

## Molecular Physics

An International Journal at the Interface Between Chemistry and Physics

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/tmph20>

# Guaranteed convergence for a class of coupled-cluster methods based on Arponen's extended theory

Simen Kvaal , Andre Laestadius & Tilmann Bodenstein

To cite this article: Simen Kvaal , Andre Laestadius & Tilmann Bodenstein (2020) Guaranteed convergence for a class of coupled-cluster methods based on Arponen's extended theory, Molecular Physics, 118:19-20, e1810349, DOI: [10.1080/00268976.2020.1810349](https://doi.org/10.1080/00268976.2020.1810349)

To link to this article: <https://doi.org/10.1080/00268976.2020.1810349>



© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 26 Aug 2020.



Submit your article to this journal [↗](#)



Article views: 174



View related articles [↗](#)



View Crossmark data [↗](#)

# Guaranteed convergence for a class of coupled-cluster methods based on Arponen's extended theory

Simen Kvaal, Andre Laestadius and Tilmann Bodenstern

Hylleraas Centre for Quantum Molecular Sciences, Department of Chemistry, University of Oslo, Oslo, Norway

## ABSTRACT

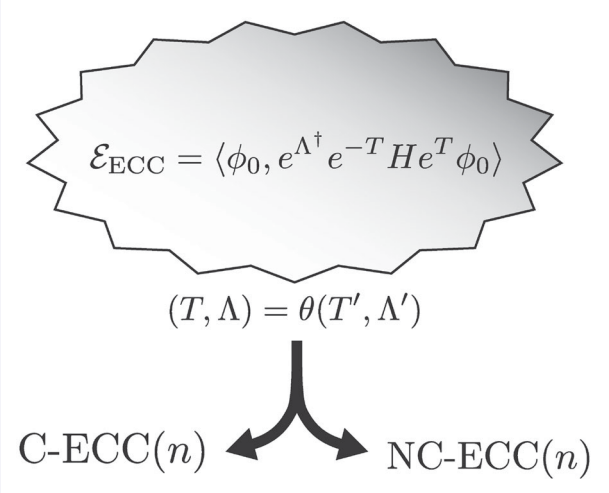
A wide class of coupled-cluster methods is introduced, based on Arponen's extended coupled-cluster theory. This class of methods is formulated in terms of a coordinate transformation of the cluster operators. The mathematical framework for the error analysis of coupled-cluster methods based on Arponen's bivariational principle is presented, in which the concept of local strong monotonicity of the flipped gradient of the energy is central. A general mathematical result is presented, describing sufficient conditions for coordinate transformations to preserve the local strong monotonicity. The result is applied to the presented class of methods, which include the standard and quadratic coupled-cluster methods, and also Arponen's canonical version of extended coupled-cluster theory. Some numerical experiments are presented, and the use of canonical coordinates for diagnostics is discussed.

## ARTICLE HISTORY

Received 13 March 2020  
Accepted 29 June 2020

## KEYWORDS

Coupled-cluster method;  
extended coupled-cluster  
method; error analysis;  
electronic-structure theory


$$\mathcal{E}_{\text{ECC}} = \langle \phi_0, e^{\Lambda^\dagger} e^{-T} H e^T \phi_0 \rangle$$
$$(T, \Lambda) = \theta(T', \Lambda')$$

C-ECC( $n$ )      NC-ECC( $n$ )



## 1. Introduction

It is with delight that the authors dedicate this work to Professor Jürgen Gauß on the occasion of his sixtieth birthday. In the spirit of his pursuit of scientific rigour, especially the attention to detail in coupled-cluster (CC) theory, we here present a mathematical study of some alternative formulations based on Arponen's extended CC (ECC) method [1,2]. The ECC method is defined in terms of the critical points of an energy functional,

$$\mathcal{E}_{\text{ECC}}(T, \Lambda) = \langle \phi_0, e^{\Lambda^\dagger} e^{-T} H e^T \phi_0 \rangle, \quad (1)$$

where  $T$  and  $\Lambda$  are cluster operators in the usual sense of CC theory. The well-known CC Lagrangian introduced by Helgaker and Jørgensen [3] is obtained by expanding  $e^{\Lambda^\dagger}$  to first order. In this sense, the standard CC approach is an approximation to the ECC method.

We will study a collection of methods that generalises this idea, defined by substitution of  $e^{\Lambda^\dagger}$  by a Taylor polynomial of fixed degree  $n$ . This can be formulated as a coordinate transformation of the cluster amplitudes. A second class of models is obtained by a further coordinate transformation introduced by Arponen

**CONTACT** Simen Kvaal  [simen.kvaal@kjemi.uio.no](mailto:simen.kvaal@kjemi.uio.no)  Hylleraas Centre for Quantum Molecular Sciences, Department of Chemistry, University of Oslo, P.O. Box 1033 Blindern, Oslo N-0315, Norway

to ensure a certain linkedness structure of the energy functional. We refer to these coordinates as *canonical*, as the time-dependent Schrödinger equation takes the form of Hamilton's equations of motion in this case. Thus, we obtain two hierarchies NC-ECC( $n$ ), using non-canonical coordinates, and C-ECC( $n$ ) using canonical coordinates. Our mathematical results imply that when the cluster operators are not truncated, all these models are exact and equivalent to the time-independent Schrödinger equation. Moreover, Galerkin approximations (i.e. generic truncation schemes that can approach the untruncated limit) will converge under certain relatively mild single-reference-type conditions. While the methods discussed here are all expensive (for  $n > 1$ ), we consider them as stepping stones towards producing competitive alternatives to standard CC theory that alleviate deficiencies of the latter, such as the inability to correctly break chemical bonds.

The various forms of CC methods are today among the most widely used for wavefunction-based calculations on manybody systems. The main idea stems from Hubbard's exponential parameterisation of the wavefunction based on cluster operators in manybody perturbation theory [4], which was taken as starting point for *ab initio* treatments by Coester and Kümmel for nuclear structure calculations in the 1950s [5,6]. The modern form of standard CC theory was developed by, among others, Sinanoğlu, Paldus and Cizek in the 1960s [7] and the CC method with singles, doubles and perturbative triples [CCSD(T)] today constitutes 'the gold standard of quantum chemistry' due to its excellent balance between computational cost and accuracy [8]. In nuclear structure calculations the same method has gained traction in the last decade, providing excellent predictive power for light to medium nuclei [9]. Coupled-cluster theory has also been applied to superconductivity [10], lattice gauge theory [11], and systems of trapped bosons such as Bose–Einstein condensates [12]. These examples and the cited works are by no means exhaustive, but serve to illustrate the flexibility of the CC formalism.

In the early 1980s, Arponen introduced a novel concept into CC theory, namely *the bivariational principle* [1,13], resulting in the ECC method [1,2,14], and an interpretation of standard CC theory and ECC theory as variational methods in a more general sense, i.e. they are bivariational. However, the ECC method has seen little use in chemistry due to its immense complexity, even for truncated versions. In physics, on the other hand, the ECC model has advantages over standard CC theory that can make it very useful. To illustrate, the ECC method correctly describes symmetry breaking in the Lipkin–Meshkov–Glick quasispin model of collective monopole vibrations in nuclei [13,15], in contrast to the

standard CC method, which cannot. For the electronic-structure problem in quantum chemistry, the standard CC model fails dramatically to reproduce dissociation curves of even simple dimers like  $N_2$ , while the ECC method performs quite well [16,17]. Thus, we are of the opinion that the ECC method is still worthwhile to study, and approximate forms may still prove to be useful in quantum chemistry.

The non-canonical and canonical hierarchies (N)C-ECC( $n$ ) introduced in this article turn out to be equivalent, and give identical predictions, when truncated with an excitation-rank complete scheme. On the other hand, the working equations are different and in fact cheaper in the canonical case, albeit marginally. An example is the NC-ECC(1)SD method, i.e. the standard CCSD approach, and the C-ECC(1)SD method, which are equivalent. We also raise the question about diagnostics for practical calculations, and show some numerical evidence that diagnostics can favourably be done using canonical coordinates, even if the computations are done in the usual manner using non-canonical variables.

Another well-known special case is NC-ECC(2), the quadratic coupled-cluster (QCC) method introduced by Van Vorhiis and Head-Gordon [18,19], whose canonical and non-canonical versions are equivalent in their doubles-only approximation. We note that the asymptotic cost of QCCD is the same as CCSD, even if it is a higher-order approximation to ECC. Furthermore, the perfect-pairing (PP) hierarchy [20] of amplitude truncation schemes can be applied to our methods. The PP hierarchy are approximations to the complete-active space self-consistent field (CASSCF) method, including only a tiny subset of even-rank amplitudes combined with orbital optimisation, the latter which we disregard here. The corresponding canonical and non-canonical formulations (N)C-ECC(1)PPH are inequivalent. The  $n > 1$  versions could also be interesting in their own right, as investigated by Byrd and coworkers in the case of the QCC method [19].

