N-representable density matrix perturbation theory

Mamy Rivo Dianzinga

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

SPÉCIALITÉ : PHYSIQUE CHIMIE DE LA MATIÈRE CONDENSÉE

Par Mamy Rivo DIANZINGA

THÉORIE DES PERTURBATIONS EN MATRICE DENSITÉ *N*-RÉPRESENTABLE

Sous la direction de Alain FRITSCH

Soutenue le 07 Décembre 2016

Membres du jury :

M. CASTET, Frédéric	Professeur Université de Bordeaux	Président
M. RÉRAT, Michel	Professeur Université de Pau et des Pays de l'Adour	Rapporteur
M. NIKLASSON, Anders M. N.	Directeur de recherche Los Alamos National Laboratory	Rapporteur
M. BOWLER, David	Professeur University College London	Examinateur
M. HAYN, Roland	Professeur Université d'Aix-Marseille	Examinateur
M. FRITSCH, Alain	Professeur Université de Bordeaux	Examinateur
M. TRUFLANDIER, Lionel	Maître de Conférences Université de Bordeaux	Invité

Titre : Théorie des perturbations en matrice densité *N*-représentable

Résumé :

Alors que les approches standards de résolution de la structure électronique présentent un coût de calcul à la puissance 3 par rapport à la complexité du problème, des solutions permettant d'atteindre un régime asymptotique linéaire, O(*N*), sont maintenant bien connues pour le calcul de l'état fondamental. Ces solutions sont basées sur la "myopie" de la matrice densité et le développement d'un cadre théorique permettant de contourner le problème aux valeurs propres. La théorie des purifications de la matrice densité constitue une branche de ce cadre théorique. Comme pour les approches de type O(*N*) appliquées à l'état fondamental, la théorie des perturbations nécessaire aux calculs des fonctions de réponse électronique doit être révisée pour contourner l'utilisation des routines coûteuses. L'objectif est de développer une méthode robuste basée uniquement sur la recherche de la matrice densité perturbée, pour laquelle seulement des multiplications de matrices creuses sont nécessaires. Dans une première partie, nous dérivons une méthode de purification canonique qui respecte les conditions de *N*-representabilité de la matrice densité à une particule. Nous montrons que le polynôme de purification obtenu est auto-cohérent et converge systématiquement vers la bonne solution. Dans une seconde partie, en utilisant une approche de type Hartree-Fock, nous appliquons cette méthode aux calculs des tenseurs de réponses statiques non-linéaires pouvant être déterminés par spectroscopie optique. Au delà des calculs à croissance linéaire réalisés, nous démontrons que les conditions *N*-representabilité constituent un prérequis pour garantir la fiabilité des résultats.

Mots clés :

Matrice densité, réponses électroniques, champ auto-cohérent, croissance linéaire

Title: *N*-representable density matrix perturbation theory

Abstract:

Whereas standard approaches for solving the electronic structures present a computer effort scaling with the cube of the number of atoms, solutions to overcome this cubic wall are now well established for the ground state properties, and allow to reach the asymptotic linear-scaling, O(*N*). These solutions are based on the nearsightedness of the density matrix and the development of a theoretical framework allowing bypassing the standard eigenvalue problem to directly solve the density matrix. The density matrix purification theory constitutes a branch of such a theoretical framework. Similarly to earlier developments of O(*N*) methodology applied to the ground state, the perturbation theory necessary for the calculation of response functions must be revised to circumvent the use of expensive routines, such as matrix diagonalization and sum-over-states. The key point is to develop a robust method based only on the search of the perturbed density matrix, for which, ideally, only sparse matrix multiplications are required. In the first part of this work, we derive a canonical purification, which respects the *N*-representability conditions of the one-particle density matrix for both unperturbed and perturbed electronic structure calculations. We show that this purification polynomial is self-consistent and converges systematically to the right solution. As a second part of this work, we apply the method to the computation of static non-linear response tensors as measured in optical spectroscopy. Beyond the possibility of achieving linear-scaling calculations, we demonstrate that the *N*-representability conditions are a prerequisite to ensure reliability of the results.

Keywords:

Density matrix, electronic response functions, self-consistent field, linear scaling

---

Unité de recherche

Institut des Sciences Moléculaires (ISM) – Université de Bordeaux, CNRS UMR 5255
Bâtiment A12, 351 Cours de la Libération
33405 Talence cedex

N-representable density matrix perturbation theory

Mamy Rivo Dianzinga

Chemistry Department
University of Bordeaux

This dissertation is submitted for the degree of
Doctor of Theoretical Chemistry and Physics

December 2016

I dedicate this thesis to my loving and beloved parents for all their sacrifices ...

Declaration

I declare that this thesis is the presentation of the outcome of my original research work, after three years of hard labor. However, I clearly certify that contributions from other persons are involved. This dissertation contains equations, tables, figures, algorithms, appendices and a bibliography. This work is officially recognized under the supervision of Pr. Alain Fristch, but technically under the guidance of Dr. Lionel Truflandier at the Institut des Sciences Moléculaires of Bordeaux in France.

Mamy Rivo Dianzinga
December 2016

Acknowledgements

This thesis would not be achieved without the help of some persons. For this reason, I can not fail to thank a number of people. First of all, I deeply thank Dr. Lionel Truflandier who was involved in an important part in this thesis. Dr. Lionel Truflandier was not only my instructor who supported me during these three years. But he was also a wise friend trying to teach me the professional habits necessary to carry on scientific work.

I thank Pr. David Bowler, Pr. Anders Niklasson, Pr. Matt Challacombe, Dr. Jörg Kussmann and Pr. Christian Ochsenfeld for their interesting explanations of the linear scaling methods. Once more, I thank Pr. David Bowler and Pr. Anders Niklasson for agreeing to be members of the thesis defense committee.

I can not forget to thank Pr. Ionel Navon and Pr. Vladimir Ivanov who have taken some of their time to explain to me methodological details of their papers.

I also thank the Theoretical Chemistry & Modelling group of the Institut des Sciences Moléculaires which have accepted to invite Pr. Anders Niklasson. I also thank all the group members, especially Pr. Alain Fritsch and the computer engineer Philippe Aurel. I thank all the staff members of the Institut des Sciences Moléculaires, Professors and students, who came to my presentations.

I also thank Pr. Roland Hayn who has been my internship supervisor during my Master degree. Pr. Roland Hayn was my first professional relationship I had in France. I thank him for all he has done for me since I met him, for having helped me during my internship and for accepting to be in the thesis defense committee.

Of course, I thank my whole family, especially my father and my mother who both have always done everything for making me an educated person. Once more, I thank my parents who have always wanted me to obtain a Doctor of Philosphy degree. I deeply thank my parents, from the bottom of my heart, for calling me at least once a week to make sure I was doing well and to mentally and psychologically supported me in my work.

Abstract

Whereas standard approaches for solving the electronic structures present —at least— a computer effort scaling with the cube of the number of atoms, solutions to overcome this cubic wall are now well established for the ground state properties, and allow to reach the asymptotic linear-scaling —$O(N)$. They are based on the nearsightedness of the density matrix and the development of a theoretical framework allowing to bypass the standard eigenvalue problem, to directly solve the density matrix. The density matrix purification theory constitutes a branch of such a theoretical framework. Similarly to earlier developments of $O(N)$ methodology applied to the ground state, the perturbation theory necessary for the calculation of response functions must be revised to circumvent the use of expensive routines, such as matrix diagonalization and sum-over-states. The key point is to develop a robust method based only on the search of the perturbed density matrix, for which, ideally, only sparse matrix multiplications are required. In the first part of this work, we derive a canonical purification which respects the N-representability conditions of the one-particle density matrix for both unperturbed and perturbed electronic structure calculations. We show that this purification polynomial is self-consistent and converges systematically to the right solution. As a second part of this work, using a Hartree-Fock model, we apply the method to the computation of static non-linear response tensors as measured in optical spectroscopy. Beyond the possibility of achieving linear-scaling calculations, we demonstrate that the N-representability conditions are a prerequisite to ensure reliability of the results.

Résumé

Le Chapitre 1 résume les postulats et notions de la mécanique quantique nécessaires à l'introduction de la matrice densité à une particule comme variable fondamentale dans la résolution des équations de champ moyen auto-cohérent (*self-consistent field* – SCF), héritées de l'approximation mono-déterminantale de la fonction d'onde électronique. Nous montrons qu'il est possible de contourner la résolution de l'équation de Schrödinger —qui revient à résoudre un problème aux valeurs propres du type : $HC - CE = 0$, $(H, C) \in \mathbb{R}^{M\times M}$, $H = H^t$, $C^tC = CC^t = I$, $E = \text{diag}\{\epsilon_1\epsilon_2 \cdots \epsilon_M\}$— par la résolution directe de l'équation de Liouville-von Neumann du type : $HD - DH = 0$, $D \in \mathbb{R}^{M\times M}$, $D = D^t$, $\text{Tr}\{D\} = N$, $D^2 = D$, $0 < N < M$, dont la seule variable est la matrice densité D. La détermination de la matrice densité *via* les vecteurs et valeurs propres, respectivement C et E, de la matrice hamiltonienne, H, est numériquement coûteuse puisqu'elle nécessite une étape de diagonalisation. Les ressources de calcul nécessaires à la réalisation de cette étape présentent une croissance à la puissance 3 par rapport à la complexité du problème, cette dernière étant généralement définie par le nombre d'états occupés, N, ou le nombre d'atomes, ou encore la taille des matrices M. La détermination directe de la matrice densité —sans passer par le calcul des états propres— peut être explicitée sur la base d'un principe de minimisation —sous contrainte— de l'énergie du système, qui dans le cadre de cette thèse, correspond à l'énergie Hartree-Fock. En se référant aux procédures standards de minimisation lagrangienne, il est montré que les propriétés d'idempotence et de conservation de la trace de la matrice densité sont nécessaires et suffisantes pour garantir l'unicité de la solution. Ces propriétés sont regroupées sous le terme générique de conditions de N-representabilité de la matrice densité à une particule. Dans ce travail, tous les calculs sont réalisés sur la base de l'approche Hartree-Fock semi-empirique de Pariser, Parr et Pople (PPP). Les techniques d'accélération de la procédure SCF, notamment l'interpolation à paramètre constant (*damping*) et l'extrapolation par inversion directe du sous-espace des itérations (*direct inversion of the iterative subspace* – DIIS) sont également discutées. Un exemple d'application montre clairement les avantages de la méthode DIIS.

Le Chapitre 2 présente les méthodes les plus couramment utilisées pour la résolution directe de la matrice densité. Elles peuvent être regroupées en deux familles : (i) les méthodes de minimisation, et (ii) les méthodes de purification, chacune s'appuyant partiellement sur les conditions de N-representabilité évoquées précédemment. Dans le premier cas, le polynôme de purification de McWeeny dérive du principe de minimisation des moindres carrés de la contrainte d'idempotence, couplé à un algorithme de descente de gradient. Une alternative proposée par Li, Nunes et Vanderbilt (LNV) est basée sur la minimisation de la fonctionnelle de l'énergie sous une contrainte faible d'idempotence, couplée à un algorithme de gradient conjugué. Dans leur formulation grand canonique, ces deux approches sont en mesure de garantir les conditions de N-representabilité si et seulement si le potentiel chimique est connu à l'avance. En d'autres termes, les énergies correspondant au dernier état occupé et premier état inoccupé doivent être déterminées au préalable. Notons que pour un taux d'occupation, $\theta = N/M = 50\%$, le potentiel chimique peut être évalué avec une certaine précision à partir des limites supérieure et inférieure du spectre des valeurs propres. Une autre catégorie de purification, qualifiée de *canonique*, est également présentée. Dans cet ensemble, il n'est plus nécessaire d'évaluer les valeurs propres internes du spectre de l'hamiltonien. Néanmoins, leur application implique de prendre en compte plusieurs facteurs de stabilité, ce qui peut limiter leur efficacité et par conséquent, aussi complexifier les algorithmes de calcul. C'est dans ce cadre qu'est développée la première originalité de notre travail : en introduisant une méthode de purification canonique simple et robuste qui s'affranchit des considérations heuristiques de ces prédécesseurs. Cette nouvelle variante est basée sur la reformulation lagrangienne du principe de minimisation de l'idempotence de McWeeny en introduisant une contrainte explicite sur la trace de la matrice densité de faible idempotence. De cette façon, la méthode de purification est auto-cohérente —l'ajustement *a posteriori* du pôlynome n'est plus nécessaire— et vérifie les conditions de N-représentabilité à chaque itération. Dans le cadre de l'approximation des liaisons fortes, une étude détaillée des différentes méthodes de purification canoniques est réalisée. Il est prouvé que les performances de cette nouvelle approche sont comparables aux méthodes heuristiques tout en montrant des propriétés intéressantes de convergence monotonique et variationnelle. Toujours dans ce même chapitre, après avoir résumé les approximations permettant l'application de l'algèbre linéaire creuse aux méthodes de minimisation et purification, des calculs SCF-HF-PPP à croissance linéaire sont réalisés sur une série de nanotubes de carbone. Les difficultés liées à l'application des méthodes de troncatures numérique et radiale sont étudiées.

Dans le Chapitre 3, une généralisation des méthodes de variation/perturbation de la matrice densité à une particule est proposée, puis étendue à la résolution des équations couplées-perturbées du champ auto-cohérent (*coupled-perturbed self-consistent field* – CPSCF). Il en résulte trois approches : (i) la méthode standard basée sur la résolution spectrale de la matrice de Fock qui sera considérée par la suite comme la référence, (ii) l'approche proposée par Kussmann et Ochsenfeld basée sur les relations de commutations généralisées et l'utilisation du gradient conjugué, et (iii) la méthode de Niklasson basée sur le développement par récursion de l'opérateur de Fermi-Dirac perturbé. Dans le cadre du troisième formalisme, deux méthodes de purification sont utilisées, dont notre pôlynome canonique auto-cohérent. Il est à noter que pour les deux premières approches, le calcul des fonctions de réponse d'ordre supérieur nécessite la connaissance des fonctions de réponse associées aux ordres inférieurs, alors que dans les cas des purifications toutes les matrices densités perturbées sont calculées simultanément. En d'autres termes, l'ordre zéro (non-perturbé) et l'ordre supérieur (cible), ainsi que tous les ordres intermédiaires (si nécessaires), sont déterminés au cours du même processus de purification. Pour toutes les approches mentionnées la résolution des équations CPSCF est accélérée par l'algorithme de la dérivée du DIIS (D-DIIS) introduit par Weber et Daul. En fin de chapitre, nous généralisons cet algorithme pour n'importe quel ordre de perturbation.

Le Chapitre 4 est consacré à l'application des différentes méthodes de perturbation de la matrice densité pour le calcul des fonctions de réponse électronique induites par l'application d'un champ électrique externe et statique. Ces fonctions de réponse permettent, entre autres, de déterminer des grandeurs accessibles *via* des mesures d'optique non-linéaire, comme la polarisabilité, et/ou la première et seconde hyperpolarisabilité. Dans ce même chapitre, le principe des méthodes de différence de champ fini (*finite field difference* – FFD) est résumé. Dans un premier temps, toutes les méthodes sont appliquées aux calculs des propriétés optiques d'une série de petites molécules π-conjuguées en utilisant l'approche CPSCF-HF-PPP. L'accord remarquable observé démontre la fiabilité des différentes implémentations. Dans un deuxième temps, les calculs sont étendus à des systèmes facilement réplicables tels que des hydrocarbures insaturés. L'erreur de chaque méthode de perturbation basée sur la matrice densité (ii-iii) par rapport à la méthode de référence (i) est évaluée en fonction de la taille du polymère. Les résultats démontrent que la purification canonique est la seule à conserver une précision remarquable, et ce quelque soit l'ordre de perturbation. Ce résultat, qui constitue la deuxième originalité de ce travail, est directement relié au respect des précieuses conditions de N-representabilité discutées dans les chapitres précédents. Toutefois il est à noter que pour les approches ne vérifiant pas explicitement ces contraintes, l'erreur observée reste acceptable puisqu'elle

ne dépasse pas 0.1 % de la valeur attendue. Dans un troisième temps, les performances des différentes méthodes sont comparées grâce au décompte du nombre d'itérations CPSCF réalisé pour le calcul des fonctions de réponse en considérant des polymères de taille croissante. De cette comparaison, les approches perturbatives par purification se montrent être les plus stables avec un nombre moyen d'une douzaine d'itérations. La méthode basée sur le gradient conjugué requiert un nombre de cycles six fois plus élevé pour atteindre le même degré de convergence. En termes de temps de calcul global les méthodes perturbatives par purification sont les plus avantageuses. En analysant la variation du temps de calcul en fonction de la taille du polymère, on observe clairement que la méthodes de résolution des équations CPSCF basée uniquement sur la matrice densité sont plus efficaces de près d'un ordre de grandeur. Dans une dernière partie, l'étude précédente est répétée en applicant la troncature numérique. Les résultats obtenus dans le régime de croissance linéaire sont très similaires à ceux de l'analyse précédente.

Table of contents

List of figures

List of tables

Nomenclature

Subscripts

$\alpha, \beta, \gamma, ...$ indices for atomic orbitals

$i, j, k, ...$ indices for molecular orbitals

Other Symbols

$[A, B]$ Commutator $AB - BA$

$\{A, B\}$ Anticommutator $AB + BA$

$\mathrm{Tr}\{X\}$ Trace of a matrix X

$\|X\|$ Frobenuis norm of a matrix or a vector X

Acronyms / Abbreviations

AO Atomic orbitals

CPU Central processing unit

CSC Compressed sparse column

D-DIIS Derivative direct inversion of iterative subspace

DIIS Direct inversion of iterative subspace

DM Density matrix

DMM Density matrix minimization

DMPE Density matrix polynomial expansion

DMPT Density matrix perturbation theory

FFD Finite field difference

HPCP Hole particle canonical purification

LNV Li Nunes Vanderbilt

McW McWeeny

MM Matrix-matrix multiplication

MO Molecular orbitals

NLO Non Linear Optical

PMCP Palser Manolopoulos canonical purification

PPP Pariser-Parr-Pople

RHF Restricted Hartree-Fock

SCF Self consistent field

TB Tight binding

TC2 Trace correcting

TRS4 Trace resetting

Introduction

Sustained by the fast increase of investments in high performance computing technologies, quantum mechanics accuracy, as found in standard electronic structure methods, is about to reach the mesoscale within the next century. This implies the development of advanced parallelized programs including adapted theoretical frameworks and efficient numerical algorithms. At the same time, spectroscopies are improving in resolution and increasing in complexity with respect to the size of the probed systems, causing new difficulties for interpreting spectra, and new challenges for the theoreticians. There is no doubt in the importance of probing structure of matter at the atomic scale to establish clear relationships between the macroscopic properties and the atoms' arrangement in a sample. For that purpose, electromagnetic spectroscopies are banal experiments to obtain molecular fingerprint in physico-chemical analyses. Beyond their standard use, when we have no *a priori* knowledge on the system —but let's say the chemical composition for a material sample— analysis of the spectra may become tedious, or very difficult when dealing with disordered, amorphous or soft matter. Even if spectrocospies are continuously increasing their possibilities in resolution, support of theoretical prediction remains of primary interest for spectrum assignment and structure elucidation. On this way, the accuracy of the theoretical methods and the size of the investigated system consitute interrelated bottlenecks that need to be addressed.

Accurate quantum methods based on the explicit resolution of the many-electron Schrödinger equation, where dynamic and/or static electron correlation can selectively be accounted for remain limited to a few dozens of atoms or less depending on the level of theory. If now, we are willing to sacrifice accuracy in order to resolve electronic structure for larger systems where the number of electrons is above a few hundreds, single-determinant theories, such as Hartree-Fock[1] (HF) or Kohn-Sham (KS) density functional theory[2] are, until now, the only relevant methods. In that case, the many-electron Schrödinger equation is reduced to a mean-field one-electron equation, whose variational solutions are obtained by solving an eigenvalue problem. As a result, solutions are obtained by minimizing the electronic energy by means of the self-consistent field

(SCF) procedure. Whatever the basis set used to expand the wave functions, SCF methods have a common limitation on the size of the problem in such way that they require calculation of the full or partial set of eigenstates of the Hamiltonian matrix at each iteration of the SCF. The computational effort related to the direct/iterative diagonalization techniques[3, 4] or state-by-state conjugate gradient (CG) algorithm,[5, 6] increases with the cube of the number of electrons. Over the last two decades, alternative methods which scale linearly with the size of the problem were proposed as solution to these standard energy minimizations.[7–10] These methods are based on the Kohn's principle of electronic structure nearsightedness,[11, 12] which under certain conditions, eg. non-vanishing electronic gap, shows an exponential decay of the density matrix (DM) elements with respect to the distance.[13, 14] On exploiting this natural property, that is by enforcing sparsity of the matrices using a truncation scheme, $O(N)$ can be achieved by replacing the diagonalization step with DM solvers along with sparse-matrix multiply (SpMM) algorithms.[4, 15, 16]

Predictions of spectroscopic observables for molecules and solids rely on the solid approximation that the strengths of the electromagnetic radiations are negligible with respect to magnitude of the electron bonding allowing the safely use of the Rayleigh-Schrödinger wave function perturbation theory to compute the electronic response at any order. The first applications of this theory to SCF methods based on molecular-orbital (MO) wave functions were introduced during the 60s for the computation of molecular properties such as magnetic susceptibility,[17] static polarizabilities and force constants,[18, 19] which are all related to second-order energy derivatives through the calculation of the first-order change of the wavefunctions with respect to the small perturbation. Similarly to the unperturbed case, variational solutions of the perturbed MOs are obtained by solving the so-called coupled-perturbed self-consistent field (CPSCF) equations.[20–22] These early developments based either on the perturbed MOs or mixed perturbed AOs-MOs are well-known to involve cumbersome matrix transformations.[23, 24] In 1962, McWeeny had already introduced the elegant formalism of the density matrix perturbation theory (DMPT),[25] which was extended to the CPSCF equations resolution by Diercksen and McWeeny[26] for the evaluation of π-electron polarizabilities using the Pariser-Parr-Pople model. This work has first inspired Moccia to generalize the McW-CPSCF equations resolution to non-orthogonal basis.[27, 28] Perturbation-dependent non-orthogonal basis implementation was then proposed by Dodds, McWeeny, Sadlej and Wolinski[29–31] for the calculation of atomic (hyper)-polarizabilities using HF method and gaussian-type orbitals. The advantages of the McWeeny's approach over MO/AO-CPSCF have been clearly outlined, for instance, in the seminal article of

Wolinski, Hinton, and Pulay[32] dealing with the calculation of magnetic shieldings as measured in nuclear magnetic resonance (NMR) spectroscopy.

All the methods mentioned above have in common two limitations which narrow their applicability to few hundred atoms system at most, which are: (i) the unperturbed eigenstates are required prior the evaluation of the perturbed quantities, (ii) CPSCF equations resolution involves dense matrix multiplications which scale as M^3. Considering the specific case of AO/GTO basis, the construction of the effective Hamiltonian matrix —and the corresponding derivatives— should also be considered as an additional rate-limiting step, although robust linear scaling methods are nowadays well recognized.[33–36] Disregarding by now this specific feature, linear scaling can be achieved only if the two following conditions are fullfilled: (i) the theoretical framework involves the density matrix as the unique variable, that is no wavefunctions enter anymore in the formalism, (ii) perturbed density matrices must preserve some locality pattern allowing for SpMM algebra. Whereas, to the authors knowledge, analytical demonstration of the former point has not yet been proposed, raw numerical analysis have already shown that first-order perturbed density matrices for insulating systems present an approximate exponential decay of the elements[37, 38] which apply also, to a lesser extent, to higher orders.[39] Current methods dealing with condition (i) are intrinsically related to the DM solver used for the unperturbed case. Concerning the schemes, Ochsenfeld and Head-Gordon first reformulated the CPSCF equations in terms of the density matrix only[37] (referred as CG-CPSCF by the authors) starting from the Li-Nunes-Vanderbilt (LNV) unconstrained energy functional[40] where the McWeeny purification polynomial[41, 42] is used as input DM. Later, Kussmann and Ochsenfeld recognized important deficiencies in this initial version which were corrected in the alternative derivation of Ref. [43, 44].

In this manuscript we present the necessary and sufficient materials for developing a one-particle density matrix solver for electronic structure perturbation theory, especially within the framework of single-determinant theory. Chapter 1 introduces the main quantum mechanical fondations necessary to approach the density as the main variable. Chapter 2 presents the current methods applied to solve for the density matrix. In Chapter 3, a comprehensive description of the density matrix perturbation theory is presented. In Chapter 4, a large set of numerical experiments are performed to compare the various schemes.

Chapter 1

Density matrices and electronic structure

1.1 Density operator and stationary condition

We shall start from an ensemble of particles within some external potential. From quantum mechanics first postulate, information on this ensemble —eg. positions and momenta— is completely specified by a mathematical object: the wavefunction Ψ, which relates the probability amplitude of finding the system in a state —usually symbolized by the ket $|\Psi\rangle$— to its physical observation. The physical observation can only be realized through the scope of an operator $\hat{\mathcal{O}}$, for which, when applied to $|\Psi\rangle$ and integrated over the space of the possibilities, results in the most probable value, that is, the expectation value of the observable (operator). Obviously, depending on what we want to observe, the operator is chosen accordingly. Whatever is this observable, at the end, the process is always the same, that is, bring $|\Psi\rangle$ to the space specified by the operator, ie. $|\hat{\mathcal{O}}\Psi\rangle$, and integrate over that space, $\langle\Psi|\hat{\mathcal{O}}\Psi\rangle$. The time evolution of the ket is governed by the time-dependent Schrödinger equation[45]

$$i\hbar \frac{\partial |\Psi(t)\rangle}{\partial t} = \hat{\mathcal{H}}(t) |\Psi(t)\rangle \tag{1.1}$$

where $\hat{\mathcal{H}}$ is the Hamiltonian operator which describes the energy of the system. For an unperturbed and closed system, where the Hamiltonian operator does not depend explicitely on time, the time dependence of the wave function can be separated assuming

$$|\Psi(t)\rangle = e^{-i\hat{\mathcal{H}}t/\hbar} |\Psi\rangle \tag{1.2}$$

The expectation value of the time-independent Hamiltonian, ie. the energy $\mathcal{E}$, is then given by

$$\mathcal{E} = \frac{\langle\Psi(t)|\hat{\mathcal{H}}|\Psi(t)\rangle}{\langle\Psi(t)|\Psi(t)\rangle} = \frac{\langle\Psi|\hat{\mathcal{H}}|\Psi\rangle}{\langle\Psi|\Psi\rangle} \tag{1.3}$$

where the second equality emphasizes that $\mathcal{E}$ is independent on the time, that is, the eigenstate $\{\mathcal{E}, \Psi\}$ is stationary. Given a properly normalized state,

$$\langle\Psi|\Psi\rangle = 1 \tag{1.4}$$

evaluating $\mathcal{E}$ from the definition of Eq. (1.3), requires to solve an eigenvalue problem: the time-independent Schrödinger equation

$$\hat{\mathcal{H}} |\Psi\rangle = \mathcal{E} |\Psi\rangle \tag{1.5}$$

Let us now introduce another object, the density operator, which from a mathematical point of view represents an (orthogonal) projection from a vector space to the same vector space. The density operator is defined according to

$$\hat{\mathcal{D}} := |\Psi\rangle \langle\Psi| \tag{1.6}$$

such that, from the above definition and the normalization condition (1.4), $\hat{\mathcal{D}}$ verifies the following properties:

$$\text{hermicity:} \quad \hat{\mathcal{D}} = \hat{\mathcal{D}}^\dagger \tag{1.7a}$$

$$\text{idempotency:} \quad \hat{\mathcal{D}}^2 = \hat{\mathcal{D}} \tag{1.7b}$$

$$\text{normalization:} \quad \mathrm{Tr}\{\hat{\mathcal{D}}\} = 1 \tag{1.7c}$$

For the last equality, we made use of the property of the projection operators: the trace of the projection matrix is equal to the inner product of its constitutive eigenvectors. From here, we can search for the equation of motion of $\hat{\mathcal{D}}$ retaining the Schrödinger picture of Eq. (1.1). This gives rise to the Liouville-von Neumann equation[46]

$$i\hbar \frac{\partial \hat{\mathcal{D}}(t)}{\partial t} = \left[\hat{\mathcal{H}}(t), \hat{\mathcal{D}}(t)\right] \tag{1.8}$$

where $[\cdot, \cdot]$ denotes a commutator. Again, if we consider a conservative system, the solution of Eq. (1.8) is found to be

$$\hat{\mathcal{D}}(t) = e^{-i\hat{\mathcal{H}}t/\hbar}\, \hat{\mathcal{D}}\, e^{i\hat{\mathcal{H}}t/\hbar} \tag{1.9}$$

In that case, the stationary condition (1.5) can be recast in an operator form, according to

$$\hat{\mathcal{H}}\hat{\mathcal{D}} = \hat{\mathcal{D}}\hat{\mathcal{H}} \tag{1.10}$$

We may call it the time-independent Liouville-von Neumann equation. On multipying on the left (or on the right) by $\hat{\mathcal{D}}$ and assuming that conditions (1.7) are respected, we obtain

$$\hat{\mathcal{D}}\hat{\mathcal{H}}\hat{\mathcal{D}} = \hat{\mathcal{D}}^2\hat{\mathcal{H}} \quad \Leftrightarrow \quad \hat{\mathcal{H}}\hat{\mathcal{D}}^2 = \hat{\mathcal{D}}\hat{\mathcal{H}}\hat{\mathcal{D}} \tag{1.11a}$$

$$|\Psi\rangle \mathcal{E} \langle\Psi| = \hat{\mathcal{D}}\hat{\mathcal{H}} \quad \Leftrightarrow \quad \hat{\mathcal{H}}\hat{\mathcal{D}} = |\Psi\rangle \mathcal{E} \langle\Psi| \tag{1.11b}$$

$$\mathcal{E} = \mathrm{Tr}\{\hat{\mathcal{D}}\hat{\mathcal{H}}\} \quad \Leftrightarrow \quad \mathrm{Tr}\{\hat{\mathcal{H}}\hat{\mathcal{D}}\} = \mathcal{E} \tag{1.11c}$$

where me made explicit the evaluation of the energy of the stationary state. Therefore, we found that the expectation value of $\hat{\mathcal{H}}$ does not necessarily require to resolve the Schrödinger equation (1.5). An alternative route is offered by the calculation of the

matrix representation of the density operator. It is worth noticing that there is a one-to-one correspondence between $(\mathcal{E}, \Psi)$ and $(\hat{\mathcal{H}}, \hat{\mathcal{D}})$ such that, for non-degenerate cases, the unique solution of the Schrödinger equation leads to a unique definition of the density operator. This remark can, in principle, be extended to the calculation of (time-dependent/independent) properties based on Rayleigh-Schrödinger perturbation theory.

If associated with $|\Psi\rangle$ we defined an (abstract) vector space as being a separable Hilbert space of elements $\{|u_i\rangle\}_{i=1}^{\infty}$, such that, $\langle u_i|u_j\rangle = \delta_{ij}$, the ket is expanded into this basis according to

$$|\Psi\rangle = \sum_i \langle u_i|\Psi\rangle \, |u_i\rangle \tag{1.12}$$

On inserting the above definition into Eq. (1.6), the density operator transforms to

$$\hat{\mathcal{D}} = \sum_{i,j} \langle u_i|\Psi\rangle \, |u_i\rangle \, \langle u_j| \, \langle \Psi|u_j\rangle \tag{1.13a}$$

$$= \sum_{i,j} |u_i\rangle \, \langle u_i|\Psi\rangle \, \langle \Psi|u_j\rangle \, \langle u_j| \tag{1.13b}$$

$$= \sum_{i,j} |u_i\rangle \, \mathcal{D}_{ij} \, \langle u_j| \tag{1.13c}$$

$$\text{with:} \quad \mathcal{D}_{ij} := \langle u_i|\hat{\mathcal{D}}|u_j\rangle \tag{1.13d}$$

As a result, $\hat{\mathcal{D}}$ can be expressed as a superposition of basis projectors. In that case, it is easy to show that the definition (1.13c) also verifies the properties (1.7), the trace of the density operator being defined by

$$\mathrm{Tr}\{\hat{\mathcal{D}}\} = \sum_i \mathcal{D}_{ii} \tag{1.14}$$

If now, instead of enforcing the system to be described by a single pure state, we allow for a statistical description. On introducing a probability distribution over all the possible pure states, the ensemble ($\mathscr{S}$) density operator turns to be

$$\hat{\mathcal{D}}_{\mathscr{S}} := \sum_i p_i \, |\Psi_i^{\mathscr{S}}\rangle \, \langle \Psi_i^{\mathscr{S}}| \tag{1.15a}$$

subject to:

$$p_i \geq 0 \tag{1.15b}$$

$$\sum_i p_i = 1 \tag{1.15c}$$

where p_i is the probability of the system being found in the microstate $|\Psi_i^{\mathscr{S}}\rangle$. Depending on the statistical ensemble, eg. canonical ($\mathscr{S} = NVT$) or grand canonical ($\mathscr{S} = \mu VT$), the particle number of each $|\Psi_i^{\mathscr{S}}\rangle$ may vary. Within the NVT ensemble, subject to the constraints of Eq. (1.15) we can show that $\hat{\mathcal{D}}_{\mathscr{S}}$ verifies the hermiticity and the normalization conditions of Eq. (1.7), whereas idempotency is lost. This observation allows to distinguish a *mixed state* from a *pure state*. By induction, we can state that a pure state is well-defined within the microcanonical ensemble ($\mathscr{S} = NVE$) at $T = 0$.

1.2 Density matrix for fermion systems

1.2.1 Generalities

In this work, we are mainly interested in computing the energy of an ensemble of N electrons within the external potential created by K nuclei, and latter in the manuscript, its variation(s) with respect to some external perturbation(s). Since electrons are fermions, we must insure that the ket $|\Psi\rangle$ —as defined in some vector space– respects the anti-symmetry principle. As a consequence, the space of representation of $|\Psi\rangle$ is reduced to the anti-symmetric Hilbert space, such that, given the state vectors $\{|u_i\rangle\}_{i=1}^{\infty}$ allowing to define the N–particle symmetric state,

$$|u_1 u_2 \cdots u_N) := |u_1\rangle\,|u_2\rangle \cdots |u_N\rangle \tag{1.16}$$

any of the N-particle anti-symmetric state is obtained from the following definition:

$$\begin{aligned} |u_1 u_2 \cdots u_N\rangle &:= \mathscr{A}|u_1 u_2 \cdots u_N) \\ \text{with:} \quad \mathscr{A} &:= \frac{1}{\sqrt{N!}} \sum_{p \in \mathcal{S}_N} (-1)^p \mathscr{P}_p \end{aligned} \tag{1.17}$$

where $\mathscr{A}$ is the anti-symmetrization operator, $\mathscr{P}$ is the permutation operator of two particles, and $(-1)^p$ relates the parity of the permutation.[1] It is customary in chemistry to solve the time-independent Schrödinger equation [Eq. (1.5)] —or the time-independent Liouville-von Neumann [Eq. (1.10)]— within the position–spin space. In that context, the N–particle wavefunction is expressed in an abstract basis of continuous position

[1] For for the ensemble of N particles, there exist $N!$ possible permutations p, which constitute the elements of the (symmetric) group of permutation $\mathcal{S}_N$. As a result, we can show that: $\mathscr{A} = \mathscr{A}^\dagger$ and $\mathscr{A}^2 = \sqrt{N!}\mathscr{A}$.

vectors, according to

$$\Psi(\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_N) = \langle \mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_N|\mathscr{A}|\Psi\rangle \tag{1.18}$$

where $\{|\mathbf{x}_i\rangle := |\mathbf{r}_i, \sigma_i\rangle\}$ stands for the space and spin coordinates. In this new vector space, the matrix representation of the density operator as given in Eq. (1.13c) reads

$$\mathcal{D}(\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_N, \mathbf{x}'_1\mathbf{x}'_2\cdots\mathbf{x}'_N) = \Psi(\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_N)\Psi^*(\mathbf{x}'_1\mathbf{x}'_2\cdots\mathbf{x}'_N) \tag{1.19}$$

Note that by using Eq. (1.18) and the resolution of identity in a continuous basis,

$$\int d\mathbf{x}\, |\mathbf{x}\rangle \langle \mathbf{x}| = I \tag{1.20}$$

where $\int d\mathbf{x} := \int\int d\mathbf{r} d\sigma$, we can easily demonstrate that the hermiticity and imdempotency properties still hold for the density matrix defined in Eq. (1.19). Looking especially to the trace, we obtain

$$\begin{aligned}
&\mathrm{Tr}\{\mathcal{D}(\mathbf{x}_1\cdots\mathbf{x}_N, \mathbf{x}'_1\cdots\mathbf{x}'_N)\} \\
&\qquad = \int d\mathbf{x}_1\cdots d\mathbf{x}_N\ \mathcal{D}(\mathbf{x}_1\cdots\mathbf{x}_N, \mathbf{x}_1\cdots\mathbf{x}_N) &\text{(1.21a)}\\
&\qquad = \int d\mathbf{x}_1\cdots d\mathbf{x}_N\ \Psi(\mathbf{x}_1\cdots\mathbf{x}_N)\Psi^*(\mathbf{x}_1\cdots\mathbf{x}_N) &\text{(1.21b)}
\end{aligned}$$

On multiplying Eq. (1.19) by infinitesimal space-spin elements centered on each particle coordinate, keeping only the diagonal elements,

$$\mathcal{D}(\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_N, \mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_N)\ d\mathbf{x}_1 d\mathbf{x}_2\cdots d\mathbf{x}_N \tag{1.22}$$

we obtain the probability of an electron is in the space-spin volume element $d\mathbf{x}_1$ located at $\mathbf{x}_1$ with spin state s_1, while simultaneously another electron is in $d\mathbf{x}_2$ at $\mathbf{x}_2$ with spin state s_2 and so on.

1.2.2 Reduced density matrices

Since one- and two-electron operators are necessary to fully describe the electronic Hamiltonian, the generalized expression of Eq. (1.19) describing the so-called Nth–order density matrix can be reduced following the reduced density matrix theory. For the second–order reduced density matrix D_2 this gives

$$D_2(\mathbf{x}_1\mathbf{x}_2, \mathbf{x}'_1\mathbf{x}'_2) := \frac{N(N-1)}{2}\int d\mathbf{x}_3\cdots d\mathbf{x}_N\ \mathcal{D}(\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_N, \mathbf{x}'_1\mathbf{x}'_2\cdots\mathbf{x}_N) \tag{1.23}$$

and for the first–order density matrix D_1,

$$D_1(\mathbf{x}_1, \mathbf{x}_1') := N \int d\mathbf{x}_2 \cdots d\mathbf{x}_N \; \mathcal{D}(\mathbf{x}_1\mathbf{x}_2 \cdots \mathbf{x}_N, \mathbf{x}_1'\mathbf{x}_2 \cdots \mathbf{x}_N) \tag{1.24}$$

It is worth to emphasize that D_2 fully determines D_1. This is apparent by noting that

$$D_1(\mathbf{x}_1, \mathbf{x}_1') = \frac{2}{N-1} \int d\mathbf{x}_2 \; D_2(\mathbf{x}_1\mathbf{x}_2, \mathbf{x}_1'\mathbf{x}_2) \tag{1.25}$$

From the above definitions, we found that the trace of the second–order density matrix leads to the number of particle pairs,

$$\mathrm{Tr}\{D_2(\mathbf{x}_1\mathbf{x}_2, \mathbf{x}_2'\mathbf{x}_2')\} = \int d\mathbf{x}_1 d\mathbf{x}_2 \; D_2(\mathbf{x}_1\mathbf{x}_2, \mathbf{x}_1\mathbf{x}_2) = \frac{N(N-1)}{2} \tag{1.26}$$

whereas, the first–order density matrix is normalized to the number of particles

$$\mathrm{Tr}\{D_1(\mathbf{x}_1, \mathbf{x}_1')\} = \int d\mathbf{x}_1 \; D_1(\mathbf{x}_1, \mathbf{x}_1) = N \tag{1.27}$$

Indeed by integrating out the spin variable, the diagonal of $\mathcal{D}_1$ is recognized as the one-electron density function, —usually identified as $\rho(\mathbf{r})$—, that is, the probability of finding one electron in $d\mathbf{r}$ at position $\mathbf{r}$ assuming that the others are anywhere else, the indistinguishability of the fermions being properly accounted for in Eq. (1.27). In this work, we have reduced the scope of our investigations to closed-shell systems where the total (electron) spin momentum is zero. Thereafter, we shall integrate out the spin variables. Within the context of continuous reduced density matrices, the expectation values of any one-particle ($\hat{\mathcal{O}}_1$) and two-particle ($\hat{\mathcal{O}}_2$) operators are given by

$$\mathrm{Tr}\{\mathcal{O}_1\mathcal{D}\} = \int\int d\mathbf{r}_1 d\mathbf{r}_1' \; \mathcal{O}_1(\mathbf{r}_1, \mathbf{r}_1') D_1(\mathbf{r}_1', \mathbf{r}_1) \tag{1.28}$$

$$\mathrm{Tr}\{\mathcal{O}_2\mathcal{D}\} = \int\int d\mathbf{r}_1 d\mathbf{r}_2 \; d\mathbf{r}_1' d\mathbf{r}_2' \; \mathcal{O}_2(\mathbf{r}_1\mathbf{r}_2, \mathbf{r}_1'\mathbf{r}_2') D_2(\mathbf{r}_1'\mathbf{r}_2', \mathbf{r}_1\mathbf{r}_2) \tag{1.29}$$

If we assume that those operators are local, ie.

$$\mathcal{O}_1(\mathbf{r}_1, \mathbf{r}_1') = \mathcal{O}_1(\mathbf{r}_1)\delta(\mathbf{r}_1 - \mathbf{r}_1') \tag{1.30}$$

$$\mathcal{O}_2(\mathbf{r}_1\mathbf{r}_2, \mathbf{r}_1'\mathbf{r}_2') = \mathcal{O}_2(\mathbf{r}_1\mathbf{r}_2)\delta(\mathbf{r}_1 - \mathbf{r}_1')\delta(\mathbf{r}_2 - \mathbf{r}_2') \tag{1.31}$$

Eqs. (1.28) and (1.29) simplify to

$$\mathrm{Tr}\{\mathcal{O}_1\mathcal{D}\}_{\mathbf{r}_1=\mathbf{r}'_1} = \int d\mathbf{r}_1 \; \mathcal{O}_1(\mathbf{r}_1) D_1(\mathbf{r}_1, \mathbf{r}_1) \tag{1.32}$$

$$\mathrm{Tr}\{\mathcal{O}_2\mathcal{D}\}_{\mathbf{r}_1=\mathbf{r}'_1,\mathbf{r}_2=\mathbf{r}'_2} = \int d\mathbf{r}_1 d\mathbf{r}_2 \; \mathcal{O}_2(\mathbf{r}_1\mathbf{r}_2) D_2(\mathbf{r}_1\mathbf{r}_2, \mathbf{r}_1\mathbf{r}_2) \tag{1.33}$$

The non-relativistic time-independent Hamiltonian operator for an ensemble of N fermions within the external potential (V_{ext}) created by K nuclei —considered fixed in positions $\{\mathbf{R}_A\}_{A=1}^K$ with charges $\{Z_A\}_{A=1}^K$ — can be expressed as the sum over the electron kinetic energy operator ($\hat{T}$), the electron-electron Coulomb interaction ($\hat{G}$), and the aforementioned external potential

$$\hat{\mathcal{H}} := \hat{T} + \hat{V}_{\text{ext}} + \hat{G} \tag{1.34}$$

with the following explicit definitions[2]

$$\hat{T} = -\frac{1}{2}\sum_i \nabla_i^2, \qquad \hat{G} = \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}$$

and

$$\hat{V}_{\text{ext}} = \sum_i v(\mathbf{r}_i) \quad \text{with:} \quad v(\mathbf{r}_i) = -\sum_A \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} \tag{1.35}$$

where i runs over the electrons, $(i < j)$ the electron pairs, and A the nuclei. By recognizing that $\hat{T}$ and $\hat{V}_{\text{ext}}$ are one-electron operators, and $\hat{G}$ is a two-electron operator, in virtue of Eqs. (1.32) and (1.32), the electronic energy is given by

$$\mathcal{E}\,[D_2] = T\,[D_1] + V_{\text{ext}}\,[D_1] + G\,[D_2] \tag{1.36}$$

according to the following definitions

$$T\,[D_1] := \mathrm{Tr}\{\hat{T}\mathcal{D}\} = \int d\mathbf{r}_1 \; \delta(\mathbf{r}_1 - \mathbf{r}'_1) \left(-\frac{1}{2}\nabla^2_{\mathbf{r}}\right) D_1(\mathbf{r}_1, \mathbf{r}'_1) \tag{1.37}$$

$$V_{\text{ext}}\,[D_1] := \mathrm{Tr}\{\hat{V}_{\text{ext}}\mathcal{D}\} = \int d\mathbf{r}_1 \; v(\mathbf{r}_1) D_1(\mathbf{r}_1, \mathbf{r}_1) \tag{1.38}$$

$$G\,[D_2] := \mathrm{Tr}\{\hat{G}\mathcal{D}\} = \int d\mathbf{r}_1 d\mathbf{r}_2 \; \frac{D_2(\mathbf{r}_1\mathbf{r}_2, \mathbf{r}_1\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{1.39}$$

[2]In atomic units: $\hbar = m_e = e = 1$, where $\hbar$, m_e and e are the reduced Planck constant, the electron mass and elementary charge, respectively.

where, for the kinetic energy component, we made explicit the fact that, first, the Laplacien is applied to $D_1(\mathbf{r}_1, \mathbf{r}'_1)$, and then, integration over space coordinate is performed for $\mathbf{r}' = \mathbf{r}$. The square braket $[\cdot]$ in Eqs. (1.37)-(1.39) indicates the functional dependence of the energy contribution over the density matrix. As previously noted, since D_1 is determined by D_2, it is necessary and sufficient to know the second–order reduced density matrix. Starting from an initial guess for D_2 without any prior knowledge of the electronic wavefunction, and solving a variational principle related to Eq. (1.36) in order to evalute the energy of an ensemble of interacting electrons is an extraordinarily difficult task that we leave to the specialists of the reduced density matrix theory (RDMT).[47–52] At this stage, it is important to note that the RDMT embrasses, at some point, the formalism of the (orbital-free) density functional theory[2, 53–56] (DFT) and Kohn-Sham DFT (KS-DFT),[57] in the sense where, for a given $V_{\text{ext}}[D_1]$, both of them are trying to approximate $T[D_1]$ and $G[D_2]$ —in terms of $D_1(\mathbf{r}_1, \mathbf{r}_1)$ for DFT and $D_1(\mathbf{r}_1, \mathbf{r}'_1)$ for KS-DFT—, without requiring the support of Ψ, which is also a tedious challenge.

Interestingly for our work, addressing the difficulties mentioned aboved involved introducing a set of constraints which must be fullfiled during the search of the solution to guarantee that, at convergence, the density matrix corresponds to an acceptable anti-symmetrized wavefunction. These constraints are called the N–representability conditions[58–60] for RDMT, in addition to the v–representability conditions,[2, 61, 62] a specificity of the DFT.

1.2.3 Density matrix for a single determinant

Let us now consider a more standard Hilbert space built from a discrete set of square integrable functions $\{|\psi_i\rangle\}_{i=1}^{\infty}$, such that, $\psi_i(\mathbf{r}) = \langle \mathbf{r}|\psi_i\rangle$ and $\langle \psi_i|\psi_j\rangle = \delta_{ij}$. If we impose that the N–wavefuntion is approximated by a single anti-symmetrized product [Eq. (1.17)] —also called a Slater determinant— of a subset of these functions, that is $\{|\psi_i\rangle\}_{i=1}^{N}$, the first-order density matrix in the coordinate representation, reads

$$D_1(\mathbf{r}_1, \mathbf{r}'_1) = \sum_{i=1}^{N} \psi_i(\mathbf{r}_1)\psi_i^*(\mathbf{r}'_1) \tag{1.40}$$

We can show that there exists a one-to-one mapping between D_1 and a single anti-symmetrized product of the form (1.17). As a consequence of the density operator properties: (i) the trace of D_1 is equal to the number of electrons [Eq. (1.27)] and (ii) D_1 is idempotent, ie. $\int d\mathbf{r}''\, D_1(\mathbf{r}, \mathbf{r}'')D_1(\mathbf{r}'', \mathbf{r}') = D_1(\mathbf{r}, \mathbf{r}')$, constitute *necessary and sufficient* conditions for the first-order density matrix to correspond to a *pure state* approximated

by a single Slater determinant.[2] In that context, the second-order density matrix can be defined in terms of D_1 according to

$$D_2(\mathbf{r}_1\mathbf{r}_2, \mathbf{r}'_2\mathbf{r}'_2) = \frac{1}{2}\left(D_1(\mathbf{r}_1, \mathbf{r}'_1)D_1(\mathbf{r}_2, \mathbf{r}'_2) - D_1(\mathbf{r}_1, \mathbf{r}'_2)D_1(\mathbf{r}_2, \mathbf{r}'_1)\right) \tag{1.41}$$

On introducing the Eqs. (1.26) and (1.41) into Eq. (1.36), we obtain the definition of the Hartree-Fock (HF) energy expressed in density matrix form:

$$\mathcal{E}_{\text{HF}}[D_1] := T[D_1] + V_{\text{ext}}[D_1] + G[D_1] \tag{1.42}$$
$$\text{with:} \quad G[D_1] := J[D_1] + K[D_1]$$

such that,

$$J[D_1] := +\frac{1}{2}\int d\mathbf{r}_1 d\mathbf{r}_2 \frac{D_1(\mathbf{r}_1, \mathbf{r}_1)D_1(\mathbf{r}_2, \mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{1.43}$$

$$K[D_1] := -\frac{1}{2}\int d\mathbf{r}_1 d\mathbf{r}_2 \frac{D_1(\mathbf{r}_1, \mathbf{r}_2)D_1(\mathbf{r}_2, \mathbf{r}_1)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{1.44}$$

where the electron-electron energy is a sum over the classical Coulomb repulsion J, and the quantum exchange energy K arising from the anti-symmetry principle of Eq. (1.17). In textbooks it is also common to find the following condensed expression for the Hartree-Fock energy,[3]

$$\mathcal{E}_{\text{HF}}[\rho(\mathbf{r}, \mathbf{r}')] = \int_{\mathbf{r}=\mathbf{r}'} d\mathbf{r}\left(-\frac{1}{2}\nabla^2 + v(\mathbf{r})\right)\rho(\mathbf{r}, \mathbf{r}') + \frac{1}{2}\int d\mathbf{r}d\mathbf{r}'\left(\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} - \frac{\rho(\mathbf{r}, \mathbf{r}')\rho(\mathbf{r}', \mathbf{r})}{|\mathbf{r} - \mathbf{r}'|}\right) \tag{1.45}$$

1.2.4 Density matrix representation in finite non-orthogonal basis

Practical calculation of the HF energy and other properties requires a closed and separable Hilbert subspace, although the accuracy of result is closely linked with the dimension of such space through the variational principle. In chemistry, it is rather natural to use local atomic orbitals (AO) to describe chemical bonds in molecules. When properly parametrized, these orbitals permit to reach a high level of accuracy with a limited set of variational parameters. Unfortunatly, these orbitals form a non-orthogonal basis for the representation of the operators, increasing the computational complexity.

[3] Using the following substitutions: $\rho(\mathbf{r}) := D_1(\mathbf{r}, \mathbf{r})$ and $\rho(\mathbf{r}, \mathbf{r}') := D_1(\mathbf{r}, \mathbf{r}')$.

Let us introduce a set of M non-orthogonal atomic-like basis functions $\{|\phi_\mu\rangle\}_{\mu=1}^{M}$ to expand the one-electron states[4] $\{|\psi_a\rangle\}_{a=1}^{M}$ —the sp-called molecular orbitals (MO). According to Eq. (1.12), we have

$$|\psi_a\rangle = \sum_{\mu=1}^{M} \langle\phi_\mu|\psi_a\rangle\, |\psi_a\rangle \tag{1.46}$$

which, in coordinate representation, transforms to

$$\langle\mathbf{r}|\psi_a\rangle = \sum_{\mu} \langle\phi_\mu|\psi_a\rangle\, \langle\mathbf{r}|\phi_\mu\rangle \quad \Leftrightarrow \quad \psi_a(\mathbf{r}) = \sum_{\mu} c_{\mu a}\phi_\mu(\mathbf{r}) \tag{1.47}$$

where the inner products, $\{\langle\phi_\mu|\psi_a\rangle\}_\mu$, are generally identified as the linear combination of atomic orbital (LCAO) coefficients $c_{\mu a}{}_\mu$. These coefficients constitute the set of variational parameters to be optimized. By making use of Eqs. (1.13c) and (1.40), we obtain for the one-particle density operator the following expression

$$\hat{D} := \sum_{\mu,\nu} |\mu\rangle\, D_{\mu\nu}\, \langle\nu|\,, \quad \text{with:} \quad D_{\mu\nu} := \sum_{i=1}^{N} c_{\mu i} c^*_{\nu i} \tag{1.48}$$

where i runs over the N occupied states, that is, in restricted Hartree-Fock theory: $N = N/2$. The set of elements $\{D_{\mu\nu}\}$ constitutes the HF one-particule density matrix in the atomic orbitals basis. In coordinate representation Eq. (1.48) transforms to

$$D(\mathbf{r},\mathbf{r}') := \langle\mathbf{r}|\hat{D}|\mathbf{r}'\rangle = \sum_{\mu,\nu} \phi_\mu(\mathbf{r}) D_{\mu\nu} \phi_\nu(\mathbf{r}') \tag{1.49}$$

Note in passing, for such kind of basis set we need to introduce the overlap matrix S to respect the idempotency relation, ie.

$$D(\mathbf{r},\mathbf{r}') = \int d\mathbf{r}''d\mathbf{r}'''\; D(\mathbf{r},\mathbf{r}'')S(\mathbf{r}'',\mathbf{r}''')D(\mathbf{r}''',\mathbf{r}') \tag{1.50}$$

For a properly normalized basis, the elements of the S matrix are determined by

$$S_{\mu\nu} = \langle\mu|\nu\rangle = \int d\mathbf{r}\; \phi^*_\mu(\mathbf{r}-\mathbf{R}_\mu)\phi_\nu(\mathbf{r}-\mathbf{R}_\nu) \quad \left\{ \begin{array}{l|l} S_{\mu\mu} = 1 & \\ 0 < |S_{\mu\nu}| < 1 & \mathrm{Tr}\{S\} = M \end{array} \right\} \tag{1.51}$$

[4] Greek indices refer to atomic orbitals, whereas Roman indices refer to molecular orbitals.

