beta \vee_m \alpha \vee_m \gamma}$ from the commutativity of $\vee_m$. It remains to prove $\sigma(\alpha, \beta \vee_m \gamma)\sigma(\beta, \gamma) = \sigma(\beta, \alpha \vee_m \gamma)\sigma(\alpha, \gamma)$. Let π_1, π_2 and τ_1, τ_2 be the permutations that put $((\overline{\beta} \cup \overline{\gamma})^<, \overline{\alpha}^<)$, $(\overline{\gamma}^<, \overline{\beta}^<)$ and $((\overline{\alpha} \cup \overline{\gamma})^<, \overline{\beta}^<)$, $(\overline{\gamma}^<, \overline{\alpha}^<)$, respectively, in lexicographic order. Then $\pi_1 \circ \pi_2 = \tau_1 \circ \tau_2 = \sigma$, where σ is the permutation that puts $(\overline{\alpha}, \overline{\beta}, \overline{\gamma})$ in lexicographic order. The claim follows from the multiplicativity of the sgn function on permutations.

For (ii), suppose that G_m is transitive and that $\alpha, \beta \in \Xi(G_m)$. Using (i), either $X_\alpha^{(m)} X_\beta^{(m)} = \pm X_{\alpha \vee_m \beta}^{(m)}$ or $X_\alpha^{(m)} X_\beta^{(m)} = 0$. In the former case, $(\beta \vee_m \gamma, \alpha \vee_m \beta \vee_m \gamma)$, $(\gamma, \beta \vee_m \gamma) \in E_m$ implies that $(\gamma, \alpha \vee_m \beta \vee_m \gamma) \in E_m$ by the transitivity of G_m, so $\alpha \vee_m \beta \in \Xi(G_m)$.

For (iii), it is enough to notice that $(\alpha \vee_m \gamma, \alpha \vee_m \gamma) = (\alpha \vee_m \alpha \vee_m \gamma, \alpha \vee_m \gamma) \notin E_m^{\text{full}}$, because G_m^{full} does not contain loop edges by definition. ∎

It is important to note that in general *excitation operators corresponding to different reference states do not commute*: $X_\alpha^{(m)} X_\beta^{(\ell)} \neq X_\beta^{(m)} X_\alpha^{(\ell)}$ for $m \neq \ell$, again, because of Remark 3.13.

Remark 3.18. The excitation operators are traditionally expressed using the language of second quantization. Let $a_p^\dagger$ and a_p denote the fermionic creation and annihilation operators. Then $\Phi_\beta = a_{\beta_1}^\dagger \cdots a_{\beta_N}^\dagger |\text{vac}\rangle$, where $\beta = \{\beta_1 < \ldots < \beta_N\}$, and

$$X_\alpha = a_{p_1}^\dagger a_{q_1} \cdots a_{p_n}^\dagger a_{q_n}.$$

Here $\{q_1, \ldots, q_n\} = 0 \smallsetminus \underline{\alpha}$ and $\{p_1, \ldots, p_n\} = \overline{\alpha}$ with $q_1 < \ldots < q_n$ and $p_1 < \ldots < q_n$. In other words, X_α changes the orbitals $0 \smallsetminus \underline{\alpha}$ to $\overline{\alpha}$, as expected. Although the excitation operators commute with each other, they *do not* commute in general with the Hamilton operator.

We now define a family of operators which "reverse" the action of $X_\alpha^{(m)}$.

Definition 3.19. *Let $G_m = (L_m, E_m)$ be a subgraph of G_m^{full}. For all $\alpha \in \Xi(G_m)$, the linear operators $(X_\alpha^{(m)})^\dagger : \mathfrak{H}_K^1 \to \mathfrak{H}_K^1$ defined via*

$$(X_\alpha^{(m)})^\dagger \Phi_\beta = \begin{cases} \sigma(\alpha, \alpha^\perp \wedge_m \beta) \Phi_{\alpha^\perp \wedge_m \beta} & (\alpha^\perp \wedge_m \beta, \beta) \in E_m \\ 0 & (\alpha^\perp \wedge_m \beta, \beta) \notin E_m \end{cases}$$

for any $\beta \in L$, and extended boundedly and linearly to the whole space $\mathfrak{H}_K^1$, are called de-excitation operators on G_m.

It is easy to see using Lemma 3.5 and Proposition 3.3 that

$$\text{rk}_m(\alpha^\perp \wedge_m \beta) = \text{rk}_m(\beta) - \text{rk}_m(\alpha), \tag{3.5}$$

whenever $(\alpha^\perp \wedge_m \beta, \beta) \in E_m$. Therefore, we may roughly say that the de-excitation operator $(X_\alpha^{(m)})^\dagger$ decreases the rank by $\text{rk}_m(\alpha)$. Of course, the notation $\dagger$ is not coincidental, and $(X_\alpha^{(m)})^\dagger$ is in fact the $\mathfrak{L}^2$-adjoint of $X_\alpha^{(m)}$.

Theorem 3.20. *Suppose that $\{X_\alpha^{(m)}\}$ and $\{(X_\alpha^{(m)})^\dagger\}$ are the set of excitation and de-excitation operators corresponding to the excitation graph G_m. Then*

$$\langle (X_\alpha^{(m)})^\dagger \Phi, \Psi \rangle = \langle \Phi, X_\alpha^{(m)} \Psi \rangle \quad \textit{for all} \quad \Phi, \Psi \in \mathfrak{H}_K^1 \textit{ and } \alpha \in \Xi(G_m).$$

Proof. It is enough to prove the relation for $\Phi = \Phi_\gamma$ and $\Psi = \Phi_\beta$, as the general statement follows by linearity. Suppose that $(\alpha^\perp \wedge_m \gamma, \gamma) \in E_m$, then

$$\begin{aligned}\langle (X_\alpha^{(m)})^\dagger \Phi_\gamma, \Phi_\beta\rangle &= \sigma(\alpha, \alpha^\perp \wedge_m \gamma)\langle \Phi_{\alpha^\perp \wedge_m \gamma}, \Phi_\beta\rangle \\ &= \sigma(\alpha,\beta)\langle \Phi_\gamma, \Phi_{\alpha\vee_m\beta}\rangle = \langle \Phi_\gamma, X_\alpha^{(m)}\Phi_\beta\rangle,\end{aligned}$$

where we used that $\alpha^\perp \wedge_m \gamma = \beta \in L_m$ if and only if $\alpha \vee_m \beta = \gamma \in L_m$ (Lemma 3.5). ∎

THEOREM 3.21. *Let $G_m = (L_m, E_m)$ be a consistent subgraph of G_m^{full} and let $\{X_\alpha^{(m)}\}_{\alpha\in\Xi(G_m)}$ and $\{(X_\alpha^{(m)})^\dagger\}_{\alpha\in\Xi(G_m)}$ denote the set of excitation-, and deexcitation operators on G_m. Then the following properties hold true.*

(i) *(commutativity) For all $\alpha, \beta \in \Xi(G_m)$, there holds*

$$(X_\alpha^{(m)})^\dagger (X_\beta^{(m)})^\dagger = (X_\beta^{(m)})^\dagger (X_\alpha^{(m)})^\dagger.$$

(ii) *For any $\alpha, \beta \in \Xi(G_m)$ and $\gamma \in L$, the following formula holds true:*

$$(X_\alpha^{(m)})^\dagger X_\beta^{(m)}\Phi_\gamma = \sigma(\alpha, \alpha^\perp \wedge_m (\beta \vee_m \gamma))\sigma(\beta,\gamma)\Phi_{\alpha^\perp\wedge(\beta\vee\gamma)}$$

if $(\gamma, \beta \vee_m \gamma) \in E_m$ and $(\alpha^\perp \wedge_m (\beta \vee_m \gamma), \beta \vee_m \gamma) \in E_m$ both hold true. Otherwise, $(X_\alpha^{(m)})^\dagger X_\beta^{(m)}\Phi_\gamma = 0$. In particular, $(X_\alpha^{(m)})^\dagger \Phi_\alpha = \Phi_{0_m}$.

(iii) *$(X_\alpha^{(m)})^\dagger \Phi_{0_\ell} = 0$ for any $m \neq \ell$ and $\alpha \in \Xi(G_m)$.*

(iv) *(nilpotency) $((X_\alpha^{(m)})^\dagger)^2 = 0$ for any $\alpha \in \Xi(G_m)$.*

Proof. Part (i) follows from Theorem 3.20 combined with Theorem 3.17 (i). Part (ii) follows directly from the definitions. Part (iii) comes from the fact that there are no edges between different 0_m's. Part (iv) follows since $(\alpha^\perp \wedge_m \alpha, \alpha) = (0, \alpha) \in E_m$ for every $\alpha \in \Xi(G_m)$. ∎

It is highly important to stress that in general excitation-, and de-excitation operators do not commute with each other:

$$X_\alpha^{(m)}(X_\alpha^{(m)})^\dagger \neq (X_\alpha^{(m)})^\dagger X_\alpha^{(m)},$$

in other words, the $X_\alpha^{(m)}$'s are *nonnormal* operators. Also, $[(X_\alpha^{(m)})^\dagger, X_\beta^{(m)}] \neq 0$ in general. This fact is the source of many technical obstacles in the analysis of the CC method, primarily because it implies that the similarity-transformed Hamilton operator (2.12) is nonnormal.

3.4. Cluster operators. From now on, we omit the reference index m from the notations, with the understanding that the considerations hold true for every reference independently. Suppose that we constructed the set of excitation operators $\{X_\alpha\}_{\alpha\in\Xi(G)}$ for a given consistent subgraph $G = (L, E)$. The completion of their linear hull

$$\mathfrak{v}(G) = \overline{\mathrm{Span}\{X_\alpha\}_{\alpha\in\Xi(G)}}^{\|\cdot\|_{\mathcal{L}(\mathfrak{H}^1,\mathfrak{H}^1)}}$$

is called the *space of cluster operators on G* endowed with operator norm $\|\cdot\|_{\mathcal{L}(\mathfrak{H}^1,\mathfrak{H}^1)}$. As mentioned earlier, if G is not the full excitation graph G^{full}, then certain excitation operators will be absent and therefore, they will be missing from $\mathfrak{v}(G)$ as well.

PROPOSITION 3.22. *For any $T \in \mathfrak{v}(G)$, we have $T^{N+1} = 0$.*

Proof. It is enough to prove that an arbitrary product of $N+1$ excitation operators is zero. In fact, by definition every excitation operator either increases the rank of a determinantal wavefunction by at least 1 or maps it to zero. But the rank cannot increase above N, so the product must be zero. ∎

It is well-known that the vector space $\mathfrak{v}(G^{\text{full}})$ constructed on the full excitation graph G^{full} forms a *commutative algebra* (see e.g. [51, Lemma 4.2]) with the usual multiplication (a subalgebra of the algebra of bounded linear operators $\mathcal{L}(\mathfrak{H}_K^1, \mathfrak{H}_K^1)$). According to Proposition 3.22, it is also *nilpotent*. More generally, we have

THEOREM 3.23. *$\mathfrak{v}(G)$ is a nilpotent, commutative algebra for any transitive excitation graph G.*

Proof. Follows from Theorem 3.17 (ii). ∎

If, however, G is not transitive, then $\mathfrak{v}(G)$ is *not* an algebra in general—for instance in $\mathfrak{v}(G(\text{SD}))$ there are no excitation operators of rank 3 and above, but the rank of the products of excitation operators can be arbitrary ($\leq N$).

Example 3.24. We observed in Example 3.11 that the CAS-subgraph $G(\text{CAS})$ corresponding to the TCC method is transitive and consistent, hence $\mathfrak{v}(G(\text{CAS}))$ forms a subalgebra of $\mathfrak{v}(G^{\text{full}})$ (cf. [28]). Similarly, for $G(\text{int})$ in Example 3.12, $\mathfrak{v}(G(\text{int}))$ also forms a subalgebra. However, in a truncated setting, where only certain low-rank edges of $E(\text{CAS})$ (or $E(\text{int})$) are retained, transitivity, hence the subalgebra property, is lost.

Let now the excitation graph $G = (L, E)$ be arbitrary. A cluster operator $C \in \mathfrak{v}(G)$ may be decomposed according to the excitation ranks of its constituent excitations as

$$C = \sum_{r=1}^{N} C_r, \quad \text{where} \quad C_r = \sum_{\text{rk}(\alpha)=r} c_\alpha X_\alpha. \tag{3.6}$$

We say that C *is of rank* r if it contains excitation operators of rank at most r. Note that the graded structure of G is compatible with this decomposition in the sense that if C and D are of ranks r and s, respectively, then CD is of rank $r + s$.

Remark 3.25. In the SR case, the cluster operators can be used to express any wavefunction in $\mathfrak{H}_K^1$ if the full excitation graph G^{full} is used for their construction. In fact, in this case, $X_\alpha \Phi_0 = \Phi_\alpha$ for *every* $\alpha \in \overline{L}$, so we may express any function in $\mathfrak{H}_K^1$ through a linear combination of the excitation operators *and* the identity I. More precisely, if

$$\Psi = \sum_{\alpha \in L} c_\alpha \Phi_\alpha = c_0 \Phi_0 + \sum_{\alpha \in \overline{L}} c_\alpha \Phi_\alpha, \quad \text{then} \quad \Psi = \left[c_0 I + \sum_{\alpha \in \overline{L}} c_\alpha X_\alpha \right] \Phi_0,$$

for some scalars $\{c_\alpha\}_{\alpha \in L}$. Recall that in subsection 2.4 we assumed the intermediate normalization condition $\langle \Psi, \Phi_0 \rangle = 1$, which implies $c_0 = 1$. There is a one-to-one correspondence between functions $\Psi \in \mathfrak{H}_K^{1,\perp}$ and the cluster operators C_Ψ defined as

$$C_\Psi = \sum_{\alpha \in \overline{L}} c_\alpha X_\alpha, \quad \text{where} \quad c_\alpha = \langle \Psi, \Phi_\alpha \rangle. \tag{3.7}$$

It is not clear, however, that $C_\Psi \in \mathcal{L}(\mathfrak{H}_K^1, \mathfrak{H}_K^1)$. See Theorem 3.26 below for the precise statement of this nontrivial fact. Also, if the excitation graph does not contain every edge of the form $(0, \alpha)$—which is typically the case if some truncation is used—then it is *not* possible to assign a cluster operator (3.7) to every $\Psi \in \mathfrak{H}_K^{1,\perp}$.

The following important result makes the aforementioned correspondence between functions and cluster operators precise.

THEOREM 3.26. *[49, Theorem 4.1 and Lemma 5.1] Fix $\Psi \in \mathfrak{H}^{1,\perp}$. Then, the following hold true.*

1. *The cluster operator C_Ψ (3.7) satisfies $C_\Psi \in \mathcal{L}(\mathfrak{H}^1, \mathfrak{H}^1)$. Furthermore, there is a constant $b > 0$ independent of Ψ such that*

$$\|\Psi\|_{\mathfrak{H}^1} \leq \|C_\Psi\|_{\mathcal{L}(\mathfrak{H}^1,\mathfrak{H}^1)} \leq b\|\Psi\|_{\mathfrak{H}^1}.$$

2. *$C_\Psi^\dagger \in \mathcal{L}(\mathfrak{H}^1, \mathfrak{H}^1)$, and there is a constant $b' > 0$ independent of Ψ such that*

$$\|C_\Psi^\dagger\|_{\mathcal{L}(\mathfrak{H}^1,\mathfrak{H}^1)} \leq b'\|\Psi\|_{\mathfrak{H}^1},$$

   and there cannot be a uniform lower bound in terms of $\|\Psi\|_{\mathfrak{H}^1}$.
3. *C_Ψ can be extended to $\mathcal{L}(\mathfrak{H}^{-1}, \mathfrak{H}^{-1})$.*

Next, we consider the so-called *exponential Ansatz*, which is the representation

$$I + C = e^T, \quad \text{where} \quad T = \sum_{\alpha \in \Xi(G^{\text{full}})} t_\alpha X_\alpha \in \mathfrak{v}(G^{\text{full}}),$$

and $C \in \mathfrak{v}(G^{\text{full}})$. Here, e^T is simply a finite sum due to the nilpotency of T, i.e.

$$e^T = I + T + \tfrac{1}{2!}T^2 + \ldots + \tfrac{1}{N!}T^N.$$

The inverse of the exponential should be the logarithm, as one would expect, and in fact the following elementary theorem holds.

THEOREM 3.27. *For any cluster operator $C \in \mathfrak{v}(G^{\text{full}})$ there exists a unique cluster operator $T \in \mathfrak{v}(G^{\text{full}})$, such that $e^T = I + C$. Furthermore,*

$$T = \log(I + C) = C - \tfrac{1}{2}C^2 + \tfrac{1}{3}C^3 - \ldots + \tfrac{(-1)^{N-1}}{N}C^N.$$

Proof. A tedious but straightforward calculation, or see [51, Theorem 4.3] for a quick proof using holomorphic functional calculus. ∎

It is important to note that if some proper subgraph $G = (L, E)$ is considered instead of G^{full}, the previous result does *not* hold. For instance, if $G(\text{SD})$ is considered, then it might not be possible to represent $I + C$ as e^T, where $C \in \mathfrak{v}(G^{\text{full}})$ and $T \in \mathfrak{v}(G)$. This in particular implies that wavefunctions of the form $e^T\Phi_0$ where $T \in \mathfrak{v}(G)$ is *not* the totality of intermediately normalized wavefunctions.

In the multireference (MR) case, the analogue of the exponential Ansatz is called the *Jeziorski–Monkhorst (JM) Ansatz*, see subsection 4.2 below. In the JM-MRCC method, M wavefunctions, say $\Psi_1, \ldots, \Psi_M$ are "targeted", and the expansion

$$\Psi_j = \sum_{m=1}^{M} a_j^{(m)} e^{T^{(m)}} \Phi_{0_m}, \quad \text{where} \quad a_j^{(m)} \in \mathbb{R}, \tag{3.8}$$

is utilized. In the untruncated case, suppose that $\Psi_j = (I + C^{(j)})\Phi_{0_j} = e^{T^{(j)}}\Phi_{0_j}$, as above, for all $j = 1, \ldots, M$. Then the JM expansion coefficients $a_j^{(m)}$ of Ψ_j are simply δ_{jm}.

3.5. Cluster amplitude spaces. The linear combination coefficients of the excitation operators making up a cluster operator are called *cluster amplitudes*. Let $\ell^2(G)$ denote Hilbert space of square summable real-, or complex-valued sequences indexed by the edge labels of the excitation graph G, i.e.

$$\ell^2(G) = \{t = (t_\alpha)_{\alpha\in\Xi(G)} : \|t\|_{\ell^2} < \infty\}.$$

The (real or complex) Hilbert space

$$\mathbb{V}(G) = \{t \in \ell^2(G) : \|T\Phi_0\|_{\mathfrak{H}^1} < \infty\},$$

endowed with the $\mathfrak{H}^1$-inner product $\langle t, s\rangle_{\mathbb{V}} = \langle T\Phi_0, S\Phi_0\rangle_{\mathfrak{H}^1}$ is called the *(cluster) amplitude space corresponding to* G. Nevertheless, from now on we use the convention that the unmarked $\langle t, s\rangle = \langle T\Phi_0, S\Phi_0\rangle_{\mathfrak{L}^2}$ and $\|\cdot\|$ refers to the ℓ^2-inner product and ℓ^2-norm. Clearly, $\|t\| \le \|t\|_{\mathbb{V}}$.

Remark 3.28. Similarly to $\mathfrak{H}^1 \hookrightarrow \mathfrak{L}^2 \hookrightarrow \mathfrak{H}^{-1}$, the spaces $\mathbb{V}(G) \hookrightarrow \ell^2(G) \hookrightarrow \mathbb{V}(G)^*$ also form a Gelfand triple. Therefore, the ℓ^2-inner product $\langle\cdot,\cdot\rangle$ may be used instead of the dual pairing $\langle\cdot,\cdot\rangle_{\mathbb{V}(G)^*\times\mathbb{V}(G)}$ with the understanding that the appropriate relations must be extended to $\mathbb{V}(G)^* \times \mathbb{V}(G)$ by density.

It is clear that the space of cluster operators $\mathfrak{v}(G)$ is canonically isomorphic to $\mathbb{V}(G)$ via

$$\mathfrak{v}(G) \ni \sum_{\alpha\in\Xi(G)} c_\alpha X_\alpha = C \mapsto c = (c_\alpha)_{\alpha\in\Xi(G)} \in \mathbb{V}(G).$$

As customary in CC theory, we will never explicitly denote this isomorphism, and instead use capital letters S, T, U, V, W, etc. to denote the cluster operators and small letters s, t, u, v, w, etc. to denote their corresponding cluster amplitudes.

Furthermore, to every amplitude space $\mathbb{V}(G)$ there corresponds a *functional amplitude space* $\mathfrak{V}(G) \subset \mathfrak{H}^{1,\perp}$ through the $(\ell^2, \mathfrak{L}^2)$-isometric isomorphism $\mathbb{V}(G) \to \mathfrak{V}(G)$ given by

$$\mathbb{V}(G) \ni c \mapsto C\Phi_0 = \sum_{\alpha\in\Xi(G)} c_\alpha \Phi_\alpha \in \mathfrak{V}(G).$$

Clearly, an appropriate subset of the determinantal basis $\mathfrak{B}$ (see (3.4)) forms a basis of the functional amplitude space $\mathfrak{V}(G)$.

Given a closed subspace $\mathfrak{U} \subset \mathfrak{V}(G)$, we will sometimes use the orthogonal projector $\Pi_{\mathfrak{U}} : \mathfrak{L}^2 \to \mathfrak{U} \subset \mathfrak{L}^2$ onto $\mathfrak{U}$, defined as

$$\langle \Pi_{\mathfrak{U}}\Psi, \Phi\rangle = \langle \Psi, \Phi\rangle, \quad \text{for all} \quad \Psi \in \mathfrak{L}^2,\ \Phi \in \mathfrak{U}.$$

Hence, the inclusion map $I_{\mathfrak{U}} : \mathfrak{U} \to \mathfrak{L}^2$, given by $I_{\mathfrak{U}}\Phi = \Phi$ for all $\Phi \in \mathfrak{U}$ satisfies $I_{\mathfrak{U}}^\dagger = \Pi_{\mathfrak{U}}$.

Remark 3.29. In the case $K < \infty$, following [51], we can introduce the norm

$$|||t|||^2 := \sum_{\alpha\in\Xi(G)} \varepsilon_\alpha |t_\alpha|^2 \quad \text{for all} \quad t \in \mathbb{V},$$

where ε_α denotes the eigenvalues of the Fock operator $\mathcal{F}$, see (2.5). It was shown in [51] that $|||\cdot|||$ and $\|\cdot\|_{\mathbb{V}}$ are equivalent norms, furthermore, under certain assumptions, the constants in the norm equivalence $|||\cdot||| \sim \|\cdot\|_{\mathbb{V}}$ are independent of K. Finally, we also define the norm $|||\cdot|||$ on $\mathfrak{V}$ via $|||T\Phi_0||| := |||t|||$ for any $t \in \mathbb{V}$.

We continue by recalling an important notion due to [51].

DEFINITION 3.30. *The excitation graph G is said to be* excitation complete*, if $\alpha^\perp \wedge \beta \in \Xi(G)$ for all $\alpha, \beta \in \Xi(G)$ with $(\alpha^\perp \wedge \beta, \beta) \in E$ and $\alpha \neq \beta$.*

It is easy to see using (3.5), that commonly used rank-truncated graphs such as $G(1, 2, \ldots, \rho)$ and $G(\mathrm{D})$ are excitation complete. The following result is used for proving that two different formulations of the CC method are equivalent Lemma 5.3.

PROPOSITION 3.31. *[51, Lemma 5.5] Suppose that G is excitation complete, let $\mathfrak{V} = \mathfrak{V}(G)$ and $\mathfrak{V}_0 = \mathrm{Span}\{\Phi_0\} \oplus \mathfrak{V}$. Fix $t \in \mathbb{V}$.*

(i) *The linear mappings $e^{\pm T^\dagger} I_{\mathfrak{V}_0} : \mathfrak{V}_0 \to \mathfrak{V}_0$ are bijective.*

(ii) *The linear mappings $\Pi_{\mathfrak{V}} e^{\pm T^\dagger} I_{\mathfrak{V}} : \mathfrak{V} \to \mathfrak{V}$ are surjective.*

The result follows easily from the next lemma.

LEMMA 3.32. *[51, Lemma 5.4] Suppose that G is excitation complete. Then, for every $\alpha, \beta \in \Xi(G)$ we have $X_\alpha^\dagger \Phi_\beta \in \mathfrak{V}(G) \cup \{\Phi_0\}$.*

Proof. From Theorem 3.21 (ii), we have

$$X_\alpha^\dagger \Phi_\beta = \sigma(\alpha, \alpha^\perp \wedge \beta) \Phi_{\alpha^\perp \wedge \beta},$$

if $(\alpha^\perp \wedge \beta, \beta) \in E$. If $\alpha \neq \beta$, then right-hand side is in $\mathfrak{V}(G)$, since G is excitation complete. If $\alpha = \beta$, then the right-hand side is simply Φ_0. ∎

Proof of Proposition 3.31. By linearity, Lemma 3.32 implies that the mapping $T^\dagger : \mathfrak{V}_0(G) \to \mathfrak{V}_0(G)$ and so $e^{\pm T^\dagger} : \mathfrak{V}_0(G) \to \mathfrak{V}_0(G)$ as well. But $(e^{T^\dagger})^{-1} = e^{-T^\dagger}$, which proves (i). ∎

4. Derivation of the Coupled-Cluster Equations. In this section, we give derivations of the SRCC-, and a variant of the MRCC equations. The approach presented here is based on [56, pp. 99–184], but it is more general and sharper. We would like to stress that the discussion only applies to the *full* (that is, untruncated) CC methods.

The essence of the following theorem seems to be well-known in the physics and in the quantum chemistry literature, and the method itself is generally attributed to C. Bloch [8], who devised it in the context of perturbation theory.

THEOREM 4.1. *Let $\mathfrak{H}$ and $\mathfrak{L}$ be (real or complex) Hilbert spaces so that they form a Gelfand triple: $\mathfrak{H} \subset \mathfrak{L} \subset \mathfrak{H}^*$. Let $\mathcal{H} : \mathfrak{H} \to \mathfrak{H}^*$ be a bounded operator. Let $\mathfrak{M}, \mathfrak{N} \subset \mathfrak{H}$ be any pair of closed subspaces so that the following* complementarity condition *holds:*

$$\mathfrak{M} \oplus \mathfrak{N}^\perp = \mathfrak{H}. \tag{4.1}$$

Then the following are equivalent.

(i) $\mathfrak{M} \subset \mathfrak{H}$ *is* weakly $\mathcal{H}$-invariant: *for every $\Phi \in \mathfrak{M}$ there exists $\widetilde{\Phi} \in \mathfrak{M}$ such that $\langle \mathcal{H}\Phi, \Phi' \rangle = \langle \widetilde{\Phi}, \Phi' \rangle$ for all $\Phi' \in \mathfrak{H}$.*

(ii) *(weak Bloch equation) There holds*

$$\langle \mathcal{H}\Xi\Phi, (I - \Xi^\dagger)\Phi' \rangle = 0 \quad \text{for all} \quad \Phi \in \mathfrak{N}, \Phi' \in \mathfrak{N}^\perp, \tag{4.2}$$

where $\Xi : \mathfrak{H} \to \mathfrak{H}$ denotes the (oblique) projector onto $\mathfrak{M}$ along $\mathfrak{N}^\perp$, i.e. $\operatorname{ran} \Xi = \mathfrak{M}$ *and* $\ker \Xi = \mathfrak{N}^\perp$.

Furthermore, if

$$\mathfrak{M} = \mathrm{Span}\{\Psi_j \in \mathfrak{H} : j = 1, \ldots, J\}, \quad \text{where} \quad \langle \mathcal{H}\Psi_j, \overline{\Phi} \rangle = \mathcal{E}_j \langle \Psi_j, \overline{\Phi} \rangle \quad (\overline{\Phi} \in \mathfrak{H}) \tag{4.3}$$

for some $\mathcal{E}_j \in \mathbb{C}$*, then with the* effective Hamiltonian $\mathcal{H}^{\text{eff}} : \mathfrak{N} \to \mathfrak{N}$*, given by* $\langle \mathcal{H}^{\text{eff}}\Phi, \Phi'\rangle = \langle \mathcal{H}\Xi\Phi, \Phi'\rangle$ *for all* $\Phi, \Phi' \in \mathfrak{N}$*, we have*

$$\langle \mathcal{H}^{\text{eff}}\Pi\Psi_j, \Phi\rangle = \mathcal{E}_j\langle \Pi\Psi_j, \Phi\rangle \quad \textit{for all} \quad \Phi \in \mathfrak{N}, \tag{4.4}$$

where $\Pi : \mathfrak{H} \to \mathfrak{H}$ *denotes the* $\mathfrak{L}$-orthogonal *projector onto* $\mathfrak{N}$*, i.e.* $\operatorname{ran}\Pi = \mathfrak{N}$ *and* $\ker\Pi = \mathfrak{N}^\perp$.

Proof. For (i)$\Longrightarrow$(ii), note that using $\ker\Xi = \mathfrak{N}^\perp$ and $\operatorname{ran}\Xi = \mathfrak{M}$, it follows from (i) that for every $\Phi \in \mathfrak{N}$ there exists $\widetilde{\Phi} \in \mathfrak{M}$ such that $\langle \mathcal{H}\Xi\Phi, \overline{\Phi}\rangle = \langle \widetilde{\Phi}, \overline{\Phi}\rangle$ for all $\overline{\Phi} \in \mathfrak{H}$. Put $\overline{\Phi} = (I - \Xi^\dagger)\Phi'$ to obtain

$$\langle \mathcal{H}\Xi\Phi, (I - \Xi^\dagger)\Phi'\rangle = \langle \widetilde{\Phi}, (I - \Xi^\dagger)\Phi'\rangle = 0 \quad \text{for all} \quad \Phi \in \mathfrak{N}, \Phi' \in \mathfrak{H},$$

where we used that $\widetilde{\Phi} \in \mathfrak{M}$ and $\operatorname{ran}(I - \Xi^\dagger) = \mathfrak{M}^\perp$. From this, (4.2) follows.

