Convergence analysis of adaptive DIIS algorithms with application to electronic ground state calculations

Maxime Chupin, Mi-Song Dupuy, Guillaume Legendre, Eric Séré

Convergence analysis of adaptive DIIS algorithms with application to electronic ground state calculations

Maxime Chupin*, Mi-Song Dupuy†, Guillaume Legendre‡and Éric Séré§

March 13, 2020

Abstract

This paper deals with a general class of algorithms for the solution of fixed-point problems that we refer to as *Anderson–Pulay acceleration*. This family includes the DIIS technique and its variant sometimes called commutator-DIIS, both introduced by Pulay in the 1980s to accelerate the convergence of self-consistent field procedures in quantum chemistry, as well as the related Anderson acceleration, which dates back to the 1960s, and the wealth of methods it inspired. Such methods aim at accelerating the convergence of any fixed-point iteration method by combining several previous iterates in order to generate the next one at each step. The size of the set of stored iterates is characterised by its *depth*, which is a crucial parameter for the efficiency of the process. It is generally fixed to an empirical value in most applications.

In the present work, we consider two parameter-driven mechanisms to let the depth vary along the iterations. One way to do so is to let the set grow until the stored iterates (save for the last one) are discarded and the method "restarts". Another way is to "adapt" the depth by eliminating some of the older, less relevant, iterates at each step. In an abstract and general setting, we prove under natural assumptions the local convergence and acceleration of these two types of Anderson–Pulay acceleration methods and demonstrate how to theoretically achieve a superlinear convergence rate. We then investigate their behaviour in calculations with the Hartree–Fock method and the Kohn–Sham model of density functional theory. These numerical experiments show that the restarted and adaptive-depth variants exhibit a faster convergence than that of a standard fixed-depth scheme, and require on average less computational effort per iteration. This study is complemented by a review of known facts on the DIIS, in particular its link with the Anderson acceleration and some multisecant-type quasi-Newton methods.

1 Introduction

The DIIS (for *Direct Inversion in the Iterative Subspace*) technique, introduced by Pulay [48] and also known as the *Pulay mixing*, is a locally convergent method widely used in computational quantum chemistry for accelerating the numerical solution of self-consistent equations intervening in the context of the Hartree–Fock method and of the Kohn–Sham model of density functional theory, among others. As a complement to available globally convergent methods, like the *optimal damping algorithm* (ODA) [10] or its energy-DIIS (EDIIS) variant [38], it remains a method of choice, in a large part due to its simplicity and nearly unparalleled performance once a convergence region has been attained. Due to this success, variants of the technique have been proposed over the years in other types of application, like the GDIIS adaptation [14, 17] for geometry optimization or the *residual minimisation method–direct inversion in the iterative subspace* (RMM-DIIS) for the simultaneous computation of eigenvalues and corresponding eigenvectors, attributed to Bendt and Zunger and described in [36]. It has also been combined with other schemes to improve the rate of convergence of various types of iterative calculations (see [34] for instance).

From a general point of view, the DIIS can be seen as an acceleration technique based on extrapolation, applicable to any fixed-point iteration method for which a measure of the error at each step, in the form of a residual for instance, is (numerically) available. It was recently established that the method is a particular avatar of an older process known as the *Anderson acceleration* [3]. It was also shown that it amounts to a multisecant-type variant of a Broyden method [8] and that, when applied to solve linear problems, it is (essentially) equivalent to the *generalised minimal residual*

*CEREMADE, UMR CNRS 7534, Université Paris-Dauphine, Université PSL, Place du Maréchal De Lattre De Tassigny, 75775 Paris cedex 16, France (chupin@ceremade.dauphine.fr).

†Zentrum Mathematik, Technische Universität München, Boltzmannstraße 3, 85747 Garching, Germany (dupuy@ma.tum.de)

‡CEREMADE, UMR CNRS 7534, Université Paris-Dauphine, Université PSL, Place du Maréchal De Lattre De Tassigny, 75775 Paris cedex 16, France (guillaume.legendre@ceremade.dauphine.fr).

§CEREMADE, UMR CNRS 7534, Université Paris-Dauphine, Université PSL, Place du Maréchal De Lattre De Tassigny, 75775 Paris cedex 16, France (sere@ceremade.dauphine.fr).

(GMRES) method of Saad and Schultz [52]. On this basis, Rohwedder and Schneider [50] (for the DIIS), and later Toth and Kelley [59] (for the Anderson acceleration), analysed the method in an abstract framework and provided linear convergence results.

In the present paper, we consider a unified family of methods, which we refer to as *Anderson–Pulay acceleration*, encompassing the DIIS, its commutator-DIIS (CDIIS) variant [49], and the Anderson acceleration. In such methods, one keeps a "history" of previous iterates which are combined to generate the next one at each step with the aim of accelerating the convergence of the sequence of iterates. The size of the set of stored iterates is characterised by an integer, sometimes called the *depth* (see [59, 1, 18]), and is an important parameter for the efficiency of the process. In many applications, the depth is empirically fixed to a maximal value m. One of the main conclusions of our work is that it can be beneficial to let the depth vary along the iterations.

In order to prove this point, we propose and investigate two Anderson–Pulay acceleration methods, each of them including a different parameter-driven procedure to determine the depth at each step. The first one allows the method to "restart", based on a condition initially introduced by Gay and Schnabel for a quasi-Newton method using multiple secant equations [25] and used by Rohwedder and Schneider in the context of the DIIS [50]. In the second one, the depth is adapted using a criterion which is, as far as we know, new.

In a general framework, we mathematically analyse these methods and prove local convergence and acceleration properties. These results are obtained *without* any hypothesis (which would have to be verified *a posteriori* in practice) on the boundedness of the extrapolation coefficients, in contrast with preceding works [50, 59]. Indeed, the built-in mechanism in each of the proposed algorithms prevents linear dependency from occurring in the least-squares problem for the coefficients, allowing us to derive a theoretical *a priori* bound on these coefficients. Applications of both methods to the numerical calculation of the electronic ground state of molecules by self-consistent field methods and comparisons with their fixed-depth counterparts demonstrate their good performances, and even suggest that the adaptive depth approach gives the fastest convergence in practice.

The paper is organized as follows. The principle of the Anderson–Pulay acceleration is presented in Section 2 through an overview of the DIIS and its relation with the Anderson acceleration and a class of quasi-Newton methods based on multisecant updating. We also recall convergence results for this class of methods, in both linear and nonlinear cases, existing in the literature. A generic abstract problem is next set in Section 3 and two versions of the Anderson–Pulay acceleration are proposed to solve it, one allowing "restarts" in the course of its application, the other having an adaptive depth. For both variants, a local linear convergence result is proved, as well as an acceleration property and the possibility of obtaining a superlinear rate of convergence. In Section 4, the quantum chemistry problems we consider for the application of the methods are recalled and their mathematical properties are discussed in connection with the assumptions used in the convergence results. Finally, numerical experiments, performed on molecules in order to illustrate the convergence behaviour of the CDIIS form of both methods and to compare them to their "classical" fixed-depth counterparts, are reported in Section 5.

2 Overview of the DIIS

The class of methods named Anderson–Pulay acceleration in the present paper comprises several methods, introduced in various applied contexts but with the same goal of accelerating the convergence of fixed-point iterations by means of extrapolation, the most famous of these in the quantum chemistry community probably being the DIIS technique, introduced by Pulay in 1980 [48]. We first describe its basic principle by using the DIIS applied to the solution of an abstract problem and exploring its relation to similar extrapolation methods and to a family of multisecant-type quasi-Newton methods. We conclude this section by recalling a number of previously established results pertaining to the DIIS and the Anderson acceleration.

2.1 Presentation

Consider the numerical computation of the solution x_* in $\mathbb{R}^n$ of a given problem by a fixed-point iteration method,

$$\forall k \in \mathbb{N},\ x^{(k+1)} = g(x^{(k)}), \tag{1}$$

with g a function from $\mathbb{R}^n$ to $\mathbb{R}^n$, for which an *"error"* vector $r^{(k)}$ in $\mathbb{R}^p$, corresponding to the variable value $x^{(k)}$, can be computed at each step. Given a non-negative integer m, the DIIS paradigm assumes that a good approximation to the solution x_* can be obtained at step $k+1$ by forming a combination involving $m_k + 1$ previous guess values, with $m_k = \min\{m, k\}$, that is

$$\forall k \in \mathbb{N},\ x^{(k+1)} = \sum_{i=0}^{m_k} c_i^{(k)}\, g(x^{(k-m_k+i)}), \tag{2}$$

while requiring that the coefficients $c_i^{(k)}$, $i = 0, \ldots, m_k$, of the combination are such that the associated linearised error vector, given by $\sum_{i=0}^{m_k} c_i^{(k)}\, r^{(k-m_k+i)}$, approximates zero in the least-squares sense under the constraint[1] that

$$\forall k \in \mathbb{N},\ \sum_{i=0}^{m_k} c_i^{(k)} = 1. \tag{3}$$

As discussed by Pulay, the minimisation subject to constraint (3) may be achieved through a Lagrange multiplier technique applied to the normal equations associated to the problem. More precisely, introducing an undetermined scalar λ and defining the Lagrangian function

$$\frac{1}{2}\left\| \sum_{i=0}^{m_k} c_i^{(k)}\, r^{(k-m_k+i)} \right\|_2^2 - \lambda \left(\sum_{i=0}^{m_k} c_i^{(k)} - 1 \right) = \frac{1}{2} \sum_{i=0}^{m_k} \sum_{j=0}^{m_k} c_i^{(k)} c_j^{(k)} b_{ij}^{(k)} - \lambda \left(\sum_{i=0}^{m_k} c_i^{(k)} - 1 \right),$$

in which the coefficients $b_{ij}^{(k)}$, $i, j = 0, \ldots, m_k$ are those of the Gramian matrix associated with the set of error vectors stored at the end of step k, *i.e.* $b_{ij}^{(k)} = \left\langle r^{(k-m_k+j)}, r^{(k-m_k+i)} \right\rangle_2$, solving the Euler-Lagrange equations with respect to the coefficients $c_i^{(k)}$, $i = 0, \ldots, m_k$ and Lagrange multiplier λ leads to the following system of $m_k + 2$ linear equations

$$\begin{pmatrix} b_{00}^{(k)} & \ldots & b_{0m_k}^{(k)} & -1 \\ \vdots & & \vdots & \vdots \\ b_{m_k 0}^{(k)} & \ldots & b_{m_k m_k}^{(k)} & -1 \\ -1 & \ldots & -1 & 0 \end{pmatrix} \begin{pmatrix} c_0^{(k)} \\ \vdots \\ c_{m_k}^{(k)} \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix}. \tag{4}$$

The DIIS iterations are generally ended once an acceptable precision has been reached, for instance when the value of the norm of the current error vector lies below a prescribed tolerance.

In practice and for a large number of applications, the given integer m is small, of the order of a few units. Since system (4) is related to normal equations, it may nevertheless happen that it is ill-conditioned if some error vectors are (almost) linearly dependent. In such a case, one usually drops the oldest stored vectors, as they are most likely to be the less relevant for the rest of the computation, one by one until the condition number of the resulting system becomes acceptable. One should note that the use of an unconstrained equivalent formulation of the least-squares problem, like those derived in the next subsection, is usually advocated, as it is generally observed that it results in a better condition number for the matrix of the underlying linear system. We refer the reader to [54, 60] for more details on this topic.

For the choice of error vectors, Pulay initially suggested using the quantities

$$\forall k \in \mathbb{N},\ r^{(k)} = x^{(k+1)} - x^{(k)}, \tag{5}$$

when searching for stationary point of energy functionals by methods of the quasi-Newton type [48]. He later proposed a variant form of the procedure [49], that is sometimes known as the commutator-DIIS or simply CDIIS (see [24] for instance), for which he claimed that the choice of an erro vector involving the commutator between the Fock and density matrices is better suited to the numerical solution by *ab initio* self-consistent field (SCF) techniques of the Hartree–Fock method or the Kohn–Sham model of density functional theory (see Subsection 4.5 for a presentation).

2.2 Relation with the Anderson and nonlinear Krylov accelerations

When looking for the solution a nonlinear equation of the form

$$f(x) = 0 \tag{6}$$

by a fixed-point iteration method based on the function

$$g(x) = x + \beta\, f(x), \tag{7}$$

[1] The rationale behind this choice can be seen as follows. Writing the approximation at each step as a sum of the solution x_* and an error term, that is, $\forall k \in \mathbb{N}$, $x^{(k)} = x_* + e^{(k)}$, the DIIS approximation is given by

$$\forall k \in \mathbb{N},\ x^{(k+1)} = \sum_{i=0}^{m_k} c_i^{(k)} \left(x_* + e^{(k-m_k+i)} \right) = \left(\sum_{i=0}^{m_k} c_i^{(k)} \right) x_* + \sum_{i=0}^{m_k} c_i^{(k)}\, e^{(k-m_k+i)}.$$

Total minimisation of the error would then lead to $x^{(k+1)} = x_*$ by making the second term in the right-hand side vanish, thus imposing condition (3). Of course, the actual error at each step is not known in practice, and it is thus replaced by a quantity whose choice depends on the problem to be solved.

with a sufficiently regular, homogeneous operator β, which can be seen as a preconditioner of some sort (the simplest case being the multiplication by a (nonzero) constant – a relaxation (or damping) parameter – or, in the context of the fixed-point iteration (1), a sequence $(\beta^{(k)})_{k\in\mathbb{N}}$ of such constants[2]), one sees that the choice (5) for the error vectors corresponds to setting

$$\forall k \in \mathbb{N},\ r^{(k)} = g(x^{(k)}) - x^{(k)} = \beta\, f(x^{(k)}),$$

that is, using *residuals*. In such a case, it is possible to relate the DIIS to a structurally similar extrapolation method, the Anderson acceleration [3] (sometimes called the *Anderson mixing*), introduced for the numerical solution of discretised nonlinear integral equations and originally formulated as follows. At the $k+1$th step, given the m_k+1 most recent iterates $x^{(k-m_k)}, \ldots, x^{(k)}$ and the corresponding residuals $r^{(k-m_k)}, \ldots, r^{(k)}$, define

$$\forall k \in \mathbb{N},\ x^{(k+1)} = g(x^{(k)}) + \sum_{i=1}^{m_k} \theta_i^{(k)} \left(g(x^{(k-m_k-1+i)}) - g(x^{(k)}) \right), \tag{8}$$

the scalars $\theta_i^{(k)}$, $i = 1, \ldots, m_k$ being chosen so as to minimise the norm of the associated linearised residual, *i.e.*

$$\forall k \in \mathbb{N},\ \boldsymbol{\theta}^{(k)} = \underset{(\theta_1,\ldots,\theta_{m_k})\in\mathbb{R}^{m_k}}{\arg\min} \left\| r^{(k)} + \sum_{i=1}^{m_k} \theta_i \left(r^{(k-m_k-1+i)} - r^{(k)} \right) \right\|_2. \tag{9}$$

Setting

$$\forall k \in \mathbb{N},\ c_i^{(k)} = \theta_{i+1}^{(k)},\ i = 0, \ldots, m_k - 1,\ c_{m_k}^{(k)} = 1 - \sum_{j=1}^{m_k} \theta_j^{(k)},$$

one observes that relation (8) can be put into the form of relation (2), so that the DIIS applied to the solution of (6) by fixed-point iteration is actually a reformulation of the Anderson acceleration.

This instance of the DIIS is not the only reinvention of the Anderson acceleration. In 1997, Washio and Oosterlee [62] introduced a closely related process dubbed *Krylov subspace acceleration* for the solution of nonlinear partial differential equation problems. The extrapolation within it relies on a nonlinear extension of the GMRES method (see subsection 2.4) and is in all points identical to (8) and (9), save for the definition of the fixed-point function g which corresponds in their setting to the application of a cycle of a nonlinear multigrid scheme. Another method, also based on a nonlinear generalisation of the GMRES method, is the so-called *nonlinear Krylov acceleration* (NKA), introduced by Carlson and Miller around 1990 (see [12]) and used for accelerating modified Newton iterations in a finite element solver with moving mesh. It leads to yet another equivalent formulation of the Anderson acceleration procedure by setting

$$\forall k \in \mathbb{N},\ x^{(k+1)} = g(x^{(k)}) - \sum_{i=1}^{m_k} \alpha_i^{(k)} \left(g(x^{(k-m_k+i)}) - g(x^{(k-m_k+i-1)}) \right), \tag{10}$$

with

$$\forall k \in \mathbb{N},\ \boldsymbol{\alpha}^{(k)} = \underset{(\alpha_1,\ldots,\alpha_{m_k})\in\mathbb{R}^{m_k}}{\arg\min} \left\| r^{(k)} - \sum_{i=1}^{m_k} \alpha_i \left(r^{(k-m_k+i)} - r^{(k-m_k+i-1)} \right) \right\|_2, \tag{11}$$

which amounts to take $\alpha_i^{(k)} = \sum_{j=1}^{i} \theta_j^{(k)}$, $i = 1, \ldots, m_k$. Another form of the minimisation problem for the linearised residual is found in [15], where it is combined with a line-search to yield a so-called *nonlinear GMRES* (N-GMRES) optimisation algorithm for computing the sum of R rank-one tensors that has minimal distance to a given tensor in the Frobenius norm.

The fact that the above acceleration schemes are different forms of the same method has been recognised on several occasions in the literature, see [21, 60, 9] for instance or Anderson's own account in [4]. In [7], it is shown that the Anderson acceleration is actually part of a general framework for the so-called Shanks transformations, used for accelerating the convergence of sequences.

Like the DIIS, the Anderson acceleration has been used to improve the convergence of numerical schemes in a number of contexts, and the recent years have seen a wealth of publications dealing with a large range of applications (see [44, 23, 63, 31, 1, 45, 65, 30] for instance). Variants with restart conditions to prevent stagnation [20], periodic restarts [47] or a periodic use of the extrapolation [5] have also been proposed.

[2] In [3], it is advocated that $\beta^{(k)} > 0$ for any integer k, the choice of the constant value $\beta^{(k)} = 1$ being usually the most appropriate in the numerical experiments. On the contrary, the vanishing choice $\beta^{(k)} = 0$ is excluded to prevent stagnation in the acceleration process defined by (8) and (9).

2.3 Interpretation as a multisecant-type quasi-Newton method

An insight on the behaviour of the DIIS may be gained by seeing it as a quasi-Newton method using multiple secant equations. It was indeed observed by Eyert [19] that the Anderson acceleration procedure amounts to a modification of Broyden's *second*[3] method [8], in which a given number of secant equations are satisfied at each step. This relation was further clarified by Fang and Saad [20] (see also the paper by Walker and Ni [60]) as follows.

Keeping with the previously introduced notations[4] and considering vectors as column matrices, we introduce the matrices respectively containing the differences of successive iterates and the differences of associated successive residuals stored at step k,

$$\mathscr{Y}^{(k)} = \left[x^{(k-m_k+1)} - x^{(k-m_k)}, \dots, x^{(k)} - x^{(k-1)}\right] \text{ and } \mathscr{S}^{(k)} = \left[r^{(k-m_k+1)} - r^{(k-m_k)}, \dots, r^{(k)} - r^{(k-1)}\right],$$

in order to rewrite the recursive relation (10) as

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} + r^{(k)} - \left(\mathscr{Y}^{(k)} + \mathscr{S}^{(k)}\right) \boldsymbol{\alpha}^{(k)},$$

the minimization problem (11) becoming a simple least-squares problem,

$$\forall k \in \mathbb{N},\ \boldsymbol{\alpha}^{(k)} = \underset{\boldsymbol{\alpha} \in \mathcal{M}_{m_k,1}(\mathbb{R})}{\arg\min} \left\| r^{(k)} - \mathscr{S}^{(k)} \boldsymbol{\alpha} \right\|_2.$$

Assuming that $\mathscr{S}^{(k)}$ is a full-rank matrix, which means that the residuals are linearly independent, and characterising $\boldsymbol{\alpha}^{(k)}$ as the solution of the associated normal equations, with closed form

$$\boldsymbol{\alpha}^{(k)} = \left({\mathscr{S}^{(k)}}^\top \mathscr{S}^{(k)}\right)^{-1} {\mathscr{S}^{(k)}}^\top r^{(k)}, \tag{12}$$

one obtains the relation

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \left(-I_n + \left(\mathscr{Y}^{(k)} + \mathscr{S}^{(k)}\right) \left({\mathscr{S}^{(k)}}^\top \mathscr{S}^{(k)}\right)^{-1} {\mathscr{S}^{(k)}}^\top\right) r^{(k)},$$

which can be identified with the update formula of a quasi-Newton method of multisecant type,

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - G^{(k)} r^{(k)}, \tag{13}$$

with

$$G^{(k)} = -I_n + \left(\mathscr{Y}^{(k)} + \mathscr{S}^{(k)}\right) \left({\mathscr{S}^{(k)}}^\top \mathscr{S}^{(k)}\right)^{-1} {\mathscr{S}^{(k)}}^\top. \tag{14}$$

Here, the matrix $G^{(k)}$ is regarded as an approximate inverse of the Jacobian of the function f at point $x^{(k)}$, satisfying the inverse multiple secant condition

$$G^{(k)} \mathscr{S}^{(k)} = \mathscr{Y}^{(k)}.$$

It moreover minimises the Frobenius norm $\|G^{(k)} + I_n\|_2$ among all the matrices satisfying this condition. Thus, formula (14) can be viewed as a rank-m_k update of $-I_n$ generalising Broyden's second method, which effectively links the Anderson acceleration to a particular class of quasi-Newton methods. Apparently not aware of this connection, Rohwedder and Schneider [50] derived a similar conclusion in their analysis of the DIIS considering a full history of iterates, showing that it corresponds to a projected variant of Broyden's second method proposed by Gay and Schnabel in 1977 (see Algorithm II' in [25]).