In order to make the link between the density matrix formalism described above and programming, we shall recast Eq. (1.48) using a more suitable form. Assuming that we already know the LCAO coefficients for each $|\psi_a\rangle$, we generally defined a coefficient matrix, in our case $C \in \mathbb{R}^{M\times M}$, using the format described below:

$$C \equiv \overset{\psi_1 \quad \psi_N \; \psi_{N+1} \quad \psi_M}{\begin{pmatrix} \vert & & \vert & \vert & & \vert \\ \vert & \cdots & \vert & \vert & \cdots & \vert \\ \vert & & \vert & \vert & & \vert \end{pmatrix}} \begin{matrix} \\ \\ \\ \underbrace{\qquad\qquad}_{\text{occupied}} \; \underbrace{\qquad\qquad}_{\text{unoccupied}} \end{matrix} \tag{1.52}$$

where C collects the M coefficients (sorted in column) of the M eigenvectors (sorted in row). Consequently, the *one-particle* density matrix is expressed as

$$D = C\mathcal{O}C^\dagger \tag{1.53}$$

where $\mathcal{O}$ is the matrix of the *particle* occupation numbers,

$$\mathcal{O} := \mathrm{diag}\{I_N, 0_{\bar{N}}\} \equiv \begin{matrix} \text{occupied} \\ \\ \\ \\ \text{unocuppied} \end{matrix} \overset{\qquad}{\left(\begin{array}{ccc|ccc} 1 & & 0 & & & \\ & \ddots & & & 0 & \\ 0 & & 1 & & & \\ \hline & & & 0 & & 0 \\ & 0 & & & \ddots & \\ & & & 0 & & 0 \end{array}\right)} \tag{1.54}$$

with $\bar{N}$ the number of unoccupied states, such that: $M = N + \bar{N}$. Conversely, one can also define the *one-hole* density matrix built from the set of unoccupied eigenstates. By analogy with Eq. (1.53), this yields to introduce

$$\bar{D} = C\bar{\mathcal{O}}C^\dagger \tag{1.55}$$

where $\bar{\mathcal{O}}$ is the matrix of the *hole* occupation numbers,

$$\bar{\mathcal{O}} := \mathrm{diag}\{0_N, I_{\bar{N}}\} \equiv \begin{array}{c} \\ \text{occupied} \\ \\ \\ \\ \\ \\ \\ \text{unocuppied} \end{array} \begin{pmatrix} 0 & & 0 & & \\ & \ddots & & & \\ 0 & & 0 & & \\ & & & 1 & & 0 \\ & 0 & & & \ddots & \\ & & & 0 & & 1 \end{pmatrix} \tag{1.56}$$

It is worth to emphasize that $\mathcal{O}$ and $\bar{\mathcal{O}}$ are the matrix representations of the one-particle and one-hole density operator, respectively, in the molecular orbitals basis. Using this representation, the idempotency property of Eq. (1.50) writes:

$$\begin{aligned} D = DSD = C\mathcal{O}C^\dagger SC\mathcal{O}C^\dagger \qquad (1.57) \\ \text{subject to:} \quad C^\dagger SC = I \end{aligned}$$

1.3 Restricted Hartree-Fock energy

In restricted Hartree-Fock (RHF) theory applied to closed-shell systems, the electronic energy of Eq. (1.42) can be recast in the following form

$$\mathcal{E}_{\mathrm{HF}} = \mathrm{Tr}\{D(2h + G)\} \tag{1.58}$$

$$= \mathrm{Tr}\{D(h + F)\} \tag{1.59}$$

with h and G the one-electron and two-electron contributions. In Eq. (1.59) we have expressed the HF energy in terms of the Fock matrix, $F := h + G$ (cf. Section 1.5). The matrix elements of the one-particle Hamiltonian (also called the core hamiltonian), within the AO basis, are defined according to

$$h_{\mu\nu} = \langle \mu | \hat{h} | \nu \rangle \tag{1.60a}$$

$$\text{with:} \quad \hat{h} = -\frac{1}{2}\nabla^2 - \sum_{A=1}^{K} \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} \tag{1.60b}$$

where we recognize the kinetic energy operator and the external potential created by the K nuclei, respectively. The matrix elements of G are given by:

$$G_{\mu\nu} = \langle\mu|\hat{V}_{\text{HF}}|\nu\rangle \tag{1.61a}$$

$$\text{with:}\quad \hat{V}_{\text{HF}} = \sum_{\eta,\kappa} D_{\eta\kappa}\left[2\int d\mathbf{r}_2\,\frac{\phi_\eta^*(\mathbf{r}_2)\phi_\kappa(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|} - \int d\mathbf{r}_2\,\frac{\phi_\eta^*(\mathbf{r}_2)\mathscr{P}_{12}\phi_\kappa(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|}\right] \tag{1.61b}$$

where $\mathscr{P}_{12}$ is the 12-permutation operator[5] already introduced in Eq. (1.17). The first and second term in Eq. (1.61b) are easily recognized as the Coulomb and exchange operators:

$$\hat{J} := \sum_{\eta,\kappa}\int d\mathbf{r}_2\,\frac{\phi_\eta^*(\mathbf{r}_2)D_{\eta\kappa}\phi_\kappa(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|} \tag{1.62a}$$

$$\hat{K} := \sum_{\eta,\kappa}\int d\mathbf{r}_2\,\frac{\phi_\eta^*(\mathbf{r}_2)D_{\eta\kappa}\mathscr{P}_{12}\phi_\kappa(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|} \tag{1.62b}$$

such that:

$$\hat{V}_{\text{HF}} = 2\hat{J} - \hat{K}$$

We emphasize that the Coulomb operator of Eq. (1.62a) can be re-written in terms of the one-particle electron density $\rho(\mathbf{r})$, according to

$$2\hat{J} = \int d\mathbf{r}_2\,\frac{\rho(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|} \tag{1.63a}$$

$$\text{with:}\quad \rho(\mathbf{r}) = 2\sum_{\eta,\kappa} D_{\eta\kappa}\phi_\eta^*(\mathbf{r})\phi_\kappa(\mathbf{r}) \tag{1.63b}$$

If we look for expressions of G matrix elements, by inserting Eqs. (1.62a) and (1.62b) into Eq. (1.61b), we obtain:

$$G_{\mu\nu} = 2J_{\mu\nu} - K_{\mu\nu}$$

with:

$$J_{\mu\nu} = \langle\mu|\hat{J}|\nu\rangle = \sum_{\eta,\kappa}\int\int d\mathbf{r}_1\,d\mathbf{r}_2\,\frac{\phi_\mu^*(\mathbf{r}_1)\phi_\eta^*(\mathbf{r}_2)D_{\eta\kappa}\phi_\kappa(\mathbf{r}_2)\phi_\nu(\mathbf{r}_1)}{|\mathbf{r}_1-\mathbf{r}_2|} \tag{1.64a}$$

$$K_{\mu\nu} = \langle\mu|\hat{K}|\nu\rangle = \sum_{\eta,\kappa}\int\int d\mathbf{r}_1\,d\mathbf{r}_2\,\frac{\phi_\mu^*(\mathbf{r}_1)\phi_\eta^*(\mathbf{r}_2)D_{\eta\kappa}\phi_\kappa(\mathbf{r}_1)\phi_\nu(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|} \tag{1.64b}$$

We shall introduce some of the commonly used notations in Chemistry for the matrix elements of the Coulomb and exchange operators. They are expressed in terms of the

[5] The permutation operator interchanges the coordinate of two electrons. When applied to the right of the atomic orbitals product, this gives: $\mathscr{P}_{12}\phi_\mu(\mathbf{r}_2)\phi_\nu(\mathbf{r}_1) = \phi_\mu(\mathbf{r}_1)\phi_\nu(\mathbf{r}_2)$.

electron repulsion integrals (ERI). Following the Chemists' notation, the formal expression of an ERI is given according to

$$(\mu\nu|\eta\kappa) := \int\int d\mathbf{r}_1\, d\mathbf{r}_2\, \frac{\phi^*_\mu(\mathbf{r}_1)\phi_\nu(\mathbf{r}_1)\phi^*_\eta(\mathbf{r}_2)\phi_\kappa(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{1.65}$$

where the coordinate of the electron 1 and 2 appear side by side with respect to the vertical bar in $(\mu\nu|\eta\kappa)$. If we identify Eq. (1.65) to be the Coulomb integral as found in Eq. (1.64a), the exchange integral appearing in Eq. (1.64b) writes

$$(\mu\kappa|\eta\nu) := \int\int d\mathbf{r}_1\, d\mathbf{r}_2\, \frac{\phi^*_\mu(\mathbf{r}_1)\phi_\kappa(\mathbf{r}_1)\phi^*_\eta(\mathbf{r}_2)\phi_\nu(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{1.66}$$

As a result, the G matrix elements are defined according to

$$G_{\mu\nu} = \sum_{\eta,\kappa} D_{\eta\kappa}\left[2(\mu\nu|\eta\kappa) - (\mu\kappa|\eta\nu)\right] \tag{1.67}$$

On including the one-electron Hamiltonian in Eq. (1.67), we obtain the following expression for the Fock matrix elements:

$$F_{\mu\nu} = h_{\mu\nu} + \sum_{\eta,\kappa} D_{\eta\kappa}\left[2(\mu\nu|\eta\kappa) - (\mu\kappa|\eta\nu)\right] \tag{1.68}$$

We shall briefly review other commonly found definitions of the electronic energy since it can be confusing in literature. From Eq. (1.58), the RHF energy can be alternatively expressed as

$$\mathcal{E} = \mathrm{Tr}\{D(2H + G)\} = 2\mathrm{Tr}\{D(H + \frac{1}{2}G)\} \tag{1.69}$$

By introducing the *bond-order* matrix, $P := 2D$, we may write

$$\mathcal{E} = \mathrm{Tr}\{P(H + \frac{1}{2}G)\} = \frac{1}{2}\mathrm{Tr}\{P(H + F)\} \tag{1.70a}$$

$$\text{with:}\quad G_{\mu\nu} = \sum_{\eta,\kappa} P_{\eta\kappa}\left((\mu\nu|\eta\kappa) - \frac{1}{2}(\mu\kappa|\eta\nu)\right) \tag{1.70b}$$

If we briefly review this Section, we observed that the construction for the Fock matrix can be a serious bottleneck for large scale calculations. Since the AOs extend over the whole molecule, the amount of information in each AO is proportional to N.[63] Based on a naive analysis of Eqs. (1.60a), (1.64a) and (1.64b), we found that building the one-core Hamiltonian involved integrating the product of two AOs scales as $O(N^2)$.

While with the product of four AOs, the Coulomb and the exchange matrices scale as $O(N^4)$. Therefore, the cost of the Fock matrix construction is basically $O(N^4)$. For large molecular systems, relying on some distance criteria, the scaling can be reduced to reach asymptotically N^2. In order to calculate the Coulomb and exchange integrals in full linear scaling, specific numerical techniques for GTOs have been developped,[64, 8] such as the continuous fast multipole method[33] for the Coulomb integrals, and the LinK method[36, 65] for the exchange integrals.

1.4 Pariser-Parr-Pople method

The entire implementation and all the applications performed during this thesis were based on a semi-empirical method derived for the calculation of energetics and properties of π-conjugated systems. In regard to the all-electron HF approaches based on non-orthogonal extended local basis sets, this choice permits to focus our efforts mainly on algorithmic developments related to the density matrix solvers, the computational ressources used for evaluating the matrices being negligeable in that case. It has also the merit to overcome the intricacies encountered when trying to modify some of the routines found in standard quantum chemistry packages. Using such kind of simplified HF model, we were able to performed a fast and fair comparison between various methods, while making sure that coding was optimized for all of them. For these reasons, we have considered the most simple semi–empirical HF model, the Pariser-Parr-Pople method[66–68] (PPP), which was originally developed for treating conjugated hydrocarbons.

1.4.1 Zero-differential-overlap approximation

For atomic-like basis functions centered on atoms, ie. which are explicilty dependent on the nucleus coordinates, such as Slater-type orbital (STO) or Gaussian-type orbitals (GTO) —commonly used in quantum chemistry—, the ERI of Eq. (1.65) explicitly writes as:

$$\int\int d\mathbf{r}_1\, d\mathbf{r}_2\, \frac{\phi^*_\mu(\mathbf{r}_1-\mathbf{R}_\mu)\phi_\nu(\mathbf{r}_1-\mathbf{R}_\nu)\phi^*_\eta(\mathbf{r}_2-\mathbf{R}_\eta)\phi_\kappa(\mathbf{r}_2-\mathbf{R}_\kappa)}{|\mathbf{r}_1-\mathbf{r}_2|} \tag{1.71}$$

where $\{\mathbf{R}_\mu, \mathbf{R}_\nu, \mathbf{R}_\eta, \mathbf{R}_\kappa\}$ are the nucleus Cartesian coordinates. The zero-differential-overlap (ZDO) approximation consists in neglecting ERIs containing product $\phi_\mu(\mathbf{r}_1 - \mathbf{R}_\mu)\phi_\nu(\mathbf{r}_1 - \mathbf{R}_\nu)$ where $\mu \neq \nu$, ie. ERIs are assumed to be zero for pair-densities not centered on the same nucleus. Put in other words,

$$(\mu\nu|\eta\kappa) = (\mu\mu|\eta\eta)\,\delta_{\mu\nu}\delta_{\eta\kappa} \tag{1.72}$$

with[6]

$$(\mu\mu|\eta\eta) \equiv \int\int d\mathbf{r}_1\, d\mathbf{r}_2\, \frac{|\phi_\mu(\mathbf{r}_1 - \mathbf{R}_\mu)|^2\, |\phi_\eta(\mathbf{r}_2 - \mathbf{R}_\eta)|^2}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{1.73}$$

Note that the Kronecker functions in Eq. (1.72), implies that the basis set is orthonormal, such that:

$$\langle\mu|\nu\rangle \quad \begin{cases} 1 & \text{if } \mu = \nu \\ 0 & \text{if } \mu \neq \nu \end{cases} \tag{1.74}$$

On introducing Eq. (1.72) into (1.68), we obtain:

$$F_{\mu\nu} = h_{\mu\nu} + \sum_{\eta,\kappa} D_{\eta\kappa}\, [2(\mu\mu|\eta\eta)\delta_{\mu\nu}\delta_{\eta\kappa} - (\mu\mu|\eta\eta)\delta_{\mu\kappa}\delta_{\eta\nu}] \tag{1.75}$$

which, without loss of generality, simplifies to

$$F_{\mu\nu} = h_{\mu\nu} + 2\sum_{\eta} D_{\eta\eta}(\mu\mu|\eta\eta)\delta_{\mu\nu} - \sum_{\eta,\kappa} D_{\nu\nu}(\mu\mu|\eta\eta)\delta_{\mu\kappa}\delta_{\eta\nu} \tag{1.76a}$$

for: $\mu = \nu$

$$F_{\mu\mu} = h_{\mu\mu} + 2\sum_{\nu} D_{\nu\nu}(\mu\nu|\mu\nu) - D_{\mu\mu}(\mu\mu|\mu\mu) \tag{1.76b}$$

for: $\mu \neq \nu$

$$F_{\mu\nu} = h_{\mu\nu} - D_{\nu\mu}(\mu\nu|\mu\nu) \tag{1.76c}$$

In literature, we identify:

$$\Gamma_{\mu\mu} := (\mu\mu|\mu\mu)$$
$$\Gamma_{\mu\nu} := (\mu\nu|\mu\nu)$$

which are the one-center integral corresponding to an energy constant, and the two-electron repulsion integral which is parametrized with respect to a set of internal geometric parameters. As a result, the ZDO approach greatly simplifies the problem at the cost of a parametrization, which reduces the *ab initio* character of the Hartree-Fock method.

1.4.2 Pariser-Parr-Pople model parameterization

Several parameterizations of the PPP model can be found in literature.[69–72] In this work we haved used the Ohno's parameterization[70] using standard parameters[73, 74] collected in Table {1.1}. HF-PPP self-consistent field (SCF) calculation (*vide infra*) is initiated after a first tight-binding (TB) calculation. The TB matrix elements are the on-site energy $t_{\mu\mu}$, and the hopping term $t_{\mu\nu}$ which is assigned with respect to the C−C

[6]Note that: $(\mu\mu|\eta\eta)\delta_{\mu\nu}\delta_{\eta\kappa} = (\nu\nu|\eta\eta)\delta_{\mu\nu}\delta_{\eta\kappa} = (\nu\nu|\kappa\kappa)\delta_{\mu\nu}\delta_{\eta\kappa} = (\mu\mu|\kappa\kappa)\delta_{\mu\nu}\delta_{\eta\kappa}$

	TB	PPP	core Hamiltonian
on-site	$t_{\mu\mu} = 0$	$\Gamma_{\mu\mu} = 11.130$	$h_{\mu\mu} = t_{\mu\mu} - \sum_{\nu\neq\mu} \Gamma_{\mu\nu}$
off-site	$t_{\mu\nu} = \begin{cases} -2.568 & \text{if } r_{\mu\nu} \leq R_d \\ 0 & \text{else} \end{cases}$	$\Gamma_{\mu\nu} = \frac{\Gamma_{\mu\mu}}{\sqrt{1+\left(\frac{r_{\mu\nu}}{1.2786}\right)^2}}$	$h_{\mu\nu} = t_{\mu\nu}$

Table 1.1 Ohno parametrization. $t_{\mu\mu}$ and $t_{\mu\nu}$ are the on-site and off-site energy terms (in eV). $t_{\mu\nu}$ is assigned with respect to a distance criteria R_d.

distance of the first nearest neighbour(s). Then, the electron-nuclei interaction is added to the TB part in order to give the final one-electron Hamiltonian matrix elements $h_{\mu\nu}$. In the Table {1.1}, the electron-nuclei interaction is the second term in $h_{\mu\mu}$. Finally, the two-electron contribution is constructed with the density matrix and the elements $\Gamma_{\mu\mu}$ and $\Gamma_{\mu\nu}$, which added to the core Hamiltonian, leads to the Fock matrix elements of Eq. (1.76).

1.5 Minimization of the Hartree-Fock energy

Let us consider a Hamiltonian $\hat{\mathcal{H}}$ in an infinite dimensional Hilbert space, and let us assume that we know the eigenstates $\mathcal{S}_{\mathcal{H}} := \{\mathcal{E}, |\Psi\rangle\}$. In quantum mechanics, the variational principle tells us that for any eigenstate, $|\Psi\rangle \in \mathcal{S}_{\mathcal{H}}$, the expectation value is an upper bound of the exact ground state energy $\mathcal{E}_0$ associated with the ket $|\Psi_0\rangle$, ie. $\langle\Psi|\hat{\mathcal{H}}\Psi\rangle := \mathcal{E}[\Psi] \geq \mathcal{E}_0 =: \langle\Psi_0|\hat{\mathcal{H}}\Psi_0\rangle$. Interestingly for practical calculations, this principle can be extended to any approximate state $|\tilde{\Psi}\rangle$ in a subspace of $\mathcal{S}_{\mathcal{H}}$. For instance, within the Hartree-Fock approximation, we shall have:

$$\mathcal{E}[\tilde{\Psi}_{\text{HF}}] \geq \mathcal{E}_{\text{HF}} > \mathcal{E}_0 \tag{1.78}$$

where $\mathcal{E}_{\text{HF}}$ [cf. Section 1.2.3] is the exact energy of a single antisymmetrized product of one-electron functions, ie. Ψ_{HF}, expanded over a finite basis. As a consequence, the variational principle is the basis for the minimization principle which aims to find the best approximate wavefunction for the ground state verifying Eq. (1.78). We can easily establish a *constrained* minimization principle from the first inequality[7] of Eq. (1.78) and

[7] This is also clearly apparent from Eq. (1.3).

the normalization condition (1.4), according to

$$\min_{\Psi_{\mathrm{HF}}} \left\{ \langle \Psi_{\mathrm{HF}} | \hat{\mathcal{H}} \Psi_{\mathrm{HF}} \rangle \;\; | \;\; \langle \Psi_{\mathrm{HF}} | \Psi_{\mathrm{HF}} \rangle = 1 \right\} \tag{1.79}$$

This kind of constrained minimization problem can be solved by using the Lagrange multiplier technique. On introducing the wavefunction-based Hartree-Fock Lagrangian,

$$\mathcal{L}_{\mathrm{HF}}[\Psi_{\mathrm{HF}}] := \langle \Psi_{\mathrm{HF}} | \hat{\mathcal{H}} \Psi_{\mathrm{HF}} \rangle - \mathcal{E}_{\mathrm{HF}} \left(\langle \Psi_{\mathrm{HF}} | \Psi_{\mathrm{HF}} \rangle - 1 \right) \tag{1.80}$$

where the Hartree-Fock energy $\mathcal{E}_{\mathrm{HF}}$ is recognized as the Lagrange multiplier. Basic calculus of variations applied to this equation leads to the analogue of the time-independent Schrödinger (1.5) in matrix form [cf. Section 1.2.4]. Instead, we may choose an alternative route for solving this problem using the time-independent Liouville-von Neumann equation (1.11c) along with matrix algebra. Nevertheless, in that case, we have to define necessary and sufficient conditions in order to derive a minimization principle based uniquely on $\mathcal{D}_{\mathrm{HF}} := |\Psi_{\mathrm{HF}}\rangle \langle \Psi_{\mathrm{HF}}|$, which leads to analogue of Eq. (1.79) and ensures the uniqueness of the solution. These conditions are the N-representability conditions for a pure state introduced in Section 1.2.3. As a result, we can introduce the following minimization principle in matrix form:

$$\min_{\mathcal{D}_{\mathrm{HF}}} \left\{ \mathrm{Tr}\{\hat{\mathcal{H}} \mathcal{D}_{\mathrm{HF}}\} \;\; | \;\; \mathrm{Tr}\{\mathcal{D}_{\mathrm{HF}}^2\} = \mathrm{Tr}\{\mathcal{D}_{\mathrm{HF}}\}, \quad \mathrm{Tr}\{\mathcal{D}_{\mathrm{HF}}\} = 1 \right\} \tag{1.81}$$

which translates to density matrix-based Hartree-Fock Lagrangian, according to

$$\mathcal{L}_{\mathrm{HF}}[\mathcal{D}_{\mathrm{HF}}] := \mathrm{Tr}\{\hat{\mathcal{H}} \mathcal{D}_{\mathrm{HF}}\} - \Big(\mathrm{Tr}\{\Gamma(\mathcal{D}_{\mathrm{HF}}^2 - \mathcal{D}_{\mathrm{HF}})\} + \gamma(\mathrm{Tr}\{\mathcal{D}_{\mathrm{HF}}\} - 1) \Big) \tag{1.82}$$

where Γ ($\in \mathbb{R}^{M \times M}$) and γ ($\in \mathbb{R}$) are Lagrange multipliers. Since the HF energy is a functional of the one-particle density matrix only [cf. Eqs. (1.42) and (1.45)], from the expression (1.69) and assuming an orthonormal basis set ($S = I$), Eq. (1.82) reduces to

$$\mathcal{L}_{\mathrm{HF}}[D] := 2\mathrm{Tr}\{D(h + \frac{1}{2} G(D))\} - 2 \Big(\mathrm{Tr}\{\Lambda(D^2 - D)\} + \mu(\mathrm{Tr}\{D\} - N) \Big) \tag{1.83}$$

where the matrix of Lagrange multipliers Λ and the scalar μ have been introduced to constrained idempotency and trace conservation, respectively. From here, we shall search for minimizing such functional with respect to D. This yields to solve:

$$\nabla \mathcal{L}_{\mathrm{HF}}[D] = 2(h + G(D) - \Lambda D - D\Lambda + \Lambda - \mu I) = 0 \tag{1.84}$$

where we have used the following functional derivative properties,[8]

$$\begin{aligned}\nabla \mathrm{Tr}\{DA\} &= A^\dagger \\ \nabla \mathrm{Tr}\{D^2 A\} &= (DA + AD)^\dagger \\ \nabla \mathrm{Tr}\{DG(D)\} &= 2G(D)^\dagger\end{aligned}$$

In the last statement, we used $\mathrm{Tr}\{XG(Y)\} = \mathrm{Tr}\{YG(X)\}$[76, 77] for $X = Y = D$. On recalling that all the operators are Hermitian —more specifically in this work all the matrices are symmetric—, the following working equation is found

$$F(D) - \mu I = \Lambda D + D\Lambda - \Lambda \tag{1.85}$$

where, $F = h + G$, is the Fock matrix already introduced in Section 1.3. It should be outlined that despite the appealing form of this equation, to our knowledge, there is only one paper dealing with its resolution.[78]

To demonstrate that solving Eq. (1.85) leads to an unique solution decribing a pure state within the *NVE* ensemble, we may try to recover the famous Roothaan-Hall equation[79, 80] widely used in quantum chemistry. By taking the commutator of Eq. (1.85) with respect to D, we obtain:

$$[D, F - \mu I] = D\Lambda D + D^2\Lambda - D\Lambda - \Lambda D^2 - D\Lambda D + D\Lambda \tag{1.86}$$

If the density matrix is exactly idempotent, the above equation reduced to:

$$FD = DF \tag{1.87}$$

which is the single-determinant time-independent Liouville-von Neumann equation in matrix form. For instance, multiplying on the right by the coefficient matrix C and using the definition of Eq. (1.53), we have

$$FDC = DFC \tag{1.88}$$

$$FC\mathcal{O}C^\dagger C = C\mathcal{O}C^\dagger FC \tag{1.89}$$

$$FC\mathcal{O} = C\mathcal{O}E \tag{1.90}$$

[8] At first sight, passing from Eq. (1.83) to Eq. (1.84) might not be that straightforward. Given a functional $\mathscr{F}$, such that $\mathscr{F} : \mathbb{R}^{M\times M} \mapsto \mathbb{R}$, the variation of $\mathscr{F}(X)$ with respect to X is formally given by: $\delta\mathscr{F}(X)/\delta X \equiv \nabla\mathscr{F}(X) = f(X)^\dagger$, where f is the scalar derivative of $\mathscr{F}$.[2, 75]

where, for an orthogonal basis set, E is a diagonal matrix containing the M eigenvalues of F.[9] The presence of occupation number matrix $\mathcal{O}$ indicates that Eq. (1.90) gives access only to the eigenvalues of occupied states. Based on symmetry considerations, it is easily proved that Eq. (1.87) holds as well for the one-hole density matrix [cf. Eq. (1.55)]. As a result, the eigenvalues of the unocuppied states can be obtained by solving

$$FC\bar{\mathcal{O}} = C\bar{\mathcal{O}}E \tag{1.91}$$

On assembling Eqs. (1.90) and (1.91), we obtain the condensed matrix form

$$F(D)C - CE = 0 \tag{1.92a}$$

$$\text{subject to:} \quad C^\dagger C = I \tag{1.92b}$$

which will be referred as the Roothaan-Hall equation.[10] On multiplying Eq. (1.92a) from the right by $C^\dagger$ and using Eqs. (1.90) and (1.91), the Fock matrix reads

$$F = C\mathcal{O}EC^\dagger + C\bar{\mathcal{O}}EC^\dagger \tag{1.93}$$

Hence, the spectrum of the Fock matrix can be resolved according to Eq. (1.13c), that is

$$F = \sum_i \epsilon_i D_i + \sum_j \bar{\epsilon}_j \bar{D}_j \tag{1.94a}$$

$$\text{subject to:} \quad \sum_{i=1}^{N} D_i = D, \quad \text{and} \quad \sum_{j=1}^{\bar{N}} \bar{D}_j = \bar{D} \tag{1.94b}$$

with i and j running over the energy-weighted projectors for the occupied and unoccupied subspace, respectively.

[9] For a given symmetric (Hermitian) non-degenerate Fock martrix $F \in \mathbb{R}^{M\times M}$ ($\in \mathbb{C}^{M\times M}$), there always exists a similarity transformation, such that: X^tFX ($X^\dagger FX$) $= \text{diag}\{\epsilon_1\epsilon_2\cdots\epsilon_M\}$, where $\{\epsilon_i\}_{i=1}^M$ are the (real) eigenvalues of F, and the transformation matrix X is orthogonal (unitary), ie. $X^tX = X^{-1}X = I$ ($X^\dagger X = X^{-1}X = I$). From the definition (1.52) of the coefficient matrix C, it is obvious that $X \equiv C$.

[10] Indeed, the Roothaan-Hall equation[79, 80] corresponds to a generalized eigenvalue problem $FC = SCE$ deriving from the HF equations, expressed in a non-orthogonal basis. Nevertheless, we will retain this naming convention.

1.6 The self-consistent field procedure

Since the eigenvalue problem of Eqs. (1.92) is *non-linear* with respect to the density matrix, the Roothaan-Hall equations are resolved iteratively using a two-step approach: (i) solve a *linear* eigenvalue problem for a fixed Fock matrix, (ii) update the new Fock matrix from the previous solutions. Iterations are repeated until some convergence criteria is met. This procedure, called self-consistent field (SCF), is illustrated in Figure {1.1}. The update of the Fock matrix at iteration $n+1$, from the Fock matrix at iteration n,

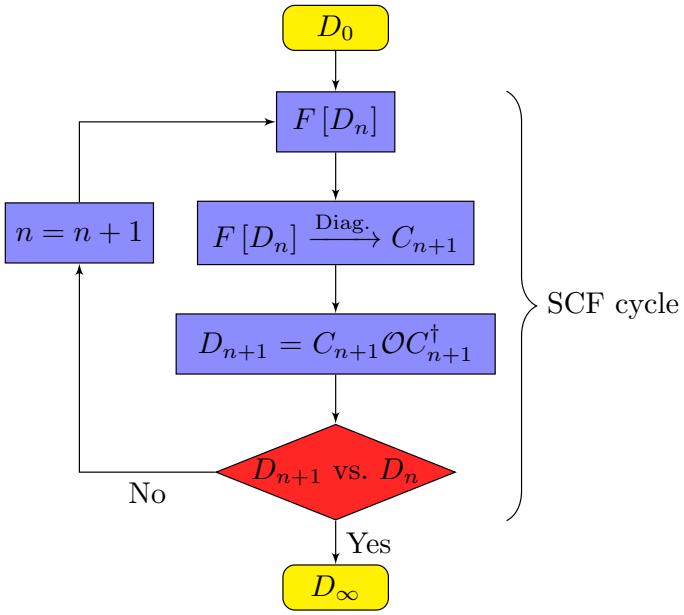


Fig. 1.1 Flow diagram of the SCF processus.

through the matrix D_n, constitutes a SCF cycle. The initial guess and the converged coefficient matrix in Figure {1.1} are designated by D_0 and D_∞, respectively.

The intial guess is one of the key steps of the SCF procedure which may have a strong impact on the convergence rate. A poor or wrong initial guess can slow down the convergence, or even worse, to a divergence. The former is generally related to oscillations when approaching the final state so that it can not be reached with a reasonable number of iterations. The latter indicates that the initial guess has no physical significance or is too far away from the expected solution. There are different ways to define the starting guess. In this work we merely start from the solutions of a tight-binding calculation, as decribed in Section 1.4.2. The fact still remains that, even a good initial guess does not prevent convergence instabilities. For that reason, numerous suggestions[81–85] have been made to solve these issues. Many of those were combined into hybrid methods

attempting to overcome the weakness of the standalone models[82, 83]. For our purpose here, we recall two schemes which consist in optimizing the Fock matrix construction in order to stabilize and accelerate the SCF convergence.

1.6.1 Constant damping algorithm

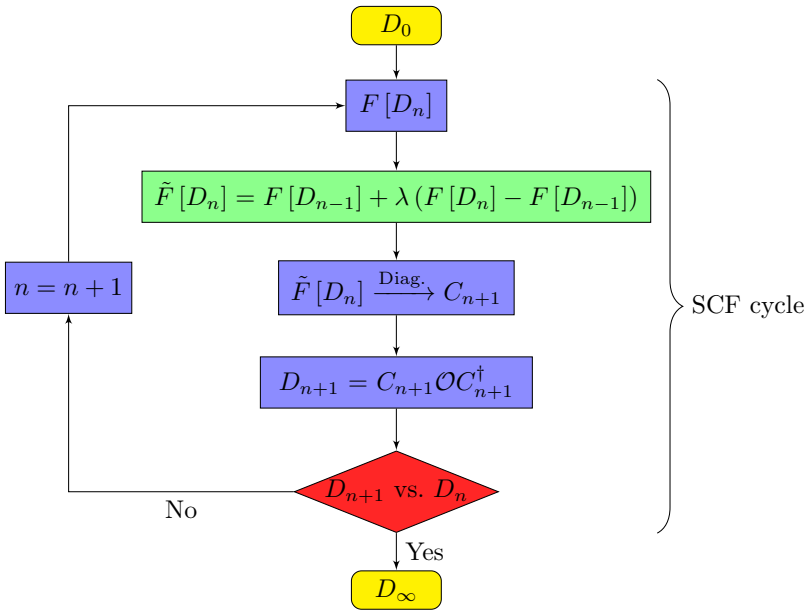


Fig. 1.2 Flow diagram of the SCF scheme including the CDA.

The constant damping algorithm[86] (CDA) is a simple method based on a linear interpolation between the *past* and *futur* events, that is, the Fock matrix at iteration n is defined according to

$$\tilde{F}_n := \lambda F_n + (1 - \lambda)F_{n-1}, \tag{1.95}$$

where λ, referring to a damping factor, is a constant chosen freely in the interval $[0, 1]$. The damping step is outlined by the green chart in Figure {1.2} which can be compared to algorithm of Figure {1.1}. The main drawback with the CDA, is that the convergence rate is now fixed by the value of λ. In other words, having chosen heuristically λ to initiate the SCF processus, if for some reasons convergence problem persists, one has to stop and restart the processus with another damping parameter. It is true that some approaches were proposed to dynamically optimize the damping factor during the SCF[87, 84]. It remains that using a constant or dynamic approach, the CDA is not always successful for solving convergence issues.

1.6.2 Direct inversion of the iterative subspace extrapolation

Another more general technique to speed up and stabilize the SCF convergence is the direct inversion of the iterative subspace (DIIS) extrapolation proposed by Pulay[88–90]. This method and several improvements[82, 83] have shown to be very efficient[82, 83] and it constitutes one of our ingredient for efficient calculation of the response properties presented in Chapter 3. The idea of the DIIS is to extrapolate the Fock matrix at iteration n, from a linear combination of Fock matrices taken in the history of the SCF procedure according to

$$\tilde{F}_n := \sum_{i=n-m}^{n} c_i F_i \tag{1.96}$$

where m is the size of the set of historical Fock matrices and n is the iteration from which the DIIS is switched on. The coefficients of the linear combination are determined using a set of Pulay's error vectors, $\{e_i\}$, corresponding to $\{F_i, D_i\}$, according to $e_i = [F_i, D_i]$. The DIIS approach assumes that this linear combination is a good approximation of the

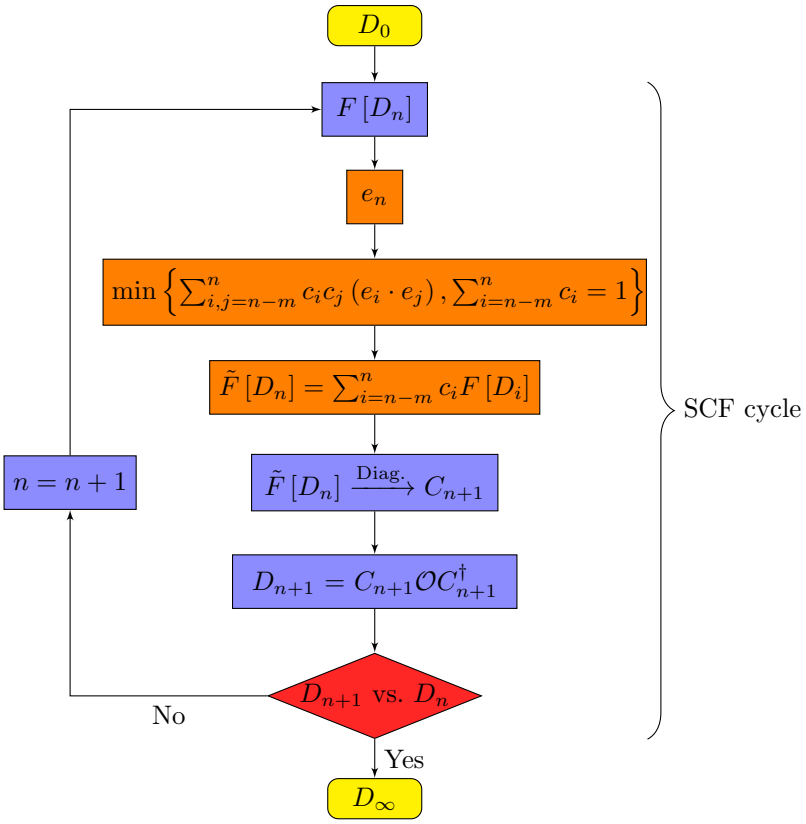


Fig. 1.3 Flow diagram of the SCF scheme including the DIIS.

converged Fock matrix, symbolized here by F_∞. Then, it supposes that each trial Fock

matrix is the sum of the exact solution and an error vector $\mathbf{e}_i$. Therefore, Eq. (1.96) becomes

$$\begin{aligned}\tilde{F}_n &= \sum_{i=n-m}^{n} c_i \left(F_\infty + e_i\right) \\ &= F_\infty \sum_{i=n-m}^{n} c_i + \sum_{i=n-m}^{n} c_i e_i \end{aligned} \tag{1.97}$$

From this equation, in order to have $\tilde{F}_n = F_\infty$, we need to minimize the norm of the second term while requiring the sum in the first term to be normalized, which gives[11]

$$\min \left\{ \| \sum_{i=n-m}^{n} c_i e_i \|^2 \,\middle|\, \sum_{i=n-m}^{n} c_i = 1 \right\} \tag{1.98}$$

where the coefficients $\{c_i\}$ are assumed real, ie. $c_i^* = c_i$. As for the minimization of the Hartree-Fock energy developped in Section 1.5, we can use the Lagrange multiplier technique to minimize the expression of Eq. (1.98). This yields to

$$\mathcal{L}_{\text{DIIS}} := \sum_{i,j=n-m}^{n} c_i c_j B_{ij} - \lambda \left(\sum_{i=n-m}^{n} c_i - 1 \right) \tag{1.99}$$

where λ is the Lagrange multiplier, and $B_{ij} = (e_i \cdot e_j)$. The first derivative of $\mathcal{L}_{\text{DIIS}}$ with respect to the coefficient c_l, gives

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{DIIS}}}{\partial c_l} &= \sum_{i,j} \frac{\partial c_i}{\partial c_l} c_j B_{ij} + \sum_{i,j} c_i \frac{\partial c_j}{\partial c_l} B_{ij} - \lambda \sum_i \frac{\partial c_i}{\partial c_l} \\ &= \sum_{i,j} \delta_{il} c_j B_{ij} + \sum_{i,j} c_i \delta_{jl} B_{ij} - \lambda \sum_i \delta_{il} \\ &= \sum_j c_j B_{lj} + \sum_i c_i B_{il} - \lambda \end{aligned} \tag{1.100}$$

where, for the last step, we used the following properties

$$\frac{\partial c_i}{\partial c_j} = \delta_{ij}, \quad \sum_i \delta_{ij} = 1, \quad \sum_i c_i \delta_{ij} = c_j \tag{1.101}$$

[11] We recall that the first term within the braces is the expression to be minimized, and the second term is the constraint ; $\| \cdot \|$ stands for the Frobenius norm, and $(\cdot)$ for the scalar product.

Substituting j by i in the first sum of Eq. (1.100), and using the fact that the matrix B of elements $\{B_{ij}\}$ is symmetric, Eq. (1.100) simplifies to

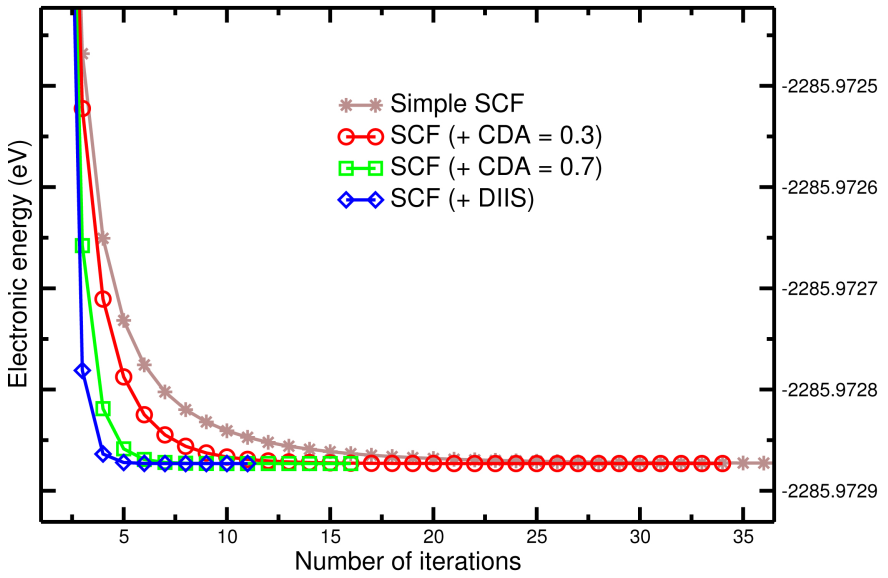
$$\frac{\partial \mathcal{L}_{\text{DIIS}}}{\partial c_l} = 2\sum_i c_i B_{il} - \lambda \tag{1.102}$$

As a result, solutions of eq. (1.98) translates to the minimization of the Lagragian $\mathcal{L}_{\text{DIIS}}$ such that

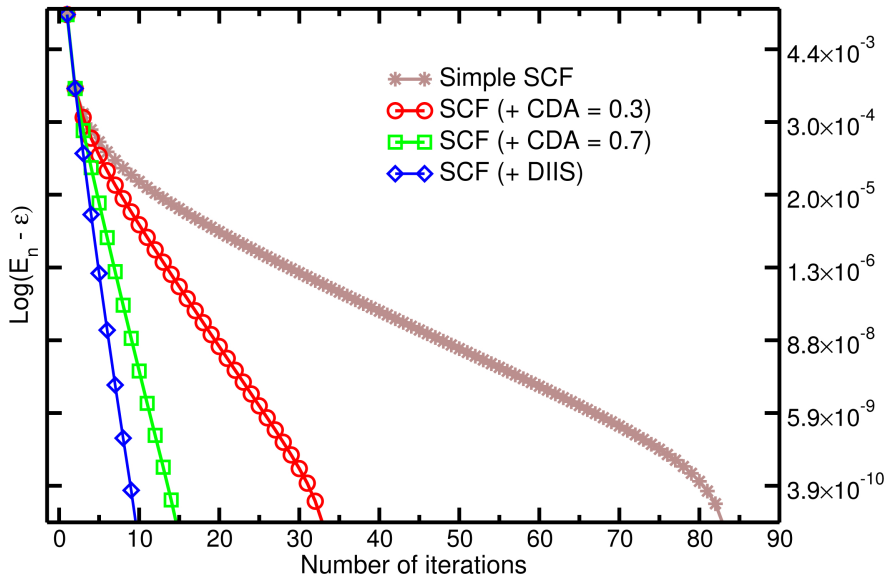
$$\sum_i c_i B_{il} - \lambda = 0 \tag{1.103}$$

The coefficients $\{c_i\}$ are finally given by the Eq. (1.103) which corresponds to a system of $(m+1)$ linear equations[91–94]. Once the coefficients are found, the Fock matrix is updated following the linear expansion of Eq. (1.96).

The DIIS steps are defined by the orange chart in the diagram of Figure {1.3}. In the particular case of $m = 2$, the DIIS corresponds to the damping scheme of Eq. (1.95) but without any restriction on λ, which means that the DIIS is a dynamical extension of the CDA where the coefficients are optimized on-the-fly. As a concrete example, in Figure {1.4} are presented the SCF convergence profiles obtained for a non-optimized procedure (the simple SCF), the CDA interpolation using two different values of the damping parameter, and the DIIS extrapolation. It is quite clear that the convergence with the DIIS is the most efficient.



(a)



(b)

Fig. 1.4 SCF convergence profiles obtained for the carbon nanotube (11,5) using a simple approach, the CDA and the DIIS optimization. In (a) is displayed the energy minimization during the iterative process. In (b) is represented the energy error during the iterative process using a logarithmic scale.

Chapter 2

Density matrix purifications and minimizations

The orthodox resolution of the SCF equations [cf. Section 1.6] as obtained from a single Slater determinant is based on the diagonalization of the Fock matrix at each iteration. The diagonalization step is well known for being an expensive task, which becomes rapidly the limiting step for large scale calculations. For this reason, it has been suggested that one can solve the SCF equations relying only on the one-particle density matrix.[41] Even if, in their native forms, density matrix (DM) methods also present an asymptotical cubic scaling, they constitute the first ingredient towards linear scaling regime. Density matrix solvers can be classified following the physical motivations they originated from. These solvers are: (i) the iterative density matrix functional minimizations[40, 95–101] where for one-determinant SCF theories, the HF or KS energy functional is minimized with respect to an auxiliary density matrix used in place of the conventional fixed DM built from the eigenvectors, (ii) the recursive density matrix polynomial expansion where the Fermi-Dirac ground state DM at the zero electronic temperature limit is obtained by a recursive application of projection polynomials —also referred to as purifications.[102–114]

2.1 Density matrix minimization principle

In continuation of Section 1.5, where we have shown that the minimization of the Hartree-Fock (HF) energy functional can be expressed in terms of the one-particle density matrix only, we can either try to minimize the HF Lagrangian of Eq. (1.83) by releasing some of the constraints, or by introducing another objective functional to minimize.

2.1.1 Idempotency error functional minimization

In order to obtain an exactly idempotent density matrix from a roughly idempotent initial guess, McWeeny has proposed to minimize the sum of the squares of the idempotency errors, that is, $\|D^2 - D\|^2$, using a steepest gradient descent method.[41, 42] This is fully equivalent to minimize the following functional:

$$\Omega_{\mathrm{McW}} := \mathrm{Tr}\{(D^2 - D)^2\} \tag{2.1}$$

Using the trace algebra summarized in Section 1.5, the gradient for this functional is given according to

$$\nabla\Omega_{\mathrm{McW}} = 2(2D^3 - 3D^2 + D) \tag{2.2}$$

The optimal step length γ for the line search of the steepest descent can be derived from the Cauchy relation:

$$\gamma := \min_{\gamma} \ \mathrm{Tr}\{(D_\gamma^2 - D_\gamma)^2\} \tag{2.3}$$

where

$$D_\gamma := D - \gamma \nabla \Omega_{\mathrm{McW}} \tag{2.4}$$

Working on Eq. (2.3), to the second order in γ, the optimum value is found to be

$$\gamma = \frac{\mathrm{Tr}\{(D^2 - D)^2\}}{\mathrm{Tr}\{(D^2 - D)^2 (2D - I)^2\}} \tag{2.5}$$

On substituting $D := D' + \delta$, where D' is trully idempotent,[41] and expanding Eq. (2.5), it can be easily shown that $\gamma \simeq 1$. As a result, for a fixed step length gradient descent, Eq. (2.4) reduces to

$$D = 3D^2 - 2D^3 \tag{2.6}$$

It is drawn from Eq. (2.6) that the fixed step gradient descent gives rise to an alternative approach to obtain an idempotent DM relying on the following recursive formula:

$$D_{n+1} = 3D_n^2 - 2D_n^3 \tag{2.7}$$

where n is the iteration index. This relation is the so-called McWeeny purification. We note that, in line with Section 1.1, the term purification clearly indicates that repeated application of the polynomial (2.7) to a mixed state —the initial guess (*vide infra*) for the density matrix— transforms it into a pure state: the idempotent one-particle density matrix. It is worth emphasizing that solution to the minimization problem of Eq. (2.1) is not restricted to the (fixed step) steepest descent. One can also consider other gradient descent based algorithms such as the conjugate gradient (CG) or Newton-Raphson method, each of them coming with their own pros and cons.[107]

2.1.2 Energy functional minimization

Another way to find the density matrix is to minimize an energy functional using a conjugate gradient routine.[40, 95–99] Within the tight-binding (TB) framework, Li, Nunes and Vanderbilt (LNV) have proposed to minimize the grand potential functional[40] at the zero temperature limit as defined below,

$$\Omega_\mu := \mathcal{E}\left[D\right] - \mu N \tag{2.8}$$

where μ is the chemical potential and $\mathcal{E}[D]$ stands for one-eletron energy functional of the one-particle density matrix. Instead of working on D directly, which would have led to consider a set of Lagrange multipliers in Eq. (2.8) —cf. Eq. (1.82) of Section 1.5 and discussion therein—, LNV have considered the auxiliary DM of Eq. (2.7) allowing to introduce variational degrees of freedom within the grand potential functional of Eq. (2.8). Considering the Fock matrix as input, the LNV energy functional reads:

$$\Omega_{\text{LNV}} := \text{Tr}\{F(3D^2 - 2D^3)\} - \mu\text{Tr}\{3D^2 - 2D^3\} \tag{2.9}$$

with:

$$\nabla\Omega_{\text{LNV}} = 3(DF' + F'D) - 2(D^2F' + DF'D + F'D^2) \tag{2.10}$$

and $F' := F - \mu I$. Latter, Xu and Scuseria[115] (XS) have proposed a slight modification of the LNV functional minimization by a damping method based on updating the chemical potential value between consecutive conjugate gradient iterations. They reported an improvement in the convergence of the CG minimization.

Unfortunately, since Ω_μ (or Ω_{LNV}) is only well-defined within the μVT ensemble, unconstrained minimization of Eq. (2.9) is not expected to yield the correct number of particles unless the chemical potential is known exactly. This poses sever problems for unsymmetric cases, ie. when μ is not in (or close to) the middle of the eigenvalue spectrum, in other words, when the one-electron one-orbital picture is abandoned (*vide infra*). To correct this drawback, a more general approach was introduced by Millam and Scuseria[96] (MS), where the update of the chemical potential is constrained *via* the trace of the gradient. The advantage for this functional is to explicitly calculate the chemical potential during the conjugate gradient iterations so that the electron number is preserved. Nevertheless, this scheme implies that the chemical potential must be zero at convergence,[96] which clearly restricts its domain of applicability unless modifications within the working equations are derived.

The other major issue of LNV functional minimization is that the idempotency of the density matrix is not guaranteed. For this reason, the density matrix has to be purified by the McWeeny polynomial of Eq. (2.7) outside the conjugate gradient.[98] Extensive analysis of the LNV shortcomings and solutions are given in Refs. [99, 98, 116]. Algorithms related to the XS, MS and the original LNV are given in Appendix B, along with the CG routine used in this work.[117]

2.2 Density matrix polynomial expansion

The other class of density matrix solver is based on the Fermi-Dirac (FD) operator expansion.[118] Within the μVT (or NVT) ensemble [cf. Section 1.1 and Eq. (1.15)] at non-zero electronic temperature, the one-particle density operator for a single anti-symmetrized product of one-electron function [cf. Section 1.2.3 and Eq. (1.40)] is given by:

$$\hat{D} = \sum_i \eta_i \left|\psi_i\right\rangle \left\langle\psi_i\right| \tag{2.11}$$

where $\{\psi_i\}$ are the set of molecular orbitals (MO), such that, $\langle\psi_i|\psi_j\rangle = \delta_{ij}$ [cf. Section 1.2.4 and Eq. (1.46)]. For the sake of demonstration, we shall assume that our molecular system can be equated with an ensemble of weakly interacting particles at the thermodynamic equilibrium and obeying the Fermi-Dirac (FD) statistic.[119] As a result, for a given fermion temperature T and chemical potential μ, we may associate the occupation numbers $\{\eta_i\}$ of Eq. (2.11) with the occupation probabilities of the single-particle energy states $\{\epsilon_i\}$ following:

$$\eta_{\mu,T}(\epsilon_i) = \frac{1}{1 + e^{\beta(\epsilon_i - \mu)}}, \quad \text{with} \quad \beta = \frac{1}{k_\mathrm{B} T} \tag{2.12}$$

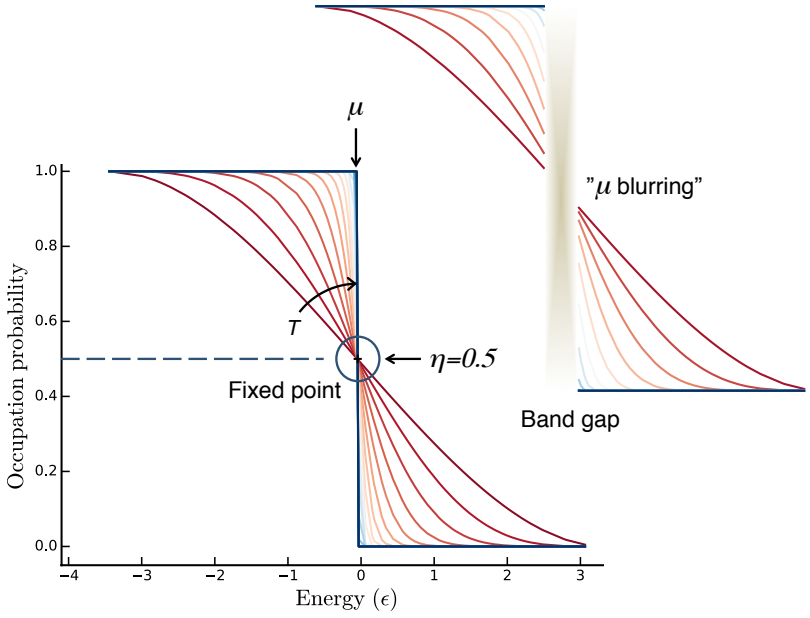


Fig. 2.1 Influence of the fermion temperature on the Fermi-Dirac distribution. The inflection point of the distribution is located at $(\mu = 0, \eta = 0.5)$

Note that the one-particle density matrix of Eq. (2.11) with the occupation probabilities as defined in Eq. (2.12) must respect the N-representability conditions:

$$\mathrm{Tr}\{\hat{D}\} = \sum_i \eta_i = N, \quad \text{and} \quad \eta_i \in [0,1] \tag{2.13}$$

Within the MO basis, the matrix representation of Eq. (2.11) is readily recognized as the matrix of the particle occupation numbers [cf. Eq. (1.54)] which, at non-zero temperature, reads: $\mathcal{O}_{\mu,T} = \mathrm{diag}\{\eta_1\eta_2\cdots\eta_M\}$, where M is the size of the basis set. Therefore, in the atomic orbitals (AO) representation, the DM is expressed as:

$$D_{\mu,T}(F) = \left(1 + e^{\beta(F-\mu I)}\right)^{-1} \tag{2.14}$$

The equation above indicates that there exists a one-to-one *non-linear* correspondence between the Fock and the density matrix. In other terms, at a given temperature, the statistical density matrix is determined by μ and F. By expanding the right-hand-side of Eq. (2.14) using appropriate polynomials, one can expect to obtain the N-representable one-particle density matrix, provided that conditions (2.13) are fulfilled. This constitutes the framework of the density matrix polynomial expansion (DMPE) theory.[118] Since the pioneering works of Goedecker and Colombo,[102] several variations of the DMPE have been proposed. These variations can be differentiated by: (i) the polynomials used for the expansion, and (ii) the statistical ensemble chosen for describing the system. In any case, solving Eq. (2.14), implies to proceed by iteration.

Within the μVT or NVT canonical ensemble, the statistical mixture of one-electron states described in Eq. (2.14) can be purified following three different ways, depending on the variable we choose to operate on.

- In μVT, for fixed chemical potential and temperature: the FD distribution of Eq. (2.14) is expanded in terms of Chebyshev polynomials.[120, 103]. Despite the good performance of the approach,[106, 121] its application requires a precise knowledge of the chemical potential, that is for isolated molecular system, knowledge of interior eigenvalues. Linear scaling algorithms addressing such task along with some improvements were proposed,[122–124] but one may wonder if this step is really necessary.

- In NVT, for a fixed number of particles N: we might try to cool down the system towards the zero temperature limit, as depicted on Figure {2.1}. This was first proposed by Daw in his seminal paper[125] where he demonstrated that the

McWeeny purification of Eq. (2.7) is related to a "temperature-driven" density matrix equation of motion. The notion of statistical ensembles and the possibility of solving Eq. (2.14) in the NVT were rationalized by Palser and Manolopoulos[104] (PM) —without establishing direct relationships with Daw's proposal. In the same work, PM introduced a way of enforcing the McWeeny polynomial to preserve the N-representability conditions throughout the purification process. Later in the manuscript, the PM approach will be referred to as canonical purification (CP).

- In $\mu(N)VT$, at the zero temperature: Niklasson[110] proposed to approach the FD distribution at zero temperature, ie. the Heaviside step function,[126] by varying the number of occupied states around the exact N, that is, adjusting the polynomial dynamically during the recursion without enforcing requirements of Eq. (2.13), such that the N-representable ground state DM is obtained only at convergence. As a result, the Niklasson's method implicitly assumes that the system is coupled to a bath of particles, which are added or withdrawn with respect to the target value. This family of polynomials constitutes the basis for trace-correcting (TC) purification.

It should be emphasized that the purification methods mentioned in the last two points can be easily adapted to the grand canonical ensemble.[104, 110] Calculations of the ground state density matrix using the CP and TC methods are based on the recursive application of projection polynomials $\{P_n\}$ to evaluate the step function, $\Theta(\mu I - F)$, centered at the (unknown) chemical potential. This can be formally written as follows:

$$\Theta(\mu I - F) = \lim_{n\to\infty} P_n(\, P_{n-1}(...P_2(\, P_1(\, D_0(F;\mu)\,)\,)...)\,) \tag{2.15}$$

$$\text{with} \quad D_0 = \alpha_1 I - \alpha_2(\mu I - F) \tag{2.16}$$

The density matrix purification is initialized by performing the linear transformation of Eq. (2.16) where $\{\alpha_1, \alpha_2\}$ are parameters judiciously chosen to: (i) map the eigenvalues of the Fock matrix into the $[0, 1]$ interval, and (ii) depending on the purification method, to verify: $\mathrm{Tr}\{D_0\} = N$. An example of purification process for a (2×2) mixed state is presented in Figure {2.2}. At this stage, the problem of the chemical potential remains a serious bottleneck since, for a given chemical potential or number of occupied states, *fixed inflection point* polynomials, eg. the McWeeny polynomial of Eq. (2.7), yield the N-representable ground state density matrix, if and only if, the inflection point is located at the middle of the eigenvalue spectrum as depicted in Figure {2.1}. As a result, there exist two possibilities: (i) determine the value of μ to shift the eigenvalues of the initial

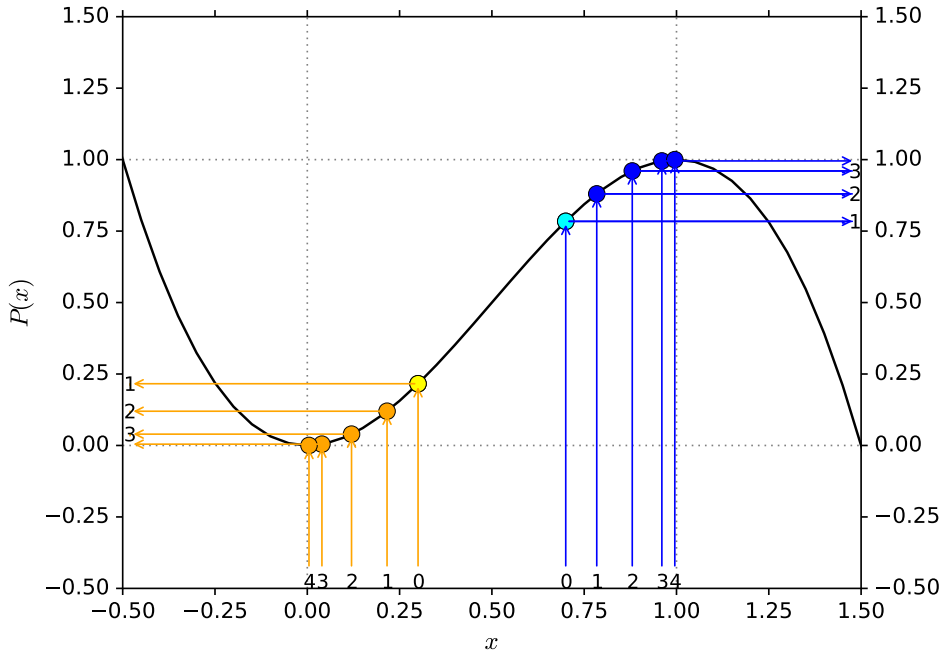


Fig. 2.2 McWeeny purification polynomial $P(x) = 3x^2 - 2x^3$. The initial (2×2) mixed state, with occupation numbers $\{\eta_1, \eta_2\}_{n=0} = \{0.3, 0.7\}$, is purified to eventually reach $\{\eta_1, \eta_2\}_{n=4} = \{0, 1\}$.

mixed state of Eq. (2.16) towards the left and the right of μ, respectively, and purify according to Eq. (2.15), or (ii) define an approximate guess without the support of μ and purify using *flexible inflection point* polynomials. This work is dealing with the second approach.