To see (ii)$\Longrightarrow$(i), fix $\Phi \in \mathfrak{M}$ and note that (4.2) implies $F_\Phi(\Phi') = 0$ for all $\Phi' \in \mathfrak{M}^\perp$, where $F_\Phi(\Phi') := \langle \mathcal{H}\Phi, \Phi'\rangle$ for all $\Phi' \in \mathfrak{H}$. Here, $F_\Phi(\cdot)$ is a bounded linear functional on $\mathfrak{H} \subset \mathfrak{L}$. Extend F_Φ to a bounded linear functional $\widehat{F}_\Phi$ on $\mathfrak{L}$ using the Hahn–Banach theorem. The Riesz representation theorem implies that there is a $\widetilde{\Phi} \in \mathfrak{L}$ such that $\widehat{F}_\Phi(\Phi') = \langle \widetilde{\Phi}, \Phi'\rangle$ for all $\Phi' \in \mathfrak{L}$. But $0 = F_\Phi(\Phi') = \widehat{F}_\Phi(\Phi') = \langle \widetilde{\Phi}, \Phi'\rangle$ for all $\Phi' \in \mathfrak{M}^\perp$, so $\widetilde{\Phi} \in \mathfrak{M}^{\perp\perp} = \mathfrak{M}$. Therefore, we constructed a $\widetilde{\Phi} \in \mathfrak{M}$ such that $\langle \mathcal{H}\Phi, \Phi'\rangle = \langle \widetilde{\Phi}, \Phi'\rangle$ for all $\Phi' \in \mathfrak{H}$, which is what we wanted to prove.

To prove the "furthermore" part, first note that $\mathfrak{M}$ is weakly $\mathcal{H}$-invariant. We now claim that $\Xi\Pi = \Xi$. In fact, $\operatorname{ran}(I - \Pi) = \ker\Pi = \ker\Xi$, so $\Xi(I - \Pi) = 0$. Continuing the proof, note that the second relation of (4.3) is equivalent to

$$\langle \mathcal{H}\Xi\Pi\Psi_j, \overline{\Phi}\rangle = \mathcal{E}_j\langle \Xi\Pi\Psi_j, \overline{\Phi}\rangle \quad \text{for all} \quad \overline{\Phi} \in \mathfrak{H}.$$

Using (4.2), this can be further written as

$$\langle \mathcal{H}\Xi\Pi\Psi_j, \Xi^\dagger\overline{\Phi}\rangle = \mathcal{E}_j\langle \Pi\Psi_j, \Xi^\dagger\overline{\Phi}\rangle \quad \text{for all} \quad \overline{\Phi} \in \mathfrak{H}.$$

The desired result follows by noting that $\operatorname{ran}\Xi^\dagger = \mathfrak{N}$. ∎

In practice, $\mathfrak{M}$ (called "exact model space") is unknown and $\mathfrak{N}$ (called "model space") is chosen in a way that it provides a "reasonable approximation" to $\mathfrak{M}$, i.e. that (4.1) holds. In particular, $\mathfrak{M} \subset \mathfrak{N}^\perp$ is not permitted. Then, the unknown "wave operator" Ξ (hence $\mathfrak{M}$) can be determined by solving the weak Bloch equation (4.2). Next, the eigenvalue problem for $\mathcal{H}^{\text{eff}}$ is solved to obtain the energies $\mathcal{E}_1, \ldots, \mathcal{E}_M$ and (some of the) eigenvectors.

Remark 4.2.

(i) It is important to note that solving the Bloch equation only provides a weakly $\mathcal{H}$-invariant subspace $\mathfrak{M}$ and it might *not* be a direct sum of (weak) eigenspaces in general. In other words, $\mathfrak{M}$ might be spanned by an incomplete set of eigenvectors. Clearly, in such a situation some of the eigenvectors cannot be recovered through solving the eigenproblem for the effective Hamiltonian $\mathcal{H}^{\text{eff}}$.

(ii) The Bloch equation (4.2) is more commonly given in the "strong" form "$\Xi\mathcal{H}\Xi = \mathcal{H}\Xi$".

The situation is greatly simplified, when one considers one-dimensional subspaces $\mathfrak{N}$ and $\mathfrak{M}$, because a one-dimensional invariant subspace is always an eigenspace.

COROLLARY 4.3. *Let* $\dim \mathfrak{N} = \dim \mathfrak{M} = 1$, *and set* $\mathfrak{N} = \mathrm{Span}\{\Phi_0\}$ *for some* $\Phi_0 \in \mathfrak{H}$. *Further, let* $\mathfrak{M} = \mathrm{Span}\{\Psi\}$, *and suppose that* $\langle \Psi, \Phi_0 \rangle = 1$. *Then, the following are equivalent.*

(i) $\langle \mathcal{H}\Psi, \overline{\Phi} \rangle = \mathcal{E}\langle \Psi, \overline{\Phi} \rangle$ *for all* $\overline{\Phi} \in \mathfrak{H}$ *and some scalar* $\mathcal{E}$.

(ii) $\langle \mathcal{H}\Xi\Phi_0, (I - \Xi^\dagger)\Phi' \rangle = 0$ *for all* $\Phi' \in \mathfrak{N}^\perp$.

Furthermore, $\mathcal{E} = \langle \mathcal{H}\Xi\Phi_0, \Phi_0 \rangle$.

4.1. The SRCC method. The single-reference Coupled-Cluster method easily follows from Corollary 4.3 through the exponential parametrization of the wave operator. In the following theorem, we re-establish [49, Theorem 5.3] (see Theorem 2.3).

THEOREM 4.4. *Let* $\mathcal{H} : \mathfrak{H}^1_K \to \mathfrak{H}^{-1}$ *be a bounded operator. Fix* $\Phi_0 \in \mathfrak{H}^1_K$ *with* $\|\Phi_0\| = 1$ *and suppose that* $\Psi \in \mathfrak{H}^1_K$ *is such that* $\langle \Psi, \Phi_0 \rangle = 1$. *Then the following are equivalent.*

(i) $\langle \mathcal{H}\Psi, \Phi \rangle = \mathcal{E}\langle \Psi, \Phi \rangle$ *for all* $\Phi \in \mathfrak{H}^1_K$ *for some scalar* $\mathcal{E}$.

(ii) *(Full CC)* $\Psi = e^{T_*}\Phi_0$ *for some* $t_* \in \mathbb{V}(G^{\mathrm{full}})$ *such that*

$$\langle e^{-T_*}\mathcal{H}e^{T_*}\Phi_0, S\Phi_0 \rangle = 0 \quad \textit{for all} \quad s \in \mathbb{V}(G^{\mathrm{full}}). \tag{4.5}$$

Furthermore, $\mathcal{E} = \langle e^{-T_*}\mathcal{H}e^{T_*}\Phi_0, \Phi_0 \rangle$.

(iii) *(Full CI)* $\Psi = (I + C_*)\Phi_0$ *for some* $c_* \in \mathbb{V}(G^{\mathrm{full}})$ *such that*

$$\langle \mathcal{H}(I + C_*)\Phi_0, S\Phi_0 \rangle = \mathcal{E}_{\mathrm{CI}}\langle (I + C_*)\Phi_0, S\Phi_0 \rangle \quad \textit{for all} \quad s \in \mathbb{V}(G^{\mathrm{full}}), \tag{4.6}$$

where $\mathcal{E}_{\mathrm{CI}} = \langle \mathcal{H}(I + C_*)\Phi_0, \Phi_0 \rangle$. *Furthermore,* $\mathcal{E} = \mathcal{E}_{\mathrm{CI}}$.

Proof. Let $\mathfrak{H} = \mathfrak{H}^1_K$ and $\mathfrak{L} = \mathfrak{L}^2$. First, we prove (i)$\Longleftrightarrow$(ii). We apply Corollary 4.3 with the SRCC wave operator

$$\Xi = e^{T_*}\Pi_{\Phi_0},$$

where T_* is some cluster operator and Π_{Φ_0} is the orthogonal projector onto $\mathfrak{N} = \mathrm{Span}\{\Phi_0\}$. Note that $\mathfrak{N}^\perp = \mathfrak{V}(G^{\mathrm{full}})$. It is easy to see that Ξ is idempotent, and that $\ker \Xi = \mathfrak{N}^\perp$. By an appropriate choice of T_*, $\operatorname{ran} \Xi = \mathfrak{M}$ using $\langle \Psi, \Phi_0 \rangle = 1$ and Theorem 3.27. Furthermore, $\mathrm{Span}\{e^{T_*}\Phi_0\} = \operatorname{ran} \Xi \subset \mathfrak{H}$ due to Theorem 3.26. Applying Corollary 4.3, (i) holds if and only if $\Psi = e^{T_*}\Phi_0$ and T_* satisfies the weak Bloch equation

$$\langle \mathcal{H}e^{T_*}\Phi_0, (I - \Pi_{\Phi_0}e^{T_*^\dagger})S'\Phi_0 \rangle = 0 \quad \text{for all} \quad s' \in \mathbb{V}(G^{\mathrm{full}}).$$

Recalling Proposition 3.31 (ii), and using the change of variables $S' = e^{-T_*^\dagger}S$,

$$\langle e^{-T_*}\mathcal{H}e^{T_*}\Phi_0, S\Phi_0 \rangle = 0 \quad \text{for all} \quad s \in \mathbb{V}(G^{\mathrm{full}}).$$

Here we used that e^{-T_*} can be extended to a bounded $\mathfrak{H}^{-1} \to \mathfrak{H}^{-1}$ operator (Theorem 3.26).[9] Note that $\mathcal{H}^{\mathrm{eff}}$ is now a one-dimensional linear map (i.e. a multiplication by a scalar), so $\sigma(\mathcal{H}^{\mathrm{eff}}) = \langle e^{-T_*}\mathcal{H}e^{T_*}\Phi_0, \Phi_0 \rangle = \mathcal{E}$.

Next, we prove (i)$\Longleftrightarrow$(iii). We now apply Corollary 4.3 with the SRCI wave operator

$$\Xi = (I + C_*)\Pi_{\Phi_0},$$

where C_* is some cluster operator and the claim follows from a straightforward calculation. Further, now $\sigma(\mathcal{H}^{\mathrm{eff}}) = \langle \mathcal{H}(I + C_*)\Phi_0, \Phi_0 \rangle = \mathcal{E}$. ∎

[9] We refer the reader to the proof of [49, Theorem 5.3] for more details.

4.2. The Jeziorski–Monkhorst MRCC method. In MRCC methods the "model space" $\mathfrak{N}$ is chosen to be the space spanned by M orthonormal reference determinants,

$$\mathfrak{N} = \mathrm{Span}\{\Phi_{0_m} : m = 1, \dots, M\}.$$

The *Jeziorski–Monkhorst method* [27] uses the following Ansatz for the wave operator:

$$\Xi = \sum_{m=1}^{M} e^{T^{(m)}} \Pi_{\Phi_{0_m}}, \tag{4.7}$$

which corresponds to (3.8).

THEOREM 4.5. *Let $\mathfrak{N}$ as above and set $\mathfrak{M} = \mathrm{Span}\{\Psi_m : m = 1, \dots, M\}$, where $\{\Psi_m\}_{m=1}^M \subset \mathfrak{H}_K^1$ is $\mathfrak{L}^2$-orthogonal. Suppose that for every $m = 1, \dots, M$, $\langle \Psi_m, \Phi_{0_n} \rangle \neq 0$ for at least one $n = 1, \dots, M$. Then, the following are equivalent.*

(i) *$\mathfrak{M}$ is weakly $\mathcal{H}$-invariant: for every Ψ_m $(m = 1, \dots, M)$ there exists $\widetilde{\Psi}_m \in \mathfrak{M}$ such that $\langle \mathcal{H}\Psi_m, \Phi' \rangle = \langle \widetilde{\Psi}_m, \Phi' \rangle$ for all $\Phi' \in \mathfrak{H}_K^1$.*

(ii) *(Full JM-MRCC) $\mathfrak{M} = \mathrm{Span}\{e^{T_*^{(m)}} \Phi_{0_m} : m = 1, \dots, M\}$, where $t_*^{(m)} \in \mathbb{V}(G_m^{\mathrm{full}})$ satisfies*

$$\langle e^{-T_*^{(m)}} \mathcal{H} e^{T_*^{(m)}} \Phi_{0_m}, S^{(m)} \Phi_{0_m} \rangle = \sum_{n=1}^{M} \mathcal{H}_{mn}^{\mathrm{eff}} \langle e^{-T_*^{(m)}} e^{T_*^{(n)}} \Phi_{0_n}, S^{(m)} \Phi_{0_m} \rangle, \tag{4.8}$$

for all $s^{(m)} \in \mathbb{V}(G_m^{\mathrm{full}})$ and $m = 1, \dots, M$, where the matrix elements of the effective Hamiltonian are given by $\mathcal{H}_{mn}^{\mathrm{eff}} = \langle e^{-T_*^{(m)}} \mathcal{H} e^{T_*^{(m)}} \Phi_{0_m}, \Phi_{0_n} \rangle$.

(iii) *(Full MRCI) $\mathfrak{M} = \mathrm{Span}\{(I + C_*^{(m)}) \Phi_{0_m} : m = 1, \dots, M\}$, where $c_*^{(m)} \in \mathbb{V}(G_m^{\mathrm{full}})$ satisfies*

$$\langle \mathcal{H}(I + C_*^{(m)}) \Phi_{0_m}, S^{(m)} \Phi_{0_m} \rangle = \sum_{n=1}^{M} \widehat{\mathcal{H}}_{mn}^{\mathrm{eff}} \langle (I + C_*^{(n)}) \Phi_{0_n}, S^{(m)} \Phi_{0_m} \rangle, \tag{4.9}$$

for all $s^{(m)} \in \mathbb{V}(G_m^{\mathrm{full}})$ and $m = 1, \dots, M$, where the matrix elements of the effective Hamiltonian are given by $\widehat{\mathcal{H}}_{mn}^{\mathrm{eff}} = \langle \mathcal{H}(I + C_*^{(m)}) \Phi_{0_m}, \Phi_{0_n} \rangle$.

Furthermore, suppose that $\langle \mathcal{H}\Psi_m, \overline{\Phi} \rangle = \mathcal{E}_m \langle \Phi_m, \overline{\Phi} \rangle$ for all $\overline{\Phi} \in \mathfrak{H}_K^1$ and $m = 1, \dots, M$. Then the following hold true.

(a) *Suppose $\mathfrak{M}$ is given as in (ii). Then the coefficients $a_j^{(m)}$ in the expansion $\Psi_j = \sum_{n=1}^M a_j^{(n)} e^{T_*^{(n)}} \Phi_{0_n}$ are given as the solution to the eigenvalue problem*

$$\sum_{n=1}^{M} \mathcal{H}_{nm}^{\mathrm{eff}} a_j^{(n)} = \mathcal{E}_j a_j^{(m)} \quad \textit{where} \quad m = 1, \dots, M.$$

(b) *Suppose $\mathfrak{M}$ is given as in (iii). Then the coefficients $\widehat{a}_j^{(m)}$ in the expansion $\Psi_j = \sum_{n=1}^M \widehat{a}_j^{(n)} (I + C_*^{(n)}) \Phi_{0_n}$ are given as the solution to the eigenvalue problem*

$$\sum_{n=1}^{M} \widehat{\mathcal{H}}_{nm}^{\mathrm{eff}} \widehat{a}_j^{(n)} = \mathcal{E}_j \widehat{a}_j^{(m)} \quad \textit{where} \quad m = 1, \dots, M.$$

Proof. Let $\mathfrak{H} = \mathfrak{H}_K^1$. First, we prove (i)$\Longleftrightarrow$(ii) by applying Theorem 4.4. Clearly, for the JM wave operator (4.7) we have $\Xi^2 = \Xi$ and $\ker \Xi = \mathfrak{N}^\perp$ and

$$\operatorname{ran} \Xi = \operatorname{Span}\{e^{T^{(m)}} \Phi_{0_m} : m = 1, \dots, M\}.$$

The weak Bloch equation (4.2) is equivalent to

$$\langle \mathcal{H} e^{T_*^{(m)}} \Phi_{0_m}, \Phi' \rangle = \sum_{n=1}^{M} \langle \mathcal{H} e^{T_*^{(m)}} \Phi_{0_m}, \Pi_{\Phi_{0_n}} e^{(T_*^{(n)})^\dagger} \Phi' \rangle$$

for all $\Phi' \in \mathfrak{N}^\perp$ and $m = 1, \dots, M$. Setting $\Phi' = S^{(m)} \Phi_{0_m}$, we obtain

$$\begin{aligned}
\langle \mathcal{H} e^{T_*^{(m)}} \Phi_{0_m}, S^{(m)} \Phi_{0_m} \rangle &= \sum_{n=1}^{M} \langle \mathcal{H} e^{T_*^{(m)}} \Phi_{0_m}, \Pi_{\Phi_{0_n}} e^{(T_*^{(n)})^\dagger} S^{(m)} \Phi_{0_m} \rangle \\
&= \sum_{n=1}^{M} \langle \mathcal{H} e^{T_*^{(m)}} \Phi_{0_m}, \Phi_{0_n} \rangle \langle e^{(T_*^{(n)})^\dagger} S^{(m)} \Phi_{0_m}, \Phi_{0_n} \rangle \\
&= \sum_{n=1}^{M} \langle e^{-T_*^{(m)}} \mathcal{H} e^{T_*^{(m)}} \Phi_{0_m}, \Phi_{0_n} \rangle \langle e^{T_*^{(n)}} \Phi_{0_n}, S^{(m)} \Phi_{0_m} \rangle
\end{aligned}$$

for all $s^{(m)} \in \mathbb{V}(G_m^{\text{full}})$. Here, we used that $(T^{(m)})^\dagger \Phi_{0_n} = 0$, see Theorem 3.21 (iii). The proof of (4.8) is finished by invoking Proposition 3.31 (ii) and replacing $S^{(m)}$ by $(e^{-T_*^{(m)}})^\dagger S^{(m)}$.

Next, we prove (i)$\Longleftrightarrow$(iii). The MRCI wave operator reads

$$\Xi = \sum_{m=1}^{M} (I + C_*^{(m)}) \Pi_{\Phi_{0_m}}.$$

With this choice (4.2) is equivalent to

$$\langle \mathcal{H}(I + C_*^{(m)}) \Phi_{0_m}, \Phi' \rangle = \sum_{n=1}^{M} \langle \mathcal{H}(I + C_*^{(m)}) \Phi_{0_m}, \Pi_{\Phi_{0_n}} (I + C_*^{(n)})^\dagger \Phi' \rangle$$

for all $\Phi' \in \mathfrak{N}^\perp$ and $m = 1, \dots, M$. Setting $\Phi' = S^{(m)} \Phi_{0_m}$, this can be written as

$$\begin{aligned}
\langle \mathcal{H}(I + C_*^{(m)}) \Phi_{0_m}, S^{(m)} \Phi_{0_m} \rangle &= \sum_{n=1}^{M} \langle \mathcal{H}(I + C_*^{(m)}) \Phi_{0_m}, \Pi_{\Phi_{0_n}} (I + C_*^{(n)})^\dagger S^{(m)} \Phi_{0_m} \rangle \\
&= \sum_{n=1}^{M} \langle \mathcal{H}(I + C_*^{(m)}) \Phi_{0_m}, \Phi_{0_n} \rangle \langle (I + C_*^{(n)})^\dagger S^{(m)} \Phi_{0_m}, \Phi_{0_n} \rangle \\
&= \sum_{n=1}^{M} \langle \mathcal{H}(I + C_*^{(m)}) \Phi_{0_m}, \Phi_{0_n} \rangle \langle (I + C_*^{(n)}) \Phi_{0_n}, S^{(m)} \Phi_{0_m} \rangle,
\end{aligned}$$

which is what we wanted to prove.

For the "furthermore" part of (a), expanding Ψ_j as $\Psi_j = \sum_{n=1}^{M} a_j^{(n)} e^{T_*^{(n)}} \Phi_{0_n}$, for some scalars $a_j^{(n)}$, we find that $a_j^{(m)} = \langle \Psi_j, \Phi_{0_m} \rangle$. It is easy to see that (4.4) now reads

$$\sum_{n=1}^{M} \langle \mathcal{H} e^{T_*^{(n)}} \Phi_{0_n}, \Phi_{0_m} \rangle a_j^{(n)} = \mathcal{E}_j a_j^{(m)}$$

for all $j = 1, \dots, M$. The proof of the "furthermore" part of (b) is similar. ∎

5. Analysis of the SRCC method. The rest of the article is concerned with the analysis of the nonlinear equation (4.5) and its variants.

It is known that the *truncated* CC equations have a large number of solutions, some of which exhibit unphysical behavior. Recall that the truncated CC equations are *not* equivalent to the truncated CI equations, hence, there is no obvious connection with the Schrödinger equation in general. Also, the SRCC method is typically unreliable for treating degenerate states. Since the early days of the CC method, researchers in quantum chemistry have been interested in understanding the complicated behavior of the (truncated) CC equations [60, 61, 42, 26, 29, 25, 43, 23, 24]. It is one of our main goals here to explain some of the numerical observations and provide tools for further investigations.

Readers are advised, before proceeding, to take a glance at Appendix A and Appendix B, where the relevant results of topological degree theory are collected.

We consider the finite-dimensional case only, by which we mean that

$$\boxed{K < \infty}$$

is assumed. Recall the definition of $\mathfrak{H}^1_K$ from subsection 2.1 and the interpretation of the projected Schrödinger equation from subsection 2.2.

5.1. Definitions and basic properties. Let $\mathbb{V} := \mathbb{V}(G)$ be the real amplitude space corresponding to some consistent excitation graph G. Let $\mathfrak{V}$ be the corresponding functional amplitude space. Define $\mathcal{A} : \mathbb{V} \to \mathbb{V}^*$ via the instruction

$$\langle \mathcal{A}(t), s\rangle = \langle e^{-T}\mathcal{H}e^{T}\Phi_0, S\Phi_0\rangle, \tag{5.1}$$

for any $t, s \in \mathbb{V}$, where T and S are the cluster operators corresponding to t and s, respectively—this convention will be used throughout. Interpreting $\mathcal{H}$ as a bounded map $\mathfrak{H}^1 \to \mathfrak{H}^{-1}$, we obtain that $\mathbb{V} \ni s \mapsto \langle \mathcal{A}(t), s\rangle$ defines a bounded linear functional in $\mathbb{V}^*$ for every given $t \in \mathbb{V}$. In fact, using the $\mathfrak{H}^1$-boundedness (2.1) of the Hamilton operator and Theorem 3.26,

$$|\langle \mathcal{A}(t), s\rangle| \leq M\|e^{T}\Phi_0\|_{\mathfrak{H}^1}\|e^{-T^\dagger}S\Phi_0\|_{\mathfrak{H}^1} \leq M\|e^{T}\Phi_0\|_{\mathfrak{H}^1}\|e^{-T^\dagger}\|_{\mathcal{L}(\mathfrak{H}^1,\mathfrak{H}^1)}\|S\Phi_0\|_{\mathfrak{H}^1}$$

for all $t, s \in \mathbb{V}$.

Recall the definition (2.9) of the CC energy, which can be written as

$$\mathcal{E}_{\mathrm{CC}}(t) = \langle e^{-T}\mathcal{H}e^{T}\Phi_0, \Phi_0\rangle = \langle \mathcal{H}e^{T}\Phi_0, \Phi_0\rangle,$$

where the second equality follows from $(e^{-T})^\dagger\Phi_0 = \Phi_0$. Note that $\mathcal{E}_{\mathrm{CC}}(t) \in \mathbb{R}$ since the amplitude space $\mathbb{V}$ is assumed to be real. The similarity-transformed Hamilton operator occurs often in the forthcoming discussion, so we introduce the notation

$$\mathcal{H}(t) = e^{-T}\mathcal{H}e^{T} : \mathfrak{H}^1 \to \mathfrak{H}^{-1}, \tag{5.2}$$

which is a bounded map for any fixed cluster amplitude $t \in \mathbb{V}$. Furthermore, for a given bounded map $\mathcal{T} : \mathfrak{H}^1 \to \mathfrak{H}^{-1}$, we define the operator $\mathcal{T}_{\mathfrak{V}} : \mathfrak{V} \to \mathfrak{V}$ via $\langle \mathcal{T}_{\mathfrak{V}}\Psi, \Psi'\rangle = \langle \mathcal{T}\Psi, \Psi'\rangle$ for all $\Psi, \Psi' \in \mathfrak{V}$.

We will also use a notation analogous to (5.2) for the similarity-transformed fluctuation operator $\mathcal{W}$ (see (2.4)), i.e. $\mathcal{W}(t) = e^{-T}\mathcal{H}e^{T}$. The similarity-transformed Fock operator can be given explicitly as

$$e^{-T}\mathcal{F}e^{T} = \mathcal{F} + [\mathcal{F}, T], \quad \text{and} \quad [\mathcal{F}, X_\alpha] = \varepsilon_\alpha X_\alpha, \tag{5.3}$$

for any $t \in \mathbb{V}$, see e.g. [16, Lemma 15]. In particular,

$$[[\mathcal{H}(t), U], V] = [[\mathcal{W}(t), U], V], \tag{5.4}$$

for any $t, u, v \in \mathbb{V}$.

The following simple observation shows the equivalence of the (strong) Schrödinger equation with the Full CC method.

LEMMA 5.1. *Assume that the determinantal basis functions satisfy $\Phi_\alpha \in \mathfrak{H}^2$. Suppose that $\mathbb{V} = \mathbb{V}(G^{\mathrm{full}})$, and that $\mathcal{A}(t_*) = 0$. Then the function $\Psi = (c_0 I + C)\Phi_0 \in \mathfrak{H}^2$ satisfies the Schrödinger equation $\mathcal{H}\Psi = \mathcal{E}\Psi$ if and only if $e^{-T_*}(c_0 I + C) = r_0 I + R$, where ($c_0 = r_0$ and)*

$$\left.\begin{aligned} \mathcal{E}_{\mathrm{CC}}(t_*) r_0 + \langle \mathcal{H}(t_*) R\Phi_0, \Phi_0\rangle &= \mathcal{E} r_0 \\ \mathcal{H}_{\mathfrak{V}}(t_*) R\Phi_0 &= \mathcal{E} R\Phi_0 \end{aligned}\right\} \tag{5.5}$$

Furthermore, with $\mathfrak{V}_0 = \mathrm{Span}\{\Phi_0\} \oplus \mathfrak{V}$,

$$\sigma(\mathcal{H}_{\mathfrak{V}_0}) = \{\mathcal{E}_{\mathrm{CC}}(t_*)\} \cup \sigma(\mathcal{H}_{\mathfrak{V}}(t_*)). \tag{5.6}$$

Proof. Using the splitting $\mathfrak{V}_0 = \mathrm{Span}\{\Phi_0\} \oplus \mathfrak{V}$, the similarity-transformed Hamilton operator is block upper triangular in the determinantal basis,

$$\mathcal{H}(t_*) = \begin{pmatrix} \mathcal{E}_{\mathrm{CC}}(t_*) & \langle \mathcal{H}(t_*)\cdot, \Phi_0\rangle \\ 0 & \mathcal{H}_{\mathfrak{V}}(t_*) \end{pmatrix},$$

due to $\langle \mathcal{H}(t_*)\Phi_0, \Phi_\alpha\rangle = 0$. The proof now follows by noting that the eigenvalues of $\mathcal{H}(t_*)$ and $\mathcal{H}$ are the same, and the eigenvectors of $\mathcal{H}(t_*)$ are of the form $e^{-T_*}\Phi$, where $\mathcal{H}\Phi = \mathcal{E}'\Phi$. Formula (5.6) follows by the fact that the spectrum of a block triangular matrix is the union of the spectra of the blocks in the diagonal. ∎

Obviously, $r_0 \neq 0$ and $R = 0$ is a solution to the system (5.5) if and only if $\mathcal{E}_{\mathrm{CC}}(t_*) = \mathcal{E}$. In this case, $\Psi = e^{T_*}\Phi_0$ is a solution, and the similarity-transformed Hamilton operator $\mathcal{H}(t_*)$ is block diagonal.

If, however, $\mathcal{E}_{\mathrm{CC}}(t_*) \neq \mathcal{E}$, then R cannot be 0 (because that would imply $\Psi = 0$). In this case, $\Psi = (r_0 I + R)e^{T_*}\Phi_0$, where $r_0 = \langle \mathcal{H}(t_*)R\Phi_0, \Phi_0\rangle / (\mathcal{E} - \mathcal{E}_{\mathrm{CC}}(t_*))$. Note that it is possible to have $r_0 = 0$, in which case $\langle \Psi, \Phi_0\rangle = 0$. We return to this latter case in Remark 5.16 and discuss the former below.

Remark 5.2. If $\mathcal{E}_{\mathrm{CC}}(t_*) = \mathcal{E}$, then (5.5) reduces to

$$\left.\begin{aligned} \langle \mathcal{H}(t_*) R\Phi_0, \Phi_0\rangle &= 0 \\ \mathcal{H}_{\mathfrak{V}}(t_*) R\Phi_0 &= \mathcal{E} R\Phi_0 \end{aligned}\right\}$$

Suppose that this system has μ linearly independent solutions $R_1, \ldots, R_\mu$. Then it is easy to see, using $\Psi = (r_0 I + R_\mu)e^{T_*}\Phi_0$, that the wavefunctions

$$\{e^{T_*}\Phi_0, R_1 e^{T_*}\Phi_0, \ldots, R_\mu e^{T_*}\Phi_0\}$$

span the eigenspace $\ker(\mathcal{H} - \mathcal{E})$. In particular, we have $\dim\ker(\mathcal{H} - \mathcal{E}) = \mu + 1$. Note also that in this case $\sigma(\mathcal{H}) = \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$.