Noticing that the generalisation of Broyden's first method aims at directly approximating the Jacobian of the function f at point $x^{(k)}$ by a matrix $B^{(k)}$ which minimises the norm $\|B^{(k)} + I_n\|_2$ subject to the multiple secant condition $B^{(k)} \mathscr{Y}^{(k)} = \mathscr{S}^{(k)}$, yielding, under the assumption that the matrix $\mathscr{Y}^{(k)}$ is full-rank, the matrix

$$B^{(k)} = -I_n + \left(\mathscr{Y}^{(k)} + \mathscr{S}^{(k)}\right) \left({\mathscr{Y}^{(k)}}^\top \mathscr{Y}^{(k)}\right)^{-1} {\mathscr{Y}^{(k)}}^\top,$$

Fang and Saad [20] defined a "type-I" Anderson acceleration (the "type-II" Anderson acceleration corresponding to the original one, as seen above). Indeed, assuming that the matrix ${\mathscr{Y}^{(k)}}^\top \mathscr{S}^{(k)}$ is invertible and applying the Sherman–Morrison–Woodbury formula, it follows that

$$(B^{(k)})^{-1} = -I_n + \left(\mathscr{Y}^{(k)} + \mathscr{S}^{(k)}\right) \left({\mathscr{Y}^{(k)}}^\top \mathscr{S}^{(k)}\right)^{-1} {\mathscr{Y}^{(k)}}^\top,$$

[3] It is sometimes also called the *bad* Broyden method (see [27]).

[4] We continue to assume that the relaxation parameter β does not vary from one iteration to the next.

and substituting $(B^{(k)})^{-1}$ to $G^{(k)}$ in (13) leads to a variant of the original method, the coefficients of which satisfy

$$\tilde{\boldsymbol{\alpha}}^{(k)} = \left(\mathscr{Y}^{(k)\top}\mathscr{S}^{(k)}\right)^{-1}\mathscr{Y}^{(k)\top}r^{(k)},$$

instead of (12).

To end this subsection, let us mention that the relation between the CDIIS applied to ground state calculations and quasi-Newton methods has also been suspected, albeit in heuristic ways, in articles originating from the computational chemistry community (see [57, 37]).

2.4 Equivalence with the GMRES method for linear problems

Consider the application of the DIIS to the solution of a system of n linear equations in n unknowns, with solution x_*, by assuming that the function f in (6) is of the form $f(x) = b - Ax$, where A is a nonsingular matrix of order n and b a vector of $\mathbb{R}^n$. In such a case, one has $g(x) = (I_n - \beta\, A)\, x + \beta\, b$ and the fixed-point iteration method (1) reduces to the stationary Richardson method. We also suppose that all the previous error vectors are kept at each iteration, that is $m = +\infty$, so that $m_k = k$ for each integer k.

A well-known iterative method for the solution of this linear system is the GMRES method [52], in which the approximate solution $x^{(k+1)}$ at the $k+1$th step is the unique vector minimising the Euclidean norm of the residual vector

$$r^{(k+1)} = b - A\, x^{(k+1)}$$

in the affine space

$$\mathcal{W}_{k+1} = \left\{ v = x^{(0)} + z \,|\, z \in \mathcal{K}_{k+1}(A, r^{(0)}) \right\},$$

where, for any integer j greater than or equal to 1, $\mathcal{K}_j(A, r^{(0)}) = \text{span}\left\{r^{(0)}, Ar^{(0)}, \ldots, A^{j-1}r^{(0)}\right\}$ is the *order-j Krylov (linear) subspace* generated by the matrix A and the residual vector $r^{(0)} = b - Ax^{(0)}$, the starting vector $x^{(0)}$ being given. Since each of these Krylov subspaces is contained in the following one, the norm of the residual decreases monotonically with the iterations and the exact solution is obtained in at most n iterations (the idea being that an already good approximation to the exact solution is reached after only a small (relatively to n) number of iterations). More precisely, the following result holds (for a proof, see for instance [46]).

Proposition 2.1 *The GMRES method converges in exactly $\nu(A, r^{(0)})$ steps, where the integer $\nu(A, r^{(0)})$ is the grade[5] of $r^{(0)}$ with respect to A.*

In [60] and [50] (see also [46] for slightly deeper results), it was shown that, the DIIS (or the Anderson acceleration) used to solve a linear system is equivalent to the GMRES method in the following sense.

Theorem 2.2 *Suppose that $x^{(0)}_{\text{DIIS}} = x^{(0)}_{\text{GMRES}} = x^{(0)}$. Then, assuming that the sequence of residual norms does not stagnate[6], that is, assuming there exists a positive integer k such that $r^{(k-1)} \neq 0$ and that $\|r^{(j-1)}\|_2 > \|r^{(j)}\|_2$ for each integer j such that $0 < j < k$, one has*

$$\forall k \in \mathbb{N},\ x^{(k)}_{\text{GMRES}} = \sum_{i=0}^{k} c_i^{(k)} x^{(i)}_{\text{DIIS}}, \tag{15}$$

and

$$\forall k \in \mathbb{N},\ x^{(k+1)}_{\text{DIIS}} = g(x^{(k)}_{\text{GMRES}}). \tag{16}$$

This equivalence can be easily established by observing that, in the linear case, relation (2) reads

$$\forall k \in \mathbb{N},\ x^{(k+1)}_{\text{DIIS}} = \sum_{i=0}^{k} c_i^{(k)} \left(\beta\, b + (I_n - \beta\, A) x^{(i)}_{\text{DIIS}}\right) = \beta\, b + (I_n - \beta\, A)\left(\sum_{i=0}^{k} c_i^{(k)} x^{(i)}_{\text{DIIS}}\right) = g\left(\sum_{i=0}^{k} c_i^{(k)} x^{(i)}_{\text{DIIS}}\right), \tag{17}$$

[5] The grade of a non-zero vector x with respect to a matrix A is the smallest integer ℓ for which there is a non-zero polynomial p of degree ℓ such that $p(A)x = 0$.

[6] As shown in [26], any nonincreasing sequence of residual norms can be produced by the GMRES method. It is then possible for the residual norm to stagnate at some point, say at the kth step, k being a nonzero natural integer, with $r^{(k)} = r^{(k-1)} \neq 0$. In such a case, the equivalence between the methods implies that $x^{(k+1)}_{\text{DIIS}} = x^{(k)}_{\text{DIIS}}$, making the least-square problem associated with the minimisation ill-posed due to the family $\{f(x^{(0)}_{\text{DIIS}}), \ldots, f(x^{(k+1)}_{\text{DIIS}})\}$ not having full rank. The DIIS method then breaks down upon stagnation before the solution has been found, whereas the GMRES method does not. As pointed out in [60], this results in the conditioning of the least-square problem being of utmost importance in the numerical implementation of the method.

and that, setting $x_{\mathrm{DIIS}}^{(i)} = x^{(0)} + z^{(i)}$, $i = 1, \ldots, k$, and using that $c_0 = 1 - \sum_{i=1}^{k} c_i$, the vector $\boldsymbol{c}^{(k)}$ of $\mathbb{R}^k$ is solution to

$$\min_{\boldsymbol{c} \in \mathbb{R}^k} \left\| b - A\, x^{(0)} - A \left(\sum_{i=1}^{k} c_i z^{(i)} \right) \right\|_2 = \min_{\boldsymbol{c} \in \mathbb{R}^k} \left\| r^{(0)} - A \left(\sum_{i=1}^{k} c_i z^{(i)} \right) \right\|_2 .$$

Now, if the set of vectors $\left\{ z^{(1)}, \ldots, z^{(j)} \right\}$ constitutes a basis of $\mathcal{K}_j(A, r^{(0)})$, $1 \leq j \leq k$, the last problem is precisely the one appearing in the GMRES method. Consequently, one finds that

$$x_{\mathrm{GMRES}}^{(k)} = x^{(0)} + \sum_{i=1}^{k} c_i^{(k)} z^{(i)} = \sum_{i=0}^{k} c_i^{(k)} x_{\mathrm{DIIS}}^{(i)},$$

implying (15), which in turn yields (16) using (17). We then just have to show that, for $1 \leq j \leq k$, the family $\left\{ z^{(1)}, \ldots, z^{(j)} \right\}$ is a basis of $\mathcal{K}_j(A, r^{(0)})$. First, one has that

$$z^{(1)} = x_{\mathrm{DIIS}}^{(1)} - x^{(0)} = g(x^{(0)}) - x^{(0)} = f(x^{(0)}) = r^{(0)},$$

which is nonzero due to the assumptions on the residuals. If $k = 1$, the proof is complete. Otherwise, suppose that $k > 1$ and set the inductive hypothesis that $\left\{ z^{(1)}, \ldots, z^{(j)} \right\}$ is a basis of $\mathcal{K}_j(A, r^{(0)})$, for some $1 \leq j < k$. It follows that, using (16),

$$z^{(j+1)} = x_{\mathrm{DIIS}}^{(j+1)} - x^{(0)} = g(x_{\mathrm{GMRES}}^{(j)}) - x^{(0)} = \beta\, b + (I_n - \beta\, A)\, x_{\mathrm{GMRES}}^{(j)} - x^{(0)} = r^{(j)} + \sum_{i=1}^{j} c_i^{(j)} z^{(i)}.$$

Since the vector $r^{(j)}$ belongs to $\mathcal{K}_{j+1}(A, r^{(0)})$ and the linear combination $\sum_{i=1}^{j} c_i^{(j)} z^{(i)}$ is in $\mathcal{K}_j(A, r^{(0)})$, one has that the vector $z^{(j+1)}$ is in $\mathcal{K}_{j+1}(A, r^{(0)})$. Using the assumption on the residuals, we also have that $r^{(j)}$ does not belongs to $\mathcal{K}_j(A, r^{(0)})$, which then shows that the vector $z^{(j+1)}$ cannot depend linearly on $\left\{ z^{(1)}, \ldots, z^{(j)} \right\}$ and that the family $\left\{ z^{(1)}, \ldots, z^{(j+1)} \right\}$ is indeed a basis of $\mathcal{K}_{j+1}(A, r^{(0)})$.

Assuming that a full history of iterates is kept and that no stagnation occurs, this equivalence directly provides convergence results in the linear case for the DIIS in view of the well-known theory for the GMRES method. It also justifies the inclusion of Krylov's name in some of the rediscoveries of the Anderson acceleration we previously mentioned. Of course, if the number of stored iterates is fixed or if "restarts" are allowed over the course of the computation, in the spirit of the restarted GMRES method, GMRES(m), these results are no longer valid. Nevertheless, the behaviour of GMRES(m) has been studied in a number of particular cases, and most of the various restart or truncation strategies (see [16] and the references therein for instance), proposed over the years in order to compensate for the loss of information by selectively retaining some of it from earlier iterations, could be worked out in the context of the restarted CDIIS introduced in section 3.

Walker and Ni [60] showed in a similar way that the type-I Anderson acceleration, recalled in the preceding subsection, is essentially equivalent to the full orthogonalisation method (FOM) based on the Arnoldi process (see section 2 of [52]) in the linear case.

2.5 Existing convergence theories in the nonlinear case

In light of the previous subsection, it appears that the convergence theory for the DIIS in the linear case is intimately linked to that of the GMRES method. On the contrary, this theory is far from being established when the method is applied to nonlinear problems. The interpretation of the DIIS as a particular instance of quasi-Newton methods does not provide an answer, as there is no general convergence theory for these methods. Nevertheless, results have emerged in the literature in the recent years.

In [50], Rohwedder and Schneider showed that the DIIS is locally linearly q-convergent when applied to the solution a general nonlinear problem, set in a Hilbert space and of the form (6), by a fixed-point iteration defined by the function (7) with the choice $\beta = -1$, under the assumptions that the underlying mapping is locally contractive and that error vectors associated with the stored iterates satisfy a linear independence condition similar to the one used in the next section. They also obtained a refined estimate for the residual at a given step that allows them to discuss the possibility of obtaining a superlinear convergence rate for the method. Their analysis is based on the equivalence between the DIIS and a projected variant of Broyden's second method, the convergence properties and an improvement of estimates appearing in the proof of convergence by Gay and Schnabel in [25] on the one hand, and an interpretation of the DIIS as an inexact Newton method on the other hand.

More recently, Kelley and Toth [59] studied the use of the Anderson acceleration on a fixed-point iteration in $\mathbb{R}^n$, also based the function (7) in which one has set $\beta = 1$, the history of error vectors being of fixed maximal depth m.

Assuming that the fixed-point mapping is Lipschitz continuously differentiable and a contraction in a neighbourhood of the solution and with the requirement that the extrapolation coefficients remain uniformly bounded in the ℓ^1-norm (a condition which can only be verified *a posteriori*), they proved that the method converges locally r-linearly (the convergence of the sequence of errors following from the convergence of the sequence of residuals). When $m = 1$, they showed that the sequence of residuals converges q-linearly if the fixed-point mapping is sufficiently contractive and the initial guess is close enough from the solution. This analysis was later extended to the case where the evaluation of the fixed point map is corrupted with errors in [58]. Chen and Kelley [13] also obtained global and local convergence results for a variant of the method in which a non-negativity constraint on the coefficients is imposed (an idea previously used in the EDIIS method [38]), generalising, under weakened hypotheses, the results in [59].

A stabilised version (which includes a regularisation ensuring the non-singularity of the approximated Jacobian, a restart strategy guaranteeing a certain linear independence of the differences of successive stored iterates, and a safeguard mechanism guaranteeing a decrease of the residual norm) of the type-I Anderson acceleration is introduced by Zhang *et al.* in [64]. A global convergence result for the corresponding acceleration applied to a Krasnoselskii–Mann iteration in $\mathbb{R}^n$ is proved under the sole assumption that the fixed-point function is non-expansive and potentially non-smooth.

Under strong assumptions on the fixed-point function and in a Hilbert space setting, Evans *et al.* proved in [18] that the Anderson acceleration increases the convergence rate, the gain at each step being measured as a quotient between the linearised residual at the current iteration and the residual at the previous one, on the condition that the extrapolation coefficients remain both bounded and bounded away from zero. Their results also show that both Anderson's acceleration and the use of damping can extend the region in which the fixed-point iteration is convergent.

3 Restarted and adaptive-depth Anderson–Pulay acceleration

A particularity of the commutator-DIIS variant proposed by Pulay for electronic ground state calculations is that, when interpreted in terms of an abstract fixed-point function g and an abstract residual function f, the functions g and f do not satisfy a relation of the form (7), as they take their values in different spaces (see Section 4). As a consequence, the convergence theories found in [50] or in [59] do not apply.

This observation leads us to introduce an abstract framework in which the class of Anderson–Pulay acceleration algorithms is defined. This general class encompasses the DIIS, the CDIIS and the other methods reviewed in the previous section. In addition, two modifications of this procedure are proposed: one including a restart condition in the spirit of Gay and Schnabel [25] (see Algorithm 2), and another in which the depth is adapted at each step according to a prescribed criterion (see Algorithm 3).

3.1 Description on an abstract problem

Consider solving the nonlinear problem (6), with solution x_* in $\mathbb{R}^n$, by the fixed-point iteration (1). Here, we do *not* assume that the fixed-point function g and the residual functions f are related by a relation of the form (7).

Denoting by $r^{(k)} = f(x^{(k)})$ the residual associated with the iterate $x^{(k)}$ at step k, the Anderson–Pulay acceleration procedure, in its classical fixed-depth form, applied to the solution of the problem is summarised in Algorithm 1.

Algorithm 1: Fixed-depth Anderson–Pulay acceleration for the solution of $f(x) = 0$ by a fixed-point method using a function g.

Data: $x^{(0)}$, tol
$r^{(0)} = f(x^{(0)})$
$k = 0$
$m_k = 0$
while $\left\|r^{(k)}\right\|_2 > tol$ **do**
  solve the constrained least-squares problem for the coefficients $\{c_i^{(k)}\}_{i=0,\dots,m_k}$
  $\widetilde{x}^{(k+1)} = \sum_{i=0}^{m_k} c_i^{(k)} x^{(k-m_k+i)}$
  $x^{(k+1)} = g(\widetilde{x}^{(k+1)})$
  $m_k = m_k + 1$
  $r^{(k+1)} = f(x^{(k+1)})$
  $k = k + 1$
end

As already discussed in Subsection 2.1, the numerical solution of constrained linear least-squares problem for the extrapolation coefficients requires some care. Indeed, the matrix of linear system (4) resulting from the use of a Lagrange multiplier accounting for the constraint has to be well-conditioned for the method to be applicable in

floating-point arithmetic. Unconstrained formulations of the problem exist, as employed in the Anderson acceleration (see (9)) or its numerous reinventions (see (11) for instance), each leading to an equivalent Anderson–Pulay acceleration algorithm, but with a possibly different condition number for the matrix of the linear system associated with the least-squares problem.

Note that a bound on the extrapolation coefficients is also needed when investigating the convergence of the method. In [59], this bound is presented as a requirement, following, for instance, from the uniform well-conditioning of the least-squares problem, which cannot be *a priori* guaranteed. Nevertheless, such a condition can be enforced *a posteriori* by diminishing the value of the integer m_k. In practice, this may be achieved by monitoring the condition number of the least-squares coefficient matrix and by dropping as many of its left-most columns as needed to have this condition number below a prescribed threshold, as proposed in [60]. Another possibility consists in ensuring that the residual vectors associated with the stored iterates fulfil a kind of linear independence condition like the one found in [25, 50]. To enforce this condition in practice, one can simply choose to "restart" the method, by resetting the value of m_k to zero and discarding the iterates previously stored, as soon as an "almost" linear dependence is detected. This approach is adopted for the first of the modifications of Algorithm 1 we propose and analyse in the present work.

More precisely, assuming that $m_k \geq 1$ at step k and setting $s^{(i)} = r^{(k-m_k+i)} - r^{(k-m_k)}$, $i = 1, \ldots, m_k$, the first condition we use *"recognizes that, in general, the projection of $s^{(i)}$ orthogonal to the subspace spanned by $s^{(1)}, \ldots, s^{(i-1)}$ must be the zero vector for some $i \leq n$"* (see [25]). The algorithm will therefore "restart" at step $k+1$ (that is, the number m_{k+1} will be set to 0) if the norm of $s^{(k+1)} - \Pi_k s^{(k+1)}$, where the operator Π_k denotes the orthogonal projector onto $\mathrm{span}\{s^{(k-m_k+1)}, \ldots, s^{(k)}\}$, is "small" compared to the norm of $s^{(k+1)}$, that is, if the inequality

$$\tau \, \|r^{(k+1)} - r^{(k-m_k)}\|_2 > \|(Id - \Pi_k)(r^{(k+1)} - r^{(k-m_k)})\|_2, \tag{18}$$

is satisfied, where the parameter τ is a real number chosen between 0 and 1. Otherwise, the integer m_k is incremented of one unit with the iteration number k. One can view condition (18) as a near-linear dependence criterion for the set of differences of stored residuals. Note that an analogue for the type-I Anderson acceleration (thus involving differences of stored iterates) of this criterion is used in [64].

Such a restart condition is particularly adapted to a specific unconstrained formulation of the linear least-squares problem

$$\boldsymbol{c}^{(k)} = \underset{\substack{(c_0,\ldots,c_{m_k}) \in \mathbb{R}^{m_k+1} \\ \sum_{i=0}^{m_k} c_i = 1}}{\arg\min} \left\| \sum_{i=0}^{m_k} c_i \, r^{(k-m_k+i)} \right\|_2, \tag{19}$$

to be effectively employed in the numerical practice. Indeed, due to the constraint on the coefficients in (19), one may write

$$\sum_{i=0}^{m_k} c_i \, r^{(k-m_k+i)} = r^{(k-m_k)} + \sum_{i=1}^{m_k} c_i \, \left(r^{(k-m_k+i)} - r^{(k-m_k)} \right) = r^{(k-m_k)} + \sum_{i=1}^{m_k} \gamma_i \, s^{(k-m_k+i)},$$

by setting

$$\forall k \in \mathbb{N}^*, \ \forall i \in \{1, \ldots, m_k\}, \ \gamma_i = c_i \text{ and } s^{(k-m_k+i)} = r^{(k-m_k+i)} - r^{(k-m_k)}.$$

The corresponding modified version of the Anderson–Pulay acceleration is given in Algorithm 2.