As discussed by Niklasson in Ref. [110] performances of the purification methods, that is the number of iterations, depends upon the location of μ in the $[\epsilon_{\min}, \epsilon_{\max}]$ interval, where $\epsilon_{\min}$ and $\epsilon_{\max}$ are the lower and upper bounds of the eigenvalue spectrum, or equivalently, on the value of the filling factor $\theta = N/M$, where M is the number of available states. Typical values of θ are about $1/2$ when dealing with one-electron one-orbital many-electron systems, and around $1/20$ for calculations based on extended basis set where, for instance, there are 10 basis functions per electron. The influence of the filling factor over the performances of the purification methods can be qualitatively understood from the fact that the preconditioning of Eq. (2.16) leads to a clustered set of eigenvalues around θ, the range of this cluster being inversely proportional to the gap of the system. As a result, for extreme values of θ —let us say $\theta < 0.1$ (or equivalently for $\theta > 0.9$)— where the initial DM eigenvalues are located around 0.1, the polynomials must be flexible enough to send a few of the eigenvalues towards the upper bound ($\eta = 1$) of the DM eigenspectrum, whereas all the others must be kept around the lower bound ($\eta = 0$), and purified accordingly. In the next section, we shall present the most popular purification polynomials.

2.2.1 Canonical purification

The Palser and Manolopoulos canonical purification[104] (PMCP) is based on the introduction of a flexible inflection point within the McWeeny polynomial of Eq. (2.7) that allows to address the issues mentioned above, and moreover, to preserve the N-representability properties of the initial guess throughout the recursive process. The density matrix is purified according to the following algorithm:

$$D_{n+1} = \begin{cases} -\dfrac{1}{1-c_n} D_n^3 + \dfrac{1+c_n}{1-c_n} D_n^2 + \dfrac{1-2c_n}{1-c_n} D_n & \text{if } c_n \leq \dfrac{1}{2} \\ -\dfrac{1}{c_n} D_n^3 + \dfrac{1+c_n}{c_n} D_n^2 & \text{if } c_n > \dfrac{1}{2} \end{cases} \tag{2.17}$$

with:

$$c_n = \frac{\mathrm{Tr}\{D_n^2 - D_n^3\}}{\mathrm{Tr}\{D_n - D_n^2\}} \tag{2.18}$$

The polynomials of Eq. (2.17) are plotted in Figure {2.3}. The coefficient c is lying

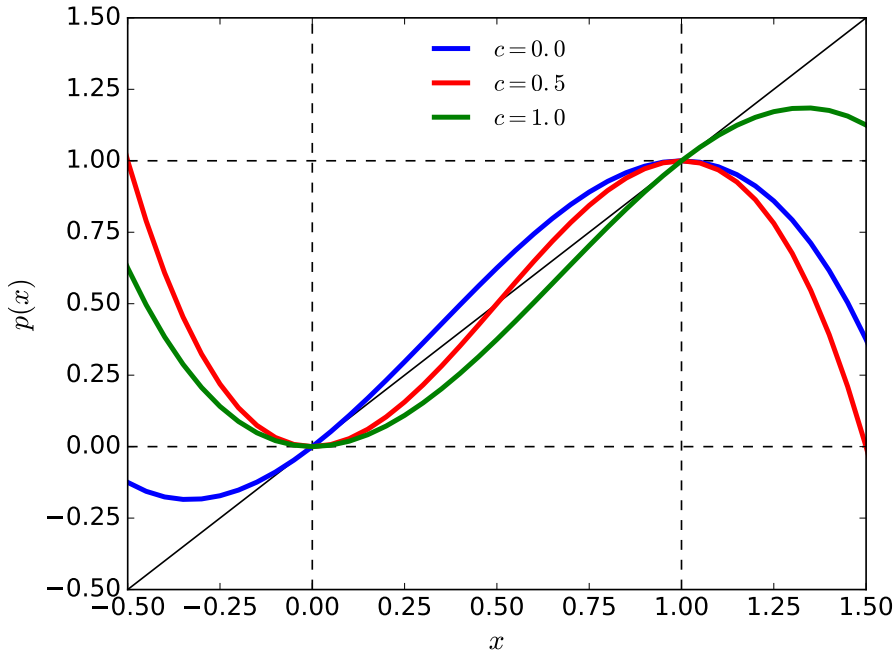


Fig. 2.3 Polynomials of the Palser and Manolopoulos canonical purification.

in the $[0, 1]$ interval. For $c = 1/2$, where the inflexion point is located at $x = 1/2$, we found that both functions behave like the McWeeny polynomial, with stationary points at $x = 0$ and $x = 1$. For the extreme value $c = 1$ ($c = 0$), the inflexion point is located at $x = 1$ ($x = 0$) with the stationary end point outside the maximum (minimum) bound of the spectrum, at $x > 1$ ($x < 0$), whereas the other stationary point remains fixed at $x = 0$ ($x = 1$). The N-representable initial guess as introduced by PM in Ref. [104] is generated from the following normalization relation:

$$D_0 = \alpha(\bar{\mu}I - F) + \theta I \tag{2.19a}$$

$$\alpha = \min\left\{\frac{N}{\tilde{\epsilon}_{\text{max}} - \bar{\mu}}, \frac{M - N}{\bar{\mu} - \tilde{\epsilon}_{\text{min}}}\right\} \tag{2.19b}$$

$$\bar{\mu} = \text{Tr}\{F\}/M \tag{2.19c}$$

where $\tilde{\epsilon}_{\text{max}}$ and $\tilde{\epsilon}_{\text{min}}$ are estimates of the highest (ϵ_{max}) and lowest (ϵ_{min}) eigenvalues of the Fock matrix, respectively. These values are usually accessed, at low cost, using the

Gershgorin's formulas[127, 128]:

$$\tilde{\epsilon}_{\max} = \max_i \left\{ F_{ii} + \sum_{j \neq i}^{M} |F_{ij}| \right\} \tag{2.20a}$$

$$\tilde{\epsilon}_{\min} = \min_i \left\{ F_{ii} - \sum_{j \neq i}^{M} |F_{ij}| \right\} \tag{2.20b}$$

such that: $\tilde{\epsilon}_{\max} > \epsilon_{\max}$, and, $\tilde{\epsilon}_{\min} < \epsilon_{\min}$. It can be demonstrated[104] that, from the N-representable initial guess built from Eq. (2.19), recursive application of the polynomials of Eq. (2.17) maintains the N-representability conditions while converging monotonically to the ground state energy associated with the ground state idempotent density matrix. As outlined by Niklasson,[110] the problem with the trace-preserving PMCP is that it slowly converges at very low or high filling factor[110, 104].

2.2.2 Trace-correcting and trace-resetting purifications

To circumvent this issue, Niklasson[110] has proposed an alternative method where the N-representability constraints are alleviated and the density matrix purification is performed using the following trace-correcting polynomials:

$$D_{n+1} = \begin{cases} P_m^{(a)}(D_n) & = I - (I - D_n)^m (I + mD_n) \quad \text{if} \quad \mathrm{Tr}\{D_n\} \leq N \\ P_m^{(b)}(D_n) & = D_n^m \, (I + m(I - D_n)) \quad \text{if} \quad \mathrm{Tr}\{D_n\} > N \end{cases} \tag{2.21}$$

where $(m+1)$ gives the order of the polynomial, and n is the index of the recursion. Note that for the special case of $m = 1$, the conditions on the trace in Eq. (2.21) have to be swapped. The set of polynomials $\{P_m^{(a)}\}$ and $\{P_m^{(b)}\}$ are plotted in Figure {2.4} for $m = \{1, 2, 3\}$. For $m = 1$ (2nd order polynomials), the stationary points are fixed for $x = 0$ and $x = 1$, for $P_1^{(a)}$ and $P_1^{(b)}$, respectively. For $m = 2$, both functions merge into the McWeeny polynomial of Figure {2.2}. Note that the computational ressources, measured in terms of matrix multiplication (MM), are given by the value of m. For instance: for $m = 1$, each purification requires one MM, for $m = 2$, two MMs are necessary, and so on. Consequently, polynomials with higher flexibililty ($m > 2$) might be costly in ressources (for large scale systems) compared to lower order polynomials if the total number of purifications needed to reach convergence is not reduced. The initial guess for the TC family of purifications is given by:

$$D_{0,m} = \frac{1 - 2\beta_m}{\tilde{\epsilon}_{\max} - \tilde{\epsilon}_{\min}} (\tilde{\epsilon}_{\max} I - F) + \beta_m I \tag{2.22}$$

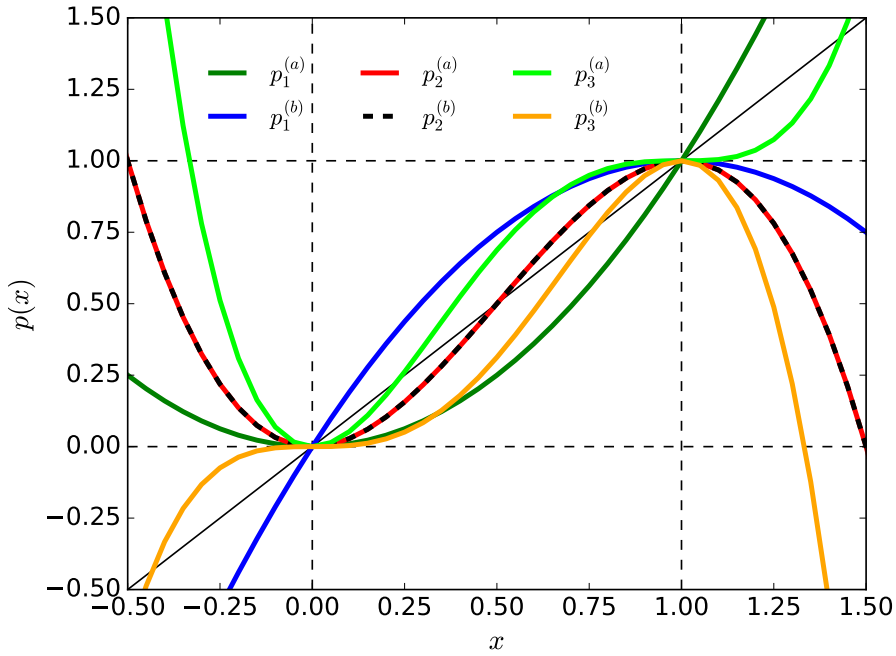


Fig. 2.4 Projection polynomials used for the trace-correcting density matrix purifications at different orders.

where $\beta_m \in [0,1]$ is the stable fixed point such that: $P_m^{(a)}(\beta_m) = \beta_m$, and $P_m^{(b)}(1-\beta_m) = 1-\beta_m$. It is worth emphasizing that normalization of Eq. (2.22) does not enforce —indeed it must not— the trace of D_0 to be equal to the correct value of N. Owing to its simplicity and efficiency, the second-order trace-correcting polynomials (TC2) is the most popular.[129, 124, 130] By setting $m = 1$ in Eq. (2.21), it writes:

$$D_{n+1} = \begin{cases} D_n^2 & \text{if } \mathrm{Tr}\{D_n\} \geq N \\ 2D_n - D_n^2 & \text{if } \mathrm{Tr}\{D_n\} < N \end{cases} \tag{2.23}$$

$$\text{with} \quad D_0 = (\tilde{\epsilon}_{\max} I - F)/(\tilde{\epsilon}_{\max} - \tilde{\epsilon}_{\min}) \tag{2.24}$$

An alternative solution to correct the CP deficiences at low or high filling factor, was brought by Niklasson, Tymczak and Challacombe through the trace-resetting mechanism.[112] This is a hybrid method involving both trace-correcting and trace-preserving polynomials. The authors proposed to subsitute the robust PMCP by a more flexible, ie. efficient, trace-preserving polynomials for which, if appearing, instabilities are controlled *via* a resetting option based on the TC2 projections. The working equations

of the trace-resetting (TRS) density matrix purification are the following,

$$D_{n+1} = \begin{cases} \mathcal{F}(D_n) + \gamma_n \mathcal{G}(D_n) & \text{for } \gamma_n \in [\gamma_{\min}, \gamma_{\max}] \\ D_n^2 & \text{if } \gamma_n < \gamma_{\min} \\ 2D_n - D_n^2 & \text{if } \gamma_n > \gamma_{\max} \end{cases} \tag{2.25}$$

$$\text{with:} \quad \gamma_n = \frac{N - \mathrm{Tr}\{\mathcal{F}(D_n)\}}{\mathrm{Tr}\{\mathcal{G}(D_n)\}}, \quad \gamma_{\min} = 0, \quad \text{and} \quad \gamma_{\max} = 6 \tag{2.26}$$

where the recipe for initializing the density matrix is identical to the TC2 purification [cf. Eq. (2.24)]. The parameter γ_n is analogous to the flexibe inflection point c_n [cf. Eq. (2.17)] but for the composite polynomial of Eq. (2.25) given by:

$$\mathcal{F}(D_n) = D_n^2(4D_n - 3D_n^2) \tag{2.27}$$

$$\mathcal{G}(D_n) = D_n^2(I - D_n)^2 \tag{2.28}$$

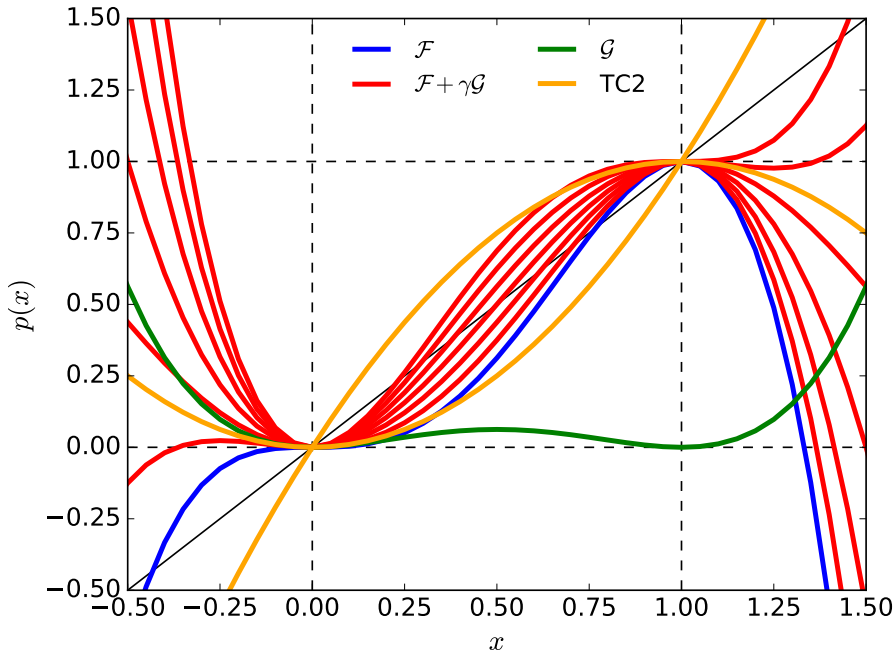


Fig. 2.5 Projection polynomials used of the trace-resetting density matrix purification for different adjustement parameters of $\gamma \in [0, 6]$.

The combination of the these two quartic polynomials (TRS4) results to a trace conserving purifiation for γ bounded in the $[0, 6]$ interval. As mentioned by Niklasson

et al., adding the function $\gamma\mathcal{G}$ to $\mathcal{F}$ and increasing γ continuously change the TRS4 polynomial from $\mathcal{F}(x)$ at $\gamma = 0$ to the mirror function $1 - \mathcal{F}(x)$ at $\gamma = 6$.[112] The variations of the TRS4 function with respect to the γ value are plotted in Figure {2.5}, along with the TC2 polynomials. Convergence is achieved when $\gamma \to 3$, that is, when the TRS4 polynomial transforms to the McWeeny purification: $D_{n+1} = 3D_n^2 - 2D_n^3$. Since the TRS4 function presents (at most), two inflection points, runaway solutions may appear when $\gamma \notin [0, 6]$. In that case the trace-resetting mechanism supplied by the TC2 projection takes over from the TRS4 and remaps the density matrix within the trace-conserving domain. Minute details including performances of the method can be found in Ref. [112].

The Table {2.1} summarizes the key points of the density matrix Fermi-Dirac polynomial expansion including the number of MMs performed for each recursive call. We emphasize that the number of MMs is governed by the order of the polynomials. For comparison, the characteristics of LNV density matrix minimization are also reported. In this case, it should be mentioned that there is an additional cost related to the use of the conjugate-gradient routine, which is indicated by the number in parenthesis. This number corresponds to the number of MMs performed during the CG line search. Algorithms used in this work are provided in Appendix C.

Density matrix solver	MM/iteration	properties
PMCP	2	(+) preserves the N-representability constraints (−) slowly converges at extreme filling factors
McW	2	(+) algorithmic simplicity (−) work only at half filling factor
TC2	1	(+) algorithmic simplicity (−) yields the correct trace only at convergence
TRS4	2	(+) preserves the trace during the last iterations (−) algorithmic complexity
LNV	6 (+5)	(+) energy functional minimization (−) algorithmic complexity (eg. idempotency and μ)

Table 2.1 Density matrix solvers and their features.

2.2.3 Hole-particle canonical purification

It is worth to mention that beyond their mathematical characteristics, some of the density matrix purifications presented above were based (more or less) on physical motivations. Since the early work of McWeeny in 1956, the increasing complexity in attempting to derive more robust and efficient polynomials has reached a stationary point, with

for example the TRS4 or other approaches introduced for instance by Mazziotti,[108] Kryachko[131] or Holas.[132] Intuitively, the NVT ensemble appears as the natural framework to derive density matrix purification for isolated system, although hybrid methods (TC2 or TRS4) relying on a ill-defined $\mu(N)VT$ ensemble are already very efficient. The PMCP was, so far, the only strictly canonical purification, in the sense that it conserves N-represensability conditions throughout the iterative process and converges systematically as the order of the recursion increases. Nevertheless, Palser and Manalopoulos did not provide any physical interpretation nor insight, for explaining their formulation. We also outlined that, all the DMPE methods presented in this chapter invoked a conditional statement with respect to the trace of the density matrix in order to adjust the polynomial accordingly.

In the work presented below, by re-considering the original proposition of McWeeny described in Section 2.1.1, we introduce a constrained minimization principle where the N-representability conditions are fulfilled from the early steps to the end of the recursion. The very simple purification polynomial emerging from it, called hole-particle canonical purification (HPCP), is given by:

$$D_{n+1} = (1-2c_n)D_n + 2(1+c_n)D_n^2 - 2D_n^3, \quad \text{with:} \quad c_n = \frac{\mathrm{Tr}\{D_n^2 - D_n^3\}}{\mathrm{Tr}\{D_n - D_n^2\}} \tag{2.29}$$

where c_n is the flexible inflexion point already introduced by Palser and Manolopoulos [cf. Eq. (2.18)]. In terms of both, the one-particle and one-hole density matrix, Eq. (2.29) can be recast as

$$D_{n+1} = D_n + 2(D_n^2\bar{D}_n - c_n D_n\bar{D}_n), \quad \text{with:} \quad c_n = \frac{\mathrm{Tr}\{D_n^2\bar{D}_n\}}{\mathrm{Tr}\{D_n\bar{D}_n\}} \tag{2.30}$$

From a conditioned D_0, the recursion relation Eq. (2.29) or Eq. (2.30) is able to deliver the exact ground-state density matrix without the need of correcting the trace, nor adjusting the polynomial during the purification process. A a consequence, the polynomial of Eq. (2.30) is self-consistent. The initial guess suitably conditioned for the HPCP approach is defined according to

$$D_0 = \alpha D_{\mathrm{min}} + (1-\alpha)\, D_{\mathrm{max}} \tag{2.31}$$

where $\alpha \in [0,1]$ is the mixing coefficient between $D_{\min}$ and $D_{\max}$, which are evaluated from the following recipe:

$$D_{\min} = \lambda_o(\mu I - F) + \theta I \tag{2.32a}$$

$$D_{\max} = \lambda_q(\mu I - F) + \theta I \tag{2.32b}$$

$$\mu = \frac{\mathrm{Tr}\{F\}}{M} \tag{2.32c}$$

$$\lambda_o = \min\{\lambda_1, \lambda_2\} \tag{2.32d}$$

$$\lambda_q = \max\{\lambda_1, \lambda_2\} \tag{2.32e}$$

$$\lambda_1 = \frac{N}{M(\epsilon_{\max} - \mu)} \tag{2.32f}$$

$$\lambda_2 = \frac{M - N}{M(\mu - \epsilon_{\min})} \tag{2.32g}$$

For the extended comparison presented in the following section, HPCP denotes for the HPCP associated with the initial guess described above.

Communication: Generalized canonical purification for density matrix minimization

Lionel A. Truflandier,[1,a)] Rivo M. Dianzinga,[1] and David R. Bowler[2]
[1]*Institut des Sciences Moléculaires, Université Bordeaux, CNRS UMR 5255, 351 cours de la Libération, 33405 Talence cedex, France*
[2]*London Centre for Nanotechnology, UCL, 17-19 Gordon St., London WC1H 0AH, United Kingdom; Department of Physics and Astronomy, UCL, Gower St., London WC1E 6BT, United Kingdom; and International Centre for Materials Nanoarchitechtonics (MANA), National Institute for Materials Science (NIMS), 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan*

(Received 6 January 2016; accepted 22 February 2016; published online 3 March 2016)

A Lagrangian formulation for the constrained search for the N-representable one-particle density matrix based on the McWeeny idempotency error minimization is proposed, which converges systematically to the ground state. A closed form of the canonical purification is derived for which no *a posteriori* adjustment on the trace of the density matrix is needed. The relationship with comparable methods is discussed, showing their possible generalization through the *hole-particle* duality. The appealing simplicity of this *self-consistent* recursion relation along with its low computational complexity could prove useful as an alternative to diagonalization in solving dense and sparse matrix eigenvalue problems. © *2016 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4943213]

As suggested 60 years ago,[1] the idempotency property of the density matrix (DM) along with a minimization algorithm would be sufficient to solve for the electronic structure without relying on the time consuming step of calculating the eigenstates of the Hamiltonian matrix. The celebrated McWeeny purification formula[2] has inspired major advances in electronic structure theory based on (conjugate-gradient) DM minimization[3–8] (DMM) or DM polynomial expansion[9,10] (DMPE), where the DM is evaluated by the recursive application of projection polynomials (commonly referred to as *purification*). DMPE resolution includes the Chebyshev polynomial recursion,[9–15] the Newton-Schultz sign matrix iteration,[16–18] the trace-correcting[19] and the trace-resetting[20] purification (TCP and TRS, respectively), and the Palser and Manolopoulos canonical purification (PMCP).[21] They constitute, with sparse matrix algebra, the principal ingredient for efficient linear-scaling tight-binding (TB) and self-consistent field (SCF) theories.[22,23] Since all these methods were originally derived within the *grand canonical ensemble*,[24] for a given total number of states (M), none of them are expected to yield the correct number of occupied states (N) unless the chemical potential (μ) is known exactly. As a result, their implementation to the *canonical ensemble* involves heuristic considerations, where the value of μ[12] or the polynomial expansion[19] is adapted *a posteriori* to reach the correct value for N, which adds irremediably to the computational complexity. Despite the remarkable performances of the DMPE approaches for solving for sparse[6,25] and dense[26–28] DMs, it remains desirable to develop an approach that overcomes the use of the chemical potential while respecting the canonical requirement of constant-N.

In this letter, we derive a rigorous and variational constrained search for the one-particle density matrix which does not rely on *ad hoc* adjustments and respects the N-representability constraint throughout the minimization process. We shall start from the McWeeny unconstrained minimization of the error in the idempotency of the density matrix,[1] given by

$$\underset{D\to\mathcal{D}_\mu}{\text{minimize}}\ \Omega_{\text{McW}}\{D;(\mathcal{H},\mu)\}, \tag{1a}$$

$$\text{with}\ \ \Omega_{\text{McW}} = \text{Tr}\{(D^2-D)^2\}, \tag{1b}$$

where for a given fixed Hamiltonian[29] $\mathcal{H}$ and chemical potential μ, the density matrix $\mathcal{D}_\mu$ is the ground-state for that Hamiltonian and chemical potential. The initial guess (D_0) is generally constructed as a function $\mathcal{H}$, suitably scaled,

$$D_0 = \beta_1 I + \beta_2(\mu I - \mathcal{H}), \tag{2}$$

where β_1 and β_2 stand for preconditioning constants such that the eigenvalues of D_0 lie within a predefined range. The double-well shape of the McWeeny function with 3 stationary points: 2 minima at $x_p = 1$ and $x_{\bar{p}} = 0$ and 1 local maximum at $x_m = \frac{1}{2}$ (see Fig. 1(a), red curve), are important features in developing robust DMM algorithms. Finding the minimum of Ω_{McW} would be easily performed by stepwise gradient descent,[1] where the DM is updated at each iteration n,

$$D_{n+1} = D_n - \sigma_n \nabla\Omega_{\text{McW}}, \tag{3a}$$

$$\text{with}\ \ \nabla\Omega_{\text{McW}} = 2\left(2D_n^3 - 3D_n^2 + D_n\right), \tag{3b}$$

and $\sigma_n \geq 0$ represents the step length in the negative direction of the gradient. Considering an optimal fixed step length descent ($\sigma = 1/2$), on inserting Eq. (3b) into Eq. (3a), the McWeeny purification formula appears,

$$D_{n+1} = 3D_n^2 - 2D_n^3, \tag{4}$$

where the right-hand side of the equation above can be view as an auxiliary DM. For a well-conditioned D_0, i.e., $\lambda(D_0) \in [-\frac{1}{2}, \frac{3}{2}]$, repeated application of the recursion

a)Electronic mail: lionel.truflandier@u-bordeaux.fr

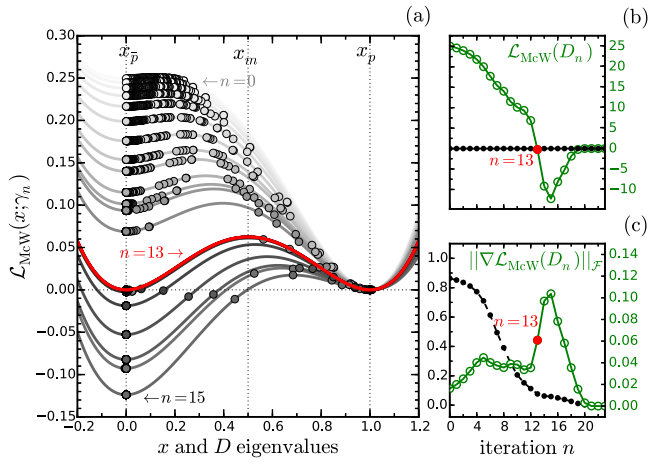


FIG. 1. (a) Convergence of the McWeeny Lagrangian and density matrix eigenvalues during the course of the minimization using a test Hamiltonian and an occupation factor $\theta = 0.10$. A grey scale is used to guide the eye during the processes of purification. Each curve is a plot of the function $\mathcal{L}_{\mathrm{McW}}(x;\gamma_n)$ computed at each iteration n. The red line corresponds to $\mathcal{L}_{\mathrm{McW}}(x;0) = \Omega_{\mathrm{McW}}$. (b) Convergence of $\mathcal{L}_{\mathrm{McW}}$ (green circles) and the trace conservation $\mathrm{Tr}\{D_n\} - N$ (black dots). (c) Convergence of $\|\nabla\mathcal{L}_{\mathrm{McW}}\|_{\mathcal{F}}$ (green circles) and $\|D_n\|_{\mathcal{F}} - N$ (black dots).

identity [Eq. (4)] naturally drives the eigenvalues of D_{n+1} towards 0 or 1. For basic TB Hamiltonians where the occupation factor ($\theta = N/M$) is close to 1/2 and μ can be determined by symmetry[21] or when the input DM is already strongly idempotent, the minimization principle (1a) is able, on its own, to deliver the correct N-representable ground-state DM ($\mathcal{D}$). Beyond these very specific cases, we have to enforce the objective function (1b) to keep N constant during the minimization. From Eq. (4), a sufficient condition would be to impose the trace of the auxiliary DM to give the correct number of occupied states. This leads us to solve a constrained optimization problem which can be formulated in terms of the McWeeny Lagrangian ($\mathcal{L}_{\mathrm{McW}}$) by

$$\underset{\substack{\{D\to\mathcal{D}|\mathrm{Tr}\{D\}=N\}\\ \gamma}}{\text{minimize}} \ \mathcal{L}_{\mathrm{McW}}\{D,\gamma;(\mathcal{H}),N\}, \tag{5a}$$

$$\text{with } \mathcal{L}_{\mathrm{McW}} = \Omega_{\mathrm{McW}} - \gamma\left(\mathrm{Tr}\{3D^2 - 2D^3\} - N\right), \tag{5b}$$

where γ is the constant-N Lagrange multiplier. The McWeeny Lagrangian can be minimized using

$$\nabla\mathcal{L}_{\mathrm{McW}} = \nabla\Omega_{\mathrm{McW}} - 6\gamma\left(D - D^2\right), \tag{6a}$$

$$\partial_\gamma\mathcal{L}_{\mathrm{McW}} = \mathrm{Tr}\{3D^2 - 2D^3\} - N. \tag{6b}$$

Taking trace Eq. (6a) we obtain the expression for γ,

$$\gamma = \frac{1}{3} - \frac{2}{3}c - \frac{1}{6}d, \tag{7a}$$

$$\text{with } c = \frac{\mathrm{Tr}\{D^2 - D^3\}}{\mathrm{Tr}\{D - D^2\}}, \tag{7b}$$

$$d = \frac{\mathrm{Tr}\{\nabla\mathcal{L}_{\mathrm{McW}}\}}{\mathrm{Tr}\{D - D^2\}}. \tag{7c}$$

Then, Eqs. (6a) and (7a) are updated at each iteration by requiring $\mathrm{Tr}\{\nabla\mathcal{L}_{\mathrm{McW}}\} = 0$, that is $d = 0$, for all D. As a result, given D_0 such that $\mathrm{Tr}\{D_0\} = N$ and $[\mathcal{H}, D_0] = 0$, from the fixed-step gradient descent minimization described above, we obtain a recursion formula,

$$D_{n+1} = D_n - \frac{1}{2}\nabla\mathcal{L}_{\mathrm{McW}}\{D_n;\gamma_n\}, \tag{8}$$

which guarantees $\mathrm{Tr}\{D_{n+1}\} = N$ and $[\mathcal{H}, D_{n+1}] = 0$, $\forall n$. Added to the preconditioning $\lambda(D_0) \in [0,1]$, the iterative process should approach the (one-particle) ground-state energy $\mathcal{E} = \mathrm{Tr}\{\mathcal{H}\mathcal{D}\}$ variationally. The parameter c [Eq. (7b)] is recognized as the unstable fixed point introduced in Ref. 21, where $c \in [0,1]$. As a result, the interval $[-\frac{1}{3}, \frac{1}{3}]$ constitutes the stable variational domain of γ.

The variation of the McWeeny Lagrangian function and the DM eigenvalues during the course of the minimization is presented in Fig. 1(a) for a test Hamiltonian with $N = 10$, $M = 100$, and a suitably conditioned initial guess (*vide infra*). The corresponding convergence profiles of $\mathcal{L}_{\mathrm{McW}}$ and $\|\nabla\mathcal{L}_{\mathrm{McW}}\|$ (green circles) are reported on Figs. 1(b) and 1(c), respectively, along with the trace conservation $\mathrm{Tr}\{D_n\} - N$ and the DM norm convergence $\|D_n\| - N$ (black dots). We may notice first that for $\gamma = 0$ (or $c = x_m = \frac{1}{2}$), $\mathcal{L}_{\mathrm{McW}}$ simplifies to Ω_{McW}. For intermediate states, $\gamma \in [-\frac{1}{3}, 0] \cup [0, \frac{1}{3}]$, the symmetry of Ω_{McW} is lost and the shape of $\mathcal{L}_{\mathrm{McW}}(x,\gamma_n)$ drives the eigenvalues in the *hole* (left) or in the *particle* (right) well. From the grey scale in Fig. 1(a), we observe how γ_n influences $\mathcal{L}_{\mathrm{McW}}$ (along the y-axis) at $x_{\bar{p}}$ and the abscissa of the second stationary point x_m which is free to move in $[x_{\bar{p}}, x_p]$. This yields to transform the *hole* well from a local ($n = 0$) to a global ($n = 15$) minimum (or conversely the *particule* well from a global to a local minimum). At the boundary values $\gamma = \{-\frac{1}{3}, \frac{1}{3}\}$, $x_{\bar{p}}$ and x_m merged to a saddle point in such a way that only one global minimum left at x_p. Notice that, for situations where $\gamma \notin [-\frac{1}{3}, \frac{1}{3}]$, the saddle point transforms to a maximum and runaway solutions may appear. Nevertheless, as long as D_0 is well conditioned, such kind of critical problem should not be encountered.

Figs. 1(b) and 1(c) highlight the minimization mechanism: (i) from iterate $n = 0$ to 12; $\gamma \to 0^+$, $\mathcal{L}_{\mathrm{McW}}$ follows the search direction and decreases monotonically. (ii) At iterate $n = 13$; $\gamma \simeq 0$, $\mathcal{L}_{\mathrm{McW}}$ is close to the target value but the gradient residual is nonzero. (iii) From $n = 14$ to 15; $\gamma < 0$, the search direction is inverted. (iv) At iterate $n = 16$, all the eigenvalues are trapped in their respective wells. (iii) From iterate $n = 17$ to 23, $\gamma \to 0^-$, we are in the McWeeny regime [Eq. (4)] and $\mathcal{L}_{\mathrm{McW}}$ eventually reaches the global minimum.

Taking advantage of the closure relation,

$$\bar{D} + D = I, \tag{9}$$

where $\bar{D}$ stands for the *hole* density matrix,[30] a more appealing form for the McWeeny canonical purification [Eq. (8)] can be derived by reformulating Eqs. (6a) and (7b) in terms of D and $\bar{D}$,

$$D_{n+1} = D_n + 2\left(D_n^2\bar{D}_n - \frac{\mathrm{Tr}\{D_n^2\bar{D}_n\}}{\mathrm{Tr}\{D_n\bar{D}_n\}}D_n\bar{D}_n\right). \tag{10}$$

Notice that since at convergence $D\bar{D} = 0$, $\mathrm{Tr}\{D\bar{D}\}$ must be chosen as the termination criterion in the recursion of Eq. (10) to avoid numerical instabilities when approaching the minima. The closed-form of this recurrence relation is remarkable: providing N and $\mathcal{H}$ used to build D_0 [Eq. (2)], we have

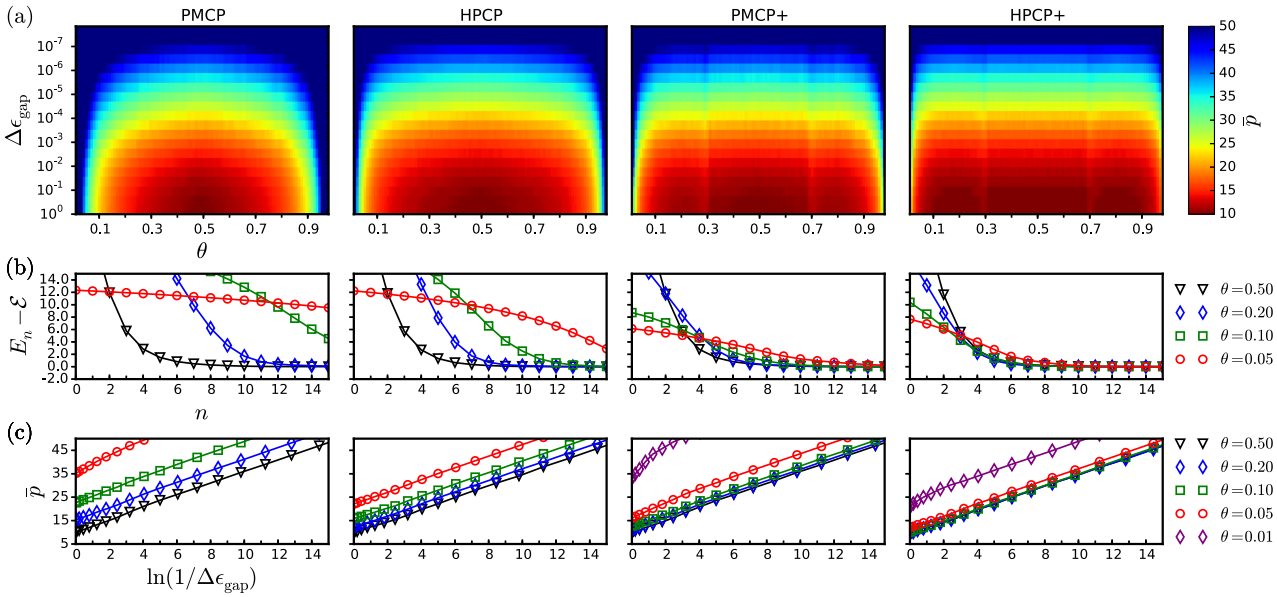


FIG. 2. (a) Color maps displaying the average number of purifications ($\bar{p}$) as the function of the filling factor (θ) and energy gap ($\Delta\epsilon_{\text{gap}}$). Results obtained from the PMCP and HPCP methods using the initial guess of Eqs. (2)-(11) and (2)-(14) (notated PMCP+ and HPCP+). Each pixel on the maps corresponds to an average over 32 test Hamiltonians. (b) Energy convergence profiles with respect to the first 15 iterations for selected values of θ. (c) Average number of purifications as a function of $\ln(1/\Delta\epsilon_{\text{gap}})$.

a self-consistent purification transformation which should converge to $\mathcal{D}$ without any support of heuristic adjustments. Indeed, Eq. (10) can also be derived from the PMCP relations by working on both D and $\bar{D}$ and enforcing relation (9) at each iteration (see the Appendix). Consequently, we can also demonstrate[31] that the *hole-particle* canonical purification (HPCP) of Eq. (10) converges quadratically on $\mathcal{D}$ as shown in Fig. 2(b).

To assess the efficiency and limitations of the HPCP, we have investigated the dependence of the number of purifications (p) on the occupation factor (θ) and the energy gap ($\Delta\epsilon_{\text{gap}} = \epsilon_{N+1} - \epsilon_N$), defined by the higher-occupied (ϵ_N) and lower-unoccupied (ϵ_{N+1}) states. Similarly to the protocol of Niklasson,[15,19] sequences of $M \times M$ dense Hamiltonian matrices ($M = 100$) with vanishing off-diagonal elements were generated, having eigenvalues randomly distributed in the range $[-2.5, \epsilon_N] \cup [\epsilon_{N+1}, 2.5]$ for various $\Delta\epsilon_{\text{gap}} \in [10^{-7}, 1.0]$. As a first test, results are compared to the PMCP,[21] along with the original initial guess [Eq. (2)], where $\beta_1 = \theta$ and $\beta_2 = \min\{\beta, \bar{\beta}\}$, with

$$\beta = \frac{\theta}{\widetilde{\mathcal{H}}_{\max} - \mu}, \quad \bar{\beta} = \frac{\bar{\theta}}{\mu - \widetilde{\mathcal{H}}_{\min}}, \quad \mu \simeq \widetilde{\mu} = \frac{\text{Tr}\{\mathcal{H}\}}{M}, \tag{11}$$

and $\bar{\theta} = 1 - \theta = \bar{N}/M$, $\bar{N}$ being the number of unoccupied states. The lower and upper bounds of the Hamiltonian eigenspectrum ($\widetilde{\mathcal{H}}_{\min}$ and $\widetilde{\mathcal{H}}_{\max}$, respectively) were estimated from to the Geršgorin's disc theorem.[32] The preconditioning of D_0 given in Eq. (11) guarantees that the DM eigenvalues lie in the interval $[0, 1]$ and gives rise to the following additional constraints:

$$\text{Tr}\{D_0\} = N, \tag{12a}$$
$$\text{Tr}\{D_0\} > \text{Tr}\{D_0^2\} > \text{Tr}\{D_0^3\}, \tag{12b}$$
$$\text{Tr}\{D_0^3\} > 2\text{Tr}\{D_0^2\} - \text{Tr}\{D_0\}, \tag{12c}$$

which are also necessary and sufficient conditions for $c \in [0, 1]$ at the first iteration. Convergence was achieved with respect to the idempotency property, such that $\text{Tr}\{D_n\bar{D}_n\} \leq 10^{-6}$ for all the calculations. Additional tests on the Frobenius norm[33] and the eigenvalues of the converged density matrix (D_∞) were performed, using

$$\|D_\infty\|_{\mathcal{F}} - \sqrt{\text{Tr}\{D_\infty\}} < 10^{-6}, \tag{13a}$$
$$\|D_\infty\|_{\mathcal{F}} - N < 10^{-6}, \tag{13b}$$
$$\|\text{diag}\{D_\infty\} - \text{diag}\{I_N, 0_{\bar{N}}\}\|_{\mathcal{F}} < 10^{-6}, \tag{13c}$$

which ensures that, at convergence, the representation of D_∞ is orthogonal, and D_∞ corresponds to $\mathcal{D}$.

The variation of the average number of purifications ($\bar{p}$) with respect to θ and $\Delta\epsilon_{\text{gap}}$ is displayed in Fig. 2(a) using a color map for $\bar{p} \in [10, 50]$. For a given energy gap, the HPCP shows a net improvement over the PMCP approach regarding moderate low and high occupation factors. Nevertheless, as previously noted by Niklasson and Mazziotti,[19,30] the extreme values of θ remain pathological for the original canonical purification and to a lesser extent for the HPCP. One solution would be to break the symmetry of the McWeeny function by moving x_m towards x_p or $x_{\bar{p}}$ depending on the θ value. Basically, this requires a higher polynomial degree for Ω_{McW}, i.e., $\text{Tr}\{(D^n - D)^2\}_{n>2}$, resulting in a higher computational complexity. Assuming optimal programming, we emphasize that the PMCP and HPCP involved only two matrix multiplications per iteration. As already proved in Ref. 21 and highlighted by the energy convergence profiles in Fig. 2(b), the PMCP and HPCP approach $\mathcal{E}$ monotonically.

The dependence of $\bar{p}$ on the band gap plotted in Fig. 2(c) confirms the early numerical experiments,[19,25] where $\bar{p}$ increases linearly with respect to $\ln(1/\Delta\epsilon_{\text{gap}})$. The influence of θ is clearly apparent if we compare the minimum number of purifications as required for the wider band gap (y-axis

intercept), where for example, with $\theta = 0.5$, both canonical purifications reach the ideal value of about 10 purifications, whereas for $\theta = 0.05$, $\bar{p}_{\mathrm{HPCP}} = 23$ and $\bar{p}_{\mathrm{PMCP}} = 37$.

Let us consider how to improve the performance of the canonical purifications by working on the initial guess, regarding the *hole-particle* equivalence (or duality[30]). Instead of searching for D, we may choose to purify $\bar{D}$, which simply requires replacing D with $\bar{D}$ in relation (10). In that case, the initial hole density matrix, satisfying $\lambda(\bar{D}_0) \in [0,1]$, would be given by Eqs. (2) and (11), with $\beta_1 = \bar{\theta}$ and $\beta_2 = -\max\{\beta, \bar{\beta}\}$. Then, intuitively, the guess for the particle density matrix should be improved by using this additional information. Therefore, a more general preconditioning is proposed,

$$D_0^+ = \alpha D_0 + (1-\alpha)(I - \bar{D}_0), \tag{14}$$

where α can be viewed as a mixing coefficient.[34] Results obtained with this new preconditioning are plotted in Fig. 2 (notated PMCP+ and HPCP+). As evident from Fig. 2(a), the naive value of $\alpha = 0.5$ leads to a net improvement of the PMCP and HPCP performances over the range $0.3 < \theta < 0.7$, inside of which the number of purifications becomes independent of θ. Outside this interval, runaway solutions were encountered due to the ill-conditioning of c, where either of the constraints in Eq. (12b) or (12c) is violated. The solution to this problem is to perform a constrained search of α in Eq. (14), such that the first inequality of Eq. (12b) is respected, that is,

$$\underset{\substack{0\le\alpha\le1\\ \delta>0}}{\mathrm{search}} \left\{ \mathrm{Tr}\{D_0^2\} = \begin{cases} N - \delta N, & \text{if } \theta < (1-\delta) \\ N - \delta\bar{N}, & \text{if } \theta > (1-\delta) \end{cases} \right\}, \tag{15}$$

which leads to solve a second-order polynomial equation in α, at the extra cost of only one matrix multiplication. Obviously, the parameter δ has to be carefully chosen such that the second equality of Eq. (12b) and condition (12c) are also respected. We found $\delta \simeq 2/3$ as the optimal value.[31] From Fig. 2, the benefits of this optimized preconditioning are clear when focussing within the range $[0.0, 0.3] \cup [0.7, 1.0]$, albeit with one or two extra purifications around the poles $\theta = \{0.3, 0.7\}$. These benefits are even clearer in Fig. 2(c), where we also show the plots of $\bar{p}$ as a function of $\ln(1/\Delta\epsilon_{\mathrm{gap}})$ for the test case $\theta = 0.01$. At the intercept, we find $\bar{p}_{\mathrm{PMCP}} \simeq 38$ compared to $\bar{p}_{\mathrm{HPCP}} \simeq 21$, showing the improvement bring by the hole-particle equivalence. We have also compared our method against the most efficient of the trace updating methods, TRS4,[20] and find that for non-pathological fillings, the two are comparable in efficiency. For the pathological cases, where TRS4 adjusts the polynomial, we found it more efficient, but at the expense of non-variational behaviour in the early iterations.

To conclude, we have shown how, by considering both electron and hole occupancies, the density matrix for a given system can be found efficiently while preserving N-representability. This opens the door to a more robust, stable ground state minimisation algorithm, with application to standard and linear scaling DFT approaches.

L.A.T. would like to acknowledge D. Hache for his unwavering support and midnight talks about how to move beads along a double-well potential.

APPENDIX: ALTERNATIVE DERIVATION OF THE HOLE-PARTICLE CANONICAL PURIFICATION

We demonstrate that by symmetrizing the Palser and Manolopoulos equations with respect to $\bar{D}$, the closed-form of Eq. (10) appears naturally. Let us start from Eq. (16) of Ref. 21,

$$\text{for } c_n \le \frac{1}{2}, \tag{A1a}$$

$$D_{n+1} = -\frac{1}{1-c_n}D_n^3 + \frac{1+c_n}{1-c_n}D_n^2 + \frac{1-2c_n}{1-c_n}D_n,$$

$$\text{for } c_n > \frac{1}{2},\ D_{n+1} = -\frac{1}{c_n}D_n^3 + \frac{1+c_n}{c_n}D_n^2, \tag{A1b}$$

with c_n given in Eq. (7b). We may search for purification relations *dual* to Eq. (A1), i.e., function of $\bar{D}$. We obtain

$$\text{for } \bar{c}_n \ge \frac{1}{2}, \tag{A2a}$$

$$\bar{D}_{n+1} = -\frac{1}{1-\bar{c}_n}\bar{D}_n^3 + \frac{1+\bar{c}_n}{1-\bar{c}_n}\bar{D}_n^2 + \frac{1-2\bar{c}_n}{1-\bar{c}_n}\bar{D}_n,$$

$$\text{for } \bar{c}_n < \frac{1}{2},\ \bar{D}_{n+1} = -\frac{1}{\bar{c}_n}\bar{D}_n^3 + \frac{1+\bar{c}_n}{\bar{c}_n}\bar{D}_n^2, \tag{A2b}$$

with $\bar{c}_n = 1 - c_n$. Instead of purifying either D or $\bar{D}$, we shall try to take advantage of the closure relation [Eq. (9)] in such a way that, if we choose to work within the subspace of occupied states, the purification of D [Eq. (A1)] is constrained to verify $D = I - \bar{D}$. By inserting this constraint in Eq. (A2), we obtain

$$\text{for } c_n \le \frac{1}{2},\quad D_{n+1} = I - \left(-\frac{1}{c_n}(I-D_n)^3 + \frac{2-c_n}{c_n}(I-D_n)^2 - \frac{1-2c_n}{c_n}(I-D_n)\right), \tag{A3a}$$

$$\text{for } c_n > \frac{1}{2},\quad D_{n+1} = I - \left(-\frac{1}{1-c_n}(I-D_n)^3 + \frac{2-c_n}{1-c_n}(I-D_n)^2\right). \tag{A3b}$$

On multiplying Eqs. (A1a) and (A3a) by $(1-c_n)$ and c_n, respectively [or multiplying Eqs. (A1b) and (A3b) by c_n and $(1-c_n)$], and adding, we obtain

$$D_{n+1} = D_n + 2\left(D_n^2\bar{D}_n - c_n D_n \bar{D}_n\right). \tag{A4a}$$

[1]R. McWeeny, Proc. R. Soc. A **235**, 496 (1956); **237**, 355 (1956); **241**, 239 (1957).
[2]R. McWeeny, Rev. Mod. Phys. **32**, 335 (1960).
[3]X.-P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10891 (1993).
[4]A. D. Daniels, J. M. Millam, and G. E. Scuseria, J. Chem. Phys. **107**, 425 (1997).

[5]J. M. Millam and G. E. Scuseria, J. Chem. Phys. **106**, 5569 (1997).
[6]A. D. Daniels and G. E. Scuseria, J. Chem. Phys. **110**, 1321 (1999).
[7]D. Bowler and M. Gillan, Comput. Phys. Commun. **120**, 95 (1999).
[8]M. Challacombe, J. Chem. Phys. **110**, 2332 (1999).
[9]S. Goedecker and L. Colombo, Phys. Rev. Lett. **73**, 122 (1994).
[10]S. Goedecker and M. Teter, Phys. Rev. B **51**, 9455 (1995).
[11]R. Baer and M. Head-Gordon, Phys. Rev. Lett. **79**, 3962 (1997).
[12]R. Baer and M. Head-Gordon, J. Chem. Phys. **107**, 10003 (1997).
[13]K. R. Bates, A. D. Daniels, and G. E. Scuseria, J. Chem. Phys. **109**, 3308 (1998).
[14]W. Liang, C. Saravanan, Y. Shao, R. Baer, A. T. Bell, and M. Head-Gordon, J. Chem. Phys. **119**, 4117 (2003).
[15]A. M. N. Niklasson, Phys. Rev. B **68**, 233104 (2003).
[16]K. Németh and G. E. Scuseria, J. Chem. Phys. **113**, 6035 (2000).
[17]G. Beylkin, N. Coult, and M. J. Mohlenkamp, J. Comput. Phys. **152**, 32 (1999).
[18]C. Kenney and A. Laub, SIAM J. Matrix Anal. Appl. **12**, 273 (1991).
[19]A. M. N. Niklasson, Phys. Rev. B **66**, 155115 (2002).
[20]A. M. N. Niklasson, C. J. Tymczak, and M. Challacombe, J. Chem. Phys. **118**, 8611 (2003).
[21]A. H. R. Palser and D. E. Manolopoulos, Phys. Rev. B **58**, 12704 (1998).
[22]D. R. Bowler and T. Miyazaki, Rep. Prog. Phys. **75**, 036503 (2012).
[23]S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999).
[24]R. G. Parr and Y. Weitao, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York; Oxford, England, 1994).
[25]E. Rudberg and E. H. Rubensson, J. Phys.: Condens. Matter **23**, 075502 (2011).
[26]E. Chow, X. Liu, M. Smelyanskiy, and J. R. Hammond, J. Chem. Phys. **142**, 104103 (2015).
[27]M. J. Cawkwell, E. J. Sanville, S. M. Mniszewski, and A. M. N. Niklasson, J. Chem. Theory Comput. **8**, 4094 (2012).
[28]M. J. Cawkwell, M. A. Wood, A. M. N. Niklasson, and S. M. Mniszewski, J. Chem. Theory Comput. **10**, 5391 (2014).
[29]We will restrict our discussion to effective one-electron Hamiltonian operators expressed in a finite orthonormal Hilbert space.
[30]D. A. Mazziotti, Phys. Rev. E **68**, 066701 (2003).
[31]L. A. Truflandier, M. R. Dianzinga, and D. R. Bowler, "On grand canonical and canonical ensemble approaches for the one-particle density matrix minimization" (unpublished).
[32]S. Geršgorin, Proc. USSR Acad. Sci. **51**, 749 (1931); available at http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=im&paperid=5235&option_lang=eng.
[33]The Frobenius norm is defined by $\|D\|_{\mathcal{F}} = (\sum_{i,j} |D_{ij}|^2)^{1/2} = \sqrt{\mathrm{Tr}\{D^2\}}$. Notice that $\forall D$, such that $D^2 = D$, then $\|D\|_{\mathcal{F}} = \sqrt{\mathrm{Tr}\{D\}}$.
[34]In that case, it can be shown that $\lambda(D_0^+) \in [-\frac{1}{2}, \frac{3}{2}]$.

2.2.4 Extended comparison of density matrix purifications

For this extended comparison of purification methods we shall consider the same set of Hamiltonian matrices used for the numerical experiment of Section 2.2.3 [cf. Figure 2 of the article and text therein]. The set of 32 Hamiltonians, ie. $H \in \mathbb{R}^{M\times M}$ with $M = 100$, was constructed using the following recipe:

1. $H := \text{random}\{N \times N\}$, such that: $H_{ij} \in [-2, +2]$
2. $H := (H + H^t)/2$, such that: $H_{ij} = H_{ji}$
3. $H_{ij} := H_{ij}/|i-j|^2 \quad \forall i \neq j$, such that: $\lim_{|i-j|\to 0} H_{ij} \to 0$
4. $(E, C) := \text{diagonalize}\{H\}$, such that: $E = \text{diag}\{\epsilon_1\epsilon_2\cdots\epsilon_M\}$ and $C^tC = I$
5. $\tilde{E} := \text{shift}\{E \mid N, \Delta\epsilon_{\text{gap}}\}$, such that: $\Delta\epsilon_{\text{gap}} = \epsilon_{N+1} - \epsilon_N$
6. $\tilde{H} := C\tilde{E}C^t$

In step 5, given a filling factor, $\theta = N/M$, and a band gap, $\Delta\epsilon_{\text{gap}}$, a shift operator is applied to the eigenvalues located at the middle of the band gap in order to verify: $\Delta\epsilon_{\text{gap}} = \epsilon_{N+1} - \epsilon_N$. The final dense Hamiltonian matrix is recovered in step 6. Given a fixed energy band gap of 1.0 au and eigenspectrum width of about 6 au, examples of eigenvalues distributions obtained from a chunk of test Hamiltonians are represented in the Figure {2.6}.

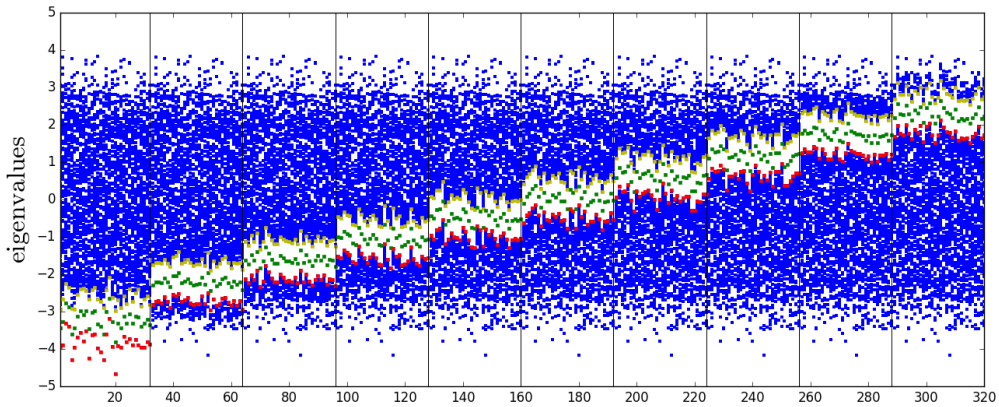


Fig. 2.6 Scatter plot of the eigenvalues (in blue) from a chunk of the test Hamiltonians. The red, yellow and green pixels correspond to ϵ_N, ϵ_{N+1} and middle of the band gap, respectively. Each panel correponds to a set 32 randomized symmetric matrices for filling factor $\theta \in \{0.01, 0.90\}$.

