SIAM J. NUMER. ANAL.
Vol. 56, No. 2, pp. 660–683




# ANALYSIS OF THE EXTENDED COUPLED-CLUSTER METHOD IN QUANTUM CHEMISTRY*

ANDRE LAESTADIUS† AND SIMEN KVAAL†

**Abstract.** The mathematical foundation of the so-called extended coupled-cluster method for the solution of the many-fermion Schrödinger equation is here developed. We prove an existence and uniqueness result, both in the full infinite-dimensional amplitude space as well as for discretized versions of it. The extended coupled-cluster method is formulated as a critical point of an energy function using a generalization of the Rayleigh–Ritz principle: the bivariational principle. This gives a quadratic bound for the energy error in the discretized case. The existence and uniqueness results are proved using a type of monotonicity property for the flipped gradient of the energy function. A comparison to the analysis of the standard coupled-cluster method is made, and it is argued that the bivariational principle is a useful tool, both for studying coupled-cluster type methods and for developing new computational schemes in general.

**Key words.** quantum chemistry, coupled-cluster method, extended coupled-cluster method, bivariational principle, uniqueness and existence, error estimates

**AMS subject classifications.** 65Z05, 81-08, 81V55

**DOI.** 10.1137/17M1116611

## 1. Introduction.

The coupled-cluster (CC) method is today the *de facto* standard wavefunction-based method for electronic-structure calculations and has a complex and interesting history [14, 11, 4, 2]. To cut a long story short, it was invented by Coester and Kümmel in the 1950s as a method for dealing with the strong correlations inside an atomic nucleus [5, 6]. From nuclear physics, the idea migrated to the field of quantum chemistry in the 1960s due to the seminal work of researchers such as Sinanoğlu, Čížek, Paldus, and Shavitt [19, 3, 15]. An interesting turn of events is that the method returned to nuclear physics in the 1990s, when Dean and Hjorth-Jensen applied the now mature methodology to nuclear structure calculations [7].

The main feature of the CC method is the use of an exponential parametrization of the wavefunction. This ensures proper scaling of the computed energy with system size (number of particles), i.e., the method is *size extensive*. At the same time, the CC method is only *polynomially scaling* with respect to system size. These factors have led to the popularity of the method.

However, the theory does not satisfy the (Rayleigh–Ritz) variational principle, i.e., the computed CC energy is not guaranteed to be an upper bound to the exact energy. This has traditionally been the main criticism of CC calculations, as an error estimate is not readily available. Furthermore, in the original formulation it was not *variational* in the sense that the solution was not formulated as a stationary point of some function(al).

*Received by the editors February 14, 2017; accepted for publication (in revised form) January 16, 2018; published electronically March 15, 2018.
http://www.siam.org/journals/sinum/56-2/M111661.html
**Funding:** This work has received funding from ERC-STG-2014 under grant agreement 639508 and from the Research Council of Norway (RCN) under CoE grant 179568/V30 (CTCC).
†Hylleraas Centre for Quantum Molecular Sciences, Department of Chemistry, University of Oslo, P.O. Box 1033 Blindern, N-0315 Oslo, Norway (andre.laestadius@kjemi.uio.no, simen.kvaal@kjemi.uio.no).

Helgaker and Jørgensen later formulated the CC method in terms of a Lagrangian [9, 10], viewing the solution of the CC amplitude equations as a constrained optimization of the energy, the set of cluster amplitude equations becoming constraints. This is today the standard formulation of the CC method.

Already in 1983, Arponen [1] had derived the so-called extended CC (ECC) method from a generalization of the Rayleigh–Ritz variational principle, the *bivariational principle*. This principle formally relaxes the condition of the Hamiltonian being symmetric and thus introduces the left eigenvector as a variable as well as the right eigenvector. Arponen noted that the standard CC method can be viewed as an approximation to the ECC method and continued to write down the standard CC Lagrangian. In the bivariational interpretation, Helgaker and Jørgensen's Lagrange multipliers are actually wavefunction parameters on equal footing with the cluster amplitudes. No distinction is being made.

Both Helgaker and Jørgensen's CC Lagrangian and Arponen's bivariational formulation cast CC theory in a *variational* (stationary point) setting. However, only the bivariational point of view allows, at least formally, systematic improvement by adding other degrees of freedom than the cluster amplitudes to the ansatz. The bivariational principle is therefore of potential great use when developing novel wavefunction-based methods; see, for example, [12], where the single-particle functions are introduced as (bi)variational parameters in a time-dependent setting. However, while the bivariational principle is rigorous, it is not known how to introduce *approximations* by parameterizations of the wavefunctions, such that one can obtain existence and uniqueness results as well as error estimates.

In this article, we will provide a rigorous analysis of a version of the ECC method. The idea is, starting from the bivariational quotient, to choose a function $\mathcal{F}$ (see (7)) that is (locally and strongly) monotone and where $\mathcal{F} = 0$ is equivalent to a critical point of the bivariational quotient. Until now, the ECC method has not been turned into a practical tool in chemistry due to its complexity. On the other hand, the analysis herein is a step toward obtaining a rigorous foundation for the application of the bivariational principle. We believe that the approach taken, by showing the monotonicity of the flipped gradient $\mathcal{F}$, is an approach that may allow existence and uniqueness results in much more general settings.

We build our analysis on articles by Rohwedder and Schneider, who fairly recently put the standard CC method on sound mathematical ground [18, 16, 17]. They proved, among other important results, a uniqueness and existence result of the solution of the CC amplitude equations. The result rests on a certain monotonicity property of the CC equations. Moreover, in [16] the boundedness of cluster operators (as operators on a Hilbert space that guarantees finite kinetic energy) was established, which turns out to be a rather subtle matter. They also provided error estimates for the energy using the stationarity condition of the Lagrangian.

This article is structured as follows. In section 2 we discuss the solution of the Schrödinger equation (SE) by employing an exponential ansatz. We here present relevant results needed for this work. In particular Lemma 8 is the motivation for our choice of ECC variables and links the ECC energy function to the bivariational principle. Theorem 9 formulates the continuous ECC equations and equates the solution of these equations with the solution of the SE.

In section 3 we analyze the flipped gradient of the ECC energy function and prove strong and local monotonicity for this entity. This is achieved for two complementary situations. Theorem 16 proves this property under assumptions on the structure of the solution, whereas Theorem 17 proves it under assumptions on the Hamiltonian. Along

the lines of the analysis of Rohwedder and Schneider for the CC theory, we prove existence and uniqueness for the solution of the (continuous) ECC equation and truncated (discrete) versions of it; see Theorem 19. This theorem also guarantees convergence toward the full solution as the truncated amplitude spaces tend to the continuous ones. Theorem 22 formulates a sufficient condition for the truncated amplitude spaces to grant a unique solution of the discrete ECC equation. Again the monotonicity is used for the flipped gradient. Last, in Theorem 24 we obtain error estimates for the truncated ECC energy. The energy estimates are obtained without the use of a Lagrangian and are instead based on the bivariational formulation of the theory.

## 2. Solving the Schrödinger equation using the exponential ansatz.

**2.1. Traditional CC theory in a rigorous manner.** In this section we consider the exponential parametrization for the $N$-electron ground-state wavefunction $\psi_*$ satisfying the $N$-electron SE

$$H\psi_* = E_*\psi_*.$$

Here, $E_*$ is the ground-state energy and $H$ is the Hamiltonian of a molecule in the Born–Oppenheimer approximation. We assume that $\psi_*$ exists and that it is non-degenerate, and we denote by $\gamma_* > 0$ the spectral gap (for a definition see section 3.2).

The set of admissible wavefunctions is a Hilbert space $\mathcal{H} \subset \mathcal{L}^2_N$ of finite kinetic energy wavefunctions, with norm $\|\psi\|^2_{\mathcal{H}} = \|\psi\|^2 + \|\nabla\psi\|^2$. Here, $\mathcal{L}^2_N$ is the space of totally antisymmetric square-integrable functions $\psi : (\mathbb{R}^3 \times \{\uparrow, \downarrow\})^N \to \mathbb{R}$, with norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$. In this work, we restrict our attention to the *real* space $\mathcal{L}^2_N$ and thus real Hamiltonians.

We will furthermore assume that the ground-state wavefunction $\psi_*$ is non-orthogonal to a (fixed) reference determinantal wavefunction $\phi_0 \in \mathcal{H}$, and thus, using intermediate normalization, we have $\psi_* = \phi_0 + \psi_\perp$, where $\langle\phi_0, \psi_\perp\rangle = 0$.

The molecular Hamiltonian has a set of useful properties that make the SE well-posed [20]. The operator $H : \mathcal{H} \to \mathcal{H}'$ is a bounded (continuous) operator into the dual $\mathcal{H}'$, i.e., there exists a constant $C \geq 0$ such that for all $\psi, \psi' \in \mathcal{H}$,

$$|\langle\psi', H\psi\rangle| \leq C\|\psi'\|_{\mathcal{H}}\|\psi\|_{\mathcal{H}}. \tag{1a}$$

Moreover, $H$ is below bounded by a constant $e \in \mathbb{R}$ such that $H + e$ is $\mathcal{H}$-coercive, i.e., there exists a constant $c > 0$ such that for all $\psi \in \mathcal{H}$,

$$\langle\psi, (H + e)\psi\rangle \geq c\|\psi\|^2_{\mathcal{H}}. \tag{1b}$$

The latter inequality is often referred to as a Gårding estimate and it is immediate that $e > -E_*$. Finally, $H$ is symmetric,

$$\langle\psi, H\psi'\rangle = \langle\psi', H\psi\rangle. \tag{1c}$$

Equations (1a)–(1c) form assumptions on $H$ that will be used frequently.

In a standard fashion, we introduce a basis for $\mathcal{H}$ of determinantal wavefunctions built from the $N$ "occupied" functions $\chi_i$ (forming $\phi_0$) as well as "virtual" functions $\chi_a$, $a = N+1, N+2, \ldots$ Assuming that $\{\chi_p : p = 1, 2, \ldots\}$ is an $\mathcal{L}^2_1$-orthonormal basis, the corresponding determinantal basis $\{\phi_\mu\}$ is $\mathcal{L}^2_N$-orthonormal. Additionally, we must require $\|\nabla\chi_p\| < +\infty$.

Each $\phi_\mu$ can be written on the form $\phi_\mu = X_\mu \phi_0$, where $X_\mu$ is an operator that creates up to $N$ particle-hole pairs, i.e., $\{X_\mu\}_{\mu\neq 0}$ are excitation operators, and for an arbitrary $\psi \in \mathcal{H}$ with $\langle \phi_0, \psi\rangle = 1$ we have

$$\psi = \phi_0 + \sum_{\mu\neq 0} c_\mu \phi_\mu = (I + C)\phi_0$$

with $C = \sum_{\mu\neq 0} c_\mu X_\mu$ being a *cluster operator*. The sequence $c = \{c_\mu\}_{\mu\neq 0}$ consists of the corresponding *cluster amplitudes*. One says that $\phi_0$ spans the "reference space" $\mathcal{P} := \operatorname{span}\{\phi_0\}$, while $\{\phi_\mu\}_{\mu\neq 0}$ forms a basis for $\mathcal{Q} = \mathcal{P}^\perp$, the "excluded space." It is clear that $\mathcal{P} \oplus \mathcal{Q} = \mathcal{H}$. (Here $\mathcal{P}^\perp$ denotes the $\mathcal{L}^2_N$ orthogonal complement of $\mathcal{P}$, i.e., with respect to the inner product $\langle\cdot,\cdot\rangle$.)

We introduce the convention that to each cluster amplitude sequence $c = \{c_\mu\}_{\mu\neq 0}$, $t = \{t_\mu\}_{\mu\neq 0}$, etc., the corresponding *cluster operator* is denoted by the capital letter, i.e., $C = \sum_\mu c_\mu X_\mu$, $T = \sum_\mu t_\mu X_\mu$, etc. Cluster operators by definition excludes $\mu = 0$, so unless otherwise specified, in what follows, all sums over $\mu$ run over excited determinants only. Moreover, we group the excitations according to the number of "particle-hole pairs" they create, i.e., $T = T_1 + T_2 + \cdots + T_N$, etc.

We follow [17] and introduce a Banach space of cluster amplitudes (in fact it is a Hilbert space). We say that $t \in \mathcal{V}$ if and only if $\|t\|_{\mathcal{V}} := \|T\phi_0\|_{\mathcal{H}} < +\infty$. Thus, $t \in \mathcal{V}$ if and only if $\{t_\mu\}$ are the amplitudes of a wavefunction of finite kinetic energy in the excluded space, i.e., $T\phi_0 \in \mathcal{Q}$. We remark that the space of cluster operators corresponding to amplitudes from $\mathcal{V}$ only depends on the choice of the reference $\phi_0$ (i.e., the space $\mathcal{P}$) and not on the choice of the virtual orbitals $\{\chi_a\}$, as long as $\{\phi_\mu\}$ is an orthonormal basis of $\mathcal{Q}$.

If the Hilbert space was finite dimensional, every linear operator would be bounded, and the exponential map $T \mapsto e^T$ would always be well-defined. A cornerstone of formal CC theory is therefore the well-definedness of the exponential map for general Hilbert spaces and cluster operators (see Lemma 2.3 in [17]).

THEOREM 1 (Rohwedder and Schneider, the exponential mapping). *$T$ and $T^\dagger$ are bounded operators on $\mathcal{H}$ if and only if $t \in \mathcal{V}$. Moreover, the exponential map $T \mapsto e^T$ is a (Fréchet) $\mathcal{C}^\infty$ isomorphism between $\mathcal{C} := \{T : t \in \mathcal{V}\}$ and $\mathcal{C}_0 := \{I + T : t \in \mathcal{V}\}$. For $\psi \in \mathcal{H}$ such that $\langle \phi_0, \psi\rangle = 1$ there exists a unique $t \in \mathcal{V}$ such that $\psi = e^T\phi_0$, depending smoothly on $\psi$. In particular the exponential map and its inverse are locally Lipschitz, i.e., for $s, t \in \mathcal{V}$ inside some ball, there exist constants $D, D'$ such that*

$$\|s - t\|_{\mathcal{V}} \leq D\|e^S\phi_0 - e^T\phi_0\|_{\mathcal{H}} \leq D'\|s - t\|_{\mathcal{V}}. \tag{2}$$

*Remark* 2. Note that the above theorem does not hold for a general subspace (truncation) $\mathcal{V}_d \subset \mathcal{V}$. To see this, let $\{\chi_p\}$ be an orthonormal set but not necessarily a (complete) basis and consider a subset $\mathcal{V}_d$ corresponding to only single excitations ($T = T_1$, $S = S_1$, etc.) and assume $N > 1$. Then the relation $e^T = I + S$ implies $T_1 + T_1^2/2 + \cdots + T_1^N/N! = S_1$. Thus, we can choose $T_1 \neq 0$ such that $e^{T_1} \neq I + S_1$ for any single excitation $S_1$.