Let us now recall that the CC equation $\mathcal{A}(t_*) = 0$ can be cast in a form that closely resembles the CI eigenvalue problem (2.7) (although it is *not* equivalent to it in general).

LEMMA 5.3. *[51, Theorem 5.6] Let G be excitation complete, and let $\mathbb{V} = \mathbb{V}(G)$ be the corresponding amplitude space. Then the "linked" CC equation $\mathcal{A}(t_*) = 0$ is equivalent to the "unlinked" (a.k.a. "energy-dependent") CC equation*

$$\langle \mathcal{H} e^{T_*} \Phi_0, S\Phi_0 \rangle = \mathcal{E}_{\mathrm{CC}}(t_*) \langle e^{T_*} \Phi_0, S\Phi_0 \rangle \quad \textit{for all} \quad s \in \mathbb{V}. \tag{5.7}$$

Proof. For part (i), we have

$$\begin{aligned}
\langle (\mathcal{H} - \mathcal{E}_{\mathrm{CC}}(t_*)) e^{T_*} \Phi_0, S\Phi_0 \rangle &= \langle e^{-T_*} (\mathcal{H} - \mathcal{E}_{\mathrm{CC}}(t_*)) e^{T_*} \Phi_0, (e^{T_*})^\dagger S\Phi_0 \rangle \\
&= \langle e^{-T_*} (\mathcal{H} - \mathcal{E}_{\mathrm{CC}}(t_*)) e^{T_*} \Phi_0, \Pi_{\mathfrak{V}} (e^{T_*})^\dagger S\Phi_0 \rangle \\
&+ \langle e^{-T_*} (\mathcal{H} - \mathcal{E}_{\mathrm{CC}}(t_*)) e^{T_*} \Phi_0, \underbrace{\Pi_{\Phi_0} (e^{T_*})^\dagger S\Phi_0}_{\mathrm{const} \cdot \Phi_0} \rangle \\
&= \langle e^{-T_*} \mathcal{H} e^{T_*} \Phi_0, \Pi_{\mathfrak{V}} (e^{T_*})^\dagger S\Phi_0 \rangle,
\end{aligned}$$

where second term on the right-hand side of the penultimate equality vanishes by the definition of $\mathcal{E}_{\mathrm{CC}}(t_*)$. The proof is completed by recalling that $\Pi_{\mathfrak{V}} (e^{T_*})^\dagger : \mathfrak{V} \to \mathfrak{V}$ is surjective due to Proposition 3.31. ∎

The "unlinked" form is less useful in practice, because the expansion of $\mathcal{H} e^T$ does not terminate like the Baker–Campbell–Hausdorff series (2.12) for $\mathcal{H}(t)$,

$$\mathcal{H}(t) = \sum_{j=0}^{4} \frac{1}{j!} [\mathcal{H}, T]_{(j)}. \tag{5.8}$$

More generally, the *doubly*[10] similarity-transformed Hamilton operator $\mathcal{H}(t+s) = e^{-S} \mathcal{H}(t) e^S$ can also be expanded using the Baker–Campbell–Hausdorff series but in this case

$$\mathcal{H}(t+s) = \sum_{j=0}^{2N} \frac{1}{j!} [\mathcal{H}(t), S]_{(j)}, \tag{5.9}$$

i.e. the series terminates at $2N$. To see this, simply note that $[\mathcal{H}(t), S]_{(2N+1)}$ consists of terms of the form $S^i \mathcal{H}(t) S^k$, where $i + k = 2N + 1$, so $i, k \geq N + 1$, which, using Proposition 3.22 implies that all terms for $j \geq 2N + 1$ vanish.

5.2. Local properties—real case. Next, we look at the local behavior of the CC mapping $\mathcal{A} : \mathbb{V} \to \mathbb{V}^*$ for general (real) amplitude spaces $\mathbb{V}$. For fixed $t \in \mathbb{V}$, define the *modified similarity-transformed Hamilton operator*,

$$\widehat{\mathcal{H}}(t) = \mathcal{H}(t) - \sum_{\alpha \in \Xi(G)^c} \langle \mathcal{H}(t) \Phi_0, \Phi_\alpha \rangle X_\alpha, \tag{5.10}$$

where $\Xi(G)^c = \Xi(G^{\mathrm{full}}) \setminus \Xi(G)$ and $\Xi(G)$ was defined in (3.3).

DEFINITION 5.4. *The amplitude space $\mathbb{V}(G)$ is said to be* rank-regular*, if*

$$\langle X_\alpha \Phi_\beta, \Phi_\gamma \rangle = 0 \quad \textit{for all } \beta, \gamma \in \Xi(G) \textit{ and } \alpha \in \Xi(G)^c.$$

We immediately get that $\widehat{\mathcal{H}}_{\mathfrak{V}}(t) = \mathcal{H}_{\mathfrak{V}}(t)$ if $\mathbb{V}(G)$ is rank-regular. The next proposition shows that the truncated subgraphs typically used in practice are rank-regular.

[10] Note that the doubly similarity-transformation above differs from the one considered in Arponen's Extended CC (ECC) theory.

PROPOSITION 5.5. *Suppose that the excitation graph G is a rank-truncated subgraph of the form $G = G(1, 2, \dots, \rho)$, for some $\rho = 1, \dots, N$ or $G = G(\mathrm{D})$. Then $\mathbb{V}(G)$ is rank-regular.*

Proof. The set $\Xi(G)^c$ consists of elements of rank $\rho + 1, \dots, N$ (or empty), so that $\langle X_\alpha \Phi_\beta, \Phi_\gamma \rangle = 0$ for all $\mathrm{rk}\,\alpha \notin \{1, 2, \dots, \rho\}$ and all $\mathrm{rk}\,\beta, \mathrm{rk}\,\gamma \in \{1, 2, \dots, \rho\}$, due to the fact that $X_\alpha \Phi_\beta$ is of rank $\mathrm{rk}(\alpha) + \mathrm{rk}(\beta) \notin \{1, 2, \dots, \rho\}$. The proof of the case $G = G(\mathrm{D})$ is similar. ∎

LEMMA 5.6. *Let t_* be a zero of $\mathcal{A} : \mathbb{V} \to \mathbb{V}^*$. Then the derivative $\mathcal{A}'(t_*) \in \mathcal{L}(\mathbb{V}, \mathbb{V}^*)$ is given by*

$$\langle \mathcal{A}'(t_*)u, v \rangle = \langle (\widehat{\mathcal{H}}(t_*) - \mathcal{E}_{\mathrm{CC}}(t_*))U\Phi_0, V\Phi_0 \rangle \tag{5.11}$$

for all $u, v \in \mathbb{V}$.

Proof. The derivative $\mathcal{A}' : \mathbb{V} \to \mathcal{L}(\mathbb{V}, \mathbb{V}^*)$ is readily computed as

$$\begin{aligned} \frac{d}{dh}\langle \mathcal{A}(t + hu), v \rangle \Big|_{h=0} &= \frac{d}{dh}\langle e^{-T-hU}\mathcal{H}e^{T+hU}\Phi_0, V\Phi_0 \rangle \Big|_{h=0} \\ &= \langle e^{-T-hU}(\mathcal{H}U - U\mathcal{H})e^{T+hU}\Phi_0, V\Phi_0 \rangle \big|_{h=0} \\ &= \langle e^{-T}(\mathcal{H}U - U\mathcal{H})e^{T}\Phi_0, V\Phi_0 \rangle, \end{aligned}$$

so using the commutativity of the cluster operators, we get

$$\langle \mathcal{A}'(t)u, v \rangle = \langle [\mathcal{H}(t), U]\Phi_0, V\Phi_0 \rangle \tag{5.12}$$

for all $t, u, v \in \mathbb{V}$. Expanding $U^\dagger V\Phi_0 \in \mathfrak{H}^1_K$ in the $\mathfrak{L}^2$-orthonormal basis $\{\Phi_\alpha\}_{\alpha \in L} \subset \mathfrak{H}^1_K$,

$$U^\dagger V\Phi_0 = \sum_{\alpha \in L} \langle U\Phi_\alpha, V\Phi_0 \rangle \Phi_\alpha,$$

we obtain using $\langle \mathcal{H}(t_*)\Phi_0, \Phi_\alpha \rangle = 0$ for all $\alpha \in \Xi(G)$,

$$\begin{aligned} \langle U\mathcal{H}(t_*)\Phi_0, V\Phi_0 \rangle &= \langle \mathcal{H}(t_*)\Phi_0, U^\dagger V\Phi_0 \rangle \\ &= \mathcal{E}_{\mathrm{CC}}(t_*)\langle U\Phi_0, V\Phi_0 \rangle + \sum_{\alpha \in \Xi(G)^c} \langle \mathcal{H}(t_*)\Phi_0, \Phi_\alpha \rangle \langle X_\alpha U\Phi_0, V\Phi_0 \rangle \end{aligned}$$

for all $u, v \in \mathbb{V}$. Inserting this into (5.12) with $t = t_*$, we obtain the stated formula. ∎

As we noted in subsection 1.1, previous analyses of the CC mapping assumed the local strong monotonicity at a zero t_*, i.e. that there is a $\delta > 0$ and a constant $C_{\mathrm{SM}}(t_*, \delta) > 0$ such that

$$\langle \mathcal{A}(t) - \mathcal{A}(s), t - s \rangle \geq C_{\mathrm{SM}}(t_*, \delta)\|t - s\|^2_{\mathbb{V}}, \quad \text{for all} \quad t, s \in B_{\mathbb{V}}(t_*, \delta). \tag{5.13}$$

The following elementary theorem makes the observations in [50] more precise.

THEOREM 5.7. *Let $t_* \in \mathbb{V}$ be a zero of $\mathcal{A} : \mathbb{V} \to \mathbb{V}^*$.*

(i) *If $\mathcal{A}$ is strongly monotone in $B_{\mathbb{V}}(t_*, \delta)$ for some $\delta > 0$, then there exists $\delta' > 0$ such that $\mathcal{A}'(t_* + u)$ is $\mathbb{V}$-coercive for all $\|u\|_{\mathbb{V}} < \delta'$ with some constant $0 < \gamma \leq C_{\mathrm{SM}}(t_*, \delta)$, i.e.*

$$\langle \mathcal{A}'(t_* + u)v, v \rangle \geq \gamma\|v\|^2_{\mathbb{V}} \quad \textit{for all} \quad v \in \mathbb{V} \textit{ and } \|u\|_{\mathbb{V}} < \delta'. \tag{5.14}$$

(ii) *Conversely, if* (5.14) *holds with* $u = 0$*, then* (5.13) *holds true with* $\delta > 0$ *chosen so that* $C_{\mathrm{SM}}(t_*, \delta) := \gamma - M_\delta \delta > 0$*, where*

$$M_\delta = \sup_{\|\zeta\|_{\mathbb{V}} \le \delta} \|\mathcal{A}''(t_* + \zeta)\|_{\mathcal{L}(\mathbb{V}\times\mathbb{V}, \mathbb{V}^*)}. \tag{5.15}$$

Proof. To see (i), fix $\delta' > 0$ and $\|u\|_{\mathbb{V}} < \delta'$ and write for any $\|r\|_{\mathbb{V}} < \delta' < \delta$,

$$\begin{aligned} C_{\mathrm{SM}}(t_*, \delta)\|r - u\|_{\mathbb{V}}^2 &\le \langle \mathcal{A}(t_* + r) - \mathcal{A}(t_* + u), r - u\rangle \\ &\le \langle \mathcal{A}'(t_* + u)(r - u), r - u\rangle + \frac{1}{2} M_\delta \|r - u\|_{\mathbb{V}}^3. \end{aligned}$$

This implies

$$(C_{\mathrm{SM}}(t_*, \delta) - M_\delta \delta')\|r - u\|_{\mathbb{V}}^2 \le \langle \mathcal{A}'(t_* + u)(r - u), r - u\rangle.$$

Any vector $v \in \mathbb{V}$ can be expressed as $v = \alpha(r - u)$ for some $\alpha > 0$ and $\|r\|_{\mathbb{V}} < \delta'$, from which $\mathbb{V}$-coercivity follows with $\gamma = C_{\mathrm{SM}}(t_*, \delta) - M_\delta \delta'$, by choosing δ' sufficiently small.

Next, to prove (ii), write the Taylor expansions of $\mathcal{A}$ at t_*,

$$\begin{aligned} \mathcal{A}(t_* + r) &= \mathcal{A}'(t_*)r + \mathcal{R}_2(t_*; r), \\ \mathcal{A}(t_* + r') &= \mathcal{A}'(t_*)r' + \mathcal{R}_2(t_*; r'), \end{aligned}$$

for any $\|r\|_{\mathbb{V}}, \|r'\|_{\mathbb{V}} < \delta$ for some $\delta > 0$, from which we obtain

$$\begin{aligned} \langle \mathcal{A}(t_* + r) - \mathcal{A}(t_* + r'), r - r'\rangle &= \langle \mathcal{A}'(t_*)(r - r'), r - r'\rangle + \langle \mathcal{R}_2(t_*; r) - \mathcal{R}_2(t_*; r'), r - r'\rangle \\ &\ge \gamma \|r - r'\|_{\mathbb{V}}^2 + \langle \mathcal{R}_2(t_*; r) - \mathcal{R}_2(t_*; r'), r - r'\rangle. \end{aligned}$$

Using the intermediate value inequality, we have

$$\|\mathcal{R}_2(t_*; r) - \mathcal{R}_2(t_*; r')\|_{\mathbb{V}^*} \le M_{r,r'} \|r - r'\|_{\mathbb{V}},$$

where

$$\begin{aligned} M_{r,r'} &= \max_{\xi \in [r,r']} \|\partial_2 \mathcal{R}_2(t_*; \xi)\| = \max_{\xi \in [r,r']} \|\mathcal{A}'(t_* + \xi) - \mathcal{A}'(t_*)\| \\ &\le \Big(\max_{\xi \in [r,r']} \max_{\zeta \in [0,\xi]} \|\mathcal{A}''(t_* + \zeta)\|_{\mathbb{V}\times\mathbb{V}} \Big) \delta \\ &= \Big(\sup_{\|\zeta\| \le \delta} \|\mathcal{A}''(t_* + \zeta)\|_{\mathbb{V}\times\mathbb{V}} \Big) \delta = M_\delta \delta \end{aligned}$$

This implies

$$\langle \mathcal{A}(t_* + r) - \mathcal{A}(t_* + r'), r - r'\rangle \ge (\gamma - M_\delta \delta)\|r - r'\|_{\mathbb{V}}^2$$

for all $\|r\|_{\mathbb{V}}, \|r'\|_{\mathbb{V}} < \delta$. Setting $t = t_* + r$ and $s = t_* + r'$ proves the claim. ∎

Remark 5.8. Let $\mathbb{V}^0 \subset \mathbb{V}$ be a subspace and consider the *projected CC mapping* $\mathcal{A}^0 : \mathbb{V}^0 \to (\mathbb{V}^0)^*$ via

$$\langle \mathcal{A}^0(t^0), s^0\rangle = \langle \mathcal{A}(t^0), s^0\rangle \quad \text{for all} \quad t^0, s^0 \in \mathbb{V}^0.$$

Clearly, if $\mathcal{A}$ is strongly monotone on $B_{\mathbb{V}}(t_*, \delta)$ with a constant $C_{\mathrm{SM}} > 0$ at a zero t_*, then $\mathcal{A}^0$ is strongly monotone on $B_{\mathbb{V}^0}(t_*^0, \sqrt{\delta^2 - \|t_*^\perp\|_{\mathbb{V}}^2})$ with the same constant C_{SM}, where we have set $t_*^0 = \Pi_{\mathbb{V}^0} t_*$ and $t_*^\perp = (I - \Pi_{\mathbb{V}^0}) t_*$. Note that t_*^0 is *not*, in general, a zero of $\mathcal{A}^0$, hence the preceding theorem is not applicable to $\mathcal{A}^0$.

Remark 5.9. The quantity M_δ contains the second derivative of $\mathcal{A}$. It is easy to see that

$$\langle \mathcal{A}''(t)(u,v), w\rangle = \langle [[\mathcal{H}(t), U], V]\Phi_0, W\Phi_0\rangle$$

for all $u, v, w \in \mathbb{V}$. Using (5.4), we have

$$\langle \mathcal{A}''(t)(u,v), w\rangle = \langle [[\mathcal{W}(t), U], V]\Phi_0, W\Phi_0\rangle,$$

so that $\mathcal{A}''(t)$ only involves the fluctuation operator $\mathcal{W}$.

Remark 5.10 (Perturbative regime). Consider the case when $t_* \approx 0$, which is the case considered in [51, 50]. Then roughly speaking, we have $\mathcal{H}(t_*) \approx \mathcal{H}$. Note that

$$\langle \mathcal{A}'(t_*)r, r\rangle = \langle (\mathcal{H} - \mathcal{E}_{\mathrm{CC}}(t_*))R\Phi_0, R\Phi_0\rangle + \mathcal{O}(\|t_*\|_{\mathbb{V}}).$$

Consequently, if

$$\langle (\mathcal{H} - \mathcal{E}_{\mathrm{CC}}(t_*))R\Phi_0, R\Phi_0\rangle \geq c(t_*)\|r\|_{\mathbb{V}}^2,$$

where $c(t_*) > 0$, then local strong monotonicity holds with constant $C_{\mathrm{SM}} = c(t_*) - M'\|t_*\|_{\mathbb{V}} - 2M_\delta\delta$ for t_* sufficiently close to 0. In [50, Lemma 3.5], it is shown that such a $c(t_*)$ exists under the assumption that $\mathcal{H}$ has a spectral gap and that Φ_0 is a sufficiently good approximation of the ground state $e^{T_*}\Phi_0$ (i.e. that t_* is sufficiently close to 0).

PROPOSITION 5.11. *If $\mathcal{A}$ is locally strongly monotone near a zero t_*, then t_* is non-degenerate.*

Proof. Suppose that $\ker \mathcal{A}'(t_*) \neq \{0\}$ and that $\mathcal{A}$ is strongly monotone near t_*. Then for any $0 \neq r \in \ker \mathcal{A}'(t_*)$ sufficiently close to 0, we have

$$C_{\mathrm{SM}}(\delta)\|r\|_{\mathbb{V}}^2 \leq \langle \mathcal{A}(t_* + r), r\rangle = \frac{1}{2}\langle \mathcal{A}''(t_*)(r,r), r\rangle + o(\|r\|_{\mathbb{V}}^4).$$

Rescaling r by $\alpha > 0$ small, and letting $\alpha \to 0$ we obtain that $C_{\mathrm{SM}} = 0$, a contradiction. ∎

Remark 5.12. When applying topological degree theory, we will view $\mathcal{A} : \mathbb{V} \to \mathbb{V}^*$ as a mapping $\mathbb{R}^n \to \mathbb{R}^n$ by identifying $\mathbb{V}$ and $\mathbb{V}^*$ with $\mathbb{R}^n$. Following [14, Chapter 6], we fix a basis $\{\tau_\alpha\}_{\alpha\in\Xi(G)}$ of $\mathbb{V}$ and define the linear homeomorphism $h : \mathbb{V} \to \mathbb{R}^n$ with

$$\mathbb{V} \ni t = \sum_{\alpha\in\Xi(G)} \widehat{t}_\alpha \tau_\alpha \mapsto h(t) = \sum_{\alpha\in\Xi(G)} \widehat{t}_\alpha e_\alpha \in \mathbb{R}^n,$$

where $\{e_\alpha\}_{\alpha\in\Xi(G)}$ is the standard (ordered) basis in $\mathbb{R}^n$. Also, fix a basis $\{\tau_\alpha^*\}_{\alpha\in\Xi(G)}$ of $\mathbb{V}^*$ and define the linear homeomorphism $g : \mathbb{V}^* \to \mathbb{R}^n$ analogously. Then

$$\widehat{\mathcal{A}} := g \circ \mathcal{A} \circ h^{-1} : \mathbb{R}^n \to \mathbb{R}^n$$

gives the desired mapping. Now suppose that two other bases $\{\widetilde{\tau}_\alpha\}_{\alpha\in\Xi(G)} \subset \mathbb{V}$ and $\{\widetilde{\tau}_\alpha^*\}_{\alpha\in\Xi(G)} \subset \mathbb{V}^*$ are given and let $\widetilde{h} : \mathbb{V} \to \mathbb{R}^n$ and $\widetilde{g} : \mathbb{V}^* \to \mathbb{R}^n$ be the corresponding linear homeomorphism. But then

$$g^{-1} \circ g \circ \mathcal{A} \circ h^{-1} \circ h = \mathcal{A} = \widetilde{g}^{-1} \circ \widetilde{g} \circ \mathcal{A} \circ \widetilde{h}^{-1} \circ \widetilde{h},$$

which implies $\widetilde{\mathcal{A}} := \widetilde{g} \circ \mathcal{A} \circ \widetilde{h}^{-1} = m \circ \widehat{\mathcal{A}} \circ \widetilde{m}$, where $m = \widetilde{g} \circ g^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ and $\widetilde{m} = h \circ \widetilde{h}^{-1} : \mathbb{R}^n \to \mathbb{R}^n$. Using [14, Lemma 6.1.1], we obtain

$$\deg(\widetilde{\mathcal{A}}, \widetilde{h}(D), \widetilde{g}(0)) = (\operatorname{sgn}\det m)(\operatorname{sgn}\det \widetilde{m}) \deg(\widehat{\mathcal{A}}, h(D), g(0))$$

for any open and bounded set $D \subset \mathbb{V}$ with $0 \notin \mathcal{A}(\partial D)$. We can conclude that the topological degree is independent of the choice of the basis if $\mathbb{V}$ and $\mathbb{V}^*$ are oriented the same.

Next, we determine the topological index of a zero of $\mathcal{A}$. The fact that the topological index of t_* is related to its CC energy $\mathcal{E}_{\mathrm{CC}}(t_*)$ and the eigenvalues of the operator $\widehat{\mathcal{H}}_{\mathfrak{V}}(t_*)$ is interesting on its own right.

THEOREM 5.13 (Index formula for SRCC—non-degenerate case). *Let t_* be a zero of the CC mapping $\mathcal{A} : \mathbb{V} \to \mathbb{V}^*$. Then t_* is non-degenerate if and only if $\mathcal{E}_{\mathrm{CC}}(t_*) \notin \sigma(\widehat{\mathcal{H}}_{\mathfrak{V}}(t_*))$, and in this case t_* is an isolated zero and the topological index of $\mathcal{A}$ at t_* is given by*

$$i(\mathcal{A}, t_*) = (-1)^\nu,$$

where

$$\nu = |\{j : \mathcal{E}_j(\widehat{\mathcal{H}}_{\mathfrak{V}}(t_*)) \in \mathbb{R},\ \mathcal{E}_j(\widehat{\mathcal{H}}_{\mathfrak{V}}(t_*)) < \mathcal{E}_{\mathrm{CC}}(t_*)\}|.$$

Proof. It is trivial to see that if t_* is non-degenerate, then it is isolated: assume that $\ker \mathcal{A}'(t_*) = \{0\}$ and write

$$\langle \mathcal{A}(t_* + r), \mathcal{J}(\mathcal{A}'(t_*)r)\rangle = \|\mathcal{A}'(t_*)r\|_{\mathbb{V}}^2 + o(\|r\|_{\mathbb{V}}^3), \tag{5.16}$$

for all $\|r\|_{\mathbb{V}} = \varepsilon$, where $\varepsilon > 0$ is sufficiently small. Here, $\mathcal{J} : \mathbb{V} \to \mathbb{V}^*$ denotes the (normalized) duality mapping. This implies that $\mathcal{A}(t_* + r) \neq 0$ for all $0 < \|r\|_{\mathbb{V}} < \varepsilon$.

We can apply Theorem A.6 with the mappings $h : \mathbb{V} \to \mathbb{R}^n$ and $g : \mathbb{V}^* \to \mathbb{R}^n$ defined in Remark 5.12. Using the notations of the said remark and (5.11), we have

$$\begin{aligned}\langle \widehat{\mathcal{A}}'(h^{-1}(t_*))e_\alpha, e_\beta\rangle &= \langle g\mathcal{A}'(t_*)h^{-1}(e_\alpha), e_\beta\rangle = \langle \mathcal{A}'(t_*)h^{-1}(e_\alpha), g^\dagger(e_\beta)\rangle \\ &= \langle (\widehat{\mathcal{H}}(t_*) - \mathcal{E}_{\mathrm{CC}}(t_*))U_\alpha\Phi_0, V_\beta\Phi_0\rangle,\end{aligned}$$

where $u_\alpha = h^{-1}(e_\alpha)$ and $v_\beta = g^\dagger(e_\beta)$ and $\alpha, \beta \in \Xi(G)$. Here, $g^\dagger : \mathbb{R}^n \to \mathbb{V}$ is the adjoint of g. Therefore, using an appropriate basis transformation

$$i(\mathcal{A}, t_*) = \operatorname{sgn}\det \widehat{\mathcal{A}}'(h^{-1}(t_*)) = \operatorname{sgn}\prod_{j\geq 0}(\mathcal{E}_j(\widehat{\mathcal{H}}_{\mathfrak{V}}(t_*)) - \mathcal{E}_{\mathrm{CC}}(t_*)).$$

The proof is completed by noting that the elements of the matrix $\widehat{\mathcal{H}}_{\mathfrak{V}}(t_*)$ are real, so its complex eigenvalues come in conjugate pairs, hence only real eigenvalues contribute to the product above. ∎

PROPOSITION 5.14. *If $\mathcal{A}$ is locally strongly monotone near a zero t_*, then we have $i(\mathcal{A}, t_*) = 1$.*

Proof. We have that in particular $\widehat{\mathcal{A}} : \mathbb{R}^n \to \mathbb{R}^n$ is monotone near $h(t_*)$, so according to Theorem A.14, $\widehat{\mathcal{A}}$ is orientation-preserving near $h(t_*)$. But then Theorem A.13 (i) implies that $i(\mathcal{A}, t_*) = i(\widehat{\mathcal{A}}, h(t_*)) > 0$. ∎

Using the "unlinked" form, we can determine the topological index in the FCC case (see (2.10)). Recall that the eigenvalues $\mathcal{E}_n(\mathcal{H})$, $n = 0, 1, \ldots$, are assumed to be increasingly ordered.

THEOREM 5.15 (Index formula for FCC—non-degenerate case). *Let $\mathbb{V} = \mathbb{V}(G^{\mathrm{full}})$ and assume that $t_* \in \mathbb{V}$ is a zero of the FCC mapping $\mathcal{A} : \mathbb{V} \to \mathbb{V}^*$. Then $e^{T_*}\Phi_0 \in \mathfrak{H}^2$ is an (intermediately normalized) eigenfunction corresponding to some non-degenerate eigenvalue $\mathcal{E}_\nu(\mathcal{H})$ if and only if t_* is non-degenerate, and in this case $i(\mathcal{A}, t_*) = (-1)^\nu$.*

Proof. First, note that $\mathbb{V}(G^{\mathrm{full}})$ is rank-regular so $\widehat{\mathcal{H}}_{\mathfrak{V}}(t) = \mathcal{H}_{\mathfrak{V}}(t)$. We have $\mathcal{E}_{\mathrm{CC}}(t_*) = \mathcal{E}_\nu(\mathcal{H})$ by the equivalence of FCC and FCI (see Theorem 2.3).

According to Lemma 5.1, we have that $e^{T_*}\Phi_0$ is a non-degenerate intermediately normalized eigenfunction if and only if $\mathcal{E}_{\mathrm{CC}}(t_*) \notin \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$. In fact, $\mathcal{E}_{\mathrm{CC}}(t_*) \in \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$ if and only if there exists $R\Phi_0 \in \mathfrak{V}$ nonzero, such that

$$\langle e^{-T_*}\mathcal{H}e^{T_*}R\Phi_0, S\Phi_0\rangle = \mathcal{E}_{\mathrm{CC}}(t_*)\langle R\Phi_0, S\Phi_0\rangle,$$

for all $s \in \mathbb{V}$. Since $\mathbb{V}(G^{\mathrm{full}})$ is excitation complete, according to Lemma 5.3 the preceding equation is equivalent to

$$\langle \mathcal{H}Re^{T_*}\Phi_0, S\Phi_0\rangle = \mathcal{E}_{\mathrm{CC}}(t_*)\langle Re^{T_*}\Phi_0, S\Phi_0\rangle, \tag{5.17}$$

for all $s \in \mathbb{V}$. But this precisely means that the FCI eigenstate $e^{T_*}\Phi_0$ is degenerate, because $Re^{T_*}\Phi_0$ is another eigenvector corresponding to the same eigenvalue $\mathcal{E}_{\mathrm{CC}}(t_*)$.

We also conclude from (5.6) that $\sigma(\mathcal{H}_{\mathfrak{V}}(t_*)) = \sigma(\mathcal{H}) \setminus \{\mathcal{E}_{\mathrm{CC}}(t_*)\}$. Applying Theorem 5.13, we obtain that t_* is non-degenerate and $i(\mathcal{A}, t_*) = (-1)^\nu$. ∎

It is worth noting that, in the FCC case, the zero t_* representing the intermediately normalized, non-degenerate *ground state* (i.e. $\mathcal{E}_{\mathrm{CC}}(t_*) = \mathcal{E}_0(\mathcal{H})$) has $i(\mathcal{A}, t_*) = 1$. Note that this is not necessarily true in the truncated case. While the CC method is most commonly aimed at the ground state, it can also be used to find other intermediately normalized eigenfunctions as well. Furthermore, it can also be used to obtain eigenfunctions which are orthogonal to the reference Φ_0 according to the remark below.