As represented on Figure {2.7}a, this protocol was repeated by varying the filling factor (the x-dimension) in the range $]0, 1[$, and the energy gap in the range $[10^{-7}, 1.0]$ (the y-dimension). This figure clearly highlights the performances of the HPCP with respect to the TC2, TRS4 and PMCP, where each pixel represents the average number of purifications over 32 randomized Hamiltonians. The number of purifications required to achieve convergence for the TC2 is much more higher than for all the other methods (about a factor 2). The TRS4 and HPCP have similar color maps, which are constant along the θ direction, although at the very low or high θ, HPCP looses in efficiency. As discussed in the original article, the impact of the energy gap over the number of purifications is clearly observed.

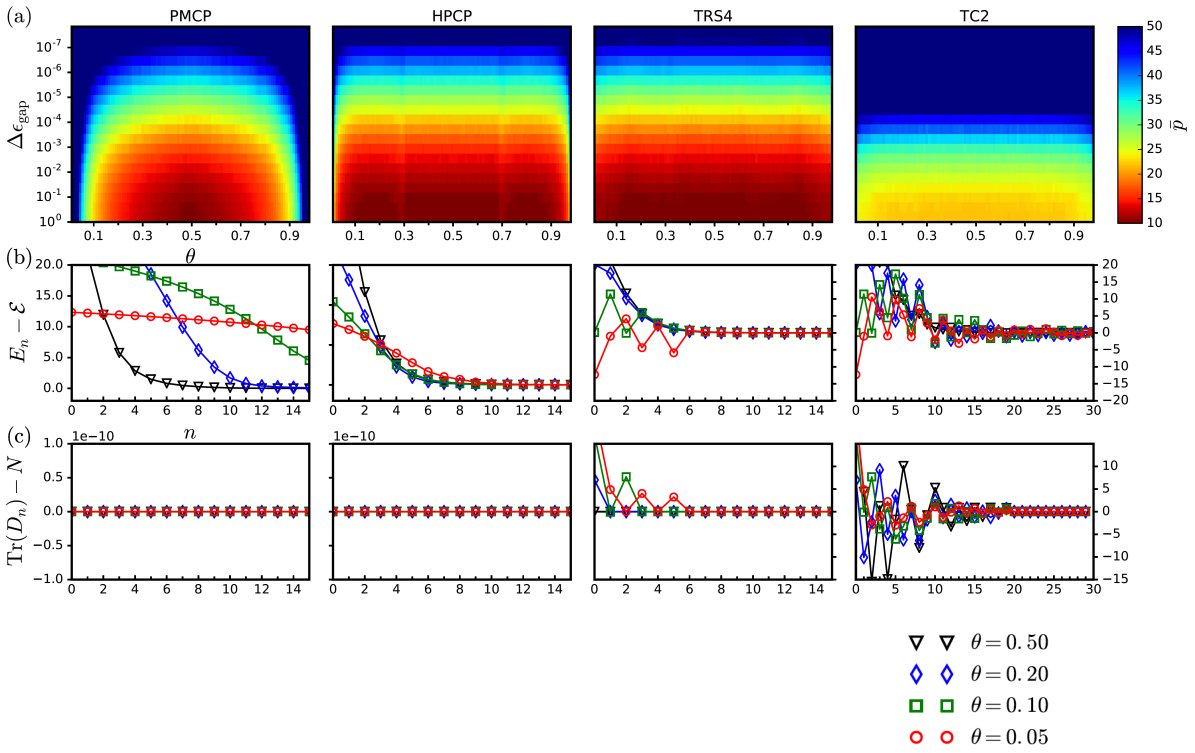


Fig. 2.7 (a) Color maps displaying the average number of purifications ($\bar{p}$) as the function of the filling factor (θ) and energy gap ($\Delta\epsilon_{\mathrm{gap}}$). Results obtained from the PMCP, HPCP, TRS4 and TC2 methods. (b) Energy convergence profiles with respect to the first 15 iterations for selected values of θ, and (c) the corresponding density matrix trace conservation profiles.

Figure {2.7(b,c)} sheds some light on the HPCP trace-preserving property regarding the strong fluctuation of the number of occupied states observed for the trace-correcting method, and to a lesser extent, the trace-resetting mechanism. It is noteworthy that

fluctuations of the trace of the density matrix are attenuated when θ is different from half filling.

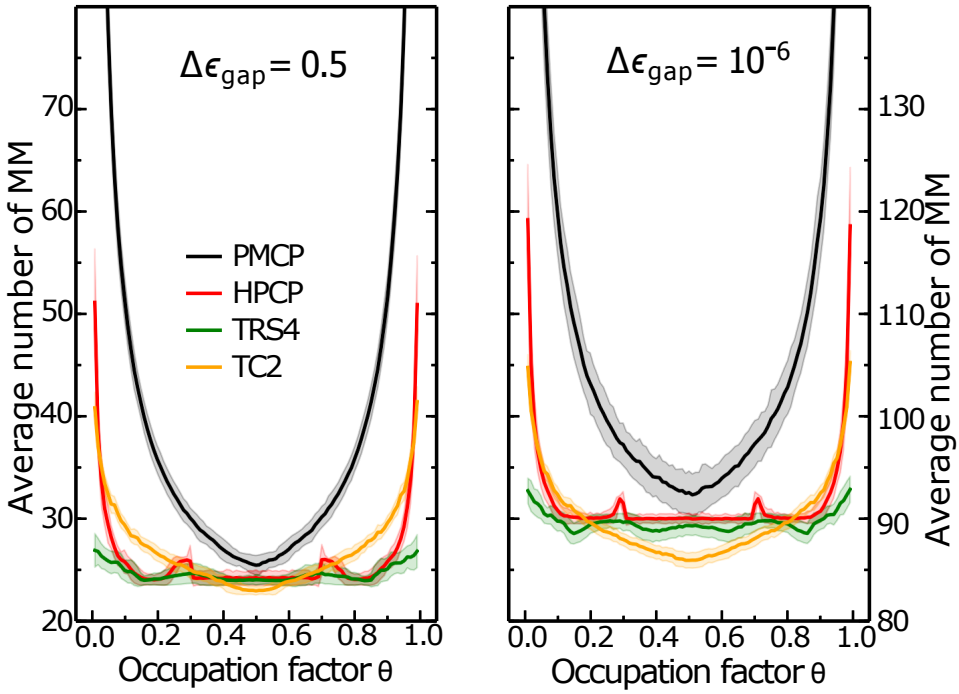


Fig. 2.8 Comparison of four density matrix purifications in terms of matrix multiplications (MMs) for varying filling factors, and two band gap values $\Delta\epsilon_{\text{gap}} = 0.5$ and $\Delta\epsilon_{\text{gap}} = 10^{-6}$. Heavy lines represent averages over 256 random Hamiltonians and shaded areas are the corresponding standard deviations.

In order to emphasize on the computational performance of the purification methods, the number of matrix multiplications (MMs) is a parameter to consider. From an optimal programming perspective, the numbers of MMs per iteration summarized in Table {2.1} allow to compute the total number of MMs realized to reach convergence. Results obtained for a set of 256 random test Hamiltonians —for two different values of the band gap— as a function of θ are plotted in Figure {2.8}. We found that, while the HPCP and TRS4 are comparable for $\theta \in [0.1, 0.9]$, TRS4 is more efficient when dealing with pathological cases. Note that the TC2 also presents the inversed bell-shape distribution. Regarding the influence of the gap, we should note that for the HPCP, the standard deviation is almost negligeable, indicating that the HPCP performances do not depend on the eigenvalue distributions. For wide and low gaps, the TC2 remains the most performant purification method when $\theta \in [0.2, 0.8]$.

2.3 Linear scaling strategies

The density matrix for insulator contains naturally small elements which can be considered as zero with respect to some threshold. In order to accelerate the calculation, one can get rid of zero elements and work only with significant (also referred to as non-zero elements, nnz). Removing the zero elements involves to truncate the density matrix. Working only with non-zero elements requires a structure representation for the truncated density matrix. Performing density matrix minimizations and purifications with these two ingredients enables to achieve a linear scaling calculation.

2.3.1 Density matrix truncations

The most popular density matrix truncations that we use in this thesis are the following:

Numerical truncation

Since there are small elements in the density matrix, a simple and direct scheme to truncate the density matrix is to drop these small elements by setting them to zero if the absolute value is below a predefined threshold τ[96, 133]. With respect to the chosen numerical threshold, the density matrix elements are said to be filtered such as

$$\tilde{D} = \text{FILTER}(\ D\ ,\ \tau\ ) = \begin{cases} D_{ij} & \text{if } |D_{ij}| > \tau \\ 0 & \text{otherwise} \end{cases} \tag{2.33}$$

$\tilde{D}$ is the truncated density matrix. In a minimization or purification algorithm, we apply the truncation (2.33) after each MM.

Radial truncation

The magnitude of density matrix elements depends on the distance between the basis function centers of atom centers. Therefore, one way to truncate the density matrix is to neglect the matrix elements that correspond to distances between basis function centers larger than a predefined cutoff radius R_c[40, 101, 134, 105, 106, 104]. Hence, the filtered matrix elements are such as

$$\tilde{D} = \text{FILTER}(\ D\ ,\ R_c\ ) = \begin{cases} D_{ij} & \text{if } |\mathbf{r}_i - \mathbf{r}_j| < R_c \\ 0 & \text{otherwise} \end{cases} \tag{2.34}$$

In a minimization or purification algorithm, we apply the truncation (2.34) only on the density matrix D as shown in Algorithm {10} (step (9)). We apply the truncation above all else (any polynomial, function or gradient) depending on the density matrix.

2.3.2 Sparse matrix representations

Using the truncation reinforces the matrix sparsity. The access to only non-zero elements is a great advantage for the minimizations and purifications algorithms. Instead of using the standard algebra for matrices in dense format, it is better to employ sparse matrices algebra. In the latter, the sparse matrix is compressed into some matrix data structures. A more detailed discussion of data structures for sparse matrices can be found in Ref. [135]. There are several matrix data structures for sparse matrices[136–139, 16]. For instance, the tool package SPARSKIT[140] presents a multitude of sparse matrices formats such as the compressed sparse column (CSC) format. The CSC representation is the simplest representation for the sparse matrices, that we use in this thesis.

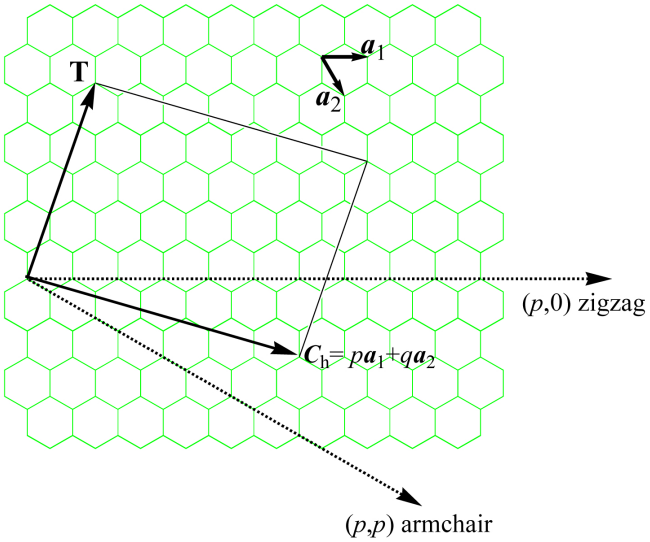
2.4 Applications to carbon nanotubes

In this section we shall compare the performances of $O(N)$ density matrix energy functional minimizations and polynomial expansions presented in Sections 2.1.2 and 2.2, with respect to the truncation schemes. For that study, the Pariser-Parr-Pople (PPP) semi-empiral method described in Section 1.4 will be applied to a set of π-conjugated systems: the carbon nanotubes. We recall that, within the SCF-PPP framework, the ideal case of the one-electron one-orbital picture is imposed. For neutral π-conjugated network, this implies that the filling factor is automatically set up to 1/2.

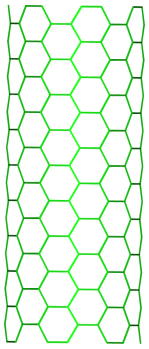
2.4.1 Carbon nanotubes

The π-conjugated systems considered in this part are the carbon nanotubes (CNTs). A CNT is a compound which has the shape of a cylindrical tube made of a rolled single layer of carbon atoms. This single layer of carbon atoms is known as graphene.[141, 142] A detailed description on CNTs and their properties can be found, for instance, in the Refs. [143–145]. The rolling of the graphene to form CNTs is modulated by the chirality vector, $\vec{C}_{\mathrm{h}}$, reproduced on Figure {2.9a}. This vector has two components, such that:

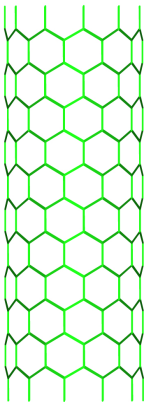
$$\vec{C}_{\mathrm{h}} = p\,\vec{a}_1 + q\,\vec{a}_2, \quad (p,q) \in \mathbb{Z} \tag{2.35}$$



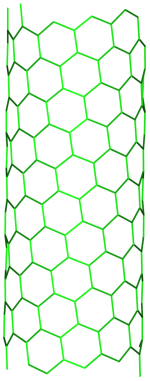
(a) Illustration of the chiral vector for a the graphene sheet. T denotes the tube axis. $\vec{a}_1$ and $\vec{a}_2$ are the unit vectors of graphene in real space.



(b) Armchair type (p, p).



(c) Zigzag type $(p, 0)$.



(d) Chiral type (p, q).

Fig. 2.9 Rolling of a graphene sheet to generate a carbon nanotube.

where $\{\vec{a}_1, \vec{a}_2\}$ are the unit vectors of the graphene sheet. The values of the pair of indices (p, q) define the way the graphene sheet is wrapped. From the definition of $\vec{C}_{\mathrm{h}}$ and the symmetry of translation, three types of CNTs can be generated:

- the armchair type, where $p = q$,
- the zigzag type, where $p = 0$ or $q = 0$,
- the chiral type, where $p \neq q$.

They are represented in Figure {2.9}. Note that p and q are generally given as positive integers with $p > q$. The indices (p, q) do not only determine the carbon atoms' arrangement in the tube, but they can also provide information on the properties of the CNT.[143] For instance:

- the armchair CNTs are all metallic,
- the zigzag CNTs are metallic when p is a multiple of 3,
- the chiral CNTs are metallic when $(p - q)$ is a multiple of 3.

CNTs	# atoms	$a(\mathring{A})$	# cells
$(8, 0)$	32	4.27	33
$(17, 0)$	68	4.26	15
$(12, 0)$	48	4.27	21
$(5, 4)$	244	33.31	3
$(15, 5)$	260	15.37	3
$(11, 5)$	268	20.15	3

Table 2.2 Carbon nanotubes investigated in this work.

Metallic systems are challenging when using linear-scaling methods.[63] Consequently, we have only considered the zigzag and chiral carbon nanotubes described in Table {2.3}. These CNTs were generated from the nanotube structure generator website TUBGEN[146], using a carbon-carbon bond length of 1.421 $\mathring{A}$. Semi-empirical approaches such as the PPP model are generally used to have a qualitative picture of electronic structure and related properties. Nevertheless, it would be interesting to compare it to the state-of-the-art Kohn-Sham DFT calculations in order to appreciate the robustness of the parametrization described in Section 1.4.2.

The Table {2.3} reports the energy gaps of the CNTs using two different PPP parametrizations, along with DFT calculations. We note a good agreement between the

Table 2.3 Energy gap for the π–π^* (in eV) at the Γ point of the first Brillouin zone.

CNTs	# atoms	$a(\mathring{A})$	# cells	Zhang[a]	Ohno[a]	B3LYP[b]
(8,0)	32	4.27	33	1.2481	1.3238	1.283
(17,0)	68	4.26	15	0.6939	0.6940	0.734
(12,0)	48	4.27	21	0.0488	0.0413	0.041
(5,4)	244	33.31	3	0.9192	0.8069	–
(15,5)	260	15.37	3	0.5860	0.6836	0.66
(11,5)	268	20.15	3	0.0481	0.0664	0.00

[a]Present work using the Ohno[70] and Zhang[72] PPP parametrizations.
[b]KS-DFT calculations using the B3LYP exchange-correlation functional reported from Ref. [147].

semi-empirical and the first-principles approaches, with differences which do not exceed 5% (for the Ohno parametrization only). This demonstrates that the major part of the two-electron π interactions is well reproduced by the PPP model.

2.4.2 Numerical truncation for SCF calculations

SCF calculation based on density matrix solvers is controlled via two threshold parameters: (i) the first one is related to the density matrix (DM) convergence inside each SCF cycle, ie. $\tau_D = \|D_{n+1} - D_n\|$,[1] (ii) the second one is related to the convergence of the SCF procedure itself, and based on an energy criterion, ie. $\tau_{\text{SCF}} = |\mathcal{E}_{n+1} - \mathcal{E}_n|$. For the results presented in the next section, we have used $\tau_D = 10^{-3}$ and $\tau_{\text{SCF}} = 10^{-6}$. In order to achieve the linear scaling regime, the density matrix purifications and minimizations must be supported by techniques which reinforce the density matrix sparsity (cf. Section 2.3.1).

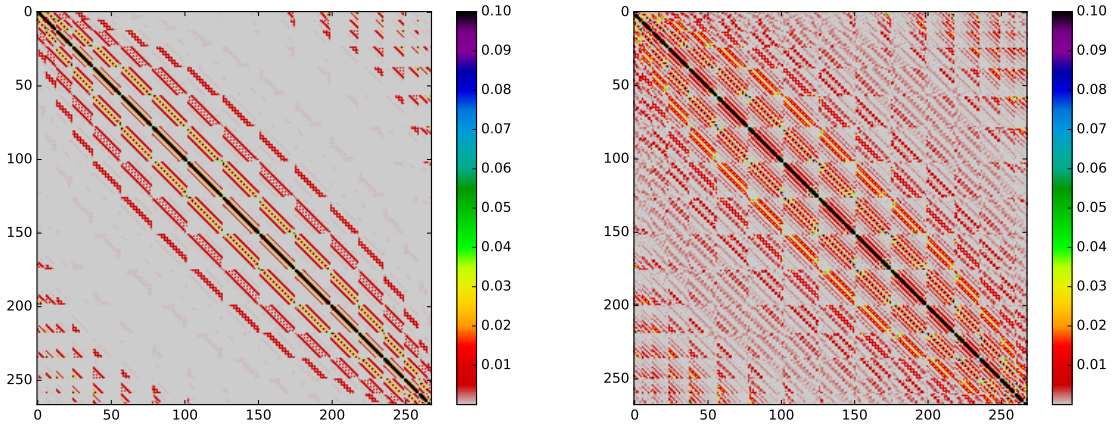
Using the numerical truncation scheme of Eq. (2.33), the sparsity of the DM is controlled by the numerical threshold τ. In Table {2.4} are reported the SCF energies obtained for the set of CNTs described in Table {2.3} using the hole-particle canonical purification (HPCP) as density matrix solver. Different values of τ were considered within the range $[10^{-5}, 10^{-7}]$. $\Delta\mathcal{E}$ is the error between the exact SCF energy obtained (without truncation) and the truncated density matrix purification. N_{SCF} is the number of SCF iterations achieved to reach convergence, and nnz is the number of non-zero DM elements. As expected, by reading Table {2.4} (along the rows), for all the CNTs, we observe that the energy approaches the exact value as τ decreases, ie. the density of the non-zero elements increases. Interestingly, by reading the same table (along the columns), for a given threshold, and given the π-π^* gaps reported in Table {2.3}, we

[1]This parameter correponds to the keyword "tolerance" introduced in the algorithms of the Appendices B and C.

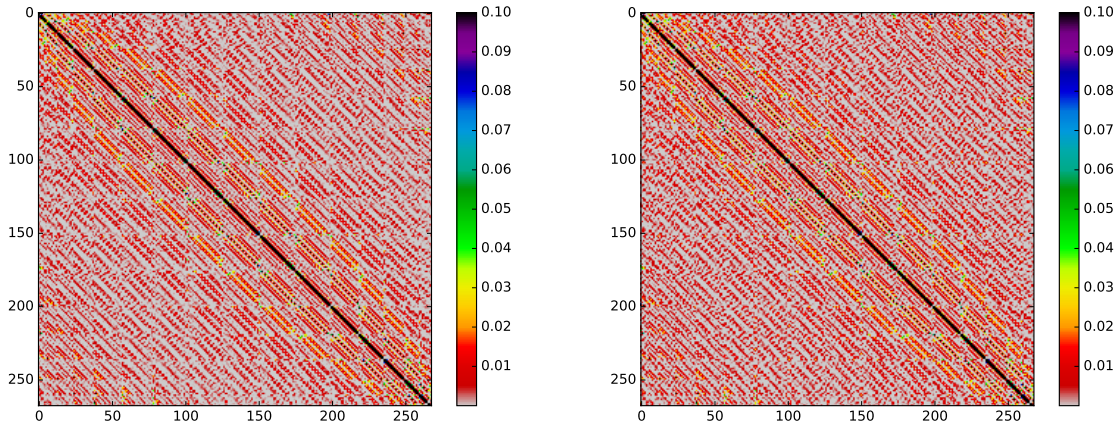
CNTs \ τ	10^{-5}	10^{-6}	10^{-7}	10^{-8}
(8,0)	$\Delta\mathcal{E} = 1.54 \times 10^{-3}$ $N_{\mathrm{SCF}} = 15$ $nnz = 35\%$	$\Delta\mathcal{E} = 6.74 \times 10^{-4}$ $N_{\mathrm{SCF}} = 6$ $nnz = 41\%$	$\Delta\mathcal{E} = 9.27 \times 10^{-6}$ $N_{\mathrm{SCF}} = 5$ $nnz = 49\%$	$\Delta\mathcal{E} = 5.89 \times 10^{-7}$ $N_{\mathrm{SCF}} = 5$ $nnz = 60\%$
(17,0)	$\Delta\mathcal{E} = 7.31 \times 10^{-3}$ $N_{\mathrm{SCF}} = 14$ $nnz = 45\%$	$\Delta\mathcal{E} = 4.86 \times 10^{-5}$ $N_{\mathrm{SCF}} = 6$ $nnz = 50\%$	$\Delta\mathcal{E} = 6.11 \times 10^{-6}$ $N_{\mathrm{SCF}} = 6$ $nnz = 61\%$	$\Delta\mathcal{E} = 4.95 \times 10^{-7}$ $N_{\mathrm{SCF}} = 6$ $nnz = 70\%$
(12,0)	$\Delta\mathcal{E} = 1.05 \times 10^{-3}$ $N_{\mathrm{SCF}} = 9$ $nnz = 50\%$	$\Delta\mathcal{E} = 9.89 \times 10^{-4}$ $N_{\mathrm{SCF}} = 7$ $nnz = 59\%$	$\Delta\mathcal{E} = 2.50 \times 10^{-6}$ $N_{\mathrm{SCF}} = 6$ $nnz = 65\%$	$\Delta\mathcal{E} = 7.50 \times 10^{-7}$ $N_{\mathrm{SCF}} = 6$ $nnz = 77\%$
(5,4)	$\Delta\mathcal{E} = 5.97 \times 10^{-4}$ $N_{\mathrm{SCF}} = 5$ $nnz = 48\%$	$\Delta\mathcal{E} = 1.16 \times 10^{-5}$ $N_{\mathrm{SCF}} = 5$ $nnz = 60\%$	$\Delta\mathcal{E} = 1.72 \times 10^{-7}$ $N_{\mathrm{SCF}} = 5$ $nnz = 69\%$	$\Delta\mathcal{E} = 2.57 \times 10^{-7}$ $N_{\mathrm{SCF}} = 5$ $nnz = 80\%$
(15,5)	$\Delta\mathcal{E} = 3.71 \times 10^{-4}$ $N_{\mathrm{SCF}} = 4$ $nnz = 51\%$	$\Delta\mathcal{E} = 2.33 \times 10^{-5}$ $N_{\mathrm{SCF}} = 5$ $nnz = 62\%$	$\Delta\mathcal{E} = 1.57 \times 10^{-6}$ $N_{\mathrm{SCF}} = 6$ $nnz = 75\%$	$\Delta\mathcal{E} = 2.89 \times 10^{-7}$ $N_{\mathrm{SCF}} = 6$ $nnz = 85\%$
(11,5)	$\Delta\mathcal{E} = 9.90 \times 10^{-5}$ $N_{\mathrm{SCF}} = 13$ $nnz = 54\%$	$\Delta\mathcal{E} = 8.61 \times 10^{-6}$ $N_{\mathrm{SCF}} = 7$ $nnz = 67\%$	$\Delta\mathcal{E} = 1.98 \times 10^{-7}$ $N_{\mathrm{SCF}} = 6$ $nnz = 80\%$	$\Delta\mathcal{E} = 5.96 \times 10^{-8}$ $N_{\mathrm{SCF}} = 6$ $nnz = 89\%$

Table 2.4 SCF convergence with respect to the numerical truncation threshold τ, for a set of carbone nanotubes. $\Delta\mathcal{E}$, N_{SCF} and nnz correspond to the energy error, the number of SCF iterations, and the number of non-zero density matrix elements after having applied the truncation.

note that the sparsity of the DM decreases for increasing energy gap. For reasonable values of the threshold ($\tau \geq 10^{-6}$), the convergence is roughly achieved after a number of 6 cycles, whatever the CNT. This is to be compared with the trend obtained for highly sparse DM where, depending on the CNT, twice as many SCF cycles are required.



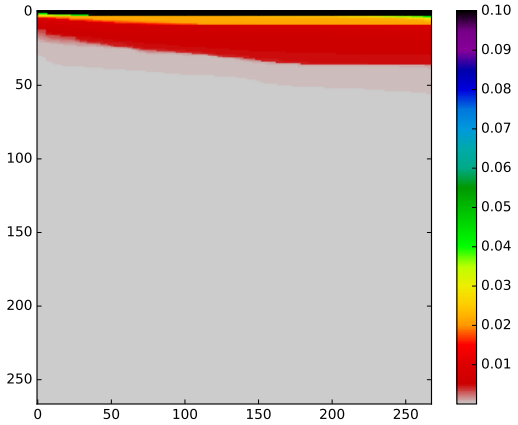
(a) $nnz = 26\%$ (b) $nnz = 72\%$



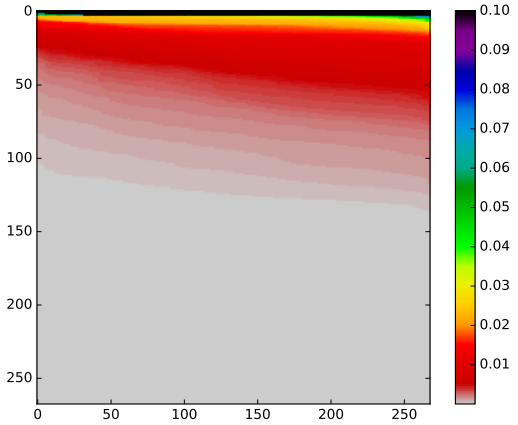
(c) $nnz = 86\%$ (d) $nnz = 89\%$

Fig. 2.10 First form of illustration of the sparsity pattern of the density matrix truncated at $\tau = 10^{-8}$ during the SCF iterations, following the sequence (a) to (d). Results obtained for the CNT (11,5).

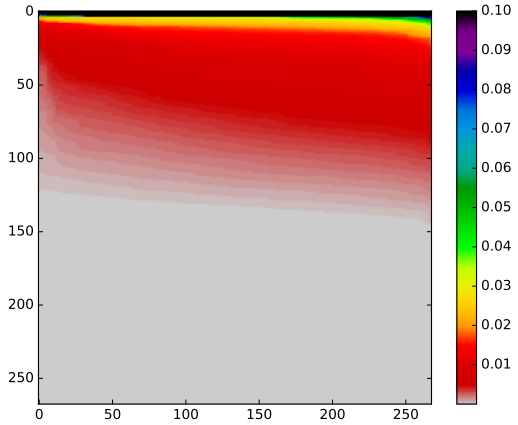
To illustrate the influence of the numerical truncation on the density matrix during the SCF, sparsity patterns were represented in Figure {2.10} for the CNT (11,5). These sparsity patterns correspond to an average of the density matrices over the six SCF cycles summarized into four sequences. In Figure {2.10}, each colored map represents the



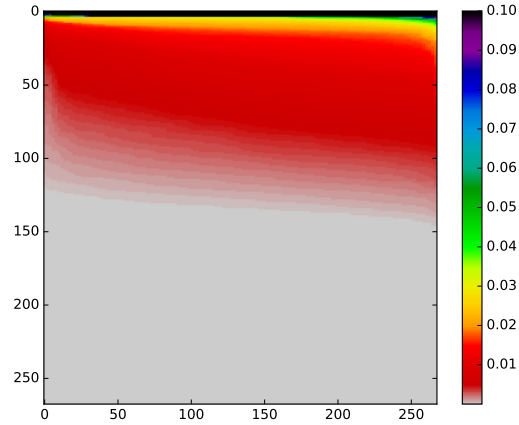
(a) $nnz = 26\%$



(b) $nnz = 72\%$



(c) $nnz = 86\%$



(d) $nnz = 89\%$

Fig. 2.11 Second form of illustration of the sparsity pattern of the density matrix truncated at $\tau = 10^{-8}$ during the SCF iterations, following the sequence (a) to (d). Results obtained for the CNT (11,5).

density matrix during the iterative procedure where, according to the legend color bar, the largest matrix elements are located on the diagonal. Starting from a sparse initial guess,[2] we observe how the DM is gradually filled up from step (a) to (d), to eventually reach convergence at $nnz = 89\%$. The alternative representation of Figure {2.11}, where the elements (in absolute value) are sorted in decreasing order (from the top to the bottom), emphasizes on how really sparse are the matrices we are dealing with, and the amplitude of variations of the nnz. Each colored map in Figure {2.11} presents an important decaying constrast from black (the largest elements) to grey (the smallest elements). From (a) to (b), one can see that the nnz region is progressively widening to reach half of the picture. Then, from (c) to (d), the amplitude of those elements —the intensity of the red color— is increasing.

2.4.3 Radial truncation for SCF calculations

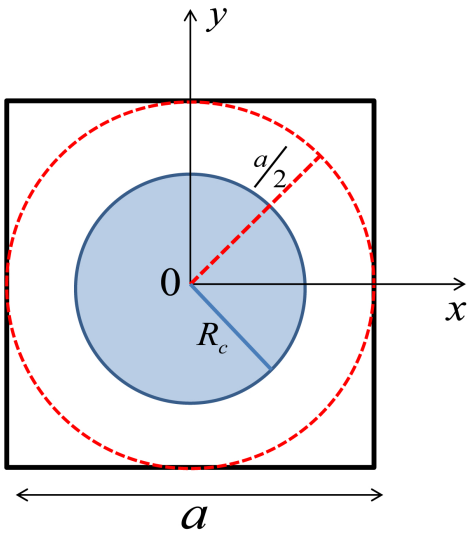


Fig. 2.12 Chart for the radial truncation scheme displaying the circle (blue solid line) of radius R_c. The largest circle (red dashed line) is of radius equal to $a/2$ inscribed in the square unit cell.

Whereas the numerical truncation operates directly on the DM elements without consideration on the density matrix topology, the radial truncation of Eq. (2.34) assumes that the relevant elements are localized within spheres of a certain radius centered on each atom. As a result, controlling the radius of the spheres, controls the number of non-zero elements which, compared to numerical truncation, has fixed positions within

[2] In this work the sparsity of the initial guess is dictated by the sparsity of the semi-empirical ZDO Fock matrix [cf. Section 1.4].

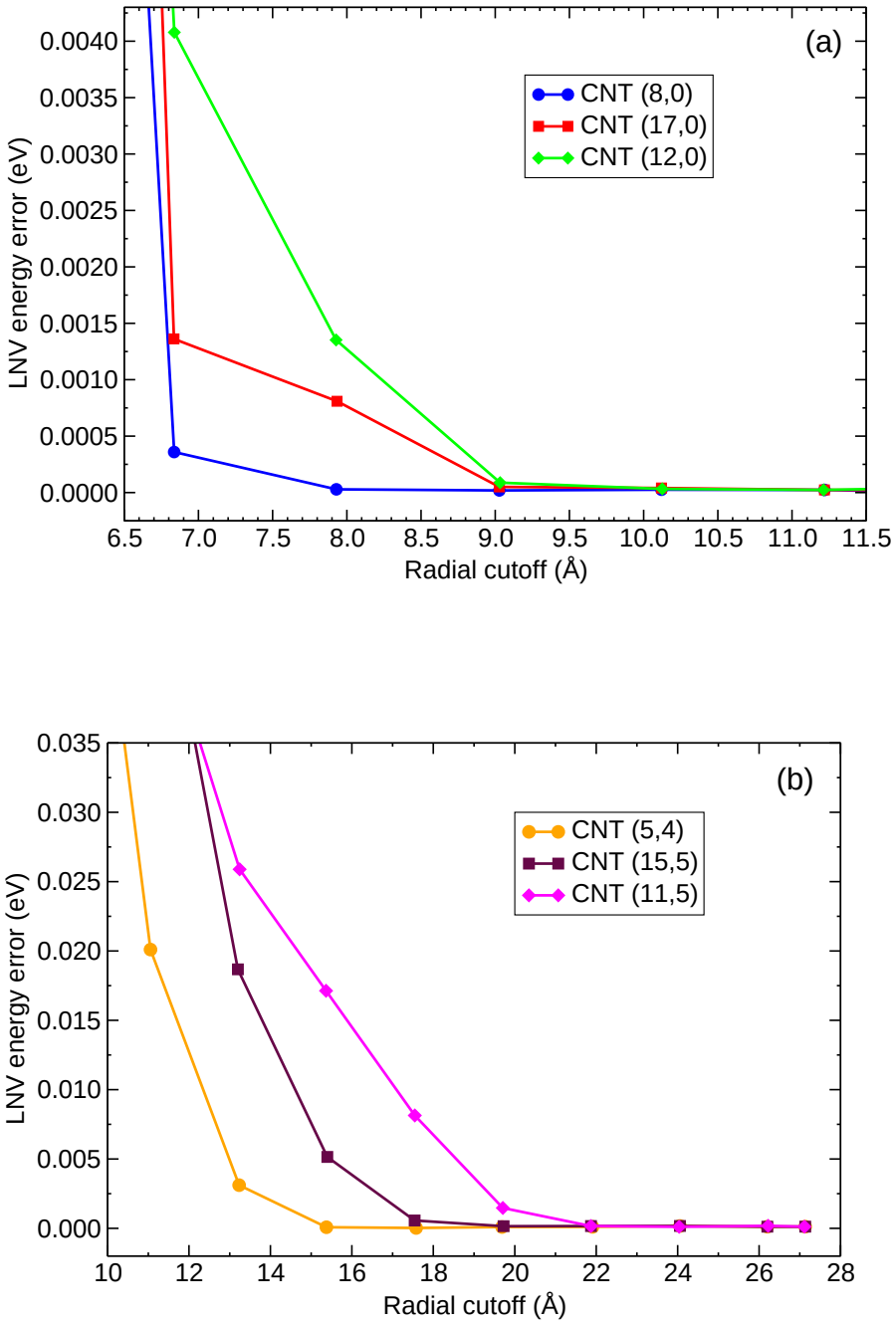


Fig. 2.13 Convergence of the LNV energy with respect to the exact value obtained from diagonalization (no truncation) as a function of the radial cutoff R_c. (a) Convergence profile obtained for the zigzag CNT. (b) Convergence profile obtained for the chiral CNT.

the DM network. As a result, the cutoff radius R_c fixes the superior limit beyond which the density matrix elements are enforced to be zero. In this work, the upper bound of R_c is defined by the lattice parameter a in the longitudinal-like direction, such that: $R_c \leq a/2$ [cf. Figure {2.12}]. It is clear that larger is the radial cutoff, higher is the

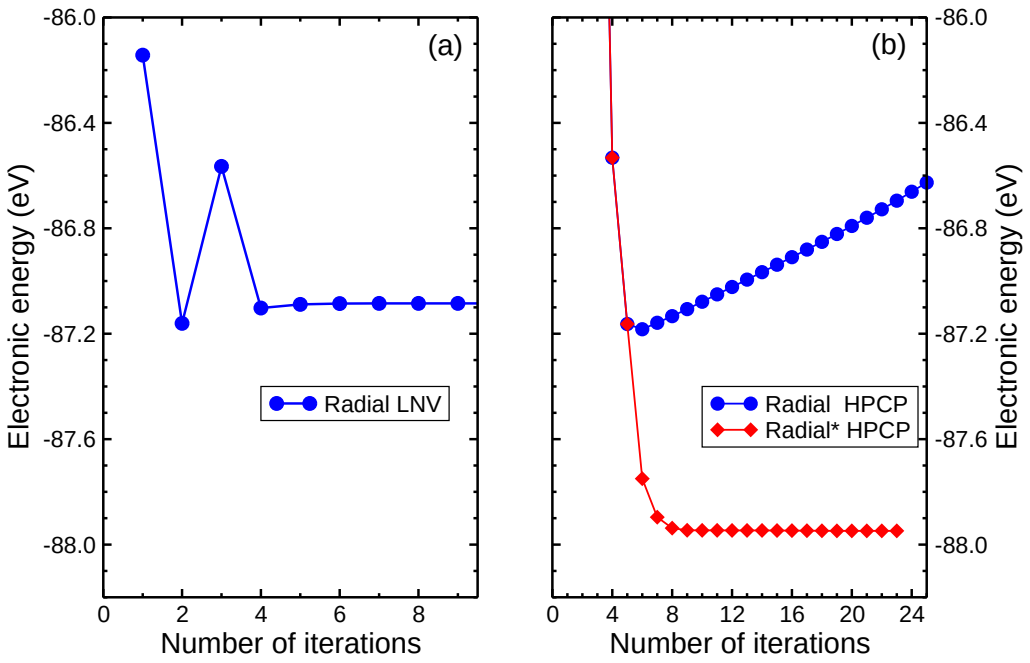
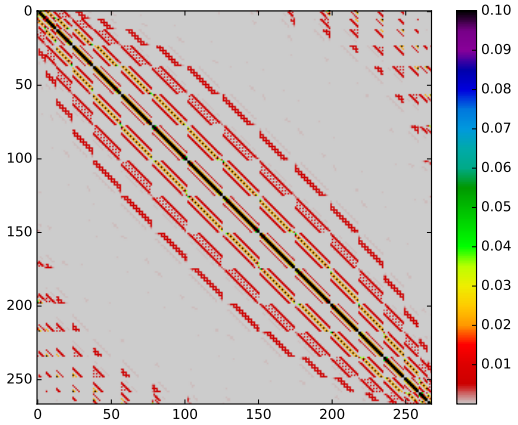
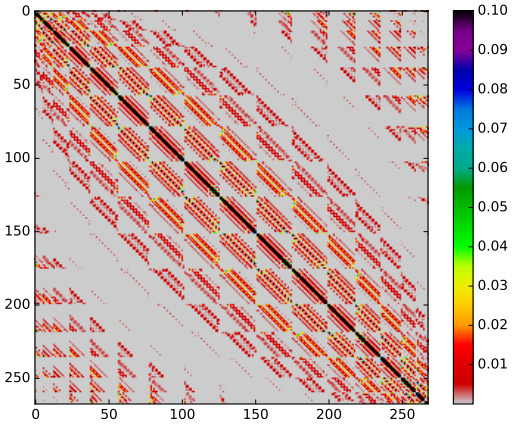


Fig. 2.14 Convergence of energy during the density matrix (a) minimization (LNV), and (b) purification (HPCP) for the CNT (8,0). A cutoff radius of 15 $\mathring{A}$ have been used (cf. text for more details).

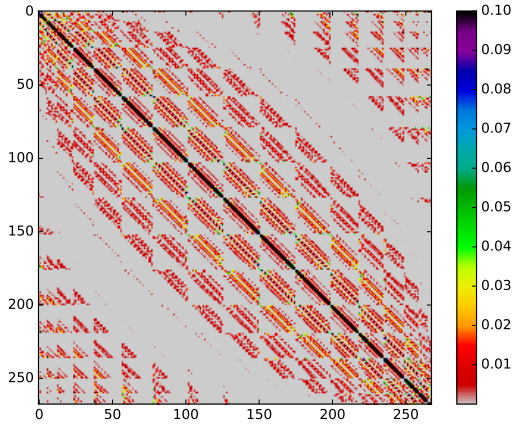
number of significant nnz of the density matrix. We have investigated the convergence of the SCF energy with respect to R_c. Results are displayed in Figure {2.13}. For each type of CNT —zigzag or chiral— we observe a variational-like convergence in agreement with the earlier works of Refs. [148, 149]. However, the purifications used with the radial truncation can be problematic for some cutoff radii. For instance, in Figure {2.14} is represented the convergence of the energy within a single SCF cycle (ie. for a fixed Fock matrix). For the variational trace-conserving purification method, we observe an increase of the energy after the 5th iteration [blue curve on panel (b)], which indicates that the (local) minimum can not be reached. Note that the relation $\mathcal{E}_{n+1} > \mathcal{E}_n$ was already proposed by Palser and Manolopoulos as the termination criterion when radial



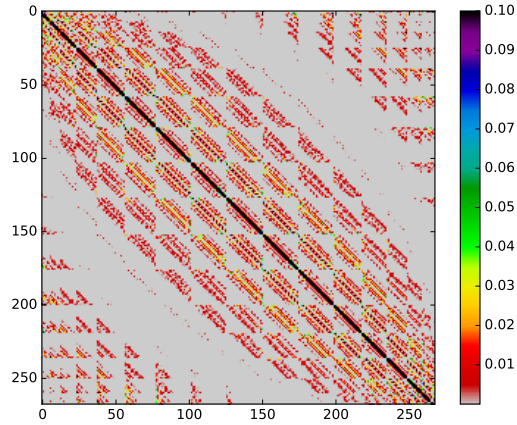
(a) $nnz = 26\%$



(b) $nnz = 38\%$

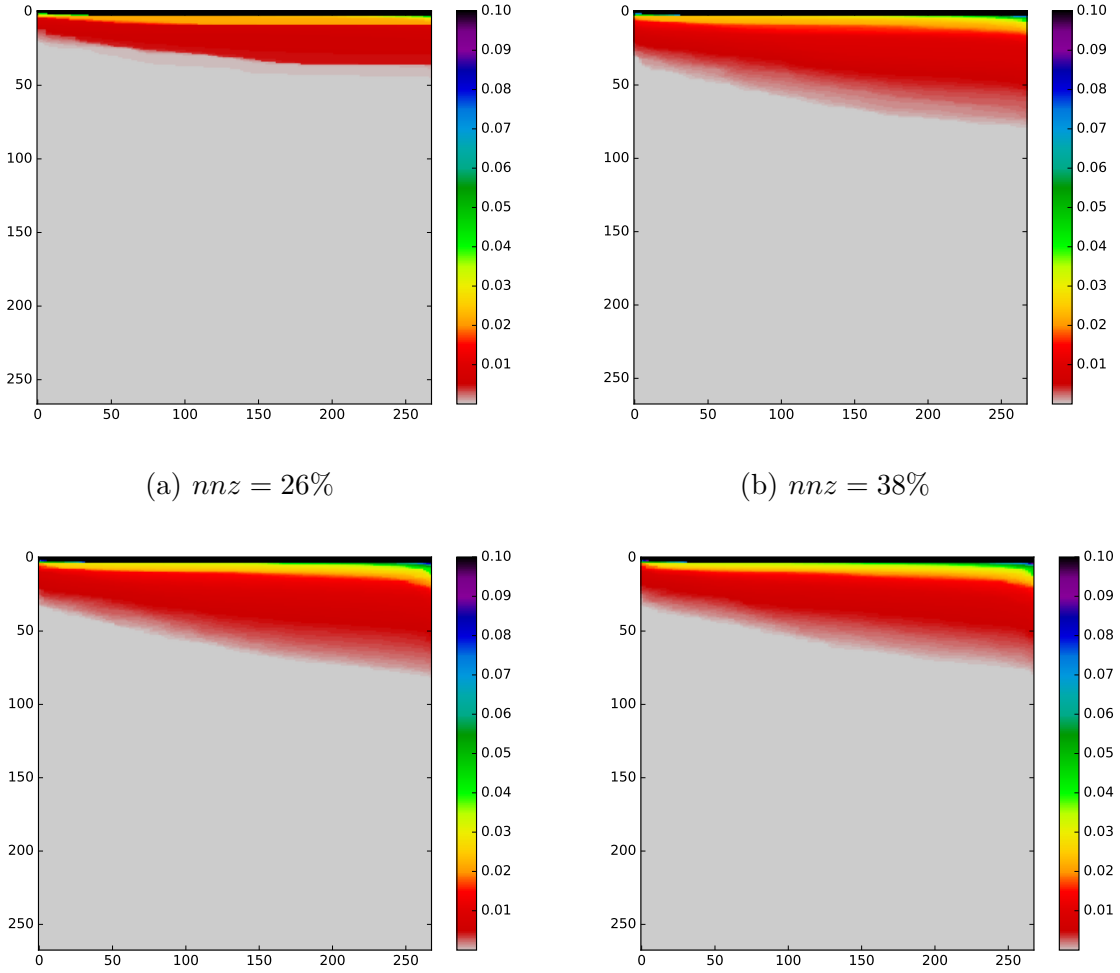


(c) $nnz = 39\%$



(d) $nnz = 39\%$

Fig. 2.15 First form of illustration for the progression of the density matrix truncated at $R_c = 10$ $\mathring{A}$ during the iterations, throughout the four respective sequences (a), (b), (c) and (d).



(a) $nnz = 26\%$ (b) $nnz = 38\%$

(c) $nnz = 39\%$ (d) $nnz = 39\%$

Fig. 2.16 Second form of illustration for the progression of the density matrix truncated at $R_c = 10$ $\mathring{A}$ during the iterations, throughout the four respective sequences (a), (b), (c) and (d).

truncation is applied.[104] As a result, unlike the numerical truncation which drops small elements disregarding their locations in the density matrix, and can be used with any of the density matrix solver, the radial truncation seems to be, for some of the cases treated here, incompatible with the purifications.

As an attempt of solution, we have performed the following test: when $\mathcal{E}_{n+1} > \mathcal{E}_n$, the radial truncation is switched off until $\mathcal{E}_{n'+1} < \mathcal{E}_{n'}$, where then, the truncation is switched back on. Result (denoted by Radial*) is dispalyed in red on the Figure {2.14}. In that case, we observed that the monotonic convergence is not interrupted at the crossing point corresponding to $n = 5$. Note that compared to the LNV minimization, the output energy (as obtained at the end of the purification) is lower. Nevertheless, for all the cases, we found that the LNV is more robust than the HPCP with regard to the radial truncation. In the following, only results obtained for the purifications and radial truncation at non-problematic cutoff radii are discussed. We now illustrate the radial truncation using the CNT (11,5) at $R_c = 10$ $\mathring{A}$, as already used for the numerical truncation in Section 2.4.2. The evolution of the density matrix sparsity during the SCF iterations is summarized in Figures {2.15} and {2.16}. Since the number of non-zero elements is predifined by the radial truncation, compared to Figures {2.10} and {2.11}, we do not observe the "let-it-grow" evolution of the numerical scheme. The progression of the red color region for the non-zero elements is limited.

2.4.4 Linear scaling SCF calculations and conclusion

Applications of the radial truncation to achieve linear scaling on large systems are presented in Figure {2.17}. For that purpose, we have replicated the CNT (11,5) in the longitudinal direction up to about 10,000 atoms. The figure displays the variation of the CPU time as a function of the system size. Calculations were performed using the diagonalization (Diag) and the following density matrix solvers: the standard LNV minimization, McWeeny purification (McW), trace correcting (TC2), canonical purification (PMCP), hole particle canonical purification (HPCP), and the trace resetting (TRS4). Figure {2.17(a)} shows that purifications are more efficient than the LNV minimization. In this ideal case of half-filling, the "fixed point" McWeeny polynomials presents the best performance. These results are in agreement with other comparative studies.[150, 151] The influence of the density matrix sparsity on the linear scaling behavior can be appreciated in Figure {2.17(b)}, where the calculation time is reduced by more than a factor 2 when the truncation radius is decreased from 50 to 10 $\mathring{A}$. Note in passing that, all the density matrix methods are cleary proved more efficient than the diagonalization. Using the two previous forms of matrix illustration [cf. Section 2.4.2 or Section 2.4.3], we

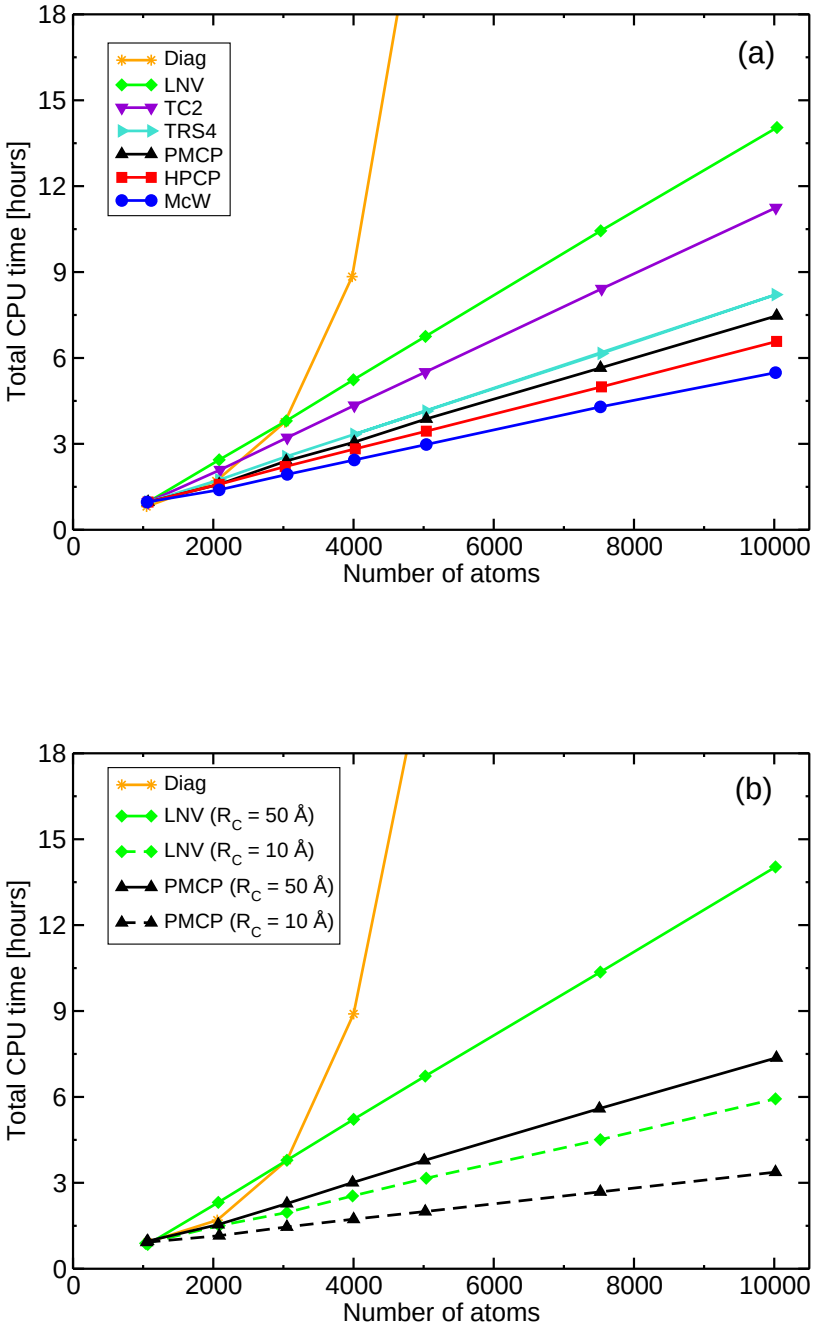
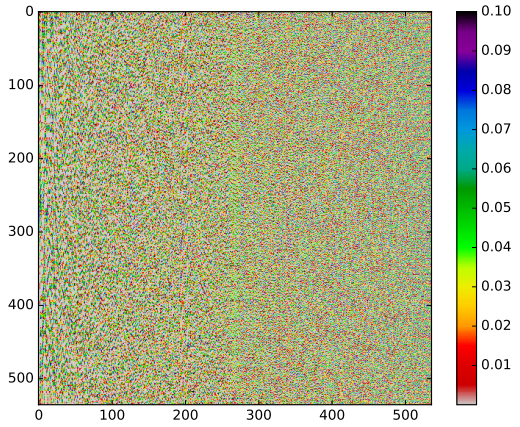
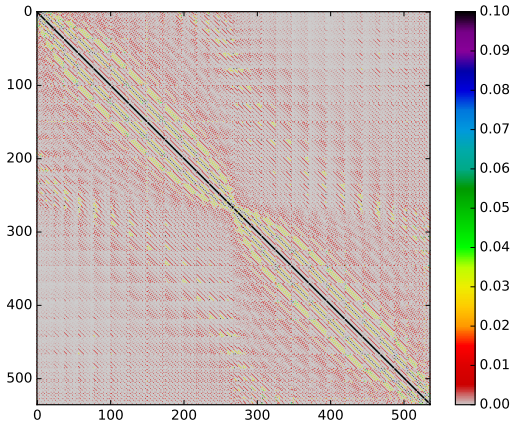


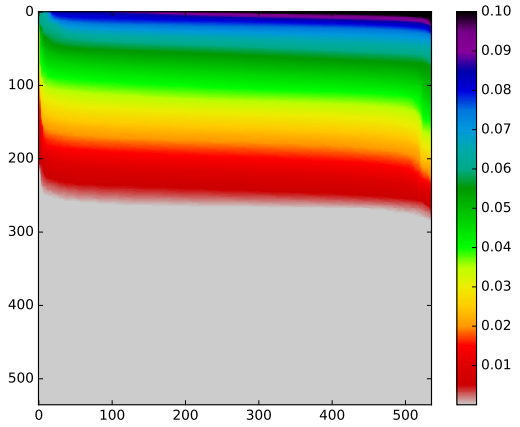
Fig. 2.17 Calculation time as a function of the number of atoms. Linear scaling regime is achieved using the radial truncation. (a) $R_c = 50$ $\AA$. (b) $R_c = 10$ and 50 $\AA$ Results were obtained for the replicated CNT (11,5).



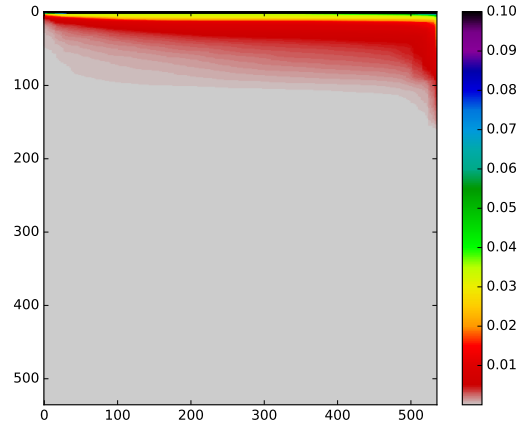
(a) C coefficients matrix



(b) D density matrix



(c) C coefficients matrix



(d) D density matrix

Fig. 2.18 Illustration of the C coefficients matrix for the diagonalization and of the untruncated D density matrix for the minimizations and purifications methods. (a) and (b) are the first form of illustration, while (c) and (d) are the second form of illustration, for the CNT (11,5).

display in Figure {2.18} the profile of the C and D matrices obtained at the end of SCF convergence, for the CNT (11,5). C is completely full, while D is sparse with decaying profile. That justifies the net difference between the diagonalization scaling and density matrix methods scaling.

In this chapter we have shown standard diagonalization can be circumvented using density matrix based solvers, which, when combined with sparse matrix algebra can lead to linear scaling regime. We found that, enforcing sparsity by the numerical approach is more robust compared to the radial truncation. Nevertheless, in that case, the number of non-zero elements is controlled *a posteriori*. For all the density matrix purifications, we found that radial truncation may cause convergence problems. Obviously, definitive conclusions on the advantages and drawbacks of the various schemes are beyond the scope of this work. A more systematic comparison using non-orthogonal basis sets of various sizes, along with a browder set of systems should be beneficial to identify and analyse limitations in accuracy and stability of $O(N)$ methods. The main aim of this dissertation is to derive and test linear scaling algorithms for density matrix perturbation theory, which will be introduced in the next chapter.

Chapter 3

Density matrix perturbation theory

In Chapters 1 and 2 of this manuscript, we have presented the resolution of time-independent Liouville-von Neumann equation in order to determine the energy of a molecular system. However, to relate the results from quantum chemical calculations to experiment, it is essential to compute quantities that are directly comparable to measurements. For this purpose, the density matrix of the ground state obtained from the resolution of the Liouville-von Neumann equation is not sufficient. It is therefore necessary to compute further quantities that characterize the molecular system of interest. Theses quantities can be classified as follow:

1. Energy differences, such as reaction energies, dissociation energies, that involve energy information at different points on the Born-Oppenheimer potential energy surface.

2. Molecular properties, like dipole moment, polarizabilities, vibrational frequencies, nuclear magnetic resonance parameters, that require information of perturbed electronic states at a single point on the potential surface.

3. Transitions energy between different electronic states, as for instance, electronic excitation energies, radiative life times, that involve information for electronic states coupling.

