



TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik



# Mathematical Analysis of Electronic Structure Models

BENEDIKT R. GRASWALD

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

**Doktors der Naturwissenschaften** (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Michael M. Wolf

Prüfer der Dissertation:

1. Prof. Dr. Gero Friesecke
2. Prof. Jianfeng Lu, Ph.D.
3. Prof. Dr. Reinhold Schneider

Die Dissertation wurde am 13.10.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 08.03.2022 angenommen.

– *Meinem Großvater gewidmet* –

# Acknowledgments

I would like to start this thesis off by expressing my gratitude to those people, without whom this thesis would never have been written in the first place. During the last months and years I probably missed the chance to do so several times.

First, I would like to thank my supervisor Prof. Gero Friesecke for introducing me to electronic structure models and giving me the opportunity to work on this project. I thank him for his patience and encouragement as well as the creative freedom he granted me. The fruitful discussions we had during my studies granted me plenty of mathematical and non-mathematical insights.

Furthermore, I am very thankful to Prof. Jianfeng Lu and Prof. Reinhold Schneider for refereeing my thesis, and Prof. Michael M. Wolf for chairing my defense committee. I want to gratefully acknowledge the IGDK1754 for funding and Prof. Fellner for hosting me during my research stay in Graz. Additionally, I want to thank my fellow IGDK graduate students for many helpful and interesting discussions.

Navigating the bureaucracy one encounters during a Ph.D. is not an easy task, thus I also want to say thank you to Frauke Bäcker, Diane Clayton-Winter and Silvia Schulz for their help in administrative matters, and for many pleasant chats.

During the time of my Ph.D., I benefited in so many ways from my colleagues, by scientific and personal discussions but also by sharing tasks. Thus, I want to thank all members of the analysis chair for making my experience at TUM much more entertaining. In particular, I thank my office mates Kyle and Mi-Song as well as Arseniy, Marco, Rufat, and Sören.

The work on a thesis is not only carried by the academic community, but also by the support and encouragement from friends. Therefore, I want to express my deepest gratitude to Matthias C. Caro, Lars Wüstrich, Alexander Kammerer, and my friends from the ZHS, in particular Lisa and Timm, for keeping me sane during the past few years and giving me strength to bring this project forward. Additionally, I am severely indebted to Matthias for his improvements of the present text, without him there would only be half as many commas in this thesis.

Last but most importantly, I want to thank my family, my siblings, Dorothee and Johannes, and especially my parents, Andrea and Stefan, and for their endless love, patience and support. Without their encouragement to pursue my goals and the wisdom they shared, I wouldn't be where I am today.

I'm truly blessed to have such wonderful parents. Thank you mom and dad!

Diese Arbeit ist meinem Großvater gewidmet, der ihre Vollendung stets interessiert verfolgt hat.

*Benedikt R. Graswald*

Munich, October 2021



*“ I seem to have been only like a boy playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me. ”*

————— ISAAC NEWTON

*“ The most beautiful thing we can experience is the mysterious. It is the source of all true art and science. He to whom the emotion is a stranger, who can no longer pause to wonder and stand wrapped in awe, is as good as dead. ”*

————— ALBERT EINSTEIN

*“ Here I stand, atoms with consciousness, matter with curiosity. A universe of atoms, an atom in the universe. ”*

————— RICHARD FEYNMAN



# Zusammenfassung

In der vorliegenden Dissertation werden Methoden zur Lösung des elektronischen Strukturproblems analysiert. Hierbei besteht das Hauptziel darin, ein besseres mathematisches Verständnis dieser Theorien zu erreichen.

Zunächst wird die Kohn-Sham-Dichtefunktionaltheorie (KS-DFT) erörtert, welche die am häufigsten verwendete elektronische Strukturmethode für große Systeme darstellt. Hierbei werden insbesondere die Existenz und Nichtexistenz elektronischer Anregungen sowie die Dissoziationsgrenze diatomarer Moleküle untersucht, beides in der lokalen Dichteapproximation (LDA).

Der nächste Teil befasst sich mit Tensormethoden in der Quantenchemie. Diese ermöglichen sehr genaue Berechnungen, sind aber aufgrund ihrer hohen Rechenkosten auf kleine Systeme beschränkt. Wir betrachten Matrix-Product-States (MPS), auch bekannt als Tensor-Trains (TT), die das Herzstück der Quantenchemie-Dichte-Matrix-Renormierungsgruppe (QC-DMRG) bilden. Wir untersuchen, wie sich Permutationen der zugrundeliegenden Basis auf die Größe der involvierten Matrizen der entsprechenden Darstellung für allgemeine Zustände auswirken und liefern eine vollständige Charakterisierung dieser für Zwei-Elektronen-Systeme unter optimalen unitären Basistransformationen.

Die Ergebnisse dieser Arbeit sind in verschiedenen Artikeln des Autors enthalten, von denen drei veröffentlicht wurden, einer angenommen und einer eingereicht wurde. Eine Liste der relevanten Artikel ist auf Seite ix zu finden.





# Abstract

This dissertation examines methods for solving the electronic structure problem with its primary objective being to provide a better mathematical understanding of these theories.

First, we discuss Kohn-Sham Density Functional Theory (KS-DFT), which is the most widely used electronic structure method for large systems. In particular, we study the existence and non-existence of electronic excitations as well as the dissociation limit of diatomic molecules, both in the local density approximation (LDA).

The next part deals with tensor methods in quantum chemistry. These allow very accurate computations, but are limited to small systems due to their high computational cost. We consider matrix product states (MPS), also known as tensor-trains (TT), which lie at the heart of the Quantum Chemistry – Density Matrix Renormalization Group method (QC-DMRG). We study how re-orderings of the underlying basis affect the bond-dimensions of the corresponding representation for general states and provide a complete characterization of the bond-dimensions for two-electron systems under optimal fermionic mode transformations.

The results of this thesis are contained in various articles by the author, of which three have been published, one is accepted and one is submitted. A list of the contributed articles is presented on Page ix.



# List of Contributed Articles

## *Core Articles as Principal Author*

- I) Gero Friesecke and Benedikt R. Graswald (2020).  
Existence and nonexistence of HOMO–LUMO excitations in Kohn–Sham density functional theory.  
*Nonlinear Analysis* 200, 111973.  
<https://doi.org/10.1016/j.na.2020.111973>  
(see also article [54] in the bibliography)
- II) Benedikt R. Graswald and Gero Friesecke (2021).  
Electronic wavefunction with maximally entangled MPS representation.  
*European Physical Journal D* (2021) 75: 176.  
<https://doi.org/10.1140/epjd/s10053-021-00189-2>  
(see also article [64] in the bibliography)
- III) Sören Behr and Benedikt R. Graswald (2021).  
Dissociation limit in Kohn–Sham density functional theory.  
*Nonlinear Analysis* 215, 112633.  
<https://doi.org/10.1016/j.na.2021.112633>  
(see also article [11] in the bibliography)

## *Further articles under review*

- IV) Gero Friesecke and Benedikt R. Graswald (2021).  
Two-electron wavefunctions are matrix product states with bond dimension Three  
*arXiv preprint* arXiv:2109.10091.  
Submitted to *Journal of Mathematical Physics*.  
(see also article [56] in the bibliography)

## *Articles as co-author*

- V) Matthias C. Caro and Benedikt R. Graswald (2021).  
Necessary criteria for Markovian divisibility of linear maps.  
*Journal of Mathematical Physics* 62, 042203.  
<https://doi.org/10.1063/5.0031760>  
(see also article [18] in the bibliography)

I, Benedikt R. Graswald, would like to reiterate that I am the principle author of the core articles of this dissertation, which are Articles I, II and III.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Outline	2
1.2	Summary and Discussion of Results	2
<b>2</b>	<b>The Quantum Many-Body Problem</b>	<b>5</b>
2.1	Heuristics and the Hydrogen Atom	5
2.2	General Molecules and the Born-Oppenheimer Approximation	8
2.3	HVZ Theorem and Bound States in the Born-Oppenheimer Approximation	13
2.4	Why Do Molecules Bind Together?	14
<b>3</b>	<b>Density Functional Theory and the Kohn-Sham Equations</b>	<b>19</b>
3.1	Density Functional Theory and its Predecessor	20
3.2	Modern Density Functional Formalism	25
3.2.1	Hohenberg-Kohn Theorem	25
3.2.2	Kohn-Sham Equations	28
3.2.3	Exchange-Correlation Functionals	30
3.3	Contributions in the Analysis of Density Functional Theory and Related Literature	32
3.3.1	Excitations in Density Functional Theory	32
3.3.2	Dissociation Limit in Kohn-Sham DFT	34
<b>4</b>	<b>Tensors and the Quantum Chemistry - Density Matrix Renormalization Group</b>	<b>37</b>
4.1	What Is a Tensor?	38
4.1.1	Symmetric and Antisymmetric Tensor Spaces	40
4.1.2	The Tensor-Train Decomposition	40
4.2	Where Does the Tensor Come from in Quantum Chemistry?	45
4.2.1	Fock Space and the Occupation Representation	45
4.2.2	Tensor Networks and Matrix Product States	46
4.3	Contributions in QC-DMRG and Related Literature	50
4.3.1	Maximally Entangled Matrix Product States	51
4.3.2	Bond Dimension in Two-Electron Systems	52
4.3.3	Markovian Divisibility for Quantum Channels	53
	<b>Bibliography</b>	<b>57</b>

CONTENTS

**Appendices**

<b>A Core Articles</b>	<b>71</b>
A.1 Existence and nonexistence of HOMO–LUMO excitations in Kohn–Sham density functional theory . . . . .	71
A.2 Electronic wavefunction with maximally entangled MPS representation . . . . .	96
A.3 Dissociation limit in Kohn–Sham density functional theory . . . . .	104
<b>B Further Articles under Review</b>	<b>141</b>
B.1 Two-electron wavefunctions are matrix product states with bond dimension Three	141
<b>C Articles as Co-author</b>	<b>165</b>
C.1 Necessary Criteria for Markovian Divisibility of Linear Maps . . . . .	165

# Chapter 1

## Introduction

*In view of all that [...], the many obstacles we appear to have surmounted, what casts the pall over our victory celebration?*

*It is the curse of dimensionality, a malediction that has plagued the scientist from earliest days.*

---

Richard E. Bellman

In quantum chemistry, the term *electronic structure* encompasses both the wavefunctions of the electrons and the energies associated with them. Its starting point is the full quantum many-body Schrödinger equation in the Born-Oppenheimer approximation: The electrons are treated as quantum particles in an electrostatic field created by clamped, i.e., stationary, nuclei.

Although the adequate mathematical description for this theory was developed in the 1920s by people like Heisenberg, Schrödinger and Dirac, it took a long time until quantum mechanics found its way into applications. The main problem here is that unfortunately the Schrödinger equation cannot be solved analytically, except for a small number of simple problems like the hydrogen atom or very elementary molecules or systems.

Furthermore, due to the curse of dimension – for  $N$  particles, the system is described by a partial differential equation (PDE) of dimension  $3N$  – as soon as the system grows slightly, it becomes numerically unfeasible. Solving this curse of dimensionality for the  $N$ -body electronic Schrödinger equation has been a central problem in physics and chemistry for over a century. This can also be seen from the fact that electronic structure calculations rank among the most computationally intensive tasks in all scientific calculations and a large number of methods exist, with their applicability varying from case to case.

Even though this plethora of different methods has played an important role in the endeavor of understanding quantum mechanical systems, especially in quantum chemistry, solid-state physics and materials science, mathematically rigorous result are quite sparse.

The overarching topic of this thesis is providing a better mathematical understanding of some of these theories. This is done through analyzing, whether or not certain desirable properties are fulfilled, and through completely characterizing certain classes of approximations.

The two state-of-art methods considered here will be the quantum chemistry – density matrix renormalization group (QC–DMRG) with the so-called matrix product states (MPS) at its heart, and the density functional theory (DFT), with its exchange-correlation functionals.

In our analysis of MPS, i.e., studying how different transformations of the underlying basis affect the size of the involved matrices, in particular, providing lower bounds, we employ (multi-)linear algebra combined with an old result in number theory by Besicovitch. The considered problems in DFT, like the existence of excitations and dissociation limits, on the other hand, rely on tools from the field of partial differential equations in unbounded domains, like the concentration-compactness method by Lions as well as the spectral theory of Schrödinger Hamiltonians.

## 1.1 Outline

In the rest of this chapter, we briefly discuss the contributed articles in this thesis.

In Chapter 2, we give an introduction to the basic concepts of quantum mechanics necessary for understanding the electronic structure problem, like the Schrödinger equation for general molecules, the Born-Oppenheimer approximation and some well known spectral results of the involved operators.

The main part of the thesis then focuses on the two electronic structure theories mentioned above: Density functional theory as well as its predecessors are reviewed in Chapter 3. We start with a historic overview to get a general idea of the different developments in the field still influencing the functionals used today and then present the more modern and mathematical formalism.

QC–DMRG, and more explicitly tensors in general, are the topic of Chapter 4. After we discuss their basic properties and the associated tensor-train decomposition, we move on to describe how they arise in quantum chemistry and how transforming the underlying basis can affect the involved tensors.

Both of these chapters are concluded with a short summary of our own contributions in these areas as well as related research articles.

After this overview, we include the contributed articles. Every article is preceded by a summary of the contributions of the respective work and a description of the individual contribution of the author of this thesis. Furthermore, we include for each article the permission to use it in this thesis.

## 1.2 Summary and Discussion of Results

The contributed articles deal with different aspects of electronic structure models and related objects. Core Articles I and III deal with questions arising in Kohn-Sham DFT. The first one investigates the existence and non-existence of excitations in the Kohn-Sham DFT setting; whereas the second one considers the question, whether molecules dissociate correctly in the local density approximation of DFT.

The behavior of the involved matrices of a matrix product state under basis transformations is the subject of Core Article II as well as Article IV. Lastly, Article V is concerned with the related



topic of quantum channels, more precisely, it examines which quantum channels correspond to Markovian time evolutions.

Note that the author of this thesis does not claim to be the principal author of the Articles IV and V.

*Core articles as principal author*

- *Article I [54]: Existence and nonexistence of HOMO–LUMO excitations in Kohn–Sham density functional theory*

In this work we investigate the mathematical status of the simplest class of excitations in Kohn-Sham density functional theory (KS-DFT), the HOMO-LUMO excitations. Employing concentration-compactness arguments, we show that such excitations, i.e., excited states of the Kohn-Sham Hamiltonian, exist for positively charged systems, i.e.,  $Z > N$ , where  $Z$  is the total nuclear charge and  $N$  is the number of electrons. The assumptions on the exchange-correlation functional under which the result is applicable are realistic and verified explicitly for the widely used PZ81 and PW92 functionals. By contrast, in the neutral case  $Z = N$ , we find, using a method of Glaser, Martin, Grosse, and Thirring that in cases of the hydrogen and helium atoms, excited states do not exist when the self-consistent KS ground state density is replaced by a realistic but easier to analyze approximation (in case of hydrogen, the true Schrödinger ground state density).

- *Article II [64]: Electronic wavefunction with maximally entangled MPS representation*

In this core article, we present an example of an electronic wavefunction with maximally entangled MPS representation, in the sense that the bond dimension is maximal and cannot be lowered by any re-ordering of the underlying one-body basis. Our construction works for any number of electrons and orbitals. Additionally, we provide numerically the singular value distribution of the matricization of the corresponding tensor, which seems to exhibit a remarkable almost-invariance under re-ordering. In contrast, for weakly correlated states re-ordering typically reduces the tail by several order of magnitude [37].

- *Article III [11]: Dissociation limit in Kohn–Sham density functional theory*

In the third core article, we consider the dissociation limit for diatomic molecules, i.e., molecules of the type  $X_2$ , in the Kohn-Sham density functional theory setting, where  $X$  can be any element with  $N$  electrons. Our main result is the following: When the two  $X$ -atoms in the system are torn infinitely far apart, the energy of the system converges to  $\min_{\alpha \in [0, N]} (I_\alpha^X + I_{2N-\alpha}^X)$ , where  $I_\alpha^X$  denotes the energy of a  $X$  atom with  $\alpha$  electrons surrounding it. Additionally, we discuss, whether or not the minimum equals the symmetric splitting  $2I_N^X$  for the Dirac exchange functional. The decisive factor turns out to be the “strength” of the exchange functional, which in the case of this paper is determined through the constant  $c_{xc}$  in front of the Dirac exchange. We provide numerical evidence that for the  $H_2$ -molecule with the correct physical value for  $c_{xc}$  this gives the expected result of twice the energy of a H-atom,  $2I_1^H$ .

*Further articles under review*

- *Article IV [56]: Two-electron wavefunctions are matrix product states with bond dimension Three*

The topic of this article is proving the statement in the title, i.e., precisely that two-electron wavefunctions can be represented, in a suitable basis, as MPS with bond dimension Three. Our analysis is carried out for arbitrary single-particle Hilbert spaces, including the infinite-dimensional space  $L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$  for electrons.

Furthermore, we show that bond dimension Three is optimal and characterize the minimal bond dimension for arbitrary states under optimal fermionic mode transformation. Lastly, we describe the implications of our results for the QC-DMRG method for computing the electronic structure of molecules. This yields a remarkable low-rank exactness.

*Articles as co-author*

- *Article V [18]: Necessary criteria for Markovian divisibility of linear maps*

Here, we study the open problem of characterizing those quantum channels that correspond to Markovian time evolutions. Whereas there is a complete characterization for infinitesimal Markovian divisible qubit channels, no necessary or sufficient criteria are known for higher dimensions, except for necessity of non-negativity of the determinant.

We start this article by describing how to extend the notion of infinitesimal Markovian divisibility from quantum channels to general linear maps and compact and convex sets of generators. After that we present a general approach towards proving necessary criteria for (infinitesimal) Markovian divisibility. Employing this procedure, we prove two independent criteria which are necessary for infinitesimal divisibility of quantum channels in any finite dimension  $d$ : an upper bound on the determinant in terms of a  $\Theta(d)$ -power of the smallest singular value, and in terms of a product of  $\Theta(d)$  smallest singular values. These allow us to analytically construct, in any given dimension, a set of channels that contains provably non-infinitesimal Markovian divisible ones. Moreover, we show that, in general, no such non-trivial criteria can be derived for the classical counterpart of this scenario. This implies that there cannot be a mapping from the classical stochastic matrices to quantum channels which both preserves infinitesimal Markovian divisibility and leaves singular values invariant.

## Chapter 2

# The Quantum Many-Body Problem

*Anyone who is not shocked by quantum theory has not understood it.*

---

Niels Bohr

In this chapter, we want to give a brief introduction to the mathematical framework required for the quantum many-body problem for atoms and molecules, which lies at the heart of this thesis.

We start off by analyzing the simplest such system, the hydrogen atom, to gain some heuristic understanding. After that, we introduce the full molecular Hamiltonian and discuss the Born-Oppenheimer approximation, which builds the foundation of quantum chemistry. This leads to the electronic Schrödinger operator whose fundamental properties we recall in the second part of this chapter. We conclude with the so-called dissociation problem and discuss why molecules bind together, using the hydrogen molecule  $H_2$  as an example. This material can be found in standard references like [16, 66, 69, 147, 170], while the first and the last section take inspiration from a course of Prof. Gero Friesecke offered at TUM in 2016.

### 2.1 Heuristics and the Hydrogen Atom

In order to gain a deeper insight into the framework presented later in this chapter, let us start with the simplest quantum chemical system, i.e., the hydrogen atom.

Here, we have a single proton with one electron surrounding it. Thus we can always change to a reference system with the proton as the origin. So, our system is described by a so-called wavefunction  $\Psi : \mathbb{R}^3 \rightarrow \mathbb{C}$  with  $\|\Psi\|_{L^2} = 1$ , representing the probability density that the electron is at position  $x \in \mathbb{R}^3$ .

As typical in nature, we want  $\Psi$  to minimize the total energy of the system consisting of the kinetic energy of the electron and the Coulomb interaction of the electron with the proton, i.e.,

$$\mathcal{E}_{hyd}[\Psi] = \frac{1}{2} \int_{\mathbb{R}^3} |\nabla \Psi|^2 dx - \int_{\mathbb{R}^3} \frac{1}{|x|} |\Psi|^2 dx, \quad (2.1)$$

subject to the normalization condition  $\|\Psi\|_{L^2} = 1$ . The kinetic term in (2.1) wants  $\Psi$  to be flat, while the electrostatic term favors a high peak at the origin, so we expect a compromise between these two giving  $\Psi$  a certain length-scale.

We want both terms to be finite, thus we restrict ourselves to the admissible functions  $\mathcal{A}_{hyd}$  with

$$\mathcal{A}_{hyd} := \{ \Psi \in H^1(\mathbb{R}^3; \mathbb{C}) \mid \|\Psi\|_{L^2} = 1 \}.$$

Therefore, we are faced with the following minimization problem:

$$\text{Minimize } \mathcal{E}_{hyd}[\Psi] \quad \text{over } \Psi \in \mathcal{A}_{hyd}. \quad (2.2)$$

This minimization problem can be solved straightforwardly with a suitable chosen ansatz. Inspired by the variation of constant approach from ordinary differential equations, we consider  $\Psi$  to be of the form

$$\Psi(x) = e^{-a|x|}\lambda(x).$$

Note that since  $\lambda$  is any arbitrary function such that  $\Psi \in H^1(\mathbb{R}^3, \mathbb{C})$ , we have not made any restriction so far. Plugging this ansatz into the energy (2.2) gives

$$\begin{aligned} \mathcal{E}_{hyd}[\Psi] &= \frac{1}{2} \int_{\mathbb{R}^3} |\nabla (e^{-a|x|}\lambda(x))|^2 dx - \int_{\mathbb{R}^3} \frac{1}{|x|} e^{-2a|x|} |\lambda(x)|^2 dx \\ &= -\frac{a^2}{2} \int_{\mathbb{R}^3} e^{-2a|x|} |\lambda(x)|^2 dx + \frac{1}{2} \int_{\mathbb{R}^3} e^{-2a|x|} |\nabla \lambda(x)|^2 dx + (a-1) \int_{\mathbb{R}^3} \frac{1}{|x|} e^{-2a|x|} |\lambda(x)|^2 dx. \end{aligned} \quad (2.3)$$

Here we first expanded out the square

$$|\nabla (e^{-a|x|}\lambda(x))|^2 = a^2 e^{-2a|x|} |\lambda|^2 - 2a \operatorname{Re} \left( \frac{x}{|x|} \cdot \lambda \nabla \lambda \right) e^{-2a|x|} + e^{-2a|x|} |\nabla \lambda|^2$$

and then used integration by parts on the middle term

$$\begin{aligned} -a \operatorname{Re} \int_{\mathbb{R}^3} \frac{x}{|x|} e^{-2a|x|} \cdot \nabla \lambda^2(x) dx &= a \int_{\mathbb{R}^3} \lambda^2(x) \operatorname{div} \left( \frac{x}{|x|} e^{-2a|x|} \right) dx \\ &= a \int_{\mathbb{R}^3} \lambda^2(x) \left( \frac{2}{|x|} e^{-2a|x|} - 2ae^{-2a|x|} \right) dx. \end{aligned}$$

Setting now  $a = 1$  and using that  $\Psi = e^{-a|x|}\lambda$  is normalized in  $L^2$  reduces (2.3) to

$$\mathcal{E}_{hyd}[\Psi] = -\frac{1}{2} + \frac{1}{2} \int_{\mathbb{R}^3} e^{-2|x|} |\nabla \lambda|^2 dx.$$

Therefore  $\Psi$  is a minimizer if and only if  $\lambda$  is constant. In particular, the energy functional is bounded from below. Employing now again the normalization of  $\Psi$  gives  $\lambda = \frac{1}{\sqrt{\pi}}\alpha$  with  $|\alpha| = 1$ . Thus the minimizer is unique up to a phase factor ( $\alpha \in \mathbb{C}, |\alpha| = 1$ ) and is given by

$$\Psi(x) = \alpha \frac{e^{-|x|}}{\sqrt{\pi}}. \quad (2.4)$$

If we reintroduce physical units, the energy functional  $\mathcal{E}_{hyd}$  from (2.1) becomes

$$\mathcal{E}_{hyd}[\Psi] = \frac{\hbar^2}{2m} \int_{\mathbb{R}^3} |\nabla \Psi|^2 dx - \frac{e^2}{4\pi\epsilon_0} \int_{\mathbb{R}^3} \frac{1}{|x|} |\Psi|^2 dx,$$

where  $\hbar$  is Planck's constant,  $e$  denotes the charge of the electron and  $m$  its mass, and  $4\pi\epsilon_0$  stands for the electric permittivity of the vacuum.

Thus, the minimizer takes the form

$$\Psi(x) = \alpha \frac{e^{-\frac{|x|}{a_0}}}{\sqrt{\pi a_0^3}} \quad \text{with} \quad a_0 = \frac{\hbar^2/m}{e^2/4\pi\epsilon_0} \approx 0.529 \cdot 10^{-10} m. \quad (2.5)$$

Here one can see that the length scale of the exponential decay  $a_0$ , which emerges naturally from the two prefactors of the minimization, behaves inversely proportional to the particle mass.

Furthermore, we want to look at the Euler-Lagrange equation of our problem. Consider, for any arbitrary  $\varphi \in H^1(\mathbb{R}^3; \mathbb{C})$ , the function

$$f : \mathbb{R} \rightarrow \mathbb{R}, \quad f(\varepsilon) = \mathcal{E}_{hyd} \left[ \frac{\Psi + \varepsilon\varphi}{\|\Psi + \varepsilon\varphi\|_{L^2}} \right].$$

Then,  $f$  has a global minimum at  $\varepsilon = 0$ , which implies

$$0 = \frac{d}{d\varepsilon} f(\varepsilon) \Big|_{\varepsilon=0} = 2 \operatorname{Re} \int_{\mathbb{R}^3} \left( -\frac{1}{2} \Delta \Psi - \frac{1}{|x|} \Psi - \mathcal{E}_{hyd}[\Psi] \Psi \right) \bar{\varphi} dx.$$

Since  $\varphi$  was arbitrary, we obtain

$$-\frac{1}{2} \Delta \Psi - \frac{1}{|x|} \Psi = \mathcal{E}_{hyd}[\Psi] \Psi,$$

which is in fact equivalent to

$$-\frac{1}{2} \Delta \Psi - \frac{1}{|x|} \Psi = \lambda \Psi \quad \text{for some } \lambda \in \mathbb{R}. \quad (2.6)$$

This equivalence can be seen by multiplying (2.6) by  $\bar{\Psi}$  and integrating, resulting in

$$\mathcal{E}_{hyd}[\Psi] = \frac{1}{2} \int_{\mathbb{R}^3} |\nabla \Psi|^2 - \frac{1}{|x|} |\Psi|^2 dx = \lambda \int_{\mathbb{R}^3} |\Psi|^2 dx = \lambda.$$

Therefore, we have arrived at the Schrödinger equation for the hydrogen atom, which is just an eigenvalue problem for a linear partial differential equation. We usually write (2.6) in the form

$$H_{hyd} \Psi = \lambda \Psi, \quad (2.7)$$

where  $H_{hyd}$  denotes the Hamiltonian of our system

$$H_{hyd} = -\frac{1}{2} \Delta - \frac{1}{|x|}.$$

In this section, we heuristically motivated two major points which will be made more precise in the following sections.