Algorithm 2: Restarted Anderson–Pulay acceleration for the solution of $f(x) = 0$ by a fixed-point method using a function g.

Data: $x^{(0)}$, tol, τ
$r^{(0)} = f(x^{(0)})$
$x^{(1)} = g(x^{(0)})$, $r^{(1)} = f(x^{(1)})$
$s^{(1)} = r^{(1)} - r^{(0)}$
$k = 1$
$m_k = 1$
while $\left\|r^{(k)}\right\|_2 > tol$ **do**
  solve the unconstrained least-squares problem for the coefficients $\{\gamma_i^{(k)}\}_{i=1,\dots,m_k}$
  $\widetilde{x}^{(k+1)} = x^{(k-m_k)} + \sum_{i=1}^{m_k} \gamma_i^{(k)} (x^{(k-m_k+i)} - x^{(k-m_k)})$
  $x^{(k+1)} = g(\widetilde{x}^{(k+1)})$, $r^{(k+1)} = f(x^{(k+1)})$
  $s^{(k+1)} = r^{(k+1)} - r^{(k-m_k)}$
  compute $\Pi_k s^{(k+1)}$ (the orthogonal projection of $s^{(k+1)}$ onto the span of the family $\{s^{(k-m_k+1)}, \dots, s^{(k)}\}$)
  if $\tau \left\|s^{(k+1)}\right\|_2 > \left\|(Id - \Pi_k) s^{(k+1)}\right\|_2$ **then**
    $m_k = 0$
  else
    $m_k = m_k + 1$
  end
  $k = k + 1$
end

Note that, with this instance of the method, all the stored iterates, except for the last one, are discarded when a restart occurs. In some cases, a temporary slowdown of the convergence is observed in practical computations (see Section 5). To try to remedy such an inconvenience, we introduce an adaptive-depth version of the procedure, based on an update of the set of stored iterates at step $k+1$, which keeps the iterate $x^{(i)}$, $i = k + 1 - m_{k+1}, \dots, k$ if the inequality

$$\delta \|r^{(i)}\|_2 < \|r^{(k+1)}\|_2$$

is satisfied, where the parameter δ is a real number chosen between 0 and 1. This criterion is, to the best of our knowledge, new, and the corresponding variant of the Anderson–Pulay acceleration is given in Algorithm 3.

Algorithm 3: Adaptive-depth Anderson–Pulay acceleration for the solution of $f(x) = 0$ by a fixed-point method using a function g.

Data: $x^{(0)}$, tol, δ
$r^{(0)} = f(x^{(0)})$
$x^{(1)} = g(x^{(0)})$, $r^{(1)} = f(x^{(1)})$
$s^{(1)} = r^{(1)} - r^{(0)}$
$k = 1$
$m_k = 1$
while $\left\|r^{(k)}\right\|_2 > tol$ **do**
  solve the unconstrained least-squares problem for the coefficients $\{\alpha_i^{(k)}\}_{i=1,\dots,m_k}$
  $\widetilde{x}^{(k+1)} = x^{(k)} - \sum_{i=1}^{m_k} \alpha_i^{(k)} (x^{(k-m_k+i+1)} - x^{(k-m_k+i)})$
  $x^{(k+1)} = g(\widetilde{x}^{(k+1)})$, $r^{(k+1)} = f(x^{(k+1)})$
  $s^{(k+1)} = r^{(k+1)} - r^{(k)}$
  set m_{k+1} the largest integer $m \leq m_k + 1$ such that for $k + 1 - m \leq i \leq k$, $\delta \|r^{(i)}\|_2 < \|r^{(k+1)}\|_2$
  $k = k + 1$
end

In the last algorithm, one may observe that the unconstrained form of the least-squares problem for the coefficients uses the differences of residuals associated with *successive* iterates, that is, $s^{(i)} = r^{(i)} - r^{(i-1)}$, $i = k - m_k + 1, \dots, k$, which is a computationally convenient choice when storing a set of iterates whose depth is adapted at each step. One then has $\alpha_i^{(k)} = \sum_{j=0}^{i-1} c_j^{(k)}$, $i = 1, \dots, m_k$.

3.2 Linear convergence

We shall now prove the convergence of the modified Anderson–Pulay acceleration methods devised in the previous subsection. To this end, we consider the following functional setting.

Let n and p be two nonzero natural integers such that $p \leq n$, Σ be a smooth submanifold in $\mathbb{R}^n$, V be an open subset of $\mathbb{R}^n$, f be a function in $\mathscr{C}^2(V, \mathbb{R}^p)$, g be a function in $\mathscr{C}^2(V, \Sigma)$, and x_* in $V \cap \Sigma$ be a fixed point of g

satisfying $f(x_*) = 0$. We moreover suppose that:

Assumption 1. There exists a constant $K \in (0,1)$ such that

$$\forall x \in V \cap \Sigma,\ \|f(g(x))\|_2 \leq K\,\|f(x)\|_2.$$

Assumption 2. There exists a constant $\sigma > 0$ such that

$$\forall x \in V \cap \Sigma,\ \sigma\|x - x_*\|_2 \leq \|f(x)\|_2.$$

Note that Assumption 1 ensures that x_* is the unique fixed point of g in $V \cap \Sigma$.

Since the functions f and g are both of class $\mathscr{C}^2$, we may assume, taking V a smaller neighbourhood of x_*, that

$$\forall (x,y) \in V^2,\ \|f(x) - f(y)\|_2 \leq 2\,\|\mathrm{D}f(x_*)\|_2\|x-y\|_2 \text{ and } \|(f\circ g)(x) - (f\circ g)(y)\|_2 \leq 2\,\|\mathrm{D}f(x_*)\circ \mathrm{D}g(x_*)\|_2\|x-y\|_2, \tag{20}$$

where $\mathrm{D}f(x_*)$ and $\mathrm{D}g(x_*)$ are the respective differentials of f and g at point x_*, and also that, for some positive constant L,

$$\forall x \in V,\ \|f(x) - \mathrm{D}f(x_*)(x - x_*)\|_2 \leq \frac{L}{2}\,{\|x - x_*\|_2}^2. \tag{21}$$

For a neighbourhood U of x_*, included in the tubular neighbourhood of $\Sigma \cap \overline{V}$, the nonlinear projection operator P_Σ onto Σ, defined as

$$\forall x \in U,\ P_\Sigma(x) = \operatorname*{arg\,min}_{y\in\Sigma} \|x - y\|_2,$$

is well-defined [55]. Let $P_{T_{x_*}\Sigma}$ be the orthogonal projector onto the tangent space $T_{x_*}\Sigma$ to Σ at x_*. For U small enough, there exists a positive constant M such that the following holds

$$\forall x \in U,\ \|P_\Sigma(x) - (x_* + P_{T_{x_*}\Sigma}(x - x_*))\|_2 \leq \frac{M}{2}{\|x - x_*\|_2}^2. \tag{22}$$

Note that the introduction of the submanifold Σ is needed in view of the applications of the methods to quantum chemistry computations, as presented in Subsection 4.5 ; the above assumptions will be discussed in such a context (see Subection 4.6). As seen in Section 2, the usual abstract framework for the DIIS or the Anderson acceleration is simpler: there is no submanifold (that is, one takes $\Sigma = \mathbb{R}^n$), the integers p and n are the same, and the function g and f are related by (7). All the results that follow are, of course, still valid and, as far as we know, new in this more restrictive setting.

Let us state the first main result of this Section, concerning the local r-linear convergence (with respect to restarts) of the restarted Anderson–Pulay acceleration.

Theorem 3.1 *Let Assumptions 1 and 2 hold and let μ be a real number such that $K < \mu < 1$. There exists a positive constant C_μ, depending on p, K, L, M, σ, μ, but not on τ, such that, if the initialisation $x^{(0)}$ is sufficiently close to x_*, in the sense that $\|r^{(0)}\|_2 \leq C_\mu\tau^{2p}$, and one runs Algorithm 2 then, as long as $\|r^{(k)}\|_2 > 0$, one has*

$$m_k \leq \min(k,p) \tag{23}$$

$$\|c^{(k)}\|_\infty \leq C_{m_k}\left(1 + \frac{1}{(\tau(1-\mu))^{m_k}}\right), \tag{24}$$

where C_{m_k} is a positive constant depending only on m_k, and

$$\|r^{(k+1)}\|_2 \leq \mu^{m_k+1}\|r^{(k-m_k)}\|_2. \tag{25}$$

The method is thus locally r-linearly convergent.

Moreover, if a restart occurs at step $k+1$, there exists a positive constant C, independent of τ and of the preceding iterates $(x^{(i)})_{0\leq i\leq k}$, such that

$$(1 - K(1+\tau))\,\|r^{(k+1)}\|_2 \leq K\tau\|r^{(k-m_k)}\|_2 + \frac{C}{\tau^{2\,m_k}}{\|r^{(k-m_k)}\|_2}^2. \tag{26}$$

The second result deals with the convergence of the adaptive-depth version of the method.

Theorem 3.2 *Let Assumptions 1 and 2 hold and let μ be a real number such that $K < \mu < 1$. Let the parameter δ in Algorithm 3 satisfy $0 < \delta < K$. There exists a constant $c_\mu > 0$, depending on p, K, L, M, σ, μ, but not on δ, such that, if $\|r^{(0)}\|_2 \leq c_\mu \delta^2$ and one runs Algorithm 3, then, as long as $\|r^{(k)}\|_2 > 0$, one has*

$$m_k \leq \min(k, p), \tag{27}$$

$$\|c^{(k)}\|_\infty \leq C_{m_k} \left(1 + \left(\frac{1-\mu}{\mu}\right)^{m_k}\right), \tag{28}$$

where C_{m_k} is a positive constant depending only on m_k, and

$$\|r^{(k+1)}\|_2 \leq \mu \|r^{(k)}\|_2. \tag{29}$$

The method is thus locally q-linearly convergent.

Moreover, if $k - m_k \geq 1$, then

$$\|r^{(k)}\|_2 \leq \delta \|r^{(k-m_k-1)}\|_2. \tag{30}$$

3.3 Acceleration and reaching superlinear convergence

In this subsection, we study from a mathematical viewpoint the local acceleration properties of the proposed variants of the Anderson–Pulay acceleration. We prove that, for any choice of a real number λ in $(0, K)$, r-linear convergence at rate λ can be achieved, meaning that $\|r^{(k)}\|_2 = \mathcal{O}(\lambda^k)$, with Algorithm 2 (for τ small enough) or with Algorithm 3 (for δ small enough). These results are consequences of Theorems 3.1 and 3.2 respectively, when one assumes that the starting point is sufficiently close to the solution.

Corollary 3.3 *Using the notations and under the assumptions of Theorem 3.1, let λ be a real number such that $0 < \lambda < K$. There are two positive constants $\varepsilon(\lambda)$ and $\tau(\lambda)$, depending on λ and also on the parameters p, K, L, M, and μ, such that, if $\tau = \tau(\lambda)$ and $\|r^{(0)}\|_2 \leq \varepsilon(\lambda)$ in Algorithm 2, then one has*

$$\forall k \in \mathbb{N},\ \|r^{(k)}\|_2 \leq \mu^{\min(k,p)} \lambda^{\max(0,k-p)} \|r^{(0)}\|_2.$$

Proof. If the constants $\varepsilon(\lambda)$ and $\tau(\lambda)$ in the statement are such that $\varepsilon(\lambda) \leq C_\mu \tau(\lambda)^{2p}$, Theorem 3.1 guarantees that $m_k \leq p$ and $\|r^{(k)}\|_2 \leq \mu^k \|r^{(0)}\|_2$ for all k. What remains to be done is to satisfy the condition

$$\|r^{(k+1)}\|_2 \leq \lambda^{p+1} \|r^{(k-m_k)}\|_2$$

when a restart occurs. To do this, let us rewrite estimate (26) of Theorem 3.1 into

$$\|r^{(k+1)}\|_2 \leq (1 - K(1+\tau))^{-1} \left(K\tau + \frac{C}{\tau^{2\,m_k}} \|r^{(k-m_k)}\|_2\right) \|r^{(k-m_k)}\|_2.$$

We impose that $\tau(\lambda) = \left(\frac{2Cp}{K}\varepsilon(\lambda)\right)^{\frac{1}{2p+1}}$, with $\varepsilon(\lambda)$ so small that $K(1+\tau(\lambda)) < \frac{1}{2}$ and $\varepsilon(\lambda) \leq C_\mu \tau(\lambda)^{2p}$. The above inequality then implies that

$$\|r^{(k+1)}\|_2 \leq 2\left(1 + \frac{1}{2p}\right) K \left(\frac{2Cp}{K}\varepsilon(\lambda)\right)^{\frac{1}{2p+1}} \|r^{(k-m_k)}\|_2,$$

and $\varepsilon(\lambda)$ just needs to be such that $2\left(1 + \frac{1}{2p}\right) K \left(\frac{2Cp}{K}\varepsilon(\lambda)\right)^{\frac{1}{2p+1}} \leq \lambda^{p+1}$. This is clearly possible and the proof is complete. □

Note that the theoretical values of $\tau(\lambda)$ and $\varepsilon(\lambda)$ in the above proof are unreasonably small and certainly not representative of what is observed in applications. We will indeed see in Section 5 that acceleration is commonly achieved in practice.

A similar result for acceleration in Algorithm 3 follows immediately from Theorem 3.2, so that we omit its proof.

Corollary 3.4 *Using the notations and under the assumptions of Theorem 3.2, let λ be a real number such that $0 < \lambda < K$. Then, setting $\delta = \lambda^{p+1}$ in Algorithm 3 and supposing that $\|r^{(0)}\|_2 \leq c_\mu \lambda^{2p+2}$, one has*

$$\forall k \in \mathbb{N},\ \|r^{(k)}\|_2 \leq \mu^{\min(k,p)} \lambda^{\max(0,k-p)} \|r^{(0)}\|_2.$$

In some references, one can find mentions of the DIIS exhibiting a "superlinear convergence behaviour", in the sense that the ratio of successive residual norms decreases as the iterations progress, generally when the iterates are close to the solution (one talks of asymptotic q-superlinear convergence). Rohwedder and Schneider discussed in [50] the circumstances under which this may occur for the DIIS method they analysed in the same paper. In the rest of this subsection, we suggest some modifications of the restarted and adaptive-depth Anderson–Pulay accelerations, in which the value of the parameter τ (for Algorithm 2) or δ (for Algorithm 3) is slowly decreased along the iteration, and establish r-superlinear convergence rates for the resulting algorithms. Once again, these results follow from Theorems 3.1 and 3.2.

Let us first consider Algorithm 2. A by-product of the linear convergence analysis, estimate (26) in Theorem 3.1 allows a local r-superlinear convergence result, by carefully changing the value of the parameter τ at each restart. More precisely, we have the following result.

Corollary 3.5 *Using the notations and under the assumptions of Theorem 3.1, suppose furthermore that* $\|r^{(0)}\|_2 \leq \varepsilon^{2p+1}$*, where* $K(1+\varepsilon) < 1$*,* $\frac{\|r^{(0)}\|_2}{C_\mu} < 1$*, and that the restart parameter used at iteration* k *in Algorithm 2 is given by* $\tau^{(k)} = \left(\frac{\|r^{(k-m_k)}\|_2}{C_\mu}\right)^{\frac{1}{2p+1}}$*, which means that the value of the parameter is modified with each restart. Then, one has*

$$\forall k \in \mathbb{N},\ k - m_k \geq 1,\ \|r^{(k)}\|_2 \leq \frac{C+K}{1-K(1+\varepsilon)} \mu^{m_k} \|r^{(k-m_k-m_{k-m_k})}\|_2^{1+\frac{1}{2p+1}}.$$

Proof. For $k = 0$, one has $k - m_k = 0$ so that $\|r^{(0)}\|_2 = C_\mu(\tau^{(0)})^{2p+1} \leq C_\mu(\tau^{(0)})^{2p}$. Theorem 3.1 then applies with τ replaced by $\tau^{(k)}$ at step k and, using that the method is r-linearly convergent between restarts, one has

$$\forall k \in \mathbb{N},\ \tau^{(k)} = \|r^{(k-m_k)}\|_2^{\frac{1}{2p+1}} \leq \|r^{(0)}\|_2^{\frac{1}{2p+1}} \leq \varepsilon.$$

In particular, if a restart occurs at step $k+1$, one obtains, by replacing τ by ε in the left-hand side of inequality (26) in Theorem 3.1 and by the value of $\tau^{(k)}$ in its right-hand side,

$$\|r^{(k+1)}\|_2 \leq \frac{C+K}{1-K(1+\varepsilon)} \|r^{(k-m_k)}\|_2^{1+\frac{1}{2p+1}}.$$

The desired inequality then follows from the r-linear convergence of the method. □

For Algorithm 3, the superlinear convergence follows straightforwardly from bound (30) in Theorem 3.2, as soon as the parameter δ is adequately chosen at each step. The result is the following.

Corollary 3.6 *Using the notations and under the assumptions of Theorem 3.2, suppose furthermore that the parameter* δ *used in Algorithm 3 is given at iteration* k *by* $\delta^{(k)} = \sqrt{\frac{\|r^{(k-m_k)}\|_2}{c_\mu}}$*. Then, one has*

$$\forall k \in \mathbb{N},\ k - m_k \geq 1,\ \|r^{(k)}\|_2 \leq \sqrt{\frac{\mu}{c_\mu}} \|r^{(k-m_k-1)}\|_2^{3/2}.$$

Proof. For $k = 0$, one has $k - m_k = 0$ so that $\|r^{(0)}\|_2 = c_\mu(\delta^{(0)})^2$. Theorem 3.2 then applies with δ replaced by $\delta^{(k)}$ at step k. Using the linear convergence of the method, it then follows from estimate (30) of this Theorem that

$$\forall k \in \mathbb{N},\ k - m_k \geq 1,\ \|r^{(k)}\|_2 \leq \delta^{(k)}\|r^{(k-m_k-1)}\|_2 = \sqrt{\frac{\|r^{(k-m_k)}\|_2}{c_\mu}}\|r^{(k-m_k-1)}\|_2 \leq \sqrt{\frac{\mu}{c_\mu}} \|r^{(k-m_k-1)}\|_2^{3/2}.$$

□

3.4 Proofs of the results

The idea behind most local convergence proofs for the DIIS (or the Anderson acceleration) is that, for functionals f and g which are linear, the process converges in a number of steps at most equal to the dimension of the codomain of f, if the fixed-point iteration (1) converges at a linear rate. In the nonlinear case, this is no longer true due to the presence of quadratic terms, but these additional terms should allow convergence at a superlinear rate. More exactly, with a sufficiently large depth and starting close enough to the solution, one should obtain linear convergence at any given rate.

To prove this claim, one has to control the quadratic errors and a bound on the size of the extrapolation coefficients at each step is needed. This amounts to quantitatively measuring the affine independence of the residuals, since the optimal coefficients are solution to a least-squares problem whose solubility is directly related to this independence.

In the existing literature, it is generally[7] assumed that the coefficients remain bounded throughout the iteration. While we do not know how to prove *a priori* such an estimate for the "classical" fixed-depth DIIS, the mechanisms employed in the restarted and adaptive-depth variants we study allow to derive one. A key theoretical ingredient is that, when the problem for the extrapolation coefficients becomes poorly conditioned, it is necessarily because the latest stored residual is much smaller than the oldest one. This phenomenon is rigorously described in Lemma 3.7 below, which constitutes the main technical tool in our convergence and acceleration proofs.

In the restarted Anderson–Pulay acceleration, the role of the parameter τ is to control the affine independence of the residuals. Lemma 3.7 shows that, as long as a restart does not occur, the least-squares problem remains well-conditioned and the size of the coefficients is bounded, whereas, when it does, the norm of the last stored residual is necessarily much smaller than that of the oldest that was kept to calculate it. Since there must be a restart after at most $p+1$ consecutive iterations, convergence with acceleration can be established.

In the adaptive-depth version of the Anderson–Pulay acceleration, the parameter δ is used to eliminate the stored iterates which are not "relevant" enough when compared with the most recent one. This criterion does not directly quantify the independence of the residuals. It is certainly not good when the initial guess is chosen far from the solution, since a large number of iterates will be kept in that case. However, starting close enough to the solution, Lemma 3.7 allows to inductively prove a bound on the extrapolation coefficients, for reasons similar to those invoked with the restarted variant. Indeed, at each step, either the stored residuals are affinely independent, and the extrapolation coefficients are bounded, or the norm of the last stored residual is smaller than δ times that of some of the oldest ones. In the latter case, the criterion will discard the corresponding iterates, restoring the conditioning of the least-squares problem at the next step. In addition, observing that a stored iterate is dismissed at least once in every $p+1$ consecutive iterations, it is inferred that accelerated convergence is possible for δ chosen small enough.