The concept of spectroscopy refers to the observation of a physical phenomenon onto an energy scale or any quantity related to an energy, like frequency or wave length. Until nowadays, the spectroscopy principle is greatly expanding in many research fields such as astronomy, biophysics, chemistry, acoustics. The basic idea of the spectroscopy consists to subject to radiation the matter and measure the response. Comparing the original radiation with the response, one can extract information related to some of its intrinsic properties, for instance: structural, electronic or magnetic. A specific radiation, ie. wavelength range, allows to probe specific properties of the system. For the second class of properties enumerated above, the perturbation theory is necessary (and sufficient) to access various spectroscopic observables. This Chapter takes a look beyond the time-independent Liouville-von Neumann equation, to derive density matrix time-independent perturbation theory.

3.1 Theoretical background

Let us consider an unperturbed system of N occupied and $\bar{N}$ unoccupied molecular states expanded onto a linear combination of M atomic basis functions, such that $N + \bar{N} = M$. Considering an orthonormalized set of basis functions [cf. Section 1.4], we recall the generalized constraints on the one-particle density matrix D for the occupied states and the one-hole density matrix $\bar{D}$ for the unoccupied eigenstates, are given as

$$\text{Idempotency:} \quad D^2 = D \quad \text{and} \quad \bar{D}^2 = \bar{D} \tag{3.1a}$$

$$\text{Trace conservation:} \quad \mathrm{Tr}\{D\} = N \quad \text{and} \quad \mathrm{Tr}\{\bar{D}\} = \bar{N} \tag{3.1b}$$

$$\text{Complementarity:} \quad D + \bar{D} = I \tag{3.1c}$$

$$\text{Orthogonality:} \quad D\bar{D} = 0 \tag{3.1d}$$

where Eqs. (3.1a) and (3.1b) are the N-representability conditions for a single Slater determinant. Along with Eq. (3.1), the ground state is guaranteed if the density matrix and the Fock matrix fulfill the SCF conditions

$$FD - DF = 0 \tag{3.2}$$

$$F\bar{D} - \bar{D}F = 0 \tag{3.3}$$

The density matrix methods such as the minimizations and the purifications, highlighted in Chapter 2, are used for solving the following unperturbed SCF equations

$$F\left(D_n\right) D_{n+1} - D_{n+1} F\left(D_n\right) = 0 \tag{3.4a}$$

$$\text{subject to:} \quad \mathrm{Tr}\{D_n\} = N \tag{3.4b}$$

$$D_n^2 = D_n, \ \forall\ n \tag{3.4c}$$

Let us now consider that this system is disturbed by an external time-independent perturbation, which takes the form of a matrix W. The perturbed Fock and one-particle density matrices formally read

$$F_\lambda := F + \lambda W_{\mathrm{F}} \tag{3.5a}$$

$$D_\lambda := D + \lambda W_{\mathrm{D}} \tag{3.5b}$$

where we have introduced a scaling parameter $\lambda \in [0, 1]$ to modulate the strength of the perturbation. Consequently, the analogue of the non-perturbed SCF equations (3.4), for

the perturbed case, are given by

$$F_\lambda\left(D_{\lambda,n}\right) D_{\lambda,n+1} - D_{\lambda,n+1} F_\lambda\left(D_{\lambda,n}\right) = 0 \tag{3.6a}$$

$$\text{subject to: } \mathrm{Tr}\{D_{\lambda,n}\} = N, \tag{3.6b}$$

$$D_{\lambda,n}^2 = D_{\lambda,n}, \ \forall\ n \tag{3.6c}$$

Since we are concerned about the response of the system to the perturbation, we need to evaluate the variation of the energy with respect to an infinitesimal variation of the perturbation strength. It is worth to note that the perturbation theory is valid if the response is small compared to the eigenspectrum of F. There exist two ways for evaluating the energy response: (i) the numerical solution based on finite differences [cf. Appendix E], or (ii) the analytical solution based on the perturbation expansion of the relevant quantities. Density matrix perturbation theory presented in this Chapter is obviously dealing with the second possiblity. In that context, one can write the matrices F_λ and D_λ down as power series in λ following the standard Taylor expansion:

$$D_\lambda = D + \lambda D^{(1)} + \lambda^2 D^{(2)} + \ldots + \lambda^k D^{(k)} \tag{3.7a}$$

$$F_\lambda = F + \lambda F^{(1)} + \lambda^2 F^{(2)} + \ldots + \lambda^k F^{(k)} \tag{3.7b}$$

with $D^{(k)}$ and $F^{(k)}$ the shorthand notations for the kth order variation of the density and Fock matrix. More explicitly, they should read

$$D^{(k)} := \frac{1}{k!}\frac{d^k D}{d\lambda^k} \tag{3.8a}$$

$$F^{(k)} := \frac{1}{k!}\frac{d^k F}{d\lambda^k} \tag{3.8b}$$

From here, the idea is to apply the perturbation theory on the stationary conditions of Eq. (3.6). On inserting the expansion of Eq. (3.7a) into (3.6c), and equating the perturbation orders of the left and right sides, we obtain the following perturbed idempotency

relations:

$$DD^{(1)} + D^{(1)}D = D^{(1)} \tag{3.9a}$$

$$DD^{(2)} + (D^{(1)})^2 + D^{(2)}D = D^{(2)} \tag{3.9b}$$

$$DD^{(3)} + D^{(1)}D^{(2)} + D^{(2)}D^{(1)} + D^{(3)}D = D^{(3)} \tag{3.9c}$$

$$\vdots$$

$$DD^{(k)} + D^{(1)}D^{(k-1)} + D^{(2)}D^{(k-2)} + ... + D^{(k)}D = D^{(k)} \tag{3.9d}$$

$$\text{or more generally:} \quad D^{(k)} = \sum_{l=0}^{k} D^{(l)}D^{(k-l)} \tag{3.9e}$$

Then, by repeating the perturbation identification on the Liouville-von Neumann equation, that is by inserting Eqs. (3.7a) and (3.7b) into Eq. (3.6a), we obtain the perturbed SCF conditions:

$$[F, D^{(1)}] + [F^{(1)}, D] = 0 \tag{3.10a}$$

$$[F, D^{(2)}] + [F^{(1)}, D^{(1)}] + [F^{(2)}, D] = 0 \tag{3.10b}$$

$$[F, D^{(3)}] + [F^{(1)}, D^{(2)}] + [F^{(2)}, D^{(1)}] + [F^{(3)}, D] = 0 \tag{3.10c}$$

$$\vdots$$

$$[F, D^{(k)}] + [F^{(1)}, D^{(k-1)}] + ... + [F^{(k)}, D] = 0 \tag{3.10d}$$

$$\text{or more generally:} \quad \sum_{l=0}^{k} [F^{(l)}, D^{(k-l)}] = 0 \tag{3.10e}$$

The perturbed idempotency constraints (3.9) and SCF conditions (3.10) are both the background of the density matrix perturbation theories which are presented in the next Section. From these relations, we may also emphasize that the evaluation of the perturbed density matrix at order (k) is based on the knowledge of the order $(k-1)$. In other words, evaluation of the perturbed quantities requires —in principle— to proceed order by order.

3.2 Wavefunction coupled perturbed self-consistent field formulation

The standard density matrix perturbation method is the atomic orbitals coupled perturbed self-consistent-field[25, 17, 26, 152] (AO-CPSCF). In order to derive it, we simply need to decompose the non-perturbed Fock matrix in terms of its eigenvalues and (one-state) projection operators built from the eigenvectors.

The density matrix perturbation theory as developed by Diercksen and McWeeny[26] is based on the partitioning of $D^{(k)}$ into four distinct contributions and their resolutions. Using the closure relation (3.1c), any operator X can be recast into the following projected components:

$$X = (D + \bar{D})X(D + \bar{D}) \tag{3.11a}$$

$$= DXD + DX\bar{D} + \bar{D}XD + \bar{D}X\bar{D} \tag{3.11b}$$

$$= X_{oo} + X_{ov} + X_{vo} + X_{vv} \tag{3.11c}$$

where the subscripts (oo) and (vv) designate the occupied-occupied, and virtual-virtual contributions related to the original orthogonal subspaces $\mathcal{H}_{\text{occ}}$ and $\mathcal{H}_{\text{virt}}$, for the occupied states and for the unoccupied states, respectively. Likewise, (ov) and (vo) stand for the perturbation induced by coupling terms associated with subspaces $\mathcal{H}_{\text{occ-virt}}$ and $\mathcal{H}_{\text{virt-occ}}$, respectively. The spectral resolution of Eq. (1.94b) already introduced in Section 1.5 allows any projected matrices of Eq. (3.11c) to be decomposed into a sum of single-projected components, following:

$$X_{oo} = DX\bar{D} = \sum_{i,j} D_i X D_j \tag{3.12a}$$

$$X_{ov} = DX\bar{D} = \sum_{i,j} D_i X \bar{D}_j \tag{3.12b}$$

$$X_{vo} = DX\bar{D} = \sum_{i,j} \bar{D}_i X D_j \tag{3.12c}$$

$$X_{vv} = DX\bar{D} = \sum_{i,j} \bar{D}_i X \bar{D}_j \tag{3.12d}$$

Note that the spectral resolution (3.12) implies the eigenstates to be known, which irremediably involves the diagonalization of the unperturbed Fock matrix.

3.2.1 First-order response

Let us start with the first-order of perturbation. On applying the projection decomposition of Eq. (3.11) to both sides of Eq. (3.9a), we obtain:

$$2D^{(1)}_{oo} + D^{(1)}_{ov} + D^{(1)}_{vo} = D^{(1)}_{oo} + D^{(1)}_{ov} + D^{(1)}_{vo} + D^{(1)}_{vv} \tag{3.13}$$

Proceeding by identification, it can be deduced that

$$D^{(1)}_{oo} = 0, \quad D^{(1)}_{vv} = 0 \tag{3.14}$$

and

$$D^{(1)} = D^{(1)}_{ov} + D^{(1)}_{vo} \tag{3.15a}$$

$$= D^{(1)}_{ov} + D^{(1)\dagger}_{ov} \tag{3.15b}$$

For the last statement (3.15b), we relied on the symmetry property of the perturbed density matrix, that is, $D^{(1)\dagger} = D^{(1)}$. The relation (3.15b) demonstrates that the determination of $D^{(1)}$ involves only the evaluation of the occupied-virtual transition matrix. For that purpose, multiplying Eq. (3.10a) by D from the left, and by $\bar{D}$ from the right, we obtain

$$F^{(1)}_{ov}(D^{(1)}_{ov,n}) = [F, D^{(1)}_{ov,n+1}] \tag{3.16a}$$

$$\text{subject to: } \mathrm{Tr}\{D^{(1)}_{ov,n}\} = 0, \ \forall n \tag{3.16b}$$

where we have re-introduced the iteration indice n of the self-consistent resolution, and the dependence of the perturbation matrix $F^{(1)}_{ov}$ over $D^{(1)}_{ov}$.[1] This relation constitutes the first-order —coupled perturbed self-consistent field (CPSCF)— density-matrix equations, analogous to the non-perturbed Eq. (3.4). Given an initial guess for the first-order perturbed density matrix $D^{(1)}_{ov}$ [cf. Eq. (3.14)], we project the first-order perturbed Fock matrix $F^{(1)}$, to build $F^{(1)}_{ov}$, and finally solve Eq. (3.16). The iterative process is repeated until convergence is achieved. By substituting Eq. (1.94a) into (3.16), one obtains

$$\sum_{i,j} \left\{ D^{(1)}_{ov,ij}(\epsilon_i - \bar{\epsilon}_j) - F^{(1)}_{ov,ij} \right\} = 0 \tag{3.17a}$$

$$\text{with: } D^{(1)}_{ov,ij} = D_i D^{(1)} \bar{D}_j \tag{3.17b}$$

$$F^{(1)}_{ov,ij} = D_i F^{(1)} \bar{D}_j \tag{3.17c}$$

which yields to the well-known sum-over-states (SOS) first-order equation:

$$\sum_{i,j} D^{(1)}_{ov,ij} = \sum_{i,j} \frac{F^{(1)}_{ov,ij}}{(\epsilon_i - \bar{\epsilon}_j)} \quad \Leftrightarrow \quad D^{(1)}_{ov} = \sum_{i,j} \frac{F^{(1)}_{ov,ij}}{(\epsilon_i - \bar{\epsilon}_j)} \tag{3.18}$$

[1] Latter in this Chapter the dependence of the perturbed Fock matrices over the perturbed density matrices will be considered as implicit.

3.2.2 Second-order response

The second-order equation can be derived by applying the resolution of identity to both side of Eq. (3.9b). Keeping the notations of Eq. (3.11), we have

$$\begin{aligned} 2D^{(2)}_{oo} + D^{(2)}_{ov} + D^{(2)}_{vo} \quad &+ \quad (D^{(1)}D^{(1)})_{oo} + (D^{(1)}D^{(1)})_{vv} + (D^{(1)}D^{(1)})_{ov} \\ &+ \quad (D^{(1)}D^{(1)})_{vo} = D^{(2)}_{oo} + D^{(2)}_{ov} + D^{(2)}_{vo} + D^{(2)}_{vv} \end{aligned} \tag{3.19}$$

By resolving the product of first-order perturbed density matrices according to

$$\begin{aligned} D^{(1)}D^{(1)} &= D^{(1)}ID^{(1)} \\ &= D^{(1)}(D + \bar{D})D^{(1)} \\ &= D^{(1)}(D^2 + \bar{D}^2)D^{(1)} \end{aligned}$$

we obtain

$$(D^{(1)}D^{(1)})_{oo} = D^{(1)}_{oo}D^{(1)}_{oo} + D^{(1)}_{ov}D^{(1)}_{vo} \tag{3.21a}$$

$$(D^{(1)}D^{(1)})_{vv} = D^{(1)}_{vv}D^{(1)}_{vv} + D^{(1)}_{vo}D^{(1)}_{ov} \tag{3.21b}$$

$$(D^{(1)}D^{(1)})_{ov} = D^{(1)}_{ov}D^{(1)}_{vv} + D^{(1)}_{oo}D^{(1)}_{ov} \tag{3.21c}$$

$$(D^{(1)}D^{(1)})_{vo} = D^{(1)}_{vo}D^{(1)}_{oo} + D^{(1)}_{vv}D^{(1)}_{vo} \tag{3.21d}$$

On inserting the right-hand side (rhs) of Eqs. (3.21) into (3.19), and using the properties (3.14), we have

$$\begin{aligned} 2D^{(2)}_{oo} + D^{(2)}_{ov} + D^{(2)}_{vo} + D^{(1)}_{ov}D^{(1)}_{vo} + D^{(1)}_{vo}D^{(1)}_{ov} \\ = D^{(2)}_{oo} + D^{(2)}_{ov} + D^{(2)}_{vo} + D^{(2)}_{vv} \end{aligned} \tag{3.22}$$

Therefore, it comes

$$D^{(2)}_{oo} = -D^{(1)}_{ov}D^{(1)}_{vo}, \quad D^{(2)}_{vv} = +D^{(1)}_{vo}D^{(1)}_{ov} \tag{3.23}$$

Unlike the 1st-order perturbation, the diagonal components of the 2nd-order perturbed density matrix are likely to be non zero and can be computed from the 1st-order perturbed density matrix. Relying furthermore on the symmetry of the perturbed density, it leaves only the occupied-virtual coupling matrix to evaluate. On resolving the 2nd-order perturbed Fock matrix using Eq. (3.10b), we obtain

$$F^{(2)}_{ov} = [F, D^{(2)}_{ov}] + [F^{(1)}, D^{(1)}]_{ov} \tag{3.24}$$

Using the spectral resolution of the non-perturbed Fock matrix and the perturbed density matrix, Eq. (3.24) transforms as

$$F^{(2)}_{ov} = \sum_{i,j} \left(D^{(2)}_{ov,ij}(\epsilon_i - \bar{\epsilon}_j) + [F^{(1)}, D^{(1)}]_{ov,ij} \right) \tag{3.25}$$

which leads to the 2nd-order SOS equation

$$D^{(2)}_{ov} = \sum_{i,j} (\epsilon_i - \bar{\epsilon}_j)^{-1} \left(F^{(2)}_{ov} - [F^{(1)}, D^{(1)}]_{ov} \right)_{ij} \tag{3.26}$$

The final 2nd-order perturbed density matrix is obtained summing over $D^{(2)}_{ov}$, its conjugate-transposed, $D^{(2)}_{vo}$, and the diagonal contributions of Eq. (3.23).

3.2.3 Third-order response

Using the same route than for the first- and second-order, the third-order response equations are derived from Eq. (3.9c). This yields to

$$\begin{aligned} 2D^{(3)}_{oo} \;+\; & D^{(3)}_{ov} + D^{(3)}_{vo} + (D^{(1)}D^{(2)})_{oo} + (D^{(1)}D^{(2)})_{vv} + (D^{(1)}D^{(2)})_{ov} \\ + \; & (D^{(1)}D^{(2)})_{vo} + (D^{(2)}D^{(1)})_{oo} + (D^{(2)}D^{(1)})_{vv} + (D^{(2)}D^{(1)})_{ov} \\ + \; & (D^{(2)}D^{(1)})_{vo} = D^{(3)}_{oo} + D^{(3)}_{ov} + D^{(3)}_{vo} + D^{(3)}_{vv} \end{aligned} \tag{3.27}$$

where

$$(D^{(1)}D^{(2)})_{oo} = D^{(1)}_{oo}D^{(2)}_{oo} + D^{(1)}_{ov}D^{(2)}_{vo} \tag{3.28a}$$
$$(D^{(1)}D^{(2)})_{vv} = D^{(1)}_{vv}D^{(2)}_{vv} + D^{(1)}_{vo}D^{(2)}_{ov} \tag{3.28b}$$
$$(D^{(1)}D^{(2)})_{ov} = D^{(1)}_{ov}D^{(2)}_{vv} + D^{(1)}_{oo}D^{(2)}_{ov} \tag{3.28c}$$
$$(D^{(1)}D^{(2)})_{vo} = D^{(1)}_{vo}D^{(2)}_{oo} + D^{(1)}_{vv}D^{(2)}_{vo} \tag{3.28d}$$
$$(D^{(2)}D^{(1)})_{oo} = D^{(2)}_{oo}D^{(1)}_{oo} + D^{(2)}_{ov}D^{(1)}_{vo} \tag{3.28e}$$
$$(D^{(2)}D^{(1)})_{vv} = D^{(2)}_{vv}D^{(1)}_{vv} + D^{(2)}_{vo}D^{(1)}_{ov} \tag{3.28f}$$
$$(D^{(2)}D^{(1)})_{ov} = D^{(2)}_{ov}D^{(1)}_{vv} + D^{(2)}_{oo}D^{(1)}_{ov} \tag{3.28g}$$
$$(D^{(2)}D^{(1)})_{vo} = D^{(2)}_{vo}D^{(1)}_{oo} + D^{(2)}_{vv}D^{(1)}_{vo} \tag{3.28h}$$

On inserting Eqs. (3.14) and (3.23) into (3.28), simplifies Eq. (3.27) to:

$$\begin{aligned} 2D^{(3)}_{oo} + D^{(3)}_{ov} + D^{(3)}_{vo} \;+\; & D^{(1)}_{ov}D^{(2)}_{vo} + D^{(2)}_{ov}D^{(1)}_{vo} + D^{(1)}_{vo}D^{(2)}_{ov} + D^{(2)}_{vo}D^{(1)}_{ov} \\ = \; & D^{(3)}_{oo} + D^{(3)}_{ov} + D^{(3)}_{vo} + D^{(3)}_{vv} \end{aligned} \tag{3.29}$$

Consequently, it follows

$$D_{oo}^{(3)} = -\left(D_{ov}^{(1)} D_{vo}^{(2)} + D_{ov}^{(2)} D_{vo}^{(1)}\right) \tag{3.30a}$$

$$D_{vv}^{(3)} = +\left(D_{vo}^{(1)} D_{ov}^{(2)} + D_{vo}^{(2)} D_{ov}^{(1)}\right) \tag{3.30b}$$

Once again, these last equations show that the 2nd-order density matrix is necessary and sufficient to compute the 3rd order diagonal components. Again, at this point, we emphasize that only the occupied-virtual transition matrix needs to be evaluated self-consistently since the perturbed density matrix is (at least) Hermitian. Using Eq. (3.10c), the 3rd-order perturbed Fock matrix reads

$$F_{ov}^{(3)} = [F, D_{ov}^{(3)}] + [F^{(1)}, D^{(2)}]_{ov} + [F^{(2)} D^{(1)}]_{ov} \tag{3.31}$$

Relying on the spectral resolution, this relation transforms to

$$F_{ov}^{(3)} = \sum_{i,j} D_{ov,ij}^{(3)} (\epsilon_i - \bar{\epsilon}_j) + \left([F^{(1)}, D^{(2)}]_{ov} + [F^{(2)} D^{(1)}]_{ov}\right)_{ij} \tag{3.32}$$

which leads to the 3rd-order SOS equation

$$D_{ov}^{(3)} = \sum_{i,j} (\epsilon_i - \bar{\epsilon}_j)^{-1} \left(F_{ov}^{(3)} - [F^{(1)}, D^{(2)}]_{ov} - [F^{(2)}, D^{(1)}]_{ov}\right)_{ij} \tag{3.33}$$

3.2.4 *k*th-order response

As a matter of fact, from the expansions (3.9) and (3.10), we can generalize the response equations to any kth-order, by proceeding in the same way and using the components from the lower orders. Then, the (diagonal) fixed components are defined according to

$$D_{oo}^{(k)} = -\sum_{i=1}^{k-1} \left(D_{oo}^{(i)} D_{oo}^{(k-i)} + D_{ov}^{(i)} D_{vo}^{(k-i)}\right) \tag{3.34}$$

$$D_{vv}^{(k)} = \sum_{i=1}^{k-1} \left(D_{vv}^{(i)} D_{vv}^{(k-i)} + D_{vo}^{(i)} D_{ov}^{(k-i)}\right) \tag{3.35}$$

whereas the (off-diagonal) transition matrices, which involve a self-consistent resolution, are defined according to

$$D_{ov}^{(k)} = \sum_{i,j} (\epsilon_i - \bar{\epsilon}_j)^{-1} \left(F_{ov}^{(k)} - M_{ov}^{(k)}\right)_{ij} \tag{3.36}$$

$$M^{(k)}_{ov,ij} = \sum_{l=1}^{k-1} \left((D^{(k)} F^{(k-l)})_{ov} - (F^{(k)} D^{(k-l)})_{ov} \right)_{ij} \tag{3.37}$$

using the shorthand notations

$$\begin{aligned} F^{(k)}_{ov,ij} &= D_i F^{(k)} \bar{D}_j \\ (D^{(k)} F^{(k-l)})_{ov,ij} &= D_i D^{(k)} F^{(k-l)} \bar{D}_j, \\ (F^{(k)} D^{(k-l)})_{ov,ij} &= D_i F^{(k)} D^{(k-l)} \bar{D}_j \end{aligned}$$

The diagonal components are used as appropriate guess to initiate the perturbed density matrix whereas the off-diagonal components are taken as the iterative part using the eigenvectors of the non perturbed Fock matrix.

3.3 Density matrix coupled perturbed self-consistent field formulation

In the AO-CPSCF formalism, calculation of the perturbed density matrix $D^{(k)}$ requires the eigenstates of the unperturbed Fock matrix, which implies a high computational effort for systems of increasing size. In this work, we also consider the density matrix-based perturbation formalism originally proposed by Oschenfeld and Head-Gordon,[37] and reformulated by Oschenfeld and Kussmann.[43, 44, 153] The approach relies on solving —self-consistently— the commutation relations (3.10) using a conjugate-gradient-based minimization. It will be referred in this manuscript to as CG-CPSCF. The simple derivation[43] of the CG-CPSCF formalism from the background equations (3.9) and (3.10) is presented below.

3.3.1 First-order response

The idea behind the Oschenfeld and Kussmann method is to constrain Eq. (3.10a) to commute with the unperturbed density matrix. That is, multiplying Eq. (3.10a) from the left and from the right separately by D, and substracting, leads to

$$\left[F, \left[D, D^{(1)}\right]\right] + 2DF^{(1)}D - \left\{D, F^{(1)}\right\} = 0 \tag{3.39}$$

This correponds to the first-order CG-CPSCF equation. Relying on the symmetry properties for the perturbed density and Fock matrices, this equation is also Hermitian. It is worthwhile to note that on multiplying the CG-CPSCF equation (3.39) from the left

by D, and from the right by $\bar{D}$, the AO-CPSCF equation Eq. (3.16) is recovered. This clearly demonstrates the relationships between the two formalisms. Nevertheless, unlike the AO-CPSCF, the resolution of the first-order CG-CPSCF equation yields directly to the first-order perturbed density matrix $D^{(1)}$.

3.3.2 Second- and third-order response

If we want to write down the second-order CG-CPSCF equations using only the perturbed density matrices, that is, without relying on the resolution of the identity given in Eq. (3.11), we need to re-formulate the diagonal components in terms of D. Then, if we start from Eqs. (3.14), (3.21a) and (3.21b), we can recast Eq. (3.23) as

$$\begin{aligned} D^{(2)}_{oo} &= -(D^{(1)}D^{(1)})_{oo} = -DD^{(1)}D^{(1)}D \\ D^{(2)}_{vv} &= +(D^{(1)}D^{(1)})_{vv} = +\bar{D}D^{(1)}D^{(1)}\bar{D} \end{aligned}$$

If we want to circumvent the use of $\bar{D}$ to define $D^{(2)}_{vv}$ and use instead D, first, we should perform the following transformations:

$$\begin{aligned} D^{(2)}_{vv} &= \bar{D}D^{(1)}D^{(1)}\bar{D} \\ &= (D^{(1)}D^{(1)})_{vv} \\ &= D^{(1)}_{vv}D^{(1)}_{vv} + D^{(1)}_{vo}D^{(1)}_{ov} \quad \text{[using Eq. (3.21b)]} \\ &= D^{(1)}_{vo}D^{(1)}_{ov} \quad \text{[using Eq. (3.14)]} \\ &= (\bar{D}D^{(1)}D)(DD^{(1)}\bar{D}) \\ &= \bar{D}(D^{(1)}DD^{(1)})\bar{D} \quad \text{[using Eq. (3.1a)]} \end{aligned} \tag{3.40}$$

Also, by noting that:

$$\begin{aligned} D^{(2)}_{vv} &= \bar{D}D^{(2)}\bar{D} \\ &= \bar{D}(\bar{D}D^{(2)}\bar{D})\bar{D} \\ &= \bar{D}(D^{(2)}_{vv})\bar{D} \end{aligned} \tag{3.41}$$

we can deduce the expression for 2nd order CG-CPSCF initial guess:

$$D^{(2)}_{oo} = -DD^{(1)}D^{(1)}D,\ D^{(2)}_{vv} = D^{(1)}DD^{(1)} \tag{3.42}$$

As for the first-order response, the commutator between the non-perturbed density matrix and Eq. (3.9b) yields to the second-order CG-CPSCF equation,[2] following:

$$\left[F,\left[D,D^{(2)}\right]\right]+2DF^{(2)}D-\left\{D,F^{(2)}\right\}=\left[D,\left[D^{(1)},F^{(1)}\right]\right] \tag{3.43}$$

Using Eqs. (3.14) and (3.28) Eq. (3.28f), the diagonal contributions for the 3rd-order response can be recast as:

$$\begin{aligned}D^{(3)}_{oo}&=-\left((D^{(2)}D^{(1)})_{oo}+(D^{(1)}D^{(2)})_{oo}\right)=-(DD^{(2)}D^{(1)}D+DD^{(1)}D^{(2)}D)\\ D^{(3)}_{vv}&=+\left((D^{(2)}D^{(1)})_{vv}+(D^{(1)}D^{(2)})_{vv}\right)=\bar{D}(D^{(2)}D^{(1)})\bar{D}+\bar{D}(D^{(1)}D^{(2)})\bar{D}\end{aligned}$$

Applying the transformations (3.40) and (3.41), we obtain:

$$D^{(3)}_{vv}=\bar{D}D^{(3)}_{vv}\bar{D}=\bar{D}(D^{(2)}DD^{(1)})\bar{D}+\bar{D}(D^{(1)}DD^{(2)})\bar{D} \tag{3.45}$$

The 3rd order CG-CPSCF initial guess is therefore given by

$$\begin{aligned}D^{(3)}_{oo}&=-D(D^{(2)}D^{(1)}+D^{(1)}D^{(2)})D\\ D^{(3)}_{vv}&=D^{(2)}DD^{(1)}+D^{(1)}DD^{(2)}\end{aligned}$$

Finally, 3rd-order CG-CPSCF equation reads:

$$\left[F,\left[D,D^{(3)}\right]\right]+2DF^{(3)}D-\left\{D,F^{(3)}\right\}=\left[D,\left[D^{(1)},F^{(2)}\right]\right]+\left[D,\left[D^{(2)},F^{(1)}\right]\right] \tag{3.47}$$

Compared to AO-CPSCF equations (3.31), the solution of this equation should lead to $D^{(3)}_{ov}$, if we perform a spectral resolution of $D^{(3)}$.

3.3.3 *k*th-order response

Consequently, we can easily generalize the CG-CPSCF equations at any kth-order, the diagonal components $D^{(k)}_{oo}$ and $D^{(k)}_{vv}$ defining the initial guess, according to

$$D^{(k)}_{oo}=-D\left(\sum_{i=1}^{k-1}D^{(k-i)}D^{(i)}\right)D \tag{3.48}$$

$$D^{(k)}_{vv}=\sum_{i=1}^{k-1}D^{(k-i)}DD^{(i)} \tag{3.49}$$

[2] Obviously, the projection of Eq. (3.43) onto the subspace $\mathcal{H}_{\text{occ-virt}}$ or $\mathcal{H}_{\text{virt-occ}}$ leads to the second order AO-CPSCF equations (3.24).

with the kth-order transition matrices being the solutions of the following equation

$$\left[F,\left[D,D^{(k)}\right]\right]+2DF^{(k)}D-\left\{D,F^{(k)}\right\}=\left[D,\sum_{i=1}^{k}\left[D^{(k-i)},F^{(i)}\right]\right] \tag{3.50}$$

In Eq. (3.50), the right-hand side changes with respect to the density and the Fock matrices at lower orders, whereas the left-hand side (lhs) depends only on the kth order density matrix to be determined. As a result, solving the CG-CPSCF equation is analogous to solve a linear system of equations $AX = B$. The description of the CG-CPSCF equation resolution using a conjugate-gradient algorithm is given in Appendix D.

3.4 Perturbed projection by trace-correcting purification

Derived from the generalized equations for the SCF conditions (3.10) and for the idempotency constraints (3.9), the AO-CPSCF method uses the eigenstates of the unperturbed Fock matrix, while the CG-CPSCF method employs uniquely the density matrix. Both approaches require to proceed order by order. In other words, given a perturbed Fock matrix we are solving a linear problem at each order. On the other hand, it is possible to compute, the perturbed density matrix at a desired order without prior calculation of the lower terms, ie. all the perturbed density matrices of lower orders, all of them, being computed on-the-fly during the CPSCF processus. In other words, we are solving a non-linear set of equations.

This approach proposed by Weber and co-workers[38, 39], is based on inserting the perturbative expansions (3.7) obtained for the density and Fock matrices within the trace correcting purification (TCP) formalism.[110] By doing so, they show that the approach is capable to purify the kth and lower orders perturbed density matrices within the SCF loop. This perturbed purification method is called the perturbed projection by the trace-correcting purification (TC2-CPSCF). The unperturbed TCP aims to purify simultaneously each perturbed density matrix following the relation:

$$D_{\lambda,n+1}=\begin{cases} D_{\lambda,n}^2 & \text{if } \mathrm{Tr}\{D_{\lambda,n}\}\geq N \\ 2D_{\lambda,n}-D_{\lambda,n}^2 & \text{if } \mathrm{Tr}\{D_{\lambda,n}\}<N \end{cases} \tag{3.51}$$

using the following initial guess:

$$D_{\lambda,0}=(\epsilon_{\max}I-F_\lambda)/(\epsilon_{\max}-\epsilon_{\min}) \tag{3.52}$$

On introducing the perturbative expansion (3.7) in Eqs. (3.51) and (3.52), the unperturbed TCP equation transforms to

$$D_{n+1} = \begin{cases} (\sum_{k=0}^{3} D_n^{(k)})^2 & \text{if } \mathrm{Tr}\{\sum_{k=0}^{3} D_n^{(k)}\} \geq N \\ 2(\sum_{k=0}^{3} D_n^{(k)}) - (\sum_{k=0}^{3} D_n^{(k)})^2 & \text{if } \mathrm{Tr}\{\sum_{k=0}^{3} D_n^{(k)}\} < N \end{cases}$$

$$D_0 = \left(\epsilon_{\max} I - \sum_{k=0}^{3} F^{(k)}\right) / \left(\epsilon_{\max} - \epsilon_{\min}\right) \tag{3.54}$$

Considering the trace,

$$\begin{aligned} \mathrm{Tr}\{(D_n + D_n^{(1)} + D_n^{(2)} + D_n^{(3)})\} &= \mathrm{Tr}\{D\} + \mathrm{Tr}\{D_n^{(1)}\} + \mathrm{Tr}\{D_n^{(2)}\} + \mathrm{Tr}\{D_n^{(3)}\} \\ &= N + \delta N^{(1)} + \delta N^{(2)} + \delta N^{(3)} \end{aligned}$$

If we start from a well-conditioned initial guess, such as: $\delta N^{(k)} < \tau, \forall k \geq 1$, where τ is some threshold parameter (typically about 10^{-3}), we may expect the perturbed density matrix to preserve the trace, that is,

$$\mathrm{Tr}\{(D_n + D_n^{(1)} + D_n^{(2)} + D_n^{(3)})\} \simeq \mathrm{Tr}\{D_n\} \tag{3.56}$$

This constraint means that the perturbation does not create nor annihilate particles. By developing and assembling terms by perturbation order, Eq. (3.53) can be recast in the following form

$$\mathrm{Tr}\{D_n\} \geq N \begin{cases} D_{n+1} = D_n^2 \\ D_{n+1}^{(1)} = \left\{D_n, D_n^{(1)}\right\} \\ D_{n+1}^{(2)} = \left(D_n^{(1)}\right)^2 + \left\{D_n, D_n^{(2)}\right\} \\ D_{n+1}^{(3)} = \left\{D_n, D_n^{(3)}\right\} + \left\{D_n^{(1)}, D_n^{(2)}\right\} \end{cases} \tag{3.57}$$

$$\mathrm{Tr}\{D_n\} < N \begin{cases} D_{n+1} = 2D_n - D_n^2 \\ D_{n+1}^{(1)} = 2D_n^{(1)} - \left\{D_n, D_n^{(1)}\right\} \\ D_{n+1}^{(2)} = 2D_n^{(2)} - \left[\left(D_n^{(1)}\right)^2 + \left\{D_n, D_n^{(2)}\right\}\right] \\ D_{n+1}^{(3)} = 2D_n^{(3)} - \left[\left\{D_n, D_n^{(3)}\right\} + \left\{D_n^{(1)}, D_n^{(2)}\right\}\right] \end{cases} \tag{3.58}$$

The initial guesses are defined accordingly,

$$\begin{aligned} D &= (\epsilon_{\max} I - F)/(\epsilon_{\max} - \epsilon_{\min}) \\ D^{(1)} &= -F^{(1)}/(\epsilon_{\max} - \epsilon_{\min}) \\ D^{(2)} &= -F^{(2)}/(\epsilon_{\max} - \epsilon_{\min}) \\ D^{(3)} &= -F^{(3)}/(\epsilon_{\max} - \epsilon_{\min}) \end{aligned}$$

We can even generalize the perturbed TCP at any order k (≥ 0)

$$D_{n+1}^{(k)} = \begin{cases} \sum_{l=0}^{k} D_n^{(l)} D_n^{(k-l)} & \text{if } \ \mathrm{Tr}\{D_n\} \geq N \\ \\ 2D_n^{(k)} - \sum_{l=0}^{k} D_n^{(l)} D_n^{(k-l)} & \text{if } \ \mathrm{Tr}\{D_n\} < N \end{cases} \tag{3.60}$$

with the initial guess as

$$D^{(k)} = -F^{(k)}/(\epsilon_{\max} - \epsilon_{\min}) \tag{3.61}$$

In principle, the perturbed projection can be derived from any purification approach. In this work, we have combined this method with the hole-particle canonical purification (HPCP).

3.5 Perturbed projection by hole-particle canonical purification

Regarding its polynomials, the TCP is the simplest method, compared to other purifications. However, let us recall that, using the TCP the density matrix trace reaches the correct value only at convergence [cf. Section 2.2.4], whereas the HPCP maintains the N-representability conditions throughout the purification process. As a result, our objective is to compare the performances of the two polynomials in terms of efficiency and reliability. From Eqs. (2.29), (2.31) and (2.32)], the HPCP perturbed projection equations are given below up to the 3rd-order.

0th order

$$\begin{aligned} D_{n+1} &= (1 - 2c_n)D_n + 2(1 + c_n)D_n^2 - 2D_n^3 \\ D &= \alpha D_{\min} + (1 - \alpha)D_{\max} \\ D_{\min} &= \lambda_o(\mu I - F) + \theta I, \quad D_{\max} = \lambda_q(\mu I - F) + \theta I \end{aligned} \tag{3.62}$$

1st order

$$
\begin{aligned}
D_{n+1}^{(1)} &= (1-2c_n)D_n^{(1)} + 2(1+c_n)\left\{D_n, D_n^{(1)}\right\} - 2\left[D_n D_n^{(1)} D_n + \left\{D_n^2, D_n^{(1)}\right\}\right] \\
D^{(1)} &= \alpha D_{\min}^{(1)} + (1-\alpha) D_{\max}^{(1)} \\
D_{\min}^{(1)} &= -\lambda_o F^{(1)},\ D_{\max}^{(1)} = -\lambda_q F^{(1)}
\end{aligned}
\tag{3.63}
$$

2nd order

$$
\begin{aligned}
D_{n+1}^{(2)} &= (1-2c_n)D_n^{(2)} + 2(1+c_n)\left[\left(D_n^{(1)}\right)^2 + \left\{D_n, D_n^{(2)}\right\}\right] \\
&\quad - 2\left[D_n D_n^{(2)} D_n + D_n^{(1)} D_n D_n^{(1)} + \left\{D_n^2, D_n^{(2)}\right\} + \left\{D_n, \left(D_n^{(1)}\right)^2\right\}\right] \\
D^{(2)} &= \alpha D_{\min}^{(2)} + (1-\alpha) D_{\max}^{(2)} \\
D_{\min}^{(2)} &= -\lambda_o F^{(2)},\ D_{\max}^{(2)} = -\lambda_q F^{(2)}
\end{aligned}
\tag{3.64}
$$

3rd order

$$
\begin{aligned}
D_{n+1}^{(3)} &= (1-2c_n)D_n^{(3)} + 2(1+c_n)\left[\left\{D_n, D_n^{(3)}\right\} + \left\{D_n^{(1)}, D_n^{(2)}\right\}\right] \\
&\quad - 2\left[\left(D_n^{(1)}\right)^3 + \left\{D_n^2, D_n^{(3)}\right\} + \left\{D_n, \left\{D_n^{(1)}, D_n^{(2)}\right\}\right\}\right] \\
&\quad - 2\left[D_n D_n^{(3)} D_n + D_n^{(1)} D_n D_n^{(2)} + D_n^{(2)} D_n D_n^{(1)}\right] \\
D^{(3)} &= \alpha D_{\min}^{(3)} + (1-\alpha) D_{\max}^{(3)} \\
D_{\min}^{(3)} &= -\lambda_o F^{(3)},\ D_{\max}^{(3)} = -\lambda_q F^{(3)}
\end{aligned}
\tag{3.65}
$$

Finally, the generalization of the perturbed HPCP at any order k ($k \geq 1$) reads

$$
D_{n+1}^{(k)} = (1-2c_n)D_n^{(k)} + 2(1+c_n)\sum_{l=0}^{k} D_n^{(l)} D_n^{(k-l)} - 2\sum_{l,j=0}^{k} D_n^{(l)} D_n^{(j)} D_n^{(k-l-j)} \tag{3.66a}
$$

$$
D^{(k)} = \alpha D_{\min}^{(k)} + (1-\alpha) D_{\max}^{(k)} \tag{3.66b}
$$

$$
D_{\min}^{(k)} = -\lambda_o F^{(k)},\ D_{\max}^{(k)} = -\lambda_q F^{(k)} \tag{3.66c}
$$

We emphasize that, the constants c_n in Eq. (2.18) and μ in Eq. (2.32c) are not expected to change since the constraint is applied on the trace of the unperturbed density matrix. Compared to the CG-based minimization [cf. Section 3.3], the generalized idempotency constraints (3.9e) constitute the kernel of the polynomials for the perturbed projections. The SCF conditions of Eq. (3.10e) are used in this case to accelerate the perturbed projection.

3.6 Derivative of direct inversion of the iterative subspace

The simultaneous SCF computation for all the perturbed density matrices (from the 1st up to 3rd order) can be accelerated by means of the D-DIIS[154] (derivative direct inversion in the Iterative subspace), which is an extension of the DIIS method presented in Chapter 1 and used for the calculation of the non-perturbed density matrix [cf. Section 1.6.2]. For perturbation orders $k \geq 1$, we have implemented the D-DIIS inside the perturbed projection algorithms. The analogue of the DIIS update [cf. Eq. (1.96)] for the derivatives of the Fock matrix is given by

$$\tilde{F}_n^{(k)} := \sum_{i=n-m}^{n} c_i^{(k)} F_i^{(k)} \tag{3.67}$$

The coefficients $\{c_i^{(k)}\}$ of the linear combination are obtained by minimizing the norm of the error vectors, $\{e_i^{(k)}\}$, defined as the commutators between perturbed density matrices and the corresponding Fock matrices [cf. Section 1.6.2]. For the D-DIIS algorithm[154], these error vectors read

$$e_i^{(1)} := \left[F_i, D_i^{(1)}\right] + \left[F_i^{(1)}, D_i\right] \tag{3.68a}$$
$$e_i^{(2)} := \left[F_i, D_i^{(2)}\right] + \left[F_i^{(1)}, D_i^{(1)}\right] + \left[F_i^{(2)}, D_i\right] \tag{3.68b}$$

which can be recognized as the SCF conditions of Eq. (3.10a) and (3.10b), respectively. As a result, from Eq. (3.10e), the error vector can be generalized at any kth-order, following:

$$e_i^{(k)} := \sum_{l=0}^{k} \left[F_i^{(l)}, D_i^{(k-l)}\right] \tag{3.69}$$

The D-DIIS procedure is identical to the DIIS, the only difference being the definition of the error vectors. The minimization of the norm of the error is performed following the constraint of normalization of the coefficients [cf. Section 1.6.2, Eq. (1.98)], such as

$$\min \left\{ f^{\text{(D-)DIIS}}, \sum_{i=n-m}^{n} c_i^{(k)} = 1 \right\} \tag{3.70}$$

with

$$f^{\text{D-DIIS}}(c_{n-m}^{(k)}, ..., c_n^{(k)}) := \sum_{l,p=n-m}^{n} c_l^{(k)} c_p^{(k)} (e_l^{(k)} \cdot e_p^{(k)}) \tag{3.71}$$

The solution to the problem (3.70) is given by the Euler-Lagrange equation (1.103), which corresponds to a system of $(m+1)$ linear equations, corresponding to the m coefficients $\{c_i^{(k)}\}$ to be determined, including the Lagrange multiplier λ.

$$\begin{pmatrix} B^{(k)} & \mathbf{1}^t \\ \mathbf{1} & 0 \end{pmatrix} \begin{pmatrix} c^{(k)} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix} \tag{3.72}$$

$$\text{with:} \quad \mathbf{0} = (0, ..., 0) \quad \text{and} \quad \mathbf{1} = (1, ..., 1) \tag{3.73}$$

$$c^{(k)} = (c_{n-m}^{(k)}, ..., c_n^{(k)}) \tag{3.74}$$

$$B_{lp}^{(k)} = (e_l^{(k)} \cdot e_p^{(k)}) \tag{3.75}$$

The pseudo-algorithm using the DIIS/D-DIIS is presented in the Appendix A.

3.7 Discussions

In this Chapter we have derived three density matrix-based perturbation methods: (i) the standard AO-CPSCF based on the diagonalization, (ii) the conjugate-gradient based method CG-CPSCF, and (iii) the perturbed projections TC2-CPSCF and HPCP-CPSCF. For all the methods, the density matrix sparsity can be controlled by means of sparse matrix algebra [Section 2.3.2] and a truncation scheme [Section 2.3.1]. This allows, in principle, to perform linear scaling CPSCF calculation. However, we noted that the linear scaling regime already achieved was based only on the numerical truncation.[39, 38, 43, 44]. In Table {3.1} is reported the number of matrix multiplication (MM) involved at each CG-

Method	1st order	2nd order	3rd order	kth-order
AO-CPSCF	5	8	14	$5 + 1.5 \times k!$, $k \geq 2$
CG-CPSCF	4	7	11	$5 + k! + (6)$, $k \geq 2$
TC2-CPSCF	1	2	4	2^{k-1}, $k \geq 1$
HPCP-CPSCF	4	8	16	2^{k+1}, $k \geq 1$

Table 3.1 Number of matrix multiplication for the density matrix methods at different perturbation orders.

step or perturbed projection for all the methods, at any kth-order. For the AO-CPSCF and CG-CPSCF, this number includes the iterative and non-iterative parts. The number given in parenthesis for the CG-CPSCF at the kth-order correponds to the number of MM during the line search of the conjugate gradient. Using these methods, the k order perturbed density matrix is computed and will be used for the evaluation of molecular properties, such as the static non-linear optical properties.

Chapter 4

Applications to non-linear optical properties of π-conjugated systems

4.1 Non linear optical properties

The Chapter 3 has discussed the density matrix perturbation methods. However, this discussion was rather general in regards of perturbation to be considered. The molecular properties, as measured by spectroscopy, are the observable responses of the molecular system to an external perturbation. In other terms, the considered perturbartion allows to probe specific properties of the molecular system. We focus here on the response(s) induced by a pertubative external static electric field(s), which define the static optical properties of the molecular system.

4.1.1 Perturbed energy expression for the PPP model

On considering the molecular system in the static electric field $\vec{\mathscr{E}}$, the interaction between the system and the field is pictured by the electric molecular dipole.[155, 1] The classical molecular dipole is defined from the sum of the nuclear contributions, added to the sum over the ponctual electronic charge times the electron position operator. Since in the present formalism nuclei position are not quantized, and we are dealing with a minimal HF-PPP approach [cf. Section 1.4], only the contributions from the π-electrons are relevant. The perturbation operator describing the coupling of the electric field ($\lambda := \vec{\mathscr{E}}$) with the electronic dipole $\vec{p}$, is defined by

$$\hat{\Delta}_\lambda := -\vec{p} \cdot \vec{\mathscr{E}} \tag{4.1}$$

with

$$\vec{p} := -\sum_i^{N_e} \mathbf{r}_i \tag{4.2}$$

Since $\hat{\Delta}_\lambda$ is a one-electron operator, when added to the electronic Hamiltonian [cf. Eq. (1.34)], its contributions appear only within the one-electron contribution of Eq. (1.60b), according to

$$h^\lambda_{\mu\nu} := \int d\mathbf{r}_1 \phi^*_\mu(\mathbf{r}_1) \left[-\frac{1}{2}\nabla^2_1 - \mathbf{r}_1 \cdot \vec{\mathscr{E}} - \sum_{A=1}^{M} \frac{Z_A}{r_{1A}} \right] \phi_\nu(\mathbf{r}_1) \tag{4.3}$$

The latter can be decomposed into two terms:

$$h^\lambda_{\mu\nu} = h_{\mu\nu} + \Delta^\lambda_{\mu\nu} \tag{4.4}$$

where h is the original non-perturbed one-electron matrix and Δ_λ is the dipole-electric field coupling matrix, whose elements are given by:

$$\Delta^{\lambda}_{\mu\nu} = \langle\mu|\mathbf{r}\cdot\vec{\mathscr{E}}|\nu\rangle = \sum_{a\in\{x,y,z\}} \mathscr{E}_a\langle\mu|a|\nu\rangle \tag{4.5}$$

In the equation above, $\mathscr{E}_a$ is the component of the electric field along the cartesian direction a. The additional term in the one-electron core hamiltonian due to the presence of the electric field is incorporated during the whole SCF procedure. Consequently, the quantities such as the density and the Fock matrices along with the energies are necessarily modified. As a result, we may express the perturbed Fock matrix and electronic energy by

$$F_\lambda = h_\lambda + G\left(D_\lambda\right) \tag{4.6a}$$

$$\mathcal{E}_\lambda = \mathrm{Tr}\{D_\lambda\left(h_\lambda + F\left(D_\lambda\right)\right)\} \tag{4.6b}$$

where we made explicit the dependence of the two-electron contribution to the perturbed density matrix D_λ. Referring to Chapter 1, the expectation value of an operator is the trace of the product between this operator and the density matrix. As a result, for each cartesian component $a \in \{x, y, z\}$, we have:

$$\langle p^{(a)}\rangle := -\mathrm{Tr}\{Dh^{(a)}\} \tag{4.7}$$

where $h^{(a)}$ is the dipole moment matrix of elements $\langle\mu|a|\nu\rangle$. Within the HF-PPP model, it nicely simplifies to

$$h^{(a)} = \langle a\rangle\delta_{\mu\nu} \tag{4.8}$$

This corresponds to a diagonal matrix, with for elements, the position vector component along the direction a.

4.1.2 Energy and response expansions

The conventional expansion[156] of the energy for a system perturbed by an external electric field $\vec{\mathscr{E}}$ is given by

$$\begin{aligned}\mathcal{E}(\vec{\mathscr{E}}) = \mathcal{E} - \sum_a \mu_a\mathscr{E}_a - \frac{1}{2!}\sum_{a,b}\alpha_{ab}\mathscr{E}_a\mathscr{E}_b + \\ - \frac{1}{3!}\sum_{a,b,c}\beta_{abc}\mathscr{E}_a\mathscr{E}_b\mathscr{E}_c - \frac{1}{4!}\sum_{a,b,c,d}\gamma_{abcd}\mathscr{E}_a\mathscr{E}_b\mathscr{E}_c\mathscr{E}_d - ...\end{aligned} \tag{4.9}$$

where $(a, b, c, d) \in \{x, y, z\}$, and $\mathcal{E}$ is the non perturbed total electronic energy. The response tensors μ (1st rank $\equiv$ vector), α (2nd rank), β (3rd rank) and γ (4th rank) refer to the dipole moment, the polarizability, the first hyperpolarizability and the second hyperpolarizability, respectively. The density matrix and the Fock matrix are also expanded in terms of the electric field,[156] according to

$$D(\vec{\mathscr{E}}) = D + \sum_a D^{(a)}\mathscr{E}_a + \frac{1}{2!}\sum_{a,b} D^{(ab)}\mathscr{E}_a\mathscr{E}_b + \\ + \frac{1}{3!}\sum_{a,b,c} D^{(abc)}\mathscr{E}_a\mathscr{E}_b\mathscr{E}_c + \frac{1}{4!}\sum_{a,b,c,d} D^{(abcd)}\mathscr{E}_a\mathscr{E}_b\mathscr{E}_c\mathscr{E}_d + ... \tag{4.10}$$

$$F(\vec{\mathscr{E}}) = F + \sum_a F^{(a)}\mathscr{E}_a + \frac{1}{2!}\sum_{a,b} F^{(ab)}\mathscr{E}_a\mathscr{E}_b + \\ + \frac{1}{3!}\sum_{a,b,c} F^{(abc)}\mathscr{E}_a\mathscr{E}_b\mathscr{E}_c + \frac{1}{4!}\sum_{a,b,c,d} F^{(abcd)}\mathscr{E}_a\mathscr{E}_b\mathscr{E}_c\mathscr{E}_d + ... \tag{4.11}$$

where $(F^{(a)}, F^{(ab)}, ..., F^{(k)})$ are the perturbed Fock matrices and correspond to kth derivative of Eq. (4.6a), which are defined, at any order, by

$$F^{(k)} = \begin{cases} h^{(k)} + 2J[D^{(k)}] - K(D^{(k)}), & k = 1\ (= a) \\ 2J[D^{(k)}] - K(D^{(k)}) & k > 1\ (= ab, abc, abcd, ...) \end{cases} \tag{4.12}$$

where $h^{(1)}$ is the dipole moment matrix as defined above. J and K are respectively the Coulomb matrix and the exchange matrix, defining the bi-electronic term G and depending on the perturbed density matrix $D^{(k)}$. On taking the first derivative of Eq. (4.9) with respect to the electric field components, leads to the dipole moment.

$$p^{(a)}(\vec{\mathscr{E}}) = -\mu_a - \frac{1}{2!}\sum_b \alpha_{ab}\mathscr{E}_b + \\ - \frac{1}{3!}\sum_{b,c} \beta_{abc}\mathscr{E}_b\mathscr{E}_c - \frac{1}{4!}\sum_{b,c,d} \gamma_{abcd}\mathscr{E}_b\mathscr{E}_c\mathscr{E}_d - ... \tag{4.13}$$

Inserting firstly Eq. (4.10) into (4.7), then comparing the resulting equations with Eq. (4.13), leads to the following definitions for the response tensors,

$$\mu_a = \mathrm{Tr}\{h^{(a)}D\} \tag{4.14a}$$
$$\alpha_{ab} = \mathrm{Tr}\{h^{(a)}D^{(b)}\} \tag{4.14b}$$
$$\beta_{abc} = \mathrm{Tr}\{h^{(a)}D^{(bc)}\} \tag{4.14c}$$
$$\gamma_{abcd} = \mathrm{Tr}\{h^{(a)}D^{(bcd)}\} \tag{4.14d}$$

That system gives the basic definition of the response tensors (up to 4th order) for a molecular system within a static electric field. Each tensor is defined by a perturbed density matrix, as already outlined for the density matrix perturbation methods discussed in Chapter 3. The alternative to compute these quantities is based on differentiating the electronic energy $\mathcal{E}$ with respect to the applied field components at the zero field limit. Formally this writes:

$$\mu_a = -\frac{\partial \mathcal{E}(\vec{\mathscr{E}})}{\partial \mathscr{E}_a}\bigg|_{\vec{\mathscr{E}}=0}, \quad \alpha_{ab} = -\frac{\partial^2 \mathcal{E}(\vec{\mathscr{E}})}{\partial \mathscr{E}_a \partial \mathscr{E}_b}\bigg|_{\vec{\mathscr{E}}=0},$$

$$\beta_{abc} = -\frac{\partial^3 \mathcal{E}(\vec{\mathscr{E}})}{\partial \mathscr{E}_a \partial \mathscr{E}_b \partial \mathscr{E}_c}\bigg|_{\vec{\mathscr{E}}=0}, \quad \gamma_{abcd} = -\frac{\partial^4 \mathcal{E}(\vec{\mathscr{E}})}{\partial \mathscr{E}_a \partial \mathscr{E}_b \partial \mathscr{E}_c \partial \mathscr{E}_d}\bigg|_{\vec{\mathscr{E}}=0} \tag{4.15}$$

The Eq. (4.1.2) has actually presented two definitions of the response tensors for a molecular system under a static electric field: (i) the analytic definition of Eq. (4.14) using the density matrix perturbation methods and, (ii) the numerical definition of Eq. (4.15) where the response tensors are evaluated by a numerical differentation using the finite field difference method (FFD). Detailed description of the FFD applied to the calculation of static NLO properties is presented in Appendix E. It is worth to mention that the idempotency relation for the density matrix in a FFD is not verified. As a result, the perturbed electronic energy can not be calculated using a purification or a minimization method. In the algorithm that we have implemented for the finite difference method [Appendix E], the perturbed density matrix is calculated by means of the diagonalization.

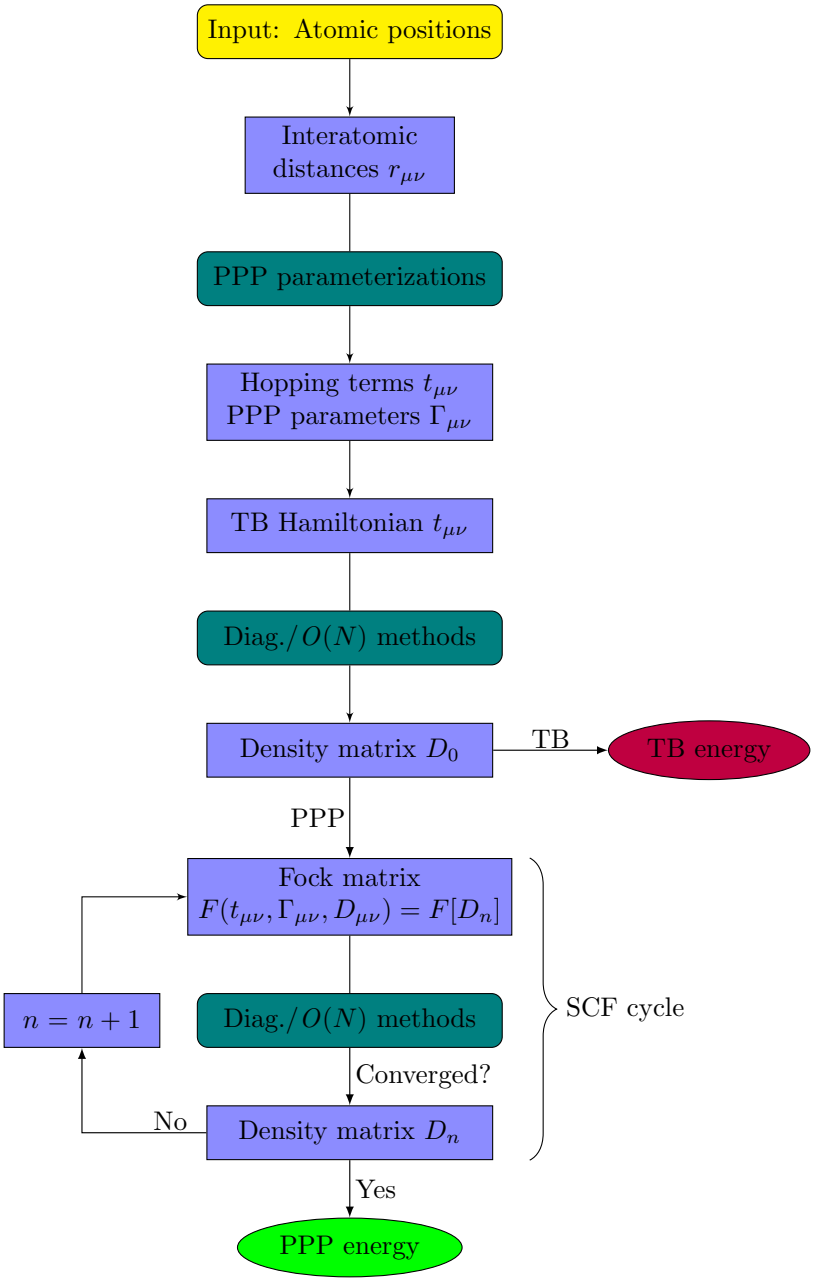


Fig. 4.1 Outline of the implementation giving the steps of the unperturbed SCF PPP calculation.