The CC ansatz uses that the exponential is a bijection between the sets $\mathcal{C}$ and $\mathcal{C}_0$ such that $\psi_* = e^{T_*}\phi_0$ for some $T_*$ satisfying $e^{T_*} = I + C_*$. We then have (see Theorem 5.3 in [16]) the following.

THEOREM 3 (Rohwedder and Schneider, continuous CC formulation). *Under the assumptions on $H$ stated in* (1a) *and* (1b), *$\psi_* = e^{T_*}\phi_0$ solves $H\psi_* = E_*\psi_*$ if and only if*

$$f(t_*) = 0 \quad \text{and} \quad E_{CC}(t_*) = E_*, \tag{3}$$

*where* $f : \mathcal{V} \to \mathcal{V}'$ *is given by*

$$f_\mu(t) := \langle \phi_\mu, e^{-T} H e^T \phi_0 \rangle,$$

*and where* $E_{CC} : \mathcal{V} \to \mathbb{R}$ *is given by*

$$E_{CC}(t) := \langle \phi_0, e^{-T} H e^T \phi_0 \rangle.$$

*Remark* 4. (i) Equation (3) is the usual *untruncated* amplitude and energy equations of CC theory, formulated in the infinite-dimensional case, with $f : \mathcal{V} \to \mathcal{V}'$. This formulation was derived and named the continuous CC method in [16], being a mathematically rigorous formulation of the electronic SE using the exponential ansatz. Continuous here means that the excluded space $\mathcal{Q}$ is not discretized.

(ii) A remark on a frequently used notation in this article is in place. Since $f(t)$ is an element of the dual space of $\mathcal{V}$, $f(t) \in \mathcal{V}'$, the pairing with any $s \in \mathcal{V}$ is continuous in $s$ and given by the infinite series $\langle f(t), s \rangle = \sum_\mu s_\mu f_\mu(t)$. It should be clear from context whether $\langle \cdot, \cdot \rangle$ refers to the $\mathcal{L}^2_N$ inner product or the just stated infinite series.

Even if Theorem 3 reformulates the SE, it is not clear that truncations of $T$, with respect to either basis set or excitation level (or both), will give discretizations that yield existence and uniqueness of solutions as well as error estimates. The main tool here is the concept of local strong monotonicity of $f : \mathcal{V} \to \mathcal{V}'$. The following theorem is basically a local application of a classical theorem by Zarantonello [21]; see also Theorem 4.1 in [17] and Theorem 25.B and Corollary 25.7 in [22]. We will have great use of this result when studying the ECC method of Arponen. Let $X$ be a Hilbert space and define for a subspace $Y \subset X$ and $x \in X$ the distance $d(Y, x)$ between $Y$ and $x$ by

$$d(Y, x) := \inf_{y \in Y} \|y - x\|_X.$$

We recall that if $Y$ is closed, then there exists a minimizer $y_m$, i.e., $d(Y, x) = \|y_m - x\|_X$. This minimizer is the orthogonal projection of $x$ onto $Y$. We now state without proof the following theorem.

THEOREM 5 (local version of Zarantonello's theorem). *Let* $f : X \to X'$ *be a map between a Hilbert space* $X$ *and its dual* $X'$*, and let* $x_* \in B_\delta$ *be a root,* $f(x_*) = 0$*, where* $B_\delta$ *is an open ball of radius* $\delta$ *around* $x_*$*.*

*Assume that* $f$ *is Lipschitz continuous in* $B_\delta$*, i.e., that for all* $x_1, x_2 \in B_\delta$*,*

$$\|f(x_1) - f(x_2)\|_{X'} \leq L \|x_1 - x_2\|_X,$$

*for a constant* $L$*. Second, assume that* $f$ *is locally strongly monotone in* $B_\delta$*, i.e., that*

$$\langle f(x_1) - f(x_2), x_1 - x_2 \rangle \geq \gamma \|x_1 - x_2\|_X^2 \qquad \text{for all } \ x_1, x_2 \in B_\delta,$$

*for some constant* $\gamma > 0$*.*

*Then, the following holds:*

(1) *The root* $x_*$ *is unique in* $B_\delta$*. Indeed, there is a ball* $C_\varepsilon \subset X'$ *with* $0 \in C_\varepsilon$ *such that the solution map* $f^{-1} : C_\varepsilon \to X$ *exists and is Lipschitz continuous, implying that the equation*

$$f(x_* + \Delta x) = y$$

*has a unique solution* $\Delta x = f^{-1}(y) - x_*$*, depending continuously on* $y$*, with norm* $\|\Delta x\|_X \leq \delta$*.*

(2) *Moreover, let $X_d \subset X$ be a closed subspace such that $x_*$ can be approximated sufficiently well, i.e., the distance $d(x_*, X_d)$ is small. Then, the projected problem $f_d(x_d) = 0$ has a unique solution $x_d \in X_d \cap B_\delta$, and*

$$\|x_* - x_d\|_X \leq \frac{L}{\gamma} d(x_*, X_d).$$

Rohwedder and Schneider proved under certain assumptions (see Theorems 3.4 and 3.7 and Assumptions A and B in [17]) that the amplitude equations $f : \mathcal{V} \to \mathcal{V}'$ are indeed locally strongly monotone. (Lipschitz continuity follows from the differentiability of $f$.) Thus, the second part of Theorem 5 then guarantees that the truncated CC equations have a unique solution and that the error tends to zero as we increase the basis size and the truncation level of $T$, if the amplitude equation map $f$ is locally strongly monotone and Lipschitz continuous.

Before addressing the ECC method we follow Helgaker and Jørgensen [9] and remark that one can view the CC method as minimization of $E_{\mathrm{CC}}(t)$ over $\mathcal{V}$ under the constraint $f(t) = 0$. The Lagrangian in this case becomes

$$\begin{aligned} \mathcal{L}(t,s) &:= \langle \phi_0, e^{-T} H e^T \phi_0 \rangle + \sum_\mu s_\mu \langle \phi_\mu, e^{-T} H e^T \phi_0 \rangle \\ &= \langle \phi_0, (I + S^\dagger) e^{-T} H e^T \phi_0 \rangle \,, \end{aligned} \tag{4}$$

where $s = (s_\mu)_{\mu \neq 0} \in \mathcal{V}$ is the multiplier, which can be gathered into an excitation operator $S = \sum_\mu s_\mu X_\mu$. Note that $D_{s_\mu} \mathcal{L} = f_\mu$ since $\mathcal{L}(t,s) = E_{\mathrm{CC}}(t) + \langle f(t), s \rangle$. We shall in the next section see that the Lagrangian formulation is contained in the bivariational formulation of CC theory.

**2.2. The extended coupled-cluster method.** To link the forthcoming discussion to the previous section, we note that Arponen [1] derived the CC Lagrangian starting from the *bivariational Rayleigh–Ritz quotient* $\mathcal{E}_{\mathrm{bivar}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$,

$$\mathcal{E}_{\mathrm{bivar}}(\psi, \psi') := \frac{\langle \psi', H\psi \rangle}{\langle \psi', \psi \rangle}.$$

Vis-á-vis the usual Rayleigh–Ritz quotient, $\psi$ and $\psi'$ are here truly independent variables (not only treated as such in a formal manner). (See also the discussion following equation (24) in [13].) The stationary condition $D\mathcal{E}_{\mathrm{bivar}} = 0$ yields the left (and right) eigenvector(s) of $H$ with eigenvalue $E_*$; in fact, by straightforward differentiation we obtain the following result.

THEOREM 6 (bivariational principle). *Let $H : \mathcal{H} \to \mathcal{H}'$ be a bounded operator. Then, $\mathcal{E}_{bivar}$ is an infinitely differentiable function at all points where $\langle \psi', \psi \rangle \neq 0$, and $D_\psi \mathcal{E}_{bivar} = D_{\psi'} \mathcal{E}_{bivar} = 0$ if and only if the left and right SE are satisfied,*

$$H\psi = E\psi, \quad H^\dagger \psi' = E\psi', \quad \langle \psi', \psi \rangle \neq 0.$$

*Here, $H^\dagger : \mathcal{H} \to \mathcal{H}'$ is defined by $\langle H^\dagger \psi', \psi \rangle := \langle \psi', H\psi \rangle$.*

*Remark* 7. If we assume that $H$ satisfies all the requirements (1a)–(1c), in particular that $H$ is symmetric, and the left and right eigenvalue problems become identical, being the weak formulation of the eigenvalue problem of a unique self-adjoint $\hat{H}$ over $\mathcal{L}^2_N$. Suppose that $\hat{H}$ is close to self-adjoint, e.g., self-adjoint up to an $\mathcal{L}^2_N$-bounded perturbation. It is then reasonable that the left and right eigenvalue problems can

be simultaneously solved (but with $\psi' \neq \psi$). Thus, the bivariational principle can be thought of as a generalization of Rayleigh–Ritz to at least certain nonsymmetric problems.

We now introduce an exponential ansatz also for the wavefunction $\tilde{\psi}$. Following Arponen [1], we eliminate the denominator by changing the normalization of $\psi'$, i.e., we set $\tilde{\psi} = \psi' / \langle \psi', \psi \rangle$. The two scalar constraints lead to a smooth submanifold $\mathcal{M} \subset \mathcal{H} \times \mathcal{H}$ of codimension 2,

$$\mathcal{M} := \left\{ (\psi, \tilde{\psi}) \in \mathcal{H} \times \mathcal{H} \mid \langle \phi_0, \psi \rangle = \langle \tilde{\psi}, \psi \rangle = 1 \right\}. \tag{5}$$

The next lemma shows that this manifold $\mathcal{M}$ can be parameterized using cluster amplitudes.

LEMMA 8 (extended CC parameterization). *Suppose $(\psi, \tilde{\psi})$ satisfies $\langle \phi_0, \psi \rangle = \langle \tilde{\psi}, \psi \rangle = 1$. Then, there exists unique $(t, \lambda) \in \mathcal{V} \times \mathcal{V}$ depending smoothly on $(\psi, \tilde{\psi}) \in \mathcal{M}$ such that*

$$\psi = e^T \phi_0 \quad \textit{and} \quad \tilde{\psi} = e^{-T^\dagger} e^\Lambda \phi_0,$$

*which is a smooth map. In other words, the map $\Phi : \mathcal{V} \times \mathcal{V} \to \mathcal{M}$, $\Phi(t, \lambda) := (\psi(t), \tilde{\psi}(t, \lambda))$ is a smooth map with a smooth inverse.*

*Proof.* By Theorem 1, $t$ exists and is unique, depending smoothly on $\psi$ and vice versa. Consider $\omega = e^{T^\dagger(\psi)} \tilde{\psi}$, which depends smoothly on $(\psi, \tilde{\psi})$. We have $\langle \phi_0, \omega \rangle = 1$, so by Theorem 1 there exists a unique $\lambda$ depending smoothly on $\omega$, and hence $(\psi, \tilde{\psi})$, such that $\omega = e^\Lambda \phi_0$. Now $\tilde{\psi} = e^{-T^\dagger} e^\Lambda \phi_0$, a smooth map of $(t, \lambda)$. ∎

We define the ECC *energy functional* $\mathcal{E} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ by $\mathcal{E} = \mathcal{E}_{\text{bivar}} \circ \Phi$, viz.,

$$\mathcal{E}(t, \lambda) = \langle \phi_0, e^{\Lambda^\dagger} e^{-T} H e^T \phi_0 \rangle. \tag{6}$$

Equation (6) defines Arponen's ECC energy functional in a continuous, infinite-dimensional formulation.

THEOREM 9 (continuous ECC equations). *Let the Hamiltonian $H : \mathcal{H} \to \mathcal{H}'$ be as before. Then,*

$$H\psi_* = E_* \psi_* \quad \textit{and} \quad H\tilde{\psi}_* = E_* \tilde{\psi}_*$$

*with normalization $\langle \phi_0, \psi_* \rangle = \langle \tilde{\psi}_*, \psi_* \rangle = 1$ if and only if $D\mathcal{E}(t_*, \lambda_*) = 0$, i.e.,*

$$D_t \mathcal{E}(t_*, \lambda_*) = 0 \quad \textit{and} \quad D_\lambda \mathcal{E}(t_*, \lambda_*) = 0,$$

*where*

$$D_{t_\mu} \mathcal{E}(t, \lambda) = \langle \phi_0, e^{\Lambda^\dagger} [e^{-T} H e^T, X_\mu] \phi_0 \rangle, \tag{7a}$$

$$D_{\lambda_\mu} \mathcal{E}(t, \lambda) = \langle \phi_\mu, e^{\Lambda^\dagger} e^{-T} H e^T \phi_0 \rangle, \tag{7b}$$

*and where $(\psi_*, \tilde{\psi}_*) = \Phi(t_*, \lambda_*)$.*

*Proof.* $\Phi$ is differentiable with a differentiable inverse on $\mathcal{M}$, which is precisely the set of function pairs satisfying the normalization constraints. Thus $D\mathcal{E}(t_*, \lambda_*) = D[\mathcal{E}_{\text{bivar}} \circ \Phi](t_*, \lambda_*) = 0$ if and only if $D\mathcal{E}_{\text{bivar}}(\psi_*, \tilde{\psi}_*) = 0$ with the side condition $\langle \tilde{\psi}_*, \psi_* \rangle = \langle \phi_0, \psi_* \rangle = 1$. Moreover, $\mathcal{E}(t_*, \lambda_*) = \mathcal{E}_{\text{bivar}}(\psi_*, \tilde{\psi}_*) = E_*$. The formulas for the partial derivatives of $\mathcal{E}$ follow by elementary differentiation strategies. ∎

As in the case of standard CC theory, the continuous ECC equations do not imply that truncations of the amplitudes or the basis set give a well-behaved approximate method. To achieve this is the goal of the next section.

*Remark* 10. (i) We note that both $\psi$ and $\psi'$ are parameterized in an explicit multiplicatively separable manner when the system is decomposed into noninteracting subsystems. This is the main advantage of the ECC parameterization. We observe that the CC Lagrangian (given by (4)) is obtained by a further change of variables $S^\dagger := e^{\Lambda^\dagger} - 1$, which destroys this property of $\psi'$. Alternatively, one can view the CC Lagrangian as a first-order approximation to the ECC functional in terms of $\lambda$.