The remainder of the article is organised as follows: In Section 2, we introduce the bivariational principle and the mathematical setting of local analysis of CC methods. The key concept of our analysis is the notion of local strong monotonicity of the flipped gradient of a smooth bivariational energy functional (see Equation (7)). The usefulness of this property is presented in Theorem 2.1, where local uniqueness and quadratic error estimates are established in a very general setting using Zangwill's Theorem from nonlinear monotone operator theory [21,22]. Next, Theorem 2.2 summarises the main results of Ref. [23], where strong monotonicity is proven for the non-canonical ECC method. For a recent review on monotonicity in CC theory we refer to [24], where

this property is linked to spectral gaps of the systems under study. Section 3 presents the idea of monotonicity-preserving coordinate transformations. Our main result, Theorem 3.1, is a change-of-coordinates result. When combined with Theorems 2.1 and 2.2, the analysis of (N)C-ECC( $n$ ) follows in Corollary 3.2. Our tools rely heavily on the functional analytic formulation of cluster operators and the Schrödinger equation developed by Rohwedder and Schneider [25–27]. The results are formulated in a qualitative manner, in the sense that they are indeed rigorous but depend on constants whose numerical (or optimal) values are unknown. We leave further quantitative investigations for future work. In Section 4, we perform some numerical experiments to elucidate some aspects of the (N)C-ECC( $n$ ) hierarchies, before we finish with some concluding remarks in Section 5. All the proofs of our results are presented in Appendix.

## 2. The non-canonical extended coupled-cluster model

### 2.1. Bivariational principle

The starting point is a generalisation of the Rayleigh–Ritz variational principle to operators that are not necessarily self-adjoint (Hermitian in the finite-dimensional case). For simplicity, we assume a real Hilbert space  $\mathcal{H}$ . Given a system Hamiltonian  $\hat{H} : D(\hat{H}) \rightarrow \mathcal{H}$ , where  $D(\hat{H}) \subset \mathcal{H}$  is dense, we define a bivariate Rayleigh quotient,  $\mathcal{E}_{\text{bivar}} : \mathcal{H} \oplus \mathcal{H} \rightarrow \mathbb{R}$ ,

$$\mathcal{E}_{\text{bivar}}(\psi, \tilde{\psi}) = \frac{\langle \tilde{\psi}, \hat{H}\psi \rangle}{\langle \tilde{\psi}, \psi \rangle}, \quad \langle \tilde{\psi}, \psi \rangle \neq 0. \quad (2)$$

Requiring the functional  $\mathcal{E}_{\text{bivar}}$  to be stationary at  $(\psi_*, \tilde{\psi}_*)$  with respect to arbitrary variations in the two wavefunctions leads to the conditions  $\langle \tilde{\psi}_*, \psi_* \rangle \neq 0$ ,  $\hat{H}\psi_* = E_*\psi_*$  and  $\hat{H}^\dagger \tilde{\psi}_* = E_*\tilde{\psi}_*$ , with  $E_* = \mathcal{E}_{\text{bivar}}(\psi_*, \tilde{\psi}_*)$ , i.e. the right and left eigenvalue problem for  $\hat{H}$ . If  $\hat{H}$  is self-adjoint, the eigenfunctions are identical up to normalisation. The introduction of two independent wavefunctions therefore might seem to complicate matters. However, the bivariate Rayleigh quotient  $\mathcal{E}$  allows *distinct* approximations of  $\psi$  and  $\tilde{\psi}$ , introducing more flexibility for approximate schemes. Moreover, the *state* defined is a (non-Hermitian) density operator, which is unique,

$$\rho = \frac{|\psi\rangle \langle \tilde{\psi}|}{\langle \tilde{\psi} | \psi \rangle}.$$

When determined variationally, the Hellmann–Feynman theorem [28] gives well-defined physical predictions in terms of  $\rho$ .

As is common in analysis of partial differential equations [29,30], we pass to a weak formulation, which in

this case is *equivalent* to the strong formulation outlined above. Under the assumption that  $\hat{H}$  is below bounded, we can introduce a unique extension  $H : \mathcal{X} \rightarrow \mathcal{X}'$  (dual space), where  $\mathcal{X} \subset \mathcal{H}$  is a dense subspace, a Hilbert space with norm  $\|\cdot\|_{\mathcal{X}}$ , continuously embedded in  $\mathcal{H}$ . It follows that  $\mathcal{H}$  is continuously embedded in  $\mathcal{X}'$ , and we have a scale of spaces with dense embeddings,  $\mathcal{X} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{X}'$ . The operator  $H$  is bounded (i.e. continuous), and satisfies a Gårding estimate, i.e. for some  $\alpha \geq 0$  and some  $\mu \in \mathbb{R}$ ,

$$\langle \psi, H\psi \rangle \geq \alpha \|\psi\|_{\mathcal{X}}^2 + \mu \|\psi\|^2$$

for all  $\psi \in \mathcal{X}$ . For the electronic-structure problem  $\mathcal{X}$  can be taken to be the space of square-integrable functions with finite kinetic energy.

If  $\mathcal{H}$  is finite-dimensional, we can set  $\mathcal{X} \equiv \mathcal{H}$ , simplifying matters a lot, and the reader may if she or he wishes stick to this picture for simplicity, where all operators are basically matrices. In the infinite-dimensional case, however,  $\hat{H}$  is typically unbounded as an operator over  $\mathcal{H}$ , and the above construction is necessary.

Under the stated conditions,  $\mathcal{E}_{\text{bivar}} : \mathcal{X} \oplus \mathcal{X} \rightarrow \mathbb{R}$  is a (Fréchet) smooth map away from the singularity  $\langle \tilde{\psi}, \psi \rangle = 0$ , and Taylor series exist and converge locally, allowing a certain degree of intuition to be borrowed from the finite-dimensional case. The right and left Schrödinger equations are then  $\partial_{\tilde{\psi}} \mathcal{E}_{\text{bivar}}(\psi_*, \tilde{\psi}_*) = 0$  and  $\partial_{\psi} \mathcal{E}_{\text{bivar}}(\psi_*, \tilde{\psi}_*) = 0$ , respectively. This is the bivariational principle.

### 2.2. Exponential ansatz and the ECC method

The standard CC method is formulated relative to a fixed reference  $\phi_0 \in \mathcal{X}$  on determinantal form. By introducing a cluster operator  $T = T_1 + T_2 + \dots$  with  $T_k$  containing all excitations of rank  $k$ , i.e. of  $k$  fermions relative to  $\phi_0$ , we have the exact parameterisation

$$\psi = e^T \phi_0,$$

assuming intermediate normalisation,  $\langle \phi_0, \psi \rangle = 1$ .

Since all excitations commute, the cluster operators form a commutative Banach algebra under suitable conditions which we now describe [25,26]. We expand the cluster operators using amplitudes and basis operators, i.e.  $T = \sum_{\mu \in \mathcal{I}} \tau_{\mu} X_{\mu}$ , where  $X_{\mu}$  excites a number  $n = n(\mu)$  of fermions in the reference into the virtual space, i.e.

$$X_{\mu} = c_{a_1}^\dagger c_{i_1} \cdots c_{a_n}^\dagger c_{i_n},$$

where the  $i_k$  are among the occupied orbitals of  $\phi_0$ , and  $a_k$  among the unoccupied orbitals. The set  $\mathcal{I}$  is the generic set of amplitude indices. We introduce a Hilbert space

$\mathcal{V}$  with norm  $\|T\| = \|T\phi_0\|_{\mathcal{X}}$ , which becomes a useful space for formulating abstract CC theory. Fundamental results include that any  $T \in \mathcal{V}$  is a bounded operator on  $\mathcal{X}$ , such that, e.g.  $\exp(T)$  also is a bounded operator. Moreover,  $T^\dagger$  is also a bounded operator, which means that we can make sense of, e.g.  $\exp(-T)H\exp(T)$ , and that we can represent any intermediately normalised  $\psi \in \mathcal{X}$  as  $\psi = e^T\phi_0$  with  $T \in \mathcal{V}$  unique. Finally, all the elements of the algebra are nilpotent. The Banach algebra structure on  $\mathcal{V}$  allows CC theory to be rigorously formulated in the full, infinite-dimensional case. This was the approach taken in Ref. [23] for a first analysis of NC-ECC theory.

Again, the finite-dimensional case may be kept in mind: In this case, cluster amplitudes are simply finite-dimensional vectors, and the existence of the exponential parameterisation is a trivial result. There is no need to introduce the norm  $\|T\|$ , instead the Euclidean norm on the amplitudes may be used.

Any  $\tilde{\psi}$  normalised according to  $\langle \tilde{\psi}, \psi \rangle = 1$  can be represented by introducing a second cluster operator  $\Lambda = \Lambda_1 + \Lambda_2 + \dots$ , viz.,

$$\tilde{\psi} = e^{-T^\dagger} e^\Lambda \phi_0.$$

Inserting the parametrization of  $\psi$  and  $\tilde{\psi}$  into the bivariate Rayleigh quotient, we obtain the energy functional  $\mathcal{E}_{\text{ECC}} : \mathcal{V} \oplus \mathcal{V} \rightarrow \mathbb{R}$  of the non-canonical ECC method, given in Equation (1). This map is everywhere smooth, and its critical points  $(T_*, \Lambda_*)$  are equivalent to the Schrödinger equation and its dual: Under the assumption that the eigenfunctions can be normalised according to  $\langle \phi_0, \psi_* \rangle = \langle \tilde{\psi}_*, \psi_* \rangle = 1$ ,  $\psi_*$  and  $\tilde{\psi}_*$  solve the Schrödinger equation and its dual if and only if

$$\frac{\partial \mathcal{E}_{\text{ECC}}(T_*, \Lambda_*)}{\partial \Lambda} = 0 \quad \text{and} \quad \frac{\partial \mathcal{E}_{\text{ECC}}(T_*, \Lambda_*)}{\partial T} = 0. \quad (3)$$

Assuming that the eigenvalue  $E_* = \mathcal{E}_{\text{ECC}}(T_*, \Lambda_*)$  is non-degenerate,  $(T_*, \Lambda_*)$  is easily seen to be locally unique.