Remark 5.16. The *Equation-of-Motion Coupled-Cluster* (EOM-CC) method [19] is aimed at calculating *excited* energies and states (i.e. $\mathcal{E}_n(\mathcal{H})$ for $n > 0$, and the corresponding eigenvectors) based on a CC ground-state solution. This is done in two steps. Let $\mathbb{V} = \mathbb{V}(G^{\mathrm{full}})$. Firstly, a conventional CC calculation determines the ground state $\Psi = e^{T_*}\Phi_0$ such that $\mathcal{A}(t_*) = 0$, i.e. $\mathcal{H}\Psi = \mathcal{E}\Psi$. Secondly, the targeted excited state is of the form $\Psi_{\mathrm{ex}} = (r_0 I + R)e^{T_*}\Phi_0 = \mathcal{R}e^{T_*}\Phi_0$, where R is a cluster operator, see Lemma 5.1. We have

$$\mathcal{H}_{\mathfrak{V}}(t_*)R\Phi_0 = \mathcal{E}_{\mathrm{ex}}R\Phi_0. \tag{5.18}$$

In other words, we need to solve the eigenproblem of the similarity-transformed projected Hamilton operator $\mathcal{H}_{\mathfrak{V}}(t_*)$. Furthermore, similarly to the proof of Lemma 5.3, it is easy to see that $\mathcal{A}(t_*) = 0$ implies

$$\langle R\mathcal{H}(t_*)\Phi_0, S\Phi_0\rangle = \mathcal{E}_{\mathrm{CC}}(t_*)\langle R\Phi_0, S\Phi_0\rangle \quad \text{for all} \quad s \in \mathbb{V}.$$

Subtracting this from (5.18), we get the "commutator form" of the EOM-CC equation:

$$\langle [\mathcal{H}(t_*), R]\Phi_0, S\Phi_0\rangle = \Delta\mathcal{E}\langle R\Phi_0, S\Phi_0\rangle, \quad \text{where} \quad \Delta\mathcal{E} = \mathcal{E}_{\mathrm{ex}} - \mathcal{E}, \tag{5.19}$$

and $\mathcal{E} = \mathcal{E}_{\mathrm{CC}}(t_*)$ is the ground-state energy as given by the CC method. Let $\mathbb{V}$ be an arbitrary amplitude space. Recalling the expression (5.12) for $\mathcal{A}'(t)$, we can rephrase the EOM-CC equation (5.19) as *the weak eigenvalue problem for* $\mathcal{A}'(t_*) : \mathbb{V} \to \mathbb{V}^*$ (cf. [21, Section 13.6.3]), i.e.

$$\langle \mathcal{A}'(t_*)R_j\Phi_0, S\Phi_0\rangle = \Delta\mathcal{E}_j\langle R_j\Phi_0, S\Phi_0\rangle, \tag{5.20}$$

for $j = 1, \dots, J$ labelling the states[11] and $s \in \mathbb{V}$ is arbitrary.[12] Notice that the $\Delta\mathcal{E}_j$'s are in general complex. Using Theorem 5.13 we can obtain the following. Suppose that $\Delta\mathcal{E}_1, \dots, \Delta\mathcal{E}_\mu$ are given by (5.20) and are all nonzero. Then

$$i(\mathcal{A}, t_*) = (-1)^\nu, \quad \nu = |\{j : \Delta\mathcal{E}_j \in \mathbb{R},\ \Delta\mathcal{E}_j < 0\}|. \tag{5.21}$$

Due to the *nonvariational property* of truncated CC (see subsection 2.4), it is not *a priori* clear whether the (real) excited energies are higher than the ground-state energy, i.e. if $\Delta\mathcal{E}_j > 0$. Therefore, (5.21) quantifies this nonvariational property through the topological index $i(\mathcal{A}, t_*)$.

Next, we draw a connection between the degeneracy of a zero t_* and the Fock-splitting (2.4) of the Hamilton operator. Define $\omega_0(t_*) = \langle \mathcal{W}(t_*)\Phi_0, \Phi_0\rangle$, which is the CC correction to the lowest eigenvalue Λ_0 of $\mathcal{F}$, so that the CC energy at t_* is obtained as $\mathcal{E}_{\mathrm{CC}}(t_*) = \Lambda_0 + \omega_0(t_*)$.

PROPOSITION 5.17. *Let $\mathbb{V}(G)$ be a rank-regular amplitude space and t_* a zero of $\mathcal{A}$. Define the linear operator $\mathcal{Q}(t_*) : \mathfrak{V} \to \mathfrak{V}$ via its matrix in the determinantal basis as*

$$[\mathcal{Q}(t_*)]_{\alpha\beta} = \varepsilon_\alpha \delta_{\alpha\beta} + \sum_{\gamma \in \Xi(G)} t_{*,\gamma} \varepsilon_\gamma \langle X_\gamma \Phi_\beta, \Phi_\alpha\rangle \quad \textit{for all} \quad \alpha, \beta \in \Xi(G).$$

Then $\omega_0(t_*) \notin \sigma(\mathcal{Q}(t_*) + \mathcal{W}_{\mathfrak{V}}(t_*))$ is equivalent to $\mathcal{E}_{\mathrm{CC}}(t_*) \notin \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$, i.e. to the fact that t_* is a non-degenerate zero of $\mathcal{A}$.

Proof. We have using (5.3),

$$\begin{aligned}
\mathcal{E}_{\mathrm{CC}}(t_*) &= \langle e^{-T_*} \mathcal{F} e^{T_*} \Phi_0, \Phi_0\rangle + \langle \mathcal{W}(t_*)\Phi_0, \Phi_0\rangle \\
&= \langle \mathcal{F}\Phi_0, \Phi_0\rangle + \langle [\mathcal{F}, T_*]\Phi_0, \Phi_0\rangle + \langle \mathcal{W}(t_*)\Phi_0, \Phi_0\rangle \\
&= \Lambda_0 + \langle \mathcal{W}(t_*)\Phi_0, \Phi_0\rangle = \Lambda_0 + \omega_0(t_*).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\langle \mathcal{H}(t_*)\Phi_\beta, \Phi_\alpha\rangle &= \langle \mathcal{F}\Phi_\beta, \Phi_\alpha\rangle + \langle [\mathcal{F}, T_*]\Phi_\beta, \Phi_\alpha\rangle + \langle \mathcal{W}(t_*)\Phi_\beta, \Phi_\alpha\rangle \\
&= (\Lambda_0 + \varepsilon_\alpha)\delta_{\alpha\beta} + \sum_{\gamma \in \Xi(G)} t_{*,\gamma} \varepsilon_\gamma \langle X_\gamma \Phi_\beta, \Phi_\alpha\rangle + \langle \mathcal{W}(t_*)\Phi_\beta, \Phi_\alpha\rangle.
\end{aligned}$$

Then, in the determinantal basis

$$\mathcal{H}_{\mathfrak{V}}(t_*) = \Lambda_0 I + \mathcal{Q}(t_*) + \mathcal{W}_{\mathfrak{V}}(t_*).$$

Hence, $\mathcal{E}_{\mathrm{CC}}(t_*) \notin \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$ is equivalent to $\omega_0(t_*) \notin \sigma(\mathcal{Q}(t_*) + \mathcal{W}_{\mathfrak{V}}(t_*))$, which finishes the proof. ∎

We now consider the case of a degenerate zero. Clearly, if $r \in \ker \mathcal{A}'(t_*) \neq \{0\}$, we have to consider higher-order terms of the Taylor polynomial of $\mathcal{A}$ at t_*,

$$\mathcal{A}(t_* + r) = \mathcal{A}'(t_*)r + \frac{1}{2}\mathcal{A}''(t_*)(r, r) + \mathcal{R}_3(t_*; r),$$

where $\mathcal{A}''(t_*) : \mathbb{V} \times \mathbb{V} \to \mathbb{V}^*$ is a bounded bilinear mapping. Here, we only consider the second-order information.

[11] Here, R_j is not to be confused with the rank-decomposition (3.6).
[12] A similar relation holds if t_* does not represent the ground state.

Assume from now on that $\mathbb{V}$ is rank-regular, so that $\widehat{\mathcal{H}}_{\mathfrak{V}}(t) = \mathcal{H}_{\mathfrak{V}}(t)$. Suppose that $\mathcal{E}_{\mathrm{CC}}(t_*) \in \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$ and that $R_1\Phi_0, \dots, R_\mu\Phi_0 \in \mathfrak{V}$ are the *right* eigenvectors of $\mathcal{H}_{\mathfrak{V}}(t_*)$ corresponding to $\mathcal{E}_{\mathrm{CC}}(t_*)$,

$$\langle \mathcal{H}(t_*)R_j\Phi_0, S\Phi_0\rangle = \mathcal{E}_{\mathrm{CC}}(t_*)\langle R_j\Phi_0, S\Phi_0\rangle \quad \text{for all } j = 1, \dots, \mu \text{ and all } s \in \mathbb{V}.$$

Also, suppose that $L_1\Phi_0, \dots, L_\mu\Phi_0 \in \mathfrak{V}$ are the *left* eigenvectors of $\mathcal{H}_{\mathfrak{V}}(t_*)$ corresponding to $\mathcal{E}_{\mathrm{CC}}(t_*)$,

$$\langle \mathcal{H}(t_*)^\dagger L_j\Phi_0, S\Phi_0\rangle = \mathcal{E}_{\mathrm{CC}}(t_*)\langle L_j\Phi_0, S\Phi_0\rangle \quad \text{for all } j = 1, \dots, \mu \text{ and all } s \in \mathbb{V}.$$

The corresponding right-, and left eigenspaces are

$$W_R = \ker \mathcal{A}'(t_*) = \mathrm{Span}\{r_1, \dots, r_\mu\},$$
$$W_L = \ker \mathcal{A}'(t_*)^\dagger = \mathrm{Span}\{\ell_1, \dots, \ell_\mu\},$$

and let $Q : \mathbb{V} \to \mathbb{V}$ be the orthogonal projector onto W_L. Further, define $\widehat{Q} : \mathfrak{V} \to \mathfrak{V}$ via $\langle \widehat{Q}U\Phi_0, V\Phi_0\rangle = \langle Qu, v\rangle$ for all $u, v \in \mathbb{V}$. We introduce the mapping $\mathcal{B} : \mathbb{V} \to \mathbb{V}^*$ via

$$\langle \mathcal{B}(t), s\rangle = \frac{1}{2}\langle \widehat{Q}[[\mathcal{H}(t_*), T], T]\Phi_0, S\Phi_0\rangle = \frac{1}{2}\langle \widehat{Q}[[\mathcal{W}(t_*), T], T]\Phi_0, S\Phi_0\rangle, \tag{5.22}$$

that is, the Q-projection of $\frac{1}{2}\mathcal{A}''(t_*)(t,t)$. In the second equality, we used (5.4). Also, note that $\mathcal{B}$ is homogeneous of degree 2, i.e. $\mathcal{B}(\alpha t) = \alpha^2\mathcal{B}(t)$. The next theorem follows esentially from Leray's second reduction formula (Theorem A.10).

THEOREM 5.18 (Index formula for SRCC—degenerate case). *Let t_* be zero of the CC mapping $\mathcal{A} : \mathbb{V} \to \mathbb{V}^*$. Suppose that $\mathcal{E}_{\mathrm{CC}}(t_*) \in \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$ and let W_R, W_L and Q be as above. Assume that*

$$\mathcal{B}(t) \neq 0 \quad \textit{for all} \quad t \in \partial B(0,1). \tag{5.23}$$

Then, t_* is an isolated zero and the topological index of $\mathcal{A}$ at t_* is given by

$$i(\mathcal{A}, t_*) = i(\mathcal{A}'(t_*) + Q, 0)\, i(\mathcal{B}|_{W_R}, 0).$$

Proof. First, we prove that t_* is isolated. When $r \notin \ker \mathcal{A}'(t_*)$ and small, it follows that $\mathcal{A}(t_* + r) \neq 0$ similarly to (5.16). If, however $r \in \ker \mathcal{A}'(t_*)$, then we may write

$$\langle \mathcal{A}(t_*+r), \mathcal{J}(\mathcal{A}''(t_*)(r,r))\rangle = \langle \underbrace{\mathcal{A}'(t_*)r}_{0}, \mathcal{J}(\mathcal{A}''(t_*)(r,r))\rangle + \frac{1}{2}\|\mathcal{A}''(t_*)(r,r)\|_{\mathbb{V}}^2 + \mathcal{O}(\|r\|_{\mathbb{V}}^5)$$

for all $r \in B(0,\varepsilon)$ for sufficiently small $\varepsilon > 0$. Condition (5.23) implies that $\mathcal{A}(t_*+r) \neq 0$ for all $r \in B^*(0,\varepsilon)$.

Next, we apply Theorem A.11 with the choice $D = B(0,\varepsilon)$, $L = \mathcal{A}'(t_*)$ and

$$\langle \mathcal{N}(t,\lambda), s\rangle = \sum_{k=2}^{2N} \frac{\lambda^{k-2}}{k!}\langle [\mathcal{H}(t_*), T]_{(k)}\Phi_0, S\Phi_0\rangle,$$

where we used (5.9). Because $\operatorname{ran} Q = \ker \mathcal{A}'(t_*)^\dagger$, it follows that

$$\ker Q = (\operatorname{ran} Q)^\perp = (\ker \mathcal{A}'(t_*)^\dagger)^\perp = \operatorname{ran} \mathcal{A}'(t_*) = \operatorname{ran} L.$$

Moreover, since t_* is an isolated zero, it is possible to choose $\delta > 0$ so that the equation

$$\lambda^{-1}\mathcal{A}(t_* + \lambda t) = \mathcal{A}'(t_*)t + \lambda\mathcal{N}(t,\lambda) = 0$$

does not admit a solution $t \in \partial B(0,\delta)$ for any $\lambda \in (0,1]$.

Note that $Q\mathcal{N}(t,0) = \mathcal{B}(t) \neq 0$ for all $t \in B^*(0,\delta)$ by assumption (5.23) and the homogenity of $\mathcal{B}$. We see that conditions (i) and (ii) of Theorem A.11 are satisfied and the result follows. ∎

The preceding theorem reduces the computation of the index to a low-dimensional problem but the zero is still degenerate. In fact, since

$$\langle \mathcal{B}'(t)u, v\rangle = \frac{1}{2}\langle \widehat{Q}([[\mathcal{H}(t_*), U], T] + [[\mathcal{H}(t_*), T], U])\Phi_0, V\Phi_0\rangle,$$

we have that $t = 0$ is a degenerate zero of $\mathcal{B}|_{W_R}$ and by assumption the only zero. Therefore, we need to apply Theorem A.9 to determine $i(\mathcal{B}|_{W_R}, 0)$.

COROLLARY 5.19. *For a degenerate, isolated zero t_* of $\mathcal{A}$, $\dim W_R = 1$, and for which (5.23) holds, we have $i(\mathcal{A}, t_*) = 0$.*

Proof. Let $W_R = \operatorname{Span}\{r\}$ and $W_L = \operatorname{Span}\{\ell\}$. We apply Theorem A.9 to the mapping $\mathcal{B}|_{W_R}$ with $D = B(0,\varepsilon)$, $\varepsilon > 0$ arbitrary. Fix $z' = \eta\ell \in \operatorname{ran} Q = W_L$ such that $0 < |\eta| < \delta$. Since $t = cr \in W_R$, the equation $\mathcal{B}(t) = z'$ in $B(0,\varepsilon) \cap W_R$ is equivalent to finding $0 \neq c \in \mathbb{R}$ such that

$$\frac{c^2}{2}\langle(\mathcal{H}(t_*) - \mathcal{E}_{\mathrm{CC}}(t_*))R^2\Phi_0, L\Phi_0\rangle = \eta\langle L\Phi_0, L\Phi_0\rangle. \tag{5.24}$$

Note that the inner product on the left-hand side is nonzero by assumption (5.23). Choose η to be of opposite sign as the inner product on the left-hand side. Then there are no solutions c, so $i(\mathcal{B}|_{W_R}, t_*) = 0$ and therefore $i(\mathcal{A}, t_*) = 0$ by Theorem 5.18. ∎

COROLLARY 5.20. *Let t_* be an isolated zero of the CC mapping $\mathcal{A} : \mathbb{V} \to \mathbb{V}^*$. Suppose that $\mathcal{E}_{\mathrm{CC}}(t_*) \in \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$ and let W_R and Q as above. Assume that $\ker \mathcal{B}'(t) = \{0\}$ for all $0 \neq t \in W_R$. If $\dim W_R = \mu$ is odd, then $i(\mathcal{A}, t_*) = 0$.*

Proof. Let $z' \in B^*(0,\delta) \cap W_L$. Then the equation $\mathcal{B}(t) = z'$ for $0 \neq t \in W_R$ is equivalent to

$$\frac{1}{2}\langle[[\mathcal{H}(t_*), T], T]\Phi_0, L\Phi_0\rangle = \langle Z'\Phi_0, L\Phi_0\rangle,$$

for all $\ell \in W_L$. Let $\mathcal{T}$ denote the set of solutions t of the preceding equation (which can be empty). By Theorem A.9, $|\mathcal{T}| = m$ for some m finite. Notice that $\mathcal{T}$ is closed under the operation $t \mapsto -t$, so m is even (we also used that $0 \notin \mathcal{T}$), and let

$$\mathcal{T} = \{t_1, \dots, t_{\frac{m}{2}}, -t_1, \dots, -t_{\frac{m}{2}}\}$$

using some appropriate indexing. From the linearity of $t \mapsto \mathcal{B}'(t)$, we get via Proposition A.3,

$$\begin{aligned}\deg(\mathcal{B}|_{W_R}, B(0,r) \cap W_R, z') &= \sum_{i=1}^{\frac{m}{2}} \operatorname{sgn}\det g\mathcal{B}'(t_i)h^{-1} + \operatorname{sgn}\det g\mathcal{B}'(-t_i)h^{-1} \\ &= \sum_{i=1}^{\frac{m}{2}} (1 + (-1)^{\mu})\operatorname{sgn}\det g\mathcal{B}'(t_i)h^{-1} = 0,\end{aligned} \tag{5.25}$$

for some sufficiently large $r > 0$. ∎

We close this section with two remarks.

Remark 5.21.

(i) Note that, according to Theorem A.9, a zero t_* of topological index 0 is "numerically unstable", because one could miss zeros altogether if the equations are solved with finite precision arithmetic, even in the FCC case. Therefore, the degenerate zeros of the SRCC mapping $\mathcal{A}$ are not robust in general. We have already seen that in the FCC case, the CC energy $\mathcal{E}_{\mathrm{CC}}(t_*)$ of a degenerate zero t_* is a degenerate eigenvalue $\mathcal{E}$ of the Hamilton operator (see Remark 5.2). Thus, we can conclude that the SRCC method is in general unsuitable for finding degenerate eigenstates and eigenvalues—an empirical fact that is well known among the practitioners of the SRCC method.

(ii) Proposition 5.14 and the preceding calculations imply that any approach that stipulates the local strong monotonicity of $\mathcal{A}$ near t_* can only provide an incomplete description of the SRCC method.

5.3. Local properties—complex case. We now discuss what happens when *complex* amplitude spaces are considered instead. We explain the complex case in detail because the differences from the real case are somewhat subtle. We will use the concepts and results of Appendix B.

Assume that $\mathbb{V}$ is a complex amplitude space and let $\mathbb{V}^*$ denote its anti-dual. It is clear that $t \mapsto \langle \mathcal{A}(t), s\rangle$ is a (complex) polynomial for fixed $s \in \mathbb{V}$, hence with the appropriate identifications, $\mathcal{A}_{\mathbb{C}} : \mathbb{V} \to \mathbb{V}^*$ is a holomorphic mapping, where we used the subscript $\mathbb{C}$ to highlight the difference.[13] Of course, a real zero $t_* \in \mathbb{V}$ to $\mathcal{A}(t_*) = 0$ is automatically a "complex" zero: $\mathcal{A}_{\mathbb{C}}(t_*) = 0$. Further, using the fact that the Hamilton operator is real (by which we mean $\langle \mathcal{H}\Phi_\alpha, \Phi_\beta\rangle \in \mathbb{R}$), $\mathcal{A}_{\mathbb{C}}(t_*) = 0$ if and only if $\mathcal{A}_{\mathbb{C}}(\bar{t}_*) = 0$. Also, $\mathcal{E}_{\mathrm{CC}}(\bar{t}) = \overline{\mathcal{E}_{\mathrm{CC}}(t)}$.

From Theorem B.3 we immediately get that $\deg(\mathcal{A}_{\mathbb{C}}, U, 0) \geq 0$ for every bounded open $U \subset \mathbb{V}$. In particular, $i(\mathcal{A}_{\mathbb{C}}, t_*) \geq 0$ for every isolated zero t_*. Notice that, even if t_* is a zero of both $\mathcal{A}$ and $\mathcal{A}_{\mathbb{C}}$, its real and complex indices $i(\mathcal{A}, t_*)$ and $i(\mathcal{A}_{\mathbb{C}}, t_*)$ may differ; for instance we know from Theorem 5.13 that $i(\mathcal{A}, t_*)$ can have a sign, while $i(\mathcal{A}_{\mathbb{C}}, t_*)$ cannot. Also, Theorem B.3 (iv) implies that $i(\mathcal{A}, t_*) \geq 2$ for an isolated, degenerate zero t_*. Moreover, the following is true.

THEOREM 5.22. *If $t_* \in \mathbb{V}$ is a* real *isolated zero of both $\mathcal{A}$ and $\mathcal{A}_{\mathbb{C}}$, then*

$$|i(\mathcal{A}, t_*)| \leq i(\mathcal{A}_{\mathbb{C}}, t_*), \quad i(\mathcal{A}, t_*) \equiv i(\mathcal{A}_{\mathbb{C}}, t_*) \mod 2.$$

Proof. The result follows from Theorem B.5 with the choice $D = B(t_*, \delta)$. ∎

For simplicity, we assume from now on that $\mathbb{V}$ is rank-regular (Definition 5.4), so that $\widehat{\mathcal{H}}_{\mathfrak{V}}(t) = \mathcal{H}_{\mathfrak{V}}(t)$. Adapting the proofs of Theorem 5.13 and Theorem 5.18 in the complex case, we have

THEOREM 5.23 (Index formula for SRCC—complex case). *Let t_* be an isolated zero of $\mathcal{A}_{\mathbb{C}} : \mathbb{V} \to \mathbb{V}^*$. Then the following hold true.*

(i) *The zero t_* is non-degenerate if and only if $\mathcal{E}_{\mathrm{CC}}(t_*) \notin \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$, and in this case $i(\mathcal{A}_{\mathbb{C}}, t_*) = 1$.*

(ii) *Suppose that $\mathcal{E}_{\mathrm{CC}}(t_*) \in \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$ and that the second-order regularity assumption (5.23) holds true. Then*

$$i(\mathcal{A}_{\mathbb{C}}, t_*) = i(\mathcal{B}|_{W_R}, 0).$$

[13] Note that the realification of $\mathcal{A}_{\mathbb{C}}$, $(\mathcal{A}_{\mathbb{C}})_{\mathbb{R}}$ does *not* equal $\mathcal{A}$ (they are mappings of different type: $(\mathcal{A}_{\mathbb{C}})_{\mathbb{R}} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$).

(iii) *Suppose that* $\mathcal{E}_{\mathrm{CC}}(t_*) \in \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$ *and that the second-order regularity assumption* (5.23) *holds true. Assume further, that* $\ker \mathcal{B}'(t) = \{0\}$ *for all* $0 \neq T\Phi_0 \in W_R$*. Then* $i(\mathcal{A}_{\mathbb{C}}, t_*) = m \geq 2$*, where* m *is the number of solutions* $0 \neq R\Phi_0 \in W_R$ *to the equation*

$$\langle (\mathcal{H}(t_*) - \mathcal{E}_{\mathrm{CC}}(t_*))R^2\Phi_0, L\Phi_0\rangle = \langle Z\Phi_0, L\Phi_0\rangle \quad (L\Phi_0 \in W_L)$$

for any $Z\Phi_0 \in W_L \cap B^*(0,\delta)$ *for sufficiently small* $\delta > 0$.

Proof. For (i), it is enough to note that $\ker \mathcal{A}'_{\mathbb{C}}(t_*) \neq \{0\}$ follows via the same calculation as in the proof of Theorem 5.13. Part (ii) follows since $i(\mathcal{A}'_{\mathbb{C}}(t_*) + Q) = 1$. For the proof of (iii), notice that (5.25) now reads

$$\deg(\mathcal{B}|_{W_R}, B(0,r) \cap W_R, z) = \sum_{i=1}^{m} \operatorname{sgn} |\det g\mathcal{B}'(t_i)h^{-1}|^2 = m. \qquad \square$$

Corollary 5.24. *For a degenerate, isolated zero* t_* *of* $\mathcal{A}_{\mathbb{C}}$*,* $\dim W_R = 1$*, and for which* (5.23) *holds, we have* $i(\mathcal{A}_{\mathbb{C}}, t_*) = 2$.

Proof. The proof is analogous to that of Corollary 5.19, but (5.24) always has exactly two nonzero complex solutions c. $\square$

We close this section with a few remarks.

Remark 5.25.

(i) Luckily, for a real zero $t_* \in \mathbb{V}$, the condition $\mathcal{E}_{\mathrm{CC}}(t_*) \notin \sigma(\mathcal{H}_{\mathfrak{V}}(t_*))$ is formally the same as in the real case, therefore a non-degenerate real zero is automatically a non-degenerate zero of $\mathcal{A}_{\mathbb{C}}$.

(ii) The degeneracy of a complex zero t_* manifests itself in numerical computations as follows. Suppose the hypotheses of Theorem 5.23 (iii) hold true. Combining this with Theorem B.4, we get that the perturbed equation $\mathcal{A}(t) = z'$ has *exactly* m solutions for almost all $z' \in \mathbb{V}$ sufficiently close to zero. This is in contrast with the real case, when one might completely "lose" solutions when the index is zero (Remark 5.21 (i)). The appearance of multiple complex zeros in degenerate situations was conjectured based on numerical observations in [44, 43].

(iii) If the Hamilton operator is real (see above), then the index of the complex conjugate zero is the same: $i(\mathcal{A}_{\mathbb{C}}, t_*) = i(\mathcal{A}_{\mathbb{C}}, \bar{t}_*)$.

(iv) The classical *Bézout theorem* states that if a polynomial system

$$\left.\begin{aligned} P_1(x_1, \ldots, x_d) &= 0 \\ P_2(x_1, \ldots, x_d) &= 0 \\ &\cdots \\ P_n(x_1, \ldots, x_d) &= 0 \end{aligned}\right\}$$

has a finite number of zeros in $\mathbb{C}^d$, then the number of zeros (counting multiplicities) is at most $\Delta = \Delta_1 \cdots \Delta_d$ (called the *Bézout number*), where Δ_k denotes the degree of P_k. As remarked earlier, according to the Baker–Campbell–Hausdorff expansion (2.12), for the polynomials constituting the system $\mathcal{A}(t) = 0$ there holds $\Delta_1 = \ldots = \Delta_d = 4$, hence the Bézout number of the CC equations is $\Delta = 4^d$, where $d = \dim \mathbb{V}$. This is typically a huge number. However, it is known that the Bézout number often grossly

overestimates the number of zeros. In fact, it was observed numerically that the number of zeros for the (truncated) CC equations is much less than the Bézout number [43].

5.4. Continuation of solutions. In this section we discuss how solutions of different CC methods can be "connected" in a systematic way. The idea is not new to this field (and certainly not new to nonlinear analysis, see e.g. [59]), and it has been a subject of both theoretical and numerical investigations in the CC literature, as we have already mentioned in the beginning of section 5.

The main theoretical tool we use to describe the aforementioned connection is a specific type of homotopy.

DEFINITION 5.26. *Let $\mathbb{V}^1$ be an amplitude space with direct sum decomposition $\mathbb{V}^1 = \mathbb{V}^0 \oplus \mathbb{V}^\angle$. Let $\mathcal{A}^j : \mathbb{V}^j \to (\mathbb{V}^j)^*$ be continuous mappings for $j = 0, 1$. A continuous map $\mathcal{K} : \mathbb{V}^1 \times [0,1] \to (\mathbb{V}^1)^*$ is said to be an* admissible homotopy*, if*

(i) $\langle \mathcal{K}(t^1, 0), s^0 \rangle = \langle \mathcal{A}^0(t^0), s^0 \rangle$ *for all* $t^1 \in \mathbb{V}^1$, $s^0 \in \mathbb{V}^0$, *and*

(ii) $\mathcal{K}(\cdot, 1) = \mathcal{A}^1$.

Furthermore, an admissible homotopy $\mathcal{K} : \mathbb{V}^1 \times [0,1] \to (\mathbb{V}^1)^*$ is said to be faithful*, if for every $t^0_{**} \in \mathbb{V}^0$ such that $\mathcal{A}^0(t^0_{**}) = 0$, there exists $t^\angle_{**} \in \mathbb{V}^\angle$ so that with $t^1_{**} = t^0_{**} + t^\angle_{**} \in \mathbb{V}^1$, there holds $\mathcal{K}(t^1_{**}, 0) = 0$.*

Let $\mathbb{V}^1 = \mathbb{V}(G^{\text{full}})$ and $\mathcal{A}^1$ be the FCC mapping, $\mathbb{V}^0$ some rank-truncated space and $\mathcal{A}^0$ the truncated CC mapping. This case is particularly important due to the equivalence of FCC and FCI (Theorem 4.4), so existence of a solution to the FCI problem (essentially the Schrödinger equation) can be exploited to infer the existence of a *truncated* CC solution. Furthermore, the topological index of the CC solution can be determined by the results of subsection 5.2 and the homotopy invariance of the topological degree can be used relate these quantities in certain situations (see (5.28) below).

The *zero set* of an (admissible) homotopy $\mathcal{K}$ is defined as

$$\mathcal{Z}(\mathcal{K}) = \{(t^1_*, \lambda) \in \mathbb{V}^1 \times [0,1] : \mathcal{K}(t^1_*, \lambda) = 0\}. \tag{5.26}$$

We omit $\mathcal{K}$ from the notation $\mathcal{Z}(\mathcal{K})$ whenever it is clear from the context. The λ-sections of $\mathcal{Z}$ are denoted as $\mathcal{Z}_\lambda = \{t^1_* \in \mathbb{V}^1 : (t^1_*, \lambda) \in \mathcal{Z}\}$. Clearly, $(\mathcal{Z}(\mathcal{K}))_1 = (\mathcal{A}^1)^{-1}(0)$ for *any* admissible homotopy $\mathcal{K}$. Furthermore, $\Pi_{\mathbb{V}^0}(\mathcal{Z}(\mathcal{K}))_0 = (\mathcal{A}^0)^{-1}(0)$ for any faithful, admissible homotopy $\mathcal{K}$. We will sometimes explicitly label t^1_* with the λ to which it corresponds as $t^1_*(\lambda)$.