Firstly, as seen in (2.5), the quantum fluctuations of objects depend greatly on their mass. Thus, it makes sense to decouple nuclei and electrons in our analysis since  $m_{nuc}/m \gtrsim 1836 - 367000$  [66]. This leads us to the so-called Born-Oppenheimer approximation [13], addressed in Section 2.2.

And secondly, the eigenvalue problem (2.6) motivates us to study the entire spectrum of the corresponding Hamiltonian, not only its lowest eigenvalue. This is exactly the content of the celebrated HVZ Theorem 2.3. Lastly, let us mention that also the uniqueness and exponential decay of the wavefunction corresponding to the lowest eigenvalue of the Hamiltonian – the so-called ground state – observed in our example will carry over to the full system.

## 2.2 General Molecules and the Born-Oppenheimer Approximation

After the short heuristical introduction above, we now want to get to the heart of the matter, and precisely present the quantum description of a molecular system. Note that we will here and in this entire thesis only discuss the non-relativistic case, but when the system at hand contains multiple heavy atoms, relativistic effects may play an important role [16].

Our starting point is a static isolated molecular system consisting of  $M$  nuclei and  $N$  electrons. In our analysis, we will always treat a nucleus as a whole, specifying its substructure only by noting the number of protons and neutrons, since this influences the total charge. The number of nucleons of a nucleus also determines additional properties, like the values its spin can take and thus the symmetry of the wavefunction, but – as we will see below – this will not be of importance to us.

The state of our system is then entirely described by a complex-valued wavefunction  $\Psi$  of the form

$$\Psi(y_1, \tau_1, \dots, y_M, \tau_M; x_1, \sigma_1, \dots, x_N, \sigma_N),$$

where  $\Psi$  depends on the position  $y_k$  and spin  $\tau_k$  of the  $k$ -th nucleus and on the position  $x_i$  and spin  $\sigma_i$  of the  $i$ -th electron. For the position variables, we have  $y_k, x_i \in \mathbb{R}^3$ , while the spin of the electrons can only take two values,  $\sigma_i \in \Sigma := \{|\uparrow\rangle, |\downarrow\rangle\}$ . For a nucleus composed of  $K$  nucleons the situation is slightly more complicated: The spin variable  $\tau_k$  can take  $\frac{1}{4}(K+2)^2$  values if  $K$  is even and  $\frac{1}{4}(K+1)(K+3)$  values if  $K$  is odd. We denote this finite set of possible spin configurations by  $\Sigma_K$ .

The bridge from the wavefunction  $\Psi$  to our real physical system is given by the fact that  $|\Psi(y_1, \tau_1, \dots, y_M, \tau_M; x_1, \sigma_1, \dots, x_N, \sigma_N)|^2$  gives the probability density to simultaneously measure the  $k$ -th nucleus at position  $y_k$  with spin  $\tau_k$  and the  $i$ -th electron at position  $x_i$  with spin  $\sigma_i$ .

This already gives us one of the important properties that a wavefunction  $\Psi$  has to satisfy in order to correspond to a physical system. The function  $\Psi$  needs to be  $L^2$ -normalized, i.e.,

$$\|\Psi\|_{L^2} = \int_{\mathbb{R}^{3(M+N)}} \sum_{\substack{\tau_1, \dots, \tau_M \\ \sigma_1, \dots, \sigma_M}} |\Psi(y_1, \tau_1, \dots, y_M, \tau_M; x_1, \sigma_1, \dots, x_N, \sigma_N)|^2 dy_1 \dots dy_M dx_1 \dots dx_N = 1. \quad (2.8)$$

A second deep physics fact is the indistinguishability of identical particles: The system has to stay independent of the arbitrary labeling that we have forced on the electrons and nuclei. This leads to the definition of bosons and fermions: The function  $\Psi$  has to be

- *symmetric* under the exchange of two identical particles, which are bosons. In our framework, the only bosons are the nuclei composed of an even number of nucleons.

or

- *antisymmetric* under the exchange of two identical particles which are fermions. In our framework, these are precisely the electrons and the nuclei composed of an odd number of nucleons.

In particular, this antisymmetric behaviour in terms of the electrons takes the form

$$\Psi(\{y_k, \tau_k\}; x_{p(1)}, \sigma_{p(1)}, \dots, x_{p(N)}, \sigma_{p(N)}) = (-1)^{\varepsilon(p)} \Psi(\{y_k, \tau_k\}; x_1, \sigma_1, \dots, x_N, \sigma_N), \quad (2.9)$$

where  $p$  denotes a permutation of the set of electron indices  $\{1, \dots, N\}$  and  $\varepsilon(p)$  its signature. This antisymmetry has three major consequence; the first being the *Pauli exclusion principle* stating that two fermions can not be in exactly the same spin state and position. If, e.g., two electrons  $i \neq j$  have  $x_i = x_j$  and  $\sigma_i = \sigma_j$ , then by (2.9)

$$\Psi(\{y_k, \tau_k\}; x_1, \sigma_1, \dots, x_N, \sigma_N) = 0. \quad (2.10)$$

Secondly, the wavefunction is orthogornal to any independent state, meaning

$$\langle \Psi, \Phi \rangle_{L^2((\mathbb{R}^3 \times \Sigma)^N)} = 0,$$

for any  $\Phi \in L^2((\mathbb{R}^3 \times \Sigma)^N)$  depending in the same way on the  $i$ - and  $j$ - component, i.e.,

$$\Phi(\{x_k, \sigma_k\}) = \varphi_*(x_i, \sigma_i) \varphi_*(x_j, \sigma_j) \chi(\{x_k, \sigma_k\}_{k \neq i, j}),$$

with  $\chi$  and  $\varphi_*$  being arbitrary. And lastly, (2.9) implies that the probability density  $|\Psi|^2$  is a symmetric function.

The Hilbert space incorporating all the above restrictions is given by

$$\mathcal{H} = \mathcal{H}_n \times \mathcal{H}_e,$$

with

$$\begin{aligned}\mathcal{H}_n &= L_x^2((\mathbb{R}^3 \times \Sigma_1) \times \dots \times (\mathbb{R}^3 \times A_M); \mathbb{C}), \\ \mathcal{H}_e &= \bigwedge_{i=1}^N L^2(\mathbb{R}^3 \times \Sigma; \mathbb{C}),\end{aligned}$$

where the subscript of  $L_x^2$  with  $x \in \{a, s\}$  indicates that certain symmetry and or antisymmetry properties need to be fulfilled, depending on the structure of the nuclei in the system.

Now, we can give the Hamiltonian describing our molecular system of  $N$  non-relativistic electrons of mass  $m$  and charge  $e$  and  $M$  atomic nuclei of masses  $m_1, \dots, m_M$  with charges  $Z_1e, \dots, Z_Me$

$$H_{mol} = -\sum_{i=1}^N \frac{\hbar^2}{2m} \Delta_{x_i} - \sum_{j=1}^M \frac{\hbar^2}{2m_j} \Delta_{y_j} + V(x, y), \quad (2.11)$$

where  $\hbar$  is Planck's constant,  $x = (x_1, \dots, x_N) \in \mathbb{R}^{3N}$  and  $y = (y_1, \dots, y_M) \in \mathbb{R}^{3M}$  stand for the electron and nuclear coordinates, respectively, and  $V(x, y)$  denotes the entire Coulomb interaction potential of the system, i.e., between the electrons themselves, between the electrons and the nuclei, and between the nuclei themselves. This potential is given by

$$4\pi\epsilon_0 V(x, y) = \frac{1}{2} \sum_{i \neq j} \frac{e^2}{|x_i - x_j|} - \sum_{i,j} \frac{Z_j e}{|x_i - y_j|} + \frac{1}{2} \sum_{i \neq j} \frac{Z_i Z_j e^2}{|y_i - y_j|}, \quad (2.12)$$

with  $4\pi\epsilon_0$  being the electric permittivity of the vacuum.

In the special case  $M = 1$ , i.e., an atom, the last term in (2.12) vanishes. Furthermore, we call a molecule neutral or neutrally charged if

$$\sum_{j=1}^M Z_j = N. \quad (2.13)$$

Note, from now on we will – as common in the literature – work with atomic units, where Planck's constant  $\hbar = 1$ , the charge and mass of an electron  $e = 1$  and  $m = 1$ , and also the electric constant  $4\pi\epsilon_0 = 1$ .

Now, the minimization problem to find the ground state of our system is given by

$$E_0^{mol} := \inf\{\langle \Psi, H_{mol} \Psi \rangle : \Psi \in \mathcal{H}, \|\Psi\|_{L^2} = 1\}. \quad (2.14)$$

Since we only want to consider physically realistic states of finite energy, we need to impose finite kinetic energy, i.e., integrability in the  $L^2$  sense of the first derivative of  $\Psi$ . Thus, we further restrict our Hilbert space  $\mathcal{H}$  to be the tensor product  $\mathcal{H} = \mathcal{H}_n \otimes \mathcal{H}_e$  with

$$\mathcal{H}_n = H_{a,s}^1((\mathbb{R}^3 \times \Sigma_1) \times \dots \times (\mathbb{R}^3 \times \Sigma_M)) \quad \text{and} \quad \mathcal{H}_{el}^N = \bigwedge_{i=1}^N H^1((\mathbb{R}^3 \times \Sigma)^N),$$

where the differentiability condition is only with respect to the continuous space variables.



In the next step, we want to motivate the procedure of decoupling the electrons and the nuclei, which is generally known as the Born-Oppenheimer approximation [13]. We follow the steps from [66].

As already mentioned above, this procedure is based on the key fact that nuclei are much heavier than electrons. Thus, the two particles live on different time scales, i.e., the electrons adjust almost instantaneously to the positions of the nuclei.

Therefore, we first assume that the nuclei are clamped at positions  $y = (R_1, \dots, R_M)$  and consider the electronic Hamiltonian depending on the parameters  $(R_1, \dots, R_M)$

$$H_N^{el}((R_1, \dots, R_M)) = -\frac{1}{2} \sum_{i=1}^N \Delta_{x_i} + \frac{1}{2} \sum_{i \neq j} \frac{1}{|x_i - x_j|} - \sum_{i,j} \frac{Z_j}{|x_i - R_j|}. \quad (2.15)$$

This operator is often referred to as the Born-Oppenheimer Hamiltonian.

Next, with the electrons in their equilibrium position, corresponding to the state of lowest possible energy  $E_{el}(y)$  of  $H_N^{el}$  for a given configuration of nuclei  $y$ , one considers the motion of the nuclei. Here,  $H_N^{el}(y)$  in (2.11) is replaced by the multiplication operator  $E_{el}(y)$  as the potential interaction energy leading to the nuclear Hamiltonian

$$H_{nuc} = -\sum_{j=1}^M \frac{1}{2m_j} \Delta_{y_j} + E_{el}(y) \Big|_{y=(R_1, \dots, R_M)} + \frac{1}{2} \sum_{i \neq j} \frac{Z_i Z_j}{|R_i - R_j|}. \quad (2.16)$$

From a physics point of view, one expects the eigenvalues of  $H_{nuc}$  to be a good approximation for the ones of the full Hamiltonian  $H_{mol}$ . Computing  $E_{el}(y)$  for a given configuration of nuclei clamped at positions  $y$  is called solving the *electronic structure problem*.

In order to precisely state the Born-Oppenheimer approximation, we define the parameter

$$\kappa = \frac{1}{\min_j m_j},$$

which, depending on the system at hand, will vary from 1/1836 to 1/367000 [66].

For simplicity, assume that the ground state energy  $E_{el}(y)$  of the electronic Hamiltonian  $H_N^{el}$  in (2.15) is non-degenerate, and denote the corresponding normalized minimizer by  $\psi_y(x)$ , i.e.,  $H_N^{el}\psi_y(x) = E_{el}\psi_y(x)$ . Then with small technical modifications, one can prove the following result.

**Theorem 2.1** (Born-Oppenheimer Approximation (see Chapter 12 in [66]))

To second order in the parameter  $\kappa$ , the ground state energy  $E_0^{mol}$  of  $H_{mol}$  is the ground state energy of the operator

$$H_{eff} := H_{nuc} + v,$$

where  $H_{nuc}$  is given in (2.16) and  $v = O(\kappa)$  is a multiplication operator of order  $\kappa$  given by

$$v = \sum_{j=1}^M \frac{1}{2m_j} \int_{\mathbb{R}^3} |\nabla_{y_j} \psi_y(x)|^2 dx.$$

Thus, henceforth we will ignore the quantum fluctuations of the nuclei and consider them as point particles with charges  $Z_1, \dots, Z_M$  clamped at positions  $R_1, \dots, R_M$ . Our wavefunction  $\Psi$  describing the system of  $N$  electrons is then a function in  $H^1((\mathbb{R}^3 \times \Sigma)^N; \mathbb{C})$  with the additional constraints of  $L^2$ -normalization

$$\sum_{\sigma_1, \dots, \sigma_N} \int_{\mathbb{R}^{3N}} |\Psi(x_1, \sigma_1, \dots, x_N, \sigma_N)|^2 dx_1 \dots x_N = 1, \quad (2.17)$$

and antisymmetry

$$\Psi(x_{p(1)}, \sigma_{p(1)}, \dots, x_{p(N)}, \sigma_{p(N)}) = (-1)^{\varepsilon(p)} \Psi(x_1, \sigma_1, \dots, x_N, \sigma_N). \quad (2.18)$$

We call functions satisfying these constraints admissible (electronic) wavefunction and write

$$\mathcal{A}_N := \{ \Psi \in H^1((\mathbb{R}^3 \times \Sigma)^N; \mathbb{C}) \mid \Psi \text{ satisfies (2.18) and (2.17)} \}. \quad (2.19)$$

The (electronic) quantum mechanical energy functional is then given by

$$\begin{aligned} \mathcal{E}^{el}[\Psi, \{R_\alpha\}] &= T[\Psi] + V_{ee}[\Psi] + V_{ne}[\Psi] \\ &= \frac{1}{2} \sum_{\sigma \in \Sigma^N} \int_{\mathbb{R}^{3N}} |\nabla \Psi(x, \sigma)|^2 + \left( \sum_{1 \leq i < j \leq N} \frac{1}{|x_i - y_j|} - \sum_{i=1}^N \sum_{j=1}^M \frac{Z_j}{|x_i - R_j|} \right) |\Psi(x, \sigma)|^2 dx, \end{aligned} \quad (2.20)$$

where the individual parts of the energy functional are  $T[\Psi]$  denoting the kinetic energy,  $V_{ee}$  describing the electron-electron interaction energy, and  $V_{ne}[\Psi]$  corresponding to the electron-nuclei interaction energy.

Then, the electronic structure problem becomes, for fixed  $(R_1, \dots, R_M)$ ,

$$\text{Minimize } \mathcal{E}^{el}[\Psi, \{R_j\}] \quad \text{over } \Psi \in \mathcal{A}_N. \quad (2.21)$$

If it is clear from the context that the  $R_j$  are fixed, we will drop the explicit dependence and denote the (electronic) ground state energy by

$$E_N^0 := \inf_{\Psi \in \mathcal{A}_N} \mathcal{E}^{el}[\Psi]. \quad (2.22)$$

In this framework, the total energy of the system is obtain via minimizing of the electrons and the clamped nuclei, i.e.,

$$\text{Minimize } \mathcal{E}^{tot}[\Psi, \{R_j\}] \quad \text{over } \mathcal{A}_N \times \mathbb{R}^{3M}, \quad (2.23)$$

where

$$\mathcal{E}^{tot}[\Psi, \{R_j\}] = \mathcal{E}^{el}[\Psi, \{R_\alpha\}] + \sum_{1 \leq i < j \leq M} \frac{Z_i Z_j}{|R_i - R_j|}. \quad (2.24)$$

The last interesting object for us is the Born-Oppenheimer potential energy surface

$$E(\{R_j\}) = \inf_{\Psi \in \mathcal{A}_N} (\mathcal{E}^{el}[\Psi, \{R_\alpha\}]) + \sum_{1 \leq i < j \leq M} \frac{Z_i Z_j}{|R_i - R_j|}, \quad (2.25)$$

which we will discuss more in Section 2.4. Let us just mention that

$$\text{Minimize } \mathcal{E}(R) \quad \text{over } R \in \mathbb{R}^{3M}, \quad (2.26)$$

is called solving the geometry optimization problem, as one searches for the minimizing configuration of nuclei positions, which yields the molecular geometry.

## 2.3 HVZ Theorem and Bound States in the Born-Oppenheimer Approximation

As described in the last section, the Hamiltonian  $H_N^{el}$  describes our system in the sense that our ground state is given as its eigenfunction corresponding to the lowest eigenvalue. Therefore, we want to recall some fundamental properties of this operator in the following section. In particular, this includes the existence of well-localized and stable states, implying that the quantum systems under consideration exist as well-localized objects, and are stable under sufficiently small perturbations [66].

The first result goes back to Kato and is essential, since it guarantees that our energies are real and bounded from below.

**Theorem 2.2** (Kato 1951 [85])

*The operator  $H_N^{el}$  is self-adjoint and bounded from below.*

Understanding the energy levels of a given quantum system is still to this day one of the major problems in physics; in our case the general form of the spectrum is known (see, e.g., [83]). This result is attributed to Hunziker [82], van Winter [178], and Zhislin [196], with their initials giving it its name: HVZ-theorem.

**Theorem 2.3** (HVZ-Theorem [147])

*The essential spectrum of  $H_N^{el}$  takes the following form*

$$\sigma_{ess}(H_N^{el}) = [\Omega_N, \infty), \quad (2.27)$$

*where the ionization threshold  $\Omega_N = \inf \sigma(H_{N-1}^{el}) \leq 0$  and each potential eigenvalue, if it exists, must lie in  $(-\infty, 0]$ .*

Theorem 2.3 can be interpreted as follows: To obtain the energy values in  $\sigma_{ess}(H_N^{el})$ , remove one electron from the system and relocate it to infinity. The electron can move freely there, while the rest of the system is placed in its ground state, then the energy of the total system takes the form

$$\Omega_N + \frac{1}{2}|k|^2, \quad \text{for all momenta } k \text{ of the electron at infinity.}$$

Varying this expression over all values  $|k|$ , gives exactly  $\sigma_{ess}(H_N^{el}) = [\Omega_N, \infty)$ . Note, that removing more than one electron just creates a more positive energy level and thus gives nothing new.

Additionally, Theorem 2.3 incorporates the condition  $E_N^{el} < \Omega_N$  for a minimizer to exist, which corresponds to the physical property of the nuclei being able to bind  $N$  electrons in their vicinity. Physical intuition suggests that this should hold, at least as long as  $N$  is not significantly larger than  $Z$ .

The last result of this section, first shown by Zhislin [196], states precisely that (for an elegant mathematical proof see [52]).

**Theorem 2.4** (Bound states [66])

For  $N < Z + 1$ ,  $H_N^{el}$  has infinitely many eigenvalues  $(E_N^{(i)})_{i \geq 1}$  below its ionization threshold  $\Omega_N$ . Additionally the corresponding eigenfunctions  $\Psi_N^{(i)}$  of  $H_N^{el}$ , called bound states, decay exponentially in the sense that

$$\int_{\mathbb{R}^3} |\Psi_N^{(i)}(x)|^2 e^{2\alpha|x|} dx < \infty, \quad \forall \alpha < \sqrt{\Omega_N - E_N^{(i)}}.$$

The eigenfunctions, corresponding to eigenvalues above  $E_N^0$ , are called excited states.

Lastly, let us mention a related open problem, the so-called *ionization conjecture*, see [162, Problem 9] or [111, Chapter 12]. It comes from the experimental observations that a neutral atom can bind at most two extra electrons and tries to prove this rigorously from the first principles of quantum mechanics. The final goal here would be to establish a bound of the form  $N \leq Z + C$  as a restriction on the existence of minimizer. So far, this is unsolved, even though this problem has been studied extensively by many authors [44, 110, 112, 129, 150, 155, 160]. These papers resulted in various bounds for the maximal number  $N$  of electrons that a nucleus of charge  $Z$  can bind. In particular, the following bounds were obtained:

$$N \leq \min\{2Z + 1, 1.22Z + 3Z^{1/3}, Z + CZ^{5/7} + C\},$$

where  $C$  denotes some universal constant.

## 2.4 Why Do Molecules Bind Together?

In this last subsection, we want to consider an important problem, namely the binding and dissociation of molecules. As done in the literature, we will only be considering the simplest molecule, i.e., the hydrogen molecule to illustrate all relevant aspects of this question.

First of all, let us make clear what we mean by *binding*: We want to prove that

$$\varepsilon_{H^2} := \lambda_{\min}(H^{H^2}) < 2\lambda_{\min}(H^H) =: 2\varepsilon_H, \quad (2.28)$$

where  $H^{H_2}$  and  $H^H$  refer to the Hamiltonian of the  $H_2$ -molecule and the  $H$ -atom, respectively. To be more precise, define the Born-Oppenheimer potential energy surface

$$E^{H_2}(R) = \inf_{\Psi \in \mathcal{A}} \langle \Psi, H_R^{H_2} \Psi \rangle, \quad (2.29)$$

where  $R := |R_1 - R_2|$  denotes the distance between the two nuclei at  $R_1$  and  $R_2$ .

In order to see that  $E^{H_2}$  only depends on  $R$  and not on  $R_1$  and  $R_2$  explicitly, note the general fact that for any Galilean transformation  $g(x) := Ox + b$ , with  $O \in \mathcal{O}(3)$  a rotation matrix and  $b \in \mathbb{R}^3$  a translation vector, we have

$$H_{BO}(g(R_1), \dots, g(R_M)) = U_g H_{BO}(R_1, \dots, R_M) U_g^{-1}, \quad (2.30)$$

where  $U_g : L^2((\mathbb{R}^3 \times \Sigma)^N)$  denotes the unitary transformation

$$U_g \Psi(x_1, \sigma_1, \dots, x_N, \sigma_N) = \Psi(g^{-1}(x_1), \sigma_1, \dots, g^{-1}(x_N), \sigma_N).$$

Therefore, both Hamiltonians in (2.30) are isospectral and thus, one can always reduce the parameter space of nuclei positions under Galilean transformation. In the case of  $H_2$ , this just so happens to give us the desired assertion.

Now, we can formulate the binding and dissociation of  $H_2$  in the following theorem, which is visualized in Figure 2.1.

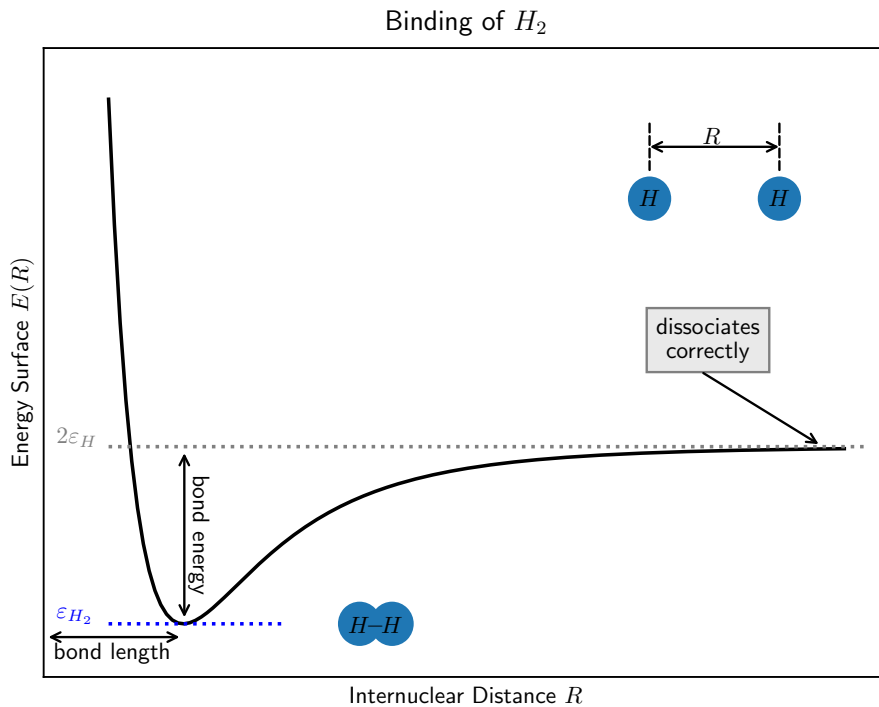


Figure 2.1: Plot of the Born-Oppenheimer potential energy surface of the  $H_2$  molecule in terms of the internuclear distance  $R$

**Theorem 2.5** (Binding of  $H_2$  – [53])

The Born-Oppenheimer potential energy surface of  $H_2$  given by (2.29) satisfies

- (i)  $\lim_{R \rightarrow \infty} E(R) = 2\varepsilon_H$ , i.e., the molecules dissociates correctly.
- (ii)  $\min_{R>0} E(R) = \varepsilon_{H_2} < 2\varepsilon_H$ , i.e.,  $H_2$  is binding.
- (iii)  $\lim_{R \rightarrow 0} E(R) = \infty$ , i.e., the internuclear repulsion dominates.

Since the equivalent statement of Theorem 2.5 in the DFT setting (explained in Chapter 3) is also the main result in one of our included articles [11] and furthermore gives a more quantitative behaviour of  $E(R)$ , let us provide a sketch of the proof. For more details, we refer to the lecture notes of Prof. Friesecke [53].