### 2.3. Truncations and monotonicity analysis

The non-canonical ECC energy is just one out of many possible parameterisations of the exact bivariate Rayleigh quotient  $\mathcal{E}_{\text{bivar}}$ . In this section, we take a more abstract approach and consider a general energy functional  $\mathcal{E} : \mathcal{V} \oplus \mathcal{V} \rightarrow \mathbb{R}$ , obtained by some exact parameterisation of  $(\psi, \tilde{\psi})$  by means of the space  $\mathcal{V} \oplus \mathcal{V}$ , i.e. by a pair of cluster operators  $(T, \Lambda)$ . We will discuss several such functionals in Section 3, obtained from the NC-ECC functional by coordinate transformations.

Only in rare cases can the amplitude equations (3) be solved exactly. Introduce therefore a discretised space

$\mathcal{V}_d \subset \mathcal{V}$  of finite dimension by truncating the amplitude index set  $\mathcal{I}_d \subset \mathcal{I}$ , that is,  $T_d \in \mathcal{V}_d$  if and only if

$$T_d = \sum_{\mu \in \mathcal{I}_d} \tau_{d,\mu} X_\mu \in \mathcal{V}_d. \quad (4)$$

The set  $\mathcal{I}_d$  is typically defined by the restriction of the excitations to a finite virtual space (a finite basis), and to a finite excitation rank (smaller than the number of electrons). In the chemistry literature, the excitation hierarchy for a given basis is traditionally denoted singles (S), doubles (D), and so on. In the ECC literature, one typically speaks of the SUB $n$  approximation, with  $n$  being the maximum rank.

When the discrete space is established, we define a discrete solution by the stationary conditions of the restricted energy function  $\mathcal{E}_d = \mathcal{E} \upharpoonright_{\mathcal{V}_d \oplus \mathcal{V}_d}$ . The stationary equations take the form

$$\frac{\partial \mathcal{E}(T_{d*}, \Lambda_{d*})}{\partial \lambda_\mu} = \frac{\partial \mathcal{E}(T_{d*}, \Lambda_{d*})}{\partial \tau_\mu} = 0, \quad (5)$$

for all  $\mu \in \mathcal{I}_d$ .

It is not *necessary* to use the traditional truncation scheme outlined here; any increasing sequence of subspaces  $\mathcal{V}_d \subset \mathcal{V}$ , with  $d$  a parameter, that can approximate elements in  $\mathcal{V}$  arbitrarily well by increasing  $d$  can be used. We let  $\text{dist}(v, \mathcal{V}_d)$  be the distance from  $v$  to  $\mathcal{V}_d$  measured with respect to the norm of  $\mathcal{V}$ . Consequently, for all  $v \in \mathcal{V}$  we have  $\text{dist}(v, \mathcal{V}_d) \rightarrow 0$  as  $d \rightarrow +\infty$ . Such a sequence of spaces is referred to as a Galerkin sequence. Other options than the traditional truncation schemes are explicitly correlated methods [31] and complete-active space methods [20,32,33] such as the PP hierarchy.

An often overlooked point in the physics literature is the fact that convergence of the *equations* does not in general imply convergence of their *solutions*. An important question is therefore whether the discrete critical points  $(T_{d*}, \Lambda_{d*})$  converge to the exact critical points  $(T_*, \Lambda_*)$  as  $d \rightarrow +\infty$ . This would imply that the energy converges too, and in a quadratic manner due to the critical point formulation.

*Monotonicity* is an important notion in connection with the local analysis of the CC method and its variations [23–27,34]. The use of monotonicity in the analysis of the standard CC method was introduced by Schneider and Rohwedder [25,27]. The intuition behind (strict) monotonicity is that of an everywhere (strictly) increasing (or decreasing) function. The function that one studies is typically a root problem such as the CC amplitude equations  $f(T) = 0$ , where  $f_\mu(T) = \langle \phi_\mu, e^{-T} He^T \phi_0 \rangle$ . Monotonicity allows the establishment of locally unique solutions of the Galerkin problem and is therefore important for the motivation of numerical

implementations. As such it is a fundamental result of the CC method's practical usage in quantum chemistry. It also connects spectral gaps, e.g. HOMO-LUMO gap, to stability constants within the analysis [24]. (See also the steerable CAS-ext gap connected to the tailored CC method [35] that treats quasi-degenerate systems [36].)

The particular monotonicity property that is key for this presentation is an even stronger version than that of strict monotonicity, and is called strong monotonicity. It is defined as follows: A finite-dimensional vector-valued function  $F(Z)$  is locally strongly monotone near some  $Z_*$  if for  $Z_1, Z_2$  in a neighbourhood of  $Z_*$  we have

$$\langle F(Z_1) - F(Z_2), Z_1 - Z_2 \rangle \geq \eta \|Z_1 - Z_2\|^2, \quad (6)$$

for some constant  $\eta > 0$ . (In the infinite-dimensional case  $\langle \cdot, \cdot \rangle$  is the dual pairing, which then becomes an infinite sum, see Ref. [23] for more details.)

Furthermore, we need the concept of Lipschitz continuity:  $F$  is locally Lipschitz continuous with constant  $L > 0$  if

$$\|F(Z_1) - F(Z_2)\| \leq L \|Z_1 - Z_2\|.$$

In particular, any (Fréchet) smooth function is locally Lipschitz continuous, and so are all its derivatives.

The map  $F$  that we will study is the *flipped gradient* of the general energy functional  $\mathcal{E} : \mathcal{V} \oplus \mathcal{V} \rightarrow \mathbb{R}$ , defined as

$$F(T, \Lambda) = (\partial_\Lambda \mathcal{E}(T, \Lambda), \partial_T \mathcal{E}(T, \Lambda)), \quad (7)$$

or more compactly  $F(T, \Lambda) = R\partial\mathcal{E}(T, \Lambda)$ , with  $R$  being the map that exchanges the partial derivatives. The motivation is as follows: If we consider the bivariate Rayleigh quotient,  $\partial\mathcal{E}_{\text{bivar}}$  is *not* locally strongly monotone, as its critical points are saddle points. On the other hand, the flipped gradient  $F_{\text{bivar}} = R\partial\mathcal{E}_{\text{bivar}}$  can be seen to be locally strongly monotone near the *ground state*, given that this ground state is non-degenerate with a nonzero spectral gap to the remaining spectrum. It is natural to expect that one can find conditions such that the flipped gradient of the energy when expressed in new coordinates is locally strongly monotone.

The following is a central result, adapting a result due to Zarantonello [21,22] (points 1 and 2) to the present notation and setting, and applied to the flipped gradient of an energy functional (point 3).

**Theorem 2.1 (General convergence and error estimates):** *Let  $F : \mathcal{V} \oplus \mathcal{V} \rightarrow \mathcal{V}' \oplus \mathcal{V}'$  be a map, and let  $U \subset \mathcal{V} \oplus \mathcal{V}$  be an open ball containing a  $Z_*$  such that  $F(Z_*) = 0$ .*

*Let  $\mathcal{V}_d \subset \mathcal{V}$  be a Galerkin sequence of subspaces with  $P_d$  being the orthogonal projector onto  $\mathcal{V}_d \oplus \mathcal{V}_d$ . Furthermore, let  $F_d : \mathcal{V}_d \oplus \mathcal{V}_d \rightarrow \mathcal{V}'_d \oplus \mathcal{V}'_d$  be the Galerkin discretisation of  $F$ , i.e.  $F_d(Z_d) = P_d F(Z_d)$ .*

*Assume that  $F$  is locally strongly monotone with constant  $\eta > 0$  and Lipschitz continuous with constant  $L > 0$  on  $U$ . Then, the following holds:*

- (1)  $Z_*$  is the only root in  $U$ .
- (2) *There is a sufficiently large  $d_0$ , such that for any  $d > d_0$ , there exists  $Z_{d*} \in \mathcal{V}_d \oplus \mathcal{V}_d$  such that  $F_d(Z_{d*}) = 0$ . This root is unique in  $U$  and we have the following error estimate (quasi-optimality of the discrete solution):*

$$\|Z_{*d} - Z_*\| \leq \frac{L}{\eta} \text{dist}(Z_*, \mathcal{V}_d \oplus \mathcal{V}_d). \quad (8)$$

*Let  $\mathcal{E} : \mathcal{V} \oplus \mathcal{V} \rightarrow \mathbb{R}$ ,  $Z \mapsto \mathcal{E}(Z)$  be a (Fréchet) smooth energy functional. Let  $R$  be the flipping map as introduced after Equation (7) and set  $F = R\partial\mathcal{E}$ , and  $E_* = \mathcal{E}(Z_*)$ .*

- (3) *For  $d > d_0$ , the discrete Galerkin equations  $\partial\mathcal{E}_d(Z_{*d}) = 0$  have locally unique solutions, and in addition to the error estimate (8), we have the energy error*

$$\begin{aligned} |\mathcal{E}(Z_{*d}) - E_*| &\leq C \|Z_{*d} - Z_*\|^2 \\ &\leq C \left(\frac{L}{\eta}\right)^2 \text{dist}(Z_*, \mathcal{V}_d \oplus \mathcal{V}_d)^2. \end{aligned} \quad (9)$$

The proof is presented in [Appendix](#). The error estimate (9) shows that for (smooth) energy functionals with a locally strongly monotone flipped gradient, the bivariational method of discretisation behaves very similar to the usual Rayleigh–Ritz variational method of discretisation. As we enlarge the Galerkin space, the discrete ground state converges, and the energy error is quadratic in the error of the state. However, we cannot guarantee convergence *from above*, but this is much less important than actually having a quadratic error.