3.4.1 A preliminary lemma

We first introduce some notations. For any integers k and m_k in $\mathbb{N}$ such that $m_k \leq \min(k,p)$, consider a set $\{r^{(k-m_k)},\dots,r^{(k)}\}$ of vectors of $\mathbb{R}^p$, and, for any integer ℓ in $\{0,\dots,m_k\}$, let us denote by $\mathcal{A}_\ell^{(k)}$ the affine span of $\{r^{(k-m_k)},\dots,r^{(k-m_k+\ell)}\}$, that is

$$\mathcal{A}_\ell^{(k)} = \mathrm{Aff}\left\{r^{(k-m_k)}, r^{(k-m_k+1)},\dots,r^{(k-m_k+\ell)}\right\} = \left\{ r = \sum_{i=0}^{\ell} c_i\, r^{(k-m_k+i)} \mid \sum_{i=0}^{\ell} c_i = 1 \right\},$$

and, for any integer ℓ in $\{1,\dots,m_k\}$, by $d_\ell^{(k)}$ the distance between $r^{(k-m_k+\ell)}$ and $\mathcal{A}_{\ell-1}^{(k)}$, that is

$$d_\ell^{(k)} = \min_{r\in\mathcal{A}_{\ell-1}^{(k)}} \|r^{(k-m_k+\ell)} - r\|_2.$$

Lemma 3.7 *Let k and m_k be positive integers such that $m_k \leq k$, $x^{(k-m_k)},\dots,x^{(k)}$ be a set of vectors in $V\cap\Sigma\setminus\{x_*\}$ and t be a positive real number. For any integer i in $\{0,\dots,m_k\}$, set $r^{(k-m_k+i)} = f(x^{(k-m_k+i)})$, and assume that*

$$\forall i \in \{1,\dots,m_k\},\ d_i^{(k)} \geq t \max_{j\in\{i,\dots,m_k\}} \|r^{(k-m_k+j)}\|_2. \tag{31}$$

Consider the unique $\tilde{c}$ in $\mathbb{R}^{m_k+1}$ such that $\sum_{i=0}^{m_k}\tilde{c}_i = 1$ and $\|\sum_{i=0}^{m_k}\tilde{c}_i r^{(k-m_k+i)}\|_2 = \mathrm{dist}_2(0,\mathcal{A}_{m_k}^{(k)})$. Then, there exists a positive constant C_{m_k}, depending only on m_k, such that

$$\|\tilde{c}\|_\infty \leq C_{m_k}\,(1+t^{-m_k}), \tag{32}$$

Moreover assuming that $\tilde{x}^{(k+1)} = \sum_{i=0}^{m_k}\tilde{c}_i\, x^{(k-m_k+i)}$ belongs to $V\cap g^{-1}(V)$, define $x^{(k+1)} = g(\tilde{x}^{(k+1)})$, $r^{(k+1)} = f(x^{(k+1)})$, and set $d_{m_k+1}^{(k)} = \mathrm{dist}_2(r^{(k+1)},\mathcal{A}_{m_k}^{(k)})$. Then, there exists a positive constant κ, depending on m_k, K, L, M and σ, such that

$$\|r^{(k+1)}\|_2 \leq K\,\mathrm{dist}_2(0,\mathcal{A}_{m_k}^{(k)}) + \kappa(1+t^{-2m_k}) \max_{i\in\{0,\dots,m_k\}} {\|r^{(k-m_k+i)}\|_2}^2. \tag{33}$$

and

$$(1-K)\,\|r^{(k+1)}\|_2 \leq K\, d_{m_k+1}^{(k)} + \kappa(1+t^{-2m_k}) \max_{i\in\{0,\dots,m_k\}} {\|r^{(k-m_k+i)}\|_2}^2. \tag{34}$$

[7] An exception is made in the paper by Zhang *et al.* [64], where a bound on the coefficients is shown for a restarted type-I Anderson acceleration method (the DIIS rather corresponds to a type-II Anderson acceleration). The authors then prove a global linear convergence result when f is the gradient of a convex functional, but the linear rate they obtain is no better than the rate of the basic iteration process.

PROOF. Set $s^{(i)} = r^{(k-m_k+i)} - r^{(k-m_k+i-1)}$, for any integer i in $\{1,\dots,m_k\}$, $q^{(1)} = s^{(1)}$ and $q^{(i)} = (Id - \Pi_{\boldsymbol{\mathcal{A}}_{i-1}^{(k)}})s^{(i)}$, for any integer i in $\{2,\dots,m_k\}$, where $\Pi_{\boldsymbol{\mathcal{A}}_{i-1}^{(k)}}$ denotes the orthogonal projector onto $\boldsymbol{\mathcal{A}}_{i-1}^{(k)}$, the underlying vector space of $\mathcal{A}_{i-1}^{(k)}$. Then, the set of vectors $\{q^{(i)}\}_{1\le i\le m_k}$ is mutually orthogonal and, for any integer i in $\{1,\dots,m_k\}$, one has $\|q^{(i)}\|_2 = d_i^{(k)}$.

We may then write

$$\sum_{i=0}^{m_k} \tilde{c}_i\, r^{(k-m_k+i)} = r^{(k)} + \sum_{i=1}^{m_k} \tilde{\zeta}_i s^{(i)} = r^{(k)} + \sum_{i=1}^{m_k} \tilde{\lambda}_i q^{(i)},$$

with $\tilde{\lambda}_i = -\frac{(q^{(i)})^\top r^{(k)}}{(d_i^{(k)})^2}$, so that $\left|\tilde{\lambda}_i\right| \le \frac{\|r^{(k)}\|_2}{d_i^{(k)}} \le \frac{1}{t}$, due to lower bound (31).

On the other hand, $\tilde{\zeta} = P\tilde{\lambda}$, where P is the change-of-basis matrix from $\{s^{(i)}\}_{1\le i\le m_k}$ to $\{q^{(i)}\}_{1\le i\le m_k}$. In the same way, for any integer j in $\{2,\dots,m_k\}$, let $P^{(j)}$ be the change-of-basis matrix from $\{q^{(1)},\dots,q^{(j-1)},s^{(j)},\dots,s^{(m_k)}\}$ to $\{q^{(1)},\dots,q^{(j)},s^{(j+1)},\dots,s^{(m_k)}\}$,. Since $q^{(1)} = s^{(1)}$, and $q^{(j)} = s^{(j)} - \sum_{i=1}^{j-1} \frac{(s^{(j)})^\top q^{(i)}}{(q^{(i)})^\top q^{(i)}}\, q^{(i)}$, we have

$$\forall i \in \{1,\dots,m_k\},\ (P^{(j)})_{ii} = 1 \text{ and, } \forall i \in \{1,\dots,j-1\},\ (P^{(j)})_{ij} = -\frac{(s^{(j)})^\top q^{(i)}}{(q^{(i)})^\top v^{(i)}},$$

all the other coefficients of the matrix being zero. It follows that

$$\forall i \in \{1,\dots,j-1\},\ \left|(P^{(j)})_{ij}\right| \le \frac{\|s^{(j)}\|_2}{d_i^{(k)}} \le \frac{\|r^{(k-m_k+j)}\|_2 + \|r^{(k-m_k+j-1)}\|_2}{d_i^{(k)}} \le \frac{2}{t}.$$

As a consequence, there holds an estimate of the form

$$\|P^{(j)}\|_2 \le C(1+t^{-1})$$

for some constant C depending only on m_k, which thus yields

$$\|P\| = \|P^{(m_k)}P^{(m_k-1)}\dots P^{(2)}\| \le C^{m_k-1}(1+t^{-1})^{m_k-1}.$$

Hence, it follows that

$$\|\tilde{\zeta}\|_\infty \le C^{m_k-1}(1+t^{-1})^{m_k-1}\|\tilde{\lambda}\|_\infty \le C^{m_k-1}(1+t^{-1})^{m_k-1}t^{-1}.$$

Finally, one has $\tilde{c}_0 = -\tilde{\zeta}_1$, $\tilde{c}_i = \tilde{\zeta}_i - \tilde{\zeta}_{i+1}$, $1 \le i \le m_k-1$, and $\tilde{c}_{m_k} = 1 + \tilde{\zeta}_{m_k}$, so that $\|\tilde{c}\|_\infty \le C_{m_k}(1+t^{-m_k})$, for some positive constant C_{m_k} depending only on m_k, thus proving bound (32) on the coefficients.

Next, setting $\tilde{r}^{(k+1)} = f(\tilde{x}^{(k+1)})$, one has

$$\|\tilde{r}^{(k+1)} - \sum_{i=0}^{m_k} \tilde{c}_i r^{(k-m_k+i)}\|_2 \le \|f(\tilde{x}^{(k+1)}) - \mathrm{D}f(x_*)(\tilde{x}^{(k+1)} - x_*)\|_2 + \|\sum_{i=0}^{m_k} \tilde{c}_i f(x^{(k-m_k+i)}) - \mathrm{D}f(x_*)(\tilde{x}^{(k+1)} - x_*)\|_2$$

so that, using inequality (21) and the fact that the coefficients $(\tilde{c}_i)_{0\le i\le m_k}$ sum to 1,

$$\begin{aligned}\|\tilde{r}^{(k+1)} - \sum_{i=0}^{m_k} \tilde{c}_i r^{(k-m_k+i)}\|_2 &\le \frac{L}{2}\|\sum_{i=0}^{m_k} \tilde{c}_i (x^{(k-m_k+i)} - x_*)\|_2^{\,2} + \|\sum_{i=0}^{m_k} \tilde{c}_i (f(x^{(k-m_k+i)}) - \mathrm{D}f(x_*)(x^{(k-m_k+i)} - x_*))\|_2 \\ &\le \frac{L}{2}(m_k+1)\|\tilde{c}\|_\infty((m_k+1)\|\tilde{c}\|_\infty + 1) \max_{i\in\{0,\dots,m_k\}} \|x^{(k-m_k+i)} - x_*\|_2^{\,2}.\end{aligned}$$

It thus follows from the definition of the coefficients $\tilde{c}_i$ that

$$\begin{aligned}\|\tilde{r}^{(k+1)}\|_2 &\le \|\tilde{r}^{(k+1)} - \sum_{i=0}^{m_k} \tilde{c}_i\, r^{(k-m_k+i)}\|_2 + \mathrm{dist}_2(0, \mathcal{A}_{m_k}^{(k)}) \\ &\le \frac{L}{2}(m_k+1)\|\tilde{c}\|_\infty((m_k+1)\|\tilde{c}\|_\infty + 1) \max_{i\in\{0,\dots,m_k\}} \|x^{(k-m_k+i)} - x_*\|_2^{\,2} + \mathrm{dist}_2(0, \mathcal{A}_{m_k}^{(k)}),\end{aligned}$$

and, using Assumption 2 and bound (32),

$$\|\tilde{r}^{(k+1)}\|_2 \le \frac{L}{2\sigma^2}(m_k+1)C_{m_k}(1+t^{-m_k})((m_k+1)C_{m_k}(1+t^{-m_k})+1) \max_{i\in\{0,\dots,m_k\}} \|r^{(k-m_k+i)}\|_2^{\,2} + \mathrm{dist}_2(0, \mathcal{A}_{m_k}^{(k)}).$$

Let us now estimate the distance between $\tilde{x}^{(k+1)}$ and the submanifold Σ. One has

$$\|\tilde{x}^{(k+1)} - P_\Sigma(\tilde{x}^{(k+1)})\|_2 \leq \|\tilde{x}^{(k+1)} - x_* - P_{T_{x_*}\Sigma}(\tilde{x}^{(k+1)} - x_*)\|_2 + \|x_* + P_{T_{x_*}\Sigma}(\tilde{x}^{(k+1)} - x_*) - P_\Sigma(\tilde{x}^{(k+1)})\|_2,$$

and bounds for both terms in the right-hand side of this inequality are needed. For the first one, we find, since $x^{(k-m_k+i)} = P_\Sigma(x^{(k-m_k+i)})$ for $i = 0, \ldots, m_k$ and using (22) and (32), that

$$\begin{aligned}
\|\tilde{x}^{(k+1)} - x_* - P_{T_{x_*}\Sigma}(\tilde{x}^{(k+1)} - x_*)\|_2 &= \|\sum_{i=0}^{m_k} \tilde{c}_i \, (I - P_{T_{x_*}\Sigma})(x^{(k-m_k+i)} - x_*)\|_2 \\
&\leq \sum_{i=0}^{m_k} |\tilde{c}_i| \, \|(I - P_{T_{x_*}\Sigma})(x^{(k-m_k+i)} - x_*)\|_2 \\
&= \sum_{i=0}^{m_k} |\tilde{c}_i| \, \|P_\Sigma(x^{(k-m_k+i)}) - x_* - P_{T_{x_*}\Sigma}(x^{(k-m_k+i)} - x_*)\|_2 \\
&\leq \frac{M}{2} \sum_{i=0}^{m_k} |\tilde{c}_i| \, {\|x^{(k-m_k+i)} - x_*\|_2}^2 \\
&\leq \frac{M}{2}(m_k+1) C_{m_k} (1 + t^{-m_k}) \max_{i \in \{0,\ldots,m_k\}} {\|x^{(k-m_k+i)} - x_*\|_2}^2.
\end{aligned}$$

For the second term, we use again (22) and (32) to obtain

$$\begin{aligned}
\|x_* + P_{T_{x_*}\Sigma}(\tilde{x}^{(k+1)} - x_*) - P_\Sigma(\tilde{x}^{(k+1)})\|_2 &\leq \frac{M}{2} {\|\tilde{x}^{(k+1)} - x_*\|_2}^2 \\
&\leq \frac{M}{2}(m_k+1)^2 {C_{m_k}}^2 (1 + t^{-m_k})^2 \max_{i \in \{0,\ldots,m_k\}} {\|x^{(k-m_k+i)} - x_*\|_2}^2.
\end{aligned}$$

Adding these two estimates and using Assumption 2 give

$$\begin{aligned}
\|\tilde{x}^{(k+1)} - P_\Sigma(\tilde{x}^{(k+1)})\|_2 &\leq \frac{M}{2}(m_k+1)\|\tilde{c}\|_\infty((m_k+1)\|\tilde{c}\|_\infty + 1) \max_{i \in \{0,\ldots,m_k\}} {\|x^{(k-m_k+i)} - x_*\|_2}^2 \\
&\leq \frac{M}{2\sigma^2}(m_k+1) C_{m_k}(1 + t^{-m_k}) \left((m_k+1) C_{m_k}(1 + t^{-m_k}) + 1\right) \max_{i \in \{0,\ldots,m_k\}} {\|r^{(k-m_k+i)}\|_2}^2.
\end{aligned}$$

Using Assumption 1 and inequalities (20), combined with the last inequalities, then yields

$$\begin{aligned}
\|r^{(k+1)}\|_2 &\leq \|f(g(P_\Sigma(\tilde{x}^{(k+1)})))\|_2 + 2\,\|\mathrm{D}f(x_*) \circ \mathrm{D}g(x_*)\|_2 \|\tilde{x}^{(k+1)} - P_\Sigma(\tilde{x}^{(k+1)})\|_2 \\
&\leq K\,\|\tilde{r}^{(k+1)}\|_2 + 2\,(K\,\|\mathrm{D}f(x_*)\|_2 + \|\mathrm{D}f(x_*) \circ \mathrm{D}g(x_*)\|_2)\,\|\tilde{x}^{(k+1)} - P_\Sigma(\tilde{x}^{(m+1)})\|_2 \\
&\leq K\,\mathrm{dist}_2(0, \mathcal{A}_{m_k}^{(k)}) + \kappa(1 + t^{-2m_k}) \max_{i \in \{0,\ldots,m_k\}} {\|r^{(k-m_k+i)}\|_2}^2,
\end{aligned}$$

where $\kappa := \frac{1}{\sigma^2}\left(KL + M(K\,\|\mathrm{D}f(x_*)\|_2 + \|\mathrm{D}f(x_*) \circ \mathrm{D}g(x_*)\|_2)\right)(m_k+1)^2 {C_{m_k}}^2$.

Finally, by the triangle inequality, one has

$$\mathrm{dist}_2(0, \mathcal{A}_{m_k}^{(k)}) \leq \|r^{(k+1)}\|_2 + d_{m_k+1}^{(k)},$$

from which the last inequality follows. □

3.4.2 Proof of Theorem 3.1

For any k in $\mathbb{N}$ such that $\|r^{(k)}\|_2 > 0$, we consider the following properties

$$\text{if } m_k \geq 1 \text{ then, } \forall \ell \in \{1, \ldots, m_k\},\ d_\ell^{(k)} \geq \tau(1-\mu)\|r^{(k-m_k)}\|_2, \tag{a_k}$$

$$\text{if } m_k \geq 1 \text{ then, } \forall j \in \{1, \ldots, m_k\},\ \|r^{(k-m_k+j)}\|_2 \leq \mu^j \|r^{(k-m_k)}\|_2, \tag{b_k}$$

$$\|r^{(k-m_k)}\|_2 \leq C_\mu \tau^{2p}. \tag{c_k}$$

Assume that these properties hold for some natural integer k. Property (a_k) implies that the vectors of the set $\{r^{(k-m_k+i)}\}_{i=0,\ldots,m_k}$, are affinely independent in $\mathbb{R}^p$, which implies (23). It then follows from properties (a_k) and (b_k) that the sequence $\{x^{(k-m_k)}, \ldots, x^{(k)}\}$ satisfies condition (31) of Lemma 3.7 with $t = \tau(1-\mu)$, since

$$\max_{0 \leq i \leq m_k} \|r^{(k-m_k+i)}\|_2 = \|r^{(k-m_k)}\|_2.$$

Bound (24) is then simply bound (32) in Lemma 3.7 with $t = \tau(1-\mu)$.

If the constant C_μ is small enough, bound (24) and properties (b_k) and (c_k) imply that the point $\tilde{x}^{(k+1)}$ belongs to $V \cap g^{-1}(V)$ so that estimate (33) in Lemma 3.7 holds with $t = \tau(1-\mu)$, and, using property (b_k), one gets

$$\begin{aligned}\|r^{(k+1)}\|_2 &\leq K \operatorname{dist}_2(0, \mathcal{A}_k^{(k)}) + \kappa \left(1 + \frac{1}{(\tau(1-\mu))^{2m_k}}\right) {\|r^{(k-m_k)}\|_2}^2 \\ &\leq \left(K\mu^{m_k} + \kappa \left(1 + \frac{1}{(\tau(1-\mu))^{2p}}\right) \|r^{(k-m_k)}\|_2\right) \|r^{(k-m_k)}\|_2.\end{aligned}$$

To establish (25), we thus need that

$$\frac{K}{\mu} + \kappa \left(1 + \frac{1}{(\tau(1-\mu))^{2p}}\right) \frac{C_\mu \tau^{2p}}{\mu^{m_k+1}} \leq 1,$$

and it suffices to choose the constant C_μ such that

$$\frac{K}{\mu} + \kappa \left(1 + \frac{1}{(1-\mu)^{2p}}\right) \frac{C_\mu}{\mu^{p+1}} \leq 1.$$

Let us now show that properties (a_k), (b_k), and (c_k) hold for all natural integers and C_μ small enough, and that they imply the statements of the Theorem.

For $k = 0$, one has $m_k = 0$. Since we assumed that $\|r^{(0)}\|_2 \leq C_\mu \tau^{2p}$, property (c_k) holds for $k = 0$. Next, assume that (a_k), (b_k) and (c_k) hold for a given natural integer k. If $m_k = 0$ (meaning there is a restart at step k), then $k + 1 - m_{k+1} = k + 1$ and, since we have previously shown that (25) holds, it follows from property (c_k) that

$$\|r^{(k+1-m_{k+1})}\|_2 \leq \mu^{m_k+1} \|r^{(k-m_k)}\|_2 \leq \mu^{m_k+1} C_\mu \tau^{2p} \leq C_\mu \tau^{2p},$$

so that property (c_{k+1}) holds. Otherwise $m_{k+1} = m_k + 1$ (meaning there is no restart at step k) so that $k+1-m_{k+1} = k - m_k$. Thus, it follows that $\|r^{(k+1-m_{k+1})}\|_2 = \|r^{(k-m_k)}\|_2 \leq C_\mu \tau^{2p}$ from property (c_k), which then also holds for $k + 1$. Since there is no restart, the following condition is verified

$$\tau \|r^{(k+1)} - r^{(k-m_k)}\|_2 \leq \|(Id - \Pi_k)(r^{(k+1)} - r^{(k-m_k)})\|_2 = d_{m_k+1}^{(k)}$$

We then have

$$\begin{aligned}d_{m_k+1}^{(k)} &\geq \tau \left(\|r^{(k-m_k)}\|_2 - \|r^{(k+1)}\|_2\right) \\ &\geq \tau \left(\|r^{(k-m_k)}\|_2 - \mu^{m_k+1} \|r^{(k-m_k)}\|_2\right) \\ &\geq \tau(1-\mu) \|r^{(k-m_k)}\|_2.\end{aligned}$$

Since $m_{k+1} = m_k + 1$, we also have that $d_i^{(k+1)} = d_i^{(k)}$ for any i in $\{1, \ldots, m_k + 1\}$ so that property (a_{k+1}) follows from property (a_k) and the above inequality.