*Proof.* One starts by getting ride of the antisymmetric condition in minimization of (2.29), compare the appendix of our Article II. Next, it is quite straightforward to obtain a good upper bound by constructing a test-function from the ground-state  $\varphi$  of the hydrogen atom (2.4)

$$\Psi(x, y) = (\varphi_{R_1} \otimes \varphi_{R_2})(x, y) = \frac{e^{-|x-R_1|}}{\sqrt{\pi}} \frac{e^{-|y-R_2|}}{\sqrt{\pi}}.$$

With this wavefunction at hand, one gets (see, e.g., [167])

$$E(R) \leq \langle \Psi, H_R^{H_2} \Psi \rangle = 2\varepsilon_H + e^{-2R} \left( \frac{1}{R} + \frac{5}{8} - \frac{3}{4}R - \frac{1}{6}R^2 \right).$$

For the lower bounds, we have to split the regimes  $R \rightarrow 0$  and  $R \rightarrow \infty$ : First, notice that

$$\begin{aligned} H_R^{H_2} &= \frac{1}{2} \left( -\frac{1}{2}\Delta_x + \frac{2}{|x-R_1|} \right) + \frac{1}{2} \left( -\frac{1}{2}\Delta_x + \frac{2}{|x-R_2|} \right) + \frac{1}{|x-y|} \\ &+ \underbrace{\frac{1}{2} \left( -\frac{1}{2}\Delta_y + \frac{2}{|y-R_1|} \right) + \frac{1}{2} \left( -\frac{1}{2}\Delta_y + \frac{2}{|y-R_2|} \right)}_{=H_{R_1, R_2}} + \frac{1}{R}, \end{aligned}$$

where the Hamiltonian  $-\frac{1}{2}\Delta - \frac{Z}{|x|}$  has, by Section 2.1, ground state energy  $\varepsilon_H Z^2$ . This yields statement (iii) as

$$E(R) \geq 8\varepsilon_H + \frac{1}{R}.$$

For the lower bound at  $R \rightarrow \infty$ , we make the idea rigorous that the wavefunction splits into two parts  $\Psi \approx \varphi_1 + \varphi_2$ , with each  $\varphi_i$  staying only near  $R_i$ . To do so, we first assume, without loss of generality by the Galilean invariance (2.30), that  $R_1 = 0$  and  $R_2 = R\hat{e}_x$ , and then take a smooth partition of unity

$$0 \leq \tilde{\xi}_i \leq 1, \quad \tilde{\xi}_1^2 + \tilde{\xi}_2^2 = 1, \quad \tilde{\xi}_1(x) = 1 \text{ for all } x \leq \frac{1}{3}, \quad \text{and } \tilde{\xi}_1(x) = 0 \text{ for all } x \geq \frac{2}{3},$$

to define  $\xi_i : \mathbb{R}^3 \rightarrow \mathbb{R}$  with  $\xi_i(x) := \tilde{\xi}_i\left(\frac{x_1}{R}\right)$ . Note that we then have the estimate  $|\nabla \xi_i(x)| \leq \frac{C}{R}$ . Now, take any  $\varphi \in H^1(\mathbb{R}^3)$  with  $\|\varphi\|_{L^2} = 1$  and define  $\varphi_i = \xi_i \cdot \varphi$ . Then, we have  $\sum_{i=1}^2 |\varphi_i|^2 = |\varphi|^2$  and a short calculation shows

$$|\nabla \varphi|^2 \geq \sum_{i=1}^2 |\nabla \varphi_i|^2 - 2 \frac{C^2}{R^2} |\varphi|^2.$$

Additionally, we have

$$\langle \varphi_i, H_{R_1, R_2} \varphi_i \rangle \geq \left(\varepsilon_H - \frac{3}{R}\right) \|\varphi_i\|^2,$$

since  $|x - R_{1/2}| \geq \frac{R}{3}$  on  $\text{supp } \varphi_{2/1}$ . Combining the last two inequalities, implies

$$\langle \varphi, H_{R_1, R_2} \varphi \rangle \geq \left(\varepsilon_H - \frac{3}{R} - 2 \frac{C^2}{R^2}\right) \|\varphi\|^2,$$

yielding the final lower bound

$$E(R) \geq 2 \left(\varepsilon_H - \frac{3}{R} - 2 \frac{C^2}{R^2}\right) + \frac{1}{R}.$$

□

Theorem 2.5 proves that in quantum mechanics  $H_2$ , as the prototypical molecule, binds, i.e., there is chemical bonding between the two nuclei, with a molecular geometry determined by the corresponding Schrödinger equation. Furthermore, it shows the intuitive fact that, if one artificially increases the internuclear distance between parts of these molecules further and further until ultimately the bond is torn infinitely far apart, the energy of the limit system is the same as the sum of the original components.





## Chapter 3

# Density Functional Theory and the Kohn-Sham Equations

*Energy is a very subtle concept. It is very, very difficult to get right.*

---

Richard Feynman

Although the mathematical description of molecular quantum mechanics, as described in Chapter 2, was already developed in the 1920s by people like Heisenberg, Schrödinger, and Dirac, it took a long time until it found its way into application. This is due to the fact that the set of admissible functions  $\mathcal{A}_N$  from (2.19) grows exponentially with the number  $N$  of electrons.

This result is the so-called curse of dimension: Since the wavefunctions, over which one minimizes, are functions on the high-dimensional space  $\mathbb{R}^{3N}$ , a discretization scheme requires  $K^N$ -grid points, if the single-particle space  $\mathbb{R}^3$  is discretized by  $K$ -grid points. Take for example a simple molecule like  $\text{CO}_2$ ; it has 22 electrons, so if we use 10 grid points for every dimension, which is not that much, then the whole system requires  $10^{66}$  grid points, which roughly is the number of particles in the entire Milky Way galaxy.

Thus one needs a way to reduce the complexity of the system, in order to obtain something computationally feasible even for large systems; this is the realm of *density functional theory* (DFT). Introduced by Hohenberg, Kohn, and Sham in two fundamental papers [79,87] in the 1960s, this theory transforms the high-dimensional Schrödinger problem (2.22) into a low-dimensional one by converting the original linear system into a non-linear one in fewer variables.

The trade-off in this approach consists in introducing the so-called exchange-correlation energy functional, which is in theory exact but in practice unknown. Therefore, many approximations exist in the literature, trying to model this intricate many-body interaction energy [9,95,137,138]. Despite the long time since its establishment, DFT is still an important and active research area in physics, chemistry, and mathematics, see, e.g., [14, 29, 48, 59, 80, 123, 145]. As evidence for its accomplishments, Walter Kohn was awarded the Nobel Prize in chemistry 1998 for his contribution to its development. Furthermore, according to [177], it is “easily the most heavily cited concept in the physical sciences [...] twelve papers on the top-100 list relate to it, including 2 of the top 10”.

Additionally, its success can be seen from the countless quantum chemistry and solid-state physics packages implementing it, like, e.g., Octopus [3], BigDFT [60, 126, 146], or the more recent DFTK [76]. The list given in [118] gives an overview over the most notable software packages for quantum chemistry, with 90% of the them utilizing DFT.

In this chapter, we begin with a short historic overview of the milestones leading up to the development of DFT, as well as its formative years spanning roughly 1980-2010. After that, we introduce the modern formalism and precise mathematical framework, and discuss exchange-correlation functionals, focusing mostly on the so-called *local density approximation* (LDA). We conclude this chapter with a detailed discussion of our own contributions to some of the open problems in this field.

### 3.1 Density Functional Theory and its Predecessor

The full historic developments of DFT, while fascinating, are far too complex and widely branched to be portrayed here in their full glory. Still, we want to give a compact overview over some of the main milestones along the way, mostly based on the reviews [10, 84].

Just three years after Schrödinger derived his famous equation, Dirac [33] wrote the following:

“The general theory of quantum mechanics is now almost complete, [. . .]. The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.”

This necessity for “approximate practical methods of applying quantum mechanics” to accurately explain complex systems is a perfect description of the motivation for density functional theory. The DFT formalism shows that all the relevant information about a many-body quantum system at or near its ground state can be expressed in terms of its one-body density  $\rho$ . The intuitive idea that the energy of the system could be locally modeled by its uniform electron density already goes back to the early days of quantum mechanics.

The earliest predecessor of modern density functional theory is considered to be Thomas-Fermi (TF) theory, introduced in 1927 by Thomas [174] and Fermi [45, 46]. In their model they recognized the basic nature of the electron density and applied it to atoms.

They assumed the electrons to form a gas satisfying Fermi statistics, with the interaction energy solely determined by the classical Coulomb potential. The kinetic energy was replaced by a local density approximation, inspired by the kinetic energy of a homogeneous electron gas; the variational formulation of it was found by Lenz [103]. This yields the energy functional

$$\mathcal{E}^{TF}[\rho] = c_{TF} \int_{\mathbb{R}^3} \rho^{5/3} dx + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|} dx dy + \int_{\mathbb{R}^3} V_{ext}\rho dx, \quad (3.1)$$

where  $V_{ext}$  is some external potential, usually the standard Coulomb potential.

This model and its extensions allow us to approximately describe a variety of quantities, like the charge density, the electrostatic potential, and the variation of the total energy with the atomic number. Also its mathematical properties have been studied extensively, see, e.g., [108, 154, 164]. Furthermore, it is exact in the large  $Z$  limit [113], i.e., it captures the leading order behavior of the ground state energy for  $Z \rightarrow \infty$ .

However, Thomas-Fermi theory has some serious deficiencies, mostly because of its poor description of the outer region of an atom, i.e., it is unable to self-consistently reproduce atomic shell structure. The most famous problem is the so-called Teller’s “no-binding theorem” [172]: It loosely speaking states that in TF theory neutral atoms or, with some restrictions, ions do not form molecules or solids. This makes the model unsuitable for chemistry or material sciences at normal temperatures and pressures.

Dirac [34] extended this approach by incorporating exchange phenomena using Hartree-Fock theory in terms of a density function. Additionally, as a leading-order correction to the kinetic energy, the von Weizsäcker’s gradient term [185] corresponding to particles very close to the nucleus was added, resulting in Thomas-Fermi-Dirac-von Weizsäcker (TFDW) theory

$$E_Z^{TFDW}(N) := \inf \left\{ \mathcal{E}^{TFDW}[\rho] : \rho \geq 0, \sqrt{\rho} \in H^1(\mathbb{R}^3), \int_{\mathbb{R}^3} \rho \, dx = N \right\},$$

with the energy functional being

$$\begin{aligned} \mathcal{E}^{TFDW}[\rho] &= c_{TF} \int_{\mathbb{R}^3} \rho^{5/3}(x) \, dx + \int_{\mathbb{R}^3} V_{ext}(x) \rho(x) \, dx + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x) \rho(y)}{|x-y|} \, dx \, dy \\ &+ c_W \int_{\mathbb{R}^3} |\nabla \sqrt{\rho(x)}|^2 \, dx - c_D \int_{\mathbb{R}^3} \rho^{4/3}(x) \, dx. \end{aligned}$$

The exact value of the physical constants  $c_{TF}$ ,  $c_W$ , and  $c_D$  can be found in, e.g., [108]. The fact that the above problem has minimizers was proven by Lions [117] for positively charged and neutral molecules  $N \leq Z$ . Le Bris [94] extended this to slightly negatively charged ions, i.e.,  $N \leq Z + \varepsilon$  for some  $\varepsilon > 0$ .

The behavior of the of the energy with respect to the particle number, while completely understood in Thomas-Fermi theory (see Figure 3.1), is here more intricate due to the concavity of the Dirac term.

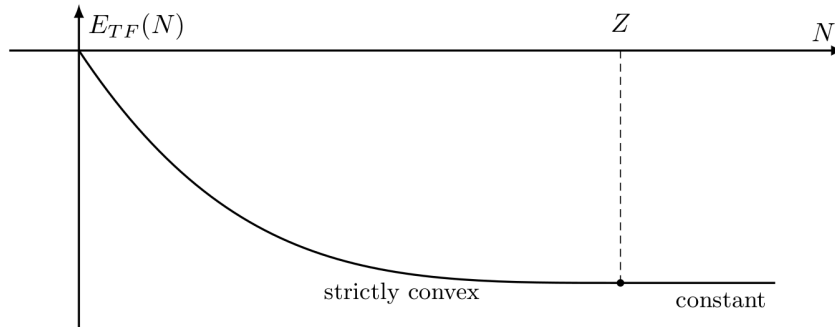


Figure 3.1: The Thomas-Fermi energy  $E_{TF}(N)$  with respect to the particle number  $N$ . For positively charged systems  $N < Z$  it is strictly convex, for  $N > Z$  it remains constant.

This can be seen from the fact that non-existence for large  $N$  has been solved only quite recently, in the special case  $Z = 0$  by Lu and Otto [119], and for  $Z > 0$  but very small by Nam and Bosch [130]. Another approximation method, which in contrast to TF satisfies the shell structure coming from the Pauli exclusion principle, is Hartree-Fock (HF) theory.

In [70, 71], Hartree introduced a scheme for calculating the wavefunction of an atom and with it the idea of a “self-consistent field”. In his approach, the wavefunction of an electron  $\psi_i$  is determined by the field of the nucleus and the other electrons. One starts with an approximate field and iterates until input and output fields for all electrons are the same. The complete  $N$ -particle wavefunction is then given by the product

$$\Psi(z_1, \dots, z_N) = \psi_1(z_1) \cdot \dots \cdot \psi_N(z_N), \quad (3.2)$$

where the  $\psi_i$  are orthonormal and each  $\psi_i$  solves a Schrödinger equation with a potential created by the average field of the other electrons.

This “Hartree approximation” was generalized to more complex systems by Fock [49] and Slater [163]. They replaced the product in (3.2) by a determinant satisfying the Pauli exclusion principle, later called Slater determinant, i.e.,  $\Psi$  took now the form

$$\Psi(z_1, \dots, z_N) = \frac{1}{\sqrt{N!}} \det \begin{pmatrix} \psi_1(z_1) & \dots & \psi_1(z_N) \\ \vdots & \ddots & \vdots \\ \psi_N(z_1) & \dots & \psi_N(z_N) \end{pmatrix}.$$

From a modern perspective, this corresponds to the orthonormal projection of the tensor product in (3.2) onto the antisymmetric-subspace of  $L^2((\mathbb{R}_\Sigma^3)^N)$ . Plugging this ansatz into the electronic energy functional  $\mathcal{E}^{el}$  from (2.20) leads to the so-called Hartree-Fock energy,

$$\begin{aligned} \mathcal{E}^{HF}[\Psi] &= \sum_{i=1}^N \frac{1}{2} \int_{\mathbb{R}^3} |\nabla \psi_i|^2 dx + \sum_{i=1}^N \int_{\mathbb{R}^3} V(x) |\psi_i|^2 dx + \frac{1}{2} \sum_{i \neq j} \int_{\mathbb{R}_\Sigma^3 \times \mathbb{R}_\Sigma^3} \frac{|\psi_i|^2(z_1) |\psi_j|^2(z_2)}{|x_1 - x_2|} dz_1 dz_2 \\ &\quad - \frac{1}{2} \sum_{i \neq j} \int_{\mathbb{R}_\Sigma^3 \times \mathbb{R}_\Sigma^3} \frac{\psi_i^*(z_1) \psi_j(z_1) \psi_i^*(z_2) \psi_j(z_2)}{|x_1 - x_2|} dz_1 dz_2. \end{aligned} \quad (3.3)$$

The self-consistent field of Hartree and the generalizations to determinants of wavefunctions by Slater, Bloch, and Fock were followed by computations of Wigner and Seitz [189, 190], who developed methods for treating the wavefunction in crystals. In the following years, many theoretical results, like the “Hellmann-Feynman theorem” [47, 75], provided more advances in the development of approximate practical methods.

The starting point of modern density functional theory though can be traced back to Hohenberg and Kohn [79] and the year 1964. In their paper, they showed a one-to-one correspondence between the external potential  $V_{\text{ext}}$  and the (non-degenerate) ground state wavefunction  $\Psi$  and also between  $\Psi$ , and the ground state density  $\rho$  of an  $N$ -electron system,

$$\rho(r) = N \int \Psi^*(r, r_2, \dots, r_N) \Psi(r, r_2, \dots, r_N) dr_2 \dots dr_N, \quad (3.4)$$

where the spin coordinates are not shown explicitly. Through the density  $\rho$ , the external potential and thus the Hamiltonian can be determined up to a constant. Hence,  $\rho$  suffices to establish the excited states as well as the ground state.

In order to apply these ideas to the total energy, they defined the universal functional  $F[\rho(r)]$ , which is valid for any external potential,

$$F[\rho] = \langle \Psi_\rho | T + V_{ee} | \Psi_\rho \rangle. \quad (3.5)$$

Here,  $T$  denotes the kinetic energy and  $V_{ee}$  the electron-electron interaction potential. Hohenberg and Kohn showed that the energy functional  $\mathcal{E}[\rho, V_{\text{ext}}]$  satisfies a variational principle

$$E_{GS} = \min_{\rho} \mathcal{E}[\rho, V_{\text{ext}}] \quad \text{with} \quad \mathcal{E}[\rho, V_{\text{ext}}] = \int V_{\text{ext}} \rho \, dx + F[\rho]. \quad (3.6)$$

The task, which now remains, is finding good approximations to the functional  $F[\rho]$ . This is the content of the famous paper [87] by Kohn and Sham. Their approach was the following:

$$F[\rho] = T_{KS}[\rho] + \frac{1}{2} \int_{\mathbb{R}^3} \rho \Phi \, dx + E_{xc}[\rho], \quad (3.7)$$

where  $T_{KS}$  is the kinetic energy corresponding to a system without electron-electron interactions,  $\Phi$  is the classical Coulomb potential for electrons, and  $E_{xc}$  describes the exchange-correlation energy. Even though  $T_{KS}$  is not the true kinetic energy, it is of comparable magnitude and hence treated here without approximation. This removes many of the deficiencies of Thomas-Fermi theory, such as the absence of chemical bonding in molecules and solids [84].

The only term in (3.7), which can not be evaluated exactly, is  $E_{xc}$ , so approximations for this term are crucial in applications. Kohn and Sham [87] proposed using the so-called local density approximation (LDA)

$$E_{xc}^{LDA} = \int_{\mathbb{R}^3} \rho(x) \varepsilon_{xc}(\rho(x)) \, dx, \quad (3.8)$$

where  $\varepsilon_{xc}$  describes the exchange-correlation energy per particle of a homogeneous electron gas. Note, in mathematics the compact notation  $e_{xc}(\rho) = \rho \varepsilon_{xc}(\rho)$  is more common and will be used from now on in this thesis.

This approximation works quite well if the density is almost constant, as well as at high densities, where the kinetic energy dominates the exchange correlation terms. The DFT was soon extended to finite temperature [124], spin-polarized systems or external magnetic fields [144, 184], and in the 1980s time dependence [120, 149, 165, 194] was brought into the picture.

More complex exchange-correlation functionals, like the local spin density approximations (LSDA) or the  $X_\alpha$  approximation followed quickly. But since these can lead to overbinding of molecules and the corresponding Kohn-Sham eigenvalues often underestimate the optical band gaps measured in experiments, improved approximations with less mathematical rigor were developed. Functionals relying on the gradient of the density, i.e., setting  $\varepsilon_{xc} = \varepsilon_{xc}(\rho, \nabla \rho)$  in (3.8), called generalized gradient approximation (GGA) [8, 95, 136], did lead to better results in most cases. Additionally, *hybrid* functionals, which included a Hartree-Fock-like exchange component, were

introduced by Becke [9]. His exchange functional has three parameters and used the correlation part from Lee, Yang, and Parr [95] leading to the name B3LYP. It is to this day the most commonly used approximation in chemical applications [14].

Over the years, many more empirical functionals have been proposed with parameters often fitted to data of particular types of molecules. Relying too much on experimental data gave some scientists the impression that DFT is semiempirical in nature [84].

To counteract this development, others proposed an alternative path. In particular, Perdew and collaborators developed a sequence of approximations without experimental input. They used the metaphor of *Jacob's Ladder*, where each *rung* builds on the experience of the lower levels and satisfies certain physical restraints. Their GGA functional PBE (Perdew, Burke and Ernzerhof [137]) incorporates LSAD from below it and the *meta-GGA* from TPSS [171] builds on both of them.

While climbing this ladder, the computational cost increases and the agreement with experiments usually improves, but the theoretical interpretation becomes less clear. As observed in [123], starting in the early 2000s, newer approximations actually become worse in predicting the electron densities. This is due to only focusing on the energies and in the process sacrificing mathematical rigor in favor of the flexibility of fitting to empirical data.

Historically it took quite some time for DFT to get widely accepted in the applied sciences like quantum chemistry or solid state physics, as can be seen in Figure 3.2.

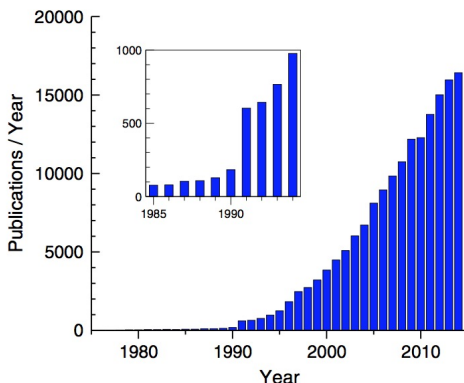


Figure 3.2: Number of publications per yeas (1975 – 2014) on topics (“density functional” or “DFT”), according to the Web of Science Core Collection (February 2015). [84]

The condensed matter theorist Heine [73] looked back on the developments like this:

“Of course at the beginning of the 1960s the big event was the Kohn Hohenberg Sham reformulation of quantum mechanics in terms of density functional theory (DFT). Well, we recognize it now as a big event, but it did not seem so at the time. That was the second big mistake of my life, not to see its importance, but then neither did the authors judging from the talks they gave, nor anyone else. Did you ever wonder why they never did any calculations with it?”

The Seventh International Congress of Quantum Chemistry in 1991 is considered by many a major turning point in the fortunes of DFT methods, particularly in chemistry. During the congress, the presentation of DFT methods led to various discussions among skeptics and proponents.

Despite the opposition in the beginning, this exchange resulted in increasing interest and research on DFT, and it thus became a more widely spread research area also in quantum chemistry.

Note that the variation in the number of citations in Figure 3.2 should be interpreted with caution. Many articles which nowadays would be associated to DFT made no reference to it. Jones in [84] formulates it as follows: “it seems that many realized around this time that they had been doing density functional calculations all along.”

## 3.2 Modern Density Functional Formalism

Now, we want to precisely define the mathematical framework needed to investigate DFT. We follow here the modern formalism introduced by Levy [104] and made rigorous by Lieb [109]. For a nice introduction to DFT from an applied mathematics point of view, see [114]. Additional standard references for this would be [134] or [42]. For an overview over DFT with the focus on the different exchange-correlation functionals, we suggest the nice review article by J. Toulouse [175].

### 3.2.1 Hohenberg-Kohn Theorem

Recall the quantum mechanical energy functional in the Born-Oppenheimer approximation

$$\mathcal{E}^{el}[\Psi] := T[\Psi] + V_{ne}[\Psi] + V_{ee}[\Psi], \quad (3.9)$$

where

$$T[\Psi] := \frac{1}{2} \int_{(\mathbb{R}_\Sigma^3)^N} \sum_{i=1}^N |\nabla_{x_i} \Psi(x_1, s_1, \dots, x_N, s_N)|^2 dz_1 \dots dz_N$$

describes the kinetic energy,

$$V_{ne}[\Psi] := \int_{(\mathbb{R}_\Sigma^3)^N} \sum_{i=1}^N V(x_i) |\Psi(x_1, s_1, \dots, x_N, s_N)|^2 dz_1 \dots dz_N$$

gives the electron-nuclei interaction energy, and

$$V_{ee}[\Psi] := \int_{(\mathbb{R}_\Sigma^3)^N} \sum_{1 \leq i < j \leq N} v_{ee}(x_i - x_j) |\Psi(x_1, s_1, \dots, x_N, s_N)|^2 dz_1 \dots dz_N$$

is the electron-electron interaction energy. Here,  $V$  denotes the external potential, usually the Coulomb potential and  $v_{ee}$  gives the interaction between the electrons through the Coulomb interaction  $v_{ee}(x) = \frac{1}{|x|}$ . Furthermore, note  $\int_{\mathbb{R}_\Sigma^3} f(z) dz = \sum_{s \in \Sigma} \int_{\mathbb{R}^3} f(x, s) dx$  has been used as

a shorthand notation. We are then interested in the exact quantum mechanical ground state energy

$$E_0^{el} = \inf_{\Psi \in \mathcal{A}_N} \mathcal{E}^{el}[\Psi]. \quad (3.10)$$

At the heart of the DFT approach lies the one-body density  $\rho_\Psi$  of an  $N$ -particle fermionic wavefunction defined by

$$\rho_\Psi(x) := N \sum_{\sigma_1, \dots, \sigma_N \in \Sigma} \int_{\mathbb{R}^3} \dots \int_{\mathbb{R}^3} |\Psi(x, \sigma_1, x_2, \sigma_2, \dots, x_N, \sigma_N)|^2 dx_2 \dots dx_N. \quad (3.11)$$

We can interpret  $\rho_\Psi$  as providing the average number of particles in space, without taking their spin component into account, hence the normalization factor  $N$  in (3.11).

The main idea now is to replace the infimum over  $\Psi$  in (3.10) by a two-step minimization of the form

$$\inf_{\Psi} \mathcal{E}^{el}[\Psi] = \inf_{\rho} \inf_{\substack{\Psi \\ \rho_\Psi = \rho}} \mathcal{E}^{el}[\Psi], \quad (3.12)$$

where on the right hand side the minimization is done first over the density  $\rho$  and then over all the wavefunctions  $\Psi$  having this prescribed density. This simple looking procedure allows to partition the energy functional into two parts:  $T + V_{ee}$  being *universal*, i.e., independent of the external potential and thus the same for all molecules, and  $V_{ne}$  being chemically specific, i.e., it is the only term incorporating the structure of the clamped nuclei. Therefore, we define the Levy-Lieb functional

$$F_{LL}[\rho] := \inf_{\substack{\Psi \in \mathcal{A}_N \\ \rho_\Psi = \rho}} \left( T[\Psi] + V_{ee}[\Psi] \right) \quad (3.13)$$

and the Hohenberg-Kohn energy functional

$$\mathcal{E}^{HK}[\rho] = F_{LL}[\rho] + \int_{\mathbb{R}^3} V(x)\rho(x) dx. \quad (3.14)$$

Note that the universal functional  $F_{LL} : \mathcal{R}_N \rightarrow \mathbb{R}$  is well-defined and the infimum in (3.13) is actually a minimum [109], when  $\rho$  belongs to the class (3.15) below.