An interesting fact is that Brouwer's fixed point theorem [37] can be used to obtain a sufficient condition for the constant  $d_0$ , where quadratic convergence sets in, namely

$$\text{dist}(Z_*, \mathcal{V}_{d_0} \oplus \mathcal{V}_{d_0}) < \frac{\delta}{1 + L/\eta}, \quad (10)$$

see Refs. [23,26]. The radius  $\delta$  of the domain  $U$  in Theorem 2.1 is unknown in general, and we see that a small monotonicity constant  $\eta$  relative to  $L$  will be severely detrimental for the convergence, as it forces  $d_0$  to be very large. On the other hand, in general one has that  $\eta$  and the Lipschitz constant  $L$  are related by  $\eta \leq L$ . The optimal value of the right-hand side is therefore  $\delta/2$ .

The following theorem summarises the main results of Ref. [23], where the proof and more details can be found:

**Theorem 2.2 (NC-ECC monotonicity):** *Assume that the system Hamiltonian  $\hat{H}$  is self-adjoint, and that the ground state of  $\hat{H}$  exists, is non-degenerate, and that there is a spectral gap  $\gamma > 0$  between the ground-state energy  $E_*$  and the rest of the spectrum. Assume that the reference  $\phi_0$  is such that it is not orthogonal to the ground-state wavefunction. Let  $Z_* = (T_*, \Lambda_*) \in \mathcal{V} \oplus \mathcal{V}$  be the corresponding critical point of  $\mathcal{E}_{\text{ECC}}$ , and assume that  $T_*$  and  $\Lambda_*$  are not too large, i.e. that  $\phi_0$  is a sufficiently good approximation to  $\psi_*$ . Then,  $F = R\partial\mathcal{E}_{\text{ECC}}$  is locally strongly monotone near  $Z_*$  with a constant  $\eta = C\gamma$ , for some  $C < 1$ .*

The consequence of Theorem 2.2, when combined with Theorem 2.1, is that the non-canonical ECC method is convergent as the discrete cluster amplitude space  $\mathcal{V}_d$  approaches the untruncated limit. As already remarked, we do not explicitly know the onset  $d_0$  of quadratic convergence.

### 3. Monotonicity-preserving coordinate transformations

#### 3.1. A class of exact coupled-cluster models

In addition to the non-canonical ECC parameterisation, Arponen also considered a second parameterisation of the bra and ket wavefunctions, which gives equations of motion for the time-dependent Schrödinger equation that are canonical in the sense of Hamiltonian mechanics [1,2]. (This must not be confused with the use of canonical Hartree–Fock orbitals, which is unrelated.) This parameterisation is given in terms of a *coordinate transformation*  $\theta_{\text{C-ECC}} : \mathcal{V} \oplus \mathcal{V} \rightarrow \mathcal{V} \oplus \mathcal{V}$  as

$$(T, \Lambda) = \theta_{\text{C-ECC}}(T', \Lambda'), \quad (11)$$

where  $\Lambda' = \Lambda$ , and  $T = S(T'; \Lambda')$  is defined by

$$QT\phi_0 = Qe^{-\Lambda'^\dagger}T'\phi_0, \quad Q = I - |\phi_0\rangle\langle\phi_0|. \quad (12)$$

This function has inverse  $QT'\phi_0 = Qe^{\tilde{\Lambda}^\dagger}T\phi_0$ . (In Arponen's work [2], the notation  $(T', \Lambda') = (\Sigma, \tilde{\Sigma}^\dagger)$  is used.) The map  $\theta_{\text{C-ECC}}$  is smooth and invertible with a smooth inverse, and we therefore obtain a new exact energy functional

$$\mathcal{E}_{\text{C-ECC}} = \mathcal{E}_{\text{ECC}} \circ \theta_{\text{C-ECC}},$$

with values

$$\mathcal{E}_{\text{C-ECC}}(T', \Lambda') = \langle\phi_0, e^{(\Lambda')^\dagger}e^{-S(T'; \Lambda')}He^{S(T'; \Lambda')}\phi_0\rangle. \quad (13)$$

A remarkable consequence of this second parameterisation is that it corresponds to retaining only those terms

in Equation (1) that can be represented by ‘doubly linked’ diagrams [1,2],

$$\mathcal{E}_{\text{C-ECC}}(T', \Lambda') = \langle\phi_0, e^{(\Lambda')^\dagger}(He^{T'})_C\phi_0\rangle_{\text{DL}}. \quad (14)$$

The phrase ‘doubly linked’ means that every power of  $(\Lambda')^\dagger$  is connected to *two*  $T'$  operators on its right, unless it is connected directly to  $H$ . The subscript ‘C’ for ‘connected’ is the usual connectedness criterion on contractions between  $H$  and powers of  $T'$  [8]. Thus, the canonical coordinates represent a more compact representation in that the resulting tensor contractions or diagrams in the energy are *identical* to those obtained in the NC-ECC energy (1), except for some diagrams that are explicitly eliminated.

Similarly, for the standard CC method, Arponen introduced the coordinate transformation  $\theta_{\text{CC}}$  given by

$$(T, \Lambda) = \theta_{\text{CC}}(T', \Lambda') = (T', e^{\Lambda'} - 1). \quad (15)$$

We obtain the energy functional  $\mathcal{E}_{\text{CC}} = \mathcal{E}_{\text{ECC}} \circ \theta_{\text{CC}}$ , where

$$\mathcal{E}_{\text{CC}}(T', \Lambda') = \langle\phi_0, (1 + (\Lambda')^\dagger)e^{-T'}He^{T'}\phi_0\rangle, \quad (16)$$

that is, the standard CC Lagrangian [3]. Incidentally, the standard CC coordinates are also canonical.

The map  $\theta_{\text{CC}}$  can be generalised to Taylor polynomials. By setting

$$e^\Lambda = (e^{\Lambda'})_n \equiv 1 + \Lambda' + \frac{1}{2}(\Lambda')^2 + \dots + \frac{1}{n!}(\Lambda')^n,$$

we can solve for  $\Lambda$  in terms of  $\Lambda'$  by, e.g. considering first the singles, then doubles, etc., giving a smooth map  $G_n : \mathcal{V} \rightarrow \mathcal{V}$  such that  $e^{G_n(\Lambda')} = (e^{\Lambda'})_n$ . In fact, since the cluster operators are nilpotent,  $G_n(\Lambda') = \ln[(e^{\Lambda'})_n]$ , where the logarithm is expanded in a (finite) Taylor series around the identity. Similarly, we can solve for  $\Lambda'$  in terms of  $\Lambda$ , demonstrating that this map has an inverse, and in fact that this inverse is smooth. We obtain a coordinate transformation  $\theta_n$  given by

$$(T, \Lambda) = \theta_n(T', \Lambda') = (T', G_n(\Lambda')), \quad (17)$$

and the corresponding energy functional

$$\mathcal{E}_{\text{NC-ECC}(n)}(T', \Lambda') = \langle\phi_0, (e^{(\Lambda')^\dagger})_n e^{-T'}He^{T'}\phi_0\rangle. \quad (18a)$$

Coordinate transformations form a group, and may thus be composed. By combining  $\theta_{\text{C-ECC}(n)} = \theta_n \circ \theta_{\text{C-ECC}}$ , we obtain an energy functional

$$\begin{aligned} \mathcal{E}_{\text{C-ECC}(n)}(T', \Lambda') \\ = \langle\phi_0, (e^{(\Lambda')^\dagger})_n e^{-S(T'; \Lambda')}He^{S(T'; \Lambda')}\phi_0\rangle. \end{aligned} \quad (18b)$$

Since, in the NC-ECC energy functional (1), an exponential  $e^{-\Lambda'^\dagger}$  can be inserted after  $e^T$  without changing

the result, both of these hierarchies correspond to truncations of a Baker–Campbell–Hausdorff expansion at order  $n$ , and are thus manifestly extensive.

### 3.2. Coordinate transformation theorem

Equations (18a) and (18b) represent two hierarchies of *exact* parameterisations of the bivariate Rayleigh quotient. It is therefore of interest to determine whether they have locally strongly monotone flipped gradients. To establish this, we study the effect on local strong monotonicity of a coordinate transformation.

**Theorem 3.1 (Coordinate transformations):** *Let  $\mathcal{E} : \mathcal{V} \oplus \mathcal{V} \rightarrow \mathbb{R}$  be a smooth energy functional, let  $Z_*$  be a critical point, and assume that  $F = R\partial\mathcal{E}$  is locally strongly monotone near  $Z_*$  with constant  $\eta > 0$ . Let a smooth  $\theta : \mathcal{V} \oplus \mathcal{V} \rightarrow \mathcal{V} \oplus \mathcal{V}$  with a smooth inverse be a given coordinate transformation, and let  $\mathcal{E}_\theta = \mathcal{E} \circ \theta$  be the energy functional expressed in the new coordinates. Let  $W_* = \theta^{-1}(Z_*)$  be the corresponding critical point for  $\mathcal{E}_\theta$ , and let  $F_\theta = R\partial\mathcal{E}_\theta$  be its flipped gradient. Let  $M_* = \partial\theta(W_*)$  be the Jacobian at  $W_*$ . Then we have the following conclusions:*

- (1) *If  $M_*R = RM_*$ , then  $F_\theta$  is locally strongly monotone near  $W_*$  with constant  $\|M_*^{-1}\|^{-2}\eta$ .*
- (2) *In the noncommuting case, if  $m_* = M_* - I$  is sufficiently small,  $F_\theta$  is locally strongly monotone near  $W_*$  with constant*

$$\eta' = \eta\|M_*^{-1}\|^{-2} - C(I + \|m_*\|)\|m_*\|,$$

where  $C$  is the constant from Theorem 2.1(3).