Finally, since $k + 1 - m_{k+1} = k - m_k$, property (b_k) is equivalent to

$$\forall j \in \{0, \ldots, m_{k+1} - 1\},\ \|r^{(k+1-m_{k+1}+j)}\|_2 \leq \mu^j \|r^{(k+1-m_{k+1})}\|_2,$$

as bound (25) is equivalent to

$$\|r^{(k+1-m_{k+1}+m_{k+1})}\|_2 \leq \mu^{m_{k+1}} \|r^{(k+1-m_{k+1})}\|_2,$$

so that property (b_{k+1}) holds.

Finally, if a restart occurs at step $k + 1$, one has, employing the same notations as in Lemma 3.7,

$$d_{m_k+1}^{(k)} < \tau \|r^{(k+1)} - r^{(k-m_k)}\|.$$

Inequality (26) then follows from estimate (34) in Lemma 3.7 and the triangle inequality by setting $C = \kappa \left(1 + (1-\mu)^{-2p}\right)$.

3.4.3 Proof of Theorem 3.2

For any natural integer k for which $\|r^{(k)}\|_2 > 0$, we consider the properties:

$$\text{if } m_k \geq 1 \text{ then, } \forall \ell \in \{1, \ldots, m_k\},\ d_\ell^{(k)} \geq \frac{1-\mu}{\mu} \|r^{(k-m_k+\ell)}\|_2, \tag{a_k}$$

$$\forall 0 \leq i < j \leq m_k, \quad \delta \|r^{(k-m_k+i)}\|_2 \leq \|r^{(k-m_k+j)}\|_2 \leq \mu^{j-i} \|r^{(k-m_k+i)}\|_2, \tag{b_k}$$

$$\|r^{(k-m_k)}\|_2 \leq c_\mu \delta^2. \tag{c_k}$$

Under the assumptions of Theorem 3.2, we will prove by induction that these properties hold for all k if c_μ is sufficiently small and that they imply the statements of the Theorem. We denote by $(\mathcal{P}_k)$ the set of properties (a_k), (b_k), and (c_k).

First, since $m_0 = 0$ and we assumed that $\|r^{(0)}\| \leq c_\mu \delta^2$, $(\mathcal{P}_0)$ holds.

Next, assume that $(\mathcal{P}_k)$ holds for some natural integer k. First, the fact that property (a_k) is verified implies that the vectors $r^{(k-m_k)}, \ldots, r^{(k)}$ are affinely independent in $\mathbb{R}^p$, so (27) necessarily holds. It also follows from properties (a_k) and (b_k) that the family $\{r^{(k-m_k)}, \ldots, r^{(k)}\}$ satisfies the assumption (31) of Lemma 3.7 for $t = \frac{1-\mu}{\mu}$, which establishes estimate (28) as (32) in Lemma 3.7.

For c_μ is small enough, estimate (28) and properties (b_k) and (c_k) imply that $\tilde{x}^{(k+1)}$ belongs to $V \cap g^{-1}(V)$, so that estimate (33) of Lemma 3.7 holds. Hence, one gets

$$\begin{aligned}
\|r^{(k+1)}\|_2 &\leq K \operatorname{dist}_2(0, \mathcal{A}_k^{(k)}) + \kappa(1 + t^{-2p}) \max_{i \in \{0,\ldots,m_k\}} \|r^{(k-m_k+i)}\|_2^2 \\
&\leq K \|r^{(k)}\|_2 + \kappa(1 + t^{-2p}) c_\mu \delta \|r^{(k)}\|_2.
\end{aligned}$$

Choosing c_μ in such a way that $K + \kappa(1 + t^{-2p}) c_\mu K \leq \mu$, we obtain (29). Assuming that $m_{k+1} = 0$, property (c_{k+1}) holds and so does $(\mathcal{P}_{k+1})$ in this case. Assuming that $m_{k+1} \geq 1$, we have by definition of the algorithm that $k - m_k \leq k + 1 - m_{k+1}$ and $\|r^{(k+1)}\|_2 \geq \delta \|r^{(k+1-m_{k+1}+i)}\|$, for any integer i in $\{0, \ldots, m_{k+1}\}$. Using property (b_k) and estimate (29), property (b_{k+1}) ensues. Next, if $k + 1 - m_{k+1}$ is equal to $k - m_k$, then property (c_{k+1}) amounts to (c_k). Otherwise, one has $k + 1 - m_{k+1} > k - m_k$, and property (c_{k+1}) follows from both (b_k) and (c_k).

Finally, estimate (34) of Lemma 3.7 gives

$$\|r^{(k+1)}\|_2 \leq \frac{K}{1-K} d_{m_k+1}^{(k)} + \frac{\kappa}{1-K} \left(1 + \left(\frac{\mu}{1-\mu}\right)^{2p}\right) c_\mu \|r^{(k+1)}\|_2, \tag{35}$$

as, for any integer i in $\{0, \ldots, m_k\}$, one has both $\|r^{(k-m_k+i)}\|_2 \leq c_\mu \delta^2$ and $\|r^{(k-m_k+i)}\|_2 \leq \frac{1}{\delta^2} \|r^{(k+1)}\|_2$, since $k \geq k + 1 - m_{k+1}$. Choosing $c_\mu > 0$ such that

$$\frac{K}{1-K} \frac{1-\mu}{\mu} + \frac{\kappa}{1-K} \left(1 + \left(\frac{\mu}{1-\mu}\right)^{2p}\right) c_\mu < 1$$

inequality (35) implies that $\|r^{(k+1)}\|_2$ is non-positive if $d_{m_k+1}^{(k)} < \frac{1-\mu}{\mu} \|r^{(k+1)}\|_2$, which is absurd. We have thus proved by contradiction that

$$d_{m_k+1}^{(k)} \geq \frac{1-\mu}{\mu} \|r^{(k+1)}\|_2.$$

Using property (a_k), we may introduce the new property:

$$\forall \ell \in \{1, \ldots, m_k + 1\}, \; d_\ell^{(k)} \geq \frac{1-\mu}{\mu} \|r^{(k-m_k+\ell)}\|_2. \tag{$\hat{a}_k$}$$

Now, it is easily seen that $\mathcal{A}_\ell^{(k+1)} \subset \mathcal{A}_{\ell+1+m_k-m_{k+1}}^{(k)}$ for any integer ℓ in $\{0, \ldots, m_{k+1} - 1\}$, so that one has $d_\ell^{(k+1)} \geq d_{\ell+1+m_k-m_{k+1}}^{(k)}$ for any integer ℓ in $\{1, \ldots, m_{k+1}\}$. Property $(\hat{a}_k)$ implying that

$$\forall \ell \in \{1, \ldots, m_{k+1}\}, \; d_{\ell+1+m_k-m_{k+1}}^{(k)} \geq \frac{1-\mu}{\mu} \|r^{(k+1-m_{k+1}+\ell)}\|_2,$$

property (a_{k+1}) holds and $(\mathcal{P}_{k+1})$ too.

We have thus proved that $(\mathcal{P}_k)$ holds for for any natural integer k if c_μ is small enough, and so do statements (27), (28), and (29) of the Theorem. It remains to prove statement (30).

Let k be an integer such that $k - p \geq 1$, with $\|r^{(k)}\|_2 > 0$. Since $m_k \leq p$, one has $k - p - 1 \leq k - m_k - 1$, so that $\|r^{(k)}\|_2 \leq \delta \|r^{(k-m_k-1)}\|_2 \leq \delta \|r^{(k-p-1)}\|_2$. This ends the proof.

4 Application to electronic ground state calculations

We now show how the Anderson–Pulay acceleration analysed in the previous section can be applied to the computation of the electronic ground state of a molecular system. In order to do so, we first recall the quantum many-body problem and present two of its most commonly used approximations in non-relativistic quantum chemistry: the *ab initio* Hartree–Fock method [29, 22], which aims at approximating the ground state electronic wavefunction, and the semi-empirical approach of Kohn and Sham [35], issued from the density functional theory.

4.1 The minimisation problem

Consider an isolated molecular system composed of M atomic nuclei and N electrons, whose state is entirely described by a wavefunction ψ valued in $\mathbb{C}$. Within the setting of the Born–Oppenheimer approximation, the motion of atomic nuclei and electrons can be separated and the nuclei are classical point-like particles with fixed positions $\overline{x_1}, \ldots, \overline{x_M}$. The state of the electrons is then represented a wavefunction ψ only depending on the time variable t and on the respective positions $x_1, \ldots, x_N$ in $\mathbb{R}^3$ and spins[8] $\sigma_1, \ldots, \sigma_N$ of the electrons.

Assuming stationariness, the minimisation problem to be solved in order to determine the ground state of the molecular system is

$$E(v) = \inf \left\{ \langle \psi, H_v \psi \rangle \mid \psi \in \mathscr{H}, \int_{\mathbb{R}^{3N}} \sum_{\sigma_1, \ldots, \sigma_N} |\psi|^2 (x_1, \ldots, x_N, \sigma_1, \ldots, \sigma_N) \, \mathrm{d}x_1 \ldots \mathrm{d}x_N = 1 \right\}, \tag{36}$$

where, having adopted the atomic unit system[9], the N-body electronic Hamiltonian H_v, derived from the stationary Schrödinger equation, is given by

$$H_v = K + \sum_{1 \leq i < j \leq N} \frac{1}{\|x_i - x_j\|} + V,$$

$K = -\frac{1}{2} \sum_{i=1}^N \Delta_{x_i}$ being the kinetic energy operator and the external potential V being corresponding to the Coulomb potential generated by the nuclei of the atoms forming the system, that is

$$V = \sum_{i=1}^N v(x_i),$$

with $v(x) = -\sum_{k=1}^M \frac{Z_k}{\|x - \overline{x_k}\|}$, Z_k being the charge the kth nucleus, where, due to both physical and mathematical considerations, the functional space $\mathscr{H}$ is

$$\mathscr{H} = \bigwedge_{i=1}^N H^1 \left(\mathbb{R}^3 \times \{|+\rangle, |-\rangle\}, \mathbb{C} \right),$$

the symbol $\bigwedge$ denoting the usual tensorial product with the additional assumption that only antisymmetrised products are considered, as required by the Pauli exclusion principle, and where the normalisation condition stemming from the fact that the quantity $|\psi(x_1, \ldots, x_N, \sigma_1, \ldots, \sigma_N)|^2$ represents the density of probability to observe the ith electron at position x_i with spin σ_i, for any integer i in $\{1, \ldots, N\}$.

Variational problem (36) is well-posed and one may envision to attack it "directly" by replacing it by a finite dimensional approximate problem to be solved numerically. Unfortunately, the intrinsically high computational complexity of this methodology renders the approach intractable for systems with more than a few electrons and one has to resort to other types of approximation to the minimisation problem, two of which will be recalled in the next subsections.

4.2 The Hartree–Fock method

The Hartree–Fock method [29, 22] for the computation of the ground-state electronic wavefunction consists in restricting the variational space in problem (36) to the set of so-called *Slater determinants*, which are antisymmetrised products of N *monoelectronic wavefunctions* (also called *molecular (or atomic) orbitals*). While this restriction only provides with an upper bound of the exact energy, the main advantage of this approximation is that the problem to be solved remains variational. There is however a price to pay in a loss of the correlation between the positions of the electrons, as an electron will evolve independently of the way the others do in this model.

For the sake of simplicity, we shall present the Hartree–Fock method by omitting the spin variables, keeping in mind that different versions of the method, which explicitly account for the spin dependence, exist[10] and are

[8] For an electron, the spin can only have two values, which are denoted by $|+\rangle$ for the spin-up orientation and $|-\rangle$ for the spin-down one.

[9] By definition, the Hartree atomic units form a system of units in which the numerical values of the electron mass m_e, the elementary charge e, the reduced Planck constant $\hbar$ and the Coulomb constant $\frac{1}{4\pi\varepsilon_0}$, ε_0 being the dielectric permittivity of the vacuum, are all unity.

[10] The general Hartree–Fock (GHF) model considers antisymmetrised products of functions in $H^1\left(\mathbb{R}^3 \times \{|+\rangle, |-\rangle\}, \mathbb{C}\right)$, but it is seldom used in practice. More popular models are derived by further restricting the space of admissible wavefunctions by prescribing some specific dependence of the one-electron wavefunctions with respect to the spin variables. In the restricted Hartree–Fock (RHF) model, which stems from the concept of Lewis electron pairs and can only be used on systems with an even number N of electrons (which are called closed-shell systems), it is assumed that each orbital is either doubly occupied or unoccupied, so that a single function of $H^1(\mathbb{R}^3, \mathbb{C})$ is associated with each of the $\frac{N}{2}$ electron pairs. Two electrons of the same pair thus share the same one-electron wavefunction, one being

used in the numerical practice. In this context, the spinless Hartree–Fock wavefunction is such that there exists $\Phi = \{\phi_i\}_{1\leq i\leq N}$ in $\mathcal{W}_N$ such that

$$\psi(x_1,\ldots,x_N) = \frac{1}{\sqrt{N!}}\begin{vmatrix}\phi_1(x_1) & \ldots & \phi_1(x_N)\\ \vdots & & \vdots\\ \phi_N(x_1) & \ldots & \phi_N(x_N)\end{vmatrix}$$

where the functions ϕ_i, $1\leq i\leq N$, are the monoelectronic wavefunctions and

$$\mathcal{W}_N = \left\{\Phi = \{\phi_i\}_{i\in\{1,\ldots,N\}} \,|\, \forall (i,j)\in\{1,\ldots,N\}^2,\ \phi_i\in H^1(\mathbb{R}^3,\mathbb{C}),\ (\phi_i,\phi_j)_{L^2(\mathbb{R}^3,\mathbb{C})} = \delta_{ij}\right\}.$$

After some computations, one is led to the minimisation problem

$$\inf\left\{\mathcal{E}^{\mathrm{HF}}(\Phi)\,|\,\Phi\in\mathcal{W}_N\right\}, \tag{37}$$

where the spinless Hartree–Fock energy functional may be written in the following compact form

$$\mathcal{E}^{\mathrm{HF}}(\Phi) = \frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3}\|\boldsymbol{\nabla}\phi_i(x)\|^2\,\mathrm{d}x + \int_{\mathbb{R}^3} v(x)\rho_\Phi(x)\,\mathrm{d}x + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho_\Phi(x)\rho_\Phi(y)}{\|x-y\|}\,\mathrm{d}x\,\mathrm{d}y - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{|\tau_\Phi(x,y)|^2}{\|x-y\|}\,\mathrm{d}x\,\mathrm{d}y,$$

by introducing the so-called *density matrix* τ_Φ, $\tau_\Phi(x,y) = \sum_{i=1}^N \phi_i(x)\phi_i(y)$, and the corresponding *electronic density* ρ_Φ, $\rho_\Phi(x) = \tau_\Phi(x,x)$, associated with the Slater determinant with Φ.

Problem (37) has been extensively studied. Notably, existence of at least one minimiser holds under the condition that the total nuclear charge $Z = \sum_{k=1}^M Z_k$ of the molecular system satisfies $N < Z+1$ (see [42]). It may be easily seen that the above energy functional is invariant under unitary transforms of the N-tuple Φ, so that any minimiser of the problem is defined up to an unitary matrix. This invariance may be used to diagonalise the Hermitian matrix of Lagrange multipliers in the Euler–Lagrange equations satisfied by the N-tuple Φ associated with a minimiser. This system of N coupled partial differential equations can then be written under the compact form known as the Hartree–Fock equations

$$\begin{cases} F_\Phi\phi_i = \lambda_i\,\phi_i,\\ \displaystyle\int_{\mathbb{R}^3}\overline{\phi_i(x)}\phi_j(x)\,\mathrm{d}x = \delta_{ij}, \end{cases} \tag{38}$$

where F_Φ is the so-called Fock operator,

$$F_\Phi\psi = -\frac{1}{2}\,\Delta\psi + V\psi + \left(\sum_{j=1}^N |\phi_j|^2 * \frac{1}{\|\cdot\|}\right)\psi - \sum_{j=1}^N\left(\psi\phi_j * \frac{1}{\|\cdot\|}\right)\phi_j,$$

the N real scalars λ_i, associated with the orthonormality constraints, are called the one-electron energies of the occupied orbitals ϕ_i, $1\leq i\leq N$, and δ_{ij} is the Kronecker symbol. Any minimiser of the Hartree–Fock problem (37) is, up to an unitary transform, a solution of the nonlinear eigenproblem (38) and it has been proved that the scalars λ_i associated with a Hartree–Fock ground state wavefunction are the N lowest eigenvalues of the Fock operator (see [43]).

4.3 The Kohn–Sham model of density functional theory

The density functional theory (DFT) of Hohenberg and Kohn [32] follows a completely different approach to the original minimisation problem and aims at including the electronic correlation missing from the Hartree–Fock method. It is based on the fact that, for any admissible wavefunction in the space $\mathscr{H}$, one can define a corresponding electronic density

$$\rho(x) = N\int_{\mathbb{R}^{3(N-1)}}|\psi|^2\,(x,x_2,\ldots,x_N)\,\mathrm{d}x_2\ldots\mathrm{d}x_N, \tag{39}$$

which belongs to the space $\mathscr{I}_N = \left\{\rho\,|\,\rho\geq 0,\ \sqrt{\rho}\in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\rho(x)\,\mathrm{d}x = N\right\}$ (note that the spin variables are again omitted hereafter). Conversely, it can be showed that, for any density in $\mathscr{I}_N$, there exists a wavefunction ψ of finite

of spin up, the other one of spin down. In the unrestricted Hartree–Fock (UHF) model, wavefunctions are built with orbitals that are either with spin up or spin down, which makes it appropriate to deal with open shell molecular systems, that is systems with an odd number of electrons or with an even number of electrons but whose ground state is not a spin singlet state. Finally, in the restricted open-shell Hartree–Fock (ROHF) model, the lowest energy orbitals are doubly occupied whereas the upper energy ones are populated with an unpaired electron.

kinetic energy satisfying (39). Following arguments by Hohenberg and Kohn, the energy functional defined in terms of the unknown wavefunction ψ can be replaced by one for the unknown density ρ, leading to the problem

$$E(v) = \inf \left\{ F_{\mathrm{LL}}(\rho) + \int_{\mathbb{R}^3} v(x)\rho(x)\,\mathrm{d}x \,|\, \rho \in \mathscr{I}_N \right\},$$

where the universal (in the sense that it does not depend on the molecular system under consideration and is thus independent of the nuclear potential v) functional F_{LL} is the Levy–Lieb density functional [40, 41],

$$F_{\mathrm{LL}}(\rho) = \inf \left\{ \langle \psi, H_0 \psi \rangle \,|\, \psi \in \mathscr{H},\ \|\psi\|_{L^2(\mathbb{R}^{3N})} = 1,\ \psi \text{ has a density } \rho \right\}$$

with $H_0 = K + \sum_{1 \le i < j \le N} \frac{1}{\|x_i - x_j\|}$ the interacting Hamiltonian of the system.