The first essential challenge is to identify the set of  $N$ -representable densities, that is, those arising from an  $N$ -particle wavefunction  $\Psi \in \mathcal{A}_N$ ,

$$\mathcal{R}_N := \left\{ \rho : \mathbb{R}^3 \rightarrow \mathbb{R} : \rho \text{ is the density of some wavefunction } \Psi \in \mathcal{A}_N \right\}. \quad (3.15)$$

One necessary condition comes from the Hoffmann-Ostenhof inequality [78]

$$\sum_{i=1}^N \int_{(\mathbb{R}_\Sigma^3)^N} |\nabla_{x_i} \Psi(x_1, \sigma_1, \dots, x_N, \sigma_N)|^2 dz_1 \dots dz_N \geq \int_{\mathbb{R}^3} |\sqrt{\nabla \rho_\Psi}(x)|^2 dx, \quad (3.16)$$

implying that  $\sqrt{\rho_\Psi} \in H^1(\mathbb{R}^3)$ .



A beautiful result by Lieb [109] proves this restriction to be optimal, i.e.,

$$\mathcal{R}_N := \left\{ \rho : \mathbb{R}^3 \rightarrow \mathbb{R} : \rho \geq 0, \sqrt{\rho} \in H^1(\mathbb{R}^3), \int_{\mathbb{R}^3} \rho(x) dx = N \right\}.$$

Note, due to the Sobolev embedding  $H^1(\mathbb{R}^3) \hookrightarrow L^6(\mathbb{R}^3)$  we have  $\rho \in L^3(\mathbb{R}^3)$ . Combining this with the dual relation

$$\left( \underbrace{L^3(\mathbb{R}^3) \cup L^1(\mathbb{R}^3)}_{\supseteq \mathcal{R}_N \ni \rho} \right)^* = L^{3/2}(\mathbb{R}^3) + L^\infty(\mathbb{R}^3),$$

we see that the chemically specific term  $\int_{\mathbb{R}^3} V(x)\rho(x) dx$  is well-defined for the entire class of potentials  $V \in L^{3/2}(\mathbb{R}^3) + L^\infty(\mathbb{R}^3)$ . In particular, this includes the Coulomb potential which is not in any  $L^p$ -space.

Therefore, the Hohenberg-Kohn ground state energy becomes

$$E_0^{HK} := \inf_{\rho \in \mathcal{R}_N} \mathcal{E}^{HK}[\rho] := \inf_{\rho \in \mathcal{R}_N} \left\{ F_{LL}[\rho] + \int_{\mathbb{R}^3} V(x)\rho(x) dx \right\}. \quad (3.17)$$

As mentioned above, this constrained-search definition of  $F_{LL}$  is due to Levy and Lieb [104, 109]; historically, in the original paper [79] by Hohenberg and Kohn, the functional was constructed in a more indirect and slightly less general way. They required  $\rho$  to be the density of some wavefunction  $\Psi$ , which is a non-degenerate ground state of  $\mathcal{E}^{el}$  for some potential  $V$ , and proved that two potentials differing by more than a constant produce different densities.

We summarize the results by Hohenberg and Kohn in the following Theorem.

**Theorem 3.1** (Hohenberg-Kohn, 1964, modern version)

*In the above setting the following two statements hold.*

(i) (ground state energy) *Both minimizations yield the same energy, i.e.,*

$$E_0^{HK} = \inf_{\rho \in \mathcal{R}_N} \mathcal{E}^{HK}[\rho] = \inf_{\Psi \in \mathcal{A}_N} \mathcal{E}^{el}[\Psi] = E_0^{el}. \quad (3.18)$$

(ii) (ground state density) *There exists a minimizer  $\rho \in \mathcal{R}_N$  of  $\mathcal{E}^{HK}$  if and only if there exists  $\Psi \in \mathcal{A}_N$  with  $\rho_\Psi = \rho$  and  $\Psi$  being a minimizer of  $\mathcal{E}^{el}$ .*

(iii) (external potential) *There exists a one-to-one correspondence between the external potential  $V(x)$  and the ground-state density  $\rho(x)$ , i.e., the external potential is a unique functional (up to an additive constant) of the ground-state density  $V[\rho](x)$ .*

*So the minimization over the lower dimensional functional  $\mathcal{E}^{HK}$  correctly predicts the exact quantum mechanical electron energy and density of the whole system.*

It is important to mention that the last point of the above Theorem, while accepted in the physics literature, is not completely settled in mathematics, in the sense that the exact necessary assumptions for its validity are not yet fully understood mathematically. This is related to the  $V$ -representability problem [42, 122], as well as the many-body unique continuation principle

[91, 106]. A significant milestone towards understanding this problem is due to Garrigue [59]. He proved the validity of the Hohenberg-Kohn theorem under the assumption that all involved potentials are in  $L^p(\mathbb{R}^d) + L^\infty(\mathbb{R}^d)$  with  $p > \max\{\frac{2d}{3}, 2\}$ , which in particular includes the Coulomb potential.

The astonishing result here is that the admissible set  $\mathcal{R}_N$  of densities carries the same information as  $\mathcal{A}_N$ , but its size does not depend on  $N$ . Thus, if one had an explicit form for  $F_{LL}$ , there would be no curse of dimension anymore.

Let us conclude this section by emphasizing that the beautiful but counter-intuitive idea here was to go from an original linear problem to a nonlinear one in fewer variables, opposite to the common strategy in undergraduate mathematics to linearize nonlinear problems.

### 3.2.2 Kohn-Sham Equations

The problem one now faces is that there is no tractable expression of  $F_{LL}$ , which could be used in practice. Ideally, one would want an expression in terms of the density  $\rho$  for the kinetic energy  $T[\rho]$  and inter-particle potential  $V_{ee}[\rho]$ .

The idea by Kohn and Sham [87] was to consider a (fictional) non-interacting system, described by some effective potential  $v_s$ , i.e.,

$$H_s = -\frac{1}{2}\Delta + \sum_{i=1}^N v_s(x_i), \quad H_s \Phi_\rho = E_s \Phi_\rho,$$

with the constraint that  $\Phi_\rho$  gives the same density  $\rho$  (and chemical potential  $\mu$ ) as  $\Psi_\rho$  which is guaranteed by the Hohenberg-Kohn theorem. Note, since we are dealing with a non-interacting system, the minimizer  $\Phi_\rho$  is a Slater determinant,  $\Phi_\rho = |\varphi_1 \dots \varphi_N\rangle$  for some orbitals  $\{\varphi_i\}_{i=1}^N$ . Furthermore, for a Slater determinant the one-body density can be computed in terms of the orbitals, to yield

$$\rho = \sum_{i=1}^N \sum_{s \in \Sigma} |\varphi_i(x, s)|^2 dx.$$

Next, since the infimum in the definition of  $F_{LL}$  is attained [109] and we have a one-to-one correspondence between density and wavefunction, we obtain

$$\begin{aligned} F_{LL}[\rho] &= T[\rho] + V_{ee}[\rho] = \langle \Psi_\rho | -\frac{1}{2}\Delta + V_{ee} | \Psi_\rho \rangle \\ &= \langle \Phi_\rho | -\frac{1}{2}\Delta + V_{ee} | \Phi_\rho \rangle + E_c[\rho], \end{aligned}$$

where the correlation energy is defined through the last equality. Since we are dealing with a non-interacting system, the kinetic energy  $T_s[\rho] = \langle \Phi_\rho | -\frac{1}{2}\Delta | \Phi_\rho \rangle$  can be written directly in terms of the orbitals

$$T_s[\rho] = \frac{1}{2} \sum_{i=1}^N \int_{\mathbb{R}_\Sigma^3} |\nabla \varphi_i|^2(z) dz.$$

To understand the approximation of the interaction potential, let us note that  $V_{ee}[\Psi]$  can be expressed explicitly through the pair density

$$\rho_2^\Psi(x, y) = \binom{N}{2} \int_{(\mathbb{R}^3)^{N-2}} \sum_{\sigma_1, \dots, \sigma_N \in \Sigma} |\Psi(x, \sigma_1, y, \sigma_2, x_3, \sigma_3, \dots, x_N, \sigma_N)|^2 dx_3 \dots dx_N,$$

in the form

$$V_{ee}[\Psi] = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_2^\Psi(x, y)}{|x - y|} dx dy.$$

Starting now with a statistical independence ansatz, i.e.,

$$\rho_2(x, y) = \frac{1}{2} \rho(x) \rho(y), \quad (3.19)$$

gives the following form for the electron-electron interaction:

$$\langle \Phi_\rho | V_{ee} | \Phi_\rho \rangle = J[\rho] + E_x[\rho], \quad J[\rho] = \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x) \rho(y)}{|x - y|} dx dy,$$

where the exchange energy  $E_x[\rho]$  is again defined such that equality holds. Finally,

$$F_{LL}[\rho] = \underbrace{T_s[\rho] + J[\rho]}_{\text{treated exactly}} + E_{xc}[\rho], \quad E_{xc}[\rho] = E_c[\rho] + E_x[\rho].$$

Now, the coupling of the fictional and the real system via the density comes into play. Consider the Euler-Lagrange equations of both systems with the chemical potential  $\mu$  as the Lagrange multiplier

$$\left. \begin{aligned} \frac{\partial T_s}{\partial \rho(x)} + v_s(x) &= \mu, \\ \frac{\partial T_s}{\partial \rho(x)} + \frac{\partial J}{\partial \rho(x)} + \frac{\partial E_{xc}}{\partial \rho(x)} + v(x) &= \mu. \end{aligned} \right\} \quad \text{same solution (fictional \& real system)}$$

From this, we obtain for the effective potential  $v_s$ :

$$v_s(x) = v(x) + \int_{\mathbb{R}^3} \frac{\rho(y)}{|x - y|} dy + v_{xc}([\rho], x), \quad v_{xc}([\rho], x) = \frac{\partial E_{xc}}{\partial \rho(x)}.$$

Plugging this back into our energy functional, and using the explicit structure of the kinetic term  $T_s$  in terms of the orbitals, gives the Kohn-Sham energy functional

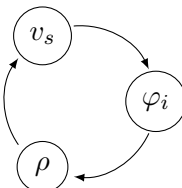
$$\mathcal{E}^{KS}[\Phi] = \sum_{i=1}^N \frac{1}{2} \int_{\mathbb{R}^3} |\nabla \varphi_i(z)|^2 dz + \int_{\mathbb{R}^3} v \rho dx + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x) \rho(y)}{|x - y|} dx dy + E_{xc}[\rho]. \quad (3.20)$$

Up until this point, no approximation has been made, since everything was absorbed into the exchange-correlation energy. Thus, we have just shifted our problem to finding a good approximation of  $E_{xc}[\rho]$ , which will be discussed in the next subsection.

Before we move on, let make some remarks. First, the orbitals in the above equation are known as the Kohn-Sham orbitals. Since these orbitals come from the fictitious non-interacting system,

they are only connected to the real system by having the same density. A direct interpretation, while widely done in practice is not completely justified, see, e.g., [166, 176].

In practice, one usually computes these orbitals by using the self-consistent Kohn-Sham scheme:

$$\underbrace{\left[ -\frac{1}{2}\Delta + v(x) + \int_{\mathbb{R}^3} \frac{\rho(y)}{|x-y|} dy + v_{xc}([\rho], x) \right]}_{H_{KS}} \varphi_i = \varepsilon_i \varphi_i, \quad \rho = \sum_{i=1}^N |\varphi_i|^2$$


As depicted by the small graphical loop, one starts with an educated guess for the orbitals, then computes the density, and from there the effective potential  $v_s$ . Then, by solving the Kohn-Sham equations one obtains an updated version of the orbitals. This scheme is performed until convergence. The eigenvalues appearing in the Kohn-Sham equations are known as the Kohn-Sham eigenvalues. There is a lot of discussion about the interpretation of the KS orbitals and eigenvalues. Since they come from the fictitious non-interacting system, and are only connected to the real system by having the same density, a direct interpretation, while sometimes loosely done in practice, is not theoretically justified, see, e.g., [166, 176].

### 3.2.3 Exchange-Correlation Functionals

Thus, the challenge becomes finding an accurate approximation for  $E_{xc}[\rho]$ . As mentioned already in Section 3.1, there is a huge variety of different exchange-correlation functionals (see, e.g., [9, 136, 140, 141]), each with its advantages and disadvantages, see, e.g., the Libxc library [102] or [121] for an overview.

In the following, we will only consider a certain class of functionals, namely the so-called local density approximation (LDA), proposed by Kohn and Sham [87]. Here, the exchange correlation functional is assumed to be of the form

$$E_{xc}[\rho] = \int_{\mathbb{R}^3} e_{xc}(\rho(x)) dx, \quad (3.21)$$

where the function  $e_{xc} : [0, \infty) \rightarrow \mathbb{R}$  has to fulfill certain properties. This usually includes some weak smoothness assumptions ( $e_{xc} \in C^1([0, \infty))$ ) as well as growth conditions.

The prototypical example for an  $E_{xc}[\rho]$  approximation stems from examining the homogeneous electron gas. In this system, one considers  $N$  non-interacting electrons in a 3 dimensional box of side length  $L$ , and then considers the thermodynamic limit  $N, L \rightarrow \infty$ , while keeping the density  $\rho = \frac{N}{L^3}$  constant.

Here, one is then able to calculate the exchange energy (note there is no correlation) exactly. It goes back to Dirac [34] (for a mathematical derivation see [52]) and is given by

$$E_{xc}[\rho] = \int_{\mathbb{R}^3} e_{xc}(\rho(x)) dx, \quad e_{xc}(\rho) = -c_{xc}\rho^{4/3}, \quad c_{xc} = \frac{3}{4} \left( \frac{3}{\pi} \right)^{1/3}. \quad (3.22)$$

This is included in almost all LDA type functionals as the exchange term with additive correlation corrections.

As mentioned in Section 3.1, functionals currently used in practice, like the ‘B3LYP’ functional of Becke, Lee, Yang, and Parr [9,15], rely on more complicated forms (e.g., local in  $\rho$ , or local in  $\rho$  and  $\nabla\rho$  with even additional terms depending non-locally on the KS orbitals), which are from a mathematical point of view questionable. These semi-empirical approaches require fitting of parameters to experimental or high-accuracy-computational data and have only lead to an improvement in accuracy over the local density approximation of approximately one order of magnitude [29].

Plugging the above ansatz (3.22) into (3.20), we obtain

$$E_0^{el} \approx E_0^{LDA} = \inf \left\{ \frac{1}{2} \sum_{i=1}^N \int_{\mathbb{R}_{\Sigma}^3} |\nabla \varphi_i|^2(z) dz + \int_{\mathbb{R}^3} V(x) \rho(x) dx + \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|} dx dy + \int_{\mathbb{R}^3} e_{xc}(\rho(x)) dx \right. \\ \left. \varphi_i \in H^1(\mathbb{R}_{\Sigma}^3), \int_{\mathbb{R}_{\Sigma}^3} \overline{\varphi_i(z)} \varphi_j(z) dz = \delta_{ij}, \rho(x) = \sum_{i=1}^N \sum_{s \in \Sigma} |\varphi_i(x, s)|^2 \right\}.$$

Although LDA leads to one of the simplest types of DFT energy functionals, it is considered “the mother of all approximations” [139] and even here the resulting mathematical properties are still far from being well understood.

Proving the existence of minimizers is made difficult by the non-convexity of the problem due to the LDA term. Using concentration-compactness techniques, introduced by Lions [115,116], it became possible to prove the existence of minimizers in several cases. Le Bris [93] proved that for a neutral or positively charged system, the Kohn-Sham problem with LDA exchange-correlation energy admits a minimizer. Anantharaman and Cancès [2] generalized this to the so-called extended Kohn-Sham model with LDA exchange-correlation energy and also GGA exchange-correlation in the one orbital case.

Additionally, the question of existence and uniqueness of finite temperature Kohn-Sham equation has been studied in [142]. For a discussion of the Kohn-Sham equations in the context of crystals, see the works of E and Lu [38–41].

Beyond these existence results, little is known in regards to, e.g., uniqueness of solutions or compactness of the various operators associated with the Kohn-Sham equation. Also, in contrast to more conventional macroscopic continuum models, it is not clear how the physical nature of the underlying material, for example, whether it is a metal or an insulator, is reflected at the mathematical level [41].

The first mathematically rigorous justification of the LDA approach, in the appropriate regime where  $\rho$  is flat in sufficiently large regions of  $\mathbb{R}^3$ , was given in [107]. Here, the authors provided a quantitative estimate on the difference between Levy-Lieb energy functional in the the grand-canonical setting of a given density and the integral over the Uniform Electron Gas energy of this density.

The following section shortly puts the results from Core Articles I and III into context with related results from the literature. For a more thorough discussion of the techniques used as well as our own contribution to these papers, see Appendices A.1 and A.3.

### 3.3 Contributions in the Analysis of Density Functional Theory and Related Literature

Our analysis concentrates on the question, whether two main properties, which we have encountered in the quantum mechanical setting in Chapter 2, carry over to the DFT setting. As we have seen in this chapter, due to the Hohenberg-Kohn theorem, if we could use the exact but unknown exchange-correlation functional, all properties of full quantum mechanics would persist in DFT, as there is no approximation.

What we are interested in is the question: If we consider the approximations used in practice, in particular the LDA approximations, can we still say the same? The two properties investigated in our work are the existence of excited states and the dissociation limit, i.e., the analogon of Theorem 2.4 and Theorem 2.5, respectively.

#### 3.3.1 Excitations in Density Functional Theory

Electronic excitations play an important role in the description of molecular properties such as absorption spectra, photoexcitation, state-to-state transition probabilities, reactivity, charge transfer processes, and reaction kinetics [30,31,77]. Therefore, improved understanding of excited states and their properties is essential in any electronic structure theory.

That said, we are not aware of any previous rigorous results on excitations in KS-DFT. Thus in Core Article I, we mathematically analyze the simplest such excitations, HOMO-LUMO transitions, defined below, in the setting of the local density approximation (LDA). Treatment of the whole excitation spectrum, as well as of many-body corrections like the Casida ansatz, lie beyond the scope of our investigation.

In this transition, an electron pair migrates from the *highest occupied molecular orbital* (HOMO) to the *lowest unoccupied molecular orbital* (LUMO). For the KS-orbitals  $\Phi = (\varphi_1, \dots, \varphi_n)$  ordered by the size of their eigenvalue, this means

$$(\varphi_1, \dots, \varphi_{n-1}, \varphi_n) \longrightarrow (\varphi_1, \dots, \varphi_{n-1}, \varphi_{n+1}), \quad (3.23)$$

where  $\varphi_n$  is the HOMO and  $\varphi_{n+1}$  – the eigenstate corresponding to the next higher eigenvalue of the KS Hamiltonian – is the LUMO.

For a systematic comparison of HOMO-LUMO excitations with experimental data, see, e.g., [4,195].

To define HOMO and LUMO in a variational way, we consider the *excitation energy functional* [57] given by the quadratic form associated with KS Hamiltonian  $H_{KS}$ , which allows for a convenient mathematical analysis of optimal excitations irrespective of degeneracies.

For positively charged systems (i.e., total nuclear charge  $Z$  greater than the number  $N$  of electrons), we prove such excitations exist, under realistic assumptions on the exchange-correlation functional, which we verify explicitly for the widely used PZ81 [141] and PW92 [140] functionals. By contrast, the neutral case  $Z = N$  holds a surprise. In the cases of the hydrogen and helium atoms, we prove that excited states do not exist when the self-consistent KS ground state density

is replaced by a realistic but easier to analyze closed-form approximation (in case of hydrogen, the true Schrödinger ground state density). Figure 3.3 summarizes these results.

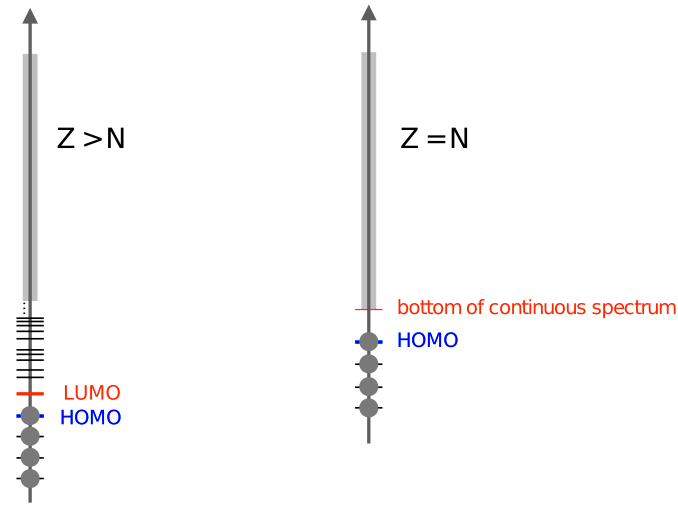


Figure 3.3: Schematic picture of the spectrum of the KS Hamiltonian [54]. Positively charged systems (left,  $Z > N$ ) have infinitely many excited states above the HOMO and below the continuous spectrum. For neutral systems (right,  $Z = N$ ), it can happen that there are no excited states, that is, the highest bound state eigenvalue is the HOMO

The quadratic functional minimized by excitations can be viewed as an approximation to the Kohn-Sham energy functional. As already pointed out in Section 3.1, the latter is closely related to the Thomas-Fermi-Dirac-von Weizsäcker functional, for which an interesting nonexistence result of minimizers was proved via completely different methods in [119].

Physically, these results indicate a significant artefact of KS-DFT. In the full  $N$ -electron Schrödinger equation, neutral systems (and even systems with  $Z > N - 1$ ) are known to possess infinitely many excited states below the bottom of the continuous spectrum, see Section 2.4. The analogous result also holds in Hartree-Fock theory: For  $Z > N$  the Fock operator associated with the Hartree-Fock ground state density possesses infinitely many bound states below the continuous spectrum [117, Lemma II.3], the latter being the interval  $[0, \infty)$ . It is also known [105] that the Hartree-Fock energy functional possesses infinitely many critical points below 0. Our results suggest that in KS-DFT, the threshold for existence of infinitely many excited states is shifted from  $Z > N - 1$  to  $Z > N$ . This is a previously unnoticed but important qualitative consequence of the (well known) incomplete cancellation of the self-interaction energy in KS-DFT.

It is interesting to interpret the nonexistence of excitations from the point of view of numerical computations in finite basis sets or mathematical analysis (as in [57]) in bounded domains. Consider a neutral system for which (exact) excitations do not exist. In a finite basis set, or a bounded domain, the spectrum of the KS Hamiltonian is purely discrete and therefore excited states exist. In the limit as the basis set approaches completeness, or the domain approaches the whole of  $\mathbb{R}^3$ ,

- (i) the LUMO energy  $\varepsilon_L$  (the lowest unoccupied eigenvalue of the KS Hamiltonian) will remain well-defined, and approaches the bottom of the continuous spectrum which equals 0,

- (ii) the LUMO (i.e., the lowest unoccupied eigenstate) will become more and more delocalized, failing to converge to a bound state.

Thus, in contrast to common (explicit or implicit) belief, restriction to finite basis sets or bounded domains may be not just a negligible technicality, but significantly alters the physical nature of LUMO excitations, from a stable bound state (i.e., invariant under the dynamics of the KS ground state Hamiltonian) to a delocalized, dispersing state associated with the continuous spectrum. Point (ii) makes it very tempting to physically interpret the HOMO-LUMO excitation in the nonexistence case as an (approximation to an) ionization process. This interpretation together with (i) yields *ionization potential*  $\approx \varepsilon_L - \varepsilon_H = 0 - \varepsilon_H$  (where  $\varepsilon_H$  is the HOMO energy, i.e. the highest occupied eigenvalue of the KS Hamiltonian), lending new theoretical support to the famous semi-empirical formula

$$-\varepsilon_H \approx \textit{ionization potential}$$

which often agrees quite well with experimental data [4, 195].

### 3.3.2 Dissociation Limit in Kohn-Sham DFT

As mentioned above, here we deal with the question whether or not the energy of a molecule dissociates correctly in KS-DFT with the LDA exchange-correlation functional. Understanding such problems and the precise electron configurations is important for further developing density functional theory [26, 27]. Our main result takes the following form.

**Theorem 3.2** (Main Theorem of [11] – Informal Version)

Let  $E_{2N,R}^{X_2}$  and  $E_\lambda^X$  be the energy of the  $X_2$ -molecule with distance  $R$  between the atoms and of the  $X$ -atom with  $\lambda$  electrons, respectively. Then we have

$$\lim_{R \rightarrow \infty} E_{2N,R}^{X_2} = \min_{\alpha \in [0, N]} (E_\alpha^X + E_{2N-\alpha}^X). \quad (3.24)$$

Thus, in contrast to the classical dissociation, here we can in general only prove that in the limit the system splits into two independent subsystems with their individual electron mass summing up to the one of the original system. One would expect from physical intuition that the optimal splitting occurs for the symmetric case, i.e., the electrons are distributed evenly over the subsystems. As we study in our paper, this is not always true, but rather which splitting is the most stable depends on the strength of the exchange. We quantify this by analyzing our result more deeply for the Dirac exchange by varying the “strength” of the exchange, i.e., the constant  $c_{xc}$ . As it turns out, if  $c_{xc}$  becomes too large, then symmetry splitting takes place, in the sense that the minimum in (3.24) is not attained at  $N$ . Figure 3.4 depicts our numerical results for the hydrogen case, obtained using the OCTOPUS package [3].

These issues in LDA-DFT and related theories like Thomas-Fermi-Dirac-von Weizsäcker are caused by the Dirac term  $-\int \rho^{4/3}$ , which to some extent makes the functional concave and can thus lead also to nonattainment, see [119].