Theorem 3.1 tells us that if we create a new method by changing coordinates in a sufficiently well-behaved manner, the monotonicity of the flipped gradient of the energy will be preserved. Hence, the new method will be convergent in the sense of Theorem 2.1, if the original method is. This gives us great flexibility in choosing new cluster-operator parameterisations, since we already know that non-canonical ECC is convergent. The interested reader can find the proof in the [Appendix](#).

### 3.3. Monotonicity of (N)C-ECC( $n$ ) models

We apply Theorem 3.1 to the maps  $\theta_n$  and  $\theta_n \circ \theta_{\text{C-ECC}}$  that define the NC-ECC( $n$ ) and C-ECC( $n$ ) models, respectively. The conclusion is as follows (with proof given in the [Appendix](#)):

**Corollary 3.2:** *For any of the NC-ECC( $n$ ) or C-ECC( $n$ ) models, the assumption that the ground-state critical point*

*$W_* = (T'_*, \Lambda'_*)$  is not too large together with a spectral gap  $\gamma > 0$  is sufficient to guarantee local strong monotonicity of the flipped gradient of the energy, and hence a quasi-optimal solution to the Galerkin problem and a quadratic error estimate for the energy.*

We note that our estimates are probably *pessimistic* for some of the models covered here. The analysis starts with a given monotonicity constant  $\eta$  for the NC-ECC scheme, depending on the spectral gap, and consistently produces an  $\eta' < \eta$  for the method obtained using the coordinate change, worsening the error estimates. It may well be that a direct analysis of the secondary method yields a better  $\eta'$ . The assumptions on the Hamiltonian and reference may also become milder. However, the important point here is that Theorem 3.1 does guarantee that the new method is convergent under *some* reasonable conditions. For example, we have now proven that QCC theory [18] is convergent if the reference is a sufficiently good approximation to the ground state, and using Equation (10) also a basic means to study which truncations or Galerkin schemes can be reasonable, at least in principle. It may be interesting to see whether truncation schemes like the PP hierarchy with orbital optimisation, also in a quadratic  $n = 2$  or higher formulation [19], can be further analysed based on our results.

In the proof of Theorem 3.1 (see [Appendix](#)), it arises naturally that the most favourable coordinate transformations are those that commute with the flipping map  $R$ , since local strong monotonicity then follows with no assumptions on the Jacobian of the map. The Jacobian commutes with  $R$  if and only if

$$\frac{\partial T}{\partial T'} = \frac{\partial \Lambda}{\partial \Lambda'}, \quad \text{and} \quad \frac{\partial T}{\partial \Lambda'} = \frac{\partial \Lambda}{\partial T'}, \quad (19)$$

and one must assume that this holds at every point in  $\mathcal{V} \oplus \mathcal{V}$ , as one does not know *a priori* where the critical point is. It is not clear what such transformations in general look like, and whether such transformations are useful reparameterisations of the energy.

### 3.4. Properties of the canonical and non-canonical schemes

The coordinate transformation  $\theta_{\text{C-ECC}}$  as represented by Equation (A7) is such that when applied to a cluster operator  $T'_k$  of rank  $k$ , it generates terms  $T_{k'}$  with  $k' \leq k$ . The same is true for the inverse map. Thus, if  $\mathcal{V}_d$  is excitation-rank complete, i.e. it contains all excitations of rank up to and including some  $k > 0$ , then the Galerkin discretisation of the NC-ECC( $n$ ) method commutes with changing



coordinates via the coordinate map  $\theta_{C-ECC}$ ,

$$\mathcal{E}_{C-ECC(n),d} = \mathcal{E}_{NC-ECC(n),d} \circ \theta_{C-ECC}.$$

By inspection, one can also see that this holds for a doubles-only truncation, since  $A(\Lambda') = I$  in this case. We obtain the following result:

**Theorem 3.3 (Equivalence of canonical and non-canonical coordinates):** *Let  $\mathcal{V}_d$  be excitation-rank complete or consist of doubles excitations only, and let  $n$  be given. Then, the discrete solutions  $(T_{*d}, \Lambda_{*d})$  and  $(T'_{*d}, \Lambda'_{*d})$  of the NC-ECC( $n$ ) and C-ECC( $n$ ) methods, respectively, are equivalent and related via  $\theta_{C-ECC}$ , i.e.  $T_{*d} = A(\Lambda'_{*d})T'_{*d}$  and  $\Lambda_{*d} = \Lambda'_{*d}$ .*

We stress that, if  $\mathcal{V}_d$  is *not* excitation-rank complete, the canonical and non-canonical parameterisations are not equivalent. This would be the case for the PP hierarchy of truncations [19,20].

According to the doubly linked structure of the energy functional  $\mathcal{E}_{C-ECC(n)}$ , see the discussion after Equation (14), the amplitude equations for the canonical case are cheaper, albeit by a small amount. Moreover, it is reasonable to expect that the canonical solution  $(T'_{*d}, \Lambda'_{*d})$  is more compact compared to  $(T_{*d}, \Lambda_{*d})$ . We investigate this claim numerically in Section 4.

## 4. Numerical results

### 4.1. Implementation

The (un-)truncated (N)C-ECC( $n$ )SD Equations (18) and (5), together with the untruncated coordinate transformation Equation (A7) have been implemented in a local full CI-based program, i.e. all intermediates are expressed as vectors in the full CI basis. To this end, the C-ECC( $n$ ) amplitudes are computed using transformed NC-ECC( $n$ ) residual expressions [17]. The C-ECC( $n$ ) amplitude equations are thus:

$$0 = \sum_{\nu \in \mathcal{I}} \langle \phi_\nu, e^{-(\Lambda')^\dagger} \phi_\mu \rangle \langle \phi_0, (e^{(\Lambda')^\dagger})_n [H^S, X_\nu] \phi_0 \rangle, \quad (20a)$$

$$0 = \langle \phi_\mu, (e^{(\Lambda')^\dagger})_{n-1} H^S \phi_0 \rangle - \sum_{\nu \in \mathcal{I}} \langle \phi_\mu, X_\nu^\dagger e^{-(\Lambda')^\dagger} T' \phi_0 \rangle \langle \phi_0, (e^{(\Lambda')^\dagger})_n [H^S, X_\nu] \phi_0 \rangle, \quad (20b)$$

where  $H^S = e^{-S(T';\Lambda')} H e^{S(T';\Lambda')}$ . The sparsity of the coordinate transformation Equation (A7) has been exploited throughout, e.g. only singles amplitudes are transformed in the case where  $\mathcal{V}_d$  contains only singles and doubles. The coupled amplitude equations (20) are solved

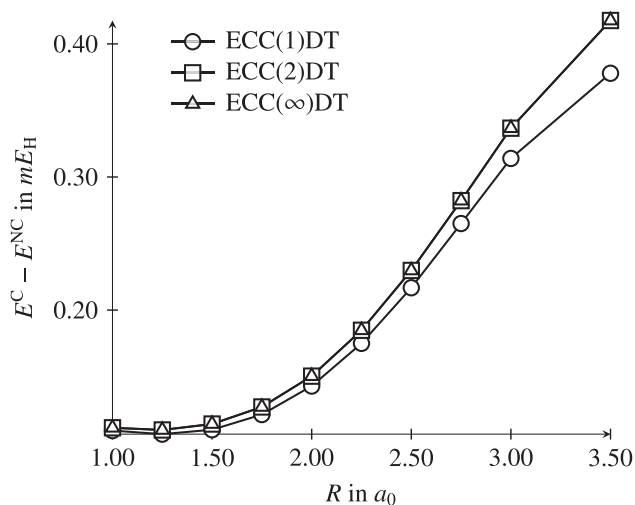
iteratively starting from an MP2-guess, using an alternating scheme and applying DIIS convergence acceleration. In all computations, residuals and energies were converged to a threshold of  $10^{-4}$  and  $10^{-6}$  a.u., respectively. The (N)C-ECC(1)SD and (N)C-ECC( $\infty$ )SD implementations are verified by reproducing the ‘CCSD’ and ‘ECCSD’ energies presented in [17].