The Levy–Lieb functional not being known explicitly for a system of N interacting electrons, suitable approximations are needed to make the DFT a practical tool for computing electronic ground states, which rely on exact (or very accurate) evaluations of the density functional of a reference system "close" to the real one. In the Kohn–Sham models [35], which are by far the most commonly used[11], the chosen reference system is a system of N *non*-interacting electrons. To introduce the analogue of the Levy–Lieb functional, the standard Kohn–Sham model defines the Kohn–Sham kinetic energy functional

$$\begin{aligned} T_{\mathrm{KS}}(\rho) &= \inf \left\{ \langle \psi, K\psi \rangle \,|\, \psi \in \mathscr{H},\ \|\psi\|_{L^2(\mathbb{R}^{3N})} = 1,\ \psi \text{ is a Slater determinant and has a density } \rho \right\} \\ &= \inf \left\{ \frac{1}{2} \sum_{i=1}^{N} \int_{\mathbb{R}^3} \|\boldsymbol{\nabla} \phi_i(x)\|^2 \,\mathrm{d}x \,|\, \Phi \in \mathcal{W}_N,\ \rho = \rho_\Phi \right\}, \end{aligned}$$

assuming that the kinetic energy admits a Slater determinant as a minimizer, and the Coulomb energy functional

$$J(\rho) = \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{\|x - y\|} \,\mathrm{d}x\mathrm{d}y,$$

representing the electrostatic energy of a classical charge distribution of density ρ, as a guess for the electronic interaction energy in a system of N electrons with density ρ. The errors in both the kinetic and electronic interaction energies are then put together in the exchange-correlation functional, defined as the difference

$$E_{\mathrm{xc}}(\rho) = F_{\mathrm{LL}}(\rho) - T_{\mathrm{KS}}(\rho) - J(\rho),$$

leading to the functional of the Kohn–Sham model,

$$E^{\mathrm{KS}}(V) = \inf \left\{ \frac{1}{2} \sum_{i=1}^{N} \int_{\mathbb{R}^3} \|\boldsymbol{\nabla} \phi_i(x)\|^2 \,\mathrm{d}x + \int_{\mathbb{R}^3} V \rho_\Phi(x)\,\mathrm{d}x + J(\rho_\Phi) + E_{\mathrm{xc}}(\rho_\Phi) \,|\, \Phi \in \mathcal{W}_N \right\}.$$

To render the numerical simulations of the above model possible, the construction of approximations to the exchange-correlation functional, like the local-density approximation (LDA) or the generalised gradient approximation (GGA), is necessary. Requiring some differentiability properties of the approximate exchange-correlation functional (which is tied to the difficult theoretical question of v-representability), the Euler–Lagrange equations associated with this last minimisation problem may be derived and written under the form of the Kohn–Sham equations

$$\begin{cases} -\frac{1}{2}\Delta\phi_i + V\phi_i + \left(\rho * \frac{1}{\|\cdot\|}\right)\phi_i + v_{\mathrm{xc}}(\rho)\phi_i = \lambda_i\,\phi_i, \\ \int_{\mathbb{R}^3} \overline{\phi_i(x)}\phi_j(x)\,\mathrm{d}x = \delta_{ij}, \end{cases}$$

where the scalars λ_i are the Lagrange multipliers associated to the orthogonality constraints (the matrix of Lagrange multipliers has been diagonalised as it was done in the Hartree–Fock setting). As the functional E_{xc} is considered as defined on $L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3)$, its derivative v_{xc} is a vector of the dual space $L^{3/2}(\mathbb{R}^3) + L^\infty(\mathbb{R}^3)$. Whereas this system of equations is formally the same has the Hartree–Fock one, observe that the functions ϕ_i which solve it cannot be interpreted as the monoelectronic wavefunctions of the ground state, as they are related to a non-interacting system having the same density as the interacting system considered (assuming such a non-interacting system exists). Moreover, under the differentiability assumption stated above and satisfied by most of the forms postulated in practice, the effective potential operator in the Kohn–Sham equations takes a local form (it is a multiplicative operator, which is not the case for the Fock operator).

As in the Hartree–Fock setting, it is necessary to resort to open-shell Kohn–Sham models when the number of electrons in the system is odd or when spin-dependant approximated exchange-correlation functional are used.

[11] The Thomas–Fermi and related models form an older class of models, in which the reference system is a homogeneous electron gas.

4.4 Galerkin approximation

In practice, the problems appearing in both the Hartree–Fock and Kohn–Sham models are approximated using a Galerkin method. Choosing to focus on the spinless Hartree–Fock framework, the monoelectronic wavefunctions are linearly expanded on a given *finite* basis set (this is the so-called *LCAO approximation*, LCAO being the acronym to *linear combination of atomic orbitals*). Due to computational complexity considerations, the basis functions are typically Slater-type or Gaussian-type orbitals, and generally form a non-orthonormal set, which we denote by $\{\chi_\upsilon\}_{1\leq\upsilon\leq d}$, which spans a vector space of finite dimension d. Writing

$$\forall j \in \{1,\dots,N\},\ \phi_j = \sum_{\upsilon=1}^{d} c_{\upsilon j}\chi_\upsilon,$$

one may introduce the rectangular matrix C of $\mathcal{M}_{d,N}(\mathbb{R})$ whose jth column contains the coefficients $c_{\upsilon j}$, $\upsilon = 1,\dots,d$, of the wavefunction ϕ_j in the finite basis, and the so-called *overlap* matrix S, which is the Gram matrix associated with the basis set, that is the Hermitian matrix S of size $d\times d$ whose coefficients are

$$\forall(\upsilon,\nu)\in\{1,\dots,d\}^2,\ s_{\upsilon\nu} = \int_{\mathbb{R}^3} \overline{\chi_\upsilon(x)}\chi_\nu(x)\,\mathrm{d}x,$$

so that the orthonormality constraints on the wavefunctions can be written as

$$\delta_{ij} = \int_{\mathbb{R}^3} \overline{\phi_i(x)}\phi_j(x)\,\mathrm{d}x = \int_{\mathbb{R}^3} \overline{\left(\sum_{\upsilon=1}^{d} c_{\upsilon i}\chi_\upsilon(x)\right)}\left(\sum_{\nu=1}^{d} c_{\nu j}\chi_\nu(x)\right)\mathrm{d}x = \sum_{\upsilon,\nu=1}^{d} \overline{c_{\upsilon i}} s_{\upsilon\nu} c_{\nu j},$$

and read in matrix form

$$C^*SC = I_N, \tag{40}$$

where I_N denotes the identity matrix of order N. Rewriting the energy functional $\mathcal{E}^{\mathrm{HF}}$ of the problem accordingly, one has

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3}\|\boldsymbol{\nabla}\phi_i(x)\|^2\,\mathrm{d}x + \int_{\mathbb{R}^3} V(x)\rho_{\boldsymbol{\phi}}(x)\,\mathrm{d}x &= \sum_{i=1}^{N}\left(\frac{1}{2}\int_{\mathbb{R}^3}\left|\sum_{\upsilon=1}^{d} c_{\upsilon i}\boldsymbol{\nabla}\chi_\upsilon(x)\right|^2\mathrm{d}x + \int_{\mathbb{R}^3} V(x)\left|\sum_{\upsilon=1}^{d} c_{\upsilon i}\boldsymbol{\nabla}\chi_\upsilon(x)\right|^2\mathrm{d}x\right)\\ &= \sum_{i=1}^{N}\sum_{\upsilon,\nu=1}^{d}\left(\int_{\mathbb{R}^3}\left(\frac{1}{2}\overline{\boldsymbol{\nabla}\chi_\upsilon(x)}\cdot\boldsymbol{\nabla}\chi_\nu(x) + V(x)\overline{\chi_\upsilon(x)}\chi_\nu(x)\right)\mathrm{d}x\right)\overline{c_{\upsilon i}}c_{\nu i} = \mathrm{tr}\left(H_{\mathrm{core}}CC^*\right),\end{aligned}$$

the matrix H_{core} of $\mathcal{M}_{d,d}(\mathbb{R})$ being the matrix of the *core Hamiltonian* $-\frac{1}{2}\Delta + V$ in the discrete basis set,

$$\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho_{\boldsymbol{\phi}}(x)\rho_{\boldsymbol{\phi}}(y)}{\|x-y\|}\,\mathrm{d}x\,\mathrm{d}y = \sum_{i,j=1}^{N}\sum_{\upsilon,\nu,\kappa,\lambda=1}^{d}\left(\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\overline{\chi_\upsilon(x)}\chi_\nu(x)\overline{\chi_\kappa(y)}\chi_\lambda(y)}{\|x-y\|}\,\mathrm{d}x\,\mathrm{d}y\right)\overline{c_{\upsilon i}}c_{\nu i}\overline{c_{\kappa j}}c_{\lambda j} = \mathrm{tr}\left(J(CC^*)CC^*\right)$$

for the Coulomb interaction contribution, and

$$\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{|\tau_{\boldsymbol{\phi}}(x,y)|^2}{\|x-y\|}\,\mathrm{d}x\,\mathrm{d}y = \sum_{i,j=1}^{N}\sum_{\upsilon,\nu,\kappa,\lambda=1}^{d}\left(\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\overline{\chi_\upsilon(x)}\chi_\nu(y)\overline{\chi_\kappa(y)}\chi_\lambda(x)}{\|x-y\|}\,\mathrm{d}x\,\mathrm{d}y\right)\overline{c_{\upsilon i}}c_{\nu i}\overline{c_{\kappa j}}c_{\lambda j} = \mathrm{tr}\left(K(CC^*)CC^*\right)$$

for the exchange interaction contribution. Defining the *discrete density matrix*

$$D = CC^*,$$

which is the (finite-dimensional) representation of the density matrix $\tau_{\boldsymbol{\phi}}$ in the basis $\{\chi_\upsilon\}_{1\leq\mu\leq d}$, and observing that it is Hermitian and, by virtue of (40), such that

$$DSD = D \text{ and } \mathrm{tr}(SD) = N,$$

the discretized problem to be solved reads

$$\inf\left\{E^{\mathrm{HF}}(D),\ D\in\mathcal{P}_N\right\},$$

where the discretised Hartree–Fock energy is given by

$$E^{\mathrm{HF}}(D) = \mathrm{tr}(H_{\mathrm{core}}D) + \frac{1}{2}\mathrm{tr}\left(J(D)D\right) - \frac{1}{2}\mathrm{tr}\left(K(D)D\right),$$

and

$$\mathcal{P}_N = \{D \in \mathcal{M}_{d,d}(\mathbb{C}),\ D^* = D,\ DSD = D,\ \mathrm{tr}(SD) = N\}\,. \tag{41}$$

The Euler–Lagrange equations associated with the last problem, also known as the *Roothaan–Hall equations* [51, 28], are then

$$\begin{cases} F^{\mathrm{HF}}(D)C = SC\Lambda \\ C^*SC = I_N \\ D = CC^* \end{cases} \tag{42}$$

where, with a slight abuse of notation, the matrix $F^{\mathrm{HF}}(D) = H_{\mathrm{core}} + J(D) - K(D)$ denotes the *Fock matrix* associated with the matrix C of orbital coefficients, and the Hermitian matrix Λ, which is by convention chosen diagonal[12], stands for a Lagrange multiplier attached to constraint (40) on the orbitals. A necessary condition for a density matrix D_* to be a "solution" of the above system is that it commutes with the associated Fock matrix $F^{\mathrm{HF}}(D_*)$ in the sense that

$$[F^{\mathrm{HF}}(D_*), D_*] = 0,$$

where the *"commutator"* $[A, B]$ between two matrices A and B is defined by $[A, B] = ABS - SBA$.

The discretisation of other Hartree–Fock models or of the Kohn–Sham models can be dealt with in the same manner. For the spinless Kohn-Sham model derived previously, the Roothaan–Hall equations read

$$\begin{cases} F^{\mathrm{KS}}(D)C = SC\Lambda \\ C^*SC = I_N \\ D = CC^* \end{cases} \tag{43}$$

with $F^{\mathrm{KS}}(D) = H_{\mathrm{core}} + J(D) + F_{\mathrm{xc}}(D)$ and $F_{\mathrm{xc}}(D) = \frac{1}{2}\nabla E_{\mathrm{xc}}(D)$, when assuming differentiability for the approximated exchange-correlation functional E_{xc} used in practice.

4.5 Self-consistent field algorithms

While based on different physical principles, problems (42) and (43) have the similar form of a nonlinear generalised eigenvalue problem. They are usually solved *"self-consistently"*, that is using an iterative fixed-point procedure, the most simple and "natural" approach being the algorithm introduced by Roothaan [51]. Given the choice of an initial density matrix $D^{(0)}$, it consists in generating a sequence of matrices $(D^{(k)})_{k\in\mathbb{N}}$ defined by

$$\forall k \in \mathbb{N},\ \begin{cases} F(D^{(k)})C^{(k+1)} = SC^{(k+1)}E^{(k+1)} \\ (C^{(k+1)})^*SC^{(k+1)} = I_N \\ D^{(k+1)} = C^{(k+1)}(C^{(k+1)})^* \end{cases}$$

where $E^{(k+1)}$ is a diagonal matrix such that $(E^{(k+1)})_{ii} = \varepsilon_i^{(k+1)}$, $i = 1, \cdots, N$, the scalars $\varepsilon_1^{(k+1)} \leq \varepsilon_2^{(k+1)} \cdots \leq \varepsilon_N^{(k+1)}$ being the N smallest eigenvalues, counted with multiplicity, of the *linear* generalised eigenproblem

$$F(D^{(k)})V = \varepsilon\, SV,$$

and the columns of the matrix $C^{(k+1)}$ are associated orthonormal (with respect to the scalar product induced by the matrix S) eigenvectors. The procedure of assembling $D^{(k+1)}$ by populating the molecular orbitals starting with those of lowest energy of the current Fock matrix is called the *Aufbau principle*[13].

Assuming the uniform well-posedness property introduced in [11], the matrix $D^{(k+1)}$ is uniquely defined at each step of the procedure and can be characterised as the minimiser of a variational problem, that is

$$D^{(k+1)} = \arg\inf \left\{ \mathrm{tr}\left(F(D^{(k)})D\right),\ D \in \mathcal{P}_N \right\} = g(D^{(k)}),$$

thus defining a fixed-point iteration process. In practice, a convergence criterion is used to end the iterations. For instance, one may compute the norm of the difference of two successive density matrices at each step and compare it to a prescribed tolerance. Another possibility is to use the norm of the commutator between the current density matrix and its associated Fock matrix.

In the context of the numerical solution of the Roothaan–Hall by the DIIS, the first choice correspond to the original form of the method [48], with $r^{(k)} = D^{(k)} - D^{(k-1)}$, while the second leads to the CDIIS [49], with $r^{(k)} =$

[12] We have seen this is indeed possible due to the orthogonal invariance of the energy functional.

[13] While it is known that the minimizers of the GHF and UHF models satisfy the Aufbau principle, this is not the case for those of the RHF or the Kohn–sham models. It is nevertheless always assumed that they do in the numerical practice.

$\big[F(D^{(k)}), D^{(k)}\big]$. This latter version of the DIIS is widely used in quantum chemistry softwares for electronic structure calculations, notably because of its simplicity with respect to implementation and its usually rapid, but not assured, convergence. Modifications of the procedure have been proposed in order to make it more robust, either by replacing the constraint on the coefficients, like in the C^2-DIIS [53], or by obtaining them by minimisation of an associated energy, like in the EDIIS[14] [38], the ADIIS[15] [33] or the LIST[16] [61].

4.6 Discussions of the assumptions 1 and 2 in such a setting

Let us now describe how the acceleration of the Roothaan algorithm by the CDIIS method recalled above enters the framework presented in Section 3.

Starting from a guess $D^{(0)}$, one first sets $\widetilde{D}^{(0)} = D^{(0)}$. After k steps of the method, given a set of $m_k + 1$ previous density matrices $D^{(k-m_k)}, \dots, D^{(k)}$, one assembles the *pseudo*-density matrix[17] $\widetilde{D}^{(k+1)}$ as

$$\widetilde{D}^{(k+1)} = \sum_{i=0}^{m_k} c_i^{(k)} D^{(k-m_k+i)},$$

where

$$\boldsymbol{c}^{(k)} = \underset{\substack{(c_0,\dots,c_{m_k}) \in \mathbb{R}^{m_k+1} \\ \sum_{i=0}^{m_k} c_i = 1}}{\arg\min} \left\| \sum_{i=0}^{m_k} c_i \left[F\left(D^{(k-m_k+i)}\right), D^{(k-m_k+i)} \right] \right\|_2,$$

the matrix norm $\|\cdot\|_2$ being the Frobenius norm. The next density matrix, $D^{(k+1)}$, is then obtained by applying the Aufbau principle to $\widetilde{D}^{(k+1)}$, that is, by diagonalising the Fock matrix[18] associated with $\widetilde{D}^{(k+1)}$ and forming $D^{(k+1)}$ from the N eigenvectors associated with its N smallest eigenvalues.

On the one hand, working with self-adjoint (symmetric) matrices of order d (d being the dimension of the vector space spanned by the basis functions used in the Galerkin approximation) with fixed trace, the integer n is equal to $\frac{1}{2}d(d+1)-1$, the function g corresponds to the application of the Aufbau principle to such a matrix, and Σ is the set $\mathcal{P}_N$ of pure state density matrices of rank N, defined in (41). On the other hand, the function f is the commutator between a given density matrix and its associated Fock matrix, $f(D) = [F(D), D]$, so that the integer p is equal to $\frac{1}{2}d(d-1)$.

In this setting, Assumption 1 simply states that the Roothaan algorithm, that is the base fixed-point iteration method, is locally convergent, which is a common assumption in the analysis of the DIIS or the Anderson acceleration (see [50] or [59] for instance). Assumption 2 amounts to a non-degeneracy assumption, which is equivalent to saying that the differential at x_* of the restriction of f to Σ is invertible. In the present context, the function f is already the gradient of the discretised Hartree–Fock (or Kohn–Sham) energy, and the assumption thus implies that the Hessian of this energy is non-degenerate.

Concerning regularity assumptions, the fact that the set $\mathcal{P}_N$ is a smooth submanifold is a consequence of the constant rank theorem. For the functions f and g being of class $\mathscr{C}^2$, this is always true for the Hartree–Fock functional, which is smooth. In the case of the Kohn–Sham model, this issue is more delicate, since most of the exchange-correlation functionals used in practice present a singularity at the origin, see *e.g.* [2]. However, it is reasonable to assume that the Kohn–Sham ground state has a non-vanishing density ρ, so that f and g are indeed regular in its neighbourhood.

[14] The difference in this method is that the coefficients of the linear expansion are such that

$$\boldsymbol{c}^{(k)} = \underset{\sum_{i=0}^{m_k} c_i = 1,\, c_i \in [0,1]}{\arg\min} E\left(\sum_{i=0}^{m_k} c_i D^{(k-m_k+i)} \right),$$

where E is the energy functional of the model under consideration.

[15] This other variant is similar to the EDIIS but uses the following augmented energy functional:

$$E^{\mathrm{A}}(D) = E(D^{(k)}) + 2\,\mathrm{tr}((D - D^{(k)})F(D^{(k)})) + \mathrm{tr}((D - D^{(k)})(F(D) - F(D^{(k)}))).$$

[16] This other variant is also similar to the EDIIS. It uses the so-called corrected Hohenberg–Kohn–Sham functional [66].

[17] This denomination stems from the fact that such a construction does not enforce the idempotency property on $\widetilde{D}^{(k+1)}$.

[18] For Hartree–Fock models, the pseudo-Fock matrix $\sum_{i=0}^{m_k} c_i^{(k)} F(D^{(k-m_k+i)})$ may as well be considered, since it holds that $F(\sum_{i=0}^{m_k} c_i^{(k)} D^{(k-m_k+i)}) = \sum_{i=0}^{m_k} c_i^{(k)} F(D^{(k-m_k+i)})$ due to constraint (3) being satisfied by the coefficients $c_i^{(k)}$, $i = 0, \dots, m_k$.

5 Numerical experiments

In this section, we report on some numerical experiments with the intent of illustrating the performances of our proposed variants of the CDIIS in the context of the electronic ground state calculations considered in Section 4. All the computations presented were performed using tools provided by the PySCF package [56]. The source code of our implementation of the CDIIS variants can be found in the following repository: https://plmlab.math.cnrs.fr/mchupin/restarted-and-adaptive-cdiis/.

5.1 Implementation details

For each variant, the least-squares problem for the extrapolation coefficients is solved using the unconstrained form involving the differences of residuals which are deemed the most convenient[19], as seen in Algorithms 2 and 3. The solution is achieved via the QR factorisation of a matrix of *tall and skinny* type (since the integer p is in general very large compared to m_k). Such a factorisation can be efficiently updated from step to step by means of a dedicated routine from the SciPy linear algebra library (namely `scipy.linalg.qr_update`), as the least-squares problem matrix is modified through the addition of a column (as in the variant with restarts) or the addition of a column and the possible removal of a set of columns (as in the adaptive-depth variant). A detailed cost analysis of the resulting factorisation algorithm is given in [31]. An added benefit of using the QR decomposition is that it directly provides the matrix of the orthogonal projector Π_k appearing in the restart condition (18), so that testing for this condition at each step in Algorithm 2 only entails a negligible cost.

Finally, we consider that numerical convergence is reached at iteration k if the residual norm $\|[F(D^{(k)}), D^{(k)}]\|_2$ is below some prescribed tolerance.

5.2 Test cases

Some of the molecular systems used in our experiments are taken from benchmarks found in [24, 33] and are considered as representative of challenging convergence tests for self-consistent field algorithms. Results of some of the tests are presented here, like the cadmium(II)-imidazole complex ($[Cd(Im)]^{2+}$) for instance, while others, like the acetaldehyde (C_2H_4O), the acetic acid ($C_2H_4O_2$), or the silane (SiH_4), are available in the online repository given above. The glycine ($C_2H_5NO_2$) test case comes from an example given in the PySCF library and the geometries for galactonolactone ($C_6H_{10}O_6$) and dimethylnitramine ($C_2H_6N_2O_2$) were found on the PubChem website (https://pubchem.ncbi.nlm.nih.gov/).