Similar observations were made in case of the  $H_2$  molecule at fixed bond-length, see [80]. They show that for fixed electron mass, the structure of the minimizing Kohn-Sham solutions change



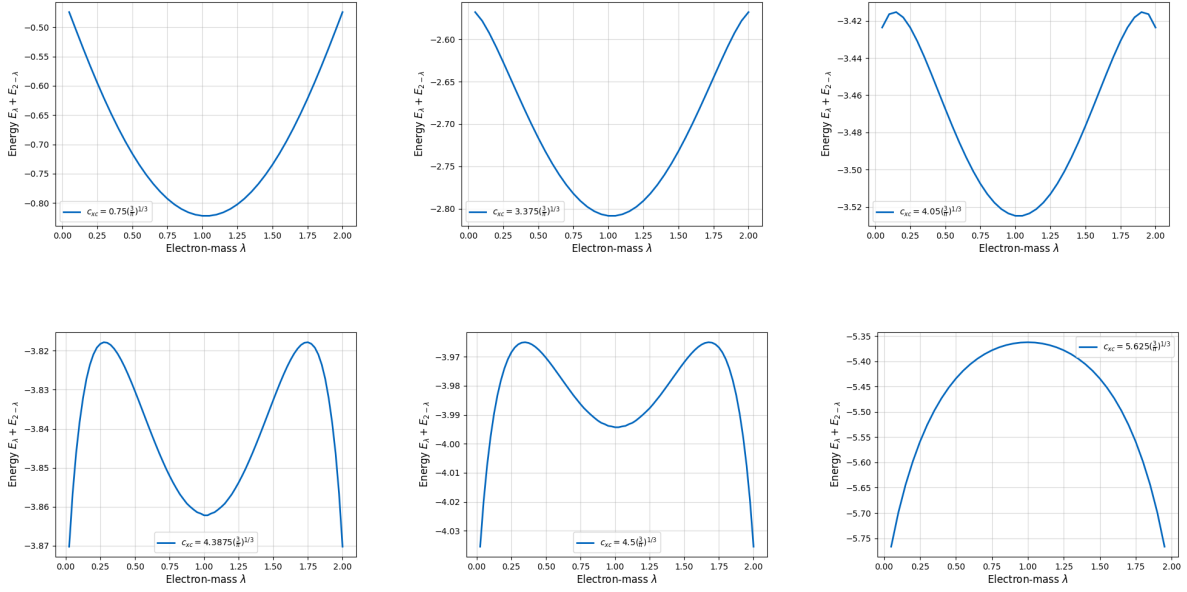


Figure 3.4: The function  $\lambda \mapsto E_\lambda^H + E_{2-\lambda}^H$  for increasing values of  $c_{xc}$ . Note that the plot in the top left corner corresponds to the physically interesting case of  $c_{xc} = \frac{3}{4} \left(\frac{3}{\pi}\right)^{1/3}$ ; here we get numerically a symmetric splitting.

character with the variation of  $c_{xc}$  as the parameters related to the relative strength of the exchange-correlation component of the functional.

Additionally, similar results for periodic systems were found in [63]. Here, they prove symmetry breaking in the Kohn-Sham model for a crystal with a large Dirac exchange coefficient. In [148] similar observations were made for the periodic TFDW model. For a discussion about the related question of binding of atoms and stability of molecules in Hartree and Thomas–Fermi type theories, see the paper series by Catto and Lions [19–22] as well as Lieb [108].

In general, symmetry breaking in quantum mechanical systems has been observed in various settings like polaron models [50, 51] or in Hartree-Fock models of atoms [61, 62]. In the physics literature, difficulties with dissociation calculations in DFT are a prominent topic [5, 135, 151, 173], but rigorous results in the general case are still lacking.

We are only aware of one other setting in the mathematical literature where dissociation is rigorously discussed in the full limit  $R \rightarrow \infty$ , namely [24]. Here they consider the strictly correlated electrons (SCE) limit [156–158]. Opposite to the independence ansatz in (3.19), the pair density is here given by

$$\rho_2^{SCE}(x, y) = \frac{1}{2N} \sum_{i \neq j} \int_{\mathbb{R}^3} \rho(z) \delta(x - T_i(z)) \delta(y - T_j(z)) dz,$$

with  $T_i : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ ,  $i = 1, \dots, N$ , being certain optimal transport maps. This connection of DFT and optimal transport has been an interesting and fruitful discovery [28, 29, 58]. Here, the system dissociates correctly, which is easier to prove since the difficult lower bound is obtained in the same way as in the quantum mechanical case, due to the only additional term being  $V_{ee}^{SCE}[\rho_2^{SCE}]$ , which is positive.



## Chapter 4

# Tensors and the Quantum Chemistry - Density Matrix Renormalization Group

*All truths are easy to understand once they are  
discovered; the point is to discover them.*

---

Galileo Galilei

This final chapter covers the second type of method for solving electronic structure problems discussed in this thesis, the *Quantum Chemistry–Density Matrix Renormalization Group* (QC-DMRG). While DFT is one of the main methods for large weakly correlated systems, in the case of many strongly correlated electrons, there has been no clear choice for a method which provides a sufficiently accurate, data-sparse representation of the exact many-body wavefunction [169]. This is where QC-DMRG seems to provide promising results. As the name suggests, this tensor method was inspired by the Density Matrix Renormalization Group (DMRG), one of the most efficient algorithms for numerical treatment of one-dimensional spin chain systems [128], introduced by White [186, 187]. The QC-DMRG method allows for very accurate computations, but is limited to small systems (around  $\sim 50$  electrons) due to its high computational cost. For some recent implementations of this algorithm see [86, 101, 131, 193].

As was discovered later on [35, 133], DMRG operates on a highly interesting class of quantum states called *matrix product states* (also known as *tensor-trains* (TT) in mathematics). The main idea for MPS consists in factorizing a tensor with  $L$  indices into a chain-like product of tensors of order 3, i.e., matrices depending on an additional physical index, see Figure 4.3 below. These physical indices correspond to the indices from the original tensor, while the other two, the so-called virtual indices, give the matrix structure and are contracted over, see (4.4). Even though it is well known that such a factorization always exists, the caveat is that, in general, the size of the involved matrices (called the bond dimension) grows exponentially with the system size  $L$  [153].

Contrary to the origin of MPS in spin physics, in quantum chemistry the sites are not identical but correspond to molecular orbitals of the system, making the situation more intricate, as will be discussed at the end of this chapter.

We start by shortly recalling what a tensor is, then describe how to obtain the tensor structure from our quantum chemical wavefunction. After precisely defining the framework needed for MPS, we discuss the problem of how to choose a good tensor network to approximate the states, which in case of a MPS boils down to how to order the underlying basis orbitals.

As in the previous chapter, we conclude with a detailed discussion of our own contributions to some open problems in this research area.

## 4.1 What Is a Tensor?

Oversimplified, tensors are just an array of numbers organized by multiple indices. Each entry of the tensor is specified by fixing values for the indices. In this thesis, we will usually denote a tensor by the letter  $C$  and its entries, specified by fixing each index  $i_1, \dots, i_d$ , by  $c_{i_1, \dots, i_d}$ . The order of a tensor is then the number of indices. For example, a scalar  $c$  has no indices, so it is a tensor of order zero. Similarly, a matrix with entries  $c_{ij}$  is a tensor of order two, see Figure 4.1. This is not limited to finite dimension, i.e., a univariate function can also be considered to be a tensor of order one, with the corresponding tensor of order  $d$  being a multivariate function of  $d$  variables.

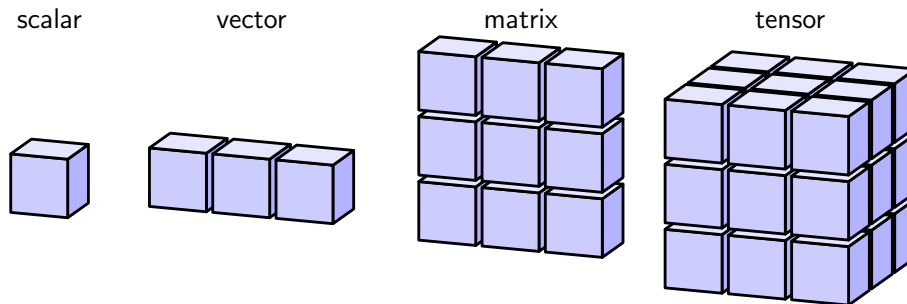


Figure 4.1: Graphical representation of a scalar, a vector, a matrix, and a tensor of order 3. Each small cube in the figure represents one entry of the tensor.

Of course, tensors are not just grids of numbers. Like matrices, they are algebraic objects with certain structures. Let us define tensors as elements of a tensor product space.

**Definition 4.1** (The tensor product)

Let  $V_i$  be vector spaces over a field  $\mathbb{K}$ . The tensor product  $\mathbf{V} := \bigotimes_{i=1}^d V_i = V_1 \otimes \dots \otimes V_d$  consists of the linear span over  $\mathbb{K}$  of all elements

$$v^{(1)} \otimes \dots \otimes v^{(d)}, \quad \text{where } v^{(i)} \in V_i. \quad (4.1)$$

The notation  $\otimes$  denotes the quotient of ordered tuples  $(v^{(1)}, \dots, v^{(d)})$ , by relations

$$\begin{aligned} (\lambda v^{(1)}) \otimes \dots \otimes v^{(d)} &= \lambda(v^{(1)} \otimes \dots \otimes v^{(d)}), \\ \text{and } (v^{(1)} + w^{(1)}) \otimes \dots \otimes v^{(d)} &= v^{(1)} \otimes \dots \otimes v^{(d)} + w^{(1)} \otimes \dots \otimes v^{(d)}, \end{aligned}$$

where the vector  $v^{(i)} \in V_i$ , the vector  $w^{(i)} \in V_1$ , and the scalar  $\lambda \in \mathbb{K}$ . The analogous equations on the other vector spaces  $V_2, \dots, V_d$ , also hold.

Any product of the form (4.1) is called an elementary tensors. From here, we can define a general tensor, see also Figure 4.2.

**Definition 4.2** (General tensor and rank)

A tensor in  $\bigotimes_{i=1}^d V_i$  has rank one if it can be written as an elementary tensor. A general tensor is a sum of rank one tensors

$$C = \sum_{i=1}^r v_i^{(1)} \otimes \cdots \otimes v_i^{(d)}, \quad \text{where } v_i^{(j)} \in V_j.$$

The smallest number of rank one tensors that sum to  $C$  is called rank of  $C$ .

Some remarks are in order: At first glance, this is just the generalization of the rank known from matrices. While this is true there are some caveats in the case  $d \geq 3$ . First, finding the rank of a given tensor and thus the above decomposition – known as CP (canonical polyadic) decomposition – is in general NP-hard [88]. Second, the rank is not lower semi-continuous, i.e., the set of tensors with rank smaller or equal to a given constant  $\tilde{r}$  is not closed. This is known as the border rank problem [67, 92]. An easy and explicit example is due to De Silv and Lim [32]: It is straightforward to see that the tensors of rank two given by

$$A_n := n(x_1 + \frac{1}{n}y_1) \otimes (x_2 + \frac{1}{n}y_2) \otimes (x_3 + \frac{1}{n}y_3) - nx_1 \otimes x_2 \otimes x_3$$

converge for  $n \rightarrow \infty$  to

$$A := x_1 \otimes x_2 \otimes y_3 + x_1 \otimes y_2 \otimes x_3 + y_1 \otimes x_2 \otimes x_3,$$

with  $A$  having rank 3.

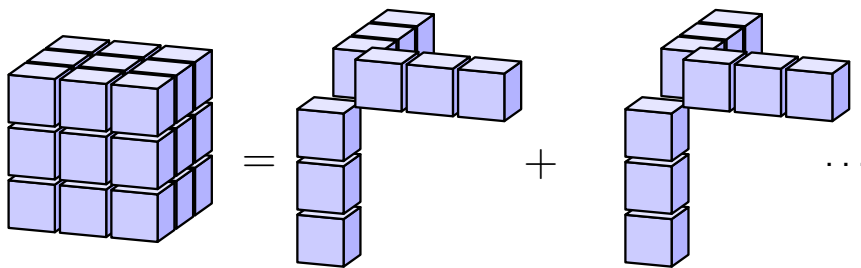


Figure 4.2: Graphical representation of a general tensor build from a sum of elementary tensors.

Furthermore, note that this definition covers algebraic tensor spaces  $\mathbf{V}_{\text{alg}}$ . In our case, the  $V_i$  will be normed spaces and we are interested in the topological tensor space  $\mathbf{V}_{\text{top}}$ , which is defined as the closure of  $\mathbf{V}_{\text{alg}}$  with respect to a chosen norm. We will only be dealing with the finite-dimensional case, where all norms are equivalent and thus both notions agree. Let us just mention that the choice of the norm in infinite dimensions is more subtle; in particular, the norm on the tensor space is not fixed by a choice for the norms of the underlying spaces  $V_i$ .

For more detail, we refer to textbooks on this topic, such as [67], [143] and [92]. The first one is most suited for the context of this thesis, as it focuses on the functional analysis of tensors with

numerical treatment of hierarchical tensor decompositions, with applications to solving partial differential equations. The second one deals with spectral theory of tensors with applications to hypergraph theory, whereas the last one presents the algebra point of view.

#### 4.1.1 Symmetric and Antisymmetric Tensor Spaces

Next, let us define the antisymmetric tensor product, which we will need to incorporate the Pauli exclusion principle (2.18). Here, tensors are associated with coinciding vector spaces  $V_j$  denoted by  $V$ :

$$V := V_1 = \dots = V_d.$$

Then, for any permutation  $\pi \in P_d$ , where  $P_d$  denotes the set of bijections of the set  $\{1, \dots, d\}$  onto itself, we obtain a linear map, again denoted by  $\pi$ ,  $\pi : \mathbf{V} \rightarrow \mathbf{V}$  via

$$\pi \left( \bigotimes_{i=1}^d v^{(i)} \right) = \bigotimes_{i=1}^d v^{(\pi^{-1}(i))}.$$

Each permutation  $\pi$  is a (possibly empty) product of transpositions:  $\pi = \pi_{\nu_1 \mu_1} \circ \dots \circ \pi_{\nu_k \mu_k}$  with  $\nu_i \neq \mu_i$  ( $1 \leq i \leq k$ ). The number  $k$  determines the parity  $\pm 1$  of the permutation  $\pi$ :  $\text{sign}(\pi) = (-1)^k$ . With this we can give the following definition.

**Definition 4.3** ((Anti-)symmetric tensor spaces)

A tensor  $v \in \mathbf{V}$  is called symmetric if  $\pi(v) = v$  for all permutations and antisymmetric if  $\pi(v) = \text{sign}(\pi)v$ . This defines the (anti)symmetric tensor space:

$$\mathbf{V}_{\text{sym}} := \{v \in \mathbf{V} : \pi(v) = v\}, \quad (4.2)$$

$$\mathbf{V}_{\text{anti}} := \{v \in \mathbf{V} : \pi(v) = \text{sign}(\pi)v\} \quad (4.3)$$

In the following, to emphasize the parameter  $d$ , we will use the notation  $\bigwedge_{j=1}^d V$  for the antisymmetric tensor space, where  $\wedge$  denotes the exterior product.

#### 4.1.2 The Tensor-Train Decomposition

In the following, we want to consider a decomposition of the tensor  $C$  of the form

$$C(i_1, \dots, i_d) = A_1[i_1]A_2[i_2] \cdots A_d[i_d] \quad (4.4)$$

$$= \sum_{\alpha_1=1}^{r_1} \sum_{\alpha_2=1}^{r_2} \cdots \sum_{\alpha_{d-1}=1}^{r_{d-1}} A_1(1, i_1, \alpha_1)A_2(\alpha_1, i_2, \alpha_2) \cdots A_d(\alpha_{d-1}, i_d, 1), \quad (4.5)$$

where the  $A_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$  are tensors of order 3. Note that for all  $k = 1, \dots, d$  we have  $n_k \geq 1$  and moreover  $1 \leq r_k \leq n_k$ , as well as the convention  $r_0 = r_d = 1$ .

The representation in (4.4) is called a tensor-train (TT) decomposition, also known as matrix product states (MPS) in physics. Moreover, the numbers  $r_k$  are called the TT-ranks. The order-3 tensors  $A_k$  are called the TT components. Figure 4.3 gives a visualization and shows the underlying ‘‘train-structure’’.

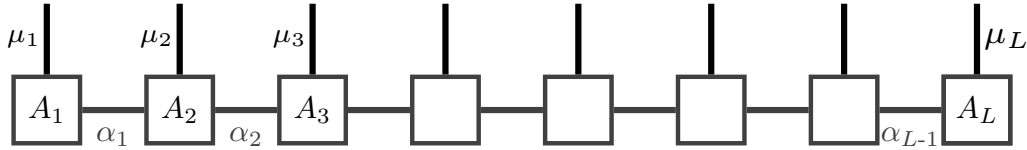


Figure 4.3: Graphical representation of a tensor-train decomposition in the finite-dimensional case; the virtual indices  $\alpha_j$  are contracted over.

Note, an important property is that, in contrast to the manifold of tensors below certain rank, the TT-manifold with TT-ranks below a certain value is closed, see [67]. Let us give an example to clarify the definition and show, why such a representation could be useful.

Consider the tensor  $C \in \mathbb{R}^{3 \times 4 \times 5}$ , defined by

$$C(i_1, i_2, i_3) = i_1 + i_2 + i_3, \quad i_1 \in [3], i_2 \in [4], i_3 \in [5],$$

where we used the shorthand notation  $[n] := \{1, \dots, n\}$ . Then we can write  $C$  in the following form

$$C(i_1, i_2, i_3) = \begin{pmatrix} 1 & i_1 \end{pmatrix} \cdot \begin{pmatrix} 1 & i_2 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} i_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & i_1 \end{pmatrix} \cdot \begin{pmatrix} i_2 + i_3 \\ 1 \end{pmatrix} = i_1 + i_2 + i_3.$$

Thus by defining the matrices

$$\begin{aligned} A_1[i_1] &= \begin{pmatrix} 1 & i_1 \end{pmatrix} \in \mathbb{R}^{1 \times 2}, \quad i_1 \in [3], \\ A_2[i_2] &= \begin{pmatrix} 1 & i_2 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2 \times 2}, \quad i_2 \in [4], \\ A_3[i_3] &= \begin{pmatrix} i_3 \\ 1 \end{pmatrix} \in \mathbb{R}^{2 \times 1}, \quad i_3 \in [5], \end{aligned}$$

we found a MPS representation of our tensor

$$C(i_1, i_2, i_3) = A_1[i_1]A_2[i_2]A_3[i_3].$$

Note that the tensor  $C$  has  $3 \cdot 4 \cdot 5 = 60$  elements, while the TT decomposition uses only  $(1 \cdot 2 \cdot 3) + (2 \cdot 2 \cdot 4) + (2 \cdot 1 \cdot 5) = 32$  elements to represent it. So, we almost need only half the storage for the TT decomposition. In general, if we define  $R := \max\{r_1, \dots, r_{d-1}\}$  and  $N := \max\{n_1, \dots, n_d\}$ , then the number of matrix elements we have to store for our MPS representation is bounded by

$$M = \sum_{k=1}^d n_k \cdot (r_{k-1} \cdot r_k) \leq d \cdot N \cdot R^2.$$

In particular, we expect that  $M \ll n_1 \cdot n_2 \cdots n_d$ .

Next, we want to define the important concept of an unfolding of a tensor.

**Definition 4.4** (Unfolding matrices)

Let  $C \in \mathbb{R}^{n_1 \times \dots \times n_d}$  be a tensor of order  $d$ , and let

$$\nu_k := \prod_{s=1}^k n_s \quad \text{and} \quad \mu_k := \prod_{s=k+1}^d n_s, \quad k = 1, \dots, d-1.$$

Then the  $k^{\text{th}}$  unfolding matrix  $C_k = C_{i_{k+1} \dots i_d}^{i_1, \dots, i_k} \in \mathbb{R}^{\nu_k \times \mu_k}$  of  $C$  is defined by

$$C((i_1, \dots, i_k); (i_{k+1}, \dots, i_d)) := C(i_1, \dots, i_d), \quad i_s \in [n_s], \quad s = 1, \dots, d,$$

where the indices  $(i_1, \dots, i_k)$  enumerate the rows, and  $(i_{k+1}, \dots, i_d)$  enumerate the columns of the matrix  $C_k$  in colexicographic order, i.e., in column-major order.

Note that this is usually implemented in programming languages like MATLAB or julia by a single call of the reshape function:

$$C_k = \text{reshape}(C, \nu_k, \mu_k).$$

As another example, consider the tensor  $C \in \mathbb{R}^{2 \times 2 \times 2}$ , defined by

$$C(i_1, i_2, i_3) = i_1 \cdot 10^2 + i_2 \cdot 10 + i_3, \quad i_1, i_2, i_3 \in [2].$$

Then, the unfolding matrices  $C_1 \in \mathbb{R}^{2 \times 4}$  and  $C_2 \in \mathbb{R}^{4 \times 2}$  are given by

$$C_1 = \begin{pmatrix} 111 & 121 & 112 & 122 \\ 211 & 221 & 212 & 222 \end{pmatrix}, \quad C_2 = \begin{pmatrix} 111 & 112 \\ 211 & 212 \\ 121 & 122 \\ 221 & 222 \end{pmatrix}.$$

**Lemma 4.5** (TT-rank equals separation rank [81], [67])

Let  $C \in \mathbb{C}^{n_1 \times \dots \times n_d}$  be an arbitrary tensor of order  $d$ . For each unfolding matrix  $C_k$  let

$$r_k = \text{rank}(C_k), \quad k = 1, \dots, d-1.$$

Then,  $C$  admits a TT-decomposition with  $A_k$  of size  $n_k \times r_k$  and  $A_{k+1}$  of size  $r_k \times n_{k+1}$ , for some  $n, m \in \mathbb{N}$ , i.e., with TT-ranks not higher than  $r_k$ . Also, there exists a TT-decomposition with all  $r_k$  being simultaneously minimal.

Before we describe the algorithm to construct a TT representation for any given tensor  $C$ , we want to define an important property of the involved tensor components  $A_k$  of order 3. Note that we only define the left-normalized version, but right-normalized tensor are defined in an analogous way. Furthermore, sometimes this property is also called left-orthogonal.



**Definition 4.6** (Left-normalized)

Let  $A_k \in \mathbb{C}^{r_{k-1} \times n_k \times r_k}$  be an tensor of order 3. Then,  $A_k$  is said to be left-normalized, if

$$\sum_{i_k=1}^{n_k} A_k[i_k]^\dagger A_k[i_k] = \text{Id}_{r_k \times r_k}.$$

Now, we present the higher-order singular value decomposition algorithm, which takes an arbitrary tensor  $C$  of order  $d$  as input and constructs a left-normalized TT decomposition, see [132]. Note that the matrix size which this algorithm produces grows in general exponentially.

**Algorithm 1** Higher-order singular value decomposition (HOSVD)

**Input:** A tensor  $C \in \mathbb{C}^{n_1 \times \dots \times n_d}$  of order  $d \geq 1$  ;

**Output:** Matrices  $A_k[i_k] \in \mathbb{C}^{r_{k-1} \times r_k}$ ,  $i_k \in [n_k]$ ,  $k = 1, \dots, d$ , such that

$$C(i_1, \dots, i_d) = A_1[i_1] \cdots A_d[i_d].$$

- 1: Compute the unfolding  $C_1 \in \mathbb{C}^{n_1 \times \mu_1}$  with  $\mu_1 = n_2 \cdots n_d$ ;
- 2: Perform a (thin) SVD of  $C_1$ :

$$C_1 = U_1 \Sigma_1 V_1^T,$$

with  $U_1 \in \mathbb{C}^{n_1 \times r_1}$ ,  $\Sigma_1 \in \mathbb{C}^{r_1 \times r_1}$ ,  $V_1 \in \mathbb{C}^{\mu_1 \times r_1}$  and  $r_1 = \text{rank}(C_1)$ .

- 3: For  $i_1 \in [n_1]$ , set  $A_1[i_1] : (1, j) \mapsto U_1(i_1, j)$ , for all  $j \in [r_1]$ , that is to say we obtain  $A_1$  by extracting the rows from  $U_1$ , i.e.

$$U_1 = \begin{pmatrix} A_1[1] \\ A_1[2] \\ \vdots \\ A_1[n_1] \end{pmatrix};$$

- 4: Compute the matrix  $R_1 = \Sigma_1 V_1^T \in \mathbb{C}^{r_1 \times \mu_1}$ ;
- 5: For  $k = 2, \dots, d - 1$ :
  - 5.1: Set  $\mu_k := n_{k+1} \cdots n_d$  and perform a reshape of the matrix  $R_{k-1}$ :

$$L_k = \text{reshape}\left(R_{k-1}, r_{k-1} \cdot n_k, \mu_k\right) \in \mathbb{C}^{r_{k-1} n_k \times \mu_k};$$

- 5.2: Perform a (thin) SVD of  $L_k$

$$L_k = U_k \Sigma_k V_k^T,$$

with  $U_k \in \mathbb{C}^{r_{k-1} n_k \times r_k}$ ,  $\Sigma_k \in \mathbb{C}^{r_k \times r_k}$ ,  $V_k \in \mathbb{C}^{\mu_k \times r_k}$  and  $r_k = \text{rank}(L_k)$ .

- 5.3: for  $i_k \in [n_k]$ , set

$$A_k[i_k] : (j_1, j_2) \mapsto U_k((i_k - 1) \cdot r_{k-1} + j_1, j_2),$$

for all  $j_1 \in [r_{k-1}]$  and  $j_2 \in [r_k]$ ;

- 5.4: Compute the matrix  $R_k = \Sigma_k V_k^T \in \mathbb{C}^{r_k \times \mu_k}$ ;

- 6: For  $i_d \in [n_d]$ , set  $A_d[i_d] : (j, 1) \mapsto R_{d-1}(j, i_d)$  for all  $j \in [r_{d-1}]$ .

**Comment:** One can obtain a left-orthogonal TT decomposition by replacing the SVD by QR decompositions.

We remark that a TT decomposition is never unique, since the product of two consecutive matrices  $A_k[i_k]$  and  $A_{k+1}[i_{k+1}]$  can always be replaced by

$$A_k[i_k]A_{k+1}[i_{k+1}] = A_k[i_k]B_kB_k^{-1}A_{k+1}[i_{k+1}],$$

where  $B_k$  is an arbitrary invertible matrix in  $\mathbb{C}^{r_k \times r_k}$ . For numerical purposes, it is thus important to set a gauge condition on the TT matrices  $A_k[i_k]$ . The following Lemma takes care of exactly that.

**Lemma 4.7** (Theorem 1 from [81])

Let  $C \in \mathbb{C}^{n_1 \times \dots \times n_d}$ . The TT decomposition (4.4) of  $C$  of minimal rank can be chosen such that the TT components are left-orthogonal for all  $k \in [d-1]$ .

Under this condition, the decomposition in (4.4) is unique up to insertion of orthogonal matrices: For any two left-normalized TT decompositions of  $C$

$$C(i_1, \dots, i_d) = A_1[i_1] \cdots A_d[i_d] = B_1[i_1] \cdots B_d[i_d],$$

there exists orthogonal matrices  $Q_1, \dots, Q_d \in \mathbb{C}^{r_k \times r_k}$  such that

$$\begin{aligned} A_1[i_1]Q_1 &= B_1[i_1], & Q_{d-1}^T A_d[i_d] &= B_d[i_d], \\ Q_{k-1}^T A_k[i_k]Q_k^T &= B_k[i_k] & \text{for } k &= 2, \dots, d-1. \end{aligned}$$

The analogous statement can be proved for right-orthogonal TT decompositions.