Recall that any topological space can be partitioned into a family of *connected components*, which are maximal (w.r.t. set inclusion), closed connected sets. We are interested in the connected components of $\mathcal{Z}(\mathcal{K})$. The extended homotopy invariance property of the degree (Theorem A.7) and the Leray–Schauder continuation principle (Theorem A.8) immediately implies the following.

THEOREM 5.27. *Suppose that $\mathcal{D} \subset \mathbb{V}^1 \times [0,1]$ is a bounded open set and let $\mathcal{K} : \mathbb{V} \times [0,1] \to \mathbb{V}^*$ be an admissible homotopy such that*

$$\mathcal{K}(t^1, \lambda) \neq 0 \quad \textit{for all} \quad (t^1, \lambda) \in \partial\mathcal{D}. \tag{5.27}$$

Then the following statements hold true.

(i) $\deg(\mathcal{K}(\cdot, 0), \mathcal{D}_0, 0) = \deg(\mathcal{A}^1, \mathcal{D}_1, 0) =: d$.

(ii) *If $\Delta \neq 0$, then there is a connected component $\mathcal{C}$ of $\mathcal{Z}(\mathcal{K})$, such that*

$$\mathcal{C} \cap (\mathcal{Z}(\mathcal{K}))_j \neq \emptyset \quad \textit{for} \quad j = 0, 1.$$

This general theorem can be used to prove existence results, which essentially consists of establishing the boundary condition (5.27). Note that (i) implies that

$$\sum_{t^1_{**}\in\mathcal{Z}_0\cap\mathcal{D}_0} i(\mathcal{K}(\cdot,0),t^1_{**}) = \sum_{t^1_{*}\in\mathcal{Z}_1\cap\mathcal{D}_1} i(\mathcal{A}^1,t^1_{*}) = d. \tag{5.28}$$

This is a necessary condition for zeros in $\mathcal{Z}_0$ and $\mathcal{Z}_1$ be in the same bounded connected component.

Remark 5.28. In [43], a specific type of admissible homotopy (which will be discussed in subsection 5.5 below) is used as the basis to distinguish "physical" truncated solutions from "unphysical" ones. Roughly speaking, they call a truncated solution "physical" if it shares a *bounded* connected component with an FCC solution (although they require more regularity of the connected component in question). They also found that some FCC (resp. truncated) solutions cannot be "connected" to any truncated (resp. FCC) solution. While this approach to the classification of solutions seems attractive at first, it has a serious conceptual drawback: it is unclear why one particular homotopy is preferred over the (infinitely many) others. For instance, it is conceivable that an admissible homotopy classifies a truncated solution as "unphysical" while another homotopy classifies it as "physical". Moreover, their homotopy itself does not admit an obvious physical interpretation. We will not pursue this line of thought any further in this work, and view homotopies as purely mathematical devices.

The most obvious example of an admissible homotopy is the *linear homotopy*. Let $\mathbb{V}^0 \subset \mathbb{V}^1$ be any subspace (a typical choice is based on some truncation: $\mathbb{V}^0 = \mathbb{V}(G(1,\dots,\rho)))$ and $\mathbb{V}^{\angle} := (\mathbb{V}^0)^{\perp}$, where the orthogonal complement is taken with respect to the $\mathbb{V}$-inner product. Also, to emphasize that $\mathbb{V}^{\angle}$ is actually given by the $\mathbb{V}$-orthogonal complement of $\mathbb{V}^0$, we shall write $t^1 = t^0 + t^{\perp}$ for some unique $t^0 \in \mathbb{V}^0$ and $t^{\perp} \in (\mathbb{V}^0)^{\perp}$, for any $t^1 \in \mathbb{V}^1$. Let $\mathcal{K}_{\mathrm{L}} : \mathbb{V}^1 \times [0,1] \to (\mathbb{V}^1)^*$ by given by

$$\langle \mathcal{K}_{\mathrm{L}}(t^1,\lambda), s^1\rangle = (1-\lambda)\left(\langle \mathcal{A}^0(t^0), s^0\rangle + \alpha|\langle t^{\perp} - u^{\perp}, s^{\perp}\rangle_{\mathbb{V}}|\right) + \lambda\langle \mathcal{A}^1(t^1), s^1\rangle$$

for any $t^1, s^1 \in \mathbb{V}^1$ and $\lambda \in [0,1]$, and some fixed constants $\alpha > 0$ and $u^{\perp} \in (\mathbb{V}^0)^{\perp}$. Then, clearly $\mathcal{K}_{\mathrm{L}}(\cdot,1) = \mathcal{A}^1$. Further, $\mathcal{K}_{\mathrm{L}}(t^1_{**},0) = 0$ is equivalent to $\mathcal{A}^0(t^0_{**}) = 0$ and $t^{\perp}_{**} = u^{\perp}$. Therefore, $\mathcal{K}_{\mathrm{L}}$ is a faithful, admissible homotopy. Also, $\mathcal{K}_{\mathrm{L}}$ has the following trivial property: if $t^1_* \in \mathbb{V}^1$ is such that $\mathcal{A}^0(t^0_*) = 0$ and $\mathcal{A}^1(t^1_*) = 0$, then $\mathcal{K}_{\mathrm{L}}(t^1_*,\lambda) = 0$ for all $\lambda \in [0,1]$.

As a simple application of the linear homotopy, we give a variant of the existence result [50, Theorem 4.1].

THEOREM 5.29. *Suppose that $t^1_* \in \mathbb{V}^1$ is a zero of $\mathcal{A}^1$. Let $\varkappa = \|t^{\perp}_*\|_{\mathbb{V}}$. Assume the following hold true.*

(i) *$\mathcal{A}^j : \mathbb{V}^j \to \mathbb{V}^j$ is C^1.*
(ii) *$L^0\varkappa = \sup_{\|u^0\|_{\mathbb{V}}=1}\langle \mathcal{A}^0(t^0_*) - \mathcal{A}^1(t^1_*), u^0\rangle$ for some $L^0 < \infty$.*
(iii) *$\mathcal{A}^j$ is strongly monotone in $B_{\mathbb{V}^j}(t^j_*,\delta_j)$ with constant $C^j_{\mathrm{SM}} = C^j_{\mathrm{SM}}(\delta_j) > 0$ for $j = 0, 1$.*
(iv) *$\varkappa < \delta\frac{C^0_{\mathrm{SM}}}{L^0}$, where $\delta = \min\{\delta_0,\delta_1\}$*

Then there exists a unique $t^0_{**} \in B_{\mathbb{V}^0}(t^0_*,\delta)$ such that $\mathcal{A}^0(t^0_{**}) = 0$. Furthermore, $i(\mathcal{K}_{\mathrm{L}}(\cdot,0),t^1_{**}) = i(\mathcal{A}^0,t^0_*) = 1$.

Proof. Set $\alpha := C^0_{\mathrm{SM}}$ and $u^{\perp} := t^{\perp}_*$ in the definition of $\mathcal{K}_{\mathrm{L}}$. We prove that the boundary condition $\mathcal{K}_{\mathrm{L}}(t^1,\lambda) \neq 0$ holds true for all $t^1 \in \partial B(t^1_*,\delta)$ and $\lambda \in [0,1]$.

Write $t^1 = t^1_* + r^1$, where $\|r^1\|_{\mathbb{V}} = \delta$ and

$$\langle \mathcal{K}_{\mathrm{L}}(t^1_* + r^1, \lambda), r^1\rangle =: (1-\lambda)A_0 + \lambda A_1.$$

Here,

$$\begin{aligned} A_0 &= \langle \mathcal{A}^0(t^0_* + r^0), r^0\rangle + C^0_{\mathrm{SM}}\langle t^\perp_* + r^\perp - t^\perp_*, r^\perp\rangle_{\mathbb{V}} \\ &= \langle \mathcal{A}^0(t^0_* + r^0) - \mathcal{A}^0(t^0_*), r^0\rangle + \langle \mathcal{A}^0(t^0_*) - \mathcal{A}^1(t^1_*), r^0\rangle + C^0_{\mathrm{SM}}\|r^\perp\|^2_{\mathbb{V}} \\ &\geq C^0_{\mathrm{SM}}\|r^0\|^2_{\mathbb{V}} - L^0\varkappa\|r^0\|_{\mathbb{V}} + C^0_{\mathrm{SM}}\|r^\perp\|^2_{\mathbb{V}} \\ &\geq (C^0_{\mathrm{SM}}\|r^1\|_{\mathbb{V}} - L^0\varkappa)\|r^1\|_{\mathbb{V}} = (C^0_{\mathrm{SM}}\delta - L^0\varkappa)\delta > 0, \end{aligned}$$

where we in the last step used that $\|r^1\|^2_{\mathbb{V}} = \|r^0\|^2_{\mathbb{V}} + \|r^\perp\|^2_{\mathbb{V}}$. Furthermore,

$$A_1 = \langle \mathcal{A}^1(t^1_* + r^1), r^1\rangle = \langle \mathcal{A}^1(t^1_* + r^1) - \mathcal{A}^1(t^1_*), r^1\rangle \geq C^1_{\mathrm{SM}}\|r^1\|^2_{\mathbb{V}} > 0.$$

Using Proposition 5.14 and the uniqueness of the zero t^1_* in $B(t^1_*, \delta)$,

$$\deg(\mathcal{K}_{\mathrm{L}}(\cdot, 1), B(t^1_*, \delta), 0) = \deg(\mathcal{A}^1, B(t^1_*, \delta), 0) = 1.$$

Since we have already seen that $\mathcal{K}_{\mathrm{L}}(t^1, \lambda) \neq 0$ for all $t^1 \in \partial B(t^1_*, \delta)$ and $\lambda \in [0, 1]$, the homotopy invariance of the degree (Theorem A.1 (iii)) can be applied to get

$$\deg(\mathcal{K}_{\mathrm{L}}(\cdot, 0), B(t^1_*, \delta), 0) = 1.$$

Corollary A.2 (i) implies that there exists $t^1_{**} \in \mathbb{V}^0$ such that $(t^0_{**}, t^\perp_*) \in B_{\mathbb{V}^1}(t^1_*, \delta)$ and $\langle \mathcal{K}_{\mathrm{L}}(t^0_{**}, 0), s^0\rangle = \langle \mathcal{A}^0(t^0_{**}), s^0\rangle = 0$ for all $s^0 \in \mathbb{V}^0$, which is what we wanted to prove. Uniqueness follows form the local strong monotonicity of $\mathcal{A}^0$. ∎

In the special case when $\mathcal{A}^0$ is given as a projection of $\mathcal{A}^1$ (see Remark 5.8), we obtain the following.

Corollary 5.30. *Suppose that $t^1_* \in \mathbb{V}^1$ is a zero of $\mathcal{A}^1$. Let $\varkappa = \|t^\perp_*\|_{\mathbb{V}}$ and suppose the following hold true.*

(i) $L^0\varkappa = \sup_{\|u^0\|_{\mathbb{V}}=1}\langle \mathcal{A}^1(t^0_*) - \mathcal{A}^1(t^1_*), u^0\rangle$ *for some* $L^0 < \infty$.
(ii) $\mathcal{A}^1$ *is strongly monotone in* $B_{\mathbb{V}^1}(t^1_*, \delta)$ *with constant* $C_{\mathrm{SM}} > 0$.
(iii) $\varkappa < \delta\frac{C_{\mathrm{SM}}}{C_{\mathrm{SM}}+L^0}$.

Then there exists a unique $t^0_{**} \in B_{\mathbb{V}^0}(t^1_*, \delta - \varkappa)$ such that $\mathcal{A}^1(t^0_{**}) = 0$.

Proof. It is enough to recall that $\mathcal{A}^0 = \Pi_{\mathbb{V}^0}\mathcal{A}^1|_{\mathbb{V}^0}$ is strongly monotone with constant C_{SM} in

$$B_{\mathbb{V}^0}(t^0_*, \sqrt{\delta^2 - \varkappa^2}) \supset B_{\mathbb{V}^0}(t^0_*, \delta - \varkappa).$$

Applying Theorem 5.29 with $C^0_{\mathrm{SM}} = C_{\mathrm{SM}}$ and $\delta_0 = \delta - \varkappa$ gives the result. ∎

5.5. Kowalski–Piecuch homotopy. In this section, we consider another homotopy that can be used to analyze the connection between the solutions to CC methods of different rank-truncation levels. The approach was pioneered by the chemists K. Kowalski and P. Piecuch [43], who conducted a comprehensive numerical study based on this idea.[14] They considered complex amplitudes and the following discussion is easily extended to that case.

Assumption. Let $\mathbb{V}^1 = \mathbb{V}^0 \oplus \mathbb{V}^\angle$ be an ℓ^2-*orthogonal* direct sum decomposition of a real amplitude space $\mathbb{V}^1 = \mathbb{V}(G^1)$, where $\mathbb{V}^0$ is an amplitude space corresponding

[14] They call it the "β–nested equations", β being the homotopy parameter.

to some lower truncation level. More precisely, assume that there is a rank $\rho \geq 1$, such that $\mathbb{V}^0$ contains all amplitudes with rank $\leq \rho$ and $\mathbb{V}^\angle := (\mathbb{V}^0)^{\perp_{\ell^2}}$ contains all amplitudes with rank $> \rho$. In the notations introduced before,

$$\mathbb{V}^0 = \mathbb{V}(G^0) = \mathbb{V}(G(1,\dots,\rho)) \cap \mathbb{V}^1, \text{ and } \mathbb{V}^\angle = \mathbb{V}(G^\perp) = \mathbb{V}(G(\rho+1,\dots,N)) \cap \mathbb{V}^1,$$

where $G^0 = G(1,\dots,\rho) \cap G^1$ and $G^\perp = G(\rho+1,\dots,N) \cap G^1$. Hence, any $t^1 \in \mathbb{V}^1$ may be uniquely decomposed as $t^1 = t^0 + t^\angle$, where $t^0 \in \mathbb{V}^0$, $t^\angle \in \mathbb{V}^\angle$ and $\langle t^0, t^\angle \rangle = 0$.

Now suppose that $\mathcal{A} : \mathbb{V}^1 \to (\mathbb{V}^1)^*$ is the SRCC mapping (5.1). Write

$$\begin{aligned}
\langle \mathcal{A}(t^1), s^1 \rangle &= \langle e^{-T^1} \mathcal{H} e^{T^1} \Phi_0, S^0 \Phi_0 \rangle + \langle e^{-T^1} \mathcal{H} e^{T^1} \Phi_0, S^\angle \Phi_0 \rangle \\
&= \langle (e^{-T^0} + e^{-T^0}(e^{-T^\angle} - I)) \mathcal{H} (e^{T^0} + e^{T^0}(e^{T^\angle} - I)) \Phi_0, S^0 \Phi_0 \rangle \\
&\quad + \langle e^{-T^1} \mathcal{H} e^{T^1} \Phi_0, S^\angle \Phi_0 \rangle \\
&= \langle \mathcal{A}^0(t^0), s^0 \rangle + \langle e^{-T^0}(e^{-T^\angle} - I) \mathcal{H} e^{T^0}(e^{T^\angle} - I) \Phi_0, S^0 \Phi_0 \rangle \\
&\quad + \langle e^{-T^0}(e^{-T^\angle} - I) \mathcal{H} e^{T^0} \Phi_0, S^0 \Phi_0 \rangle + \langle e^{-T^0} \mathcal{H} e^{T^0}(e^{T^\angle} - I) \Phi_0, S^0 \Phi_0 \rangle \\
&\quad + \langle e^{-T^1} \mathcal{H} e^{T^1} \Phi_0, S^\angle \Phi_0 \rangle \\
&= \langle \mathcal{A}(t^0), s^0 \rangle + \langle e^{-T^0} \mathcal{H} e^{T^0}(e^{T^\angle} - I) \Phi_0, S^0 \Phi_0 \rangle + \langle \mathcal{A}(t^1), s^\angle \rangle,
\end{aligned} \tag{5.29}$$

where in the last step the second and the third terms of the penultimate expression vanish because

$$(e^{-T^\angle} - I)^\dagger S^0 \Phi_0 = -(T^\angle)^\dagger S^0 \Phi_0 - \frac{1}{2} \left((T^\angle)^\dagger \right)^2 S^0 \Phi_0 - \ldots = 0,$$

due to the definition of the spaces $\mathbb{V}^0$ and $\mathbb{V}^\angle$. Note that this last relation would not hold in the case $\mathbb{V}^0 = \mathbb{V}(G(\mathrm{D})) \cap \mathbb{V}^1$, $\mathbb{V}^1 = \mathbb{V}(G^{\mathrm{full}})$, which is actually excluded by assumption.

Motivated by the preceding calculation in (5.29), we define the *Kowalski–Piecuch homotopy* $\mathcal{K}_{\mathrm{KP}} : \mathbb{V}^1 \times [0,1] \to (\mathbb{V}^1)^*$ via the instruction

$$\begin{aligned}
\langle \mathcal{K}_{\mathrm{KP}}(t^1, \lambda), s^1 \rangle &= \langle \mathcal{H}(t^0) \Phi_0, S^0 \Phi_0 \rangle + \langle \mathcal{H}(t^1) \Phi_0, S^\angle \Phi_0 \rangle \\
&\quad + \lambda \langle \mathcal{H}(t^0)(e^{T^\angle} - I) \Phi_0, S^0 \Phi_0 \rangle
\end{aligned} \tag{5.30}$$

for all $t^1, s^1 \in \mathbb{V}^1$ and $\lambda \in [0,1]$. Here, the relation $\mathcal{K}_{\mathrm{KP}}(t^1_*, \lambda) = 0$ is the same as the system [43, eqs. (90)–(91) and (93)].

It is obvious that $\mathcal{K}_{\mathrm{KP}}$ is an admissible homotopy. However, it is unclear whether it is faithful or not. In fact, $\mathcal{K}_{\mathrm{KP}}(t^1_{**}, 0) = 0$ is equivalent to the system

$$\langle \mathcal{H}(t^0_{**}) \Phi_0, S^0 \Phi_0 \rangle = 0, \tag{5.31}$$

$$\langle \mathcal{H}(t^0_{**} + t^\angle_{**}) \Phi_0, S^\angle \Phi_0 \rangle = 0, \tag{5.32}$$

for all $s^1 \in \mathbb{V}^1$. In other words, the usual SRCC equation $\mathcal{A}(t^0_{**}) = 0$, (5.31), is augmented with an additional equation for $t^\angle_{**}$, (5.32), which, in turn, depends on t^0_{**}. It is not obvious at all that (5.32) has a solution $t^\angle_{**}$ for a given zero t^0_{**}. The extensive numerical evidence in [43] clearly indicates that (5.32) admits a solution in various circumstances.

Before proving our existence result we recast the KP homotopy into a more convenient form.

LEMMA 5.31. *The following formula holds true:*

$$\langle \mathcal{K}_{\mathrm{KP}}(t^1,\lambda), s^1\rangle = \langle \mathcal{A}(t^0+\lambda t^{\angle}), s^0\rangle + \langle \mathcal{A}(t^1), s^{\angle}\rangle, \tag{5.33}$$

for all $t^1, s^1 \in \mathbb{V}^1$ *and* $\lambda \in [0,1]$.

Proof. It is enough to prove that

$$\langle \mathcal{K}_{\mathrm{KP}}(t^1,\lambda), s^1\rangle = \langle e^{-T^0}\mathcal{H}e^{T^0+\lambda T^{\angle}}\Phi_0, S^0\Phi_0\rangle + \langle e^{-T^1}\mathcal{H}e^{T^1}\Phi_0, S^{\angle}\Phi_0\rangle, \tag{5.34}$$

because $(e^{-\lambda T^{\angle}})^{\dagger}S^0\Phi_0 = S^0\Phi_0$. To see (5.34), suppose that $\mathcal{H}e^{T^0+\lambda T^{\angle}}\Phi_0 \in \mathfrak{H}^1$ and note that in the expansion $e^{T^0+\lambda T^{\angle}} = e^{T^0}(I + \lambda T^{\angle} + \frac{\lambda^2}{2!}(T^{\angle})^2 + \ldots + \frac{\lambda^N}{N!}(T^{\angle})^N)$ the quadratic and higher-order terms do not contribute because the excitation rank of $\mathcal{H}(T^{\angle})^k\Phi_0$ exceeds ρ for $k = 2,\ldots,N$, due to the fact that $\mathcal{H}$ is a two-body operator. The case $\mathcal{H}e^{T^0+\lambda T^{\angle}}\Phi_0 \in \mathfrak{H}^{-1}$ is obtained by density. ∎

To formulate the existence result, note that because $\mathbb{V}^1$ is assumed to be finite-dimensional, it is always possible to choose an $\alpha > 0$ so that

$$\langle (\mathcal{A}'(t^1_*) + \alpha I)r^1, r^1\rangle \geq \gamma_\alpha \|r^1\|^2_{\mathbb{V}} \quad \text{for all} \quad r^1 \in \mathbb{V}^1, \tag{5.35}$$

for some $\gamma_\alpha > 0$. This is also true in the complex case with a "Re" added to the left-hand side.

Define the operator $\Theta_\alpha : \mathbb{V} \to \mathbb{V}$ via[15]

$$\langle \mathcal{A}'(t^1_*)u^1, \Theta_\alpha v^1\rangle = \langle (\mathcal{A}'(t^1_*) + \alpha I)u^1, v^1\rangle \quad \text{for all} \quad u^1, v^1 \in \mathbb{V}^1. \tag{5.36}$$

Then Θ_α is well-defined, as long as $\ker \mathcal{A}'(t^1_*) = \{0\}$, i.e. that t^1_* is non-degenerate.

THEOREM 5.32 (Existence for KP). *Let* $t^1_* \in \mathbb{V}^1$ *be a non-degenerate zero of* $\mathcal{A}$. *Suppose the following.*

(i) *With* $\theta_0 = \|\Pi_0(\Theta_\alpha - I)\Pi_0\|^2_{\mathcal{L}(\mathbb{V},\mathbb{V})}$ *and* $\theta_{\angle} = \|\Pi_0(\Theta_\alpha - I)\Pi_{\angle}\|^2_{\mathcal{L}(\mathbb{V},\mathbb{V})}$, *there holds*

$$\eta := (1-g)\frac{\gamma_\alpha - \frac{1}{2}M_\delta\|\Theta_\alpha\|_{\mathcal{L}(\mathbb{V},\mathbb{V})}\delta}{\Delta(t^1_*) + M_\delta\delta} - \frac{1}{2}\max\{\varepsilon + 2(1+\varepsilon^{-1})\theta_0, 2(1+\varepsilon^{-1})\theta_{\angle}\} > \frac{1}{2},$$

for some $\varepsilon > 0$ *and* $\delta > 0$. *Here,* $\alpha > 0$ *and* $\gamma_\alpha > 0$ *satisfy* (5.35) *and* $0 < g < 1$ *is such that* $|\langle t^0, t^{\angle}\rangle_{\mathbb{V}}| \leq g\|t^0\|_{\mathbb{V}}\|t^{\angle}\|_{\mathbb{V}}$ *for all* $t^1 \in \mathbb{V}^1$.

(ii) *With* $\varkappa = \|t^{\angle}_*\|_{\mathbb{V}}$, *there holds*

$$\varkappa < \frac{2\sqrt{\eta} - \sqrt{2}}{2 - \sqrt{2} + 2\sqrt{\eta}}\delta. \tag{5.37}$$

Let $D = \{t^1_* + r^1 \in \mathbb{V}^1 : \|r^0\|^2_{\mathbb{V}} + \|r^{\angle}\|^2_{\mathbb{V}} < \frac{1}{2}(\delta - \varkappa)^2\}$. *Then, for any* $\lambda \in [0,1)$, *there exists* $t^1_{**}(\lambda) \in D$ *such that* $\mathcal{K}_{\mathrm{KP}}(t^1_{**}(\lambda), \lambda) = 0$. *Furthermore,* $\deg(\mathcal{K}_{\mathrm{KP}}(\cdot,\lambda), D, 0) \equiv d \neq 0$ *for all* $\lambda \in [0,1]$. *In particular, there exists* $t^1_{**} \in D$ *such that* $\mathcal{A}(t^0_{**}) = 0$.

Proof. We first prove that $\mathcal{K}_{\mathrm{KP}}(\cdot,\lambda) \neq 0$ on ∂D for all $\lambda \in [0,1]$. Set $t^1 = t^1_* + r^1$ and $s^1 = \Theta_\alpha r^1$ in (5.33) where $r^1 \in \partial D$, and write

$$\begin{aligned}
&\langle \mathcal{K}_{\mathrm{KP}}(t^1_* + r^1, \lambda), \Theta_\alpha r^1\rangle = \langle \mathcal{A}(t^0_* + r^0 + \lambda t^{\angle}_* + \lambda r^{\angle}), \Pi_0\Theta_\alpha r^1\rangle + \langle \mathcal{A}(t^1_* + r^1), \Pi_{\angle}\Theta_\alpha r^1\rangle \\
&= \langle \mathcal{A}(t^0_* + \lambda t^{\angle}_* + r^0 + \lambda r^{\angle}) - \mathcal{A}(t^1_* + r^1), \Pi_0\Theta_\alpha r^1\rangle + \langle \mathcal{A}(t^1_* + r^1), \Theta_\alpha r^1\rangle \\
&=: \mathrm{(I)} + \mathrm{(II)}.
\end{aligned}$$

[15] The authors learned this trick from [10, 9].