4.2 Outline of the implementation

Before presenting the results of our calculations of optical properties, this section oultines our implementation. We have modified a pre-existing code,[73] written in Fortran90. The Intel Math Kernel library[157] (MKL) routines were used for the diagonalization and CSC algebra. We note that any other package/routine can be easily adapted to the code.

The various steps for the computation of the unperturbed density matrix is outlined in Figure {4.1}. Given the atomic coordinates of the system, the distance matrix is calculated up to a predefined radial cutoff. From the chosen PPP parameterization, the hopping terms $t_{\mu\nu}$ and the PPP parameters $\Gamma_{\mu\nu}$ are evaluated accordingly. In Figure {4.1}, D_0 is the converged density matrix obtained from a TB calculation, which is used as the starting guess for the SCF calculation. In TB or PPP calculation, the density matrix is calculated using the diagonalization (Diag) or the $O(M)$ methods (Min). The $O(M)$ methods include the density matrix purifications and minimizations, combined with the numerical or radial truncations.

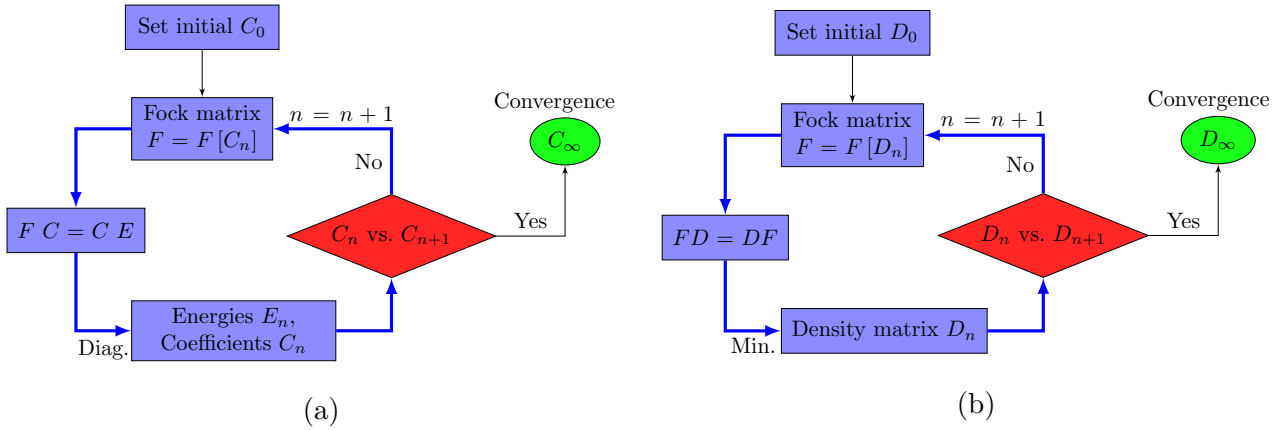
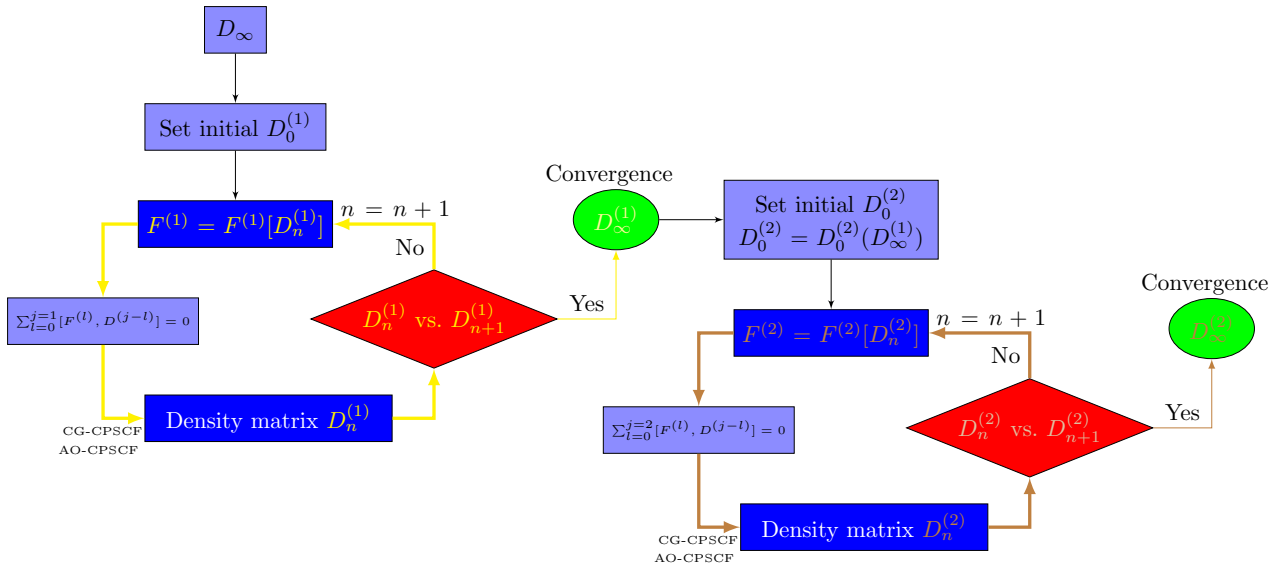


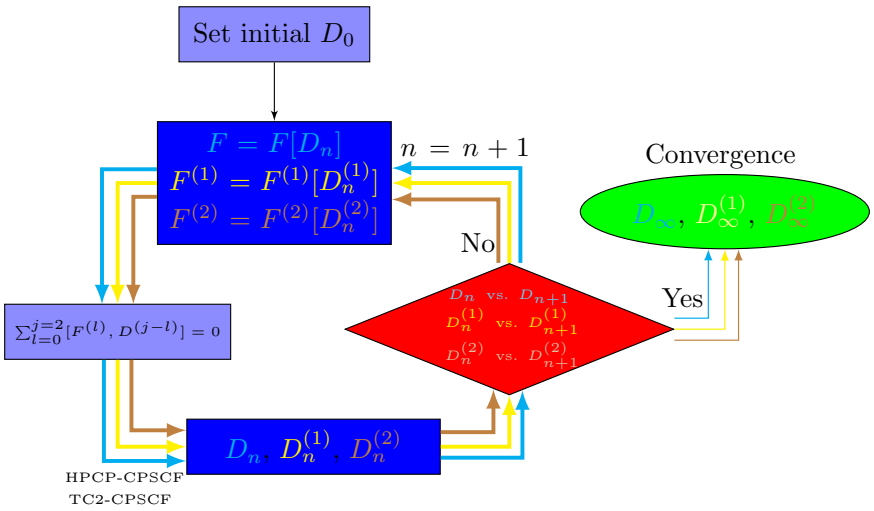
Fig. 4.2 SCF procedure using: (a) the diagonalization [Diag], and (b) the density matrix energy minimizations [Min].

4.3 Perturbed dense matrix calculation

In this Section we aim to determine the perturbed density matrices $D^{(1)}$, $D^{(2)}$ and $D^{(3)}$ for the evaluation of the static optical properties. For that purpose, we have considered a benchmark of 2D π-conjugated systems. We will first compare the reliability of the various density matrix pertubation theory (DMPT) methods, along with the FFD



(a) Procedure for AO-CPSCF and CG-CPSCF



(b) Procedure for TC2-CPSCF and HPCP-CPSCF

Fig. 4.3 CPSCF density matrix perturbation methods. (a) AO-CPSCF and CG-CPSCF, (b) TC2-CPSCF and HPCP-CPSCF.

approach.[158] In a second step, we will increase the size of the molecules and investigate the convergences.

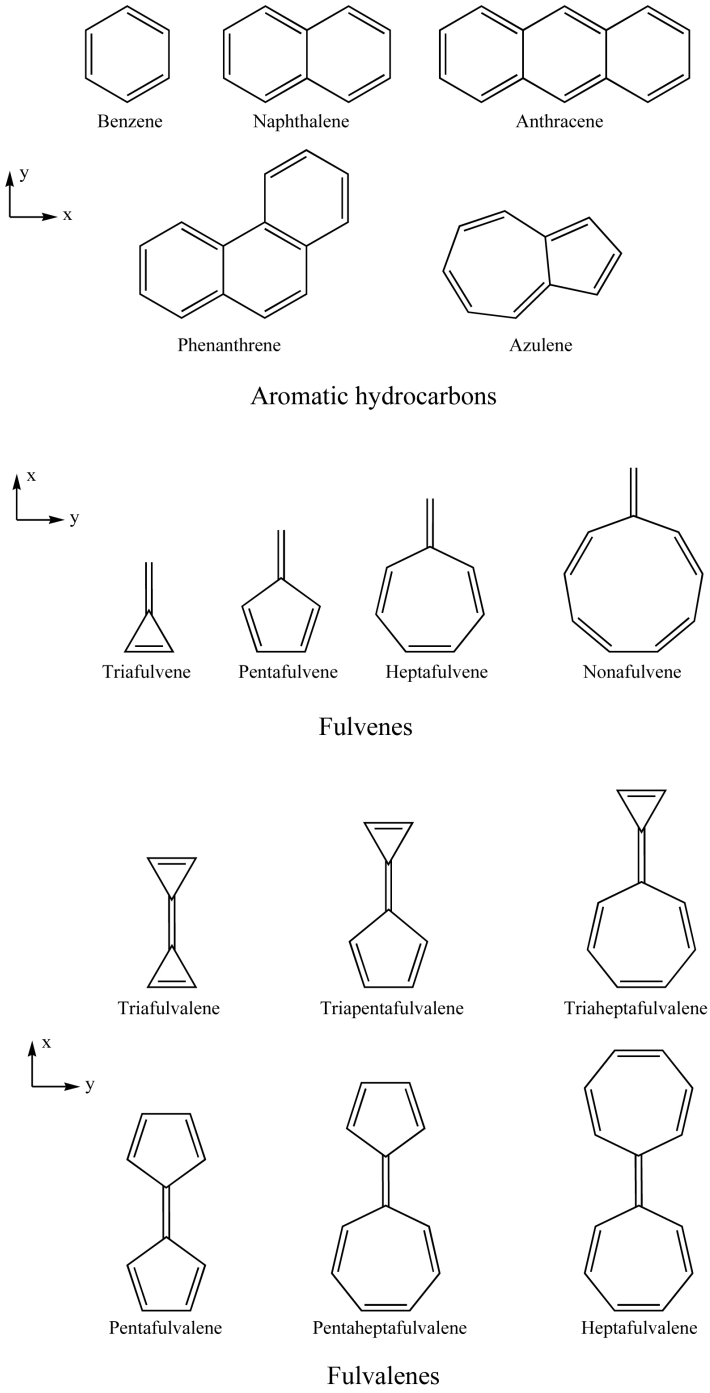


Fig. 4.4 Benchmark of molecules.

	Polarizability			1st hyperpolarizability			2nd hyperpolarizability ($\times 10^4$)			
	α_{xx}	α_{yy}	$\langle\alpha\rangle$	β_{xxx}	β_{xyy}	$\langle\beta\rangle$	γ_{xxxx}	γ_{yyyy}	γ_{xxyy}	$\langle\gamma\rangle$
					Benzene					
Ref. [158]	38.91		25.94				0.158			0.084
FFD*	38.9173	38.9173	25.9449	0	0	0	0.1583	0.1583	0.0528	0.0844
FFD	38.8585	38.8585	25.9057	0	0	0	0.1576	0.1576	0.0525	0.0841
AO-CPSCF	38.8585	38.8585	25.9057	0	0	0	0.1577	0.1576	0.0524	0.0842
CG-CPSCF	38.8585	38.8585	25.9057	0	0	0	0.1577	0.1576	0.0525	0.0842
TC2-CPSCF	38.8616	38.8616	25.9077	0	0	0	0.1576	0.1576	0.0525	0.0841
HPCP-CPSCF	38.8585	38.8585	25.9057	0	0	0	0.1576	0.1576	0.0525	0.0841
					Naphthalene					
Ref. [158]	98.57		54.59				2.73			0.643
FFD*	98.5811	65.1941	54.5917	0	0	0	2.7249	1.6631	−0.5855	0.6434
FFD	98.5899	65.1123	54.5674	0	0	0	2.7468	1.6600	−0.5767	0.6507
AO-CPSCF	98.5899	65.1123	54.5674	0	0	0	2.7468	1.6600	−0.5766	0.6507
CG-CPSCF	98.5899	65.1123	54.5674	0	0	0	2.7469	1.6601	−0.5765	0.6508
TC2-CPSCF	98.5899	65.1123	54.5674	0	0	0	2.7468	1.6600	−0.5766	0.6508
HPCP-CPSCF	98.5899	65.1123	54.5674	0	0	0	2.7468	1.6600	−0.5767	0.6507
					Anthracene					
Ref. [158]	178.37		93.62				11.95			2.180
FFD*	178.3843	102.4862	93.6235	0	0	0	11.9506	2.5780	−1.8144	2.1799
FFD	178.5480	102.3152	93.6211	0	0	0	12.0748	2.6028	−1.7922	2.2186
AO-CPSCF	178.5480	102.3152	93.6211	0	0	0	12.0748	2.6027	−1.7922	2.2185
CG-CPSCF	178.5480	102.3152	93.6211	0	0	0	12.0747	2.6027	−1.7922	2.2186
TC2-CPSCF	178.5480	102.3152	93.6211	0	0	0	12.0748	2.6028	−1.7922	2.2186
HPCP-CPSCF	178.5480	102.3152	93.6211	0	0	0	12.0750	2.6028	−1.7922	2.2186
					Phenanthrene					
Ref. [158]	159.19		83.83				6.09			2.35
FFD*	159.2031	92.2974	83.8335	0	0	0	6.0937	1.5063	2.0834	2.3533
FFD	159.2354	92.3038	83.8464	0	0	0	6.2105	1.5112	2.0920	2.3811
AO-CPSCF	159.2354	92.3038	83.8464	0	0	0	6.2106	1.5112	2.0921	2.3811
CG-CPSCF	159.2354	92.3038	83.8464	0	0	0	6.2106	1.5112	2.0621	2.3812
TC2-CPSCF	159.2354	92.3038	83.8464	0	0	0	6.2105	1.5112	2.0920	2.3812
HPCP-CPSCF	159.2354	92.3038	83.8464	0	0	0	6.2105	1.5112	2.0920	2.3811
					Azulene					
FFD*	129.2114	70.1698	66.4604	700.9358	−4.9089	417.6161	−0.5129	0.0419	1.3646	0.4516
FFD	129.3624	70.1411	66.5012	703.2813	−6.7156	417.9394	−0.5148	0.0365	1.3498	0.4443
AO-CPSCF	129.3624	70.1411	66.5012	703.2816	−6.7156	417.9396	−0.5147	0.0365	1.3498	0.4443
CG-CPSCF	129.3625	70.1420	66.5015	703.2816	−6.7156	417.9396	−0.5168	0.0367	1.3498	0.4238
TC2-CPSCF	129.3624	70.1411	66.5012	703.2817	−6.7157	417.9396	−0.5148	0.0365	1.3498	0.4443
HPCP-CPSCF	129.3625	70.1411	66.5012	703.2816	−6.7151	417.9399	−0.5148	0.0365	1.3498	0.4443

Table 4.1 Calculated π-polarizabilities, first and second π-hyperpolarizabilities of aromatic hydrocarbons in au.

	Polarizability			1st hyperpolarizability			2nd hyperpolarizability ($\times 10^4$)			
	α_{xx}	α_{yy}	$\langle\alpha\rangle$	β_{xxx}	β_{xyy}	$\langle\beta\rangle$	γ_{xxxx}	γ_{yyyy}	γ_{xxyy}	$\langle\gamma\rangle$
					Triafulvene					
Ref. [158]			20.62	−153						−0.108
FFD*	39.0978	22.7843	20.6274	−152.5390	141.4826	−6.6338	−0.4775	−0.4248	0.1805	−0.1082
FFD	38.9546	22.6620	20.5389	−154.5884	139.6921	−8.9378	−0.4656	−0.4160	0.1777	−0.1052
AO-CPSCF	38.9546	22.6620	20.5389	−154.5884	139.6721	−8.9378	−0.4656	−0.4160	0.1778	−0.1052
CG-CPSCF	38.9546	22.6620	20.5389	−154.5884	139.6721	−8.9378	−0.4657	−0.4160	0.1778	−0.1051
TC2-CPSCF	38.9546	22.6620	20.5389	−154.5880	139.6920	−8.9376	−0.4656	−0.4160	0.1777	−0.1052
HPCP-CPSCF	38.9548	22.6619	20.5389	−154.5625	139.6910	−8.9229	−0.4652	−0.4160	0.1777	−0.1052
					Pentafulvene					
Ref. [158]			36.81	−343						−0.119
FFD*	86.5924	23.8389	36.8104	−342.8855	109.2706	−140.1689	−1.2895	1.8434	−0.5733	−0.1185
FFD	86.3699	23.7416	36.7038	−337.6135	109.6390	−136.7847	−1.2628	1.8349	−0.5657	−0.1119
AO-CPSCF	86.3699	23.7416	36.7038	−337.6135	109.6390	−136.7847	−1.2628	1.8349	−0.5566	−0.1119
CG-CPSCF	86.3699	23.7416	36.7038	−337.6134	109.6390	−136.4698	−1.2627	1.8348	−0.5567	−0.1119
TC2-CPSCF	86.3699	23.7416	36.7038	−337.6134	109.6390	−136.7847	−1.2628	1.8349	−0.5657	−0.1119
HPCP-CPSCF	86.3699	23.7416	36.7038	−337.6135	109.6390	−136.7847	−1.2628	1.8349	−0.5657	−0.1119
					Heptafulvene					
Ref. [158]			57.63	78						0.015
FFD*	124.0441	48.8482	57.6308	78.2542	−71.4721	4.0692	−3.2889	0.7894	1.2868	0.0148
FFD	123.8373	48.7077	57.5150	72.2241	−72.9804	−0.4538	−3.2299	0.8156	1.2914	0.0337
AO-CPSCF	123.8473	48.7077	57.5150	72.2238	−72.9804	−0.4539	−3.2298	0.8156	1.2914	0.0337
CG-CPSCF	123.8474	48.7077	57.5150	72.2238	−72.9805	−0.4540	−3.2298	0.8156	1.2915	0.0338
TC2-CPSCF	123.8373	48.7077	57.5150	72.2238	−72.9804	−0.4539	−3.2297	0.8157	1.2914	0.0338
HPCP-CPSCF	123.8373	48.7077	57.5150	72.2238	−72.9804	−0.4539	−3.2298	0.8156	1.2914	0.0337
					Nonafulvene					
Ref. [158]			84.51	82						0.940
FFD*	171.2162	82.3483	84.5215	81.4279	117.6384	119.4398	−1.7328	4.4402	0.9958	0.9398
FFD	171.0467	82.1730	84.4066	90.3619	120.7017	126.6382	−1.6062	4.5153	1.0251	0.9919
AO-CPSCF	171.0467	82.1730	84.4066	90.3623	120.7018	126.6385	−1.6060	4.5152	1.0251	0.9919
CG-CPSCF	171.0468	82.1730	84.4066	90.3624	120.7017	126.6383	−1.6061	4.5151	1.0251	0.9918
TC2-CPSCF	171.0467	82.1730	84.4066	90.3626	120.7018	126.6386	−1.6062	4.5153	1.0251	0.9919
HPCP-CPSCF	171.0467	82.1730	84.4066	90.3524	120.7017	126.6385	−1.6060	4.5153	1.0251	0.9919

Table 4.2 Calculated π-polarizabilities, first and second π-hyperpolarizabilities of fulvenes in au.

	Polarizability			1st hyperpolarizability			2nd hyperpolarizability ($\times 10^4$)			
	α_{xx}	α_{yy}	$\langle\alpha\rangle$	β_{xxx}	β_{xyy}	$\langle\beta\rangle$	γ_{xxxx}	γ_{yyyy}	γ_{xxyy}	$\langle\gamma\rangle$
					Triafulvalene					
FFD*	56.8454	48.5862	35.1439	0	0	0	−0.5645	−1.3163	1.6566	0.2865
FFD	56.6513	48.3306	34.9940	0	0	0	−0.5437	−1.2906	1.6471	0.2920
AO-CPSCF	56.6513	48.3306	34.9940	0	0	0	−0.5436	−1.2906	1.6472	0.2920
CG-CPSCF	56.6513	48.3306	34.9940	0	0	0	−0.5437	−1.2907	1.6470	0.2919
TC2-CPSCF	56.6513	48.3306	34.9940	0	0	0	−0.5437	−1.2906	1.6471	0.2920
HPCP-CPSCF	56.6513	48.3306	34.9940	0	0	0	−0.5437	−1.2906	1.6471	0.2920
					Triapentafulvalene					
FFD*	119.4234	43.8181	54.4138	362.1458	93.8790	273.6149	−4.7450	0.8016	−0.2202	−0.8768
FFD	119.4403	43.6747	54.3717	349.9607	91.5937	264.9327	−4.7839	0.8021	−0.2220	−0.8852
AO-CPSCF	119.4403	43.6747	54.3717	349.9595	91.5935	264.9318	−4.7738	0.8021	−0.2219	−0.8851
CG-CPSCF	119.4404	43.6748	54.3717	349.9596	91.5937	264.9319	−4.7639	0.8022	−0.2220	−0.8851
TC2-CPSCF	119.4403	43.6747	54.3717	349.9594	91.5937	264.9318	−4.7839	0.8021	−0.2220	−0.8851
HPCP-CPSCF	119.4403	43.6747	54.3717	349.9594	91.5937	264.9318	−4.7839	0.8021	−0.2219	−0.8852
					Triaheptafulvalene					
FFD*	166.9462	67.7813	78.2425	−478.4360	300.1587	−106.9664	−5.8718	0.2168	2.8864	0.0235
FFD	166.8266	67.5549	78.1271	−485.2264	298.7197	−111.9040	−5.7734	0.2473	2.8941	0.0524
AO-CPSCF	166.8266	67.5549	78.1271	−485.2270	298.7197	−111.9044	−5.7734	0.2472	2.8940	0.0524
CG-CPSCF	166.8266	67.5549	78.1271	−485.2270	298.7197	−111.9044	−5.7732	0.2473	2.8940	0.0525
TC2-CPSCF	166.8266	67.5549	78.7271	−485.2269	298.7196	−111.9044	−5.7734	0.2473	2.8941	0.0524
HPCP-CPSCF	166.8266	67.5549	78.1271	−485.2270	298.7197	−111.9044	−5.7735	0.2473	2.8941	0.0524
					Pentafulvalene					
FFD*	206.8605	41.6074	82.8226	0	0	0	−5.1002	3.0939	−1.4609	−0.9856
FFD	206.6937	41.4710	82.7216	0	0	0	−4.9177	3.0797	−1.4406	−0.9439
AO-CPSCF	206.6937	41.4710	82.7216	0	0	0	−4.9176	3.0797	−1.4406	−0.9438
CG-CPSCF	206.6937	41.4710	82.7216	0	0	0	−4.9177	3.0795	−1.4404	−0.9437
TC2-CPSCF	206.6937	41.7410	82.7216	0	0	0	−4.9177	3.0797	−1.4406	−0.9438
HPCP-CPSCF	206.6937	41.4710	82.7216	0	0	0	−4.9177	3.0797	−1.4406	−0.9438
					Pentaheptafulvalene					
FFD*	267.7461	70.6512	112.7991	−2227.6425	332.3833	−1137.1555	−24.8398	1.4563	−0.0315	−4.6893
FFD	267.9667	70.5457	112.8375	−2198.9172	336.6577	−1117.3557	−25.0370	1.4722	−0.0306	−4.7252
AO-CPSCF	267.9667	70.5457	112.8375	−2198.8878	336.6577	−1117.3380	−25.0370	1.4721	−0.0306	−4.7252
CG-CPSCF	267.9667	70.5457	112.8375	−2198.8879	336.6577	−1117.3381	−25.0370	1.4720	−0.0307	−4.7521
TC2-CPSCF	267.9667	70.5457	112.8375	−2198.8878	336.6577	−1117.3381	−25.0369	1.4722	−0.0306	−4.7252
HPCP-CPSCF	267.9667	70.5457	112.8375	−2198.8878	336.6577	−1117.3381	−25.0369	1.4722	−0.0306	−4.7252
					Heptafulvalene					
FFD*	360.4462	91.5110	150.6524	0	0	0	−34.0243	0.5761	4.7407	−4.7934
FFD	360.5862	91.3361	150.6407	0	0	0	−33.5601	0.6145	4.7902	−4.6730
AO-CPSCF	360.5862	91.3361	150.6407	0	0	0	−33.5601	0.6145	4.7900	−4.6730
CG-CPSCF	360.5862	91.3361	150.6407	0	0	0	−33.5600	0.6144	4.7900	−4.6729
TC2-CPSCF	360.5862	91.3361	150.6407	0	0	0	−33.5598	0.6145	4.7902	−4.6730
HPCP-CPSCF	360.5862	91.3361	150.6407	0	0	0	−33.5598	0.6145	4.7902	−4.6730

Table 4.3 Calculated π-polarizabilities, first and second π-hyperpolarizabilities of fulvalenes in au.

4.3.1 Applications and comparison for small systems

The benchmark set of molecules is given in Figure {4.4}. All the carbon-carbon distances have been fixed to 1.4 $\AA$. All the results presented in the following are in au.[159] A very tight convergence parameter of 10^{-14} has been used for the all SCF calculations. The mean values of the (hyper)polarizability tensors,[160] have been calculated according to the standard definitions:

$$\langle\alpha\rangle = \frac{1}{3}\left(\alpha_{xx} + \alpha_{yy}\right)$$
$$\langle\beta\rangle = \frac{3}{5}\left(\beta_{xxx} + \beta_{xyy}\right) \quad (x = \text{major symmetry axis})$$
$$\langle\gamma\rangle = \frac{1}{5}\left(\gamma_{xxxx} + \gamma_{yyyy} + 2\gamma_{xxyy}\right)$$

We have performed the FFD calculation using the same parameterization as in Ref. [158], which is referred as FFD* in Table {4.1} to {4.3}. It is rather clear that our FFD* and results from Ref. [158] are in good agreement. We have also performed a FFD calculation using the conventional Ohno's parameterization [cf. Table {1.1}]. On comparing the values obtained from the FFD and DMPT approaches using the same parameterization, we found that the methods are in perfect agreement demonstrating the reliability of our implementation. Note that the 1st hyperpolarizability β is expected to be zero for the centro-symmetric structures such as Phenanthrene compared to molecules without inversion symmetry such as Triaheptafulvalene, which is related to the rank of the tensor. It is worth to note that in our implementation, no symmetry constraints were applied to the Fock matrix.

4.3.2 Methods efficiency for larger systems

In order to expand the comparison of the methods, we have investigated the convergence of (hyper)polarizabilities with respect to the system size. For this purpose, we have used dense density matrices, ie. without employing a truncation scheme. The models used in this section are the polymers presented in Figure {4.5}. Polymer A and B are the transpolyacetylene (TPA) and the polyphenylene vinylene (PPV), respectively. Polymer C (TPA+) and polymer D (PPV+), are structurally derived from polymer A and polymer B. TPA+ and PPV+ were added to the benchmark in order to bypass inversion symmetry, and obtain non zero first hyperpolarizabilities. We also specify that two sets of calculation were performed: (i) for polymers with carbon-carbon distances varying between 1.35 and 1.45 $\AA$ (Set 1), (ii) for polymers with fixed carbon-carbon

distance of 1.4 $\AA$ (Set 2). All the polymers were replicated in the longitudinal direction x. Tables {4.4} and {4.5} present the results obtained for α_{xx}, β_{xxx} and γ_{xxxx} using

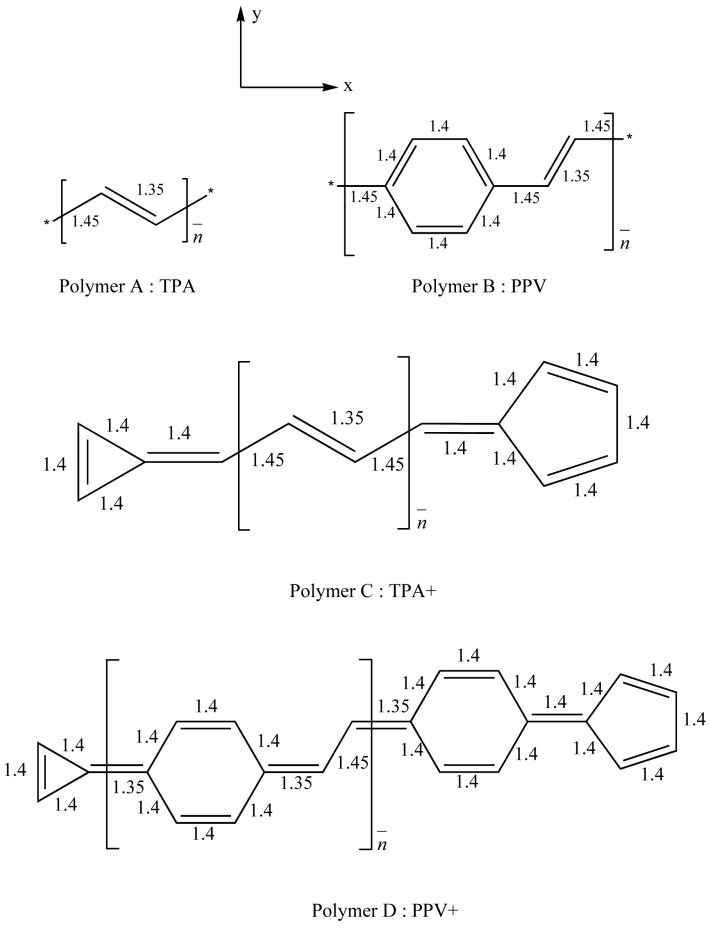


Fig. 4.5 Benchmark of polymers.

the 4 different methods. The responses using the exact method (AO-CPSCF) and the absolute errors for the same responses with respect to the AO-CPSCF (Δ(CG-CPSCF), Δ(TC2-CPSCF) and Δ(HPCP-CPSCF)) for the three other methods are reported. In both Tables, the error is evaluated with respect to the system size (number of cells) at each perturbation order. The perturbed density matrices $D^{(x)}$, $D^{(xx)}$ and $D^{(xxx)}$ are explicitly calculated using the protocol in Figure {4.3}, then the responses α_{xx}, β_{xxx} and γ_{xxxx} are deduced from Eq. (4.14). The FFD was excluded from our investigations because of the SCF instabilities. These issues might be related to unconsistencies between the size of the system with respect to the strength of the electric field.

At first glance, the results of Tables {4.4} and {4.5} are quite surprising. For both sets of polymers, the numerical accuracy is dramatically reduced for the CG-CPSCF and TC2-CPSCF, as the perturbation order and the size are increased (blue color in the table),

Order	$\bar{n}$ cells	AO-CPSCF		Δ(CG-CPSCF)		Δ(HPCP-CPSCF)		Δ(TC2-CPSCF)	
		Set 1	Set 2	Set 1	Set 2	Set 1	Set 2	Set 1	Set 2
TPA									
$1^{st}, \alpha_{xx}$									
	18	1.40×10^{3}	4.49×10^{4}	4.60×10^{-2}	4.21×10^{-3}	3.05×10^{-6}	7.97×10^{-9}	3.05×10^{-6}	7.97×10^{-4}
	26	2.07×10^{3}	1.08×10^{5}	6.65×10^{-2}	6.51×10^{-1}	4.56×10^{-6}	2.81×10^{-8}	4.56×10^{-6}	2.79×10^{-4}
	34	2.73×10^{3}	1.98×10^{5}	2.28×10^{-2}	5.52×10^{-4}	6.20×10^{-6}	1.14×10^{-6}	6.20×10^{-6}	1.27×10^{-5}
	42	3.39×10^{3}	3.11×10^{5}	2.86×10^{-2}	5.38×10^{-3}	7.82×10^{-6}	2.00×10^{-7}	7.82×10^{-6}	2.02×10^{-4}
	50	4.05×10^{3}	4.41×10^{5}	3.43×10^{-2}	1.66×10^{-2}	2.91×10^{-8}	4.94×10^{-6}	2.91×10^{-8}	4.94×10^{-3}
	114	8.69×10^{3}	5.83×10^{5}	7.45×10^{-2}	5.92×10^{-2}	8.02×10^{-5}	4.93×10^{-5}	8.02×10^{-5}	4.40×10^{-5}
	122	9.36×10^{3}	8.92×10^{5}	8.03×10^{-2}	2.69×10^{-1}	1.00×10^{-7}	6.87×10^{-7}	1.00×10^{-7}	1.59×10^{-7}
	130	1.00×10^{4}	1.05×10^{6}	8.60×10^{-2}	4.88×10^{-1}	1.09×10^{-7}	1.19×10^{-7}	1.09×10^{-7}	4.59×10^{-7}
$3^{rd}, \gamma_{xxxx}$									
	18	7.84×10^{6}	4.36×10^{10}	6.71×10^{-2}	5.46×10^{-2}	7.77×10^{-7}	4.71×10^{-7}	6.79×10^{-7}	2.97×10^{-4}
	26	1.21×10^{7}	4.41×10^{11}	1.05×10^{-2}	1.51×10^{-2}	2.01×10^{-8}	2.01×10^{-8}	4.00×10^{-7}	9.76×10^{-4}
	34	1.65×10^{7}	2.16×10^{12}	8.90×10^{-1}	2.16×10^{-1}	6.39×10^{-7}	3.64×10^{-6}	4.99×10^{-7}	1.00×10^{-2}
	42	2.09×10^{7}	6.83×10^{12}	4.53×10^{-1}	1.16×10^{-1}	1.65×10^{-6}	1.54×10^{-7}	2.60×10^{-6}	3.20×10^{-1}
	50	2.52×10^{7}	1.61×10^{13}	2.11×10^{-1}	2.65	4.17×10^{-6}	9.78×10^{-6}	2.00×10^{-6}	1.20
	114	7.32×10^{7}	3.12×10^{13}	4.17	9.08	5.78×10^{-5}	4.85×10^{-6}	3.47×10^{-3}	4.85
	122	7.75×10^{7}	7.94×10^{13}	5.86	1.93×10^{1}	9.98×10^{-7}	7.14×10^{-5}	5.06×10^{-3}	3.07×10^{1}
	130	8.19×10^{7}	1.11×10^{14}	1.74×10^{1}	2.47×10^{1}	3.13×10^{-6}	6.41×10^{-6}	6.03×10^{-3}	6.00×10^{1}
TPA+									
$1^{st}, \alpha_{xx}$									
	18	1.62×10^{3}	7.50×10^{4}	3.36×10^{-2}	1.68×10^{-1}	6.09×10^{-7}	6.82×10^{-7}	6.09×10^{-7}	2.25×10^{-6}
	26	2.28×10^{3}	1.90×10^{5}	1.35×10^{-2}	1.41×10^{-1}	7.13×10^{-7}	2.40×10^{-6}	7.12×10^{-7}	2.38×10^{-3}
	34	2.95×10^{3}	3.56×10^{5}	1.94×10^{-2}	6.21	8.63×10^{-7}	3.21×10^{-7}	8.64×10^{-7}	1.80×10^{-6}
	42	3.61×10^{3}	5.38×10^{5}	2.52×10^{-2}	1.04×10^{1}	9.89×10^{-7}	2.14×10^{-6}	9.89×10^{-7}	2.14×10^{-2}
	50	4.27×10^{3}	7.10×10^{5}	3.09×10^{-2}	1.55×10^{1}	3.39×10^{-6}	3.36×10^{-5}	3.39×10^{-6}	2.49×10^{-5}
	114	7.58×10^{3}	1.48×10^{6}	5.97×10^{-2}	1.62×10^{1}	1.10×10^{-5}	1.63×10^{-7}	1.10×10^{-5}	8.51×10^{-2}
	122	8.25×10^{3}	1.64×10^{6}	6.54×10^{-2}	1.46×10^{1}	1.57×10^{-6}	1.32×10^{-7}	1.58×10^{-6}	6.38×10^{-2}
	130	8.91×10^{3}	1.81×10^{6}	7.12×10^{-2}	1.25×10^{1}	1.98×10^{-6}	3.49×10^{-6}	1.99×10^{-6}	2.00
$2^{nd}, \beta_{xxx}$									
	18	2.946×10^{4}	-4.80×10^{7}	2.61	1.63	9.03×10^{-6}	7.68×10^{-6}	9.05×10^{-6}	7.62×10^{-4}
	26	2.958×10^{4}	-3.89×10^{8}	1.23	4.94	3.58×10^{-7}	6.52×10^{-6}	3.51×10^{-7}	6.52
	34	2.966×10^{4}	-1.56×10^{9}	1.25	7.65	3.32×10^{-6}	1.55×10^{-7}	3.30×10^{-6}	1.41×10^{-3}
	42	2.972×10^{4}	-3.60×10^{9}	1.26	1.11	8.90×10^{-6}	3.43×10^{-6}	8.92×10^{-6}	3.43×10^{2}
	50	2.976×10^{4}	-5.61×10^{9}	1.27	2.82×10^{1}	5.18×10^{-5}	4.00×10^{-5}	5.18×10^{-4}	3.98×10^{-1}
	114	2.986×10^{4}	-7.96×10^{9}	1.28	3.41×10^{1}	5.21×10^{-5}	8.54×10^{-7}	5.21×10^{-4}	3.64×10^{-3}
	122	2.987×10^{4}	-8.55×10^{9}	1.28	5.74×10^{1}	3.19×10^{-6}	7.54×10^{-7}	2.76×10^{-6}	4.54×10^{1}
	130	2.988×10^{4}	-8.96×10^{9}	1.28	3.08×10^{1}	1.95×10^{-5}	4.29×10^{-6}	1.94×10^{-4}	9.42×10^{1}
$3^{rd}, \gamma_{xxxx}$									
	18	1.41×10^{7}	-1.75×10^{10}	8.84×10^{-2}	5.46×10^{-1}	7.24×10^{-6}	1.34×10^{-7}	3.31×10^{-3}	6.90×10^{-3}
	26	1.84×10^{7}	1.25×10^{12}	3.05×10^{-1}	1.51×10^{-1}	1.66×10^{-5}	7.48×10^{-5}	3.68×10^{-5}	1.89×10^{-1}
	34	2.27×10^{7}	1.69×10^{13}	4.17×10^{-1}	2.16	4.54×10^{-5}	6.21×10^{-6}	6.01×10^{-5}	1.30
	42	2.71×10^{7}	7.12×10^{13}	2.93	1.16×10^{1}	2.38×10^{-7}	9.04×10^{-6}	2.44×10^{-4}	3.15×10^{1}
	50	3.14×10^{7}	1.49×10^{14}	5.27	2.65	9.54×10^{-6}	7.77×10^{-7}	1.14×10^{-3}	7.10×10^{1}
	114	3.58×10^{7}	3.10×10^{14}	6.31×10^{-1}	9.08×10^{2}	2.71×10^{-5}	3.22×10^{-8}	5.91×10^{-3}	1.78×10^{1}
	122	4.01×10^{7}	3.45×10^{14}	9.62×10^{-1}	1.93×10^{2}	3.44×10^{-6}	4.01×10^{-6}	5.46×10^{-3}	2.11×10^{2}
	130	4.88×10^{7}	3.81×10^{14}	1.23	2.47×10^{2}	1.33×10^{-6}	7.87×10^{-5}	4.83×10^{-2}	7.01×10^{2}

Table 4.4 Numerical accuracy Δ with respect to the AO-CPSCF for the CG-CPSCF, TC2-CPSCF and HPCP-CPSCF, at each perturbation order and for increasing molecular size. Results are obtained for TPA and TPA+.

Order	$\bar{n}$ cells	AO-CPSCF		Δ(CG-CPSCF)		Δ(HPCP-CPSCF)		Δ(TC2-CPSCF)	
		Set 1	Set 2	Set 1	Set 2	Set 1	Set 2	Set 1	Set 2
PPV									
$1^{st}, \alpha_{xx}$									
	4	5.59×10^{2}	1.18×10^{3}	2.48×10^{-6}	5.96×10^{-7}	7.55×10^{-8}	6.09×10^{-8}	7.56×10^{-8}	6.08×10^{-8}
	6	8.32×10^{2}	1.80×10^{3}	3.88×10^{-6}	1.57×10^{-6}	9.58×10^{-8}	4.86×10^{-8}	9.58×10^{-8}	4.86×10^{-8}
	8	1.10×10^{3}	2.42×10^{3}	9.15×10^{-7}	2.68×10^{-6}	6.89×10^{-7}	6.69×10^{-8}	6.89×10^{-7}	6.71×10^{-8}
	10	1.37×10^{3}	3.05×10^{3}	1.47×10^{-6}	3.90×10^{-6}	7.53×10^{-7}	1.06×10^{-7}	7.53×10^{-7}	1.06×10^{-7}
	12	1.65×10^{3}	3.67×10^{3}	1.99×10^{-6}	5.22×10^{-6}	8.80×10^{-7}	1.78×10^{-7}	8.80×10^{-7}	1.77×10^{-7}
	28	3.56×10^{3}	9.29×10^{3}	6.53×10^{-6}	1.93×10^{-5}	1.16×10^{-8}	1.82×10^{-8}	1.16×10^{-8}	1.81×10^{-8}
	30	3.83×10^{3}	9.92×10^{3}	7.18×10^{-6}	2.11×10^{-5}	1.24×10^{-8}	1.34×10^{-8}	1.25×10^{-8}	1.33×10^{-8}
	32	4.10×10^{3}	1.11×10^{4}	6.95×10^{-6}	2.46×10^{-5}	1.34×10^{-8}	5.80×10^{-9}	1.34×10^{-8}	6.00×10^{-9}
$3^{rd}, \gamma_{xxxx}$									
	4	1.32×10^{6}	1.17×10^{7}	7.01×10^{-4}	4.15×10^{-5}	9.05×10^{-8}	8.74×10^{-9}	8.00×10^{-8}	1.99×10^{-7}
	6	2.10×10^{6}	2.05×10^{7}	2.31×10^{-4}	1.52×10^{-4}	3.30×10^{-8}	6.01×10^{-8}	2.06×10^{-6}	1.40×10^{-6}
	8	2.89×10^{6}	2.94×10^{7}	2.16×10^{-3}	7.04×10^{-3}	5.18×10^{-7}	2.61×10^{-7}	1.76×10^{-5}	1.97×10^{-5}
	10	3.69×10^{6}	3.83×10^{7}	9.48×10^{-3}	6.21×10^{-3}	1.93×10^{-8}	5.03×10^{-7}	7.78×10^{-5}	2.01×10^{-5}
	12	4.48×10^{6}	4.72×10^{7}	5.04×10^{-2}	9.81×10^{-2}	5.21×10^{-6}	9.10×10^{-6}	2.49×10^{-4}	7.29×10^{-5}
	28	1.08×10^{7}	1.36×10^{8}	4.42×10^{-1}	6.23×10^{-1}	1.94×10^{-6}	1.17×10^{-7}	3.42×10^{-2}	3.72×10^{-2}
	30	1.16×10^{7}	1.54×10^{8}	6.90×10^{-1}	5.78×10^{-1}	3.19×10^{-7}	7.78×10^{-7}	4.99×10^{-2}	7.30×10^{-2}
	32	1.23×10^{7}	1.63×10^{8}	8.09×10^{-1}	8.46×10^{-1}	3.74×10^{-7}	8.36×10^{-7}	7.09×10^{-2}	9.89×10^{-2}
PPV+									
$1^{st}, \alpha_{xx}$									
	4	7.43×10^{2}	1.55×10^{3}	3.73×10^{-3}	2.88×10^{-3}	1.48×10^{-8}	4.97×10^{-7}	6.10×10^{-11}	5.09×10^{-7}
	6	1.01×10^{3}	2.17×10^{3}	1.03×10^{-3}	1.06×10^{-2}	5.90×10^{-7}	2.62×10^{-7}	7.18×10^{-9}	2.63×10^{-7}
	8	1.28×10^{3}	2.79×10^{3}	1.40×10^{-3}	3.05×10^{-7}	7.53×10^{-7}	2.61×10^{-7}	2.28×10^{-9}	2.64×10^{-7}
	10	1.56×10^{3}	3.41×10^{3}	3.86×10^{-3}	1.68×10^{-7}	8.33×10^{-7}	4.21×10^{-8}	6.42×10^{-8}	3.80×10^{-8}
	12	1.83×10^{3}	4.04×10^{3}	6.37×10^{-3}	5.13×10^{-8}	2.90×10^{-6}	5.52×10^{-8}	3.93×10^{-6}	5.88×10^{-8}
	28	3.19×10^{3}	9.66×10^{3}	1.93×10^{-2}	6.19×10^{-6}	6.52×10^{-9}	2.10×10^{-6}	2.02×10^{-7}	2.10×10^{-6}
	30	3.47×10^{3}	1.02×10^{4}	2.20×10^{-2}	4.91×10^{-6}	1.12×10^{-6}	2.91×10^{-7}	1.26×10^{-6}	2.88×10^{-7}
	32	3.74×10^{3}	1.09×10^{4}	2.47×10^{-2}	5.54×10^{-6}	6.57×10^{-9}	4.06×10^{-7}	1.87×10^{-7}	4.05×10^{-7}
$2^{nd}, \beta_{xxx}$									
	4	-1.097×10^{4}	-5.970×10^{4}	1.60	7.66×10^{-2}	6.83×10^{-8}	1.16×10^{-5}	1.21×10^{-6}	4.97×10^{-4}
	6	-1.099×10^{4}	-5.981×10^{4}	1.57	1.85×10^{-1}	7.20×10^{-7}	1.05×10^{-5}	2.88×10^{-5}	1.05×10^{-5}
	8	-1.102×10^{4}	-6.014×10^{4}	1.58	9.57×10^{-4}	5.70×10^{-8}	2.13×10^{-7}	1.86×10^{-4}	2.21×10^{-7}
	10	-1.103×10^{4}	-6.040×10^{4}	1.58	1.92×10^{-3}	4.94×10^{-6}	1.17×10^{-6}	6.33×10^{-4}	1.15×10^{-6}
	12	-1.105×10^{4}	-6.058×10^{4}	1.58	6.89×10^{-4}	4.31×10^{-5}	1.09×10^{-8}	2.28×10^{-3}	5.07×10^{-8}
	28	-1.108×10^{4}	-6.105×10^{4}	1.56	1.23×10^{-3}	9.72×10^{-7}	3.56×10^{-7}	2.21×10^{-2}	2.86×10^{-7}
	30	-1.1093×10^{4}	-6.110×10^{4}	1.55	1.10×10^{-3}	1.47×10^{-5}	6.21×10^{-7}	3.17×10^{-2}	4.51×10^{-7}
	32	-1.1096×10^{4}	-6.115×10^{4}	1.53	1.42×10^{-3}	1.11×10^{-6}	6.67×10^{-7}	8.71×10^{-2}	8.24×10^{-7}
$3^{rd}, \gamma_{xxxx}$									
	4	3.18×10^{6}	3.10×10^{7}	6.64×10^{-1}	5.20×10^{-2}	9.05×10^{-8}	8.03×10^{-8}	5.46×10^{-2}	2.05×10^{1}
	6	3.94×10^{6}	3.83×10^{7}	9.07	7.76×10^{-2}	3.30×10^{-7}	2.34×10^{-7}	1.77	9.59×10^{-6}
	8	4.72×10^{6}	4.69×10^{7}	5.15	2.26×10^{-1}	5.8×10^{-6}	6.00×10^{-7}	1.51×10^{1}	3.14×10^{-4}
	10	5.50×10^{6}	5.56×10^{7}	6.01×10^{1}	3.04×10^{-1}	3.33×10^{-6}	5.74×10^{-6}	6.01×10^{1}	2.60×10^{-3}
	12	6.29×10^{6}	6.45×10^{7}	2.10×10^{2}	9.48	1.01×10^{-6}	1.49×10^{-5}	2.16×10^{2}	1.14×10^{-2}
	28	1.02×10^{7}	1.09×10^{8}	3.08×10^{3}	7.41	2.74×10^{-8}	3.08×10^{-8}	9.08×10^{3}	9.84×10^{-4}
	30	1.10×10^{7}	1.17×10^{8}	6.42×10^{4}	6.64×10^{-1}	5.21×10^{-6}	8.25×10^{-7}	1.42×10^{4}	9.95×10^{-4}
	32	1.18×10^{7}	1.53×10^{8}	7.90×10^{4}	1.42×10^{-1}	4.24×10^{-6}	9.41×10^{-7}	1.01×10^{4}	3.21×10^{-3}

Table 4.5 Numerical accuracy Δ with respect to the AO-CPSCF for the CG-CPSCF, TC2-CPSCF and HPCP-CPSCF, at each perturbation order and for increasing molecular size. Results are obtained for PPV and PPV+.

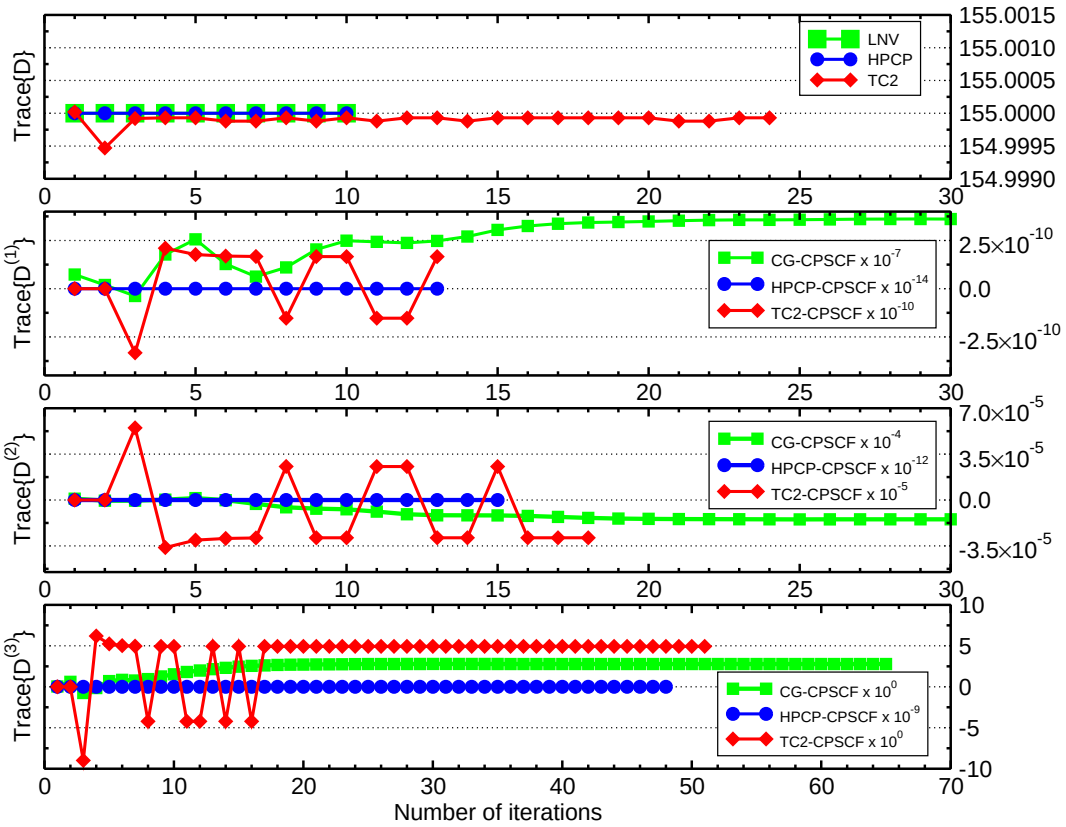


Fig. 4.6 Trace of the density matrices during the SCF iterations for the HPCP-CPSCF, TC2-CPSCF and CG-CPSCF, at zero (D), first ($D^{(1)} = D^{(x)}$), second ($D^{(2)} = D^{(xx)}$) and third ($D^{(3)} = D^{(xxx)}$) orders. Results obtained for PPV+ of Set 1 with $\bar{n} = 28$.

while the HPCP-CPSCF always conserves a remarkable numerical accuracy, whatever the order of perturbation. In order to understand these results, we have probed the density matrix during the SCF iterations, for each method at all the orders. Figures {4.6} and {4.7} display the trace and the idempotency characters of the density matrix, respectively, the latter being evaluated by

$$\Delta_{\text{Idemp}} = \| D^{(k)} - \sum_{l=0}^{k} D^{(l)} D^{(k-l)} \| \tag{4.17}$$

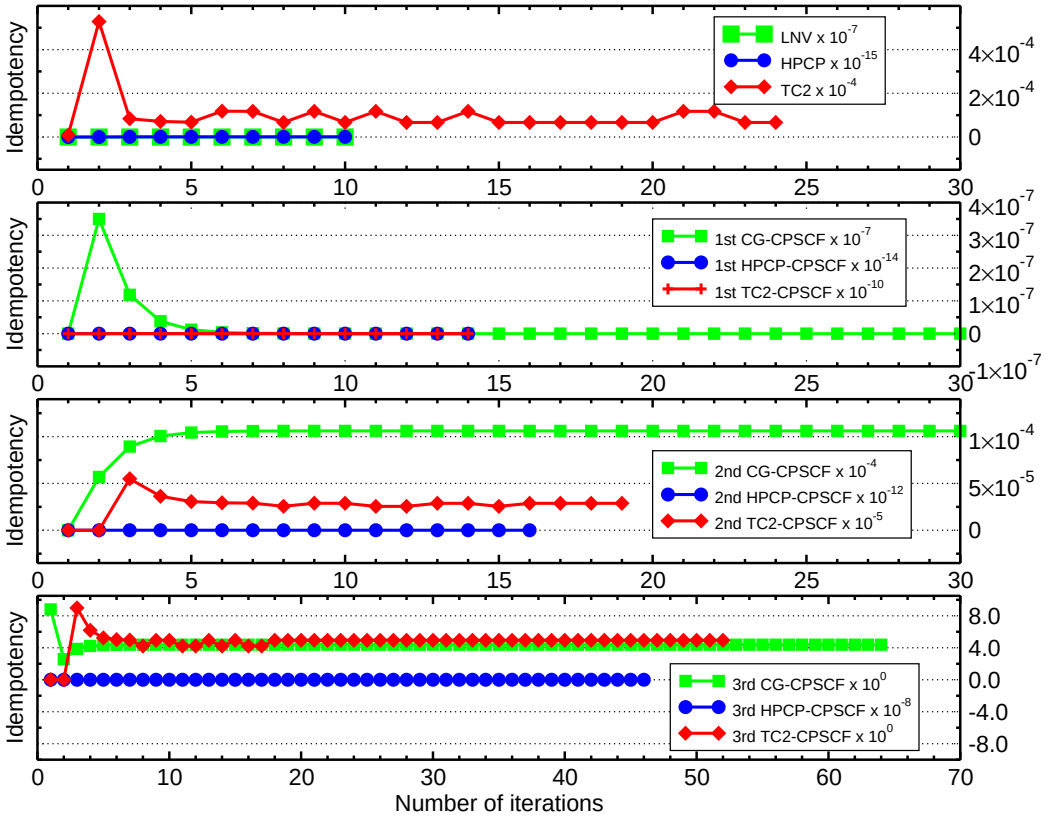


Fig. 4.7 Idempotency of the density matrices during the SCF iterations for the HPCP-CPSCF, TC2-CPSCF and CG-CPSCF, at zero (D), first ($D^{(1)} = D^{(x)}$), second ($D^{(2)} = D^{(xx)}$) and third ($D^{(3)} = D^{(xxx)}$) orders. Results obtained for PPV+ of Set 1 with $\bar{n} = 28$.