It can sometimes be advantageous to deal with TT representations that have a mixed left and right orthogonalisation: e.g., a TT decomposition

$$C(i_1, \dots, i_d) = A_1[i_1] \cdots A_d[i_d],$$

where  $A_1, \dots, A_k$  are left-normalized and  $A_{k+1}, \dots, A_d$  are right-normalized.

Finally, there is a way to write the tensor-train representation of a tensor in order to retrieve easily a left- or right-normalized TT decomposition. Namely, by keeping the singular value matrices of the HOSVD (Algorithm 1) of the tensor  $C$ , one can achieve a decomposition of the following type:

$$C(i_1, \dots, i_d) = \Gamma_1[i_1]\Sigma_1\Gamma_2[i_2]\Sigma_2 \cdots \Gamma_d[i_d]\Sigma_d,$$

where the matrices  $\Gamma_k \in \mathbb{R}^{r_{k-1} \times r_k}$  satisfy

$$\begin{aligned} \sum_{i_1=1}^{n_1} \Gamma_1[i_1]^T \Gamma_1[i_1] &= \text{Id}_{r_1 \times r_1}, & \sum_{i_d=1}^{n_d} \Gamma_d[i_d] \Gamma_d[i_d]^T &= \text{Id}_{r_d \times r_d}, \\ \sum_{i_k=1}^{n_k} \Gamma_k[i_k]^T \Sigma_{k-1}^2 \Gamma_k[i_k] &= \text{Id}_{r_k \times r_k}, & \sum_{i_k=1}^{n_k} \Gamma_k[i_k] \Sigma_k^2 \Gamma_k[i_k]^T &= \text{Id}_{r_{k-1} \times r_{k-1}}, \quad k \in 2, \dots, d-1. \end{aligned}$$

This representation is called the standard representation or the HSVD (hierarchical SVD) representation of the tensor  $C$ . In physics, it is attributed to Vidal and called the Vidal representation of the tensor  $C$  [183].

We want to mention that it is possible to convert between all these different representations. But since we will only be using the left-normalized decomposition in this thesis, we simply refer the reader to [153] for these intertranslations.

## 4.2 Where Does the Tensor Come from in Quantum Chemistry?

After this more abstract introduction, let us now come back to quantum mechanics and discuss how the tensor arises from the wavefunction. Effectively, we will expand the full wavefunction into a linear combination of Slater determinants and then make a cutoff in the number of used orbitals, the associated coefficients will build the entries of our tensor. We will here present the general approach for arbitrary Hilbert spaces following [56].

### 4.2.1 Fock Space and the Occupation Representation

We start our analysis by recalling the fermionic Fock space. First, consider a finite dimensional single-particle Hilbert space  $\mathcal{H}_L$ . When the particles are electrons,  $\mathcal{H}_L$  would correspond to a subspace of  $L^2(\mathbb{R}_\Sigma^3)$  spanned by  $L$  spin orbitals, see Chapter 2. Then, the associated state space for a system of  $N$  fermions is the  $N$ -fold antisymmetric tensor product  $\mathcal{V}_{N,L} := \bigwedge_{i=1}^N \mathcal{H}_L$ . Finally, the resulting Fock space is defined as the direct sum of the  $N$ -particle spaces,

$$\mathcal{F}_L := \bigoplus_{N=0}^L \mathcal{V}_{N,L}, \quad (4.6)$$

where  $\mathcal{V}_{0,L} \cong \mathbb{C}$  is spanned by the vacuum state  $|\Omega\rangle$ .

If the orbitals are the lowest eigenstates of the Hartree-Fock Hamiltonian, resulting from the Euler-Lagrange equation of the Hartree-Fock energy, compare (3.3),  $\mathcal{V}_{N,L}$  is known in physics as the full configuration interaction (full CI) space, see, e.g., [74].

Now given an orthonormal basis  $\{\varphi_1, \dots, \varphi_L\}$  of  $\mathcal{H}_L$ , we can write any element  $\Psi \in \mathcal{F}_L$  in the form

$$\Psi = c_0|\Omega\rangle + \sum_{i=1}^L c_i|\varphi_i\rangle + \sum_{1 \leq i < j \leq L} c_{ij}|\varphi_i\varphi_j\rangle + \sum_{1 \leq i < j < k \leq L} c_{ijk}|\varphi_i\varphi_j\varphi_k\rangle + \dots, \quad (4.7)$$

with  $|\varphi_{i_1} \dots \varphi_{i_N}\rangle$  denoting the antisymmetric tensor product, alias Slater determinant,

$$|\varphi_{i_1} \dots \varphi_{i_N}\rangle = \varphi_{i_1} \wedge \dots \wedge \varphi_{i_N} \in \mathcal{V}_{N,L}. \quad (4.8)$$

Instead of the above 'first quantized' representation, in QC-DMRG one considers a 'second quantized' representation by occupation numbers of orbitals in Fock space. A Slater determinant  $|\varphi_{i_1} \dots \varphi_{i_N}\rangle \in \mathcal{V}_{N,L}$  is represented by a binary string  $(\mu_1, \dots, \mu_L) \in \{0, 1\}^L$ , with  $\mu_i$  indicating whether or not the orbital  $\varphi_i$  is present (occupied) or absent (unoccupied). An example with  $N = 4$  and  $L = 8$  looks like this

$$|\varphi_2\varphi_3\varphi_6\varphi_8\rangle \longleftrightarrow (0, 1, 1, 0, 0, 1, 0, 1). \quad \begin{array}{cccccccc} \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\ \bigcirc & \bullet & \bullet & \bigcirc & \bigcirc & \bullet & \bigcirc & \bullet \\ \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\ \varphi_1 & \varphi_2 & & & & & & \varphi_L \end{array}$$

Here,  $\varphi_1$  is unoccupied,  $\varphi_2$  is occupied,  $\varphi_3$  is occupied, and so on. The Slater determinant (4.8), indexed by its binary label, is in the following denoted  $\Phi_{\mu_1\dots\mu_L}$ , that is to say

$$\Phi_{\mu_1\dots\mu_L} := |\varphi_{i_1} \dots \varphi_{i_N}\rangle \text{ if } \mu_i = 1 \text{ exactly when } i \in \{i_1, \dots, i_N\}, \quad i_1 < \dots < i_N. \quad (4.9)$$

The coefficients in the expansion (4.7), indexed by the corresponding binary label, are called  $C_{\mu_1\dots\mu_L}$ , that is to say

$$C_{\mu_1\dots\mu_L} = c_{i_1\dots i_N} \text{ if } \mu_i = 1 \text{ precisely when } i \in \{i_1, \dots, i_N\}, \quad i_1 < \dots < i_N, \quad (4.10)$$

yielding the occupation representation of the state  $\Psi$  from (4.7)

$$\Psi = \sum_{\mu_1, \dots, \mu_L=0}^1 C_{\mu_1\dots\mu_L} \Phi_{\mu_1\dots\mu_L}. \quad (4.11)$$

This representation might seem more abstract, but it not only allows for a more compact notation, it also is more precise in the sense that in first quantization  $\varphi_1 \wedge \varphi_2$  and  $-\varphi_2 \wedge \varphi_1$  are actually redundant names. Note that if  $\Psi$  is an  $N$ -electron wavefunction, we will use the shorthand notation  $\mathcal{N}\Psi = N\Psi$ , where  $\mathcal{N}$  is the number operator. If this is the case, the coefficients  $C_{\mu_1\dots\mu_L}$  are zero whenever  $\sum_{j=1}^L \mu_j \neq N$ . This simple fact is crucial for the structure of the unfoldings of our coefficient tensor.

## 4.2.2 Tensor Networks and Matrix Product States

The idea behind tensor networks states is to factorize the large coefficient tensor  $C$  into (smaller) tensors of lower order, where the structure of the multiplication is described by a graph or network, see Figure 4.4. We will denote internal or virtual indices, which are just summed over, by  $\alpha_j$ , while for the physical indices we use  $\mu_j$ .

We only consider here matrix product states, they build the simplest subclass of tensor network states (TNS), where the underlying graph is a subset of  $\mathbb{Z}$ . They correspond to the tensor-train decomposition from Section 4.1.2. As already mentioned, the name tensor-train is used in mathematics, while matrix product states is the preferred term in physics. In this thesis, we will usually use MPS if there is some physical context, i.e., if orbital functions are involved.

Note though that if one changes the underlying graph to include cycles, then problems arise and closedness is lost, see, e.g., [7]. For more advanced networks we refer to [67, 169]. For an easy to read full introduction into MPS see [153, 161].

A matrix product state (MPS) with respect to the basis  $\{\varphi_i\}_{i=1}^L$  with size parameters ('bond dimensions')  $r_i$  ( $i = 1, \dots, L - 1$ ) is a state of the form

$$\Psi = \sum_{\mu_1, \dots, \mu_L=0}^1 A_1[\mu_1]A_2[\mu_2]\dots A_L[\mu_L] \Phi_{\mu_1\dots\mu_L} \in \mathcal{F}_L, \quad (4.12)$$

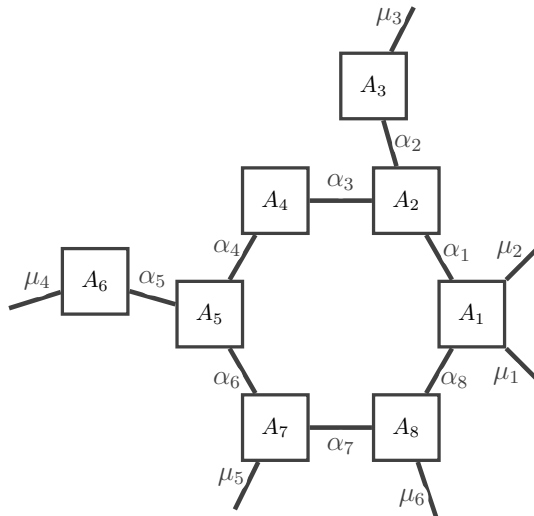


Figure 4.4: A general tensor network representation of a tensor of order 6.

where for every tuple of labels  $(\mu_1, \dots, \mu_L) \in \{0, 1\}^L$ ,  $A_i[\mu_i]$  is a  $r_{i-1} \times r_i$  matrix, with the convention  $r_0 = r_L = 1$ . Hence, the  $A_i$  can be viewed as tensors of order 3 (depending on three indices  $\alpha_{i-1}, \mu_i, \alpha_i$ ) in  $\mathbb{C}^{r_{i-1} \times 2 \times r_i}$ . The name 'bond dimensions' for the  $r_i$  has nothing to do with chemical bonds, but is related to the standard graphical representation of MPS in Figure 4.3, in which each contraction index  $\alpha_i$  is represented by a horizontal 'bond'.

The minimal bond dimensions with which a given state can be represented, have a well known meaning as ranks of matricizations of the coefficient tensor  $C$ , as recalled in Lemma 4.5 above.

The set of matrix product states (MPS) with respect to the basis  $\{\varphi_i\}_{i=1}^L$  with bond dimensions  $r_i$  ( $i = 1, \dots, L - 1$ ) is denoted by

$$\text{MPS}(L, \{r_i\}_i, \{\varphi_i\}_i) \subseteq \mathcal{F}_L. \quad (4.13)$$

Representing arbitrary states in  $\mathcal{F}_L$  as MPS is possible, but requires bond dimensions  $2^{L/2}$  (assuming  $L$  is even.), i.e., bond dimensions growing exponentially with  $L$ , see [153] or Section 4.1.2.

Now, QC-DMRG takes the MPS ansatz and simply computes corresponding approximation to the energy by minimizing the Rayleigh quotient

$$E_0^{QC-DMRG} = \min \left\{ \frac{\langle \Psi, H\Psi \rangle}{\langle \Psi, \Psi \rangle} \mid \Psi \in \text{MPS}(L, \{r_i\}_i, \{\varphi_i\}_i), \Psi \neq 0, \mathcal{N}\Psi = N\Psi \right\}, \quad (4.14)$$

where usually one takes some truncation limit  $M$  for the bond-dimensions, i.e.,  $r_i \leq M$  for all  $i \in \{1, \dots, L\}$ . Note that this parameter  $M$  allows to interpolate between Hartree-Fock ( $M = 1$ ) and full CI ( $M = 2^{L/2}$ ). In state-of-the-art computations,  $M$  usually is chosen to be of the order  $\sim 2000 - 5000$ , see, e.g., [56] for more details.

One major issue in the context of tensor networks is: *How to choose a suitable topology of the underlying network?* In our special case of MPS with fixed one-particle basis  $\{\varphi_1, \dots, \varphi_L\}$ , this boils down to: *How to order the basis?* We want to illustrate this question by means of an example, following [37].



Figure 4.5: Schematic picture of a MPS before and after reordering the orbitals [64].

We want to look at the minimal basis  $H_2$  setting, i.e., take  $N = 2$ ,  $L = 4$ , and consider an  $H_2$  molecule with nuclei clamped at  $R_A$  and  $R_B$ . For the underlying single-particle Hilbert space  $\mathcal{H}_L$ , we take the span of the four functions  $\chi_A(r)\delta_{\uparrow/\downarrow}(\sigma)$  and  $\chi_B(r)\delta_{\uparrow/\downarrow}(\sigma)$ , corresponding to the 1s orbitals of hydrogen (compare Section 2.1), just translated to the clamped nuclei respectively, i.e.,

$$\chi_A(r) = \frac{1}{\sqrt{\pi}}e^{-|r-R_A|}, \quad \chi_B(r) = \frac{1}{\sqrt{\pi}}e^{-|r-R_B|}.$$

To obtain the orthonormal basis of  $\mathcal{H}_L$  which we want to consider, we define the associated bonding respectively antibonding orbitals,

$$\varphi_A(r) = \frac{\chi_A + \chi_B}{\sqrt{2 + 2S_{AB}}}, \quad \varphi_B(r) = \frac{\chi_A - \chi_B}{\sqrt{2 - 2S_{AB}}},$$

where the overlap integral  $S_{AB} = \langle \chi_A, \chi_B \rangle$  is just used for normalization. The corresponding single-particle basis is then  $\{\varphi_A \uparrow, \varphi_A \downarrow, \varphi_B \uparrow, \varphi_B \downarrow\}$ , where we used the obvious short-hand notation to indicate the spin component. Our state of interest is now given by the Slater determinant

$$\Psi = |(c\varphi_A + s\varphi_B) \uparrow, (c'\varphi_A + s'\varphi_B) \downarrow\rangle, \quad (4.15)$$

for some coefficients  $c, s, c', s' \in \mathbb{R}$  with  $c^2 + s^2 = c'^2 + s'^2 = 1$ . As pointed out in [37], this is an interesting state, as the unrestricted Hartree-Fock (UHF) ground state of minimal-basis  $H_2$  has the above form, for any bondlength  $R = |R_A - R_B|$ ; moreover  $(c, s) \neq (c', s')$  when  $R$  is large [168]. The occupation representation of  $\Psi$  takes then the form

$$\Psi = cc'\Psi_{1100} + cs'\Psi_{1001} - sc'\Phi_{0110} + ss'\Phi_{0011} \in \mathcal{F}_L. \quad (4.16)$$

Here, we implicitly used the ordering  $\{\varphi_A \uparrow, \varphi_A \downarrow, \varphi_B \uparrow, \varphi_B \downarrow\}$ . In the following we shortly want to remark on this. As in practice usually the one body basis consists of the low-lying eigenfunctions of the Hartree-Fock operator, the simplest method, which was also used in the early days of QC-DMRG, is to order the orbitals according to their Hartree-Fock eigenvalues. This ordering is known as *canonical order*.

The pioneering article [6] introduced the Fiedler ordering, which combines concepts from quantum information theory and spectral graph theory to achieve significant improvements in the approximations. Here, one starts by computing the mutual information matrix IM and then (as all entries are nonnegative, see [17]) interprets it as a weighted adjacency matrix of the complete graph of the tensor network, in the following way: The second eigenvector of the graph Laplacian  $\mathcal{L}$  is computed, this is the so-called *Fiedler-vector*, and ordering its entries according to its values gives the so-called Fiedler ordering. Another more recent ordering scheme is the so-called best

(weighted) prefactor ordering [37]. Here, the authors show an interesting inversion symmetry for the distribution of singular values, which they utilize to improve the decay of the singular value distribution of the associated unfolding.

These orderings are important as they can significantly lower the necessary bond-dimensions (and thus the truncation limit  $M$  in (4.14)) of the state of interest. We shortly illustrate this via the minimal basis  $H_2$  example, taken from [37]. For more details, especially the explicit calculations, we refer to [37]. To avoid the degenerate case, we assume that the constants  $c, s, c', s'$  in (4.15) are all nonzero. The canonical order is just the ordering in which the bonding orbital with either spin comes first, i.e.,  $\{\varphi_A \uparrow, \varphi_A \downarrow, \varphi_B \uparrow, \varphi_B \downarrow\}$ . Writing out the unfolding of our state in occupation representation (4.16) gives

$$\psi_{\mu_3, \mu_4}^{\mu_1, \mu_2} = \begin{array}{cc} & \begin{array}{cccc} & 00 & 01 & 10 & 11 \end{array} \\ \begin{array}{c} 00 \\ 01 \\ 10 \\ 11 \end{array} & \left[ \begin{array}{cccc} & & & & ss' \\ & & & -sc & \\ & & cs' & & \\ cc' & & & & \end{array} \right]. \end{array}$$

Since all entries are nonvanishing, the matrix re-shape has full rank 4 and due to Lemma 4.5, so does the corresponding bond-dimension.

Applying now the Fiedler or the best (weighted) prefactor ordering, gives in this case the same new labeling for the single-particle basis, namely  $\{\varphi_A \uparrow, \varphi_B \uparrow, \varphi_A \downarrow, \varphi_B \downarrow\}$ . With this re-labeling of the basis, one obtains for the unfolding of our state in occupation representation (4.16)

$$\psi_{\mu_3, \mu_4}^{\mu_1, \mu_2} = \begin{array}{cc} & \begin{array}{cccc} & 00 & 01 & 10 & 11 \end{array} \\ \begin{array}{c} 00 \\ 01 \\ 10 \\ 11 \end{array} & \left[ \begin{array}{cccc} & & & & \\ & & ss' & sc' & \\ & & cs' & cc' & \\ & & & & \end{array} \right]. \end{array}$$

Since the middle block can be written as  $\begin{pmatrix} c \\ s \end{pmatrix} \begin{pmatrix} c' & s' \end{pmatrix}$ , it is just a rank-1 matrix, so our required bond-dimension is minimal. Thus, in conclusion, both re-orderings dramatically improve the bond-dimension necessary to represent the state of interest.

Lastly, let us shortly introduce the MPS setting also for states of the full Fock space, i.e., utilizing infinitely many orbitals. As we will see, this calls for half-infinite matrix product states, which we will introduce in a rigorous manner below. Graphically, this corresponds to a half-infinite chain, see Figure 4.6. This representation is part of our Article IV [56].

So let  $\mathcal{H}$  be an infinite-dimensional separable Hilbert space spanned by orthonormal orbitals  $\{\varphi_i\}_{i=1}^{\infty}$ , let  $\mathcal{V}_N$  be the  $N$ -fold antisymmetric product  $\bigwedge_{i=1}^N \mathcal{H}$ , and let  $\mathcal{F}$  be the ensuing Fock space,

$$\mathcal{F} := \bigoplus_{N=0}^{\infty} \mathcal{V}_N.$$

Analogously to (4.12), we define a matrix product state (MPS) with respect to the basis  $\{\varphi_i\}_{i=1}^{\infty}$  with size parameters ('bond dimensions')  $\{r_i\}_{i=1}^{\infty}$  to be a state of the form

$$\Psi = \lim_{L \rightarrow \infty} \sum_{\mu_1, \dots, \mu_L=0}^1 A_1[\mu_1] A_2[\mu_2] \dots A_L[\mu_L] \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \Phi_{\mu_1 \dots \mu_L} \in \mathcal{F}, \quad (4.17)$$

where the  $A_i[\mu_i]$  ( $i = 1, 2, \dots$ ) are  $r_{i-1} \times r_i$  matrices,  $r_0 = 1$ , the column vector above has length  $r_L$  (so as to make the coefficient of  $\Phi_{\mu_1 \dots \mu_L}$  scalar), and the  $A_i$  are such that the above limit exists as a strong limit in the Fock space  $\mathcal{F}$ . The key point about (4.17) is that the  $A_i$  are *fixed* matrices, which only depend on the *exact* infinite-dimensional quantum state  $\Psi$  and encode its true entanglement structure, whereas first truncating the one-body Hilbert space to dimension  $L$  and then MPS-factorizing the ensuing approximation would lead to  $L$ -dependent  $A_i$ 's.

The vector  $(0, \dots, 0, 1)$ , appearing in (4.17), may look arbitrary at first, but as we showed in [55], every normalized state  $\Psi$  in the Fock space  $\mathcal{F}$  can be represented in the form (4.17) with left-normalized  $A_i$  if the  $r_i$  are allowed to grow exponentially (i.e.,  $r_i = 2^i$ ).

The set of tensor-trains (TT) or matrix product states (MPS) with respect to the basis  $\{\varphi_i\}_{i=1}^{\infty}$  with bond dimensions  $\{r_i\}_{i=1}^{\infty}$  is denoted by

$$\text{MPS}(\infty, \{r_i\}_i, \{\varphi_i\}_i) \subseteq \mathcal{F}. \quad (4.18)$$

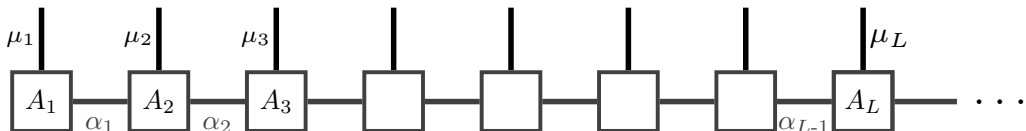


Figure 4.6: Graphical representation of a matrix product state in the infinite-dimensional case.

### 4.3 Contributions in QC-DMRG and Related Literature

In this section, we want to give an overview over the active research areas which we have contributed to and put our results into perspective.

As we have seen, in the context of MPS and also of TNS in general, one main question is how to choose a good network to approximate the state of interest, or even more generally, how to optimize the underlying one-body basis. Even though mathematical articles addressing related topics like the singular values of tensors are emerging [65, 68, 159], a good theoretical understanding of these fundamental questions is still lacking.

We split our investigation into two aspects, concurring with our Articles II and IV. The first will consider pure network optimizations, which reduces in the context of MPS to re-orderings of the basis, whereas the second part considers arbitrary unitary basis transformations, also known



as fermionic mode transformation in physics. Lastly, we shortly comment on the connection between MPS and quantum channels and our work in this related research area.

### 4.3.1 Maximally Entangled Matrix Product States

It has long been recognized that the topology of the underlying tensor network, i.e., the ordering of the orbitals, strongly influences the size of the matrices in the factorization as well as the overall approximation quality [6, 12, 98–100]. It is also of conceptual importance because entanglement of different parts of quantum systems can be viewed as the result of a coupling across their common interface [96].

In almost all QC-DMRG codes, only reorderings instead of the more general fermionic mode transformations of the underlying basis functions are considered. This stems from the fact that the chosen orbitals are carefully crafted from theoretical and empirical knowledge. As mentioned above, different ordering schemes exist, like the widely used Fiedler ordering [6] based on an entanglement analysis of the basis, or newer ones like the best weighted prefactor ordering [37], which is more tailored to quantum chemistry by utilizing an inversion symmetry of singular values for Slater determinants. For weakly correlated states, re-ordering typically reduces the tail by several order of magnitude [37]. To demonstrate the effect of orderings, let us present the example of fermionic Bell-states from our Core article II [64]:

For  $N$  electrons occupying  $L = 2N$  orbitals  $\{\varphi_1, \dots, \varphi_L\}$ , define  $\psi_k := (\varphi_k + \varphi_{k+N})/\sqrt{2}$  for  $k = 1, \dots, N$ . Then consider the Slater determinant given by  $\Psi := |\psi_1, \dots, \psi_N\rangle$ . It is quite straightforward to check that its minimal MPS representation in the basis  $(\varphi_k)_k$  has maximal bond dimension  $2^N$ , see, e.g., [37]. Now applying a re-ordering, which puts paired-up orbitals next to each other,

$$(\tilde{\varphi}_1, \tilde{\varphi}_2, \dots, \tilde{\varphi}_{L-1}, \tilde{\varphi}_L) = (\varphi_1, \varphi_{N+1}, \dots, \varphi_N, \varphi_L),$$

reduces the bond dimension to just 2 in the new basis  $(\tilde{\varphi}_k)_k$ . Indeed,

$$\Psi = \sum_{\mu_1, \dots, \mu_L=0}^1 A_1[\mu_1] \cdots A_L[\mu_L] \tilde{\Phi}_{\mu_1, \dots, \mu_L}, \quad (4.19)$$

where  $\tilde{\Phi}$  is specified as in (4.9) with the new basis  $\{\tilde{\varphi}_k\}_k$  and the matrices  $A_k$  are

$$\begin{aligned} A_1[\mu_1] &= \begin{pmatrix} \delta_{\mu_1}^0 & \delta_{\mu_1}^1 \end{pmatrix}, \quad A_L[\mu_L] = \begin{pmatrix} \delta_{\mu_L}^1 & \delta_{\mu_L}^0 \end{pmatrix}^T, \\ A_{2\ell}[\mu_{2\ell}] &= \begin{pmatrix} \delta_{\mu_{2\ell}}^1 & 0 \\ 0 & \delta_{\mu_{2\ell}}^0 \end{pmatrix}, \quad A_{2\ell+1}[\mu_{2\ell+1}] = \begin{pmatrix} \delta_{\mu_{2\ell+1}}^0 & \delta_{\mu_{2\ell+1}}^1 \\ \delta_{\mu_{2\ell+1}}^0 & \delta_{\mu_{2\ell+1}}^1 \end{pmatrix}. \end{aligned}$$

Here,  $\ell = 1, \dots, N-1$  and  $\delta_v^k$  denotes the Kronecker delta.