For (I), Taylor expansion around t_*^1 leads to

$$\begin{aligned}
\text{(I)} &= (\lambda-1)\langle \mathcal{A}'(t_*^1)(t_*^{\angle}+r^{\angle}), \Pi_0\Theta_\alpha r^1\rangle \\
&\quad + \langle \mathcal{R}_2(t_*^1,(\lambda-1)t_*^{\angle}+r^0+\lambda r^{\angle}) - \mathcal{R}_2(t_*^1,r^1), \Pi_0\Theta_\alpha r^1\rangle \\
&\geq (\lambda-1)\langle \mathcal{A}'(t_*^1)(t_*^{\angle}+r^{\angle}), \Pi_0\Theta_\alpha r^1\rangle - M\|t_*^{\angle}+r^{\angle}\|\|\Pi_0\Theta_\alpha r^1\| \\
&\geq -(\Delta(t_*^1)+M)\|t_*^{\angle}+r^{\angle}\|_{\mathbb{V}}\|\Pi_0\Theta_\alpha r^1\|_{\mathbb{V}},
\end{aligned}$$

where we used the intermediate value inequality. Here, letting $\Psi^0 \in \mathfrak{V}^0$ correspond to the amplitude $\Pi_0\Theta_\alpha r^1$,

$$\begin{aligned}
\langle \mathcal{A}'(t_*^1)(t_*^{\angle}+r^{\angle}), \Pi_0\Theta_\alpha r^1\rangle &= \langle (\mathcal{H}(t_*^1) - \mathcal{E}_{\mathrm{CC}}(t_*^1))(T_*^{\angle}+R^{\angle})\Phi_0, \Psi^0\rangle \\
&= \langle \mathcal{H}(t_*^1)(T_*^{\angle}+R^{\angle})\Phi_0, \Psi^0\rangle \\
&= \langle (\mathcal{H}+[\mathcal{H},T_*^1]+\ldots)(T_*^{\angle}+R^{\angle})\Phi_0, \Psi^0\rangle \\
&\leq \Delta(t_*^1)\|t_*^{\angle}+r^{\angle}\|_{\mathbb{V}}\|\Pi_0\Theta_\alpha r^1\|_{\mathbb{V}},
\end{aligned}$$

where the first equality is (5.11), and in the last inequality we exploited that $\mathcal{H}$ decreases the excitation rank at most by 2 (so that the single amplitudes $(t_*^1)_1$ of t_*^1 only contribute). Also, the following estimate was used,

$$\begin{aligned}
M &= \max_{\xi\in[(\lambda-1)t_*^{\angle}+r^0+\lambda r^{\angle},r^1]} \|\partial_2\mathcal{R}_2(t_*^1,\xi)\|_{\mathcal{L}(\mathbb{V},\mathbb{V}^*)} \\
&= \max_{\xi\in[(\lambda-1)t_*^{\angle}+r^0+\lambda r^{\angle},r^1]} \|\mathcal{A}'(t_*+\xi)-\mathcal{A}'(t_*)\|_{\mathcal{L}(\mathbb{V},\mathbb{V}^*)} \\
&\leq \max_{\xi\in[(\lambda-1)t_*^{\angle}+r^0+\lambda r^{\angle},r^1]} \|\xi\|_{\mathbb{V}} \max_{\zeta\in[0,\xi]} \|\mathcal{A}''(t_*^1+\zeta)\|_{\mathcal{L}(\mathbb{V}\times\mathbb{V},\mathbb{V}^*)} \\
&\leq M_\delta \max_{\xi\in[(\lambda-1)t_*^{\angle}+r^0+\lambda r^{\angle},r^1]} \|\xi\|_{\mathbb{V}} \leq M_\delta(\|r^0\|_{\mathbb{V}}+\|r^{\angle}\|_{\mathbb{V}}+\varkappa) \leq M_\delta\delta.
\end{aligned}$$

For (II), we have using (5.35), (5.36) and Taylor's theorem,

$$\begin{aligned}
\text{(II)} &= \langle \mathcal{A}'(t_*^1)r^1, \Theta_\alpha r^1\rangle - \langle \mathcal{R}_2(t_*^1,r^1), \Theta_\alpha r^1\rangle \geq \gamma_\alpha\|r^1\|_{\mathbb{V}}^2 - \tfrac{1}{2}M_\delta\|r^1\|_{\mathbb{V}}^2\|\Theta_\alpha r^1\|_{\mathbb{V}} \\
&\geq (\gamma_\alpha - \tfrac{1}{2}M_\delta\|\Theta_\alpha\|_{\mathcal{L}(\mathbb{V},\mathbb{V})}\|r^1\|_{\mathbb{V}})\|r^1\|_{\mathbb{V}}^2 \geq (\gamma_\alpha - \tfrac{1}{2}M_\delta\|\Theta_\alpha\|_{\mathcal{L}(\mathbb{V},\mathbb{V})}\delta)\|r^1\|_{\mathbb{V}}^2.
\end{aligned}$$

In summary, using the definitions of θ_0 and $\theta_{\angle}$, and setting $\widetilde{\gamma} = \gamma_\alpha - \frac{1}{2}M_\delta\|\Theta_\alpha\|_{\mathcal{L}(\mathbb{V},\mathbb{V})}\delta$,

$$\begin{aligned}
&\langle \mathcal{K}_{\mathrm{KP}}(t_*^1+r^1,\lambda), \Theta_\alpha r^1\rangle \\
&\geq \widetilde{\gamma}\|r^1\|_{\mathbb{V}}^2 - (\Delta(t_*^1)+M_\delta\delta)\|t_*^{\angle}+r^{\angle}\|_{\mathbb{V}}\|\Pi_0\Theta_\alpha r^1\|_{\mathbb{V}} \\
&\geq \widetilde{\gamma}\|r^1\|_{\mathbb{V}}^2 - \frac{\Delta(t_*^1)+M_\delta\delta}{2}(\|t_*^{\angle}+r^{\angle}\|_{\mathbb{V}}^2 + \|\Pi_0\Theta_\alpha r^1\|_{\mathbb{V}}^2) \\
&\geq (1-g)\widetilde{\gamma}(\|r^0\|_{\mathbb{V}}^2+\|r^{\angle}\|_{\mathbb{V}}^2) - \frac{\Delta(t_*^1)+M_\delta\delta}{2}\left(\varkappa^2+2\varkappa\|r^{\angle}\|_{\mathbb{V}}+\|r^{\angle}\|_{\mathbb{V}}^2+\|\Pi_0\Theta_\alpha r^1\|_{\mathbb{V}}^2\right) \\
&\geq \left((1-g)\widetilde{\gamma} - \frac{\Delta(t_*^1)+M_\delta\delta}{2}\Big(1+\max\{\varepsilon+2(1+\varepsilon^{-1})\theta_0, 2(1+\varepsilon^{-1})\theta_{\angle}\}\Big)\right) \\
&\quad \times (\|r^0\|_{\mathbb{V}}^2+\|r^{\angle}\|_{\mathbb{V}}^2) - \frac{\Delta(t_*^1)+M_\delta\delta}{2}(\varkappa^2+\sqrt{2}\varkappa(\delta-\varkappa)) \\
&\geq \left((1-g)\widetilde{\gamma} - \frac{\Delta(t_*^1)+M_\delta\delta}{2}\max\{\varepsilon+2(1+\varepsilon^{-1})\theta_0, 2(1+\varepsilon^{-1})\theta_{\angle}\}\right)\frac{(\delta-\varkappa)^2}{2} \\
&\quad - \frac{\Delta(t_*^1)+M_\delta\delta}{2}\left(\varkappa+\frac{\sqrt{2}}{2}(\delta-\varkappa)\right)^2.
\end{aligned}$$

The positivity of the last expression follows from (5.37). We also used the bound

$$\begin{aligned}
&\|r^{\angle}\|_{\mathbb{V}}^2 + \|\Pi_0\Theta_\alpha r^1\|_{\mathbb{V}}^2 \leq (1+\varepsilon^{-1})\|\Pi_0(\Theta_\alpha - I)r^1\|_{\mathbb{V}}^2 + (1+\varepsilon)\|r^0\|_{\mathbb{V}}^2 + \|r^{\angle}\|_{\mathbb{V}}^2 \\
&\leq 2(1+\varepsilon^{-1})(\|\Pi_0(\Theta_\alpha - I)\Pi_0\|_{\mathcal{L}(\mathbb{V},\mathbb{V})}^2\|r^0\|_{\mathbb{V}}^2 + \|\Pi_0(\Theta_\alpha - I)\Pi_{\angle}\|_{\mathcal{L}(\mathbb{V},\mathbb{V})}^2\|r^{\angle}\|_{\mathbb{V}}^2) \\
&+ (1+\varepsilon)\|r^0\|_{\mathbb{V}}^2 + \|r^{\angle}\|_{\mathbb{V}}^2 \\
&\leq (1+\max\{\varepsilon + 2(1+\varepsilon^{-1})\theta_0, 2(1+\varepsilon^{-1})\theta_{\angle}\})(\|r^0\|_{\mathbb{V}}^2 + \|r^{\angle}\|_{\mathbb{V}}^2).
\end{aligned}$$

We can now apply Theorem A.1 (iii) to conclude $\deg(\mathcal{K}_{\mathrm{KP}}(\cdot,\lambda), D, 0) \equiv d \neq 0$ for all $\lambda \in [0,1]$, with δ decreased if necessary. Invoking Corollary A.2 (i) completes the proof. ∎

Remark 5.33.

(i) A crucial difference between the linear homotopy and the KP homotopy is that while the decomposition used for $\mathcal{K}_{\mathrm{L}}$ is $\mathbb{V}$-orthogonal, the decomposition for $\mathcal{K}_{\mathrm{KP}}$ is ℓ^2-orthogonal—the computation (5.29) and Lemma 5.31 exploit this heavily. Nevertheless, we used the $\mathbb{V}$-inner product in the existence result for $\mathcal{K}_{\mathrm{KP}}$. This geometric discrepancy is reflected in condition (i) above.

(ii) Using the Cauchy–Schwarz inequality it is clear that $|\langle t^0, t^{\angle}\rangle_{\mathbb{V}}| < \|t^0\|_{\mathbb{V}}\|t^{\angle}\|_{\mathbb{V}}$ for all $t^1 \in \mathbb{V}^1$, due to the fact that $\mathbb{V}^1 = \mathbb{V}^0 \oplus \mathbb{V}^{\angle}$. The maximum of the function $t^1 \mapsto \langle t^0, t^{\angle}\rangle_{\mathbb{V}}$ is attained on $\|t^0\|_{\mathbb{V}} = 1$, $\|t^{\angle}\|_{\mathbb{V}} = 1$ and this maximum may be taken as the $0 < g < 1$ of condition (i).

(iii) In the coercive case, i.e. when (5.35) holds with $\alpha = 0$, we have $\Theta_0 = I$, so that $\theta_0 = \theta_{\angle} = 0$. Letting $\varepsilon \to 0$, condition (i) simplifies to

$$\eta = \frac{(1-g)(\gamma_0 - \frac{1}{2}M_\delta\delta)}{\Delta(t_*^1) + M_\delta\delta} > \frac{1}{2}.$$

This last condition in turn reduces to $\gamma_0 > M_\delta\delta$ as g and $\Delta(t_*)$ approaches zero.

(iv) It is interesting to note that while the linear homotopy $\mathcal{K}_{\mathrm{L}}$ involves the "targeted" solution t_*^1, the KP homotopy $\mathcal{K}_{\mathrm{KP}}$ does not. In this sense, $\mathcal{K}_{\mathrm{KP}}$ is "universal".

(v) The result is straightforward to extend to the complex case.

The constant $\Delta(t_*^1)$ also deserves some explanation. Roughly speaking, the "defect" $\Delta(t_*^1)$ measures how much the subspace $\mathfrak{V}^0 \subset \mathfrak{V}$ deviates from being an invariant subspace of the operator $(\mathcal{H} + [(T_*^1)_1^\dagger, \mathcal{H}])_{\mathfrak{V}} : \mathfrak{V} \to \mathfrak{V}$. In addition, we can invoke (5.3) to write

$$\Delta(t_*^1) = \sup_{\substack{\|u^0\|_{\mathbb{V}}=1 \\ \|v^{\angle}\|_{\mathbb{V}}=1}} |\langle (\mathcal{W} + [(T_*^1)_1^\dagger, \mathcal{W}])U^0\Phi_0, V^{\angle}\Phi_0\rangle|.$$

Note that the term $[(T_*^1)_1^\dagger, \mathcal{W}]$ can be eliminated via orbital rotations (since $(t_*^1)_1 = 0$ can be achieved according to the Thouless theorem), as it involves single excitations only, in which case the amplitude dependence of Δ is removed. Hence, Δ quantifies how much the operator $\mathcal{W}_{\mathfrak{V}}$ leaves $\mathfrak{V}^0$ invariant.

We summarize the above existence result in the corollary below that holds under the following structural assumptions.

Assumption (KPA). $\Delta(t_*^1)$ can be made sufficiently small by an appropriate choice of the orbital basis and truncation level $1 \leq \rho < N$.

Assumption (KPB). There is a $\delta_0 > 0$ such that M_{δ_0} is sufficiently small.

COROLLARY 5.34. *Suppose that $t^1_* \in \mathbb{V}^1$ is a non-degenerate zero of $\mathcal{A}$ and that Assumptions (KPA) and (KPB) hold. If $\|t^{\angle}_*\|_{\mathbb{V}}$ is sufficiently small, then there exists $t^1_{**} \in \mathbb{V}^1$ in a neigborhood of t^1_* such that $\mathcal{K}_{\mathrm{KP}}(t^1_{**}, 0) = 0$.*

Remark 5.35. Recall that M_δ only involves the second derivative $\mathcal{A}''$ near t^1_* (see (5.15)), so that (KPB) can be viewed as a "perturbative" assumption. Further, as we noted in Remark 5.9, M_δ involves the mapping $\zeta \mapsto [[\mathcal{W}(t^1_* + \zeta), \cdot], \cdot]\Phi_0$. Therefore, (KPB) may be viewed as the higher-order generalization of assumption on the smallness of the local Lipschitz constant of the mapping $\zeta \mapsto \mathcal{W}(t^1_* + \zeta)\Phi_0$ in [50, Assumption BII] (see also Remark 5.39 (v)).

In the last step of the *proof* of Theorem 5.32, we could have invoked Theorem 5.27 instead to obtain the existence of a connected component $\mathcal{C}$ of the zero set $\mathcal{Z}(\mathcal{K}_{\mathrm{KP}})$, such that $\mathcal{C} \cap (\mathcal{Z}(\mathcal{K}_{\mathrm{KP}}))_j \neq \emptyset$ for $j = 0, 1$. Under the above assumptions, this provides a theoretical basis for the "solution trajectories" observed in [43]. Further, combining Theorem 5.32 with (5.28), we get for $t^1_{**} = t^1_{**}(0)$,

$$\sum_{t^1_{**} \in \mathcal{C} \cap (\mathcal{Z}(\mathcal{K}_{\mathrm{KP}}))_0} i(\mathcal{K}_{\mathrm{KP}}(\cdot, 0), t^1_{**}) = i(\mathcal{A}, t^1_*).$$

Since we do not have uniqueness for t^1_{**} in this case, it is possible that the left-hand side may contain multiple terms which sum up to $i(\mathcal{A}, t^1_*)$. In particular, $i(\mathcal{K}_{\mathrm{KP}}(\cdot, 0), t^1_{**})$ does not need to be $i(\mathcal{A}, t^1_*)$.

To close this section, we calculate the topological index for a non-degenerate zero at the $\lambda = 0$ endpoint of the KP homotopy. Note that the index at $\lambda = 1$ is simply given by Theorem 5.13 and Theorem 5.18.

Fix $t^1 \in \mathbb{V}^1$ and define the operator $\widehat{\mathcal{H}}(t^1)$ (not to be confused with (5.10) in a different context),

$$\widehat{\mathcal{H}}(t^1) = \mathcal{H}(t^1) - \sum_{\alpha \in \Xi(G^0)} \langle \mathcal{H}(t^1)\Phi_0, \Phi_\alpha \rangle X_\alpha. \tag{5.38}$$

Notice that $\langle \widehat{\mathcal{H}}(t^1)\Phi_0, S^0\Phi_0 \rangle = 0$ for all $s^0 \in \mathbb{V}^0$. Define the linear mapping $\widehat{\mathcal{H}}_{\mathfrak{V}^0, \mathfrak{V}^{\angle}}(t^1) : \mathfrak{V}^0 \to \mathfrak{V}^{\angle}$ via $\langle \widehat{\mathcal{H}}_{\mathfrak{V}^0, \mathfrak{V}^{\angle}}(t^1)\Psi, \Psi' \rangle = \langle \widehat{\mathcal{H}}(t^1)\Psi, \Psi' \rangle$ for all $\Psi \in \mathfrak{V}^0$ and $\Psi' \in \mathfrak{V}^{\angle}$. We first calculate the derivative of $\mathcal{K}_{\mathrm{KP}}(\cdot, 0)$.

LEMMA 5.36. *The derivative $\partial_1 \mathcal{K}_{\mathrm{KP}}(t^1_{**}, 0) : \mathbb{V}^1 \to (\mathbb{V}^1)^*$ of $\mathcal{K}_{\mathrm{KP}}(\cdot, 0)$ at a zero t^1_{**} is given by*

$$\langle \partial_1 \mathcal{K}_{\mathrm{KP}}(t^1_*, 0)u^1, v^1 \rangle = \left\langle \begin{pmatrix} \mathcal{H}_{\mathfrak{V}^0}(t^0_{**}) - \mathcal{E}_{\mathrm{CC}}(t^0_{**}) & 0 \\ \widehat{\mathcal{H}}_{\mathfrak{V}^0, \mathfrak{V}^{\angle}}(t^1_{**}) & \widehat{\mathcal{H}}_{\mathfrak{V}^{\angle}}(t^1_{**}) - \mathcal{E}_{\mathrm{CC}}(t^1_{**}) \end{pmatrix} \begin{pmatrix} U^0\Phi_0 \\ U^{\angle}\Phi_0 \end{pmatrix}, \begin{pmatrix} V^0\Phi_0 \\ V^{\angle}\Phi_0 \end{pmatrix} \right\rangle$$

for all $u^1, v^1 \in \mathbb{V}^1$.

Proof. A calculation analogous to the one in the proof of Lemma 5.6 shows that $D(t^1) := \partial_1 \mathcal{K}_{\mathrm{KP}}(t^1, 0) : \mathbb{V}^1 \to (\mathbb{V}^1)^*$ is given by

$$\begin{aligned} \langle D(t^1)u^1, v^1 \rangle &= \langle [\mathcal{H}(t^0), U^0]\Phi_0, V^0\Phi_0 \rangle + \langle [\mathcal{H}(t^1), U^1]\Phi_0, V^{\angle}\Phi_0 \rangle \\ &=: D_1(t^1) + D_2(t^1) \end{aligned} \tag{5.39}$$

for all $t^1, u^1, v^1 \in \mathbb{V}^1$. We set $t^1 = t^1_{**}$ and evaluate D_1 and D_2. Write

$$(U^0)^\dagger V^0 \Phi_0 = \sum_{\alpha \in \Xi(G^1)\cup\{0\}} \langle U^0\Phi_\alpha, V^0\Phi_0\rangle \Phi_\alpha,$$

from which,

$$\begin{aligned}\langle U^0 \mathcal{H}(t^0_{**})\Phi_0, V^0\Phi_0\rangle &= \langle \mathcal{H}(t^0_{**})\Phi_0, (U^0)^\dagger V^0\Phi_0\rangle \\ &= \left(\sum_{\alpha\in\Xi(G^0)} + \sum_{\alpha\in\Xi(G^0)^c}\right)\langle \mathcal{H}(t^0_{**})\Phi_0, \Phi_\alpha\rangle\langle U^0\Phi_\alpha, V^0\Phi_0\rangle + \mathcal{E}_{\mathrm{CC}}(t^0_{**})\langle U^0\Phi_0, V^0\Phi_0\rangle.\end{aligned}$$

Here, the first sum vanishes because of (5.31) and the second by the orthogonality of $\mathbb{V}^0$ and $\mathbb{V}^\angle$. Consequently,

$$D_1(t^1_{**}) = \langle (\mathcal{H}(t^0_{**}) - \mathcal{E}_{\mathrm{CC}}(t^0_{**}))U^0\Phi_0, V^0\Phi_0\rangle.$$

Analogously, (5.32) implies that

$$\begin{aligned}\langle U^\angle \mathcal{H}(t^1_{**})\Phi_0, V^\angle\Phi_0\rangle = &\sum_{\alpha\in\Xi(G^0)} \langle \mathcal{H}(t^1_{**})\Phi_0, \Phi_\alpha\rangle\langle X_\alpha U^\angle\Phi_0, V^\angle\Phi_0\rangle \\ &+ \mathcal{E}_{\mathrm{CC}}(t^1_{**})\langle U^\angle\Phi_0, V^\angle\Phi_0\rangle,\end{aligned}$$

and

$$\langle U^0 \mathcal{H}(t^1_{**})\Phi_0, V^\angle\Phi_0\rangle = \sum_{\alpha\in\Xi(G^0)} \langle \mathcal{H}(t^1_{**})\Phi_0, \Phi_\alpha\rangle\langle X_\alpha U^0\Phi_0, V^\angle\Phi_0\rangle.$$

Thus,

$$D_2(t^1_{**}) = \langle (\widehat{\mathcal{H}}(t^1_{**}) - \mathcal{E}_{\mathrm{CC}}(t^1_{**}))U^\angle\Phi_0, V^\angle\Phi_0\rangle + \langle \widehat{\mathcal{H}}(t^1_{**})U^0\Phi_0, V^\angle\Phi_0\rangle,$$

and the stated expression now follows. ∎

The preceding Lemma implies the index formula for the KP homotopy.

THEOREM 5.37 (Index formula for KP – non-degenerate case). *Suppose that $t^1_{**} \in \mathbb{V}^1$ is a zero of $\mathcal{K}_{\mathrm{KP}}(\cdot, 0)$. Then t^1_{**} is non-degenerate if and only if the conditions*

(I) $\mathcal{E}_{\mathrm{CC}}(t^0_{**}) \notin \sigma(\mathcal{H}_{\mathfrak{V}^0}(t^0_{**}))$,

(II) $\mathcal{E}_{\mathrm{CC}}(t^1_{**}) \notin \sigma(\widehat{\mathcal{H}}_{\mathfrak{V}^\angle}(t^1_{**}))$

both hold true, and in this case the topological index of $\mathcal{K}_{\mathrm{KP}}(\cdot, 0)$ at t^1_{**} is given by

$$i(\mathcal{K}_{\mathrm{KP}}(\cdot, 0), t^1_{**}) = (-1)^{\nu^0+\nu^\angle},$$

where

$$\begin{aligned}\nu^0 &= |\{j : \mathcal{E}_j(\mathcal{H}_{\mathfrak{V}^0}(t^0_{**})) \in \mathbb{R},\ \mathcal{E}_j(\mathcal{H}_{\mathfrak{V}^0}(t^0_{**})) < \mathcal{E}_{\mathrm{CC}}(t^0_{**})\}|, \\ \nu^\angle &= |\{j : \mathcal{E}_j(\widehat{\mathcal{H}}_{\mathfrak{V}^\angle}(t^1_{**})) \in \mathbb{R},\ \mathcal{E}_j(\widehat{\mathcal{H}}_{\mathfrak{V}^\angle}(t^1_{**})) < \mathcal{E}_{\mathrm{CC}}(t^1_{**})\}|.\end{aligned}$$

Proof. The proof follows from Lemma 5.36 along similar lines as Theorem 5.13,

$$\begin{aligned}i(\mathcal{K}_{\mathrm{KP}}(\cdot, 0), t^1_{**}) = &\operatorname{sgn}\prod_{j\geq 0}(\mathcal{E}_j(\mathcal{H}_{\mathfrak{V}^0}(t^0_{**})) - \mathcal{E}_{\mathrm{CC}}(t^0_{**})) \\ &\times \operatorname{sgn}\prod_{j\geq 0}(\mathcal{E}_j(\widehat{\mathcal{H}}_{\mathfrak{V}^\angle}(t^1_{**})) - \mathcal{E}_{\mathrm{CC}}(t^1_{**})).\end{aligned}$$

∎

5.6. An energy error estimate. In this this section we derive an energy error estimate for general eigenstates for the KP homotopy using the results of Appendix E.

THEOREM 5.38 (Energy error estimate). *Let* $\mathbb{V}^1 = \mathbb{V}(G^{\mathrm{full}})$ *and suppose that* $t_*^1 \in \mathbb{V}^1$ *is a zero of* $\mathcal{A}$*, and that* $t_{**}^1 \in \mathbb{V}^1$ *is a zero of* $\mathcal{K}_{\mathrm{KP}}(\cdot, 0) = 0$*. If the nonorthogonality condition* $\langle e^{T_{**}^0}\Phi_0, e^{T_*^1}\Phi_0\rangle \neq 0$ *holds true, then*

$$|\mathcal{E}_{\mathrm{CC}}(t_{**}^1) - \mathcal{E}_{\mathrm{CC}}(t_*^1)| \leq C(t_{**}^0, t_{**}^1)\|t_{**}^{\angle}\|_{\mathbb{V}}, \tag{5.40}$$

where $C(t_{**}^0, t_{**}^1) \geq 0$ *is given by*

$$C(t_{**}^0, t_{**}^1) = (C^2 + M)\frac{\|\Pi_{\mathfrak{V}^{\angle}}(e^{T_{**}^0})^{\dagger}\Pi_{\mathfrak{V}^{\angle}}e^{T_*^1}\Phi_0\|_{\mathfrak{H}^1}}{|\langle e^{T_{**}^0}\Phi_0, e^{T_*^1}\Phi_0\rangle|}$$

and is independent of $t_{**}^{\angle}$*. Here,* C *is the norm equivalence constant from Remark 3.29 and* $M = \max_{\xi\in[t_{**}^0, t_{**}^1]}\|u^1 \mapsto \Pi_{\mathfrak{V}^{\angle}}[\mathcal{W}(\xi), U^1]\|_{\mathcal{L}(\mathbb{V},\mathfrak{H}^{-1})}$.

Proof. Setting $\Psi = e^{T_*^1}\Phi_0$ and $\lambda = 0$ in Theorem E.1 (II), we have

$$|\mathcal{E}_{\mathrm{CC}}(t_{**}^1) - \mathcal{E}_{\mathrm{CC}}(t_*^1)| = \frac{|\langle(\mathcal{H}(t_{**}^1) - \mathcal{H}(t_{**}^0))\Phi_0, \Pi_{\mathfrak{V}^{\angle}}(e^{T_{**}^0})^{\dagger}\Pi_{\mathfrak{V}^{\angle}}e^{T_*^1}\Phi_0\rangle|}{|\langle e^{T_{**}^0}\Phi_0, e^{T_*^1}\Phi_0\rangle|}. \tag{5.41}$$

Note that using (5.3) we may write

$$\begin{aligned}(\mathcal{H}(t_{**}^1) - \mathcal{H}(t_{**}^0))\Phi_0 &= [\mathcal{F}, T_{**}^{\angle}]\Phi_0 + (\mathcal{W}(t_{**}^1) - \mathcal{W}(t_{**}^0))\Phi_0 \\ &= \sum_{\gamma\in\Xi(G^{\angle})}\varepsilon_{\gamma}(t_{**}^{\angle})_{\gamma}\Phi_{\gamma} + (\mathcal{W}(t_{**}^1) - \mathcal{W}(t_{**}^0))\Phi_0.\end{aligned}$$

Hence, we can bound the numerator of the right-hand side of (5.41) as

$$\begin{aligned}&\Big|\sum_{\gamma\in\Xi(G^{\angle})}\varepsilon_{\gamma}(t_{**}^{\angle})_{\gamma}\langle\Phi_{\gamma}, \Pi_{\mathfrak{V}^{\angle}}(e^{T_{**}^0})^{\dagger}\Pi_{\mathfrak{V}^{\angle}}e^{T_*^1}\Phi_0\rangle\Big| \\ &+\Big|\langle(\mathcal{W}(t_{**}^1) - \mathcal{W}(t_{**}^0))\Phi_0, \Pi_{\mathfrak{V}^{\angle}}(e^{T_{**}^0})^{\dagger}\Pi_{\mathfrak{V}^{\angle}}e^{T_*^1}\Phi_0\rangle\Big|.\end{aligned}$$

Using the Cauchy–Schwarz inequality, the first term may be further bounded as

$$\sum_{\gamma\in\Xi(G^{\angle})}\varepsilon_{\gamma}(t_{**}^{\angle})_{\gamma}\langle\Phi_{\gamma}, \Pi_{\mathfrak{V}^{\angle}}(e^{T_{**}^0})^{\dagger}\Pi_{\mathfrak{V}^{\angle}}e^{T_*^1}\Phi_0\rangle \leq |||t_{**}^{\angle}|||\,\Big|\Big|\Big|\Pi_{\mathfrak{V}^{\angle}}(e^{T_{**}^0})^{\dagger}\Pi_{\mathfrak{V}^{\angle}}e^{T_*^1}\Phi_0\Big|\Big|\Big|,$$

where the norm $|||\cdot|||$ was defined in Remark 3.29. For the second term, we use the intermediate value inequality to obtain the bound

$$M\|t_{**}^{\angle}\|_{\mathbb{V}}\|\Pi_{\mathfrak{V}^{\angle}}(e^{T_{**}^0})^{\dagger}\Pi_{\mathfrak{V}^{\angle}}e^{T_*^1}\Phi_0\|_{\mathfrak{H}^1}. \qquad \square$$

Remark 5.39.

(i) If the nonorthogonality condition holds and $t_{**}^{\angle} = 0$, then according to (5.31)-(5.32), t_{**}^0 is an FCC solution such that $\langle e^{T_{**}^0}\Phi_0, e^{T_*^1}\Phi_0\rangle \neq 0$. In this case, (5.40) implies that the energy error is zero: $\mathcal{E}_{\mathrm{CC}}(t_{**}^0) = \mathcal{E}_{\mathrm{CC}}(t_*^1)$.

(ii) In the practically relevant case $\rho \geq 2$, using (E.1), we have that $\mathcal{E}_{\mathrm{CC}}(t_{**}^1) = \mathcal{E}_{\mathrm{CC}}(t_{**}^0)$ so the left-hand side of (5.40) does not involve $t_{**}^{\angle}$ at all.

(iii) The appearance of the quantity $\|t_{**}^{\angle}\|_{\mathbb{V}}$ allows us to view the auxiliary equation (5.32) as providing an *a posteriori* error estimate.

(iv) If the nonorthogonality condition does not hold, i.e. $\langle e^{T^0_{**}}\Phi_0, e^{T^1_*}\Phi_0\rangle = 0$, then $e^{T^0_{**}}\Phi_0$ and $e^{T^1_*}\Phi_0$ represent different eigenstates if t^1_* is assumed to be non-degenerate. While $e^{T^0_{**}}\Phi_0$ itself does not satisfy the Schrödinger equation, it must be viewed as an approximation to an eigenstate *different* from $e^{T^1_*}\Phi_0$.

(v) Note that a local Lipschitz assumption (on a ball including t^1_{**} and t^0_{**}) with constant L on the mapping $t \mapsto \mathcal{W}(t)\Phi_0$ can be used to obtain the result of the theorem with M replaced by L. Such an assumption is akin to Assumption B.II in [50] where it is used for guaranteeing the local strong monotonicity of the CC mapping.

6. Conclusions and further work. In the first half of the paper we proposed a framework to describe the discretization scheme involved in CC-like methods. At the core of the description is the concept of the excitation graph (Definition 3.4), which completely determines all necessary building blocks such as excitation operators (subsection 3.3), cluster operators (subsection 3.4) and amplitude spaces (subsection 3.5). The excitation graph concept admits a straightforward extension to the multireference case (Definition 3.14). Another advantage of our approach is that it avoids the use of second-quantized formalism and hence allowed us to prove the basic results (such as Theorem 3.17 and Theorem 3.21) in a more transparent manner. Besides these, we also pointed out a number of structural properties of the excitation graph in subsection 3.2. It is important to note that some of these graph-theoretic properties are reflected in the algebraic structure of the excitation operators (Theorem 3.16 and Theorem 3.23). Some relevant combinatorial quantities have been calculated in Appendix C. Furthermore, we proposed an algorithm to determine the reference states in an optimal fashion for the multirefence case in Appendix D.

In section 4, we provided rigorous derivations of both the single-reference- (subsection 4.1), and a multireference (subsection 4.2) CC method. The derivations used a general theorem (Theorem 4.1) motivated by a known method based on perturbation theory.

The rest of the paper was concerned with the analysis of the SRCC method. We began with the definitions and elementary considerations in subsection 5.1. Then we considered the local properties of the SRCC mapping in subsection 5.2, mainly in the finite-dimensional case. It turned out that the topological index of the CC mapping (Theorem 5.13) is connected with the nonvariational property of the CC method (Remark 5.16), and the eigenvalues of the Fock operator (Proposition 5.17). In the degenerate case, the classic Leray reduction formula provided the topological index (Theorem 5.18). We also discussed the case when the cluster amplitudes are allowed to be complex in subsection 5.3.

In subsection 5.4, we discussed how certain homotopies can be used to analyze the CC method, in particular to prove the existence of a truncated CC solution through the use of topological degree theory. This was done using an idea well-known both in nonlinear analysis and in quantum chemistry: that an appropriate homotopy "connects" the truncated problem with the exact problem (essentially the Schrödinger equation), therefore one is able to infer (homotopy-invariant) information regarding the former problem from the latter. As an introductory example, we considered the linear homotopy (Theorem 5.29).

Next, in subsection 5.5, motived by the works of the chemists Kowalski and Piecuch, we considered a homotopy that connects CC mappings corresponding to different truncation levels. Using this, we proved an existence result for the said homotopy (Theorem 5.32), which also implies the existence of a truncated CC solution

under certain assumptions. The index formula for the KP homotopy was also derived in the non-degenerate case (Theorem 5.37). Using an known result about the KP homotopy (Appendix E), we also derived an energy error estimate in subsection 5.6.

Finally, let us discuss some possible directions of research. Clearly, it would be interesting to extend our analysis to the infinite-dimensional case. An obvious next step would be the analysis of the JM-MRCC method (see subsection 4.2). It would be also interesting to look at the Extended CC (ECC) [32] and the Unitary CC (UCC) methods [5].