Basically, the trace of the unperturbed density matrix must correspond to the number of occupied states. On the other hand, at any perturbation order, the traces of the perturbed density matrices must be zero. For the HPCP-CPSCF, from first to third order, the trace of the perturbed density matrices is always zero during the iterations. This should be compared to the CG-CPSCF and TC2-CPSCF approaches, where the

trace oscillates, to eventually reach non-zero values. The same behaviour is also observed for the idempotency in Figure {4.7}. In view of these results, the conservation of the numerical accuracy especially at third order for HPCP-CPSCF, is likely due to the conservation of the trace of the unperturbed density matrix D at each iteration step. Even though a good preconditioning may prevent convergence instabilities, the fact that the TC2-CPSCF and CG-CPSCF do not enforce N-representability conditions, they can not avoid any departure from the physical requirements. However, it is worthwhile to note that, in terms of error percentage CG-CPSCF and TC2-CPSCF remain valid. For

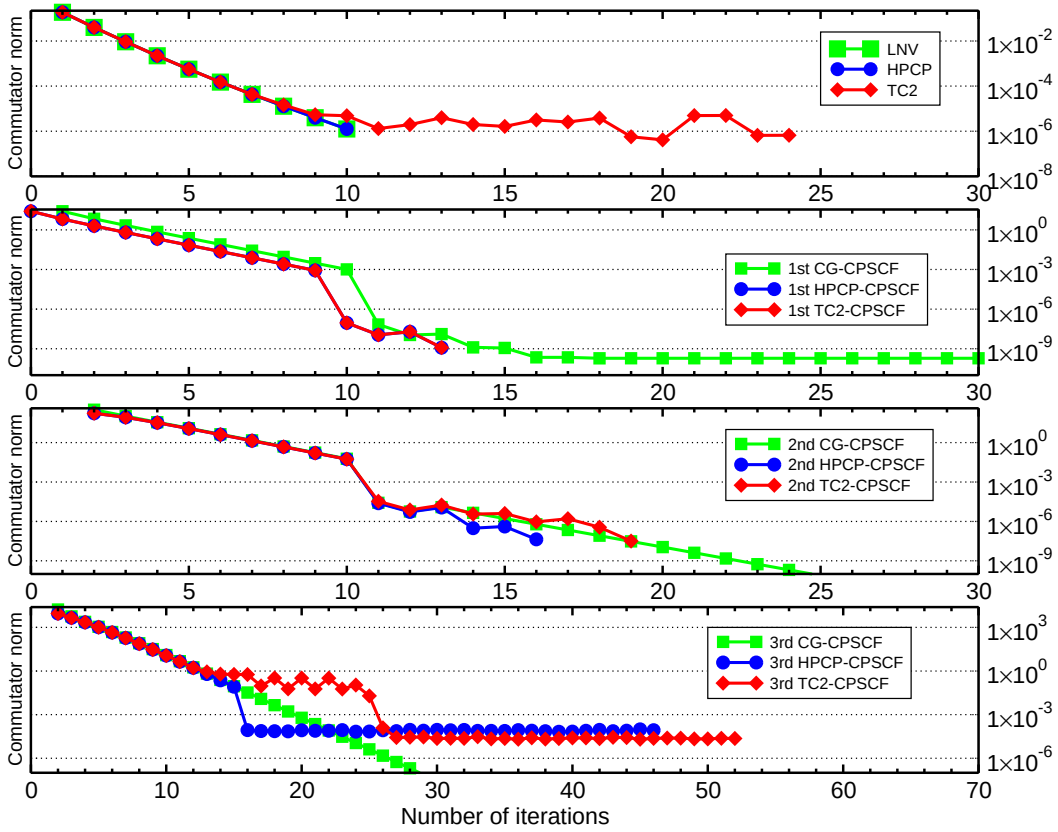
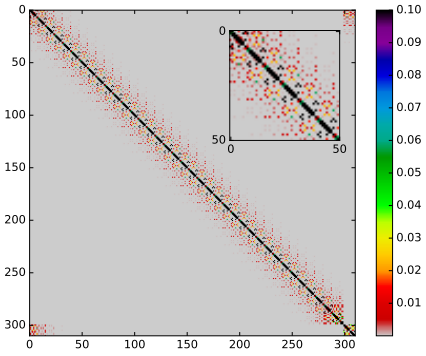
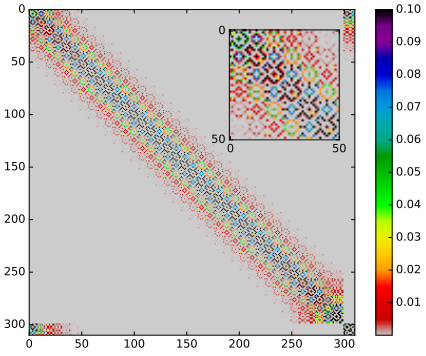


Fig. 4.8 Frobenius norm of the error vector [cf. Eq. (3.69)] of the density matrices during the SCF iterations for the HPCP-CPSCF, TC2-CPSCF and CG-CPSCF, at zero (D), first ($D^{(1)} = D^{(x)}$), second ($D^{(2)} = D^{(xx)}$) and third ($D^{(3)} = D^{(xxx)}$) orders. This example is for PPV+ of Set 1 with $\bar{n} = 28$.

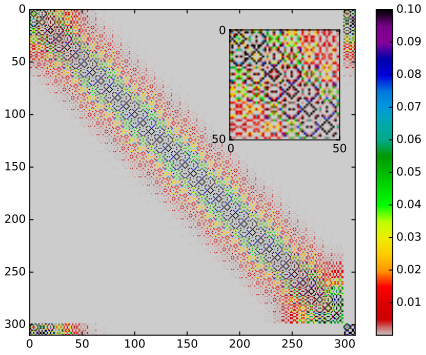
example, for the largest error (PPV+ of Set 1, with $\bar{n}$=32), %error $\sim 0.1\%$ which is insignificant. Using the two forms of matrix illustration already used in Section 2.4.2, the Figures {4.9} and {4.10}, display the profile of the density matrix obtained at the end of the SCF procedure. As the order is increasing, the number nnz of significant elements in the density matrix is also increasing, whereas their magnitude decreases.



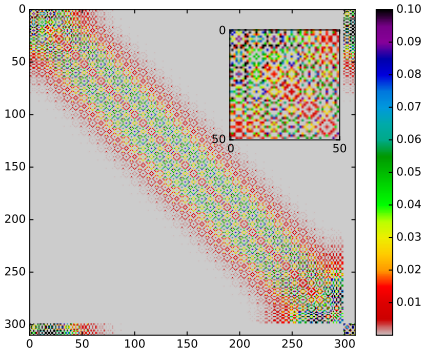
(a) D with $nnz = 7\%$



(b) $D^{(1)}$ with $nnz = 20\%$

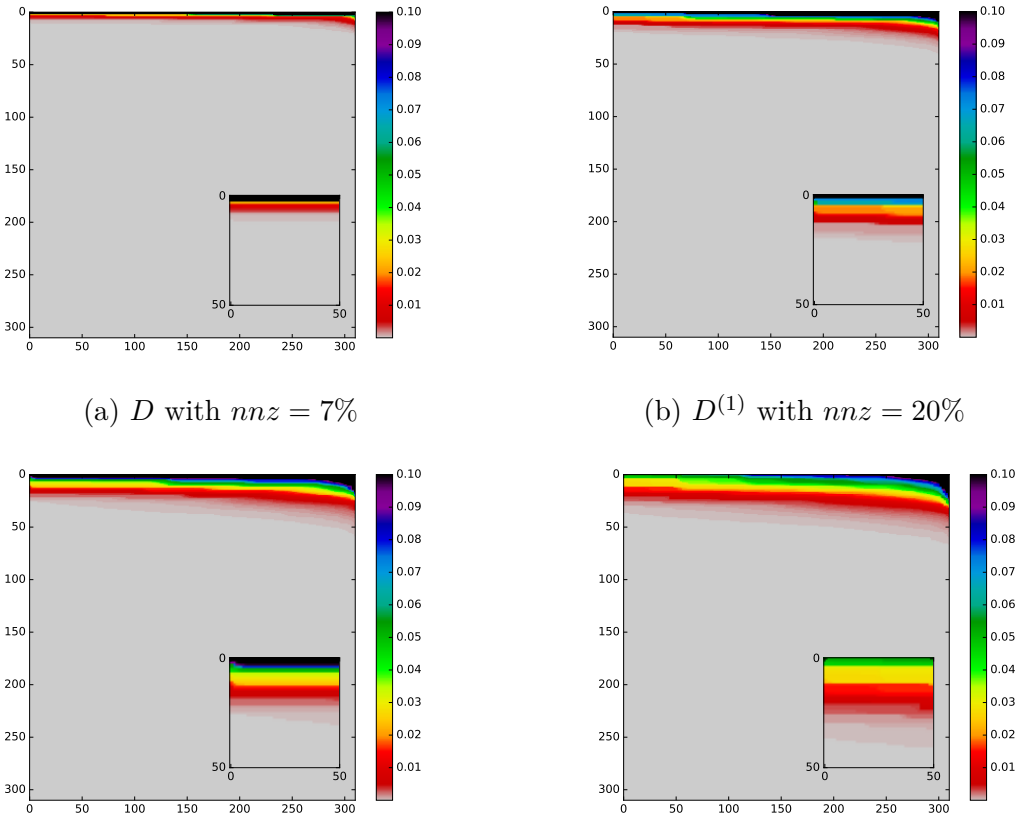


(c) $D^{(2)}$ with $nnz = 45\%$



(d) $D^{(3)}$ with $nnz = 67\%$

Fig. 4.9 First form of illustration for the converged density matrices using the HPCP, from zero to third order. nnz is the number of non zero elements at 10^{-3}. This example is for PPV+ of Set 1 with $\bar{n} = 28$. $D^{(1)} = D^{(x)}$, $D^{(2)} = D^{(xx)}$, $D^{(3)} = D^{(xxx)}$.



(a) D with $nnz = 7\%$

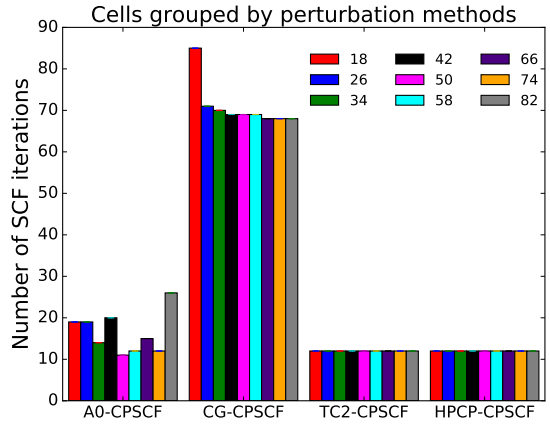
(b) $D^{(1)}$ with $nnz = 20\%$

(c) $D^{(2)}$ with $nnz = 45\%$

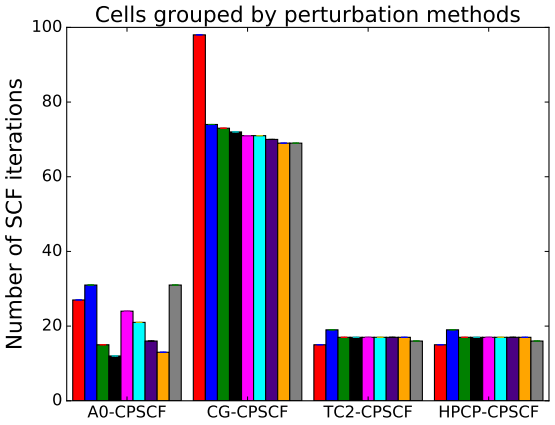
(d) $D^{(3)}$ with $nnz = 67\%$

Fig. 4.10 Second form of illustration for the converged density matrices using the HPCP, from zero to third order. nnz is the number of non zero elements at 10^{-3}. This example is for PPV+ of Set 1 with $\bar{n} = 28$. $D^{(1)} = D^{(x)}$, $D^{(2)} = D^{(xx)}$, $D^{(3)} = D^{(xxx)}$.

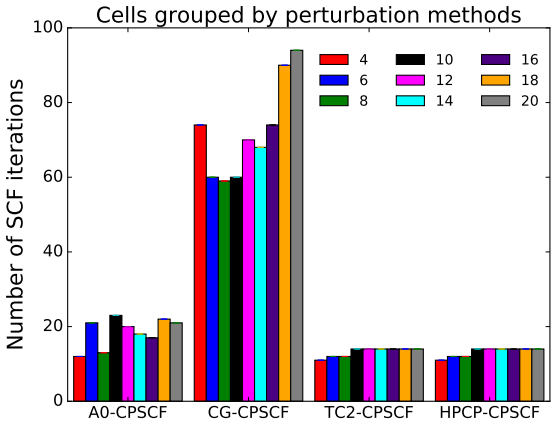
From Figures {4.7}, we note the number of iterations for both TC2-CPSCF and HPCP-CPSCF at third order is large compared at first and second orders. This result is explained by the norm of the error vector [cf. Eq. (3.69)] represented in Figure {4.8}. The value of $\sim 10^3$ at the early steps of the 3rd-order calculation compared to ~ 1 at first and second order demonstrates that we are beyond the domain of applicability of the D-DIIS[83, 81]. As a result, in our implementation, we found a convergence instability for TC2-CPSCF and HPCP-CPSCF at third order. However, as shown on Figure {4.11}, the number of iterations obtained for lower orders and for each of the density matrix perturbation methods, shows clearly that the perturbed projections are the most efficient in terms of SCF iterations.



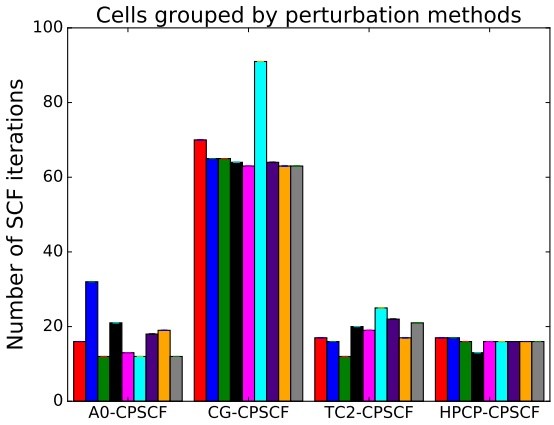
(a) 1st order, TPA+



(b) 2nd order, TPA+



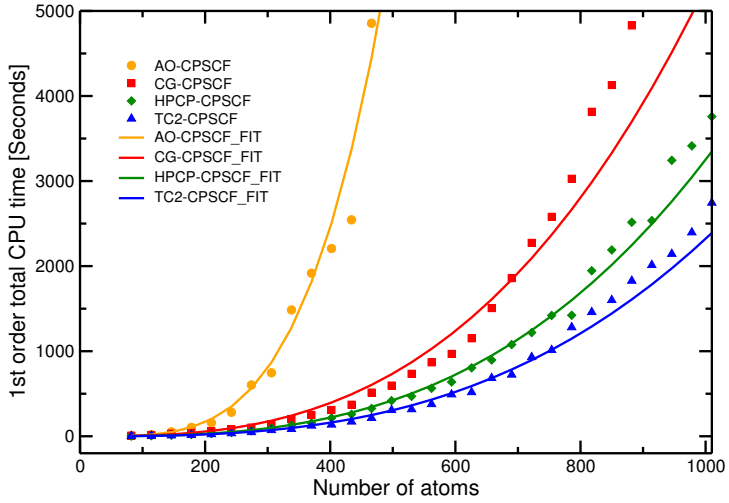
(c) 1st order, PPV+



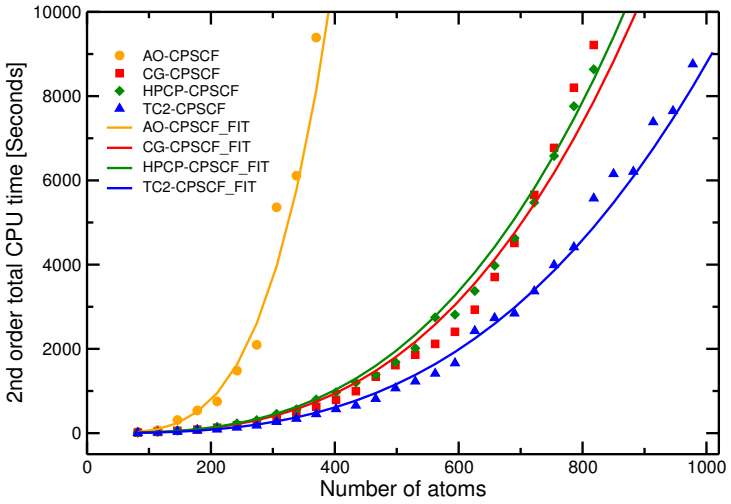
(d) 2nd order, PPV+

Fig. 4.11 Histogram of number of SCF iterations with respect to the size of the systems, for the four density matrix perturbation methods at first and second orders. The results are obtained for TPA+ and PPV+ of Set 1.

The calculation time for α_{xx} and β_{xxx} is represented as a function of number of atoms on Figure {4.12}. The TC2-CPSCF and HPCP-CPSCF are the most efficent, with a better performance for the TC2-CPSCF. Curve fitting reveals that the AO-CPSCF



(a) 1st order, TPA+

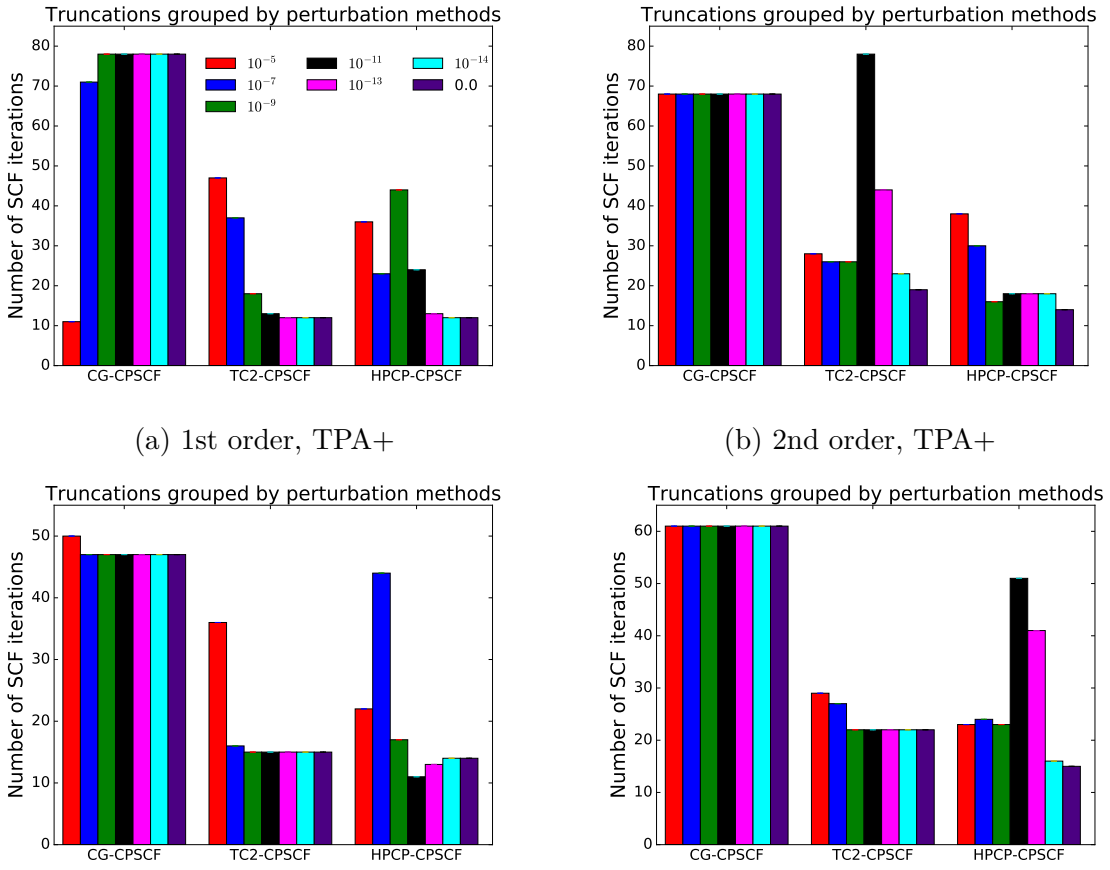


(b) 2nd order, TPA+

Fig. 4.12 Calculation time for the polarizability α_{xx} and first hyperpolarizability β_{xxx} as a function of number of atoms. Results are obtained for TPA+ (Set 1). For each density matrix perturbation method, the calculation time is pictured along with its fit.

scales as a power of 4 compared to the cubic scaling of the CG-CPSCF, TC2-CPSCF and HPCP-CPSCF. This demonstrates that the density matrix methods are already more efficient than the diagonalization.

4.4 Perturbed linear scaling calculation



(a) 1st order, TPA+

(b) 2nd order, TPA+

(c) 1st order, PPV+

(d) 2nd order, PPV+

Fig. 4.13 Histogram of number of SCF iterations with respect to the numerical threshold τ. Results are obtained for the four density matrix perturbation methods at first and second orders. The number of cells is fixed at 250 for TPA+ and 62 PPV+ of Set 1.

In Figure {4.13} is presented the number of iterations obtained for different values of numerical threshold τ (for a fixed size of polymer). We observe that the perturbed projections TC2-CPSCF and HPCP-CPSCF lead to better performances compared to the CG-CPSCF. However, the CG-CPSCF is more stable with a regular number of iterations whatever is the value of τ. The calculation time and the convergence of the

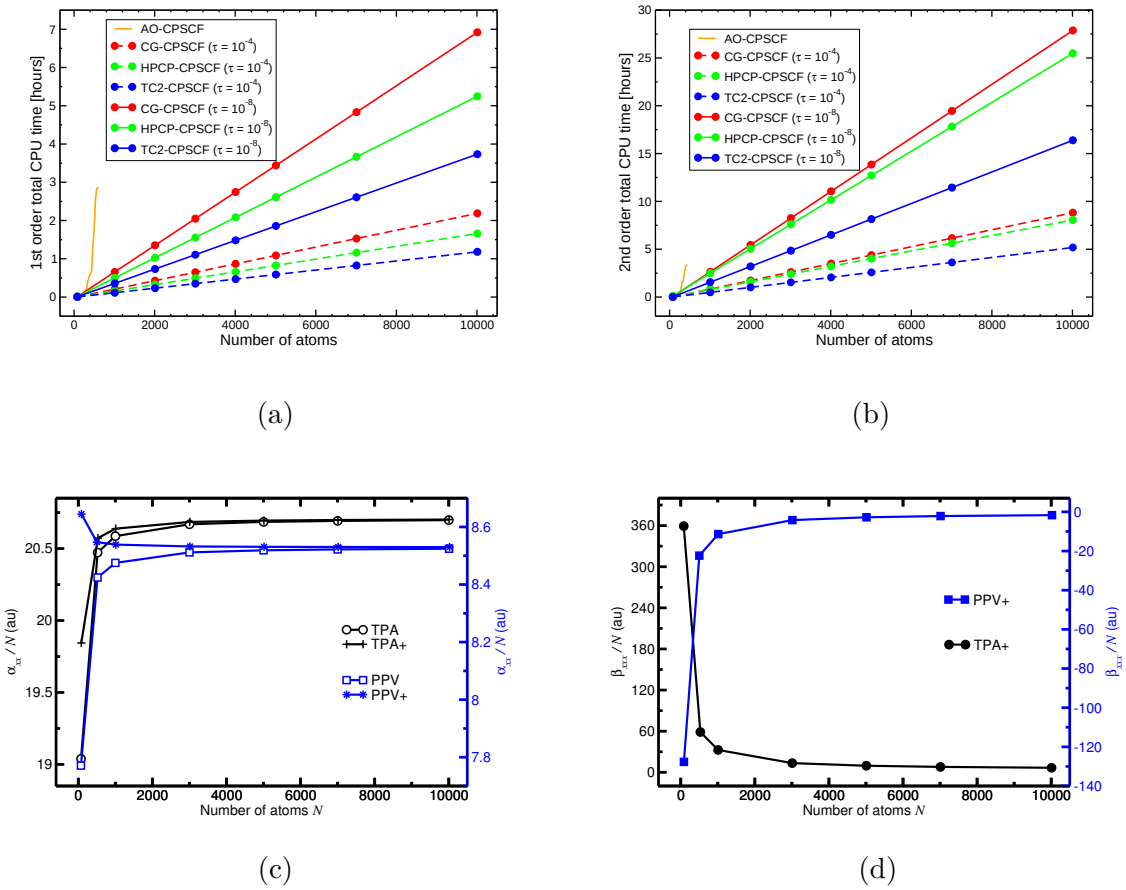


Fig. 4.14 In (a) and (b), representation of the total calculation time as a function of the number of atoms, for TPA+ (Set 1). This representation compares the AO-CPSCF, CG-CPSCF, TC2-CPSCF and HPCP-CPSCF. The density matrix methods are all truncated at $\tau = 10^{-4}$ and $\tau = 10^{-8}$. In (c) and (d), convergence for (hyper)polarizability per atom as a function of number of atoms. The polarizability α_{xx} and first hyperpolarizability β_{xxx} are calculated using the HPCP-CPSCF truncated at $\tau = 10^{-8}$ for the 4 polymers of Set 1.

responses with respect to the system size are presented in Figure {4.14}. Independently on the method, in Figure {4.14}(a) and (b), we found that the linear scaling is achieved and the TC2-CPSCF is clearly the most efficient. From Figure {4.14}(c) and (d), we observe that a number of 10,000 atoms is sufficient to reach the plateau related to the infinite chain limit.

Conclusions and outlooks

In this thesis, we have confirmed that solving electronic structure without relying on the resolution of the Schrödinger equation is a promising approach which can be extended to the perturbation theory. As a matter of fact, the alternative solution deriving from the Liouville-von Neumann equation is the kernel of this work. For single-determinant approximation, the one-particle density matrix is necessary and sufficient to access all the electronic properties of the system and allows to bypass the computational demanding task related to the eigenvalue problem resolution, ie. the diagonalization. In a first part, from the description of the general framework of the density matrix minimization and polynomial expansion, we have proposed a canonical density matrix purification which respects the N-representability constraints. We have emphasized that this purification method is self-consistent, in the sense that, it does not rely on heuristic adjustement of the polynomial during the iterative process. From numerical experiments, the purification polynomial has shown good performances with respect to the other schemes, although its efficiency degrades for the pathological cases. Furthermore, when combined with sparse-matrix algebra to reach the linear-scaling regime, this new purification method is the second most efficient in terms of CPU time. However, as the other density matrix purifications, our variant approach presents the same instability symptoms when using the radial truncation. As a solution to this issue, we have proposed to relax the radial truncation when close to critical points. It is important to specify that this solution is not fully satisfactory. The problem of the radial truncation to the density matrix purifications truly deserves more attention.

In a second part, assuming an orthogonal basis, a detailed development of the density matrix perturbation theory has been presented. One of the presented perturbative methods is related to our new purification variant, corresponding to a new canonical and non heuristic density matrix perturbation method. Comparisons with other perturbed purification and minimization methods have been performed. A detailed analysis of the results has revealed that our method is more robust with a remarkable numerical accuracy, despite its important number of matrix multiplications. The new canonical density matrix

perturbation method developed in this thesis is very promising thanks to the explicit consideration of the N-representability properties. The different density-matrix based pertubation methods discussed in this manuscript have been implemented in a code based on the Pariser-Parr-Pople Hartree-Fock model. The most likely next step is to extend this work to a general code including explicit non-orthogonal basis sets where the filling factor significantly deviates from 1/2. The idea will be then to investigate in more details the performance of the method within the linear-scaling regime, especially the influence of the sparsity on the accuracy of the perturbed quantities. Application to dynamic response calculation through the resolution of the time-dependent Liouville-von Neumann equation can also be envisaged.

From our very recent works, we finally noted that the N-representability properties for the density matrix can also be applied to the energy functional minimization. We found that it is possible to enforce these properties during the minimization, that is, the trace is conserved and the density matrix eigenvalues lied in the range 0 to 1, by using a suitably preconditioned conjugate gradient algorithm. Solving this point would be benificial for curing deficiencies observed in energy functional based density matrix perturbation theory.

References

[1] Attila Szabo and Neil S Ostlund. Modern quantum chemistry: introduction to advanced electronic structure theory. Courier Corporation, 2012.

[2] Robert G Parr. Density functional theory of atoms and molecules. Springer, 1980.

[3] Mike C Payne, Michael P Teter, Douglas C Allan, TA Arias, and JD Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. Reviews of Modern Physics, 64(4):1045, 1992.

[4] Yousef Saad. Iterative methods for sparse linear systems. Siam, 2003.

[5] Xavier Gonze. First-principles responses of solids to atomic displacements and homogeneous electric fields: Implementation of a conjugate-gradient algorithm. Physical Review B, 55(16):10337, 1997.

[6] Stefano Baroni, Stefano De Gironcoli, Andrea Dal Corso, and Paolo Giannozzi. Phonons and related crystal properties from density-functional perturbation theory. Reviews of Modern Physics, 73(2):515, 2001.

[7] Stefan Goedecker. Linear scaling electronic structure methods. Reviews of Modern Physics, 71(4):1085, 1999.

[8] Robert Zalesny, Manthos G Papadopoulos, Paul G Mezey, and Jerzy Leszczynski. Linear-scaling techniques in computational chemistry and physics: Methods and applications. Springer Science+ Business Media BV, 2011.

[9] DR Bowler and T Miyazaki. methods in electronic structure calculations. Reports on Progress in Physics, 75(3):036503, 2012.

[10] Jörg Kussmann, Matthias Beer, and Christian Ochsenfeld. Linear-scaling self-consistent field methods for large molecules. Wiley Interdisciplinary Reviews: Computational Molecular Science, 3(6):614–636, 2013.

[11] Walter Kohn. Density functional and density matrix method scaling linearly with the number of atoms. Physical Review Letters, 76(17):3168, 1996.

[12] Emil Prodan and Walter Kohn. Nearsightedness of electronic matter. Proceedings of the National Academy of Sciences of the United States of America, 102(33):11635–11638, 2005.

[13] Jacques Des Cloizeaux. Energy bands and projection operators in a crystal: analytic and asymptotic properties. Physical Review, 135(3A):A685, 1964.

[14] Sohrab Ismail-Beigi and TA Arias. Locality of the density matrix in metals, semiconductors, and insulators. Physical review letters, 82(10):2127, 1999.

[15] Matt Challacombe. A general parallel sparse-blocked matrix multiply for linear scaling scf theory. Computer physics communications, 128(1):93–107, 2000.

[16] David R Bowler, T Miyazaki, and MJ Gillan. Parallel sparse matrix multiplication for linear scaling electronic structure calculations. Computer physics communications, 137(2):255–273, 2001.

[17] RM Stevens, RM Pitzer, and WN Lipscomb. Perturbed hartree—fock calculations. i. magnetic susceptibility and shielding in the lih molecule. The Journal of Chemical Physics, 38(2):550–560, 1963.

[18] J Gerratt and Ian M Mills. Force constants and dipole-moment derivatives of molecules from perturbed hartree–fock calculations. i. The Journal of Chemical Physics, 49(4):1719–1729, 1968.

[19] J Gerratt and IM Mills. Force constants and dipole-moment derivatives of molecules from perturbed hartree–fock calculations. ii. applications to limited basis-set scf–mo wavefunctions. The Journal of Chemical Physics, 49(4):1730–1739, 1968.

[20] Knud Thomsen and Peter Swanstrøm. Calculation of molecular one-electron properties using coupled hartree-fock methods: I. computational scheme† part of this work was performed at the max-planck-institut für physik und astrophysik, münchen. Molecular Physics, 26(3):735–750, 1973.

[21] Robert Ditchfield. Self-consistent perturbation theory of diamagnetism: I. a gauge-invariant lcao method for nmr chemical shifts. Molecular Physics, 27(4):789–807, 1974.

[22] J_A Pople, R Krishnan, HB Schlegel, and J S_ Binkley. Derivative studies in hartree-fock and møller-plesset theories. International Journal of Quantum Chemistry, 16(S13):225–241, 1979.

[23] Michael Frisch, Martin Head-Gordon, and John Pople. Direct analytic scf second derivatives and electric field properties. Chemical physics, 141(2):189–196, 1990.

[24] Y Osamura, Y Yamaguchi, P Saxe, DJ Fox, MA Vincent, and HF Schaefer. Analytic second derivative techniques for self-consistent-field wave functions. a new approach to the solution of the coupled perturbed hartree-fock equations. Journal of Molecular Structure: THEOCHEM, 103:183–196, 1983.

[25] Rev McWeeny. Perturbation theory for the fock-dirac density matrix. Physical Review, 126(3):1028, 1962.

[26] G Diercksen and R McWeeny. Self-consistent perturbation theory. i. general formulation and some applications. The Journal of Chemical Physics, 44(9):3554–3560, 1966.

[27] R Moccia. Perturbed scf mo calculations. electrical polarizability and magnetic susceptibility of hf, h2o, nh3 and ch4. Theoretica chimica acta, 8(3):192–202, 1967.

[28] R Moccia. Variable bases in scf mo calculations. Chemical Physics Letters, 5(5):260–264, 1970.

[29] Janet L Dodds, Roy McWeeny, and Andrzej J Sadlej. Self-consistent perturbation theory: Generalization for perturbation-dependent non-orthogonal basis set† this research was partly supported by the institute of low temperatures and structure research of the polish academy of sciences. Molecular Physics, 34(6):1779–1791, 1977.

[30] JL Dodds, R McWeeny, WT Raynes, and JP Riley. Scf theory for multiple perturbations. Molecular Physics, 33(3):611–617, 1977.

[31] Krzysztof Woliński and Andrzej J Sadlej. Self-consistent perturbation theory: Open-shell states in perturbation-dependent non-orthogonal basis sets† this work was partly supported by the institute of low temperatures and structure research of the polish academy of sciences under contract no. mr-i. 9.4. 3/2. Molecular Physics, 41(6):1419–1430, 1980.

[32] Krzysztof Wolinski, James F Hinton, and Peter Pulay. Efficient implementation of the gauge-independent atomic orbital method for nmr chemical shift calculations. Journal of the American Chemical Society, 112(23):8251–8260, 1990.

[33] Christopher A White, Benny G Johnson, Peter MW Gill, and Martin Head-Gordon. The continuous fast multipole method. Chemical physics letters, 230(1):8–16, 1994.

[34] Matt Challacombe and Eric Schwegler. Linear scaling computation of the fock matrix. The Journal of chemical physics, 106(13):5526–5536, 1997.

[35] Eric Schwegler, Matt Challacombe, and Martin Head-Gordon. Linear scaling computation of the fock matrix. ii. rigorous bounds on exchange integrals and incremental fock build. The Journal of chemical physics, 106(23):9708–9717, 1997.

[36] Christian Ochsenfeld, Christopher A White, and Martin Head-Gordon. Linear and sublinear scaling formation of hartree–fock-type exchange matrices. The Journal of chemical physics, 109(5):1663–1669, 1998.

[37] Christian Ochsenfeld and Martin Head-Gordon. A reformulation of the coupled perturbed self-consistent field equations entirely within a local atomic orbital density matrix-based scheme. Chemical physics letters, 270(5):399–405, 1997.

[38] Valéry Weber, Anders MN Niklasson, and Matt Challacombe. Ab initio linear scaling response theory: Electric polarizability by perturbed projection. Physical review letters, 92(19):193002, 2004.

[39] Valéry Weber, Anders MN Niklasson, and Matt Challacombe. Higher-order response in o (n) by perturbed projection. The Journal of chemical physics, 123(4):044106, 2005.

[40] X-P Li, RW Nunes, and David Vanderbilt. Density-matrix electronic-structure method with linear system-size scaling. Physical Review B, 47(16):10891, 1993.

[41] R McWeeny. The density matrix in self-consistent field theory. i. iterative construction of the density matrix. In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, volume 235, pages 496–509. The Royal Society, 1956.

[42] Rev McWeeny. Some recent advances in density matrix theory. Reviews of Modern Physics, 32(2):335, 1960.

[43] Jörg Kussmann and Christian Ochsenfeld. Linear-scaling method for calculating nuclear magnetic resonance chemical shifts using gauge-including atomic orbitals within hartree-fock and density-functional theory. The Journal of chemical physics, 127(5):054103, 2007.

[44] Jörg Kussmann and Christian Ochsenfeld. A density matrix-based method for the linear-scaling calculation of dynamic second-and third-order properties at the hartree-fock and kohn-sham density functional theory levels. The Journal of chemical physics, 127(20):204103, 2007.

[45] Erwin Schrödinger. An undulatory theory of the mechanics of atoms and molecules. Physical Review, 28(6):1049, 1926.

[46] John Von Neumann. Mathematical foundations of quantum mechanics. Number 2. Princeton university press, 1955.

[47] Per-Olov Löwdin. Quantum theory of many-particle systems. i. physical interpretations by means of density matrices, natural spin-orbitals, and convergence problems in the method of configurational interaction. Physical Review, 97(6):1474, 1955.

[48] Per-Olov Löwdin. Quantum theory of many-particle systems. ii. study of the ordinary hartree-fock approximation. Physical Review, 97(6):1490, 1955.

[49] Per-Olov Löwdin. Quantum theory of many-particle systems. iii. extension of the hartree-fock scheme to include degenerate systems and correlation effects. Physical review, 97(6):1509, 1955.

[50] Roy McWeeny. The density matrix in many-electron quantum mechanics. i. generalized product functions. factorization and physical interpretation of the density matrices. In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, volume 253, pages 242–259. The Royal Society, 1959.

[51] R McWeeny and Y Mizuno. The density matrix in many-electron quantum mechanics. ii. separation of space and spin variables; spin coupling problems. In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, volume 259, pages 554–577. The Royal Society, 1961.

[52] John M Herbert. Reconstructive approaches to one-and two-electron density matrix theory. PhD thesis, University of Wisconsin–Madison, 2003.

[53] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. Physical review, 136(3B):B864, 1964.

[54] Valentin V Karasiev and Samuel B Trickey. Issues and challenges in orbital-free density functional calculations. Computer Physics Communications, 183(12):2519–2527, 2012.

[55] Junchao Xia, Chen Huang, Ilgyou Shin, and Emily A Carter. Can orbital-free density functional theory simulate molecules? The Journal of chemical physics, 136(8):084102, 2012.

[56] Mohan Chen, Xiang-Wei Jiang, Houlong Zhuang, Lin-Wang Wang, and Emily A Carter. Petascale orbital-free density functional theory enabled by small-box algorithms. Journal of chemical theory and computation, 2016.

[57] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. Physical review, 140(4A):A1133, 1965.

[58] A John Coleman. Structure of fermion density matrices. Reviews of modern Physics, 35(3):668, 1963.

[59] Hans Kummer. n-representability problem for reduced density matrices. Journal of Mathematical Physics, 8(10):2063–2081, 1967.

[60] AJ Coleman. Necessary conditions for n-representability of reduced density matrices. Journal of Mathematical Physics, 13(2):214–222, 1972.

[61] Mel Levy. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. Proceedings of the National Academy of Sciences, 76(12):6062–6065, 1979.

[62] Elliott H Lieb. Density functionals for coulomb systems. In Inequalities, pages 269–303. Springer, 2002.

[63] DR Bowler, T Miyazaki, and MJ Gillan. Recent progress in linear scaling ab initio electronic structure techniques. Journal of Physics: Condensed Matter, 14(11):2781, 2002.

[64] Christian Ochsenfeld, Jorg Kussmann, and Daniel S Lambrecht. Linear-scaling methods in quantum chemistry. Reviews in computational chemistry, 23:1, 2007.

[65] Christian Ochsenfeld. Linear scaling exchange gradients for hartree–fock and hybrid density functional theory. Chemical Physics Letters, 327(3):216–223, 2000.

[66] Rudolph Pariser and Robert G Parr. A semi-empirical theory of the electronic spectra and electronic structure of complex unsaturated molecules. i. The Journal of Chemical Physics, 21(3):466–471, 1953.

[67] Rudolph Pariser and Robert G Parr. A semi-empirical theory of the electronic spectra and electronic structure of complex unsaturated molecules. ii. The Journal of Chemical Physics, 21(5):767–776, 1953.

[68] JA Pople. Electron interaction in unsaturated hydrocarbons. Transactions of the Faraday Society, 49:1375–1385, 1953.

[69] Noboru Mataga and Kitisuke Nishimoto. Electronic structure and spectra of nitrogen heterocycles. Z. phys. Chem, 13:140, 1957.

[70] Kimio Ohno. Some remarks on the pariser-parr-pople method. Theoretica chimica acta, 2(3):219–227, 1964.

[71] K Schulten, I_ Ohmine, and M Karplus. Correlation effects in the spectra of polyenes. The Journal of Chemical Physics, 64(11):4422–4441, 1976.

[72] Dawei Zhang, Zexing Qu, Chungen Liu, and Yuansheng Jiang. Excitation energy calculation of conjugated hydrocarbons: A new pariser–parr–pople model parameterization approaching caspt2 accuracy. The Journal of chemical physics, 134(2):024114, 2011.

[73] Priya Sony and Alok Shukla. A general purpose fortran 90 electronic structure program for conjugated systems using pariser–parr–pople model. Computer Physics Communications, 181(4):821–830, 2010.

[74] Gundra Kondayya and Alok Shukla. A fortran 90 hartree–fock program for one-dimensional periodic π-conjugated systems using pariser–parr–pople model. Computer Physics Communications, 183(3):677–689, 2012.

[75] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. Technical University of Denmark, 7:15, 2008.

[76] Frank Liu. The extremum method in quantum chemistry: Part i. extremum of tr (btpc) and the trace algebra. Journal of Molecular Structure: THEOCHEM, 226(3):197–209, 1991.

[77] Frank Liu. The extremum method in quantum chemistry: Part ii. introduction to some applications. Journal of Molecular Structure: THEOCHEM, 230:47–65, 1991.

[78] Satrajit Adhikari and Roi Baer. Augmented lagrangian method for order-n electronic structure. The Journal of chemical physics, 115(1):11–14, 2001.

[79] Clemens Carel Johannes Roothaan. New developments in molecular orbital theory. Reviews of modern physics, 23(2):69, 1951.

[80] George G Hall. The molecular orbital theory of chemical valency. viii. a method of calculating ionization potentials. In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, volume 205, pages 541–552. The Royal Society, 1951.

[81] Konstantin N Kudin, Gustavo E Scuseria, and Eric Cances. A black-box self-consistent field convergence algorithm: One step closer. The Journal of chemical physics, 116(19):8255–8261, 2002.

[82] Xiangqian Hu and Weitao Yang. Accelerating self-consistent field convergence with the augmented roothaan–hall energy function. The Journal of chemical physics, 132(5):054109, 2010.

[83] Alejandro J Garza and Gustavo E Scuseria. Comparison of self-consistent field convergence acceleration techniques. The Journal of chemical physics, 137(5):054110, 2012.

[84] Eric Cancès and Claude Le Bris. Can we outperform the diis approach for electronic structure calculations? International Journal of Quantum Chemistry, 79(2):82–90, 2000.

[85] Konstantin N Kudin and Gustavo E Scuseria. Converging self-consistent field equations in quantum chemistry–recent achievements and remaining challenges. ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique, 41(2):281–296, 2007.

[86] DR Hartree. The calculation of atomic structures. Rep. Progr. Phys, 11:113, 1948.

[87] Michael C Zerner and Michael Hehenberger. A dynamical damping scheme for converging molecular scf calculations. Chemical Physics Letters, 62(3):550–554, 1979.

[88] Peter Pulay. Convergence acceleration of iterative sequences. the case of scf iteration. Chemical Physics Letters, 73(2):393–398, 1980.

[89] Peter Pulay. Improved scf convergence acceleration. Journal of Computational Chemistry, 3(4):556–560, 1982.

[90] Tracy P Hamilton and Peter Pulay. Direct inversion in the iterative subspace (diis) optimization of open-shell, excited-state, and small multiconfiguration scf wave functions. The Journal of chemical physics, 84(10):5728–5734, 1986.

[91] Jean-Paul Berrut and Lloyd N Trefethen. Barycentric lagrange interpolation. Siam Review, 46(3):501–517, 2004.

[92] Milton Abramowitz and Irene A Stegun. Handbook of mathematical functions: with formulas, graphs, and mathematical tables, volume 55. Courier Corporation, 1964.

[93] Erik Meijering. A chronology of interpolation: from ancient astronomy to modern signal and image processing. Proceedings of the IEEE, 90(3):319–342, 2002.

[94] Alfio Quarteroni and Fausto Saleri. Scientific computing with matlab, volume 2 of texts in computational science and engineering, 2003.

[95] Andrew D Daniels, John M Millam, Gustavo E Scuseria, et al. Semiempirical methods with conjugate-gradient density-matrix search to replace diagonalization for molecular-systems containing thousands of atoms. Journal of Chemical Physics, 107(2):425–431, 1997.

[96] John M Millam and Gustavo E Scuseria. Linear scaling conjugate gradient density matrix search as an alternative to diagonalization for first principles electronic structure calculations. The Journal of chemical physics, 106(13):5569–5577, 1997.

[97] Andrew D Daniels and Gustavo E Scuseria. What is the best alternative to diagonalization of the hamiltonian in large scale semiempirical calculations? The Journal of chemical physics, 110(3):1321–1328, 1999.

[98] DR Bowler and MJ Gillan. Density matrices in o (n) electronic structure calculations: theory and applications. Computer physics communications, 120(2):95–108, 1999.

[99] Matt Challacombe. A simplified density matrix minimization for linear scaling self-consistent field theory. The Journal of chemical physics, 110(5):2332–2342, 1999.

[100] Trygve Helgaker, Helena Larsen, Jeppe Olsen, and Poul Jørgensen. Direct optimization of the ao density matrix in hartree–fock and kohn–sham theories. Chemical Physics Letters, 327(5):397–403, 2000.

[101] Helena Larsen, Jeppe Olsen, Poul Jørgensen, and Trygve Helgaker. Direct optimization of the atomic-orbital density matrix using the conjugate-gradient method with a multilevel preconditioner. The Journal of Chemical Physics, 115(21):9685–9697, 2001.

[102] Stefan Goedecker and L Colombo. Efficient linear scaling algorithm for tight-binding molecular dynamics. Physical review letters, 73(1):122, 1994.

[103] Roi Baer and Martin Head-Gordon. Chebyshev expansion methods for electronic structure calculations on large molecular systems. The Journal of chemical physics, 107(23):10003–10013, 1997.

[104] Adam HR Palser and David E Manolopoulos. Canonical purification of the density matrix in electronic-structure theory. Physical Review B, 58(19):12704, 1998.

[105] Kevin R Bates, Andrew D Daniels, and Gustavo E Scuseria. Comparison of conjugate gradient density matrix search and chebyshev expansion methods for avoiding diagonalization in large-scale electronic structure calculations. The Journal of chemical physics, 109(9):3308–3312, 1998.

[106] WanZhen Liang, Chandra Saravanan, Yihan Shao, Roi Baer, Alexis T Bell, and Martin Head-Gordon. Improved fermi operator expansion methods for fast electronic structure calculations. The Journal of chemical physics, 119(8):4117–4125, 2003.

[107] Ramiro Pino and Gustavo E Scuseria. Purification of the first-order density matrix using steepest descent and newton–raphson methods. Chemical physics letters, 360(1):117–122, 2002.

[108] David A Mazziotti. Towards idempotent reduced density matrices via particle-hole duality: Mcweeny's purification and beyond. Physical Review E, 68(6):066701, 2003.

[109] Károly Németh and Gustavo E Scuseria. Linear scaling density matrix search based on sign matrices. The Journal of Chemical Physics, 113(15):6035–6041, 2000.

[110] Anders MN Niklasson. Expansion algorithm for the density matrix. Physical Review B, 66(15):155115, 2002.

[111] Anders MN Niklasson. Implicit purification for temperature-dependent density matrices. Physical Review B, 68(23):233104, 2003.

[112] Anders MN Niklasson, CJ Tymczak, and Matt Challacombe. Trace resetting density matrix purification in o (n) self-consistent-field theory. The Journal of chemical physics, 118(19):8611–8620, 2003.

[113] Dora Kohalmi, Agnes Szabados, and Peter R Surjan. Idempotency-conserving iteration scheme for the one-electron density matrix. Physical review letters, 95(1):013002–013002, 2005.

[114] Paweł Sałek, Stinne Høst, Lea Thøgersen, Poul Jørgensen, Pekka Manninen, Jeppe Olsen, Branislav Jansík, Simen Reine, Filip Pawłowski, Erik Tellgren, et al. Linear-scaling implementation of molecular electronic self-consistent field theory. The Journal of chemical physics, 126(11):114110, 2007.

[115] Chun Hui Xu and Gustavo E Scuseria. An o (n) tight-binding study of carbon clusters up to c 8640: the geometrical shape of the giant icosahedral fullerenes. Chemical physics letters, 262(3):219–226, 1996.

[116] C.M. Goringe. D. Phil Thesis. PhD thesis, University of Oxford, 1995.

[117] Jean Charles Gilbert and Jorge Nocedal. Global convergence properties of conjugate gradient methods for optimization. SIAM Journal on optimization, 2(1):21–42, 1992.

[118] Anders MN Niklasson. Density matrix methods in linear scaling electronic structure theory. In Linear-Scaling Techniques in Computational Chemistry and Physics, pages 439–473. Springer, 2011.

[119] Paul AM Dirac. On the theory of quantum mechanics. In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, volume 112, pages 661–677. The Royal Society, 1926.

[120] Stefan Goedecker and M Teter. Tight-binding electronic-structure calculations and tight-binding molecular dynamics with localized orbitals. Physical Review B, 51(15):9455, 1995.

[121] WanZhen Liang, Roi Baer, Chandra Saravanan, Yihan Shao, Alexis T Bell, and Martin Head-Gordon. Fast methods for resumming matrix polynomials and chebyshev matrix polynomials. Journal of Computational Physics, 194(2):575–587, 2004.

[122] Emanuel H Rubensson and Hans Jørgen Aa Jensen. Determination of the chemical potential and homo/lumo orbitals in density purification methods. Chemical physics letters, 432(4):591–594, 2006.

[123] Emanuel H Rubensson and Sara Zahedi. Computation of interior eigenvalues in electronic structure calculations facilitated by density matrix purification. The Journal of chemical physics, 128(17):176101, 2008.

[124] Emanuel H Rubensson and Anders MN Niklasson. Interior eigenvalues from density matrix expansions in quantum mechanical molecular dynamics. SIAM Journal on Scientific Computing, 36(2):B147–B170, 2014.

[125] Murray S Daw. Model for energetics of solids based on the density matrix. Physical Review B, 47(16):10895, 1993.

[126] Ernst Julius Berg. Heaviside's Operational Calculus as Applied to Engineering and Physics: as applied to engineering and phsyics. McGraw-Hill book company, inc., 1936.

[127] S. Geršgorin. Proc. USSR Acad. Sci., 51(6):749–754, 1931.

[128] James Hardy Wilkinson, James Hardy Wilkinson, and James Hardy Wilkinson. The algebraic eigenvalue problem, volume 87. Clarendon Press Oxford, 1965.

[129] MJ Cawkwell, EJ Sanville, SM Mniszewski, and Anders MN Niklasson. Computing the density matrix in electronic structure theory on graphics processing units. Journal of chemical theory and computation, 8(11):4094–4101, 2012.

[130] MJ Cawkwell, MA Wood, Anders MN Niklasson, and SM Mniszewski. Computation of the density matrix in electronic structure theory in parallel on multiple graphics processing units. Journal of Chemical Theory and Computation, 10(12):5391–5396, 2014.

[131] Eugene S Kryachko. Generalized idempotency purification transform in linear scaling self-consistent field theory. Chemical Physics Letters, 318(1):210–213, 2000.

[132] A Holas. Transforms for idempotency purification of density matrices in linear-scaling electronic-structure calculations. Chemical physics letters, 340(5):552–558, 2001.

[133] Paul E Maslen, Christian Ochsenfeld, Christopher A White, Michael S Lee, and Martin Head-Gordon. Locality and sparsity of ab initio one-particle density matrices and localized orbitals. The Journal of Physical Chemistry A, 102(12):2215–2222, 1998.

[134] Yihan Shao, Chandra Saravanan, Martin Head-Gordon, and Christopher A White. Curvy steps for density matrix-based energy minimization: Application to large-scale self-consistent-field calculations. The Journal of chemical physics, 118(14):6144–6151, 2003.

[135] E.H. Rubesson. Matrix Algebra for Quantum Chemistry. PhD thesis, Royal Institute of Technology in Stockholm, 2008.

[136] Emanuel H Rubensson, Elias Rudberg, and Paweł Sałek. A hierarchic sparse matrix data structure for large-scale hartree-fock/kohn-sham calculations. Journal of computational chemistry, 28(16):2531–2537, 2007.

[137] Emanuel H Rubensson, Elias Rudberg, and Paweł Sałek. Sparse matrix algebra for quantum modeling of large systems. In Applied Parallel Computing. State of the Art in Scientific Computing, pages 90–99. Springer, 2006.

[138] Daniel Langr, Ivan Šimeček, Pavel Tvrdík, Tomáš Dytrych, and Jerry P Draayer. Adaptive-blocking hierarchical storage format for sparse matrices. In Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on, pages 545–551. IEEE, 2012.

[139] Wolfgang Hackbusch. A sparse matrix arithmetic based on\ cal h-matrices. part i: Introduction to {\ Cal H}-matrices. Computing, 62(2):89–108, 1999.

[140] Youcef Saad. Sparskit: A basic tool kit for sparse matrix computations. 1990.

[141] AH Castro Neto, F Guinea, Nuno MR Peres, Kostya S Novoselov, and Andre K Geim. The electronic properties of graphene. Reviews of modern physics, 81(1):109, 2009.

[142] Andre Konstantin Geim. Graphene: status and prospects. science, 324(5934):1530–1534, 2009.

[143] Riichiro Saito, Gene Dresselhaus, Mildred S Dresselhaus, et al. Physical properties of carbon nanotubes, volume 35. World Scientific, 1998.

[144] Stephanie Reich, Christian Thomsen, and Janina Maultzsch. Carbon nanotubes: basic concepts and physical properties. John Wiley & Sons, 2008.

[145] Richard E Smalley, Mildred S Dresselhaus, Gene Dresselhaus, and Phaedon Avouris. Carbon nanotubes: synthesis, structure, properties, and applications, volume 80. Springer Science & Business Media, 2003.

[146] JT Frey and DJ Doren. Tubegen 3.3 university of delaware newark de. 2005. Available via web interface http://turin. nss. udel. edu/research/tubegenonline. html. Accessed, 16, 2010.

[147] Yuki Matsuda, Jamil Tahir-Kheli, and William A Goddard III. Definitive band gaps for single-wall carbon nanotubes. The Journal of Physical Chemistry Letters, 1(19):2946–2950, 2010.

[148] E Hernández, MJ Gillan, and CM Goringe. Linear-scaling density-functional-theory technique: the density-matrix approach. Physical Review B, 53(11):7147, 1996.

[149] T Otsuka, T Miyazaki, T Ohno, DR Bowler, and MJ Gillan. Accuracy of order-n density-functional theory calculations on dna systems using conquest. Journal of Physics: Condensed Matter, 20(29):294201, 2008.

[150] Elias Rudberg and Emanuel H Rubensson. Assessment of density matrix methods for linear scaling electronic structure calculations. Journal of Physics: Condensed Matter, 23(7):075502, 2011.

[151] Daniel K Jordan and David A Mazziotti. Comparison of two genres for linear scaling in density functional theory: Purification and density matrix minimization methods. The Journal of chemical physics, 122(8):084114, 2005.

[152] A T_ Amos and GG Hall. Self-consistent perturbation theory for conjugated molecules. Theoretica chimica acta, 5(2):148–158, 1966.

[153] Christian Ochsenfeld, Jörg Kussmann, and Felix Koziol. Ab initio nmr spectra for molecular systems with a thousand and more atoms: A linear-scaling method. Angewandte Chemie International Edition, 43(34):4485–4489, 2004.

[154] Valéry Weber and Claude Daul. Improved coupled perturbed hartree–fock and kohn–sham convergence acceleration. Chemical physics letters, 370(1):99–105, 2003.

[155] Melvin Schwartz. Principles of electrodynamics, 1987.

[156] SP Karna and M Dupuis. Frequency dependent nonlinear optical properties of molecules: formulation and implementation in the hondo program. Journal of computational chemistry, 12(4):487–504, 1991.

[157] Endong Wang, Qing Zhang, Bo Shen, Guangyong Zhang, Xiaowei Lu, Qing Wu, and Yajuan Wang. Intel math kernel library. In High-Performance Computing on the Intel® Xeon Phi™, pages 167–188. Springer, 2014.

[158] Tatyana A Klimenko, Vladimir V Ivanov, and Ludwik Adamowicz. Dipole polarizabilities and hyperpolarizabilities of the small conjugated systems in the π-electron coupled cluster theory. Molecular Physics, 107(17):1729–1737, 2009.

[159] AD McLean and M Yoshimine. Computed ground-state properties of fh and ch. The Journal of Chemical Physics, 47(9):3256–3262, 1967.

[160] GRJ Williams. Finite field calculations of molecular polarizability and hyperpolarizabilities for organic π-electron systems. Journal of Molecular Structure: THEOCHEM, 151:215–222, 1987.

[161] Edward Anderson, Zhaojun Bai, Christian Bischof, Susan Blackford, Jack Dongarra, Jeremy Du Croz, Anne Greenbaum, Sven Hammarling, A McKenney, and D Sorensen. LAPACK Users' guide, volume 9. Siam, 1999.

[162] Kendall E Atkinson. An introduction to numerical analysis. John Wiley & Sons, 2008.