This motivated us to rigorously investigate by how much orderings can in general improve the involved matrix sizes. As it turns out, the answer is not at all. Our main result are the following states. Given any basis set and any number of electrons  $N$  and orbitals  $L$ , we constructed

$$\begin{aligned}
 \Psi_{\mathcal{P}} &= \sum_{i_1 < \dots < i_N} \lambda_{i_1, \dots, i_N} |\varphi_{i_1}, \dots, \varphi_{i_N}\rangle \\
 &= \sum_{\mu_1, \dots, \mu_L=0}^1 \psi_{\mu_1, \dots, \mu_L} \Phi_{\mu_1, \dots, \mu_L},
 \end{aligned}
 \tag{4.20}$$

where the coefficients  $\lambda_{i_1, \dots, i_N}$  are pairwise distinct elements of  $\mathcal{P} := \{\sqrt{p_j} : p_j \text{ prime}\}$ . We showed that the corresponding MPS have maximal bond dimension even under any reordering of the basis. Additionally, to demonstrate that this is not a mere theoretical artifact, we investigated the singular value distribution of the unfoldings  $\psi_{\mu_{k+1}, \dots, \mu_L}^{\mu_1, \dots, \mu_k}$  and found an extremely slow decay and a remarkable almost-invariance under re-ordering. This is depicted in Figure 4.7; the plots were done using the code `tensor-train-julia` [36].

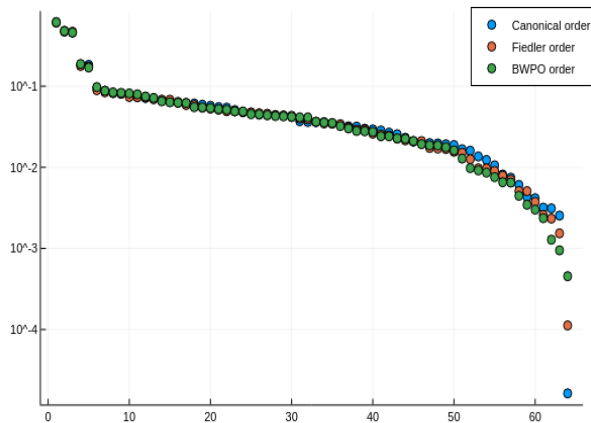


Figure 4.7: Singular value distribution of the matricization  $\psi_{\mu_7, \dots, \mu_{12}}^{\mu_1, \dots, \mu_6}$  of the state  $\Psi_{\mathcal{P}}$  from equation (4.20) with  $N = 6$  electrons and  $L = 12$  orbitals, for different orderings.

A related but less extreme observation, that the bond dimension cannot be lowered much by re-ordering, was made in an interesting numerical study of strongly correlated states in the 1D Hubbard model [96].

### 4.3.2 Bond Dimension in Two-Electron Systems

In the past decade, the Quantum Chemistry Density Matrix Renormalization Group (QC-DMRG) method [23, 97, 125, 188] has become the state-of-the-art choice for systems with up to a few dozen electrons; see [169] for a recent review.

As seen in (4.14), in QC-DMRG, one chooses a suitable finite single-particle basis, makes a matrix product state (MPS) alias tensor-train ansatz for the coefficient tensor of the many-particle wavefunction in Fock space, and optimizes the Rayleigh quotient over the matrices.

As discussed in the previous subsection, it has long been known that the accuracy strongly depends on the choice of basis, and can typically be improved by re-ordering the basis, see, e.g., [6, 37, 169], but we found extreme examples where ordering does not yield an improvement. To only consider a reordering of the basis functions instead of arbitrary fermionic mode transformation, stems from the fact that the chosen orbitals are carefully crafted from theoretical and

empirical knowledge. Nonetheless, the question which basis orbitals are best suited for the MPS representation of a given system, lies at the heart of QC-DMRG.

Additionally, Krumnow, Veis, Legeza, and Eisert [89,90] observed an interesting empirical phenomenon: going beyond ordering and *optimizing over fermionic mode transformations* (i.e., general unitary transformations of the single-particle basis) can reduce the approximation error a great deal further in systems of interest. QC-DMRG together with optimization over the single-particle basis as introduced in [89,90] can be viewed as a generalization of the classical Hartree-Fock method, to which it reduces for bond dimension 1. In particular, utilizing the size of the bond dimension as the key parameter, it interpolates between HF (bond dimension = 1) and the full configuration-interaction method (FCI) (bond dimension =  $2^{L/2}$ ).

In the absence of previous mathematical results on the influence of mode transformations on the approximation error, we investigate in article IV the simplest case  $N = 2$ . We find a dramatic effect, namely a *reduction of the bond dimension needed for exactness of the method from  $2 + \frac{L}{2}$  to 3*, where  $L$  is the number of single-particle basis functions. This is proven by showing that general two-particle wavefunctions can be represented exactly with bond dimension 3 after a (wavefunction-dependent) optimal mode transformation, with 3 being optimal. To be more precise, we prove:

**Theorem 4.8** (Characterization Two-particle case [56])

Suppose  $L \geq 4$  even,  $\Psi \in \mathcal{V}_{2,L}$ , and  $\gamma_\Psi$  has maximal rank =  $L$ . Then, for any basis  $\{\varphi_1, \dots, \varphi_L\}$  and any MPS-representation with bond dimensions  $(r_1, \dots, r_{L-1})$  we have

- $r_j \geq 2$  for every  $j \in \{1, \dots, L-1\}$
- At least one of two consecutive elements  $(r_j, r_{j+1})$  for  $j \in \{2, \dots, L-2\}$  is at least 3.

The bond dimension vector  $(r_1, \dots, r_{L-1})$  with lowest  $\ell^1$ -norm is given by  $(2, 2, 3, \dots, \underbrace{2, 3}_{L-4 \text{ times}}, 2, 2)$ .

Furthermore, there is an explicit representation with optimal bond-dimensions.

Previous exact representations in the form of low-bond-dimension MPS were, to our knowledge, limited to very special states, the prototype example being the AKLT state from spin physics [1] which arises as the ground state of a particular translation invariant Hamiltonian.

Finally, we remark that the exact bond-dimension-three representation of two-fermion wavefunctions carries over to the infinite-dimensional single-particle Hilbert space  $L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$  of full two-electron quantum mechanics, as shown in the last part of this paper.

### 4.3.3 Markovian Divisibility for Quantum Channels

As we have recalled, the idea of Matrix Product States originated in DMRG, but with its mathematical foundations in terms of the current language of tensor networks later on established by Östlund and Rommer [133]. Afterwards, these techniques were extended to more general settings, see, e.g., [127, 179, 181], and it is precisely in the field of numerical analysis where the tensor network ansatz flourished. Additionally, they were widely implemented due to DMRGs groundbreaking precision enabling a deeper understanding of the physical properties of quantum many-body systems [152].

Although the MPS formalism gained traction, the grounds of its success were not fully understood. This comprehension improved through the connection with quantum information and in particular with the theory of quantum channels or completely positive maps, which was established in [43, 182, 192], see also the works of Verstrate and Cirac [180] and Hastings [72]. For a good quantum information based review of matrix product states we refer to [25].

Due to this connection, we also included our work dealing with quantum channels. Here we discuss the open problem of characterizing those quantum channels that can arise from the solution of a (possibly time-dependent) Lindblad master equation. Endeavours towards a resolution of this problem have given rise to different notions of Markovianity for quantum evolutions. We concentrate on the definition which is based on connecting Markovianity to certain divisibility properties of quantum evolutions, in particular, to the possibility of dividing the evolution into infinitesimal pieces. Additionally, we are able to extend the approach to general sets of generators, not only the specific case of quantum channels, i.e, Lindblad generators.

**Definition 4.9** (*Infinitesimal Markovian Divisibility*)

Let  $\mathcal{G} \subset \mathcal{M}_d$  be a compact and convex set of  $d \times d$  matrices containing  $0 \in \mathcal{M}_d$ . We will refer to elements of  $\mathcal{G}$  as generators. We define the set

$$\mathcal{I}_{\mathcal{G}} := \{T \in \mathcal{M}_d \mid \forall \varepsilon > 0 \exists n \in \mathbb{N}, \text{ generators } \{G_j\}_{1 \leq j \leq n} \subset \mathcal{G} \\ \text{s.t. (i) } \|e^{G_j} - \mathbf{1}_d\| \leq \varepsilon \forall j \text{ and (ii) } \prod_{j=1}^n e^{G_j} = T\}.$$

We call the closure  $\overline{\mathcal{I}_{\mathcal{G}}}$  the set of linear maps that are infinitesimal Markovian divisible w.r.t.  $\mathcal{G}$ .

While this gives an intuitively plausible notion of time-dependent quantum Markovianity and some structural properties can be established on its basis, it has so far not given rise to easily verifiable criteria for Markovianity. Only the trivial necessary criterion of non-negativity of the determinant was known. In contrast to higher dimensions, in the qubit case, this notion is completely characterized by Wolf and Cirac [191].

In our investigation, we go beyond this characterization for the 2-dimensional case and obtain necessary criteria for a quantum channel to be divisible into infinitesimal Markovian pieces. In fact we worked with the set of Markovian divisible maps:

**Definition 4.10** (*Markovian Divisibility*)

Let  $\mathcal{G} \subset \mathcal{M}_d$  be a set of matrices, whose elements we call generators. We define the set

$$\mathcal{D}_{\mathcal{G}} := \{T \in \mathcal{M}_d \mid \exists n \in \mathbb{N}, \text{ generators } \{G_i\}_{1 \leq i \leq n} \subset \mathcal{G} \text{ s.t. } \prod_{i=1}^n e^{G_i} = T\}.$$

We call the closure  $\overline{\mathcal{D}_{\mathcal{G}}}$  the set of linear maps that are Markovian divisible w.r.t.  $\mathcal{G}$ .

But as is easy to see by continuity of the matrix exponential, if  $G \in \mathcal{G}$  implies  $\frac{1}{n}G \in \mathcal{G}$  for all  $n \in \mathbb{N}$ , then  $\mathcal{D}_{\mathcal{G}} = \mathcal{I}_{\mathcal{G}}$ . This is in particular the case if  $\mathcal{G}$  satisfies the assumptions of Definition 4.9.

Our main results for quantum channels take the following form

**Theorem 4.11** (Markovian Divisibility [18])

Let  $T$  be an Markovian divisible quantum channel, then

$$0 \leq \det(T) \leq \left( s_1^\uparrow(T) \right)^{\frac{d}{2}}.$$

Also with  $f(d) = 2d - 2\sqrt{2d} + 1$  we have

$$0 \leq \det(T) \leq \prod_{i=1}^{\lfloor f(d) \rfloor} s_i^\uparrow(T).$$

where  $s_i^\uparrow(T)$  denotes the  $i^{\text{th}}$  smallest singular value of  $T$ .

With these criteria at hand, we are able to give new examples of provably not (infinitesimal) Markovian divisible quantum channels.

Lastly, we study the classical counterpart – stochastic matrices as maps of interest and transition rate matrices as generators – and find that no analogous criterion can hold. This implies that there cannot be a mapping from  $d^2 \times d^2$  stochastic matrices to  $d$ -dimensional quantum channels that both preserves infinitesimal Markovian divisibility and leaves singular values invariant.



# Bibliography

- [1] I. Affleck, T. Kennedy, E. H. Lieb, and H. Tasaki. Rigorous results on valence-bond ground states in antiferromagnets. *Physical Review Letters*, 59:799–802, Aug 1987.
- [2] A. Anantharaman and E. Cancès. Existence of minimizers for Kohn–Sham models in quantum chemistry. *Annales de l’Institut Henri Poincaré C, Analyse non linéaire*, 26(6):2425 – 2455, 2009.
- [3] X. Andrade, D. Strubbe, U. De Giovannini, A. H. Larsen, M. J. T. Oliveira, J. Alberdi-Rodriguez, A. Varas, I. Theophilou, N. Helbig, M. J. Verstraete, L. Stella, F. Nogueira, A. Aspuru-Guzik, A. Castro, M. A. L. Marques, and A. Rubio. Real-space grids and the octopus code as tools for the development of new simulation approaches for electronic systems. *Physical Chemistry Chemical Physics*, 17:31371–31396, 2015.
- [4] E. Baerends, O. Gritsenko, and R. Van Meer. The kohn-sham gap, the fundamental gap and the optical gap: The physical meaning of occupied and virtual kohn-sham orbital energies. *Physical Chemistry Chemical Physics*, 15(39):16408–16425, 2013.
- [5] J. L. Bao, P. Verma, and D. G. Truhlar. How well can density functional theory and pair-density functional theory predict the correct atomic charges for dissociation and accurate dissociation energetics of ionic bonds? *Physical Chemistry Chemical Physics*, 20:23072–23078, 2018.
- [6] G. Barcza, O. Legeza, K. H. Marti, and M. Reiher. Quantum-information analysis of electronic states of different molecular structures. *Physical Review A*, 83:012508, Jan 2011.
- [7] T. Barthel, J. Lu, and G. Friesecke. On the closedness and geometry of tensor network state sets. *arXiv preprint arXiv:2108.00031*, 2021.
- [8] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098 – 3100, 1988.
- [9] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *Journal of Chemical Physics*, 98:5648–5652, 1993.
- [10] A. D. Becke. Perspective: Fifty years of density-functional theory in chemical physics. *Journal of Chemical Physics*, 140(18):18A301, 2014.
- [11] S. Behr and B. R. Graswald. Dissociation limit in kohn-sham density functional theory. *arXiv preprint arXiv:2010.09639*, 2020.

## BIBLIOGRAPHY

- [12] K. Boguslawski, P. Tecmer, O. Legeza, and M. Reiher. Entanglement measures for single- and multireference correlation effects. *Journal of Physical Chemistry Letters*, 3(21):3129–3135, 2012.
- [13] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20):457–484, 1927.
- [14] K. Burke. Perspective on density functional theory. *Journal of Chemical Physics*, 136(15):150901, 2012.
- [15] R. G. P. C. Lee, W. Yang. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical Review*, 37:785–789, 1988.
- [16] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Bris, and Y. Maday. Computational quantum chemistry: A primer. *Handbook of Numerical Analysis*, 10:3–270, 2003.
- [17] E. A. Carlen and E. H. Lieb. Remainder terms for some quantum entropy inequalities. *Journal of Mathematical Physics*, 55(4):042201, 2014.
- [18] M. C. Caro and B. R. Graswald. Necessary criteria for markovian divisibility of linear maps. *Journal of Mathematical Physics*, 62(4):042203, 2021.
- [19] I. Catto and P. Lions. Binding of atoms and stability of molecules in Hartree and Thomas-Fermi type theories. Part 4: Binding of neutral systems for the Hartree model. *Communications in Partial Differential Equations*, 18:1149–1159, 01 1993.
- [20] I. Catto and P.-L. Lions. Binding of atoms and stability of molecules in Hartree and Thomas-Fermi type theories. Part I: A necessary and sufficient condition for the stability of general molecular systems. *Communications in partial differential equations*, 17(7-8):1051–1110, 1992.
- [21] I. Catto and P.-L. Lions. Binding of atoms and stability of molecules in Hartree and Thomas-Fermi type theories. Part 3: Binding of neutral subsystems. *Communications in partial differential equations*, 18(3-4):381–429, 1993.
- [22] I. Catto and P.-L. Lions. Binding of atoms and stability of molecules in Hartree and Thomas-Fermi type theories: Part 2: Stability is equivalent to the binding of neutral subsystems. *Communications in partial differential equations*, 18(1-2):305–305, 1993.
- [23] G. K.-L. Chan and M. Head-Gordon. Highly correlated calculations with a polynomial cost algorithm: A study of the density matrix renormalization group. *Journal of Chemical Physics*, 116(11):4462–4476, 2002.
- [24] H. Chen, G. Friesecke, and C. B. Mendl. Numerical methods for a kohn-sham density functional model based on optimal transport. *Journal of Chemical Physics*, 10(10):4360–4368, 2014.



- [25] I. Cirac, D. Perez-Garcia, N. Schuch, and F. Verstraete. Matrix product states and projected entangled pair states: Concepts, symmetries, and theorems. *arXiv preprint arXiv:2011.12127*, 2020.
- [26] A. J. Cohen, P. Mori-Sánchez, and W. Yang. Insights into current limitations of density functional theory. *Science*, 321(5890):792–794, 2008.
- [27] A. J. Cohen, P. Mori-Sánchez, and W. Yang. Challenges for density functional theory. *Chemical Reviews*, 112(1):289–320, 2012.
- [28] M. Colombo, L. De Pascale, and S. Di Marino. Multimarginal optimal transport maps for one-dimensional repulsive costs. *Canadian Journal of Mathematics*, 67(2):350–368, 2015.
- [29] C. Cotar, G. Friesecke, and C. Klüppelberg. Density functional theory and optimal transportation with coulomb cost. *Communications on Pure and Applied Mathematics*, 66(4):548–599, 2013.
- [30] C. Cramer. *Essentials of Computational Chemistry: Theories and Models*. Wiley, 2002.
- [31] H.-L. Dai and W. Ho. *Laser Spectroscopy and Photochemistry on Metal Surfaces*. World Scientific Publishing Company, 1995.
- [32] V. De Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.
- [33] P. A. M. Dirac. Quantum mechanics of many-electron systems. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 123, pages 714–733. The Royal Society, 1929.
- [34] P. A. M. Dirac. Note on exchange phenomena in the thomas atom. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 26, pages 376–385. Cambridge University Press, 1930.
- [35] J. Dukelsky, M. A. Martín-Delgado, T. Nishino, and G. Sierra. Equivalence of the variational matrix product method and the density matrix renormalization group applied to spin chains. *EPL (Europhysics Letters)*, 43(4):457, 1998.
- [36] M.-S. Dupuy. Tensor-train-julia. <https://github.com/msdupuy/Tensor-Train-Julia>, 2021.
- [37] M.-S. Dupuy and G. Friesecke. Inversion symmetry of singular values and a new orbital ordering method in tensor train approximations for quantum chemistry. *SIAM Journal on Scientific Computing*, 43(1):B108–B131, 2021.
- [38] W. E and J. Lu. The elastic continuum limit of the tight binding model. *Chinese Annals of Mathematics, Series B*, 28(6):665–676, 2007.

## BIBLIOGRAPHY

- [39] W. E and J. Lu. Electronic structure of smoothly deformed crystals: Cauchy-Born rule for the nonlinear tight-binding model. *Communications on Pure and Applied Mathematics*, 63(11):1432–1468, 2010.
- [40] W. E and J. Lu. The electronic structure of smoothly deformed crystals: Wannier functions and the Cauchy–Born rule. *Archive for rational mechanics and analysis*, 199(2):407–433, 2011.
- [41] W. E and J. Lu. *The Kohn-Sham equation for deformed crystals*, volume 221. American Mathematical Society, 2013.
- [42] E. Engel and R. M. Dreizler. *Density functional theory*. Springer, 2013.
- [43] M. Fannes, B. Nachtergaele, and R. F. Werner. Finitely correlated states on quantum spin chains. *Communications in mathematical physics*, 144(3):443–490, 1992.
- [44] C. L. Fefferman and L. A. Seco. Asymptotic neutrality of large ions. *Communications in mathematical physics*, 128(1):109–130, 1990.
- [45] E. Fermi. Un metodo statistico per la determinazione di alcune priorieta dell’atome. *Rend. Accad. Naz. Lincei*, 6(602-607):32, 1927.
- [46] E. Fermi. Eine statistische Methode zur Bestimmung einiger Eigenschaften des Atoms und ihre Anwendung auf die Theorie des periodischen Systems der Elemente. *Zeitschrift für Physik A Hadrons and Nuclei*, 48(1):73–79, 1928.
- [47] R. P. Feynman. Forces in molecules. *Physical Review*, 56(4):340, 1939.
- [48] C. Fiolhais, F. Noqueira, and M. M. (EDS). *A Primer in Density Functional Theory*, volume 620. Springer Lecture Notes in Physics, 2003.
- [49] V. Fock. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Zeitschrift für Physik A Hadrons and Nuclei*, 61(1):126–148, 1930.
- [50] R. L. Frank, E. H. Lieb, R. Seiringer, and L. E. Thomas. Bipolaron and n-polaron binding energies. *Physical Review Letters*, 104(21):210402, 2010.
- [51] R. L. Frank, E. H. Lieb, R. Seiringer, and L. E. Thomas. Stability and absence of binding for multi-polaron systems. *Publications mathématiques de l’IHÉS*, 113:39–67, 2011.
- [52] G. Friesecke. The multiconfiguration equations for atoms and molecules: charge quantization and existence of solutions. *Archive for rational mechanics and analysis*, 169(1):35–71, 2003.
- [53] G. Friesecke. Variational principles in quantum theory. <https://www-m7.ma.tum.de/bin/view/Analysis/VariationalPrinciplesQuantum16>, 2017. Accessed: 2021–10-01.
- [54] G. Friesecke and B. Grswald. Existence and nonexistence of HOMO–LUMO excitations in Kohn–Sham density functional theory. *Nonlinear Analysis*, 200:111973, 2020.

- [55] G. Friesecke and B. Graswald. Exact reformulation of quantum mechanics by matrix product states. *in preperation*, 2021.
- [56] G. Friesecke and B. Graswald. Two-electron wavefunctions are matrix product states with bond dimension three. *arXiv preprint arXiv:2109.10091*, 2021.
- [57] G. Friesecke and M. Kniely. New optimal control problems in density functional theory motivated by photovoltaics. *Multiscale Modeling & Simulation*, 17(3):926–947, 2019.
- [58] G. Friesecke, C. B. Mendl, B. Pass, C. Cotar, and C. Klüppelberg.  $N$ -density representability and the optimal transport limit of the Hohenberg-Kohn functional. *Journal of Chemical Physics*, 139(16):164109, 2013.
- [59] L. Garrigue. Unique continuation for many-body Schrödinger operators and the Hohenberg-Kohn theorem. *Mathematical Physics, Analysis and Geometry*, 21(3):1–11, 2018.
- [60] L. Genovese, A. Neelov, S. Goedecker, T. Deutsch, S. A. Ghasemi, A. Willand, D. Caliste, O. Zilberberg, M. Rayson, A. Bergman, et al. Daubechies wavelets as a basis set for density functional pseudopotential calculations. *Journal of Chemical Physics*, 129(1):014109, 2008.
- [61] D. Gontier, C. Hainzl, and M. Lewin. Lower bound on the hartree-fock energy of the electron gas. *Physical Review A*, 99(5):052501, 2019.
- [62] D. Gontier and M. Lewin. Spin symmetry breaking in the translation-invariant hartree-fock electron gas. *SIAM Journal on Mathematical Analysis*, 51(4):3388–3423, 2019.
- [63] D. Gontier, M. Lewin, and F. Q. Nazar. The nonlinear Schrödinger equation for orthonormal functions: Existence of Ground States. *Archive for Rational Mechanics and Analysis*, pages 1–52, 2021.
- [64] B. R. Graswald and G. Friesecke. Electronic wavefunction with maximally entangled mps representation. *The European Physical Journal D*, 75(6):1–4, 2021.
- [65] M. Griebel and H. Harbrecht. Analysis of tensor approximation schemes for continuous functions. *arXiv preprint arXiv:1903.04234*, 2019.
- [66] S. J. Gustafson and I. M. Sigal. *Mathematical concepts of quantum mechanics*, volume 33. Springer, 2003.
- [67] W. Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42. Springer, 2012.
- [68] W. Hackbusch and A. Uschmajew. On the interconnection between the higher-order singular values of real tensors. *Numerische mathematik*, 135(3):875–894, 2017.
- [69] B. C. Hall. *Quantum theory for mathematicians*, volume 267. Springer, 2013.
- [70] D. R. Hartree. The wave mechanics of an atom with a non-coulomb central field. part i. theory and methods. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 89–110. Cambridge University Press, 1928.

## BIBLIOGRAPHY

- [71] D. R. Hartree. The wave mechanics of an atom with a non-Coulomb central field. part ii. some results and discussion. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 111–132. Cambridge University Press, 1928.
- [72] M. B. Hastings. An area law for one-dimensional quantum systems. *Journal of statistical mechanics: theory and experiment*, 2007(08):P08024, 2007.
- [73] V. Heine. European collaboration in ab-initio computer simulations.  $\Psi_k$  *Newsletter, Ab initio (from Electronic Structure) Calculations of Complex Processes in Materials*, 50:7–19, 2002.
- [74] T. Helgaker, P. Jørgensen, and J. Olsen. *Configuration-Interaction Theory*, chapter 11, pages 523–597. John Wiley & Sons, Ltd, 2000.
- [75] H. Hellmann. Zur Rolle der kinetischen Elektronenenergie für die zwischenatomaren Kräfte. *Zeitschrift für Physik*, 85(3-4):180–190, 1933.
- [76] M. F. Herbst, A. Levitt, and E. Cancès. Dftk: A julian approach for simulating electrons in solids. *Proceedings of the JuliaCon Conferences*, 3(26):69, 2021.
- [77] W. Ho. Reactions at metal surfaces induced by femtosecond lasers, tunneling electrons, and heating. *The Journal of Physical Chemistry*, 100(31):13050–13060, 1996.
- [78] M. Hoffmann-Ostenhof and T. Hoffmann-Ostenhof. Schrödinger inequalities and asymptotic behavior of the electron density of atoms and molecules. *Physical Review A*, 16(5):1782, 1977.
- [79] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136:B864–B871, 1964.
- [80] M. Holst, H. Hu, J. Lu, J. Marzuola, D. Song, and J. Weare. Symmetry Breaking in Density Functional Theory due to Dirac Exchange for a Hydrogen Molecule. *arXiv preprint arXiv:1902.03497*, 2019.
- [81] S. Holtz, T. Rohwedder, and R. Schneider. On manifolds of tensors of fixed TT-rank. *Numerische Mathematik*, 120(4):701–731, 2012.
- [82] W. Hunziker. On the spectra of schrödinger multiparticle hamiltonians. *Helvetica Physica Acta (Switzerland)*, 39, 1966.
- [83] W. Hunziker and I. M. Sigal. The quantum  $N$ -body problem. *Journal of Mathematical Physics*, 41(6):3448–3510, 2000.
- [84] R. O. Jones. Density functional theory: Its origins, rise to prominence, and future. *Reviews of modern physics*, 87(3):897, 2015.
- [85] T. Kato. Fundamental properties of Hamiltonian operators of Schrödinger type. *Transactions of the American Mathematical Society*, 70(2):195–211, 1951.