7. Acknowledgements. The authors would like to thank Fabian M. Faulstich and Simen Kvaal for helpful discussions and comments on the manuscript.

Appendix A. Topological degree.

We now briefly review the basic results in degree theory. We mainly follow the set of notes by Dinca and Mawhin [14], where the proofs may be found. The concept of the topological (a.k.a. mapping-, or Brouwer-) degree of a mapping goes back to Kronecker, Poincaré and Brouwer; Leray and Schauder extended it to infinite dimensions and demonstrated its usefulness in the theory of partial differential equations. (See [14, Chapter 23] for a fascinating historical perspective.) The standard textbooks on the topic are [13, 59, 41].

The first result gives an axiomatic characterization of the topological degree as the *unique* additive homotopy invariant that can be attached to continuous mappings (up to normalization). Its proof may be found in [14, Chapter 4].

THEOREM A.1 (Existence and uniqueness of the degree). *Let $D \subset \mathbb{R}^n$ be an open, bounded and nonempty subset. There exists a unique integer-valued function*

$$(\mathcal{A}, D, z) \mapsto \deg(\mathcal{A}, D, z),$$

where $\mathcal{A} \in C(\overline{D}, \mathbb{R}^n)$ and $z \notin \mathcal{A}(\partial D)$, such that the following properties hold true:

(i) *(Normalization) If $z \in D$, we have $\deg(\mathrm{id}, D, z) = 1$.*

(ii) *(Additivity) Let $D_1, D_2 \subset D$ be disjoint open subsets such that $z \notin \mathcal{A}(\overline{D} \setminus (D_1 \cup D_2))$. Then*

$$\deg(\mathcal{A}, D, z) = \deg(\mathcal{A}, D_1, z) + \deg(\mathcal{A}, D_2, z).$$

(iii) *(Homotopy invariance) If $\mathcal{A} \in C(\overline{D} \times [0,1], \mathbb{R}^n)$ and $z \notin \mathcal{A}(\partial D \times [0,1])$, then*

$$\lambda \mapsto \deg(\mathcal{A}(\cdot, \lambda), D, z) \quad \textit{is constant.}$$

Among the many useful properties of the degree we highlight the following few.

COROLLARY A.2.

(i) *(Excision property) Let $D \subset \mathbb{R}^n$ be an open, bounded and $D' \subset D$ open. If $z \notin \mathcal{A}(\overline{D} \setminus D')$, then $\deg(\mathcal{A}, D, z) = \deg(\mathcal{A}, D', z)$.*

(ii) *(Existence property) If $z \notin \mathcal{A}(\overline{D})$, then $\deg(\mathcal{A}, D, z) = 0$. Said differently, if $\deg(\mathcal{A}, D, z) \neq 0$, then there is a solution $u \in D$ to $\mathcal{A}(u) = z$.*

(iii) *(Additivity property) Suppose that $\{D_j\} \subset D$ is a sequence of open and disjoint sets. If $z \notin \mathcal{A}(\overline{D} \setminus \bigcup D_j)$, then $\deg(\mathcal{A}, D_j, z) = 0$ for all but finitely many j, and*

$$\deg(\mathcal{A}, D, z) = \sum_j \deg(\mathcal{A}, D_j, z).$$

The following formula is essential for the practical computation of the degree.

PROPOSITION A.3. *Let $\mathcal{A} \in C(\overline{D}, \mathbb{R}^n) \cap C^1(D, \mathbb{R}^n)$ such that $z \notin \mathcal{A}(\partial D)$ and $\det \mathcal{A}'(u) \neq 0$ for all $u \in \mathcal{A}^{-1}(z)$. Then*

$$\deg(\mathcal{A}, D, z) = \sum_{u \in \mathcal{A}^{-1}(z)} \operatorname{sgn} \det \mathcal{A}'(u).$$

Based on this, we call a point u *non-degenerate* if $\det \mathcal{A}'(u) \neq 0$ and *degenerate* otherwise. When talking about an *isolated solution* u of $\mathcal{A}$, i.e., a point $u \in \mathcal{A}^{-1}(z)$ such that there is a ball $B(u, r)$ with $\mathcal{A}^{-1}(z) \cap B(u, r) = \{u\}$, the concept of the *index* is useful.

DEFINITION A.4. *Let $\mathcal{A} \in C(\overline{D}, \mathbb{R}^n)$ and let u be an isolated zero. Then the* (topological) index *of $\mathcal{A}$ at u is defined as*

$$i(\mathcal{A}, u) = \deg(\mathcal{A}, B(u, r), z),$$

where $\mathcal{A}^{-1}(z) \cap B(u, r) = \{u\}$.

It follows from the excision property that $i(\mathcal{A}, u)$ is independent of r, so the definition makes sense.

PROPOSITION A.5.

(i) Let $\mathcal{A} \in C(\overline{D}, \mathbb{R}^n)$ such that $z \notin \mathcal{A}(\partial D)$ and $\mathcal{A}^{-1}(z)$ is finite. Then

$$\deg(\mathcal{A}, D, z) = \sum_{u \in \mathcal{A}^{-1}(z)} i(\mathcal{A}, u).$$

(ii) Let $u \in \mathbb{R}^n$ be a zero of $\mathcal{A} \in C(\overline{D}, \mathbb{R}^n)$, where D an open neighborhood of u. If $\mathcal{A}$ is differentiable at u with $\det \mathcal{A}'(u) \neq 0$, then

$$i(\mathcal{A}, u) = \operatorname{sgn} \det \mathcal{A}'(u).$$

The topological degree (resp. index) can be naturally extended to continuous mappings of type $\mathcal{A} : X \to Y$, where X and Y are n-dimensional oriented topological vector spaces and the degree (resp. index) is independent of the choice of the basis. We refer the reader to [14, Chapter 6] for details.

THEOREM A.6. *[14, Theorem 6.3.1] Let $\mathcal{A} : D \to Y$ be continuous, where X and Y are n-dimensional topological vector spaces and $D \subset X$ is an open neighborhood of $u \in X$. Let $h : X \to \mathbb{R}^n$ and $g : Y \to \mathbb{R}^n$ be linear homeomorphisms. Suppose that $\mathcal{A}$ is differentiable at u and that $\ker \mathcal{A}'(u) \neq \{0\}$. Then $i(\mathcal{A}, u) = \operatorname{sgn} \det g\mathcal{A}'(u)h^{-1}$.*

The homotopy invariance property may be extended as follows. Rather than working on the cylinder set $D \times [0, 1]$, we can consider a general $\mathcal{D} \subset \mathbb{R}^n \times [0, 1]$ bounded open set (open with respect to the relative topology). The λ-sections of $\mathcal{D}$ are denoted as $\mathcal{D}_\lambda = \{u : (u, \lambda) \in \mathcal{D})\}$. The λ-section of the boundary $\partial\mathcal{D}$, $(\partial\mathcal{D})_\lambda$, is defined similarly; notice that that $\partial\mathcal{D}_\lambda := \partial(\mathcal{D}_\lambda) \subset (\partial\mathcal{D})_\lambda$ and in general, the inclusion is strict. If a mapping $\mathcal{A} : \mathcal{D} \to \mathbb{R}^n$ satisfies $z \notin \mathcal{A}(\partial\mathcal{D})$, then $z \notin \mathcal{A}(\partial\mathcal{D}_\lambda)$ for all $\lambda \in [0, 1]$.

THEOREM A.7. *[14, Theorem 7.1.1] Let $\mathcal{D} \subset \mathbb{R}^n \times [0, 1]$ be a bounded open set and let $\mathcal{A} : \mathcal{D} \to \mathbb{R}^n$ be continuous. Suppose that $z \notin \mathcal{A}(\partial\mathcal{D})$. Then*

$$\lambda \mapsto \deg(\mathcal{A}(\cdot, \lambda), \mathcal{D}_\lambda, z) \quad \textit{is constant.}$$

The following result is known as the *Leray–Schauder continuation theorem*.

THEOREM A.8. *[14, Theorem 7.2.2] Let $\mathcal{D} \subset \mathbb{R}^n \times [0,1]$ be a bounded open set and let $\mathcal{A} : \mathcal{D} \to \mathbb{R}^n$ be continuous. Suppose that $\mathcal{Z}_1 \cap (\partial\mathcal{D})_1 = \emptyset$ and that*

$$\deg(\mathcal{A}(\cdot, 1), \mathcal{D}_1, z) \neq 0.$$

Then there exists a connected component $\mathcal{C} \subset \mathcal{Z}$ which intersects both $\mathcal{Z}_0 \times \{0\}$ and $(\partial\mathcal{D}) \cup \mathcal{Z}_1 \times \{1\}$.

The following result says the degree is stable under (almost all) small perturbations of the right-hand side of $\mathcal{A}(u) = 0$, and that the degree provides a lower bound for the number of solutions of the perturbed equation.

THEOREM A.9. *[13, Corollary 7.4] If $z \notin \mathcal{A}(\partial D)$, then there is a $\delta > 0$ such that if $z' \in B(z, \delta) \setminus E$, then*

$$\deg(\mathcal{A}, D, z) = \deg(\mathcal{A}, D, z'),$$

where $E = \{w \in B(z, \delta) : \det \mathcal{A}'(u) \neq 0,\ \mathcal{A}(u) = w\}$ is the set of critical values of $\mathcal{A}$ in $B(z, \delta)$. Furthermore, $\mathcal{A}^{-1}(z')$ consists of a finite number m of points, where $|\deg(\mathcal{A}, D, z)| \leq m$ and $m \equiv \deg(\mathcal{A}, D, z)$ (mod 2).

The proof is based on Sard's theorem, which says that E has n-dimensional Lebesgue measure zero.

Next, we recall a tool that is very useful for the computation of the degree in the degenerate case.

THEOREM A.10. *[14, Theorem 6.4.2] Let X and Z be n-dimensional topological vector spaces. Let $L : X \to Z$ be a linear mapping with $\ker L \neq \{0\}$ and $V \subset Z$ be a vector space such that $Z = V \oplus \operatorname{ran} L$. Let $D \subset X$ open and bounded, $r : \overline{D} \to V$ continuous with $0 \notin \mathcal{A}(\partial D)$, where $\mathcal{A} = L + r$. Then, for each invertible linear map $J : \ker L \to V$, and each projector P with $\operatorname{ran} P = \ker L$, the relation*

$$\deg(\mathcal{A}, D, 0) = i(L + JP, 0) \deg(J^{-1} r|_{\ker L}, D \cap \ker L, 0)$$

called Leray's second reduction formula *holds true.*

The preceding theorem is used in the following form.

THEOREM A.11. *[14, Theorem 8.6.1] Let $L : X \to Z$ be a linear mapping with $\ker L \neq \{0\}$ and let Q be a projector with $\ker Q = \operatorname{ran} L$. Also, let $\mathcal{N} : D \times [0,1] \to Z$ be continuous, with $D \subset X$ open and bounded. Furthermore, assume the following hold true.*

(i) *$Lu + \lambda\mathcal{N}(u, \lambda) \neq 0$ for all $u \in \partial D$ and $\lambda \in (0, 1]$.*
(ii) *$Q\mathcal{N}(u, 0) \neq 0$ for each $u \in \partial D$.*

Then

$$\deg(L + \mathcal{N}(\cdot, 1), D, 0) = i(L + Q) \deg(Q\mathcal{N}(\cdot, 0)|_{\ker L}, D \cap \ker L, 0).$$

Proof. We adapt the the proof of [14, Theorem 8.6.1]. Define the homotopy

$$\mathcal{A}(u, \lambda) = Lu + (1 - \lambda)Q\mathcal{N}(u, \lambda) + \lambda\mathcal{N}(u, \lambda).$$

Fix $\lambda \in (0, 1]$. Projecting the equation $\mathcal{A}(u, \lambda) = 0$ with Q and $I - Q$ (i.e. onto the complementary spaces $\operatorname{ran} Q$ and $\operatorname{ran}(I - Q) = \ker Q = \operatorname{ran} L$), we get that $\mathcal{A}(u, \lambda) = 0$ is equivalent to the system $Q\mathcal{N}(u, \lambda) = 0$, $Lu + \lambda(I - Q)\mathcal{N}(u, \lambda) = 0$. But this is equivalent to $Lu + \lambda\mathcal{N}(u, \lambda) = 0$. By assumption (i), $\mathcal{A}(u, \lambda) \neq 0$ for all $u \in \partial D$ and

$\lambda \in (0,1]$. Further, $\mathcal{A}(u,0) = 0$ is equivalent to the system $u \in \ker L$, $Q\mathcal{N}(u,0) = 0$, hence by assumption (ii), $\mathcal{A}(u,0) \neq 0$ for all $u \in \partial D$. Using the homotopy invariance and Leray's second reduction formula with $V = \ker L$, $J = I$ and $r = \mathcal{N}(\cdot,1)$, we get

$$\deg(\mathcal{A}(\cdot,1), D, 0) = \deg(\mathcal{A}(\cdot,0), D, 0) = i(L+Q,0)\deg(Q\mathcal{N}(\cdot,0)|_{\ker L}, D \cap \ker L, 0),$$

which finishes the proof. ∎

An important class of mappings for which the degree behaves rather nicely is the following.

DEFINITION A.12. *Let $U \subset \mathbb{R}^n$ be open.*

(i) *A linear map $L : \mathbb{R}^n \to \mathbb{R}^n$ is said to be* orientation-preserving *if* $\det L \geq 0$.
(ii) *A mapping $\mathcal{A} \in C^2(U, \mathbb{R}^n)$ is said to be* strictly orientation-preserving in U *if*
  (a) $\mathcal{A}'(u)$ is orientation-preserving for all $u \in U$, and
  (b) the set $\{u \in U : \det \mathcal{A}'(u) = 0\}$ is nowhere dense in U.
(iii) *A mapping $\mathcal{A} \in C(U, \mathbb{R}^n)$ is said to be* orientation-preserving *if for every $V \subset U$ open with $\overline{V} \subset U$ is compact, there exists a sequence $\{\mathcal{A}_j\}$ such that $\mathcal{A}_j$ is strictly orientation preserving for all j and $\mathcal{A}_j \to \mathcal{A}$ uniformly on V.*

THEOREM A.13. *[14, Theorem 19.3.1] Let $U \subset \mathbb{R}^n$ be open, $D \subset U$ an open bounded set with $\overline{D} \subset U$ and let $\mathcal{A} : U \to \mathbb{R}^n$ be orientation-preserving. If $z \notin \mathcal{A}(\partial D)$, then the following hold true.*

(i) $\deg(\mathcal{A}, D, z) \geq 0$.
(ii) $\deg(\mathcal{A}, D, z) > 0$ *if and only if* $z \in \mathcal{A}(D)$.
(iii) *If* $\deg(\mathcal{A}, D, z) = 1$, *then* $\mathcal{A}^{-1}(z) \cap D$ *is connected.*
(iv) *If $\mathcal{A}(D)$ contains a point of some component C of $\mathbb{R}^n \setminus \mathcal{A}(\partial D)$, then $C \subset \mathcal{A}(D)$.*

The preceding notions are related to the well-known monotone-type mappings.

THEOREM A.14. *[14, Proposition 19.4.1] Let $U \subset \mathbb{R}^n$ be open, and suppose that $\mathcal{A} : U \to \mathbb{R}^n$ is monotone, i.e.*

$$\langle \mathcal{A}(u) - \mathcal{A}(v), u - v \rangle \geq 0$$

for all $u, v \in U$. Then $\mathcal{A}$ is orientation-preserving.

Appendix B. Topological degree for holomorphic mappings.

Next, we consider the complex case, which, in many respects, is more "rigid" than the real case.

DEFINITION B.1. *Let $U \subset \mathbb{C}^n$ be open. A complex mapping $\mathcal{A} : U \to \mathbb{C}^n$ is said to be* holomorphic at $a \in \mathbb{C}^n$ *if there is a $\mathbb{C}$-linear mapping $L_a : \mathbb{C}^n \to \mathbb{C}^n$ and a mapping $r_a : U \setminus \{a\} \to \mathbb{C}^n$ with $r_a = o(1)$, such that*

$$\mathcal{A}(a+h) = \mathcal{A}(a) + L_a h + r_a(a+h)$$

for all $h \in \mathbb{C}$ such that $a + h \in U$. In this case, $L_a = \mathcal{A}'(a)$, where $\mathcal{A}'(a)h = h_1 \partial_{z_1} \mathcal{A}(a) + \ldots + h_n \partial_{z_n} \mathcal{A}(a)$ and $\partial_{z_k} \mathcal{A}(a)$ denotes the k^{th} complex partial derivative of $\mathcal{A}$ at a.

We now remind the reader of some elementary linear algebra [45]. Every complex vector space V is also vector space over $\mathbb{R}$. This space will be denoted as $V_{\mathbb{R}}$ and called the *realification* of V. The realification of a linear operator $A : V \to W$ over $\mathbb{C}$ is the linear map $A_{\mathbb{R}} : V_{\mathbb{R}} \to W_{\mathbb{R}}$. If $\{e_1, \ldots, e_n\}$ is a basis in V, then

$\{e_1, \dots, e_n, ie_1, \dots, ie_n\}$ is a basis in $V_{\mathbb{R}}$. Further, if $\{e'_1, \dots, e'_m\}$ is a basis in W, and $A = B + iC$ with some real matrices B and C, then the matrix of $A_{\mathbb{R}}$ with respect to the bases $\{e_1, \dots, e_n, ie_1, \dots, ie_n\}$ and $\{e'_1, \dots, e'_n, ie'_1, \dots, ie'_n\}$ is

$$\begin{pmatrix} B & -C \\ C & B \end{pmatrix}.$$

The determinant of the realification $A_{\mathbb{R}}$ obeys the following important rule:

$$\det A_{\mathbb{R}} = |\det A|^2. \tag{B.1}$$

This motivates that we define the realification of a mapping $\mathcal{A} : U \to \mathbb{C}^n$ with $\mathcal{A} = (\mathcal{A}_1, \dots, \mathcal{A}_n)$ as $\mathcal{A}_{\mathbb{R}} : U \to \mathbb{R}^{2n}$ via

$$\mathcal{A}_{\mathbb{R}}(x, y) = (\operatorname{Re} \mathcal{A}_1(z), \operatorname{Im} \mathcal{A}_1(z), \dots, \operatorname{Re} \mathcal{A}_n(z), \operatorname{Im} \mathcal{A}_n(z)),$$

where $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$ and $z = x + iy$. Notice that we do not explicitly denote the realification of the set U.

DEFINITION B.2. *Let $U \subset \mathbb{C}^n$ be open, $\mathcal{A} : U \to \mathbb{C}^n$ be a holomorphic mapping and $D \subset U$ a domain. The* degree of $\mathcal{A}$ in D *is defined as the degree of the realification,* $\deg(\mathcal{A}, D, z) := \deg(\mathcal{A}_{\mathbb{R}}, D, z)$.

It can be proved using (B.1), that if $\mathcal{A} : U \to \mathbb{C}^n$ is holomorphic, then

$$\det \mathcal{A}'_{\mathbb{R}}(x, y) = |\det \mathcal{A}'(z)|^2$$

for all $z \in U$. This innocent-looking identity has striking consequences.

THEOREM B.3. *[14, Section 19.5] Let $U \subset \mathbb{C}^n$ be open, $\mathcal{A} : U \to \mathbb{C}^n$ be a holomorphic mapping and $D \subset U$ bounded and open. Then the following statements hold true.*

(i) *$\mathcal{A}_{\mathbb{R}} : U \to \mathbb{R}^{2n}$ is orientation-preserving. In particular,* $\deg(\mathcal{A}, D, z) \geq 0$ *and* $\deg(\mathcal{A}, D, z) > 0$ *if and only if $z \in \mathcal{A}(D)$.*

(ii) *Let $\overline{D} \subset U$ and $z \notin \mathcal{A}(\partial D)$. If* $\deg(\mathcal{A}, D, z) = k$, *then $\mathcal{A}(u) = z$ has at most k solutions in D. If* $\deg(\mathcal{A}, D, z) = 1$, *then $\mathcal{A}(u) = z$ has a unique solution in D.*

(iii) *Let $\zeta \in D$ be a solution of $\mathcal{A}(\zeta) = z$ and denote by C the connected component of $\mathcal{A}^{-1}(z)$ containing ζ. Then* either *ζ is an isolated solution of $\mathcal{A}(\zeta) = z$ (and hence $C = \{\zeta\}$)* or *$C \cap G \neq \emptyset$ for any neighborhood G of ∂D.*

(iv) *Let $\overline{D} \subset U$ and $z \notin \mathcal{A}(\partial D)$. Then* $\deg(\mathcal{A}, D, z) = 1$ *if and only if there is a unique non-degenerate solution $\zeta \in D$ of $\mathcal{A}(\zeta) = z$.*

(v) *[58, Theorem 42 (b)] The number of zeros in D is finite.*

From (iv) we can conclude that if ζ is a degenerate isolated solution of $\mathcal{A}(\zeta) = z$, then necessarily $i(\mathcal{A}, \zeta) \geq 2$. Further in the holomorphic case, Theorem A.9 can sharpened as follows.

THEOREM B.4. *Let $\mathcal{A} : D \to \mathbb{C}^n$ be a holomorphic mapping and $D \subset \mathbb{C}^n$ bounded and open. If $z \notin \mathcal{A}(\partial D)$, then there is a $\delta > 0$ such that if $z' \in B(z, \delta) \setminus E$, then*

$$\deg(\mathcal{A}, D, z) = \deg(\mathcal{A}, D, z'),$$

where $E = \{w \in B(z, \delta) : \det \mathcal{A}'(u) \neq 0,\ \mathcal{A}(u) = w\}$ is the set of critical values of $\mathcal{A}$ in $B(z, \delta)$. Furthermore, $\mathcal{A}^{-1}(z')$ consists of a finite number m of points, where $m = \deg(\mathcal{A}, D, z)$.

Proof. From Theorem A.9, we have $\deg(\mathcal{A}, D, z) \le m$ and the converse inequality $m \le \deg(\mathcal{A}, D, z') = \deg(\mathcal{A}, D, z)$ follows from Theorem B.3 (ii). ∎

THEOREM B.5. *[58, Theorem 45.] Let $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$ be a real analytic mapping that can be extended to a holomorphic mapping $\widetilde{\mathcal{A}} : \mathbb{C}^n \to \mathbb{C}^n$. Let $D \subset \mathbb{R}^n$ be an open bounded set such that $0 \notin \mathcal{A}(\partial D)$. Let*

$$D_\varepsilon = \{x \in \mathbb{C}^n : \operatorname{Re} x \in D, \quad |\operatorname{Im} x| < \varepsilon\}.$$

Then, for sufficiently small $\varepsilon > 0$,

$$|\deg(\mathcal{A}, D, 0)| \le \deg(\widetilde{\mathcal{A}}, D_\varepsilon, 0), \quad \deg(\mathcal{A}, D, 0) = \deg(\widetilde{\mathcal{A}}, D_\varepsilon, 0) \mod 2.$$

Appendix C. Properties of the excitation graph.

Here, we restrict ourselves to the single-reference case ($M = 1$) and drop the subscript m's from the notation. Recall that K denotes the cardinality of the orbital set Λ. Given $\gamma \in L$, we introduce the set of paths of length n from 0 to γ in G,

$$\mathbb{P}^n(\gamma) = \{\boldsymbol{\alpha} \in L \times \ldots \times L : \text{there is a path } 0 \to \gamma \text{ in } G \text{ having edges } \boldsymbol{\alpha}\}.$$

The following theorem sheds light on the combinatorial structure of the excitation graph.

THEOREM C.1. *Let $G^{\text{full}} = (L, E^{\text{full}})$ be the full SR excitation graph with K orbitals and $2N \le K$ particles. Then the following formulas hold true.*

(i) The number of vertices in G is given by $|L| = \binom{K}{N}$.

(ii) The number of vertices of rank r is $|L(r)| = \binom{N}{r}\binom{K-N}{r}$.

(iii) There are no edges in E^{full} entirely inside $L(r)$, and the number of edges from $L(r)$ to $L(r+s)$ is given by

$$|E(r, r+s)| = \binom{K-N}{r}\binom{K-N-r}{s}\binom{N}{s+r}\binom{s+r}{r},$$

for all $r = 0, 1, \ldots, N$ and $s = 0, \ldots, N-r$, and $|E(r, r+s)| = 0$ if $s = N - r+1, \ldots, N$. Furthermore, the symmetry property $|E(r, r+s)| = |E(s, r+s)|$ holds true.

(iv) The total number of edges is given by

$$|E^{\text{full}}| = \sum_{r=1}^{N} \binom{N}{r}\binom{K-N}{r}\binom{K-2r}{N-r}.$$

(v) The number of directed paths of length $n \le r = \operatorname{rk}(\gamma)$ from 0 to γ is given by $|\mathbb{P}^n(\gamma)| = p(r, n)$, where

$$p(r, n) = \sum_{\substack{r_1+\ldots+r_n=r \\ r_1,\ldots,r_n \ge 1}} \left(\frac{r!}{r_1! \cdots r_n!}\right)^2. \tag{C.1}$$

Proof. (i) is trivial, so is (ii). As for (iii), we enumerate the pairs (α, β) in E^{full} as follows. Fix α with $\operatorname{rk}(\alpha) = r$, then β must satisfy $r + s \le N$, where $\operatorname{rk}(\beta) = s$, so that $|\underline{\alpha} \cup \underline{\beta}| = N$ is possible. In $\underline{\beta}$, we must choose the missing internal letters from $\underline{\alpha}$ and there are r of them. For the remaining $N - s - r$ elements, we may choose freely: there are $\binom{N-r}{N-s-r}$ possibilities to do this. Next, $\overline{\beta}$ must be disjoint from $\overline{\alpha}$, so there

are $M-N-r$ letters to choose from, giving $\binom{M-N-r}{s}$ possibilities. Multiplying these independent choices by the number of ways α can be chosen for fixed r, we get

$$\binom{N}{r}\binom{M-N}{r}\binom{N-r}{N-s-r}\binom{M-N-r}{s} \tag{C.2}$$

for $s = 1, \ldots, N-r$. This can be rewritten using the formula $\binom{n}{h}\binom{n-h}{k} = \binom{n}{k}\binom{n-k}{h}$ as

$$\binom{M-N}{r}\binom{M-N-r}{s}\binom{N}{s+r}\binom{s+r}{r}.$$

Using the aforementioned formula for the first two factors, we also get the desired symmetry property.

Next, to derive (iv) we sum up (C.2),

$$|E^{\text{full}}| = \sum_{r=0}^{N}\sum_{s=1}^{N-r}\binom{N}{r}\binom{M-N}{r}\binom{N-r}{N-s-r}\binom{M-N-r}{s}.$$

Using Vandermonde's identity,

$$\sum_{s=1}^{N-r}\binom{N-r}{N-s-r}\binom{M-N-r}{s} = \binom{M-2r}{N-r} - 1,$$

we get

$$|E^{\text{full}}| = \sum_{r=1}^{N}\binom{N}{r}\binom{M-N}{r}\binom{M-2r}{N-r},$$

where we used Vandermonde's identity once more.

Next, we prove (v). We need to change 0 into γ in n steps (edges). Suppose that the rank-increment of each step is $r_1, \ldots, r_n$, and are such that $r_1 + \ldots + r_n = r$. In the kth step we replace letters $(\alpha_1, \ldots, \alpha_{r_k})$ with $(\beta_1, \ldots, \beta_{r_k})$. These choices can be done independently, so there are $r!^2$ possibilities. However, the order of the α's and β's is irrelevant in each step so we have to divide by $(r_1! \cdots r_n!)^2$. Summing over all $r_1, \ldots, r_n$ gives the stated formula. ∎

Remark C.2.

(i) It follows that the vertex density per rank is hypergeometric,

$$\nu_r = \frac{\binom{N}{r}\binom{K-N}{M-N-r}}{\binom{K}{N}}, \quad \text{where} \quad r = 0, 1, \ldots, N. \tag{C.3}$$

Therefore, its mean is $\frac{N}{K}(K-N)$ and its variance is $\frac{(K-N)^2N^2}{(K-1)K^2}$.

(ii) The formula (C.1) implies that $|\mathbb{P}^n(\gamma)|$ is independent of N and M and is constant for all γ of fixed rank r.

(iii) If S truncation is in effect, we have $p_{\text{S}}(r, n) = r!^2$ if $r = n$ and 0 otherwise.

(iv) For the SD truncation, note that the number of $(r_1, \ldots, r_n)$ tuples with $r_j \in \{1, 2\}$, $r_1 + \ldots + r_n = r$ and $|\{j : r_j = 2\}| = k$ is given by $\binom{n}{k}$ if $r = n + k$ and 0 otherwise. Therefore,

$$p_{\text{SD}}(r, n) = \frac{r!^2}{4^{r-n}}\binom{n}{r-n}.$$

(v) According to the proof of [49, Lemma 4.4.],

$$|\{\beta \in L : \beta \preceq \alpha\}| = \sum_{s=1}^{r-1} \binom{r}{s}\binom{r-1}{r-s},$$

where $r = \mathrm{rk}(\alpha)$.

Appendix D. Optimal choice of multireference determinants.

In this appendix, we describe an algorithm that can be used to automatically determine an optimal set of multireference determinants. Let $J \in \mathbb{N}$ and let

$$\{\gamma_1, \dots, \gamma_J\} \subset L$$

be a fixed set of determinants. Also, fix an excitation rank truncation, e.g. S, SD, SDT, etc. We want to select a *minimal* set of reference elements $\Omega = \{0_1, \dots, 0_M\}$, so that each γ_j is reachable through a *direct* S, SD, SDT, etc. excitation from Ω, this is called "first-order interaction space" in MRCC theory.