[163] Uri M Ascher and Linda R Petzold. Computer methods for ordinary differential equations and differential-algebraic equations, volume 61. Siam, 1998.

[164] Taras I Lakoba. Simple euler method and its modifications. Lecture notes for MATH334, University of Vermont, 2012.

[165] Gordon D Smith. Numerical solution of partial differential equations: finite difference methods. Oxford university press, 1985.

[166] John C Strikwerda. Finite difference schemes and partial differential equations. Siam, 2004.

[167] Randall J LeVeque. Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems, volume 98. Siam, 2007.

[168] David Eberly. Derivative approximation by finite differences. Magic Software, Inc, 2008.

[169] Yuh-Hy Lu and Shyi-Long Lee. Semi-empirical calculations of the nonlinear optical properties of polycyclic aromatic compounds. Chemical physics, 179(3):431–444, 1994.

[170] Anton B Zakharov, Vladimir V Ivanov, and Ludwik Adamowicz. Molecular nonlinear optical parameters of π-conjugated nonalternant hydrocarbons obtained in semiempirical local coupled-cluster theory. The Journal of Physical Chemistry C, 118(15):8111–8121, 2014.

[171] Qingxu Li, Liping Chen, Qikai Li, and Zhigang Shuai. Electron correlation effects on the nonlinear optical properties of conjugated polyenes. Chemical Physics Letters, 457(1):276–278, 2008.

[172] Fiona Sim, Steven Chin, Michel Dupuis, and Julia E Rice. Electron correlation effects in hyperpolarizabilities of p-nitroaniline. The Journal of Physical Chemistry, 97(6):1158–1163, 1993.

[173] Henry A Kurtz, James JP Stewart, and Kenneth M Dieter. Calculation of the nonlinear optical properties of molecules. Journal of Computational Chemistry, 11(1):82–87, 1990.

Appendix A

Derivative direct inversion of iterative subspace

The Algorithm {1} outlines how we have implemented the DIIS/D-DIIS extrapolation for the calculations we have performed in this thesis work. In a SCF procedure, we start the DIIS/D-DIIS procedure only after ten iterations around, so that the error vector norm is about the thousandth of the initial error vector norm (step 10). First, this allows to get the solution closer to the convergence region as the norm of the error vector falls gradually. And secondly, this allows to use a sufficient number m of $c_i^{(k)}$ optimization coefficients in order to get the averaged effective Fock matrix in the convergence domain. The chosen number m of $c_i^{(k)}$ coefficients has to be reasonable (not too small, not too large) as indicated at step 11. The resolution of $c_i^{(k)}$ coefficients at step 12 can be performed using a standard linear-equation solver such as the DGESV function from LAPACK library[161]. The step 14 requires the method to be used to compute the density matrix. As a result, one can compute one density matrix (unperturbed order), and even several density matrices at the same time (perturbed propjection). In the case of the perturbed projection, each order is defined by a density matrix, a Fock matrix, and the error vector requiring the lower orders density and Fock matrices.

1: ! `Initialization`
2: $D_0^{(k)}$, $\| e_0^{(k)} \|$, $n = 0$
3: ! `Iterations`
4: **while** $\| D_{n+1} - D_n \| >$ tolerance **do**
5: $n = n + 1$
6: Build the density matrix $D_n^{(k)}$ and Fock matrix $F_n^{(k)}$.
7: Compute the error matrix $e_n^{(k)}$ using Eq. (3.69).
8: Store $F_n^{(k)}$ and $e_n^{(k)}$.
9: After ten iterations around, start the DIIS/D-DIIS extrapolations:
10: **if** $\| e_n^{(k)} \| \lesssim 10^{-3} \| e_0^{(k)} \|$
11: Keep m (6 to 8) latest error matrix $e_n^{(k)}$ to assemble $B^{(k)}$ using Eq. (3.75).
12: Resolve the $c_i^{(k)}$ coefficients from Eq. (3.72) as

$$\begin{pmatrix} B^{(k)}_{n-mn-m} & \cdots & B^{(k)}_{n-mi} & \cdots & B^{(k)}_{n-mn} & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & 1 \\ B^{(k)}_{in-m} & \cdots & B^{(k)}_{ii} & \cdots & B^{(k)}_{in} & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & 1 \\ B^{(k)}_{nn-m} & \cdots & B^{(k)}_{ni} & \cdots & B^{(k)}_{nn} & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} c^{(k)}_{n-m} \\ \vdots \\ c^{(k)}_{i} \\ \vdots \\ c^{(k)}_{n} \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

13: Assemble the average effective Fock matrix $\tilde{F}_n^{(k)}$ with Eq. (3.67).
14: Compute the new density matrix $D_n^{(k)}$ with $F_n^{(k)}$ or $\tilde{F}_n^{(k)}$.
15: **end while**
16: ! `Result: Converged density matrix`
17: $D_\infty^{(k)} = D_{n+1}^{(k)}$

Algorithm 1 Pseudo-code using the DIIS or D-DIIS extrapolations.

Appendix B

LNV minimizations and conjugate gradient routine by Jorge Nocedal

The algorithms in this appendix are for the three versions of LNV minimization. Actually, these algorithms present the key steps of the line search performed by the CGFAM routine.[117] The CGFAM routine is described below. After the line search, the density matrix is purified under the threshold parameter. In our calculations, threshold $= 10^{-2}$.

1: ! `Data:` F, D, μ, tolerance,threshold.
2: ! `Initialization`
3: $\tilde{F} = F - \mu I$
4: $D = D_0$
5: $G_0 = H_0 = -\nabla\Omega_{\rm LNV}(D_0)$
6: $n = 0$
7: ! `Line search by CGFAM routine`
8: **while** $\| \tilde{D}_{n+1} - \tilde{D}_n \| >$ tolerance **do**
9: $n = n + 1$
10: $\tilde{N}_{\rm e} = {\rm Tr}\{D_n\}$
11: $\Omega_{\rm LNV}(D_n) = {\rm Tr}\{\tilde{F}(3D_n^2 - 2D_n^3)\}$
12: $\nabla\Omega_{\rm LNV}(D_n) = 3(D_n\tilde{F} + \tilde{F}D_n) - 2(D_n^2\tilde{F} + D_n\tilde{F}D_n + \tilde{F}D_n^2)$
13: $b_n = -{\rm Tr}\{H_nG_n\}$
14: $c_n = {\rm Tr}\{3H_n^2\tilde{F} - 2\left(H_n^2D_n\tilde{F} + H_nD_nH_n\tilde{F} + D_nH_n^2\tilde{F}\right)\}$
15: $d_n = -2{\rm Tr}\{H_n^3\tilde{F}\}$
16: Minimal root of $(b_n + 2c_n\lambda_n + 3d_n\lambda_n^2) = 0$
17: $D_{n+1} = D_n + \lambda_nH_n$
18: $G_{n+1} = -\nabla\Omega_{\rm LNV}(D_{n+1})$
19: $\gamma_n = \begin{cases} \frac{G_{n+1}G_n}{G_nG_n} : \text{(FR)} \\ \text{or} \\ \frac{(G_{n+1}-G_n)G_{n+1}}{G_nG_n} : \text{(PR)} \end{cases}$
20: $H_{n+1} = G_{n+1} + \gamma_nH_n$
21: $\delta N = |\tilde{N}_{\rm e} - {\rm Tr}\{D_n\}|$
22: ! `Slight purification by McWeeny`
23: **if** $\delta N >$ threshold **then**
24: $\tilde{D}_{n+1} = 3D_{n+1}^2 - 2D_{n+1}^3$
25: **end if**
26: **end while**
27: ! `Result: Converged density matrix`
28: $D_\infty = \tilde{D}_{n+1}$

Algorithm 2 LNV minimization at constant μ (μ-LNVm)

1: ! `Data:` F, D, N, tolerance,threshold.
2: ! `Initialization`
3: $10^{-3} < \alpha_{XS} < 10^{-2}$
4: $D = D_0$
5: $G_0 = H_0 = -\nabla\Omega_{\mathrm{XS}}(D_0)$
6: $n = 0$
7: ! `Line search by CGFAM routine`
8: **while** $\| \tilde{D}_{n+1} - \tilde{D}_n \| >$ tolerance **do**
9: $n = n + 1$
10: $\tilde{N}_{\mathrm{e}} = \mathrm{Tr}\{D_n\}$
11: $\mu_{n+1} = \mu_n + \alpha_{XS}\left(N - \tilde{N}_{\mathrm{e}}\right)$
12: $\tilde{F}_n = F_n - \mu_n I$
13: $\Omega_{\mathrm{XS}}(D_n) = \mathrm{Tr}\{\tilde{F}_n(3D_n^2 - 2D_n^3)\}$
14: $\nabla\Omega_{\mathrm{XS}}(D_n) = 3(D_n\tilde{F}_n + \tilde{F}_n D_n) - 2(D_n^2\tilde{F}_n + D_n\tilde{F}_n D_n + \tilde{F}_n D_n^2)$
15: $b_n = -\mathrm{Tr}\{H_n G_n\}$
16: $c_n = \mathrm{Tr}\{3H_n^2\tilde{F}_n - 2\left(H_n^2 D_n\tilde{F}_n + H_n D_n H_n\tilde{F}_n + D_n H_n^2\tilde{F}_n\right)\}$
17: $d_n = -2\mathrm{Tr}\{H_n^3\tilde{F}\}$
18: Minimal root of $(b_n + 2c_n\lambda_n + 3d_n\lambda_n^2) = 0$
19: $D_{n+1} = D_n + \lambda_n H_n$
20: $G_{n+1} = -\nabla\Omega_{\mathrm{XS}}(D_{n+1})$
21: $\gamma_n = \begin{cases} \frac{G_{n+1}G_n}{G_n G_n} : \text{(FR)} \\ \text{or} \\ \frac{(G_{n+1}-G_n)G_{n+1}}{G_n G_n} : \text{(PR)} \end{cases}$
22: $H_{n+1} = G_{n+1} + \gamma_n H_n$
23: $\delta N = |\tilde{N}_{\mathrm{e}} - \mathrm{Tr}\{D_n\}|$
24: ! `Slight purification by McWeeny`
25: **if** $\delta N >$ threshold **then**
26: $\tilde{D}_{n+1} = 3D_{n+1}^2 - 2D_{n+1}^3$
27: **end if**
28: **end while**
29: ! `Result: Converged density matrix`
30: $D_\infty = \tilde{D}_{n+1}$

Algorithm 3 Xu-Scuseria's modified LNV minimization (XS-LNVm)

1: ! `Data:` F, D, N, tolerance, threshold.
2: ! `Initialization`
3: $D = D_0$
4: $G_0 = H_0 = -\nabla\Omega_{\rm MS}(D_0)$
5: $n = 0$
6: ! `Line search by CGFAM routine`
7: **while** $\| \tilde{D}_{n+1} - \tilde{D}_n \| >$ tolerance **do**
8: $n = n + 1$
9: $\tilde{N}_{\rm e} = {\rm Tr}\{D_n\}$
10: $\mu_n = {\rm Tr}\{2D_nFD_n - F(3D_n - 2D_n^2) - (3D_n - 2D_n^2)F\}/M$
11: $\Omega_{\rm MS}(D_n) = {\rm Tr}\{F(3D_n^2 - 2D_n^3)\} + \mu_n\,({\rm Tr}\{D_n\} - N)$
12: $\nabla\Omega_{\rm MS}(D_n) = 3(DF_n + FD_n) - 2(D_n^2F + D_nFD_n + FD_n^2) + \mu_n I$
13: $b_n = -{\rm Tr}\{H_nG_n\}$
14: $c_n = {\rm Tr}\{3H_n^2\tilde{F}_n - 2\left(H_n^2D_n\tilde{F}_n + H_nD_nH_n\tilde{F}_n + D_nH_n^2\tilde{F}_n\right)\}$
15: $d_n = -2{\rm Tr}\{H_n^3\tilde{F}\}$
16: Minimal root of $(b_n + 2c_n\lambda_n + 3d_n\lambda_n^2) = 0$
17: $D_{n+1} = D_n + \lambda_nH_n$
18: $G_{n+1} = -\nabla\Omega_{\rm MS}(D_{n+1})$
19: $\gamma_n = \begin{cases} \frac{G_{n+1}G_n}{G_nG_n} : \text{(FR)} \\ \text{or} \\ \frac{(G_{n+1}-G_n)G_{n+1}}{G_nG_n} : \text{(PR)} \end{cases}$
20: $H_{n+1} = G_{n+1} + \gamma_nH_n$
21: $\delta N = |\tilde{N}_{\rm e} - {\rm Tr}\{D_n\}|$
22: ! `Slight purification by McWeeny`
23: **if** $\delta N >$ threshold **then**
24: $\tilde{D}_{n+1} = 3D_{n+1}^2 - 2D_{n+1}^3$
25: **end if**
26: **end while**
27: ! `Result: Converged density matrix`
28: $D_\infty = \tilde{D}_{n+1}$

Algorithm 4 Millam-Scuseria's modified LNV minimization (XS-LNVm)

In order to minimize functions, the present works uses the routine CGFAM, written by Jorge Nocedal.[117] This routine briefly described below is included in the CG+ code. CG+ is a conjugate gradient code for solving large scale, unconstrained, nonlinear optimization problems. CG+ implements three different versions of the conjugate gradient method: the Fletcher-Reeves method, the Polak-Ribiere method, and the positive Polak-Ribiere method (β always non-negative). A web-based server which solves unconstrained nonlinear optimization problems using this CG code can be found at: http://users.iems.northwestern.edu/~nocedal/CG+.html

subroutine CGFAM(N, X, F, G, D, GOLD, IPRINT,
EPS, W, IFLAG, IREST, METHOD, FINISH)

Subroutine parameters:

integer: N, IPRINT(2), IFLAG, IREST, METHOD
double precision: X(N), G(N), D(N), GOLD(N), W(N), F, EPS
logical: FINISH

N (input) = Number of variables
X (output) = Iterate
F (input) = Function value
G (input) = Gradient value
GOLD (input) = Previous gradient value
IPRINT (input) = Frequency and type of printing
IPRINT(1) $<$ 0 : No output is generated
IPRINT(1) = 0 : Output only at first and last iteration
IPRINT(1) $>$ 0 : Output every iprint(1) iterations
IPRINT(2) : Specifies the type of output generated;
the larger the value (between 0 and 3),
the more information
IPRINT(2) = 0 : No additional information printed
IPRINT(2) = 1 : Initial x and gradient vectors printed
IPRINT(2) = 2 : X vector printed every iteration
IPRINT(2) = 3 : X vector and gradient vector printed
every iteration
EPS (input) = Convergence constant
W (input) = Working array

IFLAG (output) = Controls termination of code, and return to main
program to evaluate function and gradient
IFLAG = -3 : Improper input parameters
IFLAG = -2 : Descent was not obtained
IFLAG = -1 : Line search failure
IFLAG = 0 : Initial entry or
successful termination without error
IFLAG = 1 : Indicates a re-entry with new function values
IFLAG = 2 : Indicates a re-entry with a new iterate

IREST (input) = 0 (no restarts); 1 (restart every N steps)

METHOD (input) = 1 : Fletcher-Reeves
2 : Polak-ribiere
3 : Positive Polak-Ribiere (β=max$\{\beta, 0\}$)

FINISH (input) = Termination test
First initialized to .false., then must be set to .true.
when the termination test is satisfied.

Appendix C

Purification algorithms

1: ! `Data:` F, M, N, μ, m, tolerance.
2: ! `Initialization`
3: $\epsilon_{\min}, \epsilon_{\max} \leftarrow F$
4: $\alpha = \min\left\{\beta_{\mathrm{m}}\left[\epsilon_{\max} - \mu\right]^{-1}, (1-\beta_{\mathrm{m}})\left[\mu - \epsilon_{\min}\right]^{-1}\right\}$
5: $\bar{\mu} = (\mu - \epsilon_{\min})(\epsilon_{\max} - \epsilon_{\min})^{-1}$
6: $D_0 = \alpha(\mu I - F) + \beta_{\mathrm{m}} I$
7: $n = 0$
8: ! `Density matrix purification`
9: **while** $\| D_{n+1} - D_n \| >$ tolerance **do**
10: $n = n + 1$
11: **if** $\bar{\mu} \geq 0.5$ **then**
12: $D_{n+1} = I - (I - D_n)^{\mathrm{m}}(I + \mathrm{m}D_n)$
13: **else**
14: $D_{n+1} = D_n^{\mathrm{m}}\left[I + \mathrm{m}(I - D_n)\right]$
15: **end if**
16: **end while**
17: ! `Result: Converged density matrix`
18: $D_\infty = D_{n+1}$

Algorithm 5 Generalized Grand canonical purification (GCp)

! `Data:` F, M, N, tolerance.
! `Initialization`
$\epsilon_{\min}, \epsilon_{\max} \leftarrow F$
$\bar{\mu} = \mathrm{Tr}\{F\}/M$
$\alpha = \min\left\{\frac{N}{\epsilon_{\max}-\bar{\mu}}, \frac{M-N}{\bar{\mu}-\epsilon_{\min}}\right\}$
$D_0 = \alpha(\bar{\mu}I - F) + (N/M)I$
$n = 0$
! `Density matrix purification`
while $\| D_{n+1} - D_n \| >$ tolerance **do**
    $n = n + 1$
    if $\| \mathrm{Tr}\{D_n - D_n^2\} \| < 10^{-4}$ **then**
        $c_n = 0.5$
    else
        $c_n = \mathrm{Tr}\{D_n^2 - D_n^3\}/\mathrm{Tr}\{D_n - D_n^2\}$
    end if
    if $c_n \geq 0.5$ **then**
        $D_{n+1} = \left[(1 + c_n)D_n^2 - D_n^3\right]/c_n$
    else
        $D_{n+1} = \left[(1 - 2c_n)D_n + (1 + c_n)D_n^2 - D_n^3\right]/(1 - c_n)$
    end if
end while
! `Result: Converged density matrix`
$D_\infty = D_{n+1}$

Algorithm 6 Canonical purification (Cp)

! `Data:` F, N, tolerance, m(> 2).
! `Initialization`
$\epsilon_{\min}, \epsilon_{\max} \leftarrow F$
$D_0 = (1 - 2\beta_{\mathrm{m}})(\epsilon_{\max}I - F)/(\epsilon_{\max} - \epsilon_{\min}) + \beta_{\mathrm{m}}I$
$n = 0$
! `Density matrix purification`
while $\| D_{n+1} - D_n \| >$ tolerance **do**
    $n = n + 1$
    if $\mathrm{Tr}\{D_n\} < N$ **then**
        $D_{n+1} = I - (I - D_n)^{\mathrm{m}}(I + \mathrm{m}D_n)$
    else
        $D_{n+1} = D_n^{\mathrm{m}}\left[I + \mathrm{m}(I - D_n)\right]$
    end if
end while
! `Result: Converged density matrix`
$D_\infty = D_{n+1}$

Algorithm 7 Generalized Trace Correcting purification (TCp)

1: ! `Data:` F, N, tolerance.
2: ! `Initialization`
3: $\epsilon_{\min}, \epsilon_{\max} \leftarrow F$
4: $\gamma_{\min} = 0, \gamma_{\max} = 6$
5: $D_0 = (\epsilon_{\max} I - F)/(\epsilon_{\max} - \epsilon_{\min})$
6: $n = 0$
7: ! `Density matrix purification`
8: **while** $\| D_{n+1} - D_n \| >$ tolerance **do**
9: $n = n + 1$
10: $\mathcal{F}(D_n) = D_n^2(4D_n - 3D_n^2)$
11: $\mathcal{G}(D_n) = D_n^2(1 - D_n)^2$
12: **if** $\| \mathrm{Tr}\{\mathcal{G}(D_n)\} \| < 10^{-4}$ **then**
13: $\gamma_n = 3.0$
14: **else**
15: $\gamma_n = (N - \mathrm{Tr}\{\mathcal{F}(D_n)\}) / \mathrm{Tr}\{\mathcal{G}(D_n)\}$
16: **end if**
17: **if** $\gamma_n > \gamma_{\max}$ **then**
18: $D_{n+1} = 2D_n - D_n^2$
19: **else if** $\gamma_n < \gamma_{\min}$ **then**
20: $D_{n+1} = D_n^2$
21: **else**
22: $D_{n+1} = \mathcal{F}(D_n) + \gamma_n \mathcal{G}(D_n)$
23: **end if**
24: **end while**
25: ! `Result: Converged density matrix`
26: $D_\infty = D_{n+1}$

Algorithm 8 Trace Resetting purification (TRSp)

! `Data:` F, M, N, tolerance.
! `Initialization`
$\epsilon_{\min}, \epsilon_{\max} \leftarrow F$
$\bar{\mu} = \mathrm{Tr}\{F\}/M$, $\theta = N/M$, $\lambda_1 = \frac{N}{M(\epsilon_{\max} - \bar{\mu})}$, $\lambda_2 = \frac{M-N}{M(\bar{\mu} - \epsilon_{\min})}$, $0 < \alpha < 1$
$\lambda_o = \min\{\lambda_1, \lambda_2\}$, $\lambda_q = \max\{\lambda_1, \lambda_2\}$
$D_{\min} = \lambda_o(\bar{\mu}I - F) + \theta I$, $D_{\max} = \lambda_q(\bar{\mu}I - F) + \theta I$
$D_0 = \alpha D_{\min} + (1 - \alpha)D_{\max}$
$n = 0$
! `Density matrix purification`
while $\| D_{n+1} - D_n \| >$ tolerance **do**
    $n = n + 1$
    if $\| \mathrm{Tr}\{D_n - D_n^2\} \| < 10^{-4}$ **then**
        $c_n = 0.5$
    else
        $c_n = \mathrm{Tr}\{D_n^2 - D_n^3\}/\mathrm{Tr}\{D_n - D_n^2\}$
    end if
    $D_{n+1} = (1 - 2c_n)D_n + 2(1 + c_n)D_n^2 - 2D_n^3$
end while
! `Result: Converged density matrix`
$D_\infty = D_{n+1}$

Algorithm 9 Hole-particle canonical purification (HPCP)

! `Data:` F, N, tolerance, R_{c}.
! `Initialization`
$\epsilon_{\min}, \epsilon_{\max} \leftarrow F$
$D_0 = (\epsilon_{\max}I - F)/(\epsilon_{\max} - \epsilon_{\min})$
$n = 0$
! `Density matrix purification`
while $\| \tilde{D}_{n+1} - \tilde{D}_n \| >$ tolerance **do**
    $n = n + 1$
    $\tilde{D}_n = \mathrm{FILTER}(\ D_n\ ,\ R_{\mathrm{c}}\ )$
    if $\mathrm{Tr}\{\tilde{D}_n\} \geq N$ **then**
        $\tilde{D}_{n+1} = \tilde{D}_n^2$
    else
        $\tilde{D}_{n+1} = 2\tilde{D}_n - \tilde{D}_n^2$
    end if
end while
! `Result: Converged density matrix`
$D_\infty = \tilde{D}_{n+1}$

Algorithm 10 Trace correcting purification (TC2) using radial truncation

Appendix D

D-CPSCF equation solver and routine by Michael Saunders

Generalized D-CPSCF equations at k order:

$$\left[F,\left[D,D^{(k)}\right]\right]+2DF^{(k)}D-\left\{D,F^{(k)}\right\}=\left[D,\sum_{i=1}^{k-1}\left[D^{(k-i)},F^{(i)}\right]\right]$$

In order to determine $D^{(k)}$, we use a conjugate gradient solving $Ax=b$ where

$$Ax:=\left[F,\left[D,D^{(k)}\right]\right]+2DF^{(k)}D-\left\{D,F^{(k)}\right\}\quad\text{(LHS)}$$

$$b:=\left[D,\sum_{i=1}^{k-1}\left[D^{(k-i)},F^{(i)}\right]\right]\quad\text{(RHS)}$$

Ax and b are vectors while the terms of D-CPSCF equations are matrices. However, we can suppose the terms of D-CPSCF equations are reshaped in vectors. The resolution seems more technical. b is known since it involves the density matrices from lower orders. x is $D^{(k)}$. A does not need neither to be known nor explicitly extracted in some way. On the contrary, we straight need the matrix-vector product Ax which is LHS. Algorithm {11} outlines how we implement this resolution. We implement a routine (APROD) which constructs LHS. The APROD routine has only one variable, $D^{(k)}$. D and F are like parameters since they are already calculated at 0th order (unperturbed order). SYMMLQ is the routine which performs the conjugate gradient where $D^{(k)}$ is the only variable which changes during the iterations. Of course, there are other parameters required in SYMMLQ routine in order to control the convergence. The feature of SYMMLQ routine is to solve $Ax=b$ without explicitly requiring the matrix A. SYMMLQ only needs an

1: Compute D and F at 0th order
2: Assemble the density matrices from lower orders to construct b.
3: ! `Call APROD routine which constructs LHS`
4: **function** APROD($D^{(k)}$,D,F)
5: $F^{(k)} = F^{(k)}\left[D^{(k)}\right]$
6: **return** $\left[F,\left[D,D^{(k)}\right]\right] + 2DF^{(k)}D - \left\{D,F^{(k)}\right\}$
7: **end function**
8: Initialize $D^{(k)}$
9: ! `Call SYMMLQ, the Saunders routine which resolves` $Ax = b$
10: **function** SYMMLQ($D^{(k)}$,APROD,b,...)
11: b: input which is the RHS.
12: APROD: external routine required by SYMMLQ and which supplies the matrix-vector product Ax, so the LHS.
13: $D^{(k)}$: output
14: **end function**
15: ! `Result: Converged density matrix`
16: $D^{(k)}$

Algorithm 11 Resolution of D-CPSCF by $Ax = b$ solver

external routine which straight supplies the matrix-vector product Ax. SYMMLQ is the routine written by Michael Saunders. This routine can be found at:
http://web.stanford.edu/group/SOL/software/symmlq/
SYMMLQ is designed to solve the system of linear equations

$$Ax = b$$

where A is an N by N symmetric matrix and b is a given vector. The matrix A is not required to be positive definite. (If A is known to be definite, the method of conjugate gradients might be preferred, since it will require about the same number of iterations as SYMMLQ but slightly less work per iteration.) The matrix A is intended to be large and sparse. It is accessed by means of a subroutine call of the form

$$\textbf{call APROD}(\mathrm{N}, x, y)$$

which must return the product $y = Ax$ for any given vector x. More generally, SYMMLQ is designed to solve the system

$$(A - \mathrm{SHIFT}\ I_N)x = b$$

where SHIFT is a specified scalar value. If SHIFT and b are suitably chosen, the computed vector x may approximate an (unnormalized) eigenvector of A, as in the methods of inverse iteration and/or Rayleigh-quotient iteration. Again, the matrix (A - SHIFT I_N) need not be positive definite. The work per iteration is very slightly less if SHIFT = 0.

A further option is that of preconditioning, which may reduce the number of iterations required. If $M = CC^t$ is a positive definite matrix that is known to approximate (A - SHIFT I_N) in some sense, and if systems of the form $My = x$ can be solved efficiently, the parameters PRECON and MSOLVE may be used (see below). When PRECON = .true., SYMMLQ will implicitly solve the system of equations

$$P(A - \text{SHIFT } I_N)P^t\bar{x} = Pb,$$

i.e.

$$\bar{A}\bar{x} = \bar{b}$$

where

$$\begin{aligned} P &= C^{-1}, \\ \bar{A} &= P(A - \text{SHIFT } I_N)P^t, \\ \bar{b} &= Pb, \end{aligned}$$

and return the solution

$$x = P^t\bar{x}.$$

The associated residual is

$$\begin{aligned} \bar{r} &= \bar{b} - \bar{A}\bar{x} \\ &= P(b - (A - \text{SHIFT } I_N)x) \\ &= P\ r. \end{aligned}$$

EPS refers to the machine precision computed by SYMMLQ.

subroutine SYMMLQ(N, B, R1, R2, V, W, X, Y, APROD, MSOLVE,
CHECKA, GOODB, PRECON, SHIFT, NOUT, ITNLIM,
RTOL, ISTOP, ITN, ANORM, ACOND, RNORM, YNORM
)

Subroutine parameters:

external: APROD, MSOLVE
integer: N, NOUT, ITNLIM, ISTOP, ITN
logical: CHECKA, GOODB, PRECON

double precision: SHIFT, RTOL, ANORM, ACOND, RNORM, YNORM,
B(N), R1(N), R2(N), V(N), W(N), X(N), Y(N)

N (input) = The dimension of the matrix A
B(N) (input) = The right hand side vector b
R1(N) (input) = Workspace
R2(N) (input) = Workspace
V(N) (input) = Workspace
W(N) (input) = Workspace
X(N) (output) = Returns the computed solution x
Y(N) (input) = Workspace
APROD (input) = The external subroutine defining the matrix A
For a given vector x, the statement
call APROD (N, x, y)
must return the product $y = A\ x$
without altering the vector x
MSOLVE (input) = The optional external subroutine defining a
preconditioning matrix M, which should
approximate (A - SHIFT I_N) in some sense.
M must be positive definite.
For a given vector x, the statement
call MSOLVE (N, x, y)
must solve the linear system $M\ y = x$
without altering the vector x.
In general, M should be chosen so that $\bar{A}$ has
clustered eigenvalues. For example,
if A is positive definite, $\bar{A}$ would ideally
be close to a multiple of I_N.
If A or (A - SHIFT I_N) is indefinite, $\bar{A}$ might
be close to a multiple of I_N.
NOTE: The program calling SYMMLQ must declare
APROD and MSOLVE to be external.
CHECKA (input) = If CHECKA = .true., an extra call of APROD will
be used to check if A is symmetric. Also,
if PRECON = .true., an extra call of MSOLVE

will be used to check if M is symmetric.

GOODB (input) = Usually, GOODB should be .false.
If x is expected to contain a large multiple of
b (as in Rayleigh-quotient iteration),
better precision may result if GOODB = .true.
When GOODB = .true., an extra call to MSOLVE
is required.

PRECON (input) = If PRECON = .true., preconditioning will
be invoked. Otherwise, subroutine MSOLVE
will not be referenced; in this case the
actual parameter corresponding to MSOLVE may
be the same as that corresponding to APROD.

SHIFT (input) = Should be zero if the system $A\ x = b$ is to be
solved. Otherwise, it could be an
approximation to an eigenvalue of A, such as
the Rayleigh quotient $b^t Ab/(b^t b)$
corresponding to the vector b.
If b is sufficiently like an eigenvector
corresponding to an eigenvalue near shift,
then the computed x may have very large
components. When normalized, x may be
closer to an eigenvector than b.

NOUT (input) = A file number.
If NOUT > 0, a summary of the iterations
will be printed on unit NOUT.

ITNLIM (input) = An upper limit on the number of iterations.

RTOL (input) = A user-specified tolerance. SYMMLQ terminates
if it appears that $\|\bar{r}\|$ is smaller than
RTOL $\|\bar{A}\|\ \|\bar{x}\|$,
where $\bar{r}$ is the transformed residual vector,
$\bar{r} = \bar{b} - \bar{A}\ \bar{x}$.
If SHIFT = 0 and PRECON = .false., SYMMLQ
terminates if $\|b - Ax\|$ is smaller than
RTOL $\|A\|\ \|x\|$.

ISTOP (output) = An integer giving the reason for termination...

-1: $\beta = 0$ in the Lanczos iteration; i.e. the second Lanczos vector is zero. This means the RHS is very special.
If there is no preconditioner, b is an eigenvector of A.
Otherwise (if PRECON is true), let $My = b$.
If SHIFT is zero, y is a solution of the generalized eigenvalue problem $Ay = \lambda My$, with $\lambda = \alpha$ from the Lanczos vectors.
In general, (A - SHIFT I_N)$x = b$ has the solution $x = (1/\alpha)y$ where $My = b$.

0: $b = 0$, so the exact solution is $x = 0$.
No iterations were performed.

1: $\|\bar{r}\|$ appears to be less than the value RTOL $\|\bar{A}\|\ \|\bar{x}\|$.
The solution in x should be acceptable.

2: $\|\bar{r}\|$ appears to be less than the value EPS $\|\bar{A}\|\ \|\bar{x}\|$.
This means that the residual is as small as seems reasonable on this machine.

3: $\|\bar{A}\|\ \|\bar{x}\|$ exceeds $\|b\|$/EPS, which should indicate that x has essentially converged to an eigenvector of A corresponding to the eigenvalue shift.

4: ACOND (see below) has exceeded 0.1/EPS, so the matrix $\bar{A}$ must be very ill-conditioned. x may not contain an acceptable solution.

5: The iteration limit was reached before any of the previous criteria were satisfied.

6: The matrix defined by APROD does not appear to be symmetric.
For certain vectors $y = A\ v$ and $r = A\ y$, the

products $y^t y$ and $r^t v$ differ significantly.

7: The matrix defined by MSOLVE does not appear to be symmetric.
For vectors satisfying $M\ y = v$ and $M\ r = y$, the products $y^t y$ and $r^t v$ differ significantly.

8: An inner product of the form $x^t M^{-1} x$ was not positive, so the preconditioning matrix M does not appear to be positive definite.
If ISTOP ≥ 5, the final x may not be an acceptable solution.

ITN (output) = The number of iterations performed.

ANORM (output) = An estimate of the norm of the matrix operator $\bar{A} = P\ (A$ - SHIFT $I_N)\ P^t$, where $P = C^{-1}$.

ACOND (output) = An estimate of the condition of $\bar{A}$ above.
This will usually be a substantial under-estimate of the true condition.

RNORM (output) = An estimate of the norm of the final transformed residual vector,
$P\ (b$ - $(A$ - SHIFT $I_N)\ x)$.

YNORM (output) = An estimate of the norm of $\bar{x}$
This is $\sqrt{x^t M x}$. If PRECON is false,
$P\ (b$ - $(A$ - SHIFT $I_N)\ x)$.
YNORM is an estimate of $\|x\|$.

Appendix E

The principle of finite differences

Finite differences of univariate functions

Let be a scalar function f of unidimensional variable x. The derivative of f, denoted here by $f^{(1)}$, is commonly defined by

$$f^{(1)}(x) = \lim_{\xi \to 0} \frac{f(x+\xi) - f(x)}{\xi} \tag{E.1}$$

Since $\lim_{\xi \to 0}$ can not be computed, a discrete analogue is used instead,

$$f^{(1)}(x) = \frac{f(x+\xi) - f(x)}{\xi} + \mathcal{O}(\xi) \tag{E.2}$$

where ξ (> 0) is a finite small step on the discret set of points x. The relation (E.2) is known as the forward Euler difference (FED) approximation[162–164] since it uses forward differencing. There exists also the backward Euler difference (BED) approximation:

$$f^{(1)}(x) = \frac{f(x) - f(x-\xi)}{\xi} + \mathcal{O}(\xi) \tag{E.3}$$

and the centered Euler difference (CED) approximation:

$$fk1(x) = \frac{f(x+\xi) - f(x-\xi)}{2\xi} + \mathcal{O}(\xi^2) \tag{E.4}$$

The difference between these three approximations is given by their intrinsic error. Let us remind that a Taylor series is a representation of function, infinitely differentiable, by an infinite sum of terms, which are calculated from the values of the function derivatives

at a single point. For example,

$$f(x+\xi) = f(x) + \xi f^{(1)}(x) + \frac{\xi^2}{2!}f^{(2)} + ... = \sum_{k=0}^{\infty} \frac{\xi^k}{k!} f^{(k)}(x) \tag{E.5a}$$

$$f(x-\xi) = f(x) - \xi f^{(1)}(x) + \frac{\xi^2}{2!}f^{(2)} + ... = \sum_{k=0}^{\infty} (-1)^k \frac{\xi^k}{k!} f^{(k)}(x) \tag{E.5b}$$

where $f^{(k)}$ denotes the kth order derivative of f. By Rearranging for instance Eq. (E.5a), such that

$$\frac{f(x+\xi) - f(x)}{\xi} - f^{(1)}(x) = \underbrace{\frac{\xi}{2!}f^{(2)}(x) + \frac{\xi^2}{3!}f^{(3)}(x) + ...}_{\text{Truncation Error}} \tag{E.6}$$

one can observe that the FED in Eq. (E.2) corresponds to a Taylor series truncated after the second term. The rhs of Eq. (E.6) is the error in terminating the series and is referred to as the truncation error (TE).[165–167] The TE can be defined as the difference between the partial derivative and its finite difference representation.

The finite difference representation described above evaluates the function to be derived in a 2 points approximation.[158]. In order to maximize the numerical accuracy of a finite difference representation, one can evaluate the function with a higher number of points. That involves a higher order Taylor series. For example, let us first set

$$f(x+2\xi) = \sum_{k=0}^{\infty} \frac{(2\xi)^k}{k!} f^{(k)}(x) \tag{E.7a}$$

$$f(x-2\xi) = \sum_{k=0}^{\infty} (-1)^k \frac{(2\xi)^k}{k!} f^{(k)}(x) \tag{E.7b}$$

then proceeding as

$$f^{(1)}(x) = \frac{1}{2\xi}\left[-4\text{Eq. (E.5a)} + \text{Eq. (E.7a)}\right] \tag{E.8a}$$

$$f^{(1)}(x) = \frac{1}{2\xi}\left[4\text{Eq. (E.5b)} - \text{Eq. (E.7b)}\right] \tag{E.8b}$$

$$f^{(1)}(x) = \frac{1}{12\xi}\left[8\text{Eq. (E.5a)} - \text{Eq. (E.7a)} - 8\text{Eq. (E.5b)} + \text{Eq. (E.7b)}\right] \tag{E.8c}$$

leads to a FED with a second order TE,

$$f^{(1)}(x) = \frac{-f(x+2\xi) + 4f(x+\xi) - 3f(x)}{2\xi} + \mathcal{O}(\xi^2) \tag{E.9}$$

to a BED with a second order TE,

$$f^{(1)}(x) = \frac{3f(x) - 4f(x-\xi) + f(x-2\xi)}{2\xi} + \mathcal{O}(\xi^2) \tag{E.10}$$

and to a CED with a fourth order TE,

$$f^{(1)}(x) = \frac{-f(x+2\xi) + 8f(x+\xi) - 8f(x-\xi) + f(x-2\xi)}{12\xi} + \mathcal{O}(\xi^4) \tag{E.11}$$

In Eq. (E.9) and Eq. (E.10), the function is now evaluated in 3 points, while in Eq. (E.11) the function is now evaluated in 4 points. The relation (E.8) means that the Taylor series evaluated in many different points, are combined in some appropriate way so that one can generalize a relation between the d^{th} order derivative function and its finite difference representation by[168]

$$\frac{\xi^d}{d!} f^{(d)}(x) = \sum_{m=m_{\mathrm{mn}}}^{m_{\mathrm{mx}}} c_m f(x + m\xi) + \mathcal{O}(\xi^{d+p}) \tag{E.12}$$

where p (> 0) is the integer order of the TE, selected as desired. c_m are the coefficients of the finite difference represensation. In order to determine c_m, one uses the Taylor series for $f(x + m\xi)$, which is

$$f(x + m\xi) = \sum_{k=0}^{\infty} m^k \frac{\xi^k}{k!} f^{(k)}(x) \tag{E.13}$$

Introducing Eq. (E.13) into Eq. (E.12) yields

$$\begin{aligned} \frac{\xi^d}{d!} f^{(d)}(x) &= \sum_{m=m_{\mathrm{mn}}}^{m_{\mathrm{mx}}} c_m \sum_{k=0}^{\infty} m^k \frac{\xi^k}{k!} f^{(k)}(x) + \mathcal{O}(\xi^{d+p}) \\ &= \sum_{k=0}^{\infty} \left(\sum_{m=m_{\mathrm{mn}}}^{m_{\mathrm{mx}}} m^k c_m \right) \frac{\xi^k}{k!} f^{(k)}(x) + \mathcal{O}(\xi^{d+p}) \\ &= \sum_{k=0}^{d+p-1} \left(\sum_{m=m_{\mathrm{mn}}}^{m_{\mathrm{mx}}} m^k c_m \right) \frac{\xi^k}{k!} f^{(k)}(x) + \mathcal{O}(\xi^{d+p}) \end{aligned} \tag{E.14}$$

and that finally leads to

$$f^{(d)}(x) = \frac{d!}{\xi^d} \sum_{k=0}^{d+p-1} \left(\sum_{m=m_{\mathrm{mn}}}^{m_{\mathrm{mx}}} m^k c_m \right) \frac{\xi^k}{k!} f^{(k)}(x) + \mathcal{O}(\xi^p) \tag{E.15}$$

From the last equation, the simplest way to determine the c_m coefficients is to constrained c_m to have the property of the Lagrange polynomials[91–94] such that

$$\sum_{m=m_{\mathrm{mn}}}^{m_{\mathrm{mx}}} m^k c_m = \left\{ \begin{array}{ll} 0, & 0 \leq k \leq d+p-1 \text{ and } k \neq d \\ 1, & k = d \end{array} \right\} \tag{E.16}$$

in order to have a unique solution for c_m. The relation (E.16) corresponds a set of $(d+p)$ linear equations in $(m_{\mathrm{mx}} - m_{\mathrm{mn}} + 1)$ unkowns.

	FED	BED	CED
(d,p)	$m_{\mathrm{mn}} = 0$ $m_{\mathrm{mx}} = d+p-1$	$m_{\mathrm{mn}} = -(d+p-1)$ $m_{\mathrm{mx}} = 0$	$m_{\mathrm{mn}} = -(d+p-1)/2$ $m_{\mathrm{mx}} = (d+p-1)/2$

Table E.1 Indices $(m_{\mathrm{mn}},m_{\mathrm{mx}})$ for c_m given by (d,p) corresponding to the type of finite representation difference approximation.

The Table {E.1} gives the number of coefficients c_m, so the different points m in the finite difference representation from the derivative order d and the TE order p, for the differente approximation. In other words, Table {E.1} gives a relationship between the number of terms in the finite difference representation and the number of terms in the Taylor series. For this reason, in Eq. (E.15), the sum over k or the number of terms in the Taylor series is no more infinite. Note that in Table {E.1}, for CED approximation, $d+p$ is necessarily an odd number, while p can be chosen to be even regardless of the parity of d.

In order to understand Eq. (E.16) which gives the coefficients c_m of the finite difference representation, let us approximate for example $f^{(3)}(x)$ with a FED and a 1st order TE ie. $\mathcal{O}(\xi)$, so $d = 3$ and $p = 1$. Using Table {E.1} gives $m_{\mathrm{mn}} = 0$ and $m_{\mathrm{mx}} = 3$. The linear system given by Eq. (E.16) is then

$$\begin{bmatrix} (0)^0 & (1)^0 & (2)^0 & (3)^0 \\ (0)^1 & (1)^1 & (2)^1 & (3)^1 \\ (0)^2 & (1)^2 & (2)^2 & (3)^2 \\ (0)^3 & (1)^3 & (2)^3 & (3)^3 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{E.17}$$

Assuming that $0^0 = 1$, we have

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 1 & 4 & 9 \\ 0 & 1 & 8 & 27 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{E.18}$$

This equation is easily resolved by hand and its solution is $(c_0, c_1, c_2, c_3) = (-1, 3, -3, 1)/6$. In other words, using Eq. (E.12),

$$f^{(3)}(x) = \frac{-f(x) + 3f(x+\xi) - 3f(x+2\xi) + f(x+3\xi)}{\xi^3} + O(\xi) \tag{E.19}$$

Let us now approximate $f^{(3)}(x)$ with a CED and error $O(\xi^2)$. Proceeding in the same way, $d = 3$ and $p = 2$, gives $m_{\text{mx}} = -m_{\text{mn}} = 2$. The resulting linear system is

$$\begin{bmatrix} (-2)^0 & (-1)^0 & (0)^0 & (1)^0 & (2)^0 \\ (-2)^1 & (-1)^1 & (0)^1 & (1)^1 & (2)^1 \\ (-2)^2 & (-1)^2 & (0)^2 & (1)^2 & (2)^2 \\ (-2)^3 & (-1)^3 & (0)^3 & (1)^3 & (2)^3 \\ (-2)^4 & (-1)^4 & (0)^4 & (1)^4 & (2)^4 \end{bmatrix} \begin{bmatrix} c_{-2} \\ c_{-1} \\ c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \tag{E.20}$$

which corresponds to

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ 4 & 1 & 0 & 1 & 4 \\ -8 & -1 & 0 & 1 & 8 \\ 16 & 1 & 0 & 1 & 16 \end{bmatrix} \begin{bmatrix} c_{-2} \\ c_{-1} \\ c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \tag{E.21}$$

and has solution $(c_{-2}, c_{-1}, c_0, c_1, c_2) = (-1, 2, 0, -2, 1)/12$. Finally, the expression for $f^{(3)}(x)$ is

$$f^{(3)}(x) = \frac{-f(x-2\xi) + 2f(x-\xi) - 2f(x+\xi) + f(x+2\xi)}{2\xi^3} + O(\xi^2) \tag{E.22}$$

Finite differences of bivariate functions

The CED approximation is enough used for finite field difference methods applied to the response tensors[158, 169–173]. For example, we have implemented the finite field differ-

ence of Ref. [158] which uses a CED with 7 points. That involves $m = \{0, \pm 1, \pm 2, \pm 3\}$, hence $m_{\mathrm{mx}} = -m_{\mathrm{mn}} = 3$. We can now deduce the approximate expressions for the derivatives of the response tensors given in Eq. (4.15) such that

Dipole moment μ

$d = 1$, $p = 6$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & -1 & 0 & 1 & 2 & 3 \\ 9 & 4 & 1 & 0 & 1 & 4 & 9 \\ -27 & -8 & -1 & 0 & 1 & 8 & 27 \\ 81 & 16 & 1 & 0 & 1 & 16 & 81 \\ -243 & -32 & -1 & 0 & 1 & 32 & 243 \\ 729 & 64 & 1 & 0 & 1 & 64 & 729 \end{bmatrix} \begin{bmatrix} c_{-3} \\ c_{-2} \\ c_{-1} \\ c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{E.23}$$

which approximates the first order derivative of the energy with respect to the electric field as

$$\begin{aligned} \mu = \frac{\partial \mathcal{E}(\vec{\mathscr{E}})}{\partial \mathscr{E}} \approx \frac{1}{60\xi} (&\mathcal{E}_{+3} - 9\mathcal{E}_{+2} + 45\mathcal{E}_{+1} \\ &-45\mathcal{E}_{-1} + 9\mathcal{E}_{-2} - \mathcal{E}_{-3}) + O(\xi^6) \end{aligned} \tag{E.24}$$

Polarizability α

$d = 2$, $p = 5$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & -1 & 0 & 1 & 2 & 3 \\ 9 & 4 & 1 & 0 & 1 & 4 & 9 \\ -27 & -8 & -1 & 0 & 1 & 8 & 27 \\ 81 & 16 & 1 & 0 & 1 & 16 & 81 \\ -243 & -32 & -1 & 0 & 1 & 32 & 243 \\ 729 & 64 & 1 & 0 & 1 & 64 & 729 \end{bmatrix} \begin{bmatrix} c_{-3} \\ c_{-2} \\ c_{-1} \\ c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{E.25}$$

which approximates the second order derivative of the energy with respect to the electric field as

$$\begin{aligned} \alpha = \frac{\partial^2 \mathcal{E}(\vec{\mathscr{E}})}{\partial \mathscr{E}^2} \approx \frac{1}{180\xi^2} (&2\mathcal{E}_{+3} - 27\mathcal{E}_{+2} + 270\mathcal{E}_{+1} \\ &-490\mathcal{E}_0 + 270\mathcal{E}_{-1} - 27\mathcal{E}_{-2} + 2\mathcal{E}_{-3}) + O(\xi^5) \end{aligned} \tag{E.26}$$

First hyperpolarizability β

$d = 3$, $p = 4$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & -1 & 0 & 1 & 2 & 3 \\ 9 & 4 & 1 & 0 & 1 & 4 & 9 \\ -27 & -8 & -1 & 0 & 1 & 8 & 27 \\ 81 & 16 & 1 & 0 & 1 & 16 & 81 \\ -243 & -32 & -1 & 0 & 1 & 32 & 243 \\ 729 & 64 & 1 & 0 & 1 & 64 & 729 \end{bmatrix} \begin{bmatrix} c_{-3} \\ c_{-2} \\ c_{-1} \\ c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{E.27}$$

which approximates the third order derivative of the energy with respect to the electric field as

$$\begin{aligned} \beta = \frac{\partial^3 \mathcal{E}(\vec{\mathscr{E}})}{\partial \mathscr{E}^3} \approx \frac{1}{8\xi^3} (&-\mathcal{E}_{+3} + 8\mathcal{E}_{+2} - 13\mathcal{E}_{+1} \\ &+13\mathcal{E}_{-1} - 8\mathcal{E}_{-2} + \mathcal{E}_{-3}) + O(\xi^4) \end{aligned} \tag{E.28}$$

Second hyperpolarizability γ

$d = 4$, $p = 3$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & -1 & 0 & 1 & 2 & 3 \\ 9 & 4 & 1 & 0 & 1 & 4 & 9 \\ -27 & -8 & -1 & 0 & 1 & 8 & 27 \\ 81 & 16 & 1 & 0 & 1 & 16 & 81 \\ -243 & -32 & -1 & 0 & 1 & 32 & 243 \\ 729 & 64 & 1 & 0 & 1 & 64 & 729 \end{bmatrix} \begin{bmatrix} c_{-3} \\ c_{-2} \\ c_{-1} \\ c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \tag{E.29}$$

which approximates the fourth order derivative of the energy with respect to the electric field as

$$\begin{aligned} \gamma = \frac{\partial^4 \mathcal{E}(\vec{\mathscr{E}})}{\partial \mathscr{E}^4} \approx \frac{1}{6\xi^4} (&-\mathcal{E}_{+3} + 12\mathcal{E}_{+2} - 39\mathcal{E}_{+1} \\ &+56\mathcal{E}_0 - 39\mathcal{E}_{-1} + 12\mathcal{E}_{-2} - \mathcal{E}_{-3}) + O(\xi^3) \end{aligned} \tag{E.30}$$

In Eq. (E.24), Eq. (E.26), Eq. (E.28) and Eq. (E.30), ξ is the differentation step of the finite field difference representation for the energy derivative with respect to the electric field $\vec{\mathscr{E}}$. The unit of ξ is that of the electric field strength, $\xi \sim 10^{-3}$ au[158]. And $\mathcal{E}_{\pm i}$ is

$$\mathcal{E}_{\pm i} = \mathcal{E}(\vec{\mathscr{E}} \pm i\xi) \tag{E.31}$$

the energy of the system calculated for the strength of the electric field equal to $\vec{\mathscr{E}} \pm i\xi$ with $i = \{0, 1, 2, 3\}$. On the other hand, it is important to emphasize that the finite field difference representation of Eq. (E.24), Eq. (E.26), Eq. (E.28) and Eq. (E.30) is for an univariate function. That implies the electric field changes only in one direction, ie. $\mathcal{E}_{\pm i} = \mathcal{E}(\mathscr{E}_x \pm i\xi)$ in the x direction. In the case where the electric field changes in two directions x and y[158], the finite difference representation basically requires two differentiation steps ξ_x and ξ_y, respectively. Supposing that ξ_x and ξ_y can be comparable, then we can find a step ξ so that: $\xi_x = i\xi$ and $\xi_y = j\xi$. We may write the energy as

$$\mathcal{E}_{\pm i,\pm j} = \mathcal{E}(\mathscr{E}_x \pm i\xi, \mathscr{E}_y \pm j\xi) \tag{E.32}$$

This energy is associated to Eq. (4.3) which the expression is

$$h^{\lambda}_{\mu\nu} = h_{\mu\nu} + i\langle x\rangle\xi\delta_{\mu\nu} + j\langle y\rangle\xi\delta_{\mu\nu} \tag{E.33}$$

where $(i, j) = \{0, \pm1, \pm2, \pm3\}$. and $\langle x\rangle$ and $\langle y\rangle$ represent the position vector components along the x and y directions. We obtain an energy matrix such as

$$\mathcal{E}_{i,j} = \begin{pmatrix} \mathcal{E}_{-3,-3} & \mathcal{E}_{-2,-3} & \mathcal{E}_{-1,-3} & \mathcal{E}_{0,-3} & \mathcal{E}_{+1,-3} & \mathcal{E}_{+2,-3} & \mathcal{E}_{+3,-3} \\ \mathcal{E}_{-3,-2} & \mathcal{E}_{-2,-2} & \mathcal{E}_{-1,-2} & \mathcal{E}_{0,-2} & \mathcal{E}_{+1,-2} & \mathcal{E}_{+2,-2} & \mathcal{E}_{+3,-2} \\ \mathcal{E}_{-3,-1} & \mathcal{E}_{-2,-1} & \mathcal{E}_{-1,-1} & \mathcal{E}_{0,-1} & \mathcal{E}_{+1,-1} & \mathcal{E}_{+2,-1} & \mathcal{E}_{+3,-1} \\ \mathcal{E}_{-3,0} & \mathcal{E}_{-2,0} & \mathcal{E}_{-1,0} & \mathcal{E}_{0,0} & \mathcal{E}_{+1,0} & \mathcal{E}_{+2,0} & \mathcal{E}_{+3,0} \\ \mathcal{E}_{-3,+1} & \mathcal{E}_{-2,+1} & \mathcal{E}_{-1,+1} & \mathcal{E}_{0,+1} & \mathcal{E}_{+1,+1} & \mathcal{E}_{+2,+1} & \mathcal{E}_{+3,+1} \\ \mathcal{E}_{-3,+2} & \mathcal{E}_{-2,+2} & \mathcal{E}_{-1,+2} & \mathcal{E}_{0,+2} & \mathcal{E}_{+1,+2} & \mathcal{E}_{+2,+2} & \mathcal{E}_{+3,+2} \\ \mathcal{E}_{-3,+3} & \mathcal{E}_{-2,+3} & \mathcal{E}_{-1,+3} & \mathcal{E}_{0,+3} & \mathcal{E}_{+1,+3} & \mathcal{E}_{+2,+3} & \mathcal{E}_{+3,+3} \end{pmatrix} \tag{E.34}$$

In this matrix, a row(column) means that the field changes in $x(y)$ direction while is fixed in $y(x)$ direction. The fourth row(column) corresponds to the univariate case where the field exits only in $x(y)$. As a result, a derivative of this energy with respect to $\mathscr{E}_x$ ($\mathscr{E}_y$) requires the seven energies along a row (column) of this matrix. The formulas in Eq. (E.24), Eq. (E.26), Eq. (E.28) and Eq. (E.30) using the energies of the 4^{th} row (column) give the diagonal components of the tensors along x (y), since the field is applied only in a single direction. For example for the component β_{yyy}, we need to take

the formula of Eq. (E.28) with the energies

$$\begin{pmatrix} \mathcal{E}_{0,-3} \\ \mathcal{E}_{0,-2} \\ \mathcal{E}_{0,-1} \\ \mathcal{E}_{0,0} \\ \mathcal{E}_{0,+1} \\ \mathcal{E}_{0,+2} \\ \mathcal{E}_{0,+3} \end{pmatrix}$$

While for the non-diagonal components of tensors, one has to apply some combinations of formulas (E.24), (E.26), (E.28) and (E.30), as the field is applied in several directions. The value of the field changes in a direction while it is fixed in the other directions. As an example, for

$$\gamma_{xxyy} = \frac{\partial^2}{\partial \mathscr{E}_x^2}\left(\frac{\partial^2 \mathcal{E}(\mathscr{E}_x, \mathscr{E}_y)}{\partial \mathscr{E}_y^2}\right) = \frac{\partial^2 X(\mathscr{E}_x)}{\partial \mathscr{E}_x^2}$$

we first use Eq. (E.26) for each column, which leads to a row vector

$$\begin{pmatrix} X_{-3} & X_{-2} & X_{-1} & X_0 & X_{+1} & X_{+2} & X_{+3} \end{pmatrix}$$

For the elements of this vector, X_{+1} for instance means the $\mathscr{E}_y\mathscr{E}_y$ − second derivative while the field at $\mathscr{E}_x$ is fixed to $i = +1$, and so on. Then, by applying once more Eq. (E.26) to this vector, we finally obtain the non-diagonal component γ_{xxyy}. The below algorithm outines the finite field difference method which evaluates the optical properties. This method is from Ref. [158] and uses a seven points representation. The energy is calculated at different points. The resulting energies are then used in combinations of sums of energies that define the derivatives for the response tensors.

1: ! `Data:` F, ξ, tolerance.
2: ! `Initialization`
3: $D = D_0$
4: $n = 0$
5: ! `Computation of seven energy points`
6: **for** $i = -3, 3$ **do**
7: **for** $j = -3, 3$ **do**
8: **while** $\| \tilde{D}_{n+1} - \tilde{D}_n \| >$ tolerance **do**
9: $n = n + 1$
10: $\tilde{F}_n = F[D_n] + \xi\left(i\langle x\rangle + j\langle y\rangle\right) I$
11: $\tilde{D}_{n+1} \leftarrow \tilde{F}_n$
12: $\tilde{\mathcal{E}}_{n+1} \leftarrow \tilde{D}_{n+1}$
13: **end while**
14: **end for**
15: **end for**
16: $\mathcal{E}(i,j) = \tilde{\mathcal{E}}_\infty$
17: $\mathcal{E}(i,j) \xrightarrow{\text{Eq.(E.24),Eq.(E.26),Eq.(E.28),Eq.(E.30)}} \mu, \alpha, \beta, \gamma$

Algorithm 12 Finite field difference method for μ, α, β and γ (optical properties).