### 4.2. Numerical experiments

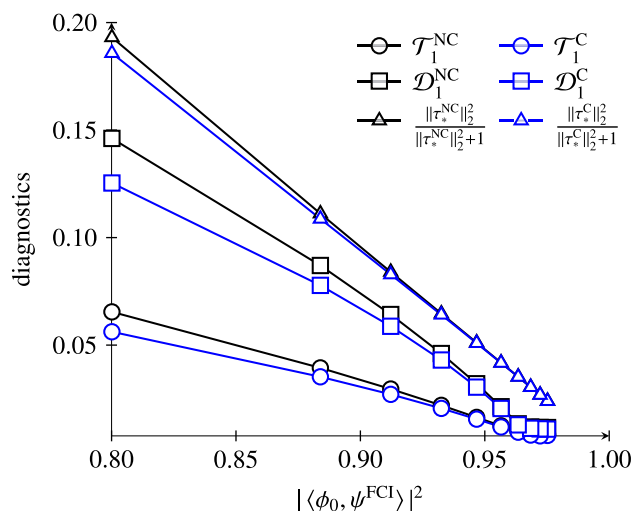
The (N)C-ECC( $n$ )SD and (N)C-ECC( $n$ )DT models have been studied numerically by investigating the potential energy curves of the hydrogen fluoride molecule with interatomic distances  $1.0 \leq R_{H-F} \leq 3.5$  ( $a_0$ ) in a DZV basis set [17] and the nitrogen molecule with  $1.5 \leq R_{N-N} \leq 4.0$  ( $a_0$ ) with a frozen core in a 6-31G basis [38]. Additionally, the  $H_8$  model system with structural parameters  $0.0001 \leq \alpha \leq 1.0$  in a minimal basis set [32,39] is considered. For large distances  $R$  and small  $\alpha$ , respectively, these systems comprise significant multireference character, i.e. the weight of the Hartree-Fock configuration in the full CI wave function,  $|\langle \phi_0, \psi^{FCI} \rangle|^2$ , is fairly small. Thus, these species are good candidates to study novel quantum chemical methods.

The energy curves of the canonical models C-ECC( $n$ )SD are identical to the non-canonical NC-ECC( $n$ )SD ones and are thus not presented here. However, the results differ if excitation-rank incomplete truncation schemes are employed. For instance, in canonical ECC( $n$ )DT, singles amplitudes are effectively generated from doubles and triples amplitudes, while these are absent in the non-canonical model. This effect has been studied on the potential curve of the HF molecule and is depicted in Figure 1: The generation of singles amplitudes entails that the canonical computation is lower in energy, in particular towards the multireference region where these contribute significantly to the wave function expansion. This depends, however, on the role of the singles amplitudes in the wave function: In test computations on the  $H_8$  model system a different trend was observed, consistent with the diminished importance of singles in the wave function (*vide infra*). Therefore, we cannot conclude that the canonical coordinates are consistently better when unconventional truncations are used.

In order to investigate the effect of using different coordinates in ECC( $n$ )SD computations, we calculated a set of CC diagnostics which are often used to assess the quality of CC computations [40,41]. These are based on the largest singular value ( $\mathcal{D}_1$ ) and Frobenius norm ( $\mathcal{T}_1$ ) of the matrix representation of the singles amplitudes. (Equivalently,  $\mathcal{T}_1^2 = \|\tau_{*1}\|_2^2/N$ , the sum of the squares of the singles amplitudes, with  $N$  the number of correlated electrons.) Although diagnostics based



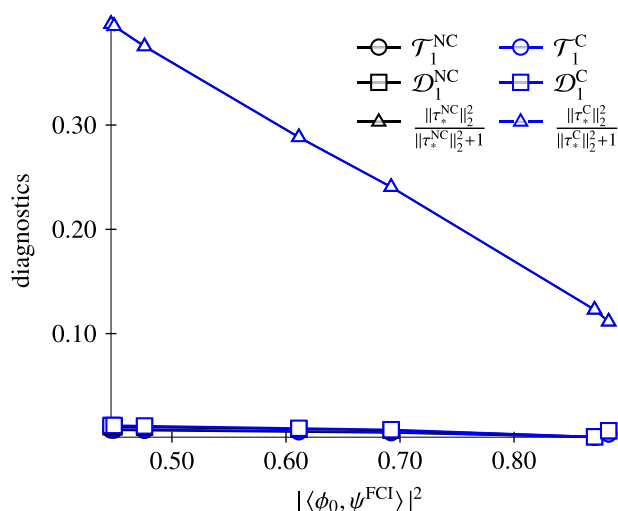
**Figure 1.** Difference between a canonical and a non-canonical ECC( $n$ )DT computation for the potential curve of HF.



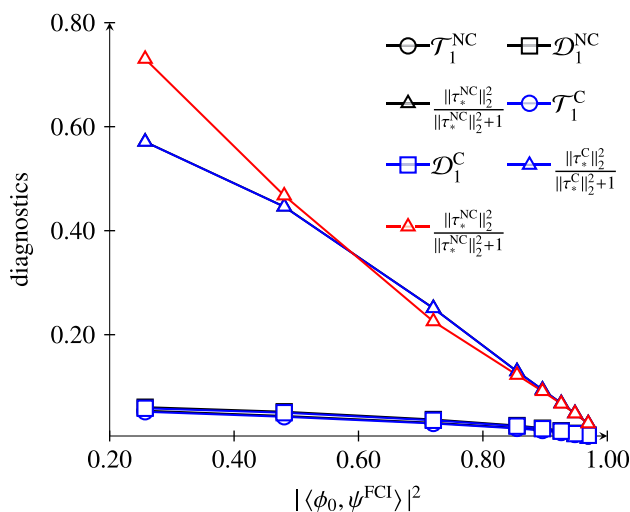
**Figure 2.** Comparison of CC diagnostics of the C-ECC(1)SD and NC-ECC(1)SD model for the HF potential curve correlated with the multireference character.

on doubles amplitudes are preferred, they are not as available in implementations as are the singles-based variants [42]. Additionally, we computed the diagnostic  $\|\tau_*\|_2^2 / (\|\tau_*\|_2^2 + 1)$  which involves all the amplitudes. This choice can be motivated from monotonicity arguments and will be discussed in a forthcoming paper.

The values have been computed for truncation schemes  $n = 1, 2, \infty$ . Since the trends are very similar, only the data for  $n = 1$  is presented for HF and  $H_8$ . For  $N_2$ , we focus on the results for  $n = 2$ , because the quadratic model is well-known to be a better approximation to full CI compared to standard CCSD in this case [16] (cf. line starting upper left with triangle markers in Figure 4). We stress, that even if the NC-ECC(1)SD and C-ECC(1)SD (which is standard CCSD) are equivalent,



**Figure 3.** Comparison of CC diagnostics of the C-ECC(1)SD and NC-ECC(1)SD model for the  $H_8$  potential curve correlated with the multireference character.



**Figure 4.** Comparison of CC diagnostics of the C-ECC(2)SD and NC-ECC(2)SD model (lower set of five curves, almost on top of each other) for the  $N_2$  potential curve correlated with the multireference character. The red curve shows diagnostics for NC-ECC(1)SD, indicating the inferiority of this model in the multireference region.

the singles amplitudes in their respective parameterisations differ, producing different values for the diagnostics. Figure 2 shows the diagnostics correlated with the multireference character for the HF potential curve, in Figures 3 and 4 values for the  $H_8$  model and the  $N_2$  molecule are shown, respectively. In both  $H_8$  and  $N_2$ , electron correlation is dominated by doubles amplitudes, as can be seen from the small values of the singles-based diagnostics. Since the reparameterisation of the amplitudes in the canonical model does not affect the amplitudes of highest excitation rank, the difference between

the NC-ECC( $n$ )SD and C-ECC( $n$ )SD amplitude vectors is negligible. This is different for the HF case. Here, the amplitude norms of the canonical models are consistently smaller than the non-canonical variants, indicating that the wave function parameterisation is more compact.

Our numerical experiments suggest, that for excitation-rank incomplete models, the canonical map generates effectively an excitation-rank complete parameterisation, but does not necessarily yield significantly better results. Concerning the *a priori* excitation-rank complete models, it has been found that the canonical parameterisation can be more compact compared to the non-canonical one, a desired property for post-Hartree–Fock methods. In particular, it may be useful to consider using canonical coordinates for *diagnostics*, even for well-established numerical codes using standard CC formalism.

## 5. Concluding remarks

In this article, we formulated basic error estimates for a class of exact models, defined in terms of replacing, in Arponen's ECC method, the exact exponential  $e^{\Lambda^\dagger}$  of the dual cluster operator with an  $n$ -th Taylor polynomial, the canonical C-NCC( $n$ ) models and the non-canonical NC-ECC( $n$ ) models. The central result was a coordinate-transformation theorem, Theorem 3.1, that gives error estimates for any method that can be described as a coordinate transformation of ECC theory. Notably, these results guarantee asymptotically quadratic error estimates for the ground-state energy of all models, under certain mild conditions.

Apart from Theorem 3.1, a basically self-contained mathematical framework for local error analysis of coupled-cluster methods was presented. This was based on Arponen's bivariational principle and basic results from nonlinear monotone operator theory, i.e. Zarattonello's theorem. Also central was our prior analysis of Arponen's extended coupled-cluster method in its non-canonical formulation.

The methods covered by our analysis include standard CC theory, quadratic CC theory [18,19], the perfect-pairing hierarchy [20] for approximating CASSCF, also in its quadratic version [19], and, Arponen's canonical ECC method. The computational cost of the (N)C-ECC( $n$ ) methods truncated with the standard singles, doubles, etc., scheme, are expensive: already the nonlinear terms in  $(\Lambda^\dagger)^2$  pushes the cost beyond standard CC theory when higher than doubles are considered. However, having obtained exact mathematical characterisations of the hierarchies of methods is an important first step in producing cheaper and more reliable methods compared to standard CC. One such approach could be similar to the

CC2 method of Christiansen and coworkers [43], where the cost of CCSD is reduced from  $\mathcal{O}(N^6)$  to  $\mathcal{O}(N^5)$ ,  $N$  being the system size.