Recall that each $\alpha \in 2^\Lambda$ can be represented as a binary characteristic vector $\overrightarrow{\alpha} \in \{0,1\}^K$ such that

$$\overrightarrow{\alpha}^t = \begin{cases} 1 & t \in \alpha \\ 0 & t \notin \alpha \end{cases}$$

The set $\{0,1\}^K$ endowed with the Hamming metric

$$d_{\mathrm{H}}(\overrightarrow{\alpha}, \overrightarrow{\beta}) = |\{t : \overrightarrow{\alpha}^t \neq \overrightarrow{\beta}^t,\ t = 1, \dots, K\}|$$

is a complete metric space, called the Hamming space. The closed balls and the spheres in this space are denoted as $B_{\mathrm{H}}(\overrightarrow{\alpha}, R)$ and $S_{\mathrm{H}}(\overrightarrow{\alpha}, R)$. Using this language, L is simply $S_{\mathrm{H}}(\overrightarrow{0}, N)$, where $\overrightarrow{0} = (0, \dots, 0)$.[16] Further,

$$\mathrm{rk}_m(\alpha) = \frac{1}{2} d_{\mathrm{H}}(\overrightarrow{0_m}, \overrightarrow{\alpha})$$

for any $m = 1, \dots, M$. Notice that $d_{\mathrm{H}}(\overrightarrow{\alpha}, \overrightarrow{\beta}) \geq 2$ for distinct $\overrightarrow{\alpha}, \overrightarrow{\beta} \in S_{\mathrm{H}}(\overrightarrow{0}, N)$.

This way, our optimization problem may be formulated as a covering problem in Hamming space. Let ρ denote the excitation rank truncation, e.g. $\rho = 1, 2, 3, \dots$ for S, SD, SDT, etc. Fix $J \in \mathbb{N}$ and $\Gamma = \{\overrightarrow{\gamma}_1, \dots, \overrightarrow{\gamma}_J\} \subset S_{\mathrm{H}}(\overrightarrow{0}, N)$. We need to find a minimal set of Hamming balls $\{B_{\mathrm{H}}(\overrightarrow{0_m}, 2\rho) : m = 1, \dots, M\}$ with $\overrightarrow{0_m} \in S_{\mathrm{H}}(\overrightarrow{0}, N)$ such that

$$\Gamma \subset \bigcup_{m=1}^{M} B_{\mathrm{H}}(\overrightarrow{0_m}, 2\rho) \cap S_{\mathrm{H}}(\overrightarrow{0}, N).$$

Obviously, $\overrightarrow{0_m} \in \Gamma_{2\rho}$, where

$$\Gamma_{2\rho} = \bigcup_{j=1}^{J} B_{\mathrm{H}}(\overrightarrow{\gamma_j}, 2\rho) \cap S_{\mathrm{H}}(\overrightarrow{0}, N).$$

In other words, it is sufficient to look for the $\overrightarrow{0_m}$'s in the much smaller set $\Gamma_{2\rho}$. Let $n = |\Gamma_{2\rho}|$, and introduce some indexing in $\Gamma_{2\rho}$, say $\Gamma_{2\rho} = \{\overrightarrow{\alpha}_1, \dots, \overrightarrow{\alpha}_n\}$. The

[16] We warn the reader that the notation $\overrightarrow{0_m}$ for the vector representation of 0_m is slightly colliding with $\overrightarrow{0}$, the actual zero vector for the Hamming space.

geometric form of the covering problem may be rephrased as a binary integer linear program (BILP) [52],

$$\left.\begin{aligned} \sum_{\nu=1}^{n} \mathbf{c}_\nu \mathbf{x}_\nu &\to \min! \\ \sum_{\substack{\overrightarrow{\gamma}_j \in B_{\mathrm{H}}(\overrightarrow{\alpha}_\nu, 2\rho) \\ 1 \le \nu \le n}} \mathbf{x}_\nu &\ge 1, \quad j = 1, \dots, J \\ \mathbf{x} &\in \{0,1\}^n \end{aligned}\right\}$$

where $\mathbf{c} \in \mathbb{Q}^n$ is a given rational cost vector.

Remark D.1. If $\mathbf{c}_\nu = 0$ for some ν, then we will automatically have $\mathbf{x}_\nu = 1$ in the solution, even if $B_{\mathrm{H}}(\overrightarrow{\alpha}_\nu, 2\rho)$ does not cover. On the other hand, assigning a larger (resp. infinite) cost $\mathbf{c}_\nu$ will likely (resp. surely) end up $\mathbf{x}_\nu = 0$ in the solution.

The above problem is called a "multidimensional knapsack problem" in the optimization community, which seems to be extensively studied. However, we just naively solve the BILP using general ILP methods available in *Mathematica*. In our experience, the BILP can be built up and solved in a small amount of time for practically relevant parameters N, K and J, even on an older machine.

The reason for the apparent efficiency might be that the number of variables n in the the BILP above is significantly less then $|L| = \binom{K}{N}$. In fact, using the binary entropy function $H(x) = -x \log_2 x - (1-x)\log_2(1-x)$, we have the rough estimate

$$\frac{n}{|L|} \le \frac{J|B_{\mathrm{H}}(\overrightarrow{0}, 2\rho)|}{|B_{\mathrm{H}}(\overrightarrow{0}, N)|} \le J\sqrt{8K\lambda'(1-\lambda')}\, 2^{-K(H(\lambda')-H(\lambda))},$$

where $\lambda = 2\rho/K$ and $\lambda' = N/K$ valid for $0 < \lambda, \lambda' < \frac{1}{2}$ [12, Lemma 2.4.4]. Notice that $H(\lambda') \ge H(\lambda)$, so $n/|L| \to 0$ as $K \to \infty$.

Appendix E. A short proof of the Kowalski–Piecuch theorem. To close this section, we present the main theorem[17] of Kowalski and Piecuch [43] in a somewhat different form. Define the *KP energy* as

$$\mathcal{E}_{\mathrm{KP}}(t^1, \lambda) = \langle \mathcal{H}(t^0 + \lambda t^{\angle})\Phi_0, \Phi_0\rangle = \langle \mathcal{H} e^{T^0 + \lambda T^{\angle}}\Phi_0, \Phi_0\rangle,$$

so that $\mathcal{E}_{\mathrm{KP}}(t^1, 0) = \mathcal{E}_{\mathrm{CC}}(t^0)$ and $\mathcal{E}_{\mathrm{KP}}(t^1, 1) = \mathcal{E}_{\mathrm{CC}}(t^1)$. If $\rho \ge 2$, then according to (2.13), we simply have

$$\mathcal{E}_{\mathrm{KP}}(t^1, \lambda) = \langle \mathcal{H}(t^0)\Phi_0, \Phi_0\rangle = \langle \mathcal{H} e^{T^0}\Phi_0, \Phi_0\rangle, \tag{E.1}$$

although we will not exploit this property in the proof.

THEOREM E.1 (Kowalski–Piecuch).

(I) *Suppose that* $\Psi = (c_0 I + C^0)\Phi_0 \in \mathfrak{H}_K^1$, *with* $c_0 \in \mathbb{R}$ *and* $c^0 \in \mathbb{V}^0$, *satisfies the (weak) Schrödinger equation*

$$\langle \mathcal{H}\Psi, \Phi\rangle = \mathcal{E}\langle \Psi, \Phi\rangle \quad \textit{for all} \quad \Phi \in \mathfrak{H}_K^1 \tag{E.2}$$

for some $\mathcal{E} \in \mathbb{R}$ *and that*

$$\langle e^{T_{**}^0(\lambda) + \lambda T_{**}^{\angle}(\lambda)}\Phi_0, \Psi\rangle \ne 0, \tag{E.3}$$

[17] They call it the "Fundamental Theorem of β-Nested Equation Formalism".

where $t^1_{**}(\lambda) \in \mathbb{V}^1$ *is a zero of* $\mathcal{K}_{\rm KP}(\cdot,\lambda)$ *for* all $\lambda \in [0,1]$. *Then*

$$\mathcal{E}_{\rm KP}(t^1_{**}(\lambda),\lambda) \equiv \mathcal{E} \quad \textit{for all} \quad \lambda \in [0,1].$$

(II) *Suppose that* $\Psi = (c_0 I + C^1)\Phi_0$, *with* $c_0 \in \mathbb{R}$ *and* $c^1 \in \mathbb{V}^1$, *satisfies* (E.2) *for some* $\mathcal{E} \in \mathbb{R}$. *If* (E.3) *holds true for* some $\lambda \in [0,1]$ *and* $t^1_{**} = t^1_{**}(\lambda) \in \mathbb{V}^1$ *with* $\mathcal{K}_{\rm KP}(t^1_{**},\lambda) = 0$, *then*

$$\begin{aligned}\mathcal{E}_{\rm KP}(t^1_{**}(\lambda),\lambda) - \mathcal{E} &= \frac{\langle (\mathcal{H}(t^1_{**}) - \mathcal{H}(t^0_{**} + \lambda t^{\angle}_{**}))\Phi_0, \Pi_{\mathfrak{Y}^{\angle}}(e^{T^0_{**}+\lambda T^{\angle}_{**}})^\dagger C^{\angle}\Phi_0\rangle}{\langle e^{T^0_{**}+\lambda T^{\angle}_{**}}\Phi_0, \Psi\rangle} \\ &= (1-\lambda)\frac{\langle \Gamma(t^1_{**},\lambda)\Phi_0, \Pi_{\mathfrak{Y}^{\angle}}(e^{T^0_{**}+\lambda T^{\angle}_{**}})^\dagger C^{\angle}\Phi_0\rangle}{\langle e^{T^0_{**}+\lambda T^{\angle}_{**}}\Phi_0, \Psi\rangle},\end{aligned}$$

where $\Gamma(t^1,\lambda)$ *is given by* (E.6) *below.*

Furthermore, in case the energy blows up, we have the following.

THEOREM E.2. *Suppose that* $\Psi = (c_0 I + C^1)\Phi_0$, *with* $c_0 \in \mathbb{R}$ *and* $c^1 \in \mathbb{V}^1$, *satisfies* (E.2) *for some* $\mathcal{E} \in \mathbb{R}$. *Furthermore, assume that* $t^1_{**}(\lambda) \in \mathbb{V}^1$ *is a zero of* $\mathcal{K}_{\rm KP}(\cdot,\lambda)$ *for all* λ *in a neighborhood of some* $\lambda_0 \in [0,1]$. *If* $|\mathcal{E}_{\rm KP}(t^1_{**}(\lambda),\lambda)| \to \infty$ *as* $\lambda \to \lambda_0$ *and*

$$\langle \Gamma(t^1_{**}(\lambda),\lambda)\Phi_0, \Pi_{\mathfrak{Y}^{\angle}}(e^{T^0_{**}+\lambda T^{\angle}_{**}})^\dagger C^{\angle}\Phi_0\rangle = \mathcal{O}\left(\frac{1}{1-\lambda}\right) \quad (\lambda \to \lambda_0), \tag{E.4}$$

then

$$\langle e^{T^0_{**}(\lambda)+\lambda T^{\angle}_{**}(\lambda)}\Phi_0, \Psi\rangle \to 0 \quad (\lambda \to \lambda_0).$$

Part (I) of Theorem E.1 says that if one can solve the Schrödinger equation *exactly* on $\mathbb{V}^0$ for an eigenvalue $\mathcal{E}$ and (E.3) holds true for a zero $t^1_{**}(\lambda) \in \mathbb{V}^1$ of $\mathcal{K}_{\rm KP}(\cdot,\lambda)$ for all λ, then the KP energy $\mathcal{E}_{\rm KP}(t^1_{**}(\lambda),\lambda)$ is identically $\mathcal{E}$. Notice that $t^1_{**}(1)$ is not required to represent Ψ, i.e. $e^{T^1_{**}(1)}\Phi_0 \neq \Psi$ is allowed. Also, no regularity of the trajectory $\lambda \mapsto t^1_{**}(\lambda)$ is demanded.

Part (II) stipulates that the Schrödinger equation can be solved on $\mathbb{V}^1$ with an eigenvalue $\mathcal{E}$ and that the nonorthogonality condition (E.3) holds true for some $t^1_{**} \in \mathbb{V}^1$ zero of $\mathcal{K}_{\rm KP}(\cdot,\lambda)$ for some λ. Then, the error in the energy can be expressed by the stated formula. If one assumes the hypothesis for all $\lambda \in [0,1]$ in a neighborhood of 1, then we can conclude that the KP energy $\mathcal{E}_{\rm KP}(t^1_{**}(\lambda),\lambda)$ tends to $\mathcal{E}$ smoothly, as $\lambda \to 1$. Again, no regularity of $\lambda \mapsto t^1_{**}(\lambda)$ is needed.

Finally, Theorem E.2 considers the case when the KP energy diverges as $\lambda \to \lambda_0$. Assuming the growth condition (E.4), we can conclude that the KP solution $e^{T^0_{**}(\lambda)+\lambda T^{\angle}_{**}(\lambda)}\Phi_0$ becomes orthogonal to the eigenstate Ψ.

For the proof, we need the following lemma which recasts the KP equations in an "unlinked" form.

LEMMA E.3. *Suppose that* $t^1_{**} = t^1_{**}(\lambda) \in \mathbb{V}^1$ *is such that* $\mathcal{K}_{\rm KP}(t^1_{**},\lambda) = 0$ *for some* $\lambda \in [0,1]$. *Then,*

$$\langle (\mathcal{H} - \mathcal{E}_{\rm KP}(t^1_{**},\lambda))e^{T^0_{**}+\lambda T^{\angle}_{**}}\Phi_0, S^1\Phi_0\rangle = -\langle \mathcal{G}(t^1_{**},\lambda)\Phi_0, \Pi_{\mathfrak{Y}^{\angle}}(e^{T^0_{**}+\lambda T^{\angle}_{**}})^\dagger S^{\angle}\Phi_0\rangle, \tag{E.5}$$

where

$$\mathcal{G}(t^1_{**},\lambda) = \mathcal{H}(t^1_{**}) - \mathcal{H}(t^0_{**} + \lambda t^{\angle}_{**}).$$

Moreover, $\mathcal{G}(t^1_{**},\lambda) = (1-\lambda)\Gamma(t^1_{**},\lambda)$*, where*

$$\Gamma(t^1_{**},\lambda) = \sum_{k=1}^{2N} \frac{(1-\lambda)^{k-1}}{k!} e^{-(T^0_{**}+\lambda T^{\angle}_{**})} [\mathcal{H}, T^{\angle}_{**}]_{(k)} e^{T^0_{**}+\lambda T^{\angle}_{**}}. \tag{E.6}$$

Proof. Assume that $\mathcal{K}_{\mathrm{KP}}(t^1_{**},\lambda) = 0$, so using (5.33) we have

$$\langle \mathcal{H}(t^0_{**} + \lambda t^{\angle}_{**})\Phi_0, S^0\Phi_0\rangle + \langle \mathcal{H}(t^1_{**})\Phi_0, S^{\angle}\Phi_0\rangle = 0 \quad \text{for all} \quad s^1 \in \mathbb{V}^1.$$

This can be rewritten as

$$\langle \mathcal{H}(t^0_{**} + \lambda t^{\angle}_{**})\Phi_0, S^1\Phi_0\rangle = -\langle (\mathcal{H}(t^1_{**}) - \mathcal{H}(t^0_{**} + \lambda t^{\angle}_{**}))\Phi_0, S^{\angle}\Phi_0\rangle \quad \text{for all} \quad s^1 \in \mathbb{V}^1,$$

or,

$$\langle \mathcal{H}(t^0_{**} + \lambda t^{\angle}_{**})\Phi_0, S^1\Phi_0\rangle = -\langle \mathcal{G}(t^1_{**},\lambda)\Phi_0, S^{\angle}\Phi_0\rangle \quad \text{for all} \quad s^1 \in \mathbb{V}^1.$$

Then we can write

$$\begin{aligned}
&\langle (\mathcal{H} - \mathcal{E}_{\mathrm{KP}}(t^1_{**},\lambda)) e^{T^0_{**}+\lambda T^{\angle}_{**}}\Phi_0, S^1\Phi_0\rangle \\
&= \langle e^{-(T^0_{**}+\lambda T^{\angle}_{**})}(\mathcal{H} - \mathcal{E}_{\mathrm{KP}}(t^1_{**},\lambda)) e^{T^0_{**}+\lambda T^{\angle}_{**}}\Phi_0, (e^{T^0_{**}+\lambda T^{\angle}_{**}})^{\dagger} S^1\Phi_0\rangle \\
&= \langle e^{-(T^0_{**}+\lambda T^{\angle}_{**})}\mathcal{H} e^{T^0_{**}+\lambda T^{\angle}_{**}}\Phi_0, \Pi_{\mathfrak{V}^1}(e^{T^0_{**}+\lambda T^{\angle}_{**}})^{\dagger} S^1\Phi_0\rangle \\
&= -\langle \mathcal{G}(t^1_{**},\lambda)\Phi_0, \Pi_{\mathfrak{V}^{\angle}}(e^{T^0_{**}+\lambda T^{\angle}_{**}})^{\dagger} S^1\Phi_0\rangle \\
&= -\langle \mathcal{G}(t^1_{**},\lambda)\Phi_0, \Pi_{\mathfrak{V}^{\angle}}(e^{T^0_{**}+\lambda T^{\angle}_{**}})^{\dagger} S^{\angle}\Phi_0\rangle
\end{aligned}$$

for all $s^1 \in \mathbb{V}^1$. In the last step we used that $\Pi_{\mathfrak{V}^{\angle}}(e^{T^0_{**}+\lambda T^{\angle}_{**}})^{\dagger}$ maps $\mathfrak{V}^0$ to zero. The "moreover" part is a simple expansion using (5.9),

$$\begin{aligned}
\mathcal{G}(t^1_{**},\lambda) &= e^{-(T^0_{**}+\lambda T^{\angle}_{**})}(e^{-(1-\lambda)T^{\angle}_{**}}\mathcal{H} e^{(1-\lambda)T^{\angle}_{**}} - \mathcal{H}) e^{T^0_{**}+\lambda T^{\angle}_{**}} \\
&= \sum_{k=1}^{2N} \frac{(1-\lambda)^k}{k!} e^{-(T^0_{**}+\lambda T^{\angle}_{**})} [\mathcal{H}, T^{\angle}_{**}]_{(k)} e^{T^0_{**}+\lambda T^{\angle}_{**}}.
\end{aligned}$$

∎

Proof of Theorem E.1. We have using Lemma E.3,

$$\begin{aligned}
0 &= \langle (\mathcal{H} - \mathcal{E}_{\mathrm{KP}}(t^1_{**}(\lambda),\lambda)) e^{T^0_{**}(\lambda)+\lambda T^{\angle}_{**}(\lambda)}\Phi_0, (c_0 I + C^0)\Phi_0\rangle \\
&= (\mathcal{E} - \mathcal{E}_{\mathrm{KP}}(t^1_{**}(\lambda),\lambda))\langle e^{T^0_{**}(\lambda)+\lambda T^{\angle}_{**}(\lambda)}\Phi_0, \Psi\rangle.
\end{aligned}$$

Similarly, for part (II)

$$\begin{aligned}
&\langle (\mathcal{H}(t^1_*) - \mathcal{H}(t^0_{**} + \lambda t^{\angle}_{**}))\Phi_0, \Pi_{\mathfrak{V}^{\angle}}(e^{T^0_{**}+\lambda T^{\angle}_{**}})^{\dagger} C^{\angle}\Phi_0\rangle \\
&\qquad = \langle (\mathcal{H} - \mathcal{E}_{\mathrm{KP}}(t^1_{**}(\lambda),\lambda)) e^{T^0_{**}+\lambda T^{\angle}_{**}}\Phi_0, (c_0 I + C^0 + C^{\angle})\Phi_0\rangle \\
&\qquad = (\mathcal{E} - \mathcal{E}_{\mathrm{KP}}(t^1_{**}(\lambda),\lambda))\langle e^{T^0_{**}+\lambda T^{\angle}_{**}}\Phi_0, \Psi\rangle.
\end{aligned}$$

Theorem E.2 also follows from the previous equality. ∎

REFERENCES

[1] R. A. Adams and J. J. Fournier, *Sobolev spaces*, Elsevier, 2003.

[2] J. S. Arponen, *Variational principles and linked-cluster exp S expansions for static and dynamic many-body problems*, Ann. Phys., 151 (1983), pp. 311–382.
[3] V. Bach, *Error bound for the Hartree-Fock energy of atoms and molecules*, Communications in mathematical physics, 147 (1992), pp. 527–548.
[4] V. Bach, E. H. Lieb, M. Loss, and J. P. Solovej, *There are no unfilled shells in unrestricted Hartree-Fock theory*, in The Stability of Matter: From Atoms to Stars, Springer, 1997, pp. 309–311.
[5] R. J. Bartlett, S. A. Kucharski, and J. Noga, *Alternative coupled-cluster ansätze II. The unitary coupled-cluster method*, Chemical physics letters, 155 (1989), pp. 133–140.
[6] R. J. Bartlett and M. Musiał, *Coupled-cluster theory in quantum chemistry*, Reviews of Modern Physics, 79 (2007), p. 291.
[7] R. Bishop, *An overview of coupled cluster theory and its applications in physics*, Theoretica chimica acta, 80 (1991), pp. 95–148.
[8] C. Bloch, *Sur la théorie des perturbations des états liés*, Nuclear Physics, 6 (1958), pp. 329–347.
[9] A. Buffa, *Remarks on the discretization of some noncoercive operator with applications to heterogeneous Maxwell equations*, SIAM Journal on Numerical Analysis, 43 (2005), pp. 1–18.
[10] A. Buffa, R. Hiptmair, T. von Petersdorff, and C. Schwab, *Boundary element methods for Maxwell transmission problems in lipschitz domains*, Numerische Mathematik, 95 (2003), pp. 459–485.
[11] E. Cances and C. Le Bris, *On the convergence of SCF algorithms for the Hartree-Fock equations*, ESAIM: Mathematical Modelling and Numerical Analysis, 34 (2000), pp. 749–774.
[12] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering codes*, Elsevier, 1997.
[13] J. Cronin, *Fixed points and topological degree in nonlinear analysis*, vol. 11, American Mathematical Soc., 1995.
[14] G. Dinca and J. Mawhin, *Brouwer degree and applications*, preprint, (2009).
[15] H. Eschrig, *The fundamentals of density functional theory*, vol. 32, Springer, 1996.
[16] F. M. Faulstich, A. Laestadius, O. Legeza, R. Schneider, and S. Kvaal, *Analysis of the tailored coupled-cluster method in quantum chemistry*, SIAM Journal on Numerical Analysis, 57 (2019), pp. 2579–2607.
[17] F. M. Faulstich, M. Máté, A. Laestadius, M. A. Csirik, L. Veis, A. Antalik, J. Brabec, R. Schneider, J. Pittner, S. Kvaal, et al., *Numerical and theoretical aspects of the DMRG-TCC method exemplified by the nitrogen dimer*, Journal of chemical theory and computation, 15 (2019), pp. 2206–2220.
[18] G. Friesecke, *The multiconfiguration equations for atoms and molecules: charge quantization and existence of solutions*, Archive for rational mechanics and analysis, 169 (2003), pp. 35–71.
[19] J. Geertsen, M. Rittby, and R. J. Bartlett, *The equation-of-motion coupled-cluster method: Excitation energies of Be and CO*, Chemical Physics Letters, 164 (1989), pp. 57–62.
[20] S. J. Gustafson and I. M. Sigal, *Mathematical concepts of quantum mechanics*, Springer Science & Business Media, 2011.
[21] T. Helgaker, P. Jorgensen, and J. Olsen, *Molecular electronic-structure theory*, John Wiley & Sons, 2014.
[22] P. D. Hislop and I. M. Sigal, *Introduction to spectral theory: With applications to Schrödinger operators*, vol. 113, Springer Science & Business Media, 2012.
[23] K. Jankowski and K. Kowalski, *Physical and mathematical content of coupled-cluster equations. II. on the origin of irregular solutions and their elimination via symmetry adaptation*, The Journal of chemical physics, 110 (1999), pp. 9345–9352.
[24] K. Jankowski and K. Kowalski, *Physical and mathematical content of coupled-cluster equations. IV. impact of approximations to the cluster operator on the structure of solutions*, The Journal of chemical physics, 111 (1999), pp. 2952–2959.
[25] K. Jankowski, K. Kowalski, I. Grabowski, and H. Monkhorst, *Correspondence between physical states and solutions to the coupled-cluster equations*, International journal of quantum chemistry, 75 (1999), pp. 483–496.
[26] K. Jankowski, K. Kowalski, and P. Jankowski, *Multiple solutions of the single-reference coupled-cluster equations. I. H4 model revisited*, International Journal of Quantum Chemistry, 50 (1994), pp. 353–367.
[27] B. Jeziorski and H. J. Monkhorst, *Coupled-cluster method for multideterminantal reference states*, Physical Review A, 24 (1981), p. 1668.

[28] K. Kowalski, *Properties of coupled-cluster equations originating in excitation sub-algebras*, The Journal of Chemical Physics, 148 (2018), p. 094104.
[29] K. Kowalski and K. Jankowski, *Full solution to the coupled-cluster equations: the H4 model*, Chemical physics letters, 290 (1998), pp. 180–188.
[30] H. Kümmel, K. H. Lührmann, and J. G. Zabolitzky, *Many-fermion theory in exps-(or coupled cluster) form*, Physics Reports, 36 (1978), pp. 1–63.
[31] A. Laestadius and F. M. Faulstich, *The coupled-cluster formalism–a mathematical perspective*, Molecular Physics, 117 (2019), pp. 2362–2373.
[32] A. Laestadius and S. Kvaal, *Analysis of the extended coupled-cluster method in quantum chemistry*, SIAM Journal on Numerical Analysis, 56 (2018), pp. 660–683.
[33] M. Lewin, *Existence of Hartree–Fock excited states for atoms and molecules*, Letters in Mathematical Physics, 108 (2018), pp. 985–1006.
[34] M. Lewin, *Semi-classical limit of the Levy–Lieb functional in density functional theory*, Comptes Rendus Mathematique, 356 (2018), pp. 449–455.
[35] M. Lewin, E. H. Lieb, and R. Seiringer, *The local density approximation in density functional theory*, Pure and Applied Analysis, 2 (2020), pp. 35–73.
[36] E. H. Lieb, *Density functionals for Coulomb systems*, International Journal of Quantum Chemistry, 24 (1983), pp. 243–277.
[37] E. H. Lieb and M. Loss, *Analysis*, in Amer. Math. Soc, 2001.
[38] E. H. Lieb and R. Seiringer, *The stability of matter in quantum mechanics*, Cambridge University Press, 2010.
[39] E. H. Lieb and B. Simon, *On solutions to the Hartree-Fock problem for atoms and molecules*, The Journal of Chemical Physics, 61 (1974), pp. 735–736.
[40] P.-L. Lions, *Solutions of Hartree-Fock equations for Coulomb systems*, Communications in Mathematical Physics, 109 (1987), pp. 33–97.
[41] D. O'Regan, Y. J. Cho, and Y.-Q. Chen, *Topological degree theory and applications*, CRC Press, 2006.
[42] J. Paldus, M. Takahashi, and B. Cho, *Degeneracy and coupled-cluster approaches*, International Journal of Quantum Chemistry, 26 (1984), pp. 237–244.
[43] P. Piecuch and K. Kowalski, *In search of the relationship between multiple solutions characterizing coupled-cluster theories*, in Computational chemistry: reviews of current trends, World Scientific, 2000, pp. 1–104.
[44] P. Piecuch, S. Zarrabian, J. Paldus, and J. Čížek, *Coupled-cluster approaches with an approximate account of triexcitations and the optimized-inner-projection technique. II. coupled-cluster results for cyclic-polyene model systems*, Physical Review B, 42 (1990), p. 3351.
[45] V. V. Prasolov, *Problems and theorems in linear algebra*, vol. 134, American Mathematical Soc., 1994.
[46] M. Reed and B. Simon, *Methods of modern mathematical physics I: Functional analysis*, vol. 1, Elsevier, 1972.
[47] M. Reed and B. Simon, *Methods of modern mathematical physics II: Fourier Analysis, Self-Adjointness*, vol. 2, Elsevier, 1975.
[48] M. Reed and B. Simon, *Methods of modern mathematical physics IV: Analysis of Operators*, vol. 4, Elsevier, 1978.
[49] T. Rohwedder, *The continuous Coupled Cluster formulation for the electronic Schrödinger equation*, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique, 47 (2013), pp. 421–447.
[50] T. Rohwedder and R. Schneider, *Error estimates for the coupled cluster method*, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique, 47 (2013), pp. 1553–1582.
[51] R. Schneider, *Analysis of the projected coupled cluster method in electronic structure calculation*, Numerische Mathematik, 113 (2009), pp. 433–471.
[52] A. Schrijver, *Theory of linear and integer programming*, John Wiley & Sons, 1998.
[53] I. Shavitt and R. J. Bartlett, *Many-body methods in chemistry and physics: MBPT and coupled-cluster theory*, Cambridge university press, 2009.
[54] J. P. Solovej, *The ionization conjecture in Hartree-Fock theory*, Annals of mathematics, (2003), pp. 509–576.
[55] J. P. Solovej, *Many body quantum mechanics*, Lecture Notes., (2007), http://www.mathematik.uni-muenchen.de/~schaub/solovej_skript_a4.pdf.
[56] S. Wilson and G. H. Diercksen, *Methods in computational molecular physics*, vol. 293, Springer Science & Business Media, 1991.
[57] H. Yserentant, *Regularity and approximability of electronic wave functions*, Springer, 2010.

[58] P. Zabrejko, *Rotation of vector fields: definition, basic properties, and calculation*, in Topological nonlinear analysis II, Springer, 1997, pp. 445–601.
[59] E. Zeidler and P. R. Wadsack, *Nonlinear functional analysis and its applications: Fixed-point theorems/transl. by Peter R. Wadsack*, Springer-Verlag, 1993.
[60] T. P. Živković, *Existence and reality of solutions of the coupled-cluster equations*, International Journal of Quantum Chemistry, 12 (1977), pp. 413–420.
[61] T. P. Živković and H. J. Monkhorst, *Analytic connection between configuration–interaction and coupled-cluster solutions*, Journal of Mathematical Physics, 19 (1978), pp. 1007–1022.