Both the restricted Hartree–Fock (RHF) model and the restricted Kohn–Sham (RKS) model are used in the experiments, the latter in conjunction with the B3LYP approximation of the exchange-correlation functional [6, 39].

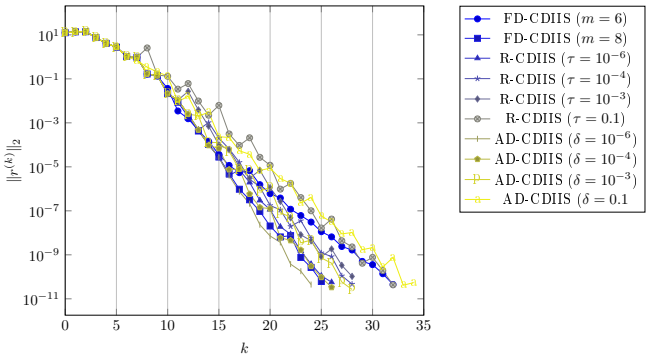
5.2.1 Global convergence behaviour

Acceleration techniques like the CDIIS are locally convergent methods and may sometimes give poor results if a mediocre initial guess is used. In practice, in order to ensure and achieve convergence in a small number of steps, one usually employs a combination of a relaxed constraint algorithm (like the ODA [10], the EDIIS [38] or the ADIIS [33].), for its global convergence properties, and of the CDIIS, for its fast local convergence (see Subsection 5.2.2). Nevertheless, our first experiment is meant to illustrate the convergence behaviour of the methods starting from the *core Hamiltonian guess*, for which the orbital coefficients are simply obtained by diagonalising the core Hamiltonian matrix in the discrete basis of the considered system.

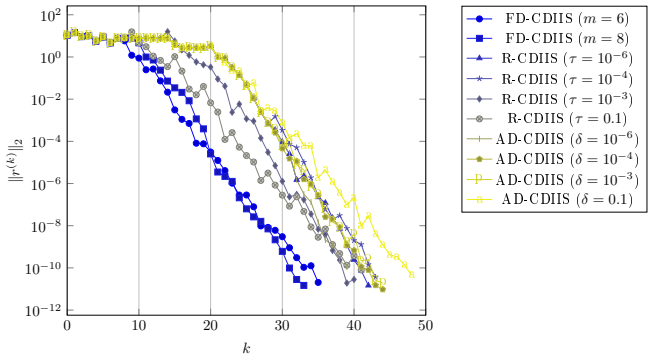
Figure 1 presents the convergence of the norm of the residual norm for the (classical) fixed-depth, restarted and adaptive-depth variants of the CDIIS with different values of their respective parameters: the fixed maximum depth m, the restart parameter τ and the adaptive-depth parameter δ. It is observed that the restarted and the adaptive-depth algorithms are efficient in most of the cases, but they may reach the convergence regime later than their fixed-depth counterpart (see Figures 1c and 1b). However, when this regime is attained, one can observe that the rate of convergence of both the restarted and adaptive-depth CDIIS is generally better than that of the fixed-depth CDIIS.

An effect of the use of a poor initial guess is the accumulation of many stored iterates in the early stages of the computation by the restarted and adaptive-depth variants, visible in Figure 2. The number of stored iterates clearly decreases as soon as a convergence regime is reached. This behaviour can be observed in all of our numerical experiments.

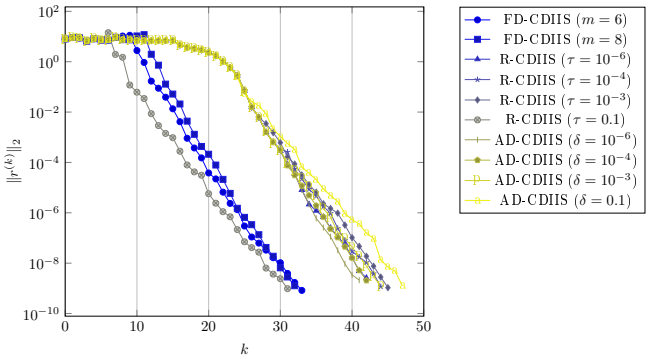
[19] For the fixed-depth CDIIS, the unconstrained formulation using successive differences is used.



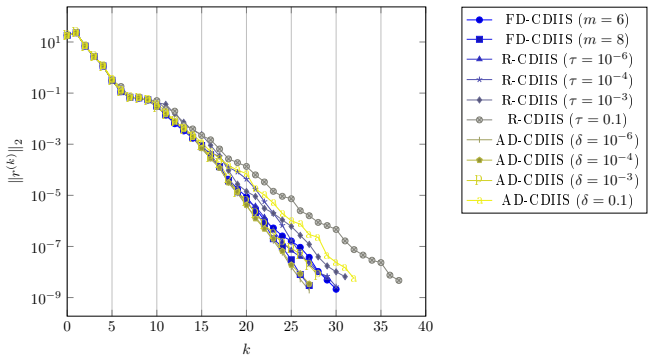
(a) Cadmium-imidazole complex in the RKS/B3LYP model with basis 3-21G.



(b) Glycine molecule in the RKS/B3LYP model with basis 6-31Gs.



(c) Dimethylnitramine molecule in the RHF model with basis 6-31G.



(d) Galactonolactone molecule in the RHF model with basis 6-31G.

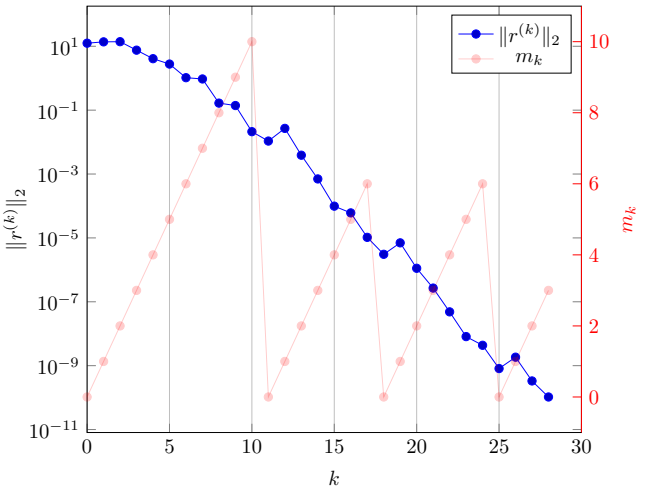
Figure 1: Residual norm convergence for the fixed-depth, restarted and adaptive-depth CDIIS on different molecular systems using an initial guess obtained by diagonalising the core Hamiltonian matrix.

5.2.2 Local convergence behaviour

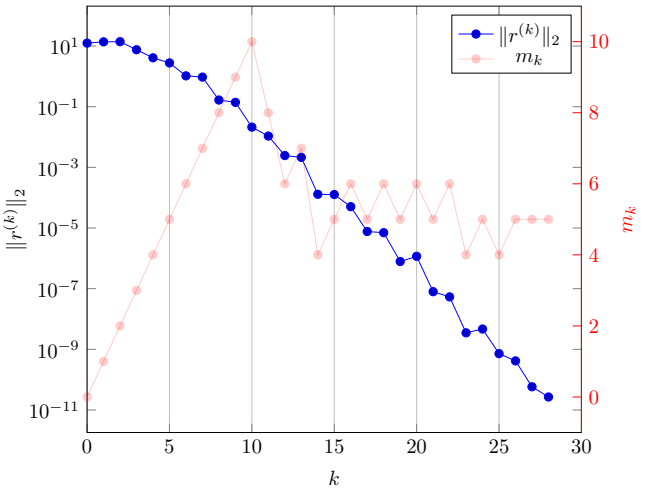
In order to properly evaluate their local convergence properties, the CDIIS variants were combined with a globally convergent method and an improved initial guess, both provided by the PySCF package. More precisely, the EDIIS with a fixed-depth equal to 8 and an initial guess generated from a superposition of atomic density matrices were used for the experiments with the RHF model, while the ADIIS with a fixed-depth equal to 8 and the same type of initial guess were used for the experiments with the RKS model. In both cases, the switch between the "global" method and the CDIIS variants was made when the residual norm was below 10^{-2}. With such an initialisation, convergence was immediately observed in every test case. Figure 3 presents the obtained results for the set of molecules already considered in Figure 1 and a smaller set of parameter values. In such a setting, the restarted and the adaptive-depth CDIIS are shown to be more efficient than the fixed-depth one when adequate values of their respective parameters are used.

Plotting the residual norm with the corresponding depth obtained in the numerical experiments, as done in Figure 4 for the dimethylnitramine and the glycine molecules, allows to witness that, as predicted by the theory for the restarted Anderson–Pulay acceleration, a restart occurs after a significant decrease of the residual norm. Unfortunately, in practice, one can notice a slowdown in the convergence (or even an increase of the residual norm) just after the restart, thus motivating the introduction of an adaptive-depth mechanism which would not suffer from such a defect.

Mean depth. The cost of the CDIIS in terms of storage and computational resource at given iteration is proportional to the value of the depth at this iteration. As a consequence, to properly compare the restarted and adaptive-depth variants with the classical fixed-depth CDIIS, we have computed the mean depth, denoted by $\bar{m}$ as the average value of m_k during an experiment. Figure 5 presents the evolution of $\bar{m}$ with respect to the values of the restart parameter τ and the adaptive-depth δ for each of the molecular systems we considered. It is seen $\bar{m}$ is a decreasing function of these parameters and that the two variants have on average lesser costs than their fixed counterparts, while their performances are comparable or better.



(a) Restarted CDIIS with $\tau = 10^{-4}$.



(b) Adaptive-depth CDIIS with $\delta = 10^{-4}$.

Figure 2: Residual norm convergence and corresponding depth value for the restarted and adaptive-depth CDIIS on the cadmium-imidazole complex in the RKS/B3LYP model with basis 3-21G, using an initial guess obtained by diagonalising the core Hamiltonian matrix.

Rate of convergence. For each of the molecular systems considered in Figure 3, the practical rate of convergence of the restarted and adaptive-depth CDIIS was computed using a linear regression and plotted against the values of the parameters τ and δ in Figure 6. As expected, it is apparent that the rate of convergence increases as the value of the parameter decreases, the evolution of the rate for the adaptive-depth variant being noticeably smoother.

Selecting values for the parameters. As previously mentioned, the decrease of the residual becomes faster and the average depth $\bar{m}$ increases as the parameters τ and δ are decreased. A closer scrutiny of both Figures 5 and 6 reveals that the convergence rate tends to slow down whereas the average depth grows at a constant slope. Thus, the gain of convergence by decreasing the parameters may not outweight the added computational cost of keeping a larger history of iterates. In this regard, a satisfying compromise is reached in our numerical tests by setting the values of δ and τ at 10^{-4}.

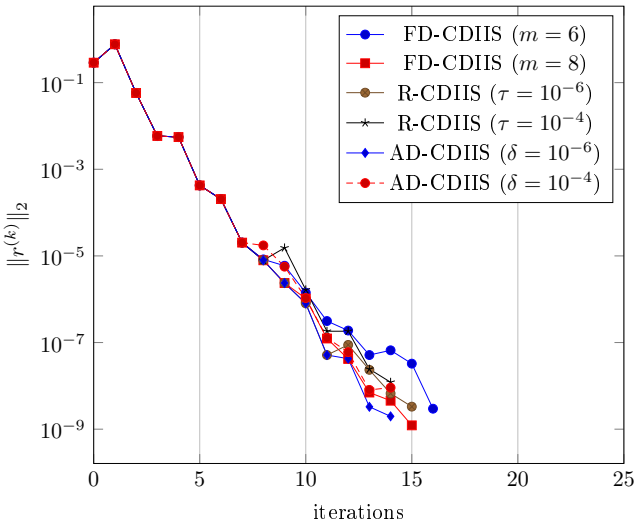
6 Conclusion

Motivated by the CDIIS technique, introduced by Pulay in 1982 [49], and its relation with other extrapolation processes for fixed-point iteration methods, we have considered a general and abstract class of acceleration methods and studied, theoretically and numerically, the local convergence properties of two of its instances: one allowing restarts, based on a condition initially introduced for a quasi-Newton using multiple secant equations [25, 50], and another one whose depth is adapted according to a criterion that appears to be new.

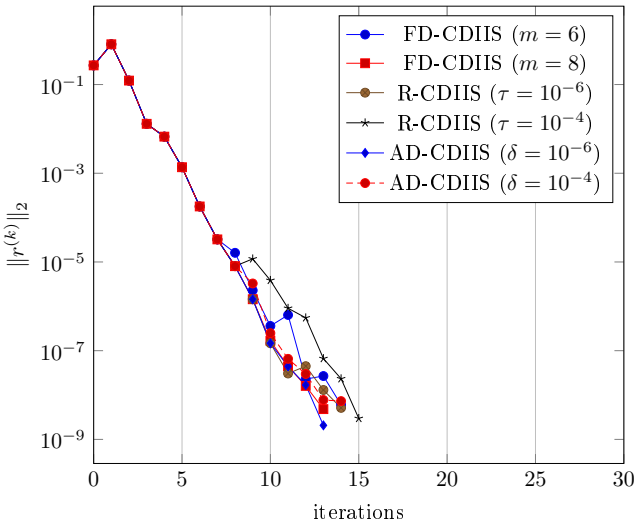
Our convergence results encompass the case of the DIIS or of the Anderson acceleration. They are obtained in a more general setting and rely on weaker assumptions than those existing in the literature [50, 59]. First, we do not impose a direct relation between the residual function f of the problem and the function g of the fixed-point iteration used to compute the solution. Second, the nondegeneracy hypothesis on the solution of the problem is weakened: it only involves the restriction of the function f to a submanifold Σ containing the range of the function g. Such generalisations are necessary in order to deal with the self-consistent field equations which are part of the numerical solution of both the Hartree–Fock method and the Kohn–Sham models, at the origin of the introduction of the CDIIS. Of course, our results also cover the case $g = Id + \beta f$ with $\Sigma = \mathbb{R}^n$.

Another novelty of our work is the absence of assumption concerning the uniform boundedness of the extrapolation coefficients. Indeed, the proposed restart and adaptive-depth mechanisms allow us to *a priori* prove such a bound. The only other work that we know of where a similar estimate can be found is [64], in which a global linear convergence analysis for a stabilized variant of the type-I Anderson acceleration is given. Our results thus provide, to the best of our knowledge, the first *complete* proof of accelerated convergence for the DIIS or the Anderson acceleration.

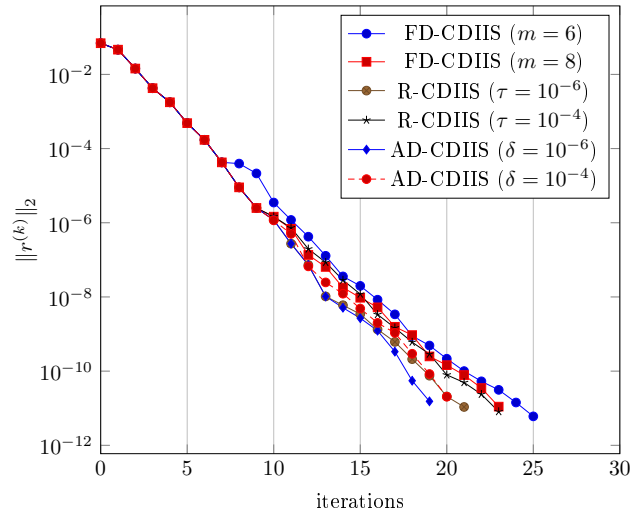
Finally, numerical experiments illustrate the good performances of the restarted and adaptive-depth acceleration algorithms applied to the numerical computation of the electronic ground state of various molecular systems. It has been observed that, with an adequate choice of their respective parameters, both variants exhibit a better convergence rate than their fixed-depth counterpart. In particular, the adaptive-depth variant shows good promise.



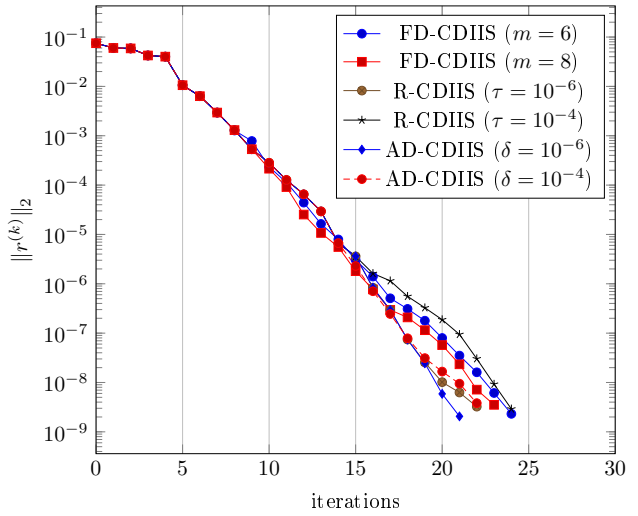
(a) Cadmium-imidazole complex in the RKS/B3LYP model with basis 3-21G.



(b) Glycine molecule in the RKS/B3LYP model with basis 6-31Gs.



(c) Dimethylnitramine molecule in the RHF model with basis 6-31G.



(d) Galactonolactone molecule in the RHF model with basis 6-31G.

Figure 3: Residual norm convergence for the fixed-depth, restarted and adaptive-depth CDIIS on different molecular systems using an initial guess provided by a globally convergent method.

It has also been noticed that acceleration occurs for values of the parameters several orders of magnitude larger than the theoretical estimates, and that the size of the set of stored iterates at each step is on average smaller than the fixed "rule of thumb" values generally found in implementations of the CDIIS. Understanding the reasons of this key fact will require further effort.

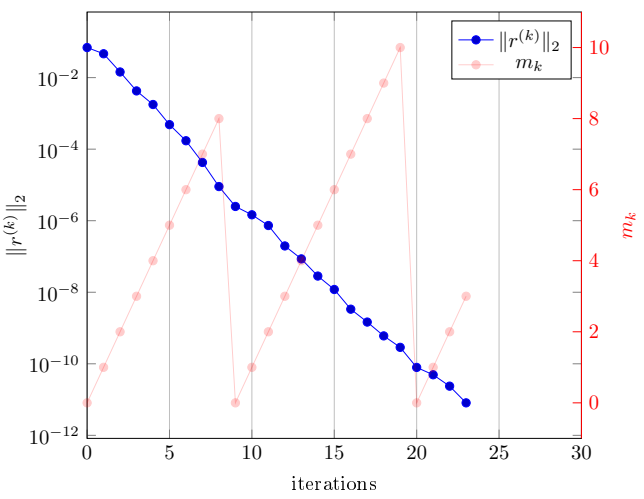
Acknowledgement

The authors wish to warmly thank Antoine Levitt, for providing them with encouragements, useful references, and insightful remarks on a first draft of the manuscript, and Qiming Sun, for his help with the PySCF package.

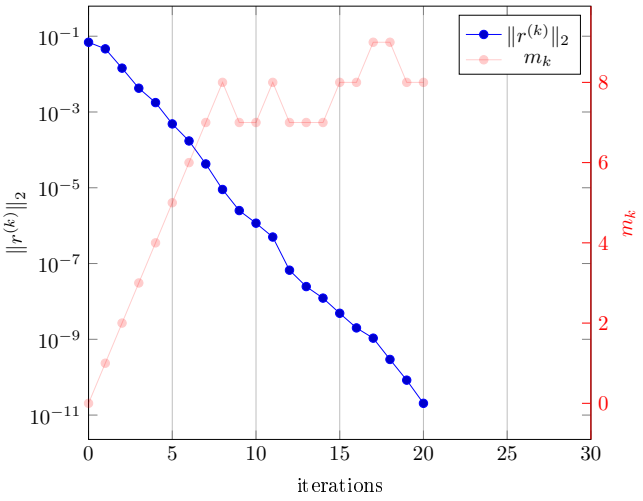
References

[1] H. An, X. Jia, and H. F. Walker. Anderson acceleration and application to the three-temperature energy equations. *J. Comput. Phys.*, 347:1–19, 2017. DOI: 10.1016/j.jcp.2017.06.031.

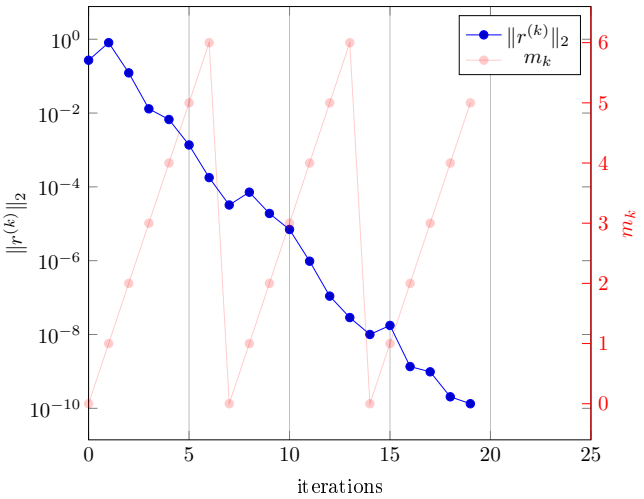
[2] A. Anantharaman and E. Cancès. Existence of minimizers for Kohn–Sham models in quantum chemistry. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 26(6):2425–2455, 2009. DOI: 10.1016/j.anihpc.2009.06.003.