The error estimates are not optimal for many methods. A direct analysis of canonical ECC would probably provide the most optimistic analysis for all the C-ECC( $n$ ) methods, due to the doubly linked structure and the equivalence of excitation-rank complete Galerkin discretisations.

Finally, we performed some simple numerical experiments, focusing on the possibility of using canonical coordinates in place of the usual CC amplitudes when doing diagnostic estimates on CC calculations on systems with multireference character. Our preliminary findings support the hypothesis that the canonical coordinates are more compact compared to the usual coordinates, providing more accurate diagnostics.

An interesting extension of the present work would be to study truncations where singles-amplitudes are replaced by orbital rotations, either unitary or biorthogonal, as in the QCC and PP approaches, or the non-orthogonal orbital optimised coupled-cluster method of Pedersen and coworkers. [44]. Moreover, the complete-active space coupled-cluster method by Adamowicz and coworkers [32] fits the present scheme. It is also known that quadratic CC and ECC in general are quite good at reproducing multireference character, while standard single-reference CC is quite poor at this. Thus, a modified analysis of the ECC method that includes multireference assumptions, such as the steerable CAS-ext gap of Ref. [34], could potentially lead to a deeper understanding of how CC methods generally behave in the presence of static correlation.

## Acknowledgments

The authors are thankful to M. A. Csirik for useful comments.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work has received funding from the Research Council of Norway (RCN) [CoE grant numbers 287906 and 262695] (Hylleraas Centre for Quantum Molecular Sciences), and from ERC-STG-2014 [grant number 639508].

## References

- [1] J.S. Arponen, *Ann. Phys.* **151** (2), 311–382 (1983). doi:10.1016/0003-4916(83)90284-1.
- [2] J.S. Arponen, R.F. Bishop and E. Pajanne, *Phys. Rev. A* **36**, 2519–2538 (1987). doi:10.1103/PhysRevA.36.2519.
- [3] T. Helgaker and P. Jørgensen, *Adv. Quant. Chem.* **19**, 183–245 (1988). doi:10.1016/S0065-3276(08)60616-4.

- [4] J. Hubbard, Proc. Roy. Soc. A: Math. Phys. Eng. Sci. **240** (1223), 539–560 (1957). doi:10.1098/rspa.1957.0106.
- [5] F. Coester, Nucl. Phys. **7**, 421–424 (1958). doi:10.1016/0029-5582(58)90280-3.
- [6] F. Coester and H. Kümmel, Nucl. Phys. **17**, 477–485 (1960). doi:10.1016/0029-5582(60)90140-1.
- [7] J. Paldus, in Theory and Applications of Computational Chemistry: The First Forty Years, edited by C.E. Dykstra, G. Frenking, K.S. Kim, and G.E. Scuseria (Elsevier, Edinburgh, London, Amsterdam, 2005), Chap. 7, p. 115.
- [8] R.J. Bartlett and M. Musiał, Rev. Mod. Phys. **79** (1), 291–352 (2007). doi:10.1103/RevModPhys.79.291.
- [9] G. Hagen, T. Papenbrock, M. Hjorth-Jensen and D.J. Dean, Rep. Prog. Phys. **77** (9), 096302 (2014). doi:10.1088/0034-4885/77/9/096302.
- [10] K. Emrich and J.G. Zabolitzky, in *Recent Progress in Many-Body Theories*, edited by H. Kümmel and M.L. Ristig (Springer, Berlin, Heidelberg, 1984), pp. 271–278.
- [11] B.H.J. McKellar, C.R. Leonard and L.C.L. Hollenberg, Int. J. Mod. Phys. B **14**, 2023–2037 (2000). doi:10.1142/S0217979200001151.
- [12] L.S. Cederbaum, O.E. Alon and A.I. Streltsov, Phys. Rev. A **73** (4) (2006). doi:10.1103/PhysRevA.73.043609.
- [13] J.S. Arponen, J. Phys. G: Nucl. Phys. **8**, L129 (1982). doi:10.1088/0305-4616/8/8/004.
- [14] J.S. Arponen, R.F. Bishop and E. Pajanne, Phys. Rev. A **36**, 2539–2549 (1987). doi:10.1103/PhysRevA.36.2539.
- [15] R.F. Bishop, Theor. Chim. Acta **80**, 95–148 (1991). doi:10.1007/BF01119617.
- [16] B. Cooper and P.J. Knowles, J. Chem. Phys. **133** (234102), 20120 (2010). doi:10.1063/1.3520564.
- [17] F.A. Evangelista, J. Chem. Phys. **134** (22), 224102 (2011). doi:10.1063/1.3598471.
- [18] T. Van Voorhis and M. Head-Gordon, Chem. Phys. Lett. **330**, 585–594 (2000). doi:10.1016/S0009-2614(00)01137-4.
- [19] E.F.C. Byrd, T. Van Voorhis and M. Head-Gordon, J. Phys. Chem. B **106** (33), 8070–8077 (2002). doi:10.1021/jp020255u.
- [20] S. Lehtola, J. Parkhill and M. Head-Gordon, J. Chem. Phys. **145**, 134110 (2016). doi:10.1063/1.4964317.
- [21] E. Zangrando, *Technical Report 160* (U.S. Army Math. Res. Centre, Madison, WI, 1960).
- [22] E. Zeidler, *Nonlinear Functional Analysis and Its Application II/B* (Springer, New York, Heidelberg, Berlin, 1990).
- [23] A. Laestadius and S. Kvaal, SIAM J. Numer. Anal. **56** (2), 660–683 (2018). doi:10.1137/17M1116611.
- [24] A. Laestadius and F.M. Faulstich, Mol. Phys. **117** (17), 2362–2373 (2019). doi:10.1080/00268976.2018.1564848.
- [25] R. Schneider, Numer. Math. **113**, 433–471 (2009). doi:10.1007/s00211-009-0237-3.
- [26] T. Rohwedder, ESAIM: Math. Mod. Num. Anal. **47**, 421–447 (2013). doi:10.1051/m2an/2012035.
- [27] T. Rohwedder and R. Schneider, ESAIM: Math. Mod. Num. Anal. **47**, 1553–1582 (2013). doi:10.1051/m2an/2013075.
- [28] R.P. Feynman, Phys. Rev. **56** (4), 340–343 (1939). doi:10.1103/PhysRev.56.340.
- [29] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I: Functional Analysis* (Academic Press, San Diego, New York, 1980).
- [30] K. Schmüdgen, *Unbounded Self-adjoint Operators on Hilbert Space*. Graduate Texts in Mathematics (Springer, Dordrecht, Heidelberg, 2012).
- [31] C. Hättig, W. Klopper, A. Köhn and D.P. Tew, Chem. Rev. **112**, 4–74 (2012). doi:10.1021/cr200168z.
- [32] L. Adamowicz, J.-P. Malrieu and V.V. Ivanov, J. Chem. Phys. **112** (23), 10075–10084 (2000). doi:10.1063/1.481649.
- [33] K. Kowalski, J. Chem. Phys. **148**, 094104 (2018). doi:10.1063/1.5010693.
- [34] F. Faulstich, A. Laestadius, S. Kvaal, Ö. Legeza and R. Schneider, SIAM J. Num. Anal. **57**, 2579 (2019). doi:10.1137/18M1171436.
- [35] T. Kinoshita, O. Hino and R.J. Bartlett, J. Chem. Phys. **123** (7), 074106 (2005). doi:10.1063/1.2000251.
- [36] F.M. Faulstich, M. Máté, A. Laestadius, M.A. Csirik, L. Veis, A. Antalik, J. Brabec, R. Schneider, J. Pittner, S. Kvaal and Ö. Legeza, J. Chem. Theor. Comp. **15** (4), 2206–2220 (2019). doi:10.1021/acs.jctc.8b00960.
- [37] E. Zeidler, *Nonlinear Functional Analysis and Its Application I* (Springer, New York Heidelberg Berlin, 1986).
- [38] A. Engels-Putzka and M. Hanrath, Mol. Phys. **107** (2), 143–155 (2009). doi:10.1080/00268970902724922.
- [39] K. Jankowski, L. Meissner and J. Wasilewski, Int. J. Quant. Chem. **28** (6), 931–942 (1985). doi:10.1002/qua.v28:6.
- [40] T.J. Lee and P.R. Taylor, Int. J. Quant. Chem. **36** (S23), 199–207 (1989). doi:10.1002/qua.v36:23+.
- [41] C.L. Janssen and I.M.B. Nielsen, Chem. Phys. Lett. **290** (4), 423–430 (1998). doi:10.1016/S0009-2614(98)00504-1.
- [42] W. Jiang, N.J. DeYonker and A.K. Wilson, J. Chem. Theor. Comp. **8** (2), 460–468 (2012). doi:10.1021/ct2006852 PMID: 26596596.
- [43] O. Christiansen, H. Koch and P. Jørgensen, Chem. Phys. Lett. **243**, 409–418 (1995). doi:10.1016/0009-2614(95)00841-Q.
- [44] T.B. Pedersen, B. Fernández and H. Koch, J. Chem. Phys. **114**, 6983 (2001). doi:10.1063/1.1358866.

## Appendix. Proofs of main results

In this appendix all our proofs have been collected and are given in the same order as the corresponding results have appeared above. We will only present a partial proof of Theorem 2.1, as the proofs of points (1) and (2) are standard, and can be found in, e.g. Ref. [22].