(ii) Arponen defined a further change of variables through $t'_\mu = \langle \phi_0, e^{\Lambda^\dagger} X_\mu^\dagger T \phi_0 \rangle$ and where the inverse $t = t(t', \lambda)$ is explicitly given by $t_\mu = \langle \phi_0, e^{-\Lambda^\dagger} X_\mu^\dagger T' \phi_0 \rangle$; see (5.6) and (5.7) in [1]. The variables $(t', \lambda)$ turn out to be *canonical* in the sense of classical Hamiltonian mechanics, i.e., the time-dependent SE is equivalent to Hamilton's equations of motion,

$$\begin{aligned} i\dot{t}'_\mu &= D_{\lambda_\mu} \mathcal{E}', \\ i\dot{\lambda}_\mu &= -D_{t'_\mu} \mathcal{E}', \end{aligned}$$

where $\mathcal{E}'(t', \lambda) := \mathcal{E}(t(t', \lambda), \lambda)$ and $\dot{t}$ (and $\dot{\lambda}$) denotes the time derivative of the amplitudes $t$ (and $\lambda$). The canonical variables have a computational advantage over the earlier defined noncanonical variables. As it turns out, they introduce cancellations in the (linked) diagram series for $E_*$ compared to when using the noncanonical $(t, \lambda)$. We shall not use the variables $(t', \lambda)$ here, as the analysis becomes considerably more complicated, and instead relegate their study to future work.

## 3. Analysis of ECC from monotonicity.

**3.1. The flipped gradient $\mathcal{F}$.** We will discuss the stationary point of $\mathcal{E}$ corresponding to the ground-state energy $E_*$ in terms of a map $\mathcal{F} : \mathcal{V} \times \mathcal{V} \to \mathcal{V}' \times \mathcal{V}'$ defined by flipping the components of the (Fréchet) derivative $D\mathcal{E} = (D_t\mathcal{E}, D_\lambda\mathcal{E})$, i.e.,

$$\mathcal{F} := (D_t\mathcal{E}, D_\lambda\mathcal{E}) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = (D_\lambda\mathcal{E}, D_t\mathcal{E}).$$

The components of the derivative are given in (7).

For the forthcoming discussion, let $B_\delta(t, \lambda)$ denote the ball of radius $\delta > 0$ centered at $(t, \lambda) \in \mathcal{V} \times \mathcal{V}$. Here the norm is $\|(\cdot, \cdot)\|^2_{\mathcal{V}\times\mathcal{V}} := \|\cdot\|^2_{\mathcal{V}} + \|\cdot\|^2_{\mathcal{V}}$. Let $(t_*, \lambda_*) \in \mathcal{V} \times \mathcal{V}$ be the optimal amplitudes corresponding to the ground-state pair $(\psi_*, \tilde{\psi}_*)$, in particular $\mathcal{F}(t_*, \lambda_*) = 0$. For the ECC function $\mathcal{F}$ we now want to establish the following:

(i) $\mathcal{F}$ is locally Lipschitz, i.e., let $(t, \lambda) \in \mathcal{V} \times \mathcal{V}$, then there exists $\delta > 0$ such that $(t_i, \lambda_i) \in B_\delta(t, \lambda)$ implies

$$\|\mathcal{F}(t_1, \lambda_1) - \mathcal{F}(t_2, \lambda_2)\|_{\mathcal{V}'\times\mathcal{V}'} \le L \|(t_1, \lambda_1) - (t_2, \lambda_2)\|_{\mathcal{V}\times\mathcal{V}}$$

for some (Lipschitz) constant $L > 0$, possibly depending only on $(t, \lambda)$ and $\delta$.

(ii) $\mathcal{F}$ is locally and strongly monotone at $(t_*, \lambda_*) \in \mathcal{V} \times \mathcal{V}$, i.e., there exists $\delta, \gamma > 0$ such that

$$\langle \mathcal{F}(t_1, \lambda_1) - \mathcal{F}(t_2, \lambda_2), (t_1, \lambda_1) - (t_2, \lambda_2) \rangle \ge \gamma(\|t_1 - t_2\|^2_{\mathcal{V}} + \|\lambda_1 - \lambda_2\|^2_{\mathcal{V}})$$

holds for all $(t_1, \lambda_1), (t_2, \lambda_2) \in B_\delta(t_*, \lambda_*)$.

Item (i) above is readily established using the fact that $\mathcal{F}$ is the flipped gradient of a smooth function. For (ii), we shall formulate two sets of assumptions (Assumptions 1 and 2 below) that each is enough to give strong monotonicity for $\mathcal{F}$ locally at $(t_*, \lambda_*)$. Having proved (i) and (ii), we can apply Theorem 5 to obtain existence and uniqueness results, also for truncated schemes.

The definition of local strong monotonicity of the map $\mathcal{F}$ reduces to the existence of a $\gamma > 0$ such that for $(t_i, \lambda_i)$ close to $t_*, \lambda_*$, the quantity

$$\begin{aligned}\Delta_1(t_1,\lambda_1,t_2,\lambda_2) + \Delta_2(t_1,\lambda_1,t_2,\lambda_2) := \langle D_\lambda \mathcal{E}(t_1,\lambda_1) - D_\lambda \mathcal{E}(t_2,\lambda_2), t_1 - t_2\rangle \\ + \langle D_t \mathcal{E}(t_1,\lambda_1) - D_t \mathcal{E}(t_2,\lambda_2), \lambda_1 - \lambda_2\rangle\end{aligned} \tag{8}$$

satisfies

$$\Delta_1(t_1,\lambda_1,t_2,\lambda_2) + \Delta_2(t_1,\lambda_1,t_2,\lambda_2) \geq \gamma(\|t_1 - t_2\|_{\mathcal{V}}^2 + \|\lambda_1 - \lambda_2\|_{\mathcal{V}}^2). \tag{9}$$

The choice of the map $\mathcal{F}$ can be motivated as follows: It is clear that $D\mathcal{E}$ cannot be locally strongly monotone, as, just like $\mathcal{E}_{\text{bivar}}$, all the critical points of $\mathcal{E}$ are intuitively saddle points (we will not prove this claim). On the other hand, in [17], the map $f(t)$ from Theorem 3 was considered and demonstrated to be locally strongly monotone under suitable assumptions. We observe that $f = D_s\mathcal{L}$, a partial derivative of the Lagrangian, which is *linear* in $s$, so that $f$ is only a function of $t$. In [17] it was demonstrated that (locally at $t_*$)

$$\Delta(t_1,t_2) = \langle [D_s\mathcal{L}](t_1) - [D_s\mathcal{L}](t_2), t_1 - t_2\rangle \geq \gamma\|t_1 - t_2\|_{\mathcal{V}}^2 \tag{10}$$

for some constant $\gamma > 0$. Thus, (10) is "half" of the inequality (9). In the ECC theory, the functional $\mathcal{E}$ is nonlinear in $\lambda$, indicating that we should include $\lambda$ in the monotonicity argument.

**3.2. Assumptions and preparation.** The analysis of Arponen's ECC method conducted here will be based on two complementary assumptions, Assumptions 1 and 2. The former deals with the accuracy of the ansatz, i.e., the accuracy of the reference $\phi_0$, while the latter considers a splitting of the Hamiltonian, e.g., the smallness of the fluctuation potential when a Hartree–Fock reference is used. We thus obtain two complementary monotonicity results applicable in different situations. However, both assumptions rest on conditions on spectral gaps. Recall the $\mathcal{P}$ denotes the reference space and moreover set $\mathcal{P}_* := \operatorname{span}\{\psi_*\}$. Let $P$ and $P_*$ denote the $\mathcal{L}^2_N$-orthogonal projections on $\mathcal{P}$ and $\mathcal{P}_*$, respectively. Essential for the analysis, we then have to assume that either there exists $\gamma_* > 0$ such that (Assumption 1)

$$\langle (I - P_*)\psi, (H - E_*)(I - P_*)\psi\rangle \geq \gamma_*\|(I - P_*)\psi\|^2 \tag{11}$$

or there exists $\gamma_0 > 0$ such that (Assumption 2)

$$\langle (I - P)\psi, (F - e_0)(I - P)\psi\rangle \geq \gamma_0\|(I - P)\psi\|^2 \tag{12}$$

for all $\psi \in \mathcal{H}$. Here $F$ is a one-body operator that has $\phi_0$ as ground state with ground-state energy $e_0$. A Hamiltonian splitting is then given by $H = F + (H - F)$ and will be dealt with below in connection with Assumption 2. We note that (11) expresses the fact that $E_*$ is the leftmost eigenvalue of $H$, that this eigenvalue exists, and that it has multiplicity 1.

We iterate that throughout the analysis we assume that the system Hamiltonian is bounded as quadratic form and additionally satisfying a Gårding estimate; see the

discussion in section 2.1, and in particular (1a)–(1c). We first state a slight upgrade of Lemma 3.5 in [17]. Note that for $\psi \in \mathcal{H}$, $(I-P)\psi \in \mathcal{Q}$. Also recall that in our notation $\|\cdot\|$ is the $\mathcal{L}^2_N$ norm.

LEMMA 11. *With $\psi_* = \phi_0 + \psi_\perp$, where $\psi_\perp \in \mathcal{Q}$ is the correction to $\phi_0$, we have the following:*

(i) *Assume that* (11) *holds with $\gamma_* > 0$ and that $\|\psi_\perp\|_{\mathcal{H}} < \varepsilon$. Then there exists a $\gamma_\varepsilon \in (0, \gamma_*]$ such that, for all $\psi \in \mathcal{Q}$,*

$$\langle \psi, (H - E_*)\psi \rangle \geq \frac{\gamma_\varepsilon}{\gamma_\varepsilon + e + E_*} c\|\psi\|^2_{\mathcal{H}}, \tag{13}$$

*where $\gamma_\varepsilon \to \gamma_*$ as $\varepsilon \to 0+$.*

(ii) *Assume $F\phi_0 = e_0\phi_0$ and that* (12) *holds with $\gamma_0 > 0$ and that $F$ satisfies the Gårding estimate given in* (1b) *(with constants $e_F$ and $c_F$). Then*

$$\langle \psi, (F - e_0)\psi \rangle \geq \frac{\gamma_0}{\gamma_0 + e_F + e_0} c_F\|\psi\|^2_{\mathcal{H}} \tag{14}$$

*for all $\psi \in \mathcal{Q}$.*

*Proof.* (i) Let $\psi \in \mathcal{Q}$. We first show that for $\gamma_\varepsilon > 0$ (and where $\gamma_\varepsilon \to \gamma_*$ as $\varepsilon \to 0+$) there holds

$$\langle \psi, (H - E_*)\psi \rangle \geq \gamma_\varepsilon \|\psi\|^2. \tag{15}$$

Following the argument in the proof of Lemma 2.4 in [17], we then have with $0 < q := \gamma_\varepsilon/(\gamma_\varepsilon + e + E_*) < 1$ (recall that $e + E_* > 0$ by necessity of the Gårding estimate)

$$\begin{aligned}\langle \psi, (H - E_*)\psi \rangle &= q\langle \psi, (H - E_*)\psi \rangle + (1-q)\langle \psi, (H - E_*)\psi \rangle \\ &\geq qc\|\psi\|^2_{\mathcal{H}} + (\gamma_\varepsilon - q(\gamma_\varepsilon + e + E_*))\|\psi\|^2.\end{aligned}$$

Thus, if (15) holds we are done.

Let $P$ and $P_*$ be as above. We use that

$$\|P - P_*\|_{\mathcal{B}(\mathcal{L}^2_N)} \leq 2\|\phi_0 - \psi'_*\|,$$

where $\psi'_* = \psi_*/\|\psi_*\|$. Since $\psi_* = \phi_0 + \psi_\perp$, with $\alpha := \|\psi_\perp\|$ we have

$$\|P - P_*\|_{\mathcal{B}(\mathcal{L}^2_N)} \leq 2\big(2 - 2(1+\alpha^2)^{-1/2}\big)^{1/2} =: j(\alpha).$$

Note that $j(\alpha)$ is an increasing function for $\alpha > 0$ and $j(\alpha) = 2\alpha + \mathcal{O}(\alpha^2)$.

Since $(H - E_*)P_*\psi = 0$ (and $H$ is symmetric), the left-hand side of (15) equals

$$\langle (I - P_*)\psi, (H - E_*)(I - P_*)\psi \rangle,$$

which by (11) is bounded from below by $\gamma_*\|(I-P_*)\psi\|^2$. Thus for $\alpha$ sufficiently small

$$\begin{aligned}\langle \psi, (H - E_*)\psi \rangle &\geq \gamma_*(\|(I-P)\psi\| - \|(P-P_*)\psi\|)^2 \\ &\geq \gamma_*(1 - j(\alpha))^2\|\psi\|^2.\end{aligned}$$

Since $\varepsilon > \|\psi_\perp\|_{\mathcal{H}} \geq \alpha$, we have that (15) holds with $\gamma_\varepsilon := \gamma_*(1 - j(\varepsilon))^2$. It is clear that $\gamma_\varepsilon \to \gamma_*$ as $\varepsilon$ tends to zero from above because $j(\varepsilon) \to 0$.

(ii) With $q_F := \gamma_0/(\gamma_0 + e_F + e_0)$ we have $0 < q_F < 1$ since $e_F > -e_0$ (equivalent to $e > -E_*$). Thus we can repeat the above scheme with $q = q_F$ to complete the proof. ∎

Because the relation $\psi_\perp = (e^{T_*} - I)\phi_0$ holds, it is immediate that $\|\psi_\perp\|_{\mathcal{H}}$ is small if and only if $\|t_*\|_{\mathcal{V}}$ is. It is a fact that the operator norm $\|T\|_{\mathcal{B}(\mathcal{H})}$ is equivalent to the norm $\|t\|_{\mathcal{V}}$; see [17]. We now state the first assumption, where the constants $e$, $c$, and $C$ are as in (1a)–(1b):

*Assumption* 1. Let $\eta_\varepsilon := \gamma_\varepsilon c/(\gamma_\varepsilon + e + E_*)$. We assume the following:
(a) Equation (11) holds with a strictly positive spectral gap $\gamma_* > 0$.
(b) The optimal amplitudes $t_*$ and $\lambda_*$ are sufficiently small in $\|\cdot\|_{\mathcal{V}}$ norm. With $C_* := C + |E_*|$ we then assume $\|\psi_\perp\|_{\mathcal{H}} < \varepsilon$, where $\varepsilon > 0$ is chosen such that

$$b_*(t_*, \lambda_*) := \|e^{-T_*^\dagger} e^{\Lambda_*} - I\|_{\mathcal{B}(\mathcal{H})} + \|e^{-T_*^\dagger} e^{\Lambda_*}\|_{\mathcal{B}(\mathcal{H})} \|e^{T_*} - I\|_{\mathcal{B}(\mathcal{H})} + K\|\phi_0\|_{\mathcal{H}} \|e^{-T_*^\dagger}\|_{\mathcal{B}(\mathcal{H})} \|e^{T_*}\|_{\mathcal{B}(\mathcal{H})} \|e^{\Lambda_*} - I\|_{\mathcal{B}(\mathcal{H})} < \frac{\eta_\varepsilon}{C_*}. \tag{16}$$

Here, $K$ is a constant such that $\|T\|_{\mathcal{B}(\mathcal{H})} \leq K\|t\|_{\mathcal{V}}$, which exists since the norms are equivalent.