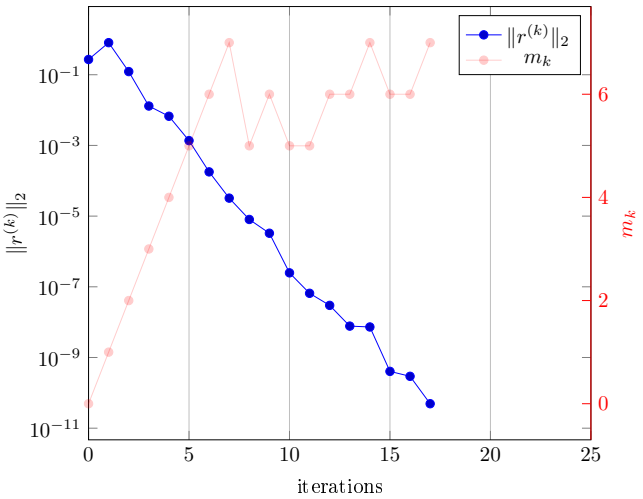
(a) Restarted CDIIS with $\tau = 10^{-4}$ for the dimethylnitramine molecule in the RHF model with basis 6-31G.



(b) Adaptive-depth CDIIS with $\delta = 10^{-4}$ for the dimethylnitramine molecule in the RHF model with basis 6-31G.



(c) Restarted CDIIS with $\tau = 10^{-4}$ for the glycine molecule in the RKS/B3LYP model with basis 6-31Gs.



(d) Adaptive-depth CDIIS with $\delta = 10^{-4}$ for the glycine molecule in the RKS/B3LYP model with basis 6-31Gs.

Figure 4: Residual norm convergence and corresponding depth for the restarted and adaptive-depth CDIIS on the dimethylnitramine and glycine molecules.

[3] D. G. Anderson. Iterative procedures for nonlinear integral equations. *J. ACM*, 12(4):547–560, 1965. doi: 10.1145/321296.321305.

[4] D. G. M. Anderson. Comments on "Anderson acceleration, mixing and extrapolation". *Numer. Algorithms*, 80(1):135–234, 2019. doi: 10.1007/s11075-018-0549-4.

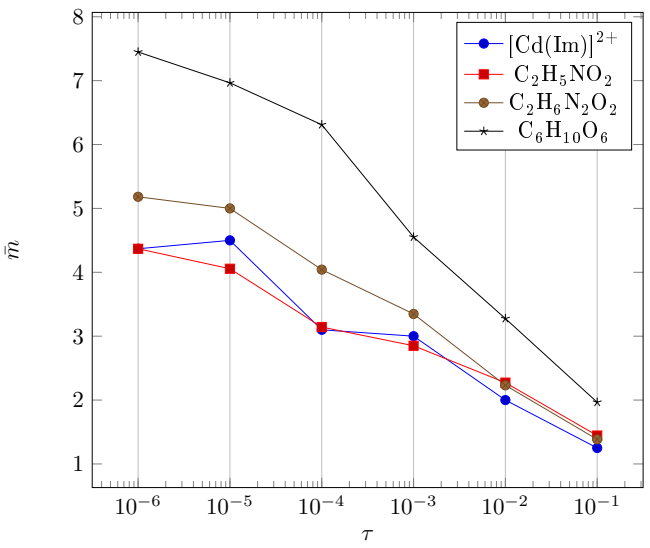
[5] A. S. Banerjee, P. Suryanarayana, and J. E. Pask. Periodic Pulay method for robust and efficient convergence acceleration of self-consistent field iterations. *Chem. Phys. Lett.*, 647:31–35, 2016. doi: 10.1016/j.cplett.2016.01.033.

[6] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38(6):3098–3100, 1988. doi: 10.1103/PhysRevA.38.3098.

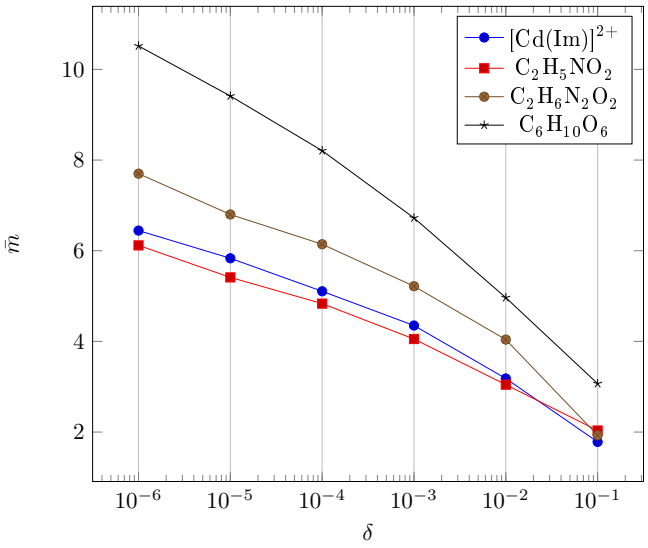
[7] C. Brezinski, M. Redivo-Zaglia, and Y. Saad. Shanks sequence transformations and Anderson acceleration. *SIAM Rev.*, 60(3):646–669, 2018. doi: 10.1137/17M1120725.

[8] C. G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Math. Comput.*, 19(92):577–593, 1965. doi: 10.1090/S0025-5718-1965-0198670-6.

[9] M. T. Calef, E. D. Fichtl, J. S. Warsa, M. Berndt, and N. N. Carlson. Nonlinear Krylov acceleration applied to a discrete ordinates formulation of the k-eigenvalue problem. *J. Comput Phys.*, 238:188–209, 2013. doi: 10.1016/j.jcp.2012.12.024.



(a) Depth mean $\bar{m}$ as a function of the restart parameter τ.



(b) Depth mean $\bar{m}$ as a function of the adaptive-depth parameter δ.

Figure 5: Evolution of the depth mean for different molecular systems and models.

[10] É. Cancès and C. Le Bris. Can we outperform the DIIS approach for electronic structure calculations? *Internat. J. Quantum Chem.*, 79(2):82–90, 2000. DOI: 10.1002/1097-461X(2000)79:2<82::AID-QUA3>3.0.CO;2-I.

[11] É. Cancès and C. Le Bris. On the convergence of SCF algorithms for the Hartree–Fock equations. *ESAIM Math. Model. Numer. Anal.*, 34(4):749–774, 2000. DOI: 10.1051/m2an:2000102.

[12] N. N. Carlson and K. Miller. Design and application of a gradient-weighted moving finite element code I: in one dimension. *SIAM J. Sci. Comput.*, 19(3):728–765, 1998. DOI: 10.1137/S106482759426955X.

[13] X. Chen and C. T. Kelley. Convergence of the EDIIS algorithm for nonlinear equations. *SIAM J. Sci. Comput.*, 41(1):A365–A379, 2019. DOI: 10.1137/18M1171084.

[14] P. Császár and P. Pulay. Geometry optimization by direct inversion in the iterative subspace. *J. Mol. Struct.*, 114:31–34, 1984. DOI: 10.1016/S0022-2860(84)87198-7.

[15] H. De Sterck. A nonlinear GMRES optimization algorithm for canonical tensor decomposition. *SIAM J. Sci. Comput.*, 34(3):A1351–A1379, 2012. DOI: 10.1137/110835530.

[16] E. De Sturler. Truncation strategies for optimal Krylov subspace methods. *SIAM J. Numer. Anal.*, 36(3):864–889, 1999. DOI: 10.1137/S0036142997315950.

[17] V. Eckert, P. Pulay, and H.-J. Werner. *Ab initio* geometry optimization for large molecules. *J. Comput. Chem.*, 18(12):1473–1483, 1997. DOI: 10.1002/(SICI)1096-987X(199709)18:12<1473::AID-JCC5>3.0.CO;2-G.

[18] C. Evans, S. Pollock, L. G. Rebholz, and M. Xiao. A proof that Anderson acceleration increases the convergence rate in linearly converging fixed point methods (but not in quadratically converging ones). *SIAM J. Numer. Anal.*, 58(1):788–810, 2020. DOI: 10.1137/19M1245384.

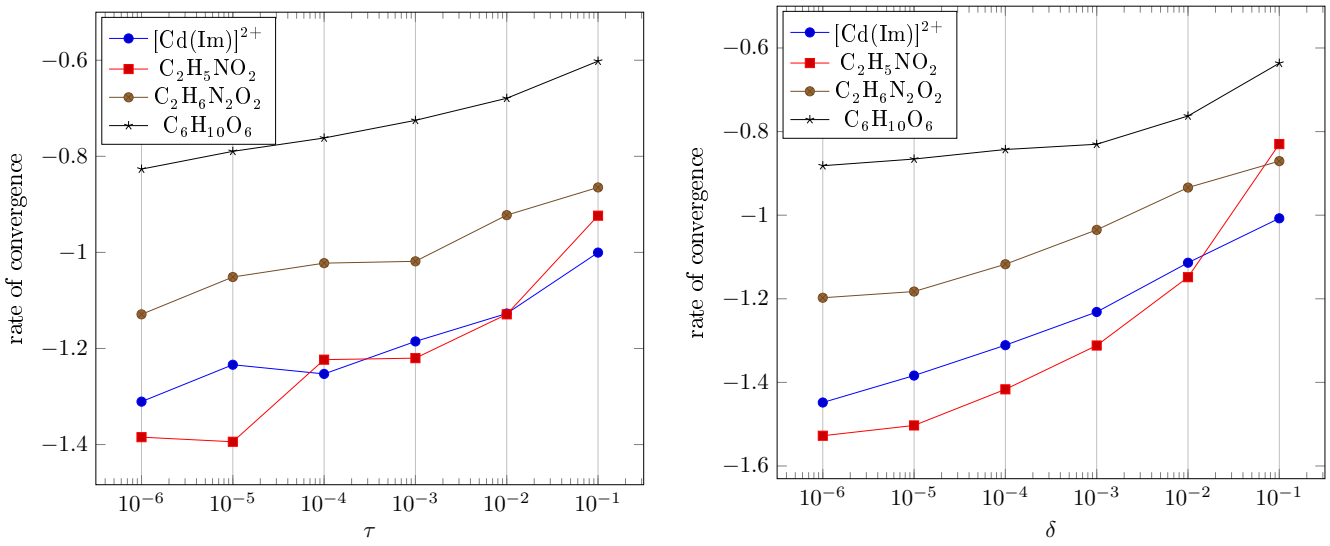
[19] V. Eyert. A comparative study on methods for convergence acceleration of iterative vector sequences. *J. Comput. Phys.*, 124(2):271–285, 1996. DOI: 10.1006/jcph.1996.0059.

[20] H.-R. Fang and Y. Saad. Two classes of multisecant methods for nonlinear acceleration. *Numer. Linear Algebra Appl.*, 16(3):197–221, 2009. DOI: 10.1002/nla.617.

[21] J.-L. Fattebert. Accelerated block preconditioned gradient method for large scale wave functions calculations in density functional theory. *J. Comput. Phys.*, 229(2):441–452, 2010. DOI: 10.1016/j.jcp.2009.09.035.

[22] V. Fock. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Z. Phys.*, 61(1-2):126–148, 1930. DOI: 10.1007/BF01340294.

[23] V. Ganine, N. J. Hills, and B. L. Lapworth. Nonlinear acceleration of coupled fluid-structure transient thermal problems by Anderson mixing. *Internat. J. Numer. Methods Fluids*, 71(8):939–959, 2013. DOI: 10.1002/fld.3689.



(a) Rate of convergence as a function of the restart parameter τ.

(b) Rate of convergence as a function of the adaptive-depth parameter δ.

Figure 6: Evolution of the rate of convergence for different molecular systems.

[24] A. J. Garza and G. E. Scuseria. Comparison of self-consistent field convergence acceleration techniques. *J. Chem. Phys.*, 137(5):054110, 2012. doi: 10.1063/1.4740249.

[25] D. M. Gay and R. B. Schnabel. Solving systems of nonlinear equations by Broyden's method with projected updates. Working Paper 169, National Bureau of Economic Research, 1977. doi: 10.3386/w0169.

[26] A. Greenbaum, V. Pták, and Z. Strakoš. Any nonincreasing convergence curve is possible for GMRES. *SIAM. J. Matrix Anal. Appl.*, 17(3):465–469, 1996. doi: 10.1137/S0895479894275030.

[27] A. Griewank. Broyden updating, the good and the bad! *Documenta Math.*, extra volume: optimization stories:301–315, 2012.

[28] G. G. Hall. The molecular orbital theory of chemical valency. VIII. A method of calculating ionization potentials. *Proc. Roy. Soc. London Ser. A*, 205(1083):541–552, 1951. doi: 10.1098/rspa.1951.0048.

[29] D. R. Hartree. The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods. *Math. Proc. Cambridge Philos. Soc.*, 24(1):89–110, 1928. doi: 10.1017/S0305004100011919.

[30] N. C. Henderson and R. Varadhan. Damped Anderson acceleration with restarts and monotonicity control for accelerating EM and EM-like algorithms. *J. Comput. Graph. Statist.*, 28(4):834–846, 2019. doi: 10.1080/10618600.2019.1594835.

[31] N. J. Higham and N. Strabić. Anderson acceleration of the alternating projections method for computing the nearest correlation matrix. *Numer. Algorithms*, 72(1):1021–1042, 2016. doi: 10.1007/s11075-015-0078-3.

[32] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864–B871, 1964. doi: 10.1103/PhysRev.136.B864.

[33] X. Hu and W. Yang. Accelerating self-consistent field convergence with the augmented Roothaan–Hall energy function. *J. Chem. Phys.*, 132(5):054109, 2010. doi: 10.1063/1.3304922.

[34] M. Kawata, C. M. Cortis, and R. A. Friesner. Efficient recursive implementation of the modified Broyden method and the direct inversion in the iterative subspace method: acceleration of self-consistent calculations. *J. Chem. Phys.*, 108(11):4426–4438, 1998. doi: 10.1063/1.475854.

[35] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133–A1138, 1965. doi: 10.1103/physrev.140.a1133.

[36] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Materials Sci.*, 6(1):15–50, 1996. doi: 10.1016/0927-0256(96)00008-0.

[37] K. N. Kudin and G. E. Scuseria. Converging self-consistent field equations in quantum chemistry – Recent achievements and remaining challenges. *ESAIM Math. Model. Numer. Anal.*, 41(2):281–296, 2007. doi: 10.1051/m2an:2007022.

[38] K. N. Kudin, G. E. Scuseria, and E. Cancès. A black-box self-consistent field convergence algorithm: one step closer. *J. Chem. Phys.*, 116(19):8255–8261, 2002. doi: 10.1063/1.1470195.

[39] C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37(2):785–789, 1988. doi: 10.1103/PhysRevB.37.785.

[40] M. Levy. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. *Proc. Nat. Acad. Sci. U.S.A.*, 76(12):6062–6065, 1979. doi: 10.1073/pnas.76.12.6062.

[41] E. H. Lieb. Density functionals for Coulomb systems. *Internat. J. Quantum Chem.*, 24(3):243–277, 1983. doi: 10.1002/qua.560240302.

[42] E. H. Lieb and B. Simon. The Hartree-Fock theory for Coulomb systems. *Comm. Math. Phys.*, 53(3):185–194, 1977. doi: 10.1007/BF01609845.

[43] P. L. Lions. Solutions of Hartree-Fock equations for Coulomb systems. *Comm. Math. Phys.*, 109(1):33–97, 1987. doi: 10.1007/BF01205672.

[44] P. A. Lott, H. F. Walker, C. S. Woodward, and U. M. Yang. An accelerated Picard method for nonlinear systems related to variably saturated flow. *Adv. in Water Res.*, 38:92–101, 2012. doi: 10.1016/j.advwatres.2011.12.013.

[45] A. L. Pavlov, G. V. Ovchinnikov, D. Y. Derbyshev, D. Tsetserukou, and I. V. Oseledets. AA-ICP: iterative closest point with Anderson acceleration. arXiv:1709.05479 [cs.RO], 2017.

[46] F. A. Potra and H. Engler. A characterization of the behavior of the Anderson acceleration on linear problems. *Linear Algebra Appl.*, 438(3):393–398, 2013. doi: 10.1016/j.laa.2012.09.008.

[47] P. P. Pratapa and P. Suryanarayana. Restarted Pulay mixing for efficient and robust acceleration of fixed-point iterations. *Chem. Phys. Lett.*, 635:69–74, 2015. doi: 10.1016/j.cplett.2015.06.029.

[48] P. Pulay. Convergence acceleration of iterative sequences. The case of SCF iteration. *Chem. Phys. Lett.*, 73(2):393–398, 1980. doi: 10.1016/0009-2614(80)80396-4.

[49] P. Pulay. Improved SCF convergence acceleration. *J. Comput. Chem.*, 3(4):556–560, 1982. doi: 10.1002/jcc.540030413.

[50] T. Rohwedder and R. Schneider. An analysis for the DIIS acceleration method used in quantum chemistry calculations. *J. Math. Chem.*, 49(9):1889–1914, 2011. doi: 10.1007/s10910-011-9863-y.

[51] C. C. J. Roothaan. New developments in molecular orbital theory. *Rev. Modern Phys.*, 23(2):69–89, 1951. doi: 10.1103/RevModPhys.23.69.

[52] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986. doi: 10.1137/0907058.

[53] H. Sellers. The C^2-DIIS convergence acceleration algorithm. *Internat. J. Quantum Chem.*, 45(1):31–41, 1993. doi: 10.1002/qua.560450106.

[54] H. Shepard and M. Minkoff. Some comments on the DIIS method. *Mol. Phys.*, 105(19-22):2839–2848, 2007. doi: 10.1080/00268970701691611.

[55] M. Spivak. *A comprehensive introduction to differential geometry*, volume one. Publish or Perish, third edition, 1999.

[56] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters, and G. K.-L. Chan. PySCF: the Python-based simulations of chemistry framework. *WIREs Comput. Mol. Sci.*, 8(1):e1340, 2017. doi: 10.1002/wcms.1340.

[57] L. Thøgersen, J. Olsen, A. Köhn, P. Jørgensen, P. Sałek, and T. Helgaker. The trust-region self-consistent field method in Kohn–Sham density-functional theory. *J. Chem. Phys.*, 123(7):074103, 2005. doi: 10.1063/1.1989311.

[58] A. Toth, J. A. Ellis, T. Evans, S. Hamilton, C. T. Kelley, R. Pawlowski, and S. Slattery. Local improvement results for Anderson acceleration with inaccurate function evaluations. *SIAM J. Sci. Comput.*, 39(5):S47–S65, 2017. doi: 10.1137/16M1080677.

[59] A. Toth and C. T. Kelley. Convergence analysis for Anderson acceleration. *SIAM J. Numer. Anal.*, 53(2):805–819, 2015. doi: 10.1137/130919398.

[60] H. F. Walker and P. Ni. Anderson acceleration for fixed-point iterations. *SIAM J. Numer. Anal.*, 49(4):1715–1735, 2011. doi: 10.1137/10078356X.

[61] Y. A. Wang, C. Y. Yam, Y. K. Chen, and G. Chen. Linear-expansion shooting techniques for accelerating self-consistent field convergence. *J. Chem. Phys.*, 134(24):241103, 2011. doi: 10.1063/1.3609242.

[62] T. Washio and C. W. Oosterlee. Krylov subspace acceleration for nonlinear multigrid schemes. *Electron. Trans. Numer. Anal.*, 6:271–290, 1997.

[63] J. Willert, W. T. Taitano, and D. Knoll. Leveraging Anderson acceleration for improved convergence of iterative solutions to transport systems. *J. Comput. Phys.*, 273:278–286, 2014. doi: 10.1016/j.jcp.2014.05.015.

[64] J. Zhang, B. O'Donoghue, and S. Boyd. Globally convergent type-I Anderson acceleration for non-smooth fixed-point iterations. arXiv:1808.03971 [math.OC], 2018.

[65] J. Zhang, Y. Yao, Y. Peng, H. Yu, and B. Deng. Fast K-Means clustering with Anderson acceleration. arXiv:1805.10638 [cs.LG], 2018.

[66] Y. A. Zhang and Y. A. Wang. Perturbative total energy evaluation in self-consistent field iterations: tests on molecular systems. *J. Chem. Phys.*, 130(14):144116, 2009. doi: 10.1063/1.3104662.