**Proof of Theorem 2.1:** For point (3), we note that  $F$  is locally Lipschitz continuous as a consequence of  $\mathcal{E}$  being smooth, which together with strong monotonicity makes points (1) and (2) applicable. The remaining argument follows [23] closely (where the case  $\mathcal{E} = \mathcal{E}_{\text{ECC}}$  was treated). First, by assumption of  $R$ ,  $F(Z_*) = 0$  and  $F_d(Z_{*d}) = 0$  are equivalent to  $\partial\mathcal{E}(Z_*) = 0$  and  $\partial\mathcal{E}_d(Z_{*d}) = 0$ , respectively (note that  $R$  commutes with  $P_d$ ). Now, Taylor expanding  $\mathcal{E}$  around  $Z_*$  and evaluating at  $Z_{*d}$  gives

$$\mathcal{E}(Z_{*d}) - E_* = \frac{1}{2} \langle Z, \partial^2 \mathcal{E}(Z_*) Z \rangle + \mathcal{O}(\|Z\|^3).$$

By the smoothness of  $\mathcal{E}$ , there exists a constant  $C'$  such that

$$\langle Z, \partial^2 \mathcal{E}(Z_*) Z \rangle \leq C' \|Z\|^2.$$

Further, the fact that on  $U$  we can control the higher-order terms by the quadratic one, we have

$$|\mathcal{E}(Z_{*d}) - E_*| \leq C \|Z_{*d} - Z_*\|^2.$$

Using Equation (8) gives the full statement in Equation (9)  $\blacksquare$

**Proof of Theorem 3.1:** With  $F : \mathcal{V} \oplus \mathcal{V} \rightarrow \mathcal{V}' \oplus \mathcal{V}'$  the flipped gradient and  $X \in \mathcal{V} \oplus \mathcal{V}$ , we obtain

$$\langle X, F(Z) \rangle_{\mathcal{V} \oplus \mathcal{V}, \mathcal{V}' \oplus \mathcal{V}'} = \langle RX, \partial \mathcal{E}(Z) \rangle_{\mathcal{V} \oplus \mathcal{V}, \mathcal{V}' \oplus \mathcal{V}' }.$$

In the sequel, we omit the specification of the spaces in the dual pairing.

Let  $Z_* \in \mathcal{V} \oplus \mathcal{V}$  be such that  $\partial \mathcal{E}(Z_*) = 0$ , i.e.  $F(Z_*) = 0$ . Since  $F$  is smooth, local strong monotonicity of  $F$  is equivalent to  $\partial F(Z_*) \in \mathcal{B}(\mathcal{V} \oplus \mathcal{V}, \mathcal{V}' \oplus \mathcal{V}')$  (the set of bounded linear operators) being coercive, i.e. there exists an  $\eta_* > 0$  such that

$$\Delta(X) = \langle X, \partial F(Z_*) X \rangle \geq \eta_* \|X\|_{\mathcal{V} \oplus \mathcal{V}}^2,$$

where  $\Delta(X)$  is defined by the first equality. (The constant  $\eta$  in Equation (6) approaches  $\eta_*$  as the ball  $U$  given in Theorem 2.1 of the local strong monotonicity approaches a point.) To see this, we find an expression for  $\Delta(h)$  in terms of the energy map,

$$\langle X, F(Z_* + \epsilon X) \rangle = \langle X, F(Z_*) \rangle + \epsilon F'(Z_*; X) + \mathcal{O}(\epsilon^2). \quad (\text{A1})$$

Here  $F'(Z_*; X)$  is the directional derivative in the direction of  $X$  such that

$$\begin{aligned} \langle X, \partial F(Z_*) X \rangle &= \langle X, F'(Z_*; X) \rangle \\ &= \frac{d}{d\epsilon} \langle RX, \partial \mathcal{E}(Z_* + \epsilon X) \rangle |_{\epsilon=0} \\ &= \langle RX, \partial^2 \mathcal{E}(Z_*) X \rangle, \end{aligned}$$

where  $\partial^2 \mathcal{E}(Z_*) \in \mathcal{B}(\mathcal{V} \oplus \mathcal{V}, \mathcal{V}' \oplus \mathcal{V}')$ . By choosing  $\epsilon$  small enough, Equation (A1) and strong monotonicity gives the coercivity claim. The logical implication also goes in the reverse direction. (This will be used below.)

Recall that  $\mathcal{E}_\theta = \mathcal{E} \circ \theta$  and that  $F_\theta$  denotes the flipped gradient of  $\mathcal{E}_\theta$ . We use that  $F_\theta$  is locally strongly monotone at  $W_* = \theta^{-1}(Z_*)$  if and only if  $\Delta_\theta$  is coercive, i.e.

$$\Delta_\theta(X) = \langle X, \partial F_\theta(W_*) X \rangle \geq \eta_\theta \|X\|^2,$$

for some  $\eta_\theta > 0$  and all  $X \in \mathcal{V} \oplus \mathcal{V}$ . A straightforward application of the chain rule now gives

$$\begin{aligned} \Delta_\theta(X) &= \langle M_* R X, \partial^2 \mathcal{E}(Z_*) (M_* X) \rangle, \\ M_* &= \partial \theta(W_*) \in \mathcal{B}(\mathcal{V} \oplus \mathcal{V}, \mathcal{V} \oplus \mathcal{V}). \end{aligned}$$

We note that this is *almost*  $\Delta(M_* X)$ . Indeed,

$$\begin{aligned} \Delta_\theta(X) &= \langle R M_* X, \partial^2 \mathcal{E}(Z_*) (M_* X) \rangle \\ &\quad + \langle [M_*, R] X, \partial^2 \mathcal{E}(Z_*) (M_* X) \rangle \\ &= \Delta(M_* X) + \langle (M_* R - R M_*) X, \partial^2 \mathcal{E}(Z_*) (M_* X) \rangle. \end{aligned} \quad (\text{A2})$$

In particular, if  $M_* R = R M_*$  then the last term vanishes in the utmost right-hand side of Equation (A2), and we obtain monotonicity of  $F_\theta$  but with a modified constant.

In the case where  $M_* R \neq R M_*$ , we write  $M_* = I + m_*$ , and note that  $M_* R - R M_* = m_* R - R m_*$ . We obtain,

$$\begin{aligned} \Delta_\theta(X) &\geq \eta \|M_* X\|^2 - \|\partial^2 \mathcal{E}(Z_*)\| \|M_*\| \|m_*\| \|X\|^2 \\ &\geq \left[ \eta \|M_*^{-1}\|^{-2} - C \|(1 + \|m_*\|)\|m_*\| \right] \|X\|^2. \end{aligned} \quad (\text{A3})$$

Here, we used that  $\theta$  has a smooth inverse, implying  $\|M_* X\| \geq \|M_*^{-1}\|^{-1} \|X\|$ , and that  $\|M_*\| \leq I + \|m_*\|$ .  $\blacksquare$

**Proof of Corollary 3.2:** We consider the Jacobian of the coordinate map, which on block form reads

$$\partial \theta(T', \Lambda') = \begin{pmatrix} \frac{\partial T}{\partial T'} & \frac{\partial T}{\partial \Lambda'} \\ \frac{\partial \Lambda}{\partial T'} & \frac{\partial \Lambda}{\partial \Lambda'} \end{pmatrix}. \quad (\text{A4})$$

For the map  $\theta_n$  (see Equation (17)), we first observe that by definition,

$$e^\Lambda = e^{\Lambda'} + \mathcal{O}(\|\Lambda'\|^{n+1}),$$

from which it follows, by taking the logarithm and expanding the logarithm around  $e^{\Lambda'}$ , which is a finite Taylor series,

$$\Lambda = \Lambda' + \mathcal{O}(\|\Lambda'\|^{n+1}).$$

We obtain

$$\partial \theta_n(T', \Lambda') = \begin{pmatrix} I & 0 \\ 0 & I + \mathcal{O}(\|\Lambda'\|^{n+1}) \end{pmatrix}. \quad (\text{A5})$$

For the map  $\theta_n \circ \theta_{\text{C-ECC}}$  we have, using the chain rule,

$$\partial \theta_{\text{C-ECC}}(T', \Lambda') = \begin{pmatrix} A(\Lambda') & \partial_{\Lambda'} A(\Lambda') T' \\ 0 & I + \mathcal{O}(\|\Lambda'\|^{n+1}) \end{pmatrix}. \quad (\text{A6})$$

Here,  $A(\Lambda')$  is the linear transformation on  $\mathcal{V}$  such that  $S(T'; \Lambda') = A(\Lambda') T'$  (see Equation (12)), i.e.  $A(\Lambda')$  can be expressed in terms of the matrix representation of  $e^{-(\Lambda')^\dagger}$ ,

$$A(\Lambda') T' = \sum_{\mu, \nu \in \mathcal{I}} X_\mu \langle \phi_\mu, e^{-(\Lambda')^\dagger} \phi_\nu \rangle \langle \phi_\nu, T' \phi_0 \rangle. \quad (\text{A7})$$

We have  $A(\Lambda') = I + \mathcal{O}(\|\Lambda'\|)$ , and  $\partial_{\Lambda'} A(\Lambda') T' = \mathcal{O}(\|T'\|)$ . For both maps, the Jacobian of the coordinate transformation at the critical point  $W_* = (T'_*, \Lambda'_*)$  becomes  $M_* = I + m_*$  with  $m_* = \mathcal{O}(\|W_*\|)$ . Applying Theorem 3.1(2), the local strong monotonicity follows, and by Theorem 2.1, quasi-optimality of the truncated solutions and a quadratic error estimate.  $\blacksquare$