*Remark* 12. It is in fact possible to choose $\varepsilon > 0$ such that (16) holds. Indeed, $\varepsilon = 0$ is equivalent to $t_* = \lambda_* = 0$, and $b_*(t_*, \lambda_*) = b(\varepsilon)$, a smooth function of $\varepsilon$. Since, $b(\varepsilon) \to 0+$ as $\varepsilon \to 0+$ and $\gamma_\varepsilon$ tends to the spectral gap $\gamma_*$, there exists a $\varepsilon_0$ such that $b_* < \eta_\varepsilon/C_*$ for $\varepsilon \leq \varepsilon_0$. Furthermore, at $\varepsilon = 0$ we have $\psi_* = \phi_0$, such that $\gamma_* = \gamma_0$ and $\mathcal{P}_* = \mathcal{P}$.

We next define the similarity transformed Hamiltonian $H_t$ and the doubly similarity transformed Hamiltonian $H_{t,\lambda}$ as given by

$$H_t := e^{-T} H e^{T}, \quad H_{t,\lambda} := e^{\Lambda^\dagger} H_t e^{-\Lambda^\dagger}.$$

Note that $(H_t)_\lambda \neq H_{t,\lambda}$. Since $e^{T_*}\phi_0$ solves the SE with eigenvalue $E_*$, $\phi_0$ is an eigenfunction of $H_{t_*}$ with the same eigenvalue. This fact and $e^{\Lambda_*^\dagger}\phi_0 = e^{-\Lambda_*^\dagger}\phi_0 = \phi_0$ make it easy to verify (i) in the following.

LEMMA 13. *Let* $f(t_*) = \mathcal{F}(t_*, \lambda_*) = 0$ *and* $E_* = \mathcal{E}(t_*, \lambda_*)$. *Then*
(i) $H_{t_*}\phi_0 = E_*\phi_0$ *and* $H_{t_*,\lambda_*}\phi_0 = E_*\phi_0$,
(ii) $H_{t_*,\lambda_*}^\dagger \phi_0 = E_*\phi_0$.

*Proof.* It remains to prove (ii). We know that (by definition of the left eigenfunction of $H$)

$$H^\dagger e^{-T_*^\dagger} e^{\Lambda_*}\phi_0 = E_* e^{-T_*^\dagger} e^{\Lambda_*}\phi_0.$$

Thus $H_{t_*}^\dagger e^{\Lambda_*}\phi_0$ equals $E_* e^{\Lambda_*^\dagger}\phi_0$. ∎

*Remark* 14. Note that Lemma 13 is valid for any critical point $(t_c, \lambda_c)$ with corresponding eigenvalue $E_c$, not only the ground state ($(t_*, \lambda_*)$ and $E_*$). Furthermore, as stated in Lemma 13, the double similarity transform makes $\phi_0$ both the left and right eigenvectors of $H_{t_*,\lambda_*}$ with the same eigenvalue.

We now move on to Assumption 2, which corresponds to an assumption made in [17], but suitable for the ECC method. Roughly speaking, instead of assuming that the reference $\phi_0$ is sufficiently accurate, in Assumption 2 we assume that we have a splitting $H = F + W$, where $F$ is a one-body operator, and where $W$ is sufficiently small in some appropriate sense. For example, $F$ can be the Fock operator and $W$ the fluctuation potential of a molecule in the Born–Oppenheimer approximation. Moreover, we assume that $F\phi_0 = e_0\phi_0$ and that (12) holds, where $\gamma_0$ is the so-called HOMO-LUMO gap.

It can be remarked that due to the structure of $H$, the Baker–Campbell–Hausdorff (BCH) expansion for $H_t$ terminates identically after four nested commutators in the case of a two-body interaction operator, i.e., $H_t$ is actually a polynomial of low order, independently of the number of particles.

The expansion for the outer similarity transform in $H_{t,\lambda}$ also truncates, albeit at a higher order. Thus, we have a finite sum

$$H_{t,\lambda} = \sum_{m,n} \frac{1}{n!m!} [[H,T]_{(n)}, -\Lambda^\dagger]_{(m)}.$$

Here $[A,B]_{(n)}$ denotes $A$ $n$-fold commutated with $B$ and $[A,B]_{(0)} := A$. For $(t,\lambda) \in \mathcal{V} \times \mathcal{V}$, we define the operator $O(t,\lambda)$ through the relation

$$H_{t,\lambda} = H + [F,T] + [\Lambda^\dagger, F] + O(t,\lambda). \tag{17}$$

The significance of $O(t,\lambda)$ is that (17) implies

$$\mathcal{E}(t,\lambda) - \langle \phi_0, H\phi_0 \rangle = \langle \phi_0, O(t,\lambda)\phi_0 \rangle\,, \tag{18}$$

i.e., $O(t,\lambda)$ gives all nontrivial contributions to $\mathcal{E}$. In the Hartree–Fock case, the right-hand side of (18) is the correlation energy functional, since the Hartree–Fock energy is given by $E_{\mathrm{HF}} = \langle \phi_0, H\phi_0 \rangle$.

The idea is that if the reference $\phi_0$ is sufficiently good, the mapping $(t,\lambda) \mapsto O(t,\lambda)$ will be well-behaved. In fact, since $O(t,\lambda)$ is a (Fréchet-)smooth map, it is locally Lipschitz: Given $(t,\lambda) \in \mathcal{V} \times \mathcal{V}$, there exist $\delta, L > 0$ such that for all $(t_i,\lambda_i) \in B_\delta(t,\lambda)$,

$$\|O(t_1,\lambda_1) - O(t_2,\lambda_2)\|_{\mathcal{B}(\mathcal{H},\mathcal{H}')} \leq L \|(t_1 - t_2, \lambda_1 - \lambda_2)\|_{\mathcal{V}\times\mathcal{V}}.$$

In our case, we assume that $L$ is sufficiently small at $(t_*,\lambda_*)$. This, in a sense, measures the smallness of $W$.

*Assumption* 2. Let $H = F + W$ and $\eta_0 := \gamma_0 c_F/(\gamma_0 + e_F + e_0)$. We assume the following:

(a) $F : \mathcal{H} \to \mathcal{H}'$ is a one-body operator that satisfies the same conditions as $H$, i.e., it is symmetric and bounded and satisfies a Gårding estimate (with constants $e_F, c_F$), as in (1a)–(1c). The constant that bounds $F$ is denoted $C_F$ and we set $C_0 := C_F + |e_0|$.

(b) $F\phi_0 = e_0\phi_0$, where $e_0$ is the smallest eigenvalue of $F$. Equation (12) holds with a $\gamma_0 > 0$, i.e., there is a strictly positive HOMO-LUMO gap. In particular, Lemma 11 gives that (14) holds for all $\psi \in \mathcal{Q}$.

(c) The Lipschitz constant $L$ at $(t_*,\lambda_*)$ and $\|\lambda_*\|_{\mathcal{V}}$ are not too large, so that the following inequality holds:

$$\begin{aligned} 0 < \gamma := \eta_0 - \frac{1}{2} L\|\phi_0\|_{\mathcal{H}} \big(3 + K\|(e^{\Lambda_*} - 1)\phi_0\|_{\mathcal{H}} + \|e^{\Lambda_*}\phi_0\|_{\mathcal{H}}/\|\phi_0\|_{\mathcal{H}} \\ + 2\|e^{\Lambda_*}\|_{\mathcal{B}(\mathcal{H})}\big) - C_0\|e^{\Lambda_*} - 1\|_{\mathcal{B}(\mathcal{H})}. \end{aligned} \tag{19}$$

Here, $K$ is a constant such that $\|T\|_{\mathcal{B}(\mathcal{H})} \leq K\|t\|_{\mathcal{V}}$, which exists since the norms are equivalent.

*Remark* 15. Assumption 2(c) does not assume that $\lambda_*$ is small compared to $\lambda_1 - \lambda_2$. However, $\lambda_*$ (and $L$) cannot be too large, since then $\gamma$ eventually becomes negative. If we do assume that $\|\lambda_*\|_{\mathcal{V}} < \delta$, we obtain some simplifications; see Corollary 18 below.

**3.3. Proof of monotonicity.** We set $\Delta := \Delta_1 + \Delta_2$, the left-hand side of (8). We then wish to prove

$$\Delta \geq \gamma\big(\|t_1 - t_2\|_{\mathcal{V}}^2 + \|\lambda_1 - \lambda_2\|_{\mathcal{V}}^2\big), \tag{20}$$

where $(t_i, \lambda_i) \in B_\delta(t_*, \lambda_*)$ and $\gamma, \delta > 0$. To simplify notation we define $\bar{T} = (T_1 + T_2)/2$ and $\delta T = T_1 - T_2$, and similarly $\bar{\Lambda} = (\Lambda_1 + \Lambda_2)/2$ and $\delta\Lambda = \Lambda_1 - \Lambda_2$. Consequently, we write $\|\delta t\|_{\mathcal{V}}$ and $\|\delta\lambda\|_{\mathcal{V}}$ for $\|t_1 - t_2\|_{\mathcal{V}}$ and $\|\lambda_1 - \lambda_2\|_{\mathcal{V}}$, respectively.

THEOREM 16. *Assume that Assumption* 1 *holds. Then $\mathcal{F}$ is strongly monotone locally at $(t_*, \lambda_*)$, $\mathcal{F}(t_*, \lambda_*) = 0$, belonging to the ground-state energy $E_* = \mathcal{E}(t_*, \lambda_*)$.*

*Proof.* Using the formulas (7) for the partial derivatives, we obtain for the two terms in (8),

$$\begin{aligned}
\Delta_1 &= \langle \delta T \phi_0, \big(e^{\Lambda_1^\dagger} H_{t_1} - e^{\Lambda_2^\dagger} H_{t_2}\big)\phi_0\rangle, \\
\Delta_2 &= \langle \phi_0, \big(e^{\Lambda_1^\dagger}[H_{t_1}, \delta\Lambda] - e^{\Lambda_2^\dagger}[H_{t_2}, \delta\Lambda]\big)\phi_0\rangle.
\end{aligned}$$

Moreover, we make use of the notation $g_i := t_i - t_*$, $k_i := \lambda_i - \lambda_*$ and define the excitation operators $G_i := \sum_\mu (g_i)_\mu X_\mu$ and $K_i := \sum_\mu (k_i)_\mu X_\mu$. Also we write $\delta G$ and $\delta K$ as for $T$ and $\Lambda$, where of course $\delta G = \delta T$ and $\delta K = \delta\Lambda$. As in [17], we note that the similarity transformed Hamiltonians $H_{t_i}$ can be expanded in terms of $H_{t_*}$ as

$$H_{t_i} = H_{t_*} + [H_{t_*}, G_i] + \mathcal{O}(\|g_i\|_{\mathcal{V}}^2). \tag{21}$$

Let $\tilde{\Delta}$ be the second-order Taylor expansion of $\Delta$ around $(t_*, \lambda_*)$, i.e., $\Delta = \tilde{\Delta} + \mathcal{O}(\|(\delta t, \delta\lambda)\|_{\mathcal{V}\times\mathcal{V}}^3)$. We will demonstrate the claim by first showing that $\tilde{\Delta}$ satisfies (20) for some $\tilde{\gamma} > 0$, using Assumption 1. Now by (21) and $\Lambda_i = K_i + \Lambda_*$, we see that

$$\begin{aligned}
\Delta_1 = \langle \delta T \phi_0, \big(&e^{K_1^\dagger} e^{\Lambda_*^\dagger}(H_{t_*} + [H_{t_*}, G_1] + \mathcal{O}(\|g_1\|_{\mathcal{V}}^2)) \\
&- e^{K_2^\dagger} e^{\Lambda_*^\dagger}(H_{t_*} + [H_{t_*}, G_2] + \mathcal{O}(\|g_2\|_{\mathcal{V}}^2))\big)\phi_0\rangle.
\end{aligned}$$

With the aid of Lemma 13 and since $e^{K_i^\dagger}\phi_0 = \phi_0$, it holds that

$$\Delta_1 = \langle \delta T \phi_0, \big(e^{K_1^\dagger} e^{\Lambda_*^\dagger}[H_{t_*}, G_1] - e^{K_2^\dagger} e^{\Lambda_*^\dagger}[H_{t_*}, G_2] + \mathcal{O}(\|g_1\|_{\mathcal{V}}^2) + \mathcal{O}(\|g_2\|_{\mathcal{V}}^2)\big)\phi_0\rangle.$$

As a next step we truncate $e^{K_i^\dagger} = I + \mathcal{O}(\|k_i\|_{\mathcal{V}})$ and there holds

$$\begin{aligned}
\Delta_1 &= \langle \delta T \phi_0, e^{\Lambda_*^\dagger}[H_{t_*}, \delta T]\phi_0\rangle + \sum_{k=0}^{3} \mathcal{O}(\|g_i\|_{\mathcal{V}}^k \|k_i\|_{\mathcal{V}}^{3-k}) \\
&= \langle \delta T \phi_0, e^{\Lambda_*^\dagger}(H_{t_*} - E_*)\delta T \phi_0\rangle + \sum_{k=0}^{3} \mathcal{O}(\|g_i\|_{\mathcal{V}}^k \|k_i\|_{\mathcal{V}}^{3-k}).
\end{aligned}$$

Again we have made use of Lemma 13. Equation (13) from Lemma 11 and (1a) give two useful bounds,

$$\langle \psi', (H - E_*)\psi\rangle \geq \eta_\varepsilon \|\psi\|_{\mathcal{H}}^2 - C_* \|\psi' - \psi\|_{\mathcal{H}} \|\psi\|_{\mathcal{H}}, \quad \psi' \in \mathcal{H}, \psi \in \mathcal{Q}, \tag{22}$$

$$\langle \psi', (H - E_*)\psi\rangle \geq -C_* \|\psi'\|_{\mathcal{H}} \|\psi\|_{\mathcal{H}}, \quad \psi', \psi \in \mathcal{H}. \tag{23}$$

Using these,

$$\begin{aligned}
\tilde{\Delta}_1 &= \langle \delta T\phi_0, e^{\Lambda_*^\dagger}(H_{t_*} - E_*)\delta T\phi_0\rangle \\
&= \langle e^{-T_*^\dagger}e^{\Lambda_*}\delta T\phi_0, (H-E_*)\delta T\phi_0\rangle + \langle e^{-T_*^\dagger}e^{\Lambda_*}\delta T\phi_0, (H-E_*)(e^{T_*}-I)\delta T\phi_0\rangle \\
&\geq \eta_\varepsilon \|\delta T\phi_0\|_{\mathcal{H}}^2 - C_*\|e^{-T_*^\dagger}e^{\Lambda_*} - I\|_{\mathcal{B}(\mathcal{H})}\|\delta T\phi_0\|_{\mathcal{H}}^2 \\
&\quad - C_*\|e^{-T_*^\dagger}e^{\Lambda_*}\|_{\mathcal{B}(\mathcal{H})}\|e^{T_*} - I\|_{\mathcal{B}(\mathcal{H})}\|\delta T\phi_0\|_{\mathcal{H}}^2 \\
&= \|\delta t\|_{\mathcal{V}}^2\big(\eta_\varepsilon - C_*(\|e^{-T_*^\dagger}e^{\Lambda_*} - I\|_{\mathcal{B}(\mathcal{H})} + \|e^{-T_*^\dagger}e^{\Lambda_*}\|_{\mathcal{B}(\mathcal{H})}\|e^{T_*} - I\|_{\mathcal{B}(\mathcal{H})})\big).
\end{aligned}$$

Next, we look at $\Delta_2$. Proceeding in similar a fashion, we compute

$$\begin{aligned}
\Delta_2 &= \langle \phi_0, (I + K_1^\dagger + \mathcal{O}(\|k_1\|_{\mathcal{V}}^2))e^{\Lambda_*^\dagger}[H_{t_*} + [H_{t_*}, G_1] + \mathcal{O}(\|g_1\|_{\mathcal{V}}^2), \delta\Lambda]\phi_0\rangle \\
&\quad - \langle \phi_0, (I + K_2^\dagger + \mathcal{O}(\|k_2\|_{\mathcal{V}}^2))e^{\Lambda_*^\dagger}[H_{t_*} + [H_{t_*}, G_2] + \mathcal{O}(\|g_2\|_{\mathcal{V}}^2), \delta\Lambda]\phi_0\rangle \\
&= \langle \phi_0, \delta\Lambda^\dagger e^{\Lambda_*^\dagger}(H_{t_*} - E_*)\delta\Lambda\phi_0\rangle + \langle \phi_0, e^{\Lambda_*^\dagger}\big[[H_{t_*}, \delta T], \delta\Lambda\big]\phi_0\rangle \\
&\quad + \sum_{k=0}^{3}\mathcal{O}(\|g_i\|_{\mathcal{V}}^k\|k_i\|_{\mathcal{V}}^{3-k}) \qquad (24) \\
&=: \tilde{\Delta}_{2,1} + \tilde{\Delta}_{2,2} + \sum_{k=0}^{3}\mathcal{O}(\|g_i\|_{\mathcal{V}}^k\|k_i\|_{\mathcal{V}}^{3-k}),
\end{aligned}$$

where the last equality defines $\tilde{\Delta}_{2,1}$ and $\tilde{\Delta}_{2,2}$. For $\tilde{\Delta}_{2,1}$ in (24), we again employ (22) and (23) to obtain

$$\begin{aligned}
\tilde{\Delta}_{2,1} &= \langle \phi_0, \delta\Lambda^\dagger e^{\Lambda_*^\dagger}(H_{t_*} - E_*)\delta\Lambda\phi_0\rangle \\
&= \langle e^{-T_*^\dagger}e^{\Lambda_*}\delta\Lambda\phi_0, (H-E_*)\delta\Lambda\phi_0\rangle + \langle e^{-T_*^\dagger}e^{\Lambda_*}\delta\Lambda\phi_0, (H-E_*)(e^{T_*}-I)\delta\Lambda\phi_0\rangle \\
&\geq \eta_\varepsilon\|\delta\Lambda\phi_0\|_{\mathcal{H}}^2 - C_*\|e^{-T_*^\dagger}e^{\Lambda_*} - I\|_{\mathcal{B}(\mathcal{H})}\|\delta\Lambda\phi_0\|_{\mathcal{H}}^2 \\
&\quad - C_*\|e^{-T_*^\dagger}e^{\Lambda_*}\|_{\mathcal{B}(\mathcal{H})}\|e^{T_*} - I\|_{\mathcal{B}(\mathcal{H})}\|\delta\Lambda\phi_0\|_{\mathcal{H}}^2 \\
&= \|\delta\lambda\|_{\mathcal{V}}^2\big(\eta_\varepsilon - C_*(\|e^{-T_*^\dagger}e^{\Lambda_*} - I\|_{\mathcal{B}(\mathcal{H})} + \|e^{-T_*^\dagger}e^{\Lambda_*}\|_{\mathcal{B}(\mathcal{H})}\|e^{T_*} - I\|_{\mathcal{B}(\mathcal{H})})\big).
\end{aligned}$$

Turning to $\tilde{\Delta}_{2,2}$ in (24), we have by Lemma 13

$$\begin{aligned}
\tilde{\Delta}_{2,2} &= \langle \phi_0, e^{\Lambda_*^\dagger}\big[[H_{t_*}, \delta T], \delta\Lambda\big]\phi_0\rangle \\
&= \langle e^{\Lambda_*}\phi_0, \big((H_{t_*}\delta T - \delta T H_{t_*})\delta\Lambda - \delta\Lambda(H_{t_*}\delta T - \delta T H_{t_*})\big)\phi_0\rangle \\
&= \langle e^{\Lambda_*}\phi_0, \big(\delta T(E_* - H_{t_*})\delta\Lambda - \delta\Lambda(H_{t_*} - E_*)\delta T\big)\phi_0\rangle.
\end{aligned}$$

Since

$$\big(\delta T(E_* - H_{t_*})\delta\Lambda - \delta\Lambda(H_{t_*} - E_*)\delta T\big)\phi_0 \in \mathcal{Q},$$

we only need to keep that part of $e^{\Lambda_*}\phi_0$ that belongs to $\mathcal{Q}$. Using (23), it holds that

$$\begin{aligned}
\tilde{\Delta}_{2,2} &= \langle e^{-T_*^\dagger}\delta T^\dagger(e^{\Lambda_*} - I)\phi_0, (E_* - H)e^{T_*}\delta\Lambda\phi_0\rangle \\
&\quad + \langle e^{-T_*^\dagger}\delta\Lambda^\dagger(e^{\Lambda_*} - I)\phi_0, (E_* - H)e^{T_*}\delta T\phi_0\rangle \\
&\geq -2C_*K\|\phi_0\|_{\mathcal{H}}\|e^{-T_*^\dagger}\|_{\mathcal{B}(\mathcal{H})}\|e^{T_*}\|_{\mathcal{B}(\mathcal{H})}\|e^{\Lambda_*} - I\|_{\mathcal{B}(\mathcal{H})}\|\delta t\|_{\mathcal{V}}\|\delta\lambda\|_{\mathcal{V}} \\
&\geq -C_*K\|\phi_0\|_{\mathcal{H}}\|e^{-T_*^\dagger}\|_{\mathcal{B}(\mathcal{H})}\|e^{T_*}\|_{\mathcal{B}(\mathcal{H})}\|e^{\Lambda_*} - I\|_{\mathcal{B}(\mathcal{H})}\big(\|\delta\lambda\|_{\mathcal{V}}^2 + \|\delta t\|_{\mathcal{V}}^2\big).
\end{aligned}$$

To summarize, collecting the lower bounds for $\tilde{\Delta}_1$ and $\tilde{\Delta}_{2,i}$ we can now conclude by means of the definition given by (16)

$$\tilde{\Delta} \geq (\eta_\varepsilon - C_* b_*(t_*, \lambda_*))\big(\|\delta t\|_{\mathcal{V}}^2 + \|\delta\lambda\|_{\mathcal{V}}^2\big).$$

By Assumption 1, $\tilde{\gamma} := \eta_\varepsilon - C_* b_*(t_*, \lambda_*) > 0$ such that

$$\tilde{\Delta} \geq \tilde{\gamma}\big(\|\delta t\|_{\mathcal{V}}^2 + \|\delta\lambda\|_{\mathcal{V}}^2\big), \quad \tilde{\gamma} > 0, \tag{25}$$

holds. To conclude the proof, we just have to note that by (25)

$$\Delta \geq \tilde{\gamma}\big(\|\delta t\|_{\mathcal{V}}^2 + \|\delta\lambda\|_{\mathcal{V}}^2\big) + \mathcal{O}(\|(\delta t, \delta\lambda)\|_{\mathcal{V}\times\mathcal{V}}^3)$$

and by choosing $\delta$ sufficiently small there holds for some $\gamma \in (0, \tilde{\gamma}]$

$$\Delta \geq \gamma\big(\|\delta t\|_{\mathcal{V}}^2 + \|\delta\lambda\|_{\mathcal{V}}^2\big)$$

for $(t_i, \lambda_i) \in B_\delta(t_*, \lambda_*)$. ∎

THEOREM 17. *Assume that Assumption* 2 *holds. Then $\mathcal{F}$ is strongly monotone locally at $(t_*, \lambda_*)$, $\mathcal{F}(t_*, \lambda_*) = 0$, belonging to the ground-state energy $E_* = \mathcal{E}(t_*, \lambda_*)$.*

*Proof.* As in the proof of Theorem 16, we study $\Delta_1$ and $\Delta_2$ separately before adding them together. We begin by noting that

$$\Delta_1 = \langle \delta T\phi_0, (e^{\Lambda_1^\dagger} H_{t_1} - e^{\Lambda_2^\dagger} H_{t_2})\phi_0\rangle = \langle \delta T\phi_0, (H_{t_1,\lambda_1} - H_{t_2,\lambda_2})\phi_0\rangle\,,$$

because any deexcitation of the reference $\phi_0$ gives zero identically. Now, using Assumption 2 and the definition (17) of the operator $O(t, \lambda)$ we immediately obtain the following lower bound for $\Delta_1$:

$$\begin{aligned}
\Delta_1 &= \langle \delta T\phi_0, (H_{t_1,\lambda_1} - H_{t_2,\lambda_2})\phi_0\rangle \\
&= \langle \delta T\phi_0, \big([F, \delta T] + [\delta\Lambda^\dagger, F] + O(t_1, \lambda_1) - O(t_2, \lambda_2)\big)\phi_0\rangle \\
&= \langle \delta T\phi_0, (F - e_0)\delta T\phi_0\rangle + \langle \delta T\phi_0, (O(t_1, \lambda_1) - O(t_2, \lambda_2))\phi_0\rangle \\
&\geq \eta_0 \|\delta T\phi_0\|_{\mathcal{H}}^2 - L\|\delta T\phi_0\|_{\mathcal{H}}\|(\delta t, \delta\lambda)\|_{\mathcal{V}\times\mathcal{V}}\|\phi_0\|_{\mathcal{H}} \\
&= \eta_0 \|\delta t\|_{\mathcal{V}}^2 - L\|\phi_0\|_{\mathcal{H}}\|\delta t\|_{\mathcal{V}}(\|\delta t\|_{\mathcal{V}}^2 + \|\delta\lambda\|_{\mathcal{V}}^2)^{1/2} \\
&\geq \eta_0 \|\delta t\|_{\mathcal{V}}^2 - L\|\phi_0\|_{\mathcal{H}}\|\delta t\|_{\mathcal{V}}(\|\delta t\|_{\mathcal{V}} + \|\delta\lambda\|_{\mathcal{V}}).
\end{aligned}$$

We next turn to $\Delta_2$. It holds that

$$e^{\Lambda_1} - e^{\Lambda_2} = e^{\bar{\Lambda}}\delta\Lambda + \mathcal{O}(\|\delta\lambda\|_{\mathcal{V}}^2). \tag{26}$$

We compute

$$\begin{aligned}
\Delta_2 &= \langle \phi_0, \big(e^{\Lambda_1^\dagger}[H_{t_1}, \delta\Lambda] - e^{\Lambda_2^\dagger}[H_{t_2}\delta\Lambda]\big)\phi_0\rangle \\
&= \langle \phi_0, \big((e^{\Lambda_1^\dagger} - e^{\Lambda_2^\dagger})[H_{t_1}, \delta\Lambda] + e^{\Lambda_2^\dagger}[H_{t_1} - H_{t_2}, \delta\Lambda]\big)\phi_0\rangle \\
&= \langle \phi_0, \big(e^{\bar{\Lambda}^\dagger}\delta\Lambda^\dagger[F + W + O(t_1, 0), \delta\Lambda] - e^{\bar{\Lambda}^\dagger}[O(t_1, 0) - O(t_2, 0), \delta\Lambda]\big)\phi_0\rangle \\
&\quad + \mathcal{O}(\|\delta\lambda\|_{\mathcal{V}}^3) + \mathcal{O}(\|\delta\lambda\|_{\mathcal{V}}\|\delta t\|_{\mathcal{V}}^2) + \mathcal{O}(\|\delta\lambda\|_{\mathcal{V}}^2\|\delta t\|_{\mathcal{V}}).
\end{aligned} \tag{27}$$

In the last equality, we exploited that the second-order nested commutator of $F$ with two excitation operators vanishes. This is so since for $\mu \neq 0$ we have that $[F, X_\mu]$

is an excitation operator and consequently $[[F,T],T'] = 0$. Moreover, we used that $O(t_1,0) - O(t_2,0) = \mathcal{O}(\|\delta t\|)$, allowing us to replace $\Lambda_2$ with $\bar{\Lambda} = \Lambda_2 + \delta\Lambda/2$, a change which affects only the higher-order terms.

Define $\tilde{\Delta}_2$ as the leading second-order term of $\Delta_2$, i.e., the first term in the last line of (27), neglecting the third-order remainders (note that these are in total $\mathcal{O}(\|(\delta t, \delta\lambda)\|^3_{\mathcal{V}\times\mathcal{V}})$). We will start by finding $\tilde{\gamma} > 0$ such that

$$\Delta_1 + \tilde{\Delta}_2 \geq \tilde{\gamma}(\|\delta t\|^2_{\mathcal{V}} + \|\delta\lambda\|^2_{\mathcal{V}}).$$

We split $\tilde{\Delta}_2$ into two contributions, $\tilde{\Delta}_{2,i}$, $i = 1, 2$.

Since $O(t,0) + W = e^{-T} W e^{T}$, the BCH formula gives

$$O(t + \delta\lambda, 0) - O(t,0) = [O(t,0) + W, \delta\Lambda] + \mathcal{O}(\|\delta\lambda\|^2).$$

This gives us the directional derivative of $O(\cdot, 0)$ in the direction $\delta\lambda$,

$$DO(t,0)(\delta\lambda) = [O(t,0) + W, \delta\Lambda].$$

On the other hand, $O$ is Lipschitz, so that

$$\|[O(t_1,0) + W, \delta\Lambda]\|_{\mathcal{B}(\mathcal{H},\mathcal{H}')} \leq \|DO(t_1,0)\|_{\mathcal{B}(\mathcal{V},B(\mathcal{H},\mathcal{H}'))}\|\delta\lambda\|_{\mathcal{V}} \leq (L + K'\delta)\|\delta\lambda\|_{\mathcal{V}}$$

for some constant $K'$.

A useful bound is obtained from (14) from Lemma 11,

$$\langle\psi', (F - e_0)\psi\rangle \geq \eta_0\|\psi\|^2_{\mathcal{H}} - C_0\|\psi' - \psi\|_{\mathcal{H}}\|\psi\|_{\mathcal{H}}. \tag{28}$$

The first contribution becomes

$$\begin{aligned}
\tilde{\Delta}_{2,1} &= \langle\phi_0, e^{\bar{\Lambda}^\dagger}\delta\Lambda^\dagger[F,\delta\Lambda]\phi_0\rangle + \langle\phi_0, e^{\bar{\Lambda}^\dagger}\delta\Lambda^\dagger[O(t_1,0) + W, \delta\Lambda]\phi_0\rangle \\
&= \langle e^{\bar{\Lambda}}\delta\Lambda\phi_0, (F - e_0)\delta\Lambda\phi_0\rangle + \langle\delta\Lambda e^{\bar{\Lambda}}\phi_0, [O(t_1,0) + W, \delta\Lambda]\phi_0\rangle \\
&\geq \eta_0\|\delta\Lambda\phi_0\|^2_{\mathcal{H}} - C_0\|(e^{\bar{\Lambda}} - 1)\delta\Lambda\phi_0\|_{\mathcal{H}}\|\delta\Lambda\phi_0\|_{\mathcal{H}} \\
&\qquad - \|e^{\bar{\Lambda}}\delta\Lambda\phi_0\|_{\mathcal{H}}(L + K'\delta)\|\delta\lambda\|_{\mathcal{V}}\|\phi_0\|_{\mathcal{H}} \\
&\geq \big(\eta_0 - C_0\|e^{\bar{\Lambda}} - 1\|_{\mathcal{B}(\mathcal{H})} - (L + K'\delta)\|\phi_0\|_{\mathcal{H}}\|e^{\bar{\Lambda}}\|_{\mathcal{B}(\mathcal{H})}\big)\|\delta\lambda\|^2_{\mathcal{V}}.
\end{aligned}$$

The second contribution is

$$\begin{aligned}
\tilde{\Delta}_{2,2} &= \langle\phi_0, e^{\bar{\Lambda}^\dagger}[O(t_1,0) - O(t_2,0), \delta\Lambda]\phi_0\rangle \\
&= \langle e^{\bar{\Lambda}}\phi_0, (O(t_1,0) - O(t_2,0))\delta\Lambda\phi_0\rangle - \langle e^{\bar{\Lambda}}\phi_0, \delta\Lambda(O(t_1,0) - O(t_2,0))\phi_0\rangle \\
&\geq -L\|e^{\bar{\Lambda}}\phi_0\|_{\mathcal{H}}\|\delta\lambda\|_{\mathcal{V}}\|\delta t\|_{\mathcal{V}} - L\|\delta\Lambda^\dagger(e^{\bar{\Lambda}} - 1)\phi_0\|_{\mathcal{H}}\|\phi_0\|_{\mathcal{H}}\|\delta t\|_{\mathcal{V}} \\
&\geq -L\|e^{\bar{\Lambda}}\phi_0\|_{\mathcal{H}}\|\delta\lambda\|_{\mathcal{V}}\|\delta t\|_{\mathcal{V}} - LK\|(e^{\bar{\Lambda}} - 1)\phi_0\|_{\mathcal{H}}\|\phi_0\|_{\mathcal{H}}\|\delta\lambda\|_{\mathcal{V}}\|\delta t\|_{\mathcal{V}} \\
&= -L(K\|(e^{\bar{\Lambda}} - 1)\phi_0\|_{\mathcal{H}} + \|e^{\bar{\Lambda}}\phi_0\|_{\mathcal{H}}/\|\phi_0\|_{\mathcal{H}})\|\phi_0\|_{\mathcal{H}}\|\delta\lambda\|_{\mathcal{V}}\|\delta t\|_{\mathcal{V}}.
\end{aligned}$$

We gather and obtain,

$$\begin{aligned}
\Delta_1 + \tilde{\Delta}_2 \geq\ & \eta_0\|\delta t\|^2_{\mathcal{V}} - L\|\phi_0\|_{\mathcal{H}}\|\delta t\|_{\mathcal{V}}(\|\delta t\|_{\mathcal{V}} + \|\delta\lambda\|_{\mathcal{V}}) \\
& + \big(\eta_0 - \|F - e_0\|_{\mathcal{B}(\mathcal{H},\mathcal{H}')}\|e^{\bar{\Lambda}} - 1\|_{\mathcal{B}(\mathcal{H})} - (L + K'\delta)\|\phi_0\|_{\mathcal{H}}\|e^{\bar{\Lambda}}\|_{\mathcal{B}(\mathcal{H})}\big)\|\delta\lambda\|^2_{\mathcal{V}} \\
& - L(K\|(e^{\bar{\Lambda}} - 1)\phi_0\|_{\mathcal{H}} + \|e^{\bar{\Lambda}}\phi_0\|_{\mathcal{H}}/\|\phi_0\|_{\mathcal{H}})\|\phi_0\|_{\mathcal{H}}\|\delta\lambda\|_{\mathcal{V}}\|\delta t\|_{\mathcal{V}}
\end{aligned}$$

$$\geq \Big(\eta_0 - \frac{1}{2}L\|\phi_0\|_{\mathcal{H}}\big(3 + K\|(e^{\bar{\Lambda}} - 1)\phi_0\|_{\mathcal{H}} + \|e^{\bar{\Lambda}}\phi_0\|_{\mathcal{H}}/\|\phi_0\|_{\mathcal{H}} + 2\|e^{\bar{\Lambda}}\|_{\mathcal{B}(\mathcal{H})}\big)$$
$$- \|F - e_0\|_{\mathcal{B}(\mathcal{H},\mathcal{H}')}\|e^{\bar{\Lambda}} - 1\|_{\mathcal{B}(\mathcal{H})}\Big)\|(\delta t, \delta\lambda)\|^2_{\mathcal{V}\times\mathcal{V}}$$
$$- K'\delta\|\phi_0\|_{\mathcal{H}}\|e^{\bar{\Lambda}}\|_{\mathcal{B}(\mathcal{H})}\|(\delta t, \delta\lambda)\|^2_{\mathcal{V}\times\mathcal{V}}$$
$$=: \tilde{\gamma}(\bar{t}, \bar{\lambda})\|(\delta t, \delta\lambda)\|^2_{\mathcal{V}\times\mathcal{V}}.$$

We now note that, by Taylor's theorem, $\tilde{\gamma}(\bar{t}, \bar{\lambda}) = \gamma + \varepsilon(\bar{t}, \bar{\lambda}) - K'\delta$, with $\gamma = \tilde{\gamma}(t_*, \lambda_*) > 0$ by (19) in Assumption 2, and $|\varepsilon| \leq C\delta$ for some $C \geq 0$. Thus,

$$\Delta_1 + \tilde{\Delta}_2 \geq (\gamma - (C + K')\delta)\|(\delta t, \delta\lambda)\|^2_{\mathcal{V}\times\mathcal{V}}.$$

Finally,

$$\Delta_1 + \Delta_2 \geq (\gamma - (C + K')\delta)\|(\delta t, \delta\lambda)\|^2_{\mathcal{V}\times\mathcal{V}} + \mathcal{O}(\|(\delta t, \delta\lambda)\|^3_{\mathcal{V}\times\mathcal{V}}).$$

Since the third-order term cannot beat the second-order term, by shrinking $\delta$, we get

$$\Delta_1 + \Delta_2 \geq (\gamma - (C + K')\delta')\|(t_1 - t_2, \lambda_1 - \lambda_2)\|^2_{\mathcal{V}\times\mathcal{V}}$$

whenever $(t_i, \lambda_i) \in B_{\delta'}(t_*, \lambda_*)$. ∎

COROLLARY 18. *Assume Assumption* 2(a)–(b) *holds and additionally that we have* $\|\lambda_*\|_{\mathcal{V}} < \delta$. *Also, assume that*

$$0 < \eta_0 - 3L\|\phi_0\|_{\mathcal{H}}. \tag{29}$$

*Then* $\mathcal{F}$ *is locally strongly monotone at the root* $(t_*, \lambda_*)$ *belonging to the ground-state energy.*

*Proof.* It is enough to observe that we need to Taylor expand $\gamma = \tilde{\gamma}(t_*, \lambda_*)$ to zeroth order, i.e., setting $\lambda_* = 0$ in (19). The reader can readily verify that this gives (29). ∎

**3.4. Existence, uniqueness, truncations, and error estimates.** Having obtained sufficient conditions for $\mathcal{F}$ to be locally strongly monotone at $(t_*, \lambda_*)$, we can now apply the local version of Zarantonello's theorem, Theorem 5, to obtain existence and local uniqueness of solutions, also for truncated versions of the ECC method.

In our setting, a (family of) truncated amplitude spaces $\mathcal{V}_d \times \mathcal{V}_d$ is such that if we let the dimension $d \to +\infty$, we can approximate $(t_*, \lambda_*)$ arbitrarily well. Of course, the usual truncation scheme defined by all excitations up to a given excitation level and additionally the restriction to a finite set of virtual orbitals conforms with this. In what follows it will be assumed that $\mathcal{V}_d$ is closed in $\mathcal{V}$.

The truncated ECC functional is the restriction $\mathcal{E}_d : \mathcal{V}_d \times \mathcal{V}_d \to \mathbb{R}$ of $\mathcal{E}$, giving the critical point problem $D\mathcal{E}_d = 0$, i.e.,

$$\text{find } (t_d, \lambda_d) \in \mathcal{V}_d \times \mathcal{V}_d \quad \text{such that} \quad \frac{\partial\mathcal{E}(t_d, \lambda_d)}{\partial t_\mu} = \frac{\partial\mathcal{E}(t_d, \lambda_d)}{\partial\lambda_\mu} = 0,$$

where $t_\mu$ ($\lambda_\mu$) are the components of $t \in \mathcal{V}_d$ ($\lambda \in \mathcal{V}_d$) in some arbitrary orthonormal basis. Since the flipping map in (3.1) commutes with projection onto $\mathcal{V}_d \times \mathcal{V}_d$, the truncated ECC equations can be written $\mathcal{F}_d(t_d, \lambda_d) = 0$.

While stated as a theorem, our main result is really a corollary of Theorems 16 and 17 and an elementary application of Theorem 5. The only point to check is that $\mathcal{F}$ is locally Lipschitz. However, $\mathcal{F}$ is (in fact infinitely) continuously differentiable in the Fréchet sense. Such functions are always locally Lipschitz.

THEOREM 19. *Assume that Assumption* 1 *or* 2 *holds such that $\mathcal{F}$ is locally strongly monotone (with constant $\gamma$) on $B_\delta(t_*, \lambda_*)$ for some $\delta > 0$. Here, $(t_*, \lambda_*)$ is the root of $\mathcal{F}$ belonging to the ground-state energy. Furthermore, let $L$ be the local Lipschitz constant of $\mathcal{F}$ at $(t_*, \lambda_*)$.*

(i) *The solution $(t_*, \lambda_*)$ of the continuous ECC equation $D\mathcal{E}(t, \lambda) = 0$ on $\mathcal{V} \times \mathcal{V}$ is locally unique.*

(ii) *For sufficiently large $d$, the projected ECC problem $D\mathcal{E}_d(t, \lambda) = 0$ has a unique solution $(t_d, \lambda_d)$ in the neighborhood $B_\delta(t_*, \lambda_*) \cap (\mathcal{V}_d \times \mathcal{V}_d)$. The truncated solution $(t_d, \lambda_d)$ satisfies the estimate*

$$\|(t_d, \lambda_d) - (t_*, \lambda_*)\|_{\mathcal{V}\times\mathcal{V}} \leq \frac{L}{\gamma} d(\mathcal{V}_d \times \mathcal{V}_d, (t_*, \lambda_*)). \tag{30}$$

*Remark* 20. (i) The local uniqueness is also a direct consequence of the assumption that the ground state is nondegenerate and Lemma 8.

(ii) By the definition of the norm on $\mathcal{V} \times \mathcal{V}$, (30) implies

$$\|t_d - t_*\|_{\mathcal{V}}^2 + \|\lambda_d - \lambda_*\|_{\mathcal{V}}^2 \leq \frac{L^2}{\gamma^2} \left( d(\mathcal{V}_d, t_*)^2 + d(\mathcal{V}_d, \lambda_*)^2 \right) \tag{31}$$

and furthermore that $(t_d, \lambda_d) \to (t_*, \lambda_*)$ as $d \to +\infty$.

Theorem 19 guarantees that for sufficiently large discrete amplitude spaces $\mathcal{V}_d$, the ECC equations actually have locally unique solutions that approximate the exact solution. However, we do not yet know what "sufficiently large" means.

By slightly adapting the proof of Theorem 4.1 in [17], we can obtain a sufficient condition on $\mathcal{V}_d$. This argument rests on Brouwer's fixed point theorem: any continuous function of a closed ball in $\mathbb{R}^n$ into itself has a fixed point. Here, we employ a version of this result [8].

LEMMA 21. *Equip $\mathbb{R}^n$ with any norm $\|\cdot\|_n$, and let $B_R$ be the closed ball of radius $R$ centered at $\vec{x} = 0$. Let $h : B_R \to \mathbb{R}^n$ be continuous and assume that on the boundary of $B_R$, $\langle h(\vec{x}), \vec{x} \rangle = h(\vec{x}) \cdot \vec{x} \geq 0$. Then $h(\vec{x}) = 0$ for some $\vec{x} \in B_R$.*

*Proof.* Assume that $h \neq 0$ everywhere. Then $f(\vec{x}) := -Rh(\vec{x})/\|h(\vec{x})\|_n$ is continuous, mapping the ball into itself (in fact, onto its boundary). Therefore, $f$ has a fixed point, say, $\vec{x}_0$, i.e., $\vec{x}_0 = -Rh(\vec{x}_0)/\|h(\vec{x}_0)\|_n$. However, this gives the contradiction $0 < \vec{x}_0 \cdot \vec{x}_0 = -R \langle h(\vec{x}_0), \vec{x}_0 \rangle / \|\vec{x}_0\|_n \leq 0$. ∎

Following [17], the idea is now to choose $h_d$ such that $\mathcal{F}_d = 0$ is equivalent to $h_d = 0$ and use the above argument.

THEOREM 22. *Let $\mathcal{V}_d$ be a finite-dimensional subspace of $\mathcal{V}$ and set*

$$\kappa_d := \min_{(t,\lambda)\in\mathcal{V}_d\times\mathcal{V}_d} \|(t, \lambda) - (t_*, \lambda_*)\|_{\mathcal{V}\times\mathcal{V}} = \|(t_m, \lambda_m) - (t_*, \lambda_*)\|_{\mathcal{V}\times\mathcal{V}}. \tag{32}$$

*Assume that $\kappa_d$ satisfies*

$$\kappa_d \leq \frac{\delta\gamma}{\gamma + L}, \tag{33}$$

*where $\gamma$ and $L$ are the monotonicity and Lipschitz constants, respectively, that hold on $B_\delta(t_*, \lambda_*)$. Then the projected ECC problem $\mathcal{F}_d(t, \lambda) = 0$ has a unique solution $(t_d, \lambda_d)$ in the neighborhood $B_\delta(t_*, \lambda_*) \cap (\mathcal{V}_d \times \mathcal{V}_d)$.*

*Proof.* Let $d := \dim \mathcal{V}_d$ and $\{b_j\}_{j=1}^d$ be an orthonormal basis of $\mathcal{V}_d$. Define the continuous vector-valued function $h_d : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ by $h_d(\vec{x}) = h_d(\vec{v}, \vec{w}) = (\vec{y}, \vec{z})$, where

$$y_j = \langle D_\lambda \mathcal{E}(t_m + v, \lambda_m + w), b_j \rangle, \quad z_j = \langle D_t \mathcal{E}(t_m + v, \lambda_m + w), b_j \rangle,$$

and $v = \sum_{j=1}^d v_j b_j$, $\vec{v} = (v_1, \ldots, v_d)$, $w = \sum_{j=1}^d w_j b_j$, $\vec{w} = (w_1, \ldots, w_d)$. Let $\|(\vec{v}, \vec{w})\|_{2d} := \|(v, w)\|_{\mathcal{V}\times\mathcal{V}}$, a norm on $\mathbb{R}^{2d}$ (a fact that can be easily checked). By definition, $h_d = 0$ is equivalent to $\mathcal{F}_d = 0$.

We now choose $R := \delta - \kappa_d \geq \delta L/(\gamma + L) > 0$ and note that $(\vec{v}, \vec{w}) \in B_R(t_m, \lambda_m)$ implies $(v, w) \in B_\delta(t_*, \lambda_*)$. For $\vec{x}$ that satisfies $\|\vec{x}\|_{2d} = R$, we have using monotonicity and the Lipschitz continuity of $\mathcal{F}$,

$$\begin{aligned}
\langle h_d(\vec{x}), \vec{x} \rangle &= \sum_{j=1}^d (y_j v_j + z_j w_j) = \langle \mathcal{F}(t_m + v, \lambda_m + w), (v, w) \rangle \\
&= \langle \mathcal{F}(t_m + v, \lambda_m + w) - \mathcal{F}(t_m, \lambda_m), (v, w) \rangle \\
&\quad + \langle \mathcal{F}(t_m, \lambda_m) - \mathcal{F}(t_*, \lambda_*), (v, w) \rangle + \langle \mathcal{F}(t_*, \lambda_*), (v, w) \rangle \\
&\geq \gamma \|(v, w)\|_{\mathcal{V}\times\mathcal{V}}^2 - L\kappa_d \|(v, w)\|_{\mathcal{V}\times\mathcal{V}}.
\end{aligned}$$

Since $\gamma R - L\kappa_d = \gamma\delta - \kappa_d(\gamma + L) \geq 0$, we can conclude $\langle h_d(\vec{x}), \vec{x} \rangle = R(\gamma R - L\kappa_d) \geq 0$. Lemma 21 now establishes that $h_d(\vec{x}_*) = 0$ for some $\vec{x}_*$ with $\|\vec{x}\|_{2d} = \|(v_*, w_*)\|_{\mathcal{V}\times\mathcal{V}} \leq R$, which is equivalent to that $(t_d, \lambda_d) := (t_m + v_*, \lambda_m + w_*)$ solves the projected problem $\mathcal{F}_d = 0$. The uniqueness follows from Theorem 19 applied to $\mathcal{F}_d$. ∎

We will next show the power of the bivariational principle as far as the ECC method is concerned. The standard variational formulation of CC theory introduces a Lagrangian. Error estimates for the CC energy then require that the dual problem has a solution. (See [17], where this nontrivial step has been done by means of the Lax–Milgram theorem.) However, the ECC method is based on the bivariational principle and the energy itself is stationary in this formulation, i.e., the solution $(t_*, \lambda_*)$ is a critical point of the bivariational energy. When $(t_d, \lambda_d)$ is close to the exact solution, we are guaranteed a quadratic error estimate for free. As our last order of business we will discuss this further.

Under the assumption that $H$ supports a ground state with ground-state energy $E_*$, the Rayleigh–Ritz variational principle states that

$$E_* \leq \mathcal{E}_{\text{var}}(\psi) := \frac{\langle \psi, H\psi \rangle}{\langle \psi, \psi \rangle}$$

for any $\psi \in \mathcal{H}$. Minimizing $\mathcal{E}_{\text{var}}$ over trial wavefunctions (say, considering $\mathcal{H}_{\text{appr}} \subset \mathcal{H}$) yields an approximate energy $E_{\text{appr}}$ that also provides an upper bound to $E_*$, i.e., $E_{\text{appr}} \geq E_*$. Furthermore, since $D_\psi \mathcal{E}_{\text{var}}(\psi_*) = 0$, we obtain a second-order error estimate of the energy (see, for instance, (1.4) in [17] and the reference given in connection for more refined estimates)

$$0 \leq E_{\text{appr}} - E_* \leq C\|\psi_{\text{appr}} - \psi_*\|_{\mathcal{H}}^2 \leq C' d(\mathcal{H}_{\text{appr}}, \psi_*)^2.$$

In a similar fashion, the critical point condition $D\mathcal{E}_{\text{bivar}}(\psi_*, \psi'_*) = 0$ of the bivariational quotient will give us a second-order error estimate of the ECC energy.

As far as truncations of the double wavefunction space $\mathcal{M} \subset \mathcal{H} \times \mathcal{H}$ is concerned (see (5)), where the bivariational pair $(\psi, \tilde{\psi})$ is an element, we will use

$$\mathcal{M}_d := \{(\psi, \tilde{\psi}) : \psi = e^T \phi_0, \tilde{\psi} = e^{-T^\dagger} e^\Lambda \phi_0, \quad t, \lambda \in \mathcal{V}_d\}.$$

Since $\mathcal{M}_d$ is closed (we assume that $\mathcal{V}_d$ is closed; see the next lemma), we define the distance

$$d(\mathcal{M}_d,(\psi_*,\tilde{\psi}_*)) := \min_{(\psi,\tilde{\psi})\in\mathcal{M}_d} \|(\psi,\tilde{\psi})-(\psi_*,\tilde{\psi}_*)\|_{\mathcal{H}\times\mathcal{H}},$$

where $\|(\cdot,\cdot)\|^2_{\mathcal{H}\times\mathcal{H}} := \|\cdot\|^2_{\mathcal{H}} + \|\cdot\cdot\|^2_{\mathcal{H}}$.

LEMMA 23. *Assume that $\mathcal{V}_d$ is closed. Then $\mathcal{M}_d$ is closed. Moreover, it holds that*

$$d(\mathcal{V}_d,t_*)^2 + d(\mathcal{V}_d,\lambda_*)^2 \le C\, d(\mathcal{M}_d,(\psi_*,\tilde{\psi}_*))^2 \tag{34}$$

*for some constant $C$.*

*Proof.* By Lemma 8, the map $\Phi : (t,\lambda) \mapsto (e^T\phi_0, e^{-T^\dagger}e^\Lambda\phi_0)$ and its inverse are smooth and $\mathcal{M}_d = \Phi(\mathcal{V}_d\times\mathcal{V}_d)$ is closed since $\mathcal{V}_d$ is.

For (34), we first note that

$$\begin{aligned} d(\mathcal{M}_d,(\psi_*,\tilde{\psi}_*))^2 = \min_{t,\lambda\in\mathcal{V}_d} \big(&\|e^T\phi_0 - e^{T_*}\phi_0\|^2_{\mathcal{H}} \\ &+ \|e^{-T^\dagger}e^\Lambda\phi_0 - e^{-T_*^\dagger}e^{\Lambda_*}\phi_0\|^2_{\mathcal{H}}\big). \end{aligned}$$

This gives (where we let $C$ be a constant that is redefined and reused at leisure)

$$\begin{aligned} d(\mathcal{V}_d,\lambda_*)^2 &\le C \min_{\lambda\in\mathcal{V}_d} \|e^\Lambda\phi_0 - e^{\Lambda_*}\phi_0\|^2_{\mathcal{H}} \\ &\le C\big(\min_{t,\lambda\in\mathcal{V}_d} \|e^{T_*^\dagger}\|^2_{\mathcal{B}(\mathcal{H})}\big(\|e^{-T^\dagger}e^\Lambda\phi_0 - e^{-T_*^\dagger}e^{\Lambda_*}\phi_0\|^2_{\mathcal{H}} \\ &\quad + \|e^{-T^\dagger}-e^{-T_*^\dagger}\|_{\mathcal{B}(\mathcal{H})}\|e^\Lambda\|^2_{\mathcal{B}(\mathcal{H})}\big)\big) \\ &\le C\big(\min_{t,\lambda\in\mathcal{V}_d} \|e^{-T^\dagger}e^\Lambda\phi_0 - e^{-T_*^\dagger}e^{\Lambda_*}\phi_0\|^2_{\mathcal{H}} + \min_{t\in\mathcal{V}_d}\|e^T\phi_0 - e^{T_*}\phi_0\|^2_{\mathcal{H}}\big) \\ &\le C\, d((\mathcal{H}\times\mathcal{H})_d,(\psi_\perp,\tilde{\psi}_\perp))^2. \end{aligned}$$

The desired inequality then follows from

$$d(\mathcal{V}_d,t_*) \le D \min_{t\in\mathcal{V}_d} ||e^T\phi_0 - e^{T_*}\phi_0||_{\mathcal{H}} \le D\, d(\mathcal{M}_d,(\psi_*,\tilde{\psi}_*)).$$

∎

THEOREM 24. *Let $\delta>0$ be such that $\mathcal{F}$ is strongly monotone (with constant $\gamma$) and Lipschitz continuous (with constant $L$) for $(t,\lambda)\in B_\delta(t_*,\delta_*)$ and assume that $\mathcal{V}_d$ is a sufficiently good approximation of $\mathcal{V}$. If $(t_d,\lambda_d)\in\mathcal{V}_d\times\mathcal{V}_d$ is the solution of $\mathcal{F}_d=0$ and $(t_*,\lambda_*)\in\mathcal{V}\times\mathcal{V}$ is the (exact) solution of $\mathcal{F}=0$, then the following hold:*

(i) *With $E_d := \mathcal{E}(t_d,\lambda_d)$ there exist constants $d_1,d_2$ such that*

$$|E_d - E_*| \le d_1\|t_d-t_*\|^2_{\mathcal{V}} + d_2\|t_d-t_*\|_{\mathcal{V}}\|\lambda_d-\lambda_*\|_{\mathcal{V}} \tag{35}$$

*and with $C_*$ as before there holds*

$$\begin{aligned} |E_d - E_*| \le (C_* + \mathcal{O}(\|t_*\|_{\mathcal{V}}) + \mathcal{O}(\|\lambda_*\|_{\mathcal{V}}))\frac{L^2}{2\gamma^2}\big(d(\mathcal{V}_d,t_*)^2 + d(\mathcal{V}_d,\lambda_*)^2\big) \\ + \mathcal{O}\big(\max(d(\mathcal{V}_d,t_*),d(\mathcal{V}_d,\lambda_*))^3\big). \end{aligned} \tag{36}$$

(ii) *Letting $\psi_* = e^{T_*}\phi_0$, $\psi_d = e^{T_d}\phi_0$, $\tilde{\psi}_* = e^{-T_*^\dagger}e^{\Lambda_*}\phi_0$ and $\tilde{\psi}_d = e^{-T_d^\dagger}e^{\Lambda_d}\phi_0$, there exist $\tilde{d}_1,\tilde{d}_2$ such that*

$$|E_d - E_*| \le \tilde{d}_1\|\psi_d-\psi_*\|^2_{\mathcal{H}} + \tilde{d}_2\|\psi_d-\psi_*\|_{\mathcal{H}}\|\tilde{\psi}_d-\tilde{\psi}_*\|_{\mathcal{H}}. \tag{37}$$

*Furthermore, there exists a constant $\tilde{C}$ such that*

$$|E_d - E_*| \le \tilde{C}\, d(\mathcal{M}_d,(\psi_*,\tilde{\psi}_*))^2 + \mathcal{O}\big(d(\mathcal{M}_d,(\psi_*,\tilde{\psi}_*))^3\big). \tag{38}$$

*Proof.* (i) Taylor expanding $\mathcal{E}(t,\lambda)$ at $(t_*,\lambda_*)$ and using the notation $g_d := t_d - t_*$ and $k_d := \lambda_d - \lambda_*$, we obtain (by Taylor's theorem)

$$E_d - E_* = \frac{1}{2}D^2\mathcal{E}(t_*,\lambda_*)((g_d,k_d)^2) + \frac{1}{2}\int_0^1 (1-r)^2 D^3\mathcal{E}((t_*,\lambda_*)+r(g_d,k_d))((g_d,k_d)^3)dr.$$

From this it is clear that

$$2|E_d - E_*| \leq |D^2\mathcal{E}(t_*,\lambda_*)((g_d,k_d)^2)| + \mathcal{O}\big(\max(d(\mathcal{V}_d,t_*),d(\mathcal{V}_d,\lambda_*))^3\big). \tag{39}$$

By straightforward differentiation with respect to the amplitudes $t_\mu$ and $\lambda_\mu$,

$$(D^2\mathcal{E}(t,\lambda))_{\mu,\nu} = \begin{bmatrix} \langle\phi_0, e^{\Lambda^\dagger}[[H_t,X_\mu],X_\nu]\phi_0\rangle & \langle\phi_\nu, e^{\Lambda^\dagger}[H_t,X_\mu]\phi_0\rangle \\ \langle\phi_\mu, e^{\Lambda^\dagger}[H_t,X_\nu]\phi_0\rangle & \langle X_\mu X_\nu\phi_0, e^{\Lambda^\dagger}H_t\phi_0\rangle \end{bmatrix}.$$

We next note that

$$\begin{aligned}
&\frac{1}{2}D^2\mathcal{E}(t_*,\lambda_*)((g_d,k_d)^2) \\
&\quad= \frac{1}{2}\big(\langle\phi_0, e^{\Lambda_*^\dagger}[[H_{t_*},G_d],G_d]\phi_0\rangle + 2\langle K_d\phi_0, e^{\Lambda_*^\dagger}[H_{t_*},G_d]\phi_0\rangle + \langle K_d^2\phi_0, e^{\Lambda_*^\dagger}H_{t_*}\phi_0\rangle\big) \\
&\quad= \frac{1}{2}\big(\langle e^{\Lambda_*}\phi_0, \big(H_{t_*}G_d^2 - 2G_dH_{t_*}G_d + G_d^2H_{t_*}\big)\phi_0\rangle + 2\langle e^{\Lambda_*}K_d\phi_0, [H_{t_*},G_d]\phi_0\rangle \\
&\qquad + \langle e^{\Lambda_*}K_d^2\phi_0, H_{t_*}\phi_0\rangle\big).
\end{aligned}$$

Using Lemma 13, specifically $H_{t_*}\phi_0 = E_*\phi_0$ and $H_{t_*}^\dagger e^{\Lambda_*}\phi_0 = E_*e^{\Lambda_*}\phi_0$, the following equality holds:

$$\begin{aligned}
&\frac{1}{2}D^2\mathcal{E}(t_*,\lambda_*)((g_d,k_d)^2) \\
&\quad= \frac{1}{2}\big(2\langle e^{\Lambda_*}\phi_0, G_d(E_*-H_{t_*})G_d\phi_0\rangle + 2\langle e^{\Lambda_*}K_d\phi_0, (H_{t_*}-E_*)G_d\phi_0\rangle\big).
\end{aligned}$$

Furthermore, since $e^{\Lambda_*}$ and $K_d$ commute, we obtain

$$\begin{aligned}
&\frac{1}{2}|D^2\mathcal{E}(t_*,\lambda_*)((g_d,k_d)^2)| \\
&\quad= |\langle G_d^\dagger e^{\Lambda_*}\phi_0, (E_*-H_{t_*})G_d\phi_0\rangle + \langle e^{\Lambda_*}K_d\phi_0, (H_{t_*}-E_*)G_d\phi_0\rangle| \\
&\quad= |\langle e^{-T_*^\dagger}\big(G_d^\dagger(e^{\Lambda_*}-I) - e^{\Lambda_*}K_d\big)\phi_0, (E_*-H)e^{T_*}G_d\phi_0\rangle| \\
&\quad\leq C_*\|e^{-T_*^\dagger}\big(G_d^\dagger(e^{\Lambda_*}-I) - e^{\Lambda_*}K_d\big)\phi_0\|_{\mathcal{H}}\|e^{T_*}G_d\phi_0\|_{\mathcal{H}} \\
&\quad\leq C_*\|e^{-T_*^\dagger}\|_{\mathcal{B}(\mathcal{H})}\big(\|G_d^\dagger\|_{\mathcal{B}(\mathcal{H})}\|e^{\Lambda_*}-I\|_{\mathcal{B}(\mathcal{H})}\|\phi_0\|_{\mathcal{H}} \\
&\qquad + \|e^{\Lambda_*}\|_{\mathcal{B}(\mathcal{H})}\|K_d\phi_0\|_{\mathcal{H}}\big)\|e^{T_*}\|_{\mathcal{B}(\mathcal{H})}\|G_d\phi_0\|_{\mathcal{H}} \\
&\quad\leq C_*\|e^{-T_*^\dagger}\|_{\mathcal{B}(\mathcal{H})}\|e^{T_*}\|_{\mathcal{B}(\mathcal{H})}\big(c\|\phi_0\|_{\mathcal{H}}\|e^{\Lambda_*}-I\|_{\mathcal{B}(\mathcal{H})}\|t_d-t_*\|_{\mathcal{V}}^2 \\
&\qquad + \|e^{\Lambda_*}\|_{\mathcal{B}(\mathcal{H})}\|t_d-t_*\|_{\mathcal{V}}\|\lambda_d-\lambda_*\|_{\mathcal{V}}\big) \\
&\quad=: D_1\|t_d-t_*\|_{\mathcal{V}}^2 + D_2\|t_d-t_*\|_{\mathcal{V}}\|\lambda_d-\lambda_*\|_{\mathcal{V}},
\end{aligned} \tag{40}$$

where in the last step we defined the constants $D_1 := D_1(t_*,\lambda_*,\phi_0)$ and $D_2 := D_2(t_*,\lambda_*)$. Thus, by (39) we can choose $d_1$ and $d_2$, under the assumption that $\max(d(\mathcal{V}_d,t_*),d(\mathcal{V}_d,\lambda_*))$ is sufficiently small, such that (35) holds.

To obtain (36), we see that (40) gives

$$\begin{aligned}
&\frac{1}{2}|D^2\mathcal{E}(t_*,\lambda_*)((g_d,k_d)^2)| \\
&\quad\le C_*\|e^{-T_*^\dagger}\|_{\mathcal{B}(\mathcal{H})}\|e^{T_*}\|_{\mathcal{B}(\mathcal{H})}\big(c\|\phi_0\|_{\mathcal{H}}\|e^{\Lambda_*}-I\|_{\mathcal{B}(\mathcal{H})}+\frac{1}{2}\|e^{\Lambda_*}\|_{\mathcal{B}(\mathcal{H})}\big) \\
&\qquad\times\big(\|t_d-t_*\|_{\mathcal{V}}^2+\|\lambda_d-\lambda_*\|_{\mathcal{V}}^2\big) \\
&\quad\le (C_*+\mathcal{O}(\|t_*\|_{\mathcal{V}})+\mathcal{O}(\|\lambda_*\|_{\mathcal{V}}))\frac{L^2}{2\gamma^2}\big(d(\mathcal{V}_d,t_*)^2+d(\mathcal{V}_d,\lambda_*)^2\big),
\end{aligned}$$

where we used (31).

(ii) Next, using Theorem 1 (equation (2)), (40) gives

$$\frac{1}{2}|D^2\mathcal{E}(t_*,\lambda_*)((g_d,k_d)^2)|\le \tilde{D}_1\|\psi_d-\psi_*\|_{\mathcal{H}}^2+\tilde{D}_2\|\psi_d-\psi_*\|_{\mathcal{H}}\|(e^{\Lambda_d}-e^{\Lambda_*})\phi_0\|_{\mathcal{H}}. \tag{41}$$

Furthermore, we use

$$\begin{aligned}
e^{\Lambda_d}-e^{\Lambda_*} &= e^{T_*^\dagger}e^{-T_*^\dagger}(e^{\Lambda_d}-e^{\Lambda_*}) \\
&= e^{T_*^\dagger}\big(e^{-T_d^\dagger}e^{\Lambda_d}-e^{-T_*^\dagger}e^{\Lambda_*}-(e^{-T_d^\dagger}-e^{-T_*^\dagger})e^{\Lambda_d}\big),
\end{aligned}$$

and we obtain

$$\begin{aligned}
\|(e^{\Lambda_d}-e^{\Lambda_*})\phi_0\|_{\mathcal{H}} &\le \|e^{T_*^\dagger}\|_{\mathcal{B}(\mathcal{H})}\big(\|\tilde{\psi}_d-\tilde{\psi}_*\|_{\mathcal{H}}+\|e^{\Lambda_d}\|_{\mathcal{B}(\mathcal{H})}\|(e^{-T_d^\dagger}-e^{-T_*^\dagger})\phi_0\|_{\mathcal{H}}\big) \\
&\le \tilde{D}\|\tilde{\psi}_d-\tilde{\psi}_*\|_{\mathcal{H}}+\tilde{D}'\|\psi_d-\psi_*\|_{\mathcal{H}}.
\end{aligned} \tag{42}$$

Inserting (42) into (41) gives

$$\frac{1}{2}|D^2\mathcal{E}(t_*,\lambda_*)((g_d,k_d)^2)|\le \tilde{D}_1'\|\psi_d-\psi_*\|_{\mathcal{H}}^2+\tilde{D}_2'\|\psi_d-\psi_*\|_{\mathcal{H}}\|\tilde{\psi}_d-\tilde{\psi}_*\|_{\mathcal{H}}.$$

Repeating the argument made in (i) for (35), we can find constants $\tilde{d}_1,\tilde{d}_2$ such that (37) holds.

To finish the proof, we use (34) in Lemma 23 that together with the proof of (i) gives (38). ∎

**4. Conclusions.** In this article we have put the formalism of Arponen's ECC method on firm mathematical ground. This has been achieved by generalizing the continuous (infinite-dimensional) formulation of standard CC theory in [16, 17] to the ECC formalism. The bivariational principle plays an important role in our analysis. With the bivariational energy $\mathcal{E}(t,\lambda)$ (and its derivatives) as the main object of study, we have derived existence and uniqueness results for the ECC equation $\mathcal{F}=0$ (the flipped gradient) and its discretizations $\mathcal{F}_d=0$. The key aspect of the analysis is the establishment of locally strong monotonicity of $\mathcal{F}$ at the exact solution $(t_*,\lambda_*)$. This has been achieved either by assuming that the reference $\phi_0$ is sufficiently good an approximation of the exact solution $\psi_*$ or by considering certain splittings of the Hamiltonian $H$.

We have formulated and proved quadratic error estimates in terms of the quality of the truncated amplitude space $\mathcal{V}_d$. The energy error has been bound in terms of $d(\mathcal{V}_d,t_*)$ and $d(\mathcal{V}_d,\lambda_*)$, or equivalently $d(\mathcal{M}_d,(\psi_*,\tilde{\psi}_*))$, where $(\psi_*,\tilde{\psi}_*)$ is the exact wavefunction pair and $\mathcal{M}_d$ the truncation of $\mathcal{H}\times\mathcal{H}$.

It is interesting to note, as ECC is variational by construction, i.e., the solution $(t_*, \lambda_*)$ is a critical point of the smooth map $\mathcal{E}$, that the error estimate is obtained basically for free. Indeed, the CC Lagrangian $\mathcal{L}$ can be thought of as a linearized formulation of ECC where the second set of amplitudes $\{\lambda_\mu\}$ are the Lagrange multipliers $\{z_\mu\}$. The dual problem of CC is, as it were, already built into the ECC theory. This again illustrates the benefit of applying the bivariational point of view.

Here, ECC has been formulated in a set of cluster amplitude coordinates that are not usually employed. A next step in the study of the ECC method would be to repeat the analysis of the monotonicity of $\mathcal{F}$ and to obtain error estimates using the so-called canonical cluster amplitudes; cf. Remark 10.

Even if ECC is currently not a practical tool in computational chemistry due to its complexity, our analysis demonstrates an important fact: The bivariational principle can be utilized to devise computational schemes that are not obtainable from the standard Rayleigh–Ritz principle but still have a quadratic error estimate. Such schemes include both the traditional CC method and the ECC method. Indeed, not being variational in the Rayleigh–Ritz sense has been the single most important critique of the CC method, precisely due to the lack of a quadratic error estimate. Moreover, we believe that the approach taken in this article, by showing the monotonicity of the flipped gradient $\mathcal{F}$, is an approach that may allow existence and uniqueness results in much more general settings.

## REFERENCES


[1] J. Arponen, *Variational principles and linked-cluster exp S expansions for static and dynamic many-body problems*, Ann. Phys., 151 (1983), pp. 311–382.

[2] R. Bishop, *An overview of coupled cluster theory and its applications in physics*, Theor. Chim. Acta, 80 (1991), pp. 95–148.

[3] J. Čížek, *On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods*, J. Chem. Phys., 45 (1966), pp. 4256–4266.

[4] J. Čížek, *Origins of the coupled cluster technique for atoms and molecules*, Theor. Chim. Acta, 80 (1991), pp. 91–94.

[5] F. Coester, *Bound states of a many-particle system*, Nuclear Phys., 7 (1958), pp. 421–424.

[6] F. Coester and H. Kümmel, *Short-range correlations in nuclear wave functions*, Nuclear Phys., 17 (1960), pp. 477–485.

[7] D. J. Dean and M. Hjorth-Jensen, *Coupled-cluster approach to nuclear physics*, Phys. Rev. C, 69 (2004), 054320.

[8] E. Emmrich, *Gewöhnliche und Operator-Differentialgleichungen*, Vieweg, Wiesbaden, Germany, 2004.

[9] T. Helgaker and P. Jørgensen, *Analytical calculation of geometrical derivatives in molecular electronic structure theory*, Adv. Quant. Chem., 19 (1988), pp. 183–245.

[10] T. Helgaker and P. Jørgensen, *Configuration-interaction energy derivatives in a fully variational formulation*, Theor. Chim. Acta, 75 (1989), pp. 111–127.

[11] H. Kümmel, *Origins of the coupled cluster method*, Theor. Chim. Acta, 80 (1991), pp. 81–89.

[12] S. Kvaal, *Ab initio quantum dynamics using coupled-cluster*, J. Chem. Phys., 136 (2012), 194109.

[13] S. Kvaal, *Variational formulations of the coupled-cluster method in quantum chemistry*, Mol. Phys., 111 (2013), pp. 1100–1108.

[14] J. Paldus, *The beginnings of coupled-cluster theory: An eyewitness account*, in Theory and Applications of Computational Chemistry: The First Forty Years, C. Dykstra, G. Frenking, K. Kim, and G. Scuseria, eds., Elsevier, New York, 2005, pp. 115–147.

[15] J. Paldus, J. Čížek, and I. Shavitt, *Correlation problems in atomic and molecular systems.* IV. *Extended coupled-pair many-electron theory and its application to the BH*3 *moleciule*, Phys. Rev. A, 5 (1972), pp. 50–67.

[16] T. Rohwedder, *The continuous coupled cluster formulation for the electronic Schrödinger equation*, ESAIM Math. Model. Numer. Anal., 47 (2013), pp. 421–447.

[17] T. Rohwedder and R. Schneider, *Error estimates for the coupled cluster method*, ESAIM Math. Model. Numer. Anal., 47 (2013), pp. 1553–1582.
[18] R. Schneider, *Analysis of the projected coupled cluster method in electronic structure calculation*, Numer. Math., 113 (2009), pp. 433–471.
[19] O. Sinanoğlu, *Many-electron theory of atoms and molecules.* I. *Shells, electron pairs vs. many-electron correlations*, J. Chem. Phys., 36 (1962), pp. 706–717.
[20] H. Yserentant, *Regularity and Approximability of Electronic Wavefunctions*, Lecture Notes In Math. 2000, Springer, New York, 2010.
[21] E. Zarantonello, *Solving Functional Equations by Contractive Averaging*, Tech. Report 160, U.S. Army Mathematics Research Center, Madison, WI, 1960.
[22] E. Zeidler, *Nonlinear Functional Analysis and Its Application* II/*B*, Springer, New York, 1990.