- [86] S. Keller, M. Dolfi, M. Troyer, and M. Reiher. An efficient matrix product operator representation of the quantum chemical hamiltonian. *Journal of Chemical Physics*, 143(24):244118, 2015.
- [87] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140:A1133–A1138, 1965.
- [88] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.
- [89] C. Krumnow, L. Veis, J. Eisert, and O. Legeza. Effective dimension reduction with mode transformations: Simulating two-dimensional fermionic condensed matter systems with matrix-product states. *Physical Review B*, 104:075137, 2021.
- [90] C. Krumnow, L. Veis, O. Legeza, and J. Eisert. Fermionic orbital optimization in tensor network states. *Physical Review Letters*, 117:210402, Nov 2016.
- [91] P. E. Lammert. In search of the Hohenberg-Kohn theorem. *Journal of Mathematical Physics*, 59(4):042110, 2018.
- [92] J. Landsberg. Tensors: geometry and applications, ser. *Graduate Studies in Mathematics*. AMS publ, 128, 2012.
- [93] C. Le Bris. *Quelques problèmes mathématiques en chimie quantique moléculaire*. PhD thesis, Palaiseau, Ecole polytechnique, 1993.
- [94] C. Le Bris. Some results on the Thomas-Fermi-Dirac-Von Weizsäcker model. *Differential and Integral Equations*, 6(2):337–353, 1993.
- [95] C. Lee, W. Yang, and R. G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785, 1988.
- [96] Ö. Legeza, F. Gebhard, and J. Rissler. Entanglement production by independent quantum channels. *Physical Review B*, 74(19):195112, 2006.
- [97] O. Legeza, J. Röder, and B. Hess. Controlling the accuracy of the density-matrix renormalization-group method: The dynamical block state selection approach. *Physical Review B*, 67(12):125114, 2003.
- [98] O. Legeza, J. Röder, and B. A. Hess. Controlling the accuracy of the density-matrix renormalization-group method: The dynamical block state selection approach. *Physical Review B*, 67:125114, Mar 2003.
- [99] O. Legeza, J. Röder, and B. A. Hess. QC-DMRG study of the ionic-neutral curve crossing of LiF. *Molecular Physics*, 101(13):2019–2028, 2003.
- [100] O. Legeza and J. Sólyom. Optimizing the density-matrix renormalization group method using quantum information entropy. *Physical Review B*, 68:195116, Nov 2003.

## BIBLIOGRAPHY

- [101] Ö. Legeza, L. Veis, and T. Mosoni. Qc-dmrg-budapest, a program for quantum chemical dmrg calculations. 2018.
- [102] S. Lehtola, C. Steigemann, M. J. Oliveira, and M. A. Marques. Recent developments in libxc — a comprehensive library of functionals for density functional theory. *SoftwareX*, 7:1–5, 2018.
- [103] W. Lenz. Über die anwendbarkeit der statistischen methode auf ionengitter. *Zeitschrift für Physik*, 77(11-12):713–721, 1932.
- [104] M. Levy. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. *Proc. Natl. Acad. Sci.*, 76(12):6062–6065, 1979.
- [105] M. Lewin. Existence of Hartree-Fock excited states for atoms and molecules. *Letters in Mathematical Physics*, 108(4):985–1006, 2018.
- [106] M. Lewin, E. H. Lieb, and R. Seiringer. Universal functionals in density functional theory. *arXiv preprint arXiv:1912.10424*, 2019 To appear in: *Density Functional Theory, E. Cancès and G. Friesecke (eds.), Springer*.
- [107] M. Lewin, E. H. Lieb, and R. Seiringer. The local density approximation in density functional theory. *Pure and Applied Analysis*, 2(1):35–73, 2019.
- [108] E. H. Lieb. Thomas-Fermi and related theories of atoms and molecules. *Reviews of Modern Physics*, 53(4):603, 1981.
- [109] E. H. Lieb. Density functionals for coulomb systems. *International Journal of Quantum Chemistry*, 24:243–277, 1983.
- [110] E. H. Lieb. Bound on the maximum negative ionization of atoms and molecules. *Physical Review A*, 29(6):3018, 1984.
- [111] E. H. Lieb and R. Seiringer. *The stability of matter in quantum mechanics*. Cambridge University Press, 2010.
- [112] E. H. Lieb, I. M. Sigal, B. Simon, and W. Thirring. Approximate neutrality of large- $Z$  ions. *Communications in mathematical physics*, 116(4):635–644, 1988.
- [113] E. H. Lieb and B. Simon. The Thomas-Fermi theory of atoms, molecules and solids. *Advances in mathematics*, 23(1):22–116, 1977.
- [114] L. Lin and J. Lu. *A Mathematical Introduction to Electronic Structure Theory*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2019.
- [115] P. L. Lions. The concentration-compactness principle in the calculus of variations. The locally compact case, part 1. *Annales de l’Institut Henri Poincaré C, Analyse non linéaire*, 1(4):109–145, 1984.

- [116] P. L. Lions. The concentration-compactness principle in the calculus of variations. The locally compact case, part 2. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 1(4):223–283, 1984.
- [117] P. L. Lions. Solutions of Hartree–Fock equations for Coulomb systems. *Communications in Mathematical Physics*, 109:33–97, 1987.
- [118] List of quantum chemistry and solid-state physics software. List of quantum chemistry and solid-state physics software — Wikipedia, the free encyclopedia, 2021. [Online; accessed 30-September-2021].
- [119] J. Lu and F. Otto. Nonexistence of a Minimizer for Thomas–Fermi–Dirac–von Weizsäcker Model. *Communications on Pure and Applied Mathematics*, 67, 10 2014.
- [120] G. D. Mahan. Modified sternheimer equation for polarizability. *Physical Review A*, 22(5):1780 – 1785, 1980.
- [121] N. Mardirossian and M. Head-Gordon. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular Physics*, 115(19):2315–2372, 2017.
- [122] R. M. Martin. *Electronic structure: basic theory and practical methods*. Cambridge university press, 2020.
- [123] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko. Density functional theory is straying from the path toward the exact functional. *Science*, 355(6320):49–52, 2017.
- [124] N. D. Mermin. Thermal properties of the inhomogeneous electron gas. *Physical Review*, 137(5A):A1441 – A1443, 1965.
- [125] A. O. Mitrushenkov, G. Fano, F. Ortolani, R. Linguerri, and P. Palmieri. Quantum chemistry using the density matrix renormalization group. *Journal of Chemical Physics*, 115(15):6815–6821, 2001.
- [126] S. Mohr, L. E. Ratcliff, P. Boulanger, L. Genovese, D. Caliste, T. Deutsch, and S. Goedecker. Daubechies wavelets for linear scaling density functional theory. *Journal of Chemical Physics*, 140(20):204110, 2014.
- [127] V. Murg, F. Verstraete, and J. I. Cirac. Exploring frustrated spin systems using projected entangled pair states. *Physical Review B*, 79(19):195119, 2009.
- [128] N. Nakatani. Matrix product states and density matrix renormalization group algorithm. In *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering*. Elsevier, 2018.
- [129] P. T. Nam. New bounds on the maximum ionization of atoms. *Communications in Mathematical Physics*, 312(2):427–445, 2012.

## BIBLIOGRAPHY

- [130] P. T. Nam and H. Van Den Bosch. Nonexistence in Thomas-Fermi-Dirac-von Weizsäcker theory with small nuclear charges. *Mathematical Physics, Analysis and Geometry*, 20(2):6, 2017.
- [131] R. Olivares-Amaya, W. Hu, N. Nakatani, S. Sharma, J. Yang, and G. K.-L. Chan. The ab-initio density matrix renormalization group in practice. *Journal of Chemical Physics*, 142(3):034102, 2015.
- [132] I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.
- [133] S. Östlund and S. Rommer. Thermodynamic limit of density matrix renormalization. *Physical Review Letters*, 75(19):3537, 1995.
- [134] R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, Oxford, 1995.
- [135] J. P. Perdew. What do the Kohn-Sham orbital energies mean? How do atoms dissociate? In *Density Functional Methods in Physics*, pages 265–308. Springer, 1985.
- [136] J. P. Perdew. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Physical Review B*, 33(12):8822 – 8824, 1986.
- [137] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77:3865–3868, 1996.
- [138] J. P. Perdew, M. Ernzerhof, and K. Burke. Rationale for mixing exact exchange with density functional approximations. *Journal of Chemical Physics*, 105(22):9982–9985, 1996.
- [139] J. P. Perdew and K. Schmidt. Jacob’s ladder of density functional approximations for the exchange-correlation energy. In *AIP Conference Proceedings*, volume 577, pages 1–20. American Institute of Physics, 2001.
- [140] J. P. Perdew and Y. Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 45(23):13244, 1992.
- [141] J. P. Perdew and A. Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Physical Review B*, 23(10):5048, 1981.
- [142] E. Prodan and P. Nordlander. On the Kohn-Sham equations with periodic background potentials. *Journal of Statistical Physics*, 111(3):967–992, 2003.
- [143] L. Qi and Z. Luo. *Tensor analysis: spectral theory and special tensors*. SIAM, 2017.
- [144] A. K. Rajagopal and J. Callaway. Inhomogeneous electron gas. *Physical Review B*, 7(5):1912 – 1919, 1973.
- [145] D. Rappoport, N. R. M. Crawford, F. Furche, and K. Burke. Approximate density functionals: Which should I choose? *Encyclopedia of Inorganic Chemistry*, 2009.



- [146] L. E. Ratcliff, W. Dawson, G. Fisicaro, D. Caliste, S. Mohr, A. Degomme, B. Videau, V. Cristiglio, M. Stella, M. D'Alessandro, S. Goedecker, T. Nakajima, T. Deutsch, and L. Genovese. Flexibilities of wavelets as a computational basis set for large-scale electronic structure calculations. *Journal of Chemical Physics*, 152(19):194110, 2020.
- [147] M. Reed and B. Simon. *Methods of Modern Mathematical Physics: Analysis of Operators*. Number Bd. 4 in *Methods of Modern Mathematical Physics*. Academic Press, 1978.
- [148] J. Ricaud. Symmetry breaking in the periodic Thomas–Fermi–Dirac–von Weizsäcker model. In *Annales Henri Poincaré*, volume 19, pages 3129–3177. Springer, 2018.
- [149] E. Runge and E. K. U. Gross. Density-functional theory for time-dependent systems. *Physical Review Letters*, 52(12):997 – 1000, 1984.
- [150] M. B. Ruskai. Absence of discrete spectrum in highly negative ions. *Communications in Mathematical Physics*, 82(4):457–469, 1982.
- [151] A. Ruzsinszky, J. P. Perdew, G. I. Csonka, O. A. Vydrov, and G. E. Scuseria. Spurious fractional charge on dissociated atoms: Pervasive and resilient self-interaction error of common density functionals. *Journal of Chemical Physics*, 125(19):194112, 2006.
- [152] M. Sanz Ruiz. *Tensor Networks in Condensed Matter*. PhD thesis, TU München, 2011.
- [153] U. Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of physics*, 326(1):96–192, 2011.
- [154] J. Schwinger. Thomas-Fermi model: The leading correction. *Physical Review A*, 22(5):1827, 1980.
- [155] L. A. Seco, I. Sigal, and J. P. Solovej. Bound on the ionization energy of large atoms. *Communications in mathematical physics*, 131(2):307–315, 1990.
- [156] M. Seidl. Strong-interaction limit of density-functional theory. *Physical Review A*, 60(6):4387, 1999.
- [157] M. Seidl, P. Gori-Giorgi, and A. Savin. Strictly correlated electrons in density-functional theory: A general formulation with applications to spherical densities. *Physical Review A*, 75(4):042511, 2007.
- [158] M. Seidl, J. P. Perdew, and M. Levy. Strictly correlated electrons in density-functional theory. *Physical Review A*, 59(1):51, 1999.
- [159] T. Shi and A. Townsend. On the compressibility of tensors. *SIAM Journal on Matrix Analysis and Applications*, 42(1):275–298, 2021.
- [160] I. M. Sigal. Geometric methods in the quantum many-body problem. nonexistence of very negative ions. *Communications in Mathematical Physics*, 85(2):309–324, 1982.

## BIBLIOGRAPHY

- [161] P. Silvi, D. Rossini, R. Fazio, G. E. Santoro, and V. Giovannetti. Matrix product state representation for Slater determinants and configuration interaction states. *International Journal of Modern Physics B*, 27(01n03):1345029, 2013.
- [162] B. Simon. Schrödinger operators in the twenty-first century. *Mathematical physics*, 2000:283–288, 2000.
- [163] J. C. Slater. Note on Hartree’s method. *Physical Review*, 35(2):210–211, 1930.
- [164] L. Spruch. Pedagogic notes on Thomas-Fermi theory (and on some improvements): atoms, stars, and the stability of bulk matter. *Reviews of Modern Physics*, 63(1):151–209, 1991.
- [165] M. J. Stott and E. Zaremba. Linear-response theory within the density-functional formalism: Application to atomic polarizabilities. *Physical Review A*, 21(1):12 – 23, 1980.
- [166] R. Stowasser and R. Hoffmann. What do the Kohn–Sham orbitals and eigenvalues mean? *Journal of the American Chemical Society*, 121(14):3414–3420, 1999.
- [167] Y. Sugiura. Über die Eigenschaften des Wasserstoffmoleküls im Grundzustande. *Zeitschrift für Physik*, 45(7):484–492, 1927.
- [168] A. Szabo and N. S. Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.
- [169] S. Szalay, M. Pfeffer, V. Murg, G. Barcza, F. Verstraete, R. Schneider, and Ö. Legeza. Tensor product methods and entanglement optimization for ab initio quantum chemistry. *International Journal of Quantum Chemistry*, 115(19):1342–1391, 2015.
- [170] L. A. Takhtajan. *Quantum mechanics for mathematicians*, volume 95. American Mathematical Society, 2008.
- [171] J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria. Climbing the density functional ladder: Nonempirical meta-generalized gradient approximation designed for molecules and solids. *Physical Review Letters*, 91(14):146401, 2003.
- [172] E. Teller. On the stability of molecules in the Thomas-Fermi theory. *Reviews of Modern Physics*, 34(4):627–631, 1962.
- [173] D. G. Tempel, T. J. Martínez, and N. T. Maitra. Revisiting molecular dissociation in density functional theory: A simple model. *Journal of Chemical Physics*, 5(4):770–780, 2009. PMID: 26609582.
- [174] L. H. Thomas. The calculation of atomic fields. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 23, pages 542–548. Cambridge University Press, 1927.
- [175] J. Toulouse. Review of approximations for the exchange-correlation energy in density-functional theory. *arXiv preprint arXiv:2103.02645*, 2021. To appear in: *Density Functional Theory*, E. Cancès and G. Friesecke (eds.), Springer.

- [176] R. van Meer, O. V. Gritsenko, and E. J. Baerends. Physical meaning of virtual kohn–sham orbitals and orbital energies: An ideal basis for the description of molecular excitations. *Journal of Chemical Physics*, 10(10):4432–4441, 2014.
- [177] R. Van Noorden, B. Maher, and R. Nuzzo. The top 100 papers: Nature explores the most-cited research of all time. *Nature*, 514(7524):550 – 553, Oktober 2014.
- [178] C. Van Winter. Theory of finite systems of particles. I. The Green function. *Det Kongelige Danske Videnskabernes Selskab, Matematisk-Fysiske Skrifter*, 2(8), 1964.
- [179] F. Verstraete and J. I. Cirac. Renormalization algorithms for quantum-many body systems in two and higher dimensions. *arXiv preprint cond-mat/0407066*, 2004.
- [180] F. Verstraete and J. I. Cirac. Matrix product states represent ground states faithfully. *Physical Review B*, 73(9):094423, 2006.
- [181] F. Verstraete, J. J. Garcia-Ripoll, and J. I. Cirac. Matrix product density operators: Simulation of finite-temperature and dissipative systems. *Physical Review Letters*, 93(20):207204, 2004.
- [182] F. Verstraete, D. Porras, and J. I. Cirac. Density matrix renormalization group and periodic boundary conditions: A quantum information perspective. *Physical Review Letters*, 93(22):227205, 2004.
- [183] G. Vidal. Efficient classical simulation of slightly entangled quantum computations. *Physical Review Letters*, 91(14):147902, 2003.
- [184] U. von Barth and L. Hedin. A local exchange-correlation potential for the spin polarized case. i. *Journal of Physics C: Solid State Physics*, 5(13):1629 – 1642, 1972.
- [185] C. v. Weizsäcker. Zur Theorie der Kernmassen. *Zeitschrift für Physik*, 96(7-8):431–458, 1935.
- [186] S. R. White. Density matrix formulation for quantum renormalization groups. *Physical Review Letters*, 69(19):2863, 1992.
- [187] S. R. White. Density-matrix algorithms for quantum renormalization groups. *Physical Review B*, 48(14):10345, 1993.
- [188] S. R. White and R. L. Martin. Ab initio quantum chemistry using the density matrix renormalization group. *Journal of Chemical Physics*, 110(9):4127–4130, 1999.
- [189] E. Wigner and F. Seitz. On the constitution of metallic sodium. *Physical Review*, 43(10):804, 1933.
- [190] E. Wigner and F. Seitz. On the constitution of metallic sodium. ii. In *Part I: Physical Chemistry. Part II: Solid State Physics*, pages 372–387. Springer, 1997.
- [191] M. M. Wolf and J. I. Cirac. Dividing quantum channels. *Communications in Mathematical Physics*, 279(1):147–168, 2008.

## BIBLIOGRAPHY

- [192] M. M. Wolf, D. Perez-Garcia, F. Verstraete, and J. I. Cirac. Matrix product state representations. *Quantum Information & Computation*, 7:401–430, 2007.
- [193] S. Wouters, W. Poelmans, P. W. Ayers, and D. Van Neck. Chemp2: A free open-source spin-adapted implementation of the density matrix renormalization group for ab initio quantum chemistry. *Computer Physics Communications*, 185(6):1501–1514, 2014.
- [194] A. Zangwill and P. Soven. Density-functional approach to local-field effects in finite systems: Photoabsorption in the rare gases. *Physical Review A*, 21(5):1561 – 1572, 1980.
- [195] G. Zhang and C. Musgrave. Comparison of dft methods for molecular orbital eigenvalue calculations. *Journal of Physical Chemistry A*, 111(8):1554–1561, 2007.
- [196] G. M. Zhislin. Discussion of the spectrum of Schrödinger operators for systems of many particles. *Trudy Moskovskogo matematicheskogo obschestva*, 9:81–120, 1960.

## Appendix A

### Core Articles

- A.1 Existence and nonexistence of HOMO–LUMO excitations in Kohn–Sham density functional theory

# Existence and nonexistence of HOMO–LUMO excitations in Kohn–Sham density functional theory

Gero Friesecke and Benedikt Graswald

---

Electronic excitations play an important role in the description of molecular properties such as absorption spectra, photoexcitation, state-to-state transition probabilities, reactivity, charge transfer processes, and reaction kinetics. The standard model being employed in numerical computations of these response properties, is Kohn–Sham density functional theory (KS-DFT), because of its good compromise between accuracy and feasibility for large systems. As the mathematical status of such excitations are still not rigorously understood, we consider the simplest such excitations, HOMO–LUMO transitions, in the setting of the local density approximation (LDA). Even in this case we are not aware of previous rigorous results.

For positively charged systems (i.e., total nuclear charge  $Z$  greater than the number  $N$  of electrons) such excitations – mathematically, excited states of the KS Hamiltonian – are rigorously proven to exist in Section 3. These results rely on standard concentration-compactness arguments. Additionally, our assumptions on the exchange-correlation functional are verified explicitly for the widely used PZ81 and PW92 functionals in the appendix.

As a corollary we also establish in Section 4 the existence of optimal excitations with respect to suitable control goals recently introduced by Friesecke and Kniely, without requiring the simplifying assumption in of bounded domains. This is done by proving compactness in a suitable topology of the set of tuples containing the KS-orbitals, the HOMO, the LUMO and the nuclear charge distribution, in addition to continuity of the involved functionals.

By contrast, Section 5 shows that in the neutral case  $Z = N$  and for the hydrogen and helium atoms, such excited states do not exist when the self-consistent KS ground state density is replaced by a realistic but easier to analyze closed-form approximation (in case of hydrogen, the true Schrödinger ground state density). This result is presented in Theorem 4 and utilizes a method by Glaser, Martin, Grosse, and Thirring, which could also be applied to numerical KS ground state densities. Gero Friesecke was the one suggesting to relax the problem in the neutral case by considering the Schrödinger density in the hydrogen case and a dilated version of the hydrogen orbital for Helium – following the ansatz of Hans Bethe.

Additionally, we give a thorough interpretation of this non-existence result from a physics perspective as well as from the point of view of numerical computations in finite basis sets, stressing the fact that in contrast to common (explicit or implicit) belief, restriction to finite basis sets or bounded domains may be not just a negligible technicality, but significantly alters the physical nature of LUMO excitations, from stable bound state to a delocalized, dispersing state associated with the continuous spectrum.

*Own contribution.* I was significantly involved in finding the ideas with the exception of the above mentioned relaxation of the problem, and carried out most of the scientific work of all parts of this article. In particular, I proved Theorems 1 through 4. Furthermore, I wrote the first draft of the article as well as all parts of the final version except the introduction, which was written jointly by both coauthors.

# Permission to include:

Gero Friesecke and Benedikt R. Graswald (2020).  
Existence and nonexistence of HOMO–LUMO excitations  
in Kohn–Sham density functional theory.  
*Nonlinear Analysis* 200, 111973.  
<https://doi.org/10.1016/j.na.2020.111973>



Home (<https://www.elsevier.com>) > About (<https://www.elsevier.com/about>)  
> Policies (<https://www.elsevier.com/about/policies>) > Copyright (<https://www.elsevier.com/about/policies/copyright>)

## Copyright

[Overview](#)   [Author rights](#)   [Institution rights](#)   [Government rights](#)   [Find out more](#)

### Overview

In order for Elsevier to publish and disseminate research articles, we need certain publishing rights from authors, which are determined by a publishing agreement between the author and Elsevier.

For articles published open access, the authors license exclusive rights in their article to Elsevier.

For articles published under the subscription model, the authors transfer copyright to Elsevier.

Regardless of whether they choose to publish open access or subscription with Elsevier, authors have many of the same rights under our publishing agreement, which support their need to share, disseminate and maximize the impact of their research.

For open access articles, authors will also have additional rights, depending on the Creative Commons end user license that they select. This Creative Commons license sets out the rights that readers (as well as the authors) have to re-use and share the article: please see here (<https://www.elsevier.com/about/policies/open-access-licenses>) for more information on how articles can be re-used and shared under these licenses.

This page aims to summarise authors' rights when publishing with Elsevier; these are explained in more detail in the [↓ publishing agreement between the author and Elsevier](#).

Irrespective of how an article is published, Elsevier is committed to protect and defend authors' works and their reputation. We take allegations of infringement, plagiarism, ethical disputes, and fraud very seriously.

### Author rights

The below table explains the rights that authors have when they publish with Elsevier, for authors who choose to publish either open access or subscription. These apply to the corresponding author and all co-authors.

Author rights in Elsevier's proprietary journals	Published open access	Published subscription
Retain patent and trademark rights	√	√
Retain the rights to use their research data freely without any restriction	√	√
Receive proper attribution and credit for their published work	√	√



Author rights in Elsevier's proprietary journals	Published open access	Published subscription
Re-use their own material in new works without permission or payment (with full acknowledgement of the original article): <ol style="list-style-type: none"> <li>1. Extend an article to book length</li> <li>2. Include an article in a subsequent compilation of their own work</li> <li>3. Re-use portions, excerpts, and their own figures or tables in other works.</li> </ol>	√	√
Use and share their works for scholarly purposes (with full acknowledgement of the original article): <ol style="list-style-type: none"> <li>1. In their own classroom teaching. Electronic and physical distribution of copies is permitted</li> <li>2. If an author is speaking at a conference, they can present the article and distribute copies to the attendees</li> <li>3. Distribute the article, including by email, to their students and to research colleagues who they know for their personal use</li> <li>4. Share and publicize the article via Share Links, which offers 50 days' free access for anyone, without signup or registration</li> <li>5. Include in a thesis or dissertation (provided this is not published commercially)</li> <li>6. Share copies of their article privately as part of an invitation-only work group on commercial sites with which the publisher has a hosting agreement</li> </ol>	√	√
Publicly share the preprint on any website or repository at any time.	√	√
Publicly share the accepted manuscript on non-commercial sites	√	√ using a CC BY-NC-ND license and usually only after an embargo period (see Sharing Policy ( <a href="https://www.elsevier.com/about/policies/sharing">https://www.elsevier.com/about/policies/sharing</a> ) for more information)
Publicly share the final published article	√ in line with the author's choice of end user license	×
Retain copyright	√	×

## Institution rights

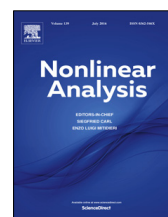
Regardless of how the author chooses to publish with Elsevier, their institution has the right to use articles for classroom teaching and internal training. Articles can be used for these purposes throughout the author's institution, not just by the author:



Contents lists available at ScienceDirect

Nonlinear Analysis

www.elsevier.com/locate/na



# Existence and nonexistence of HOMO–LUMO excitations in Kohn–Sham density functional theory



Gero Friesecke, Benedikt Graswald\*

Department of Mathematics, Technische Universität München, Germany

## ARTICLE INFO

*Article history:*

Received 26 November 2019

Accepted 14 May 2020

Communicated by Irena Lasiecka

*Keywords:*

Nonlinear eigenvalue equations

Density-functional-theory

Kohn–Sham equations

HOMO–LUMO gap

Excitations

## ABSTRACT

In numerical computations of response properties of electronic systems, the standard model is Kohn–Sham density functional theory (KS-DFT). Here we investigate the mathematical status of the simplest class of excitations in KS-DFT, HOMO–LUMO excitations. We show that such excitations, i.e. excited states of the Kohn–Sham Hamiltonian, exist for  $Z > N$ , where  $Z$  is the total nuclear charge and  $N$  is the number of electrons. The result applies under realistic assumptions on the exchange–correlation functional, which we verify explicitly for the widely used PZ81 and PW92 functionals. By contrast, and somewhat surprisingly, we find using a method of Glaser, Martin, Grosse, and Thirring (Glaser et al., 1976) that in case of the hydrogen and helium atoms, excited states do not exist in the neutral case  $Z = N$  when the self-consistent KS ground state density is replaced by a realistic but easier to analyze approximation (in case of hydrogen, the true Schrödinger ground state density). Implications for interpreting minus the HOMO eigenvalue as an approximation to the ionization potential are indicated.

© 2020 Elsevier Ltd. All rights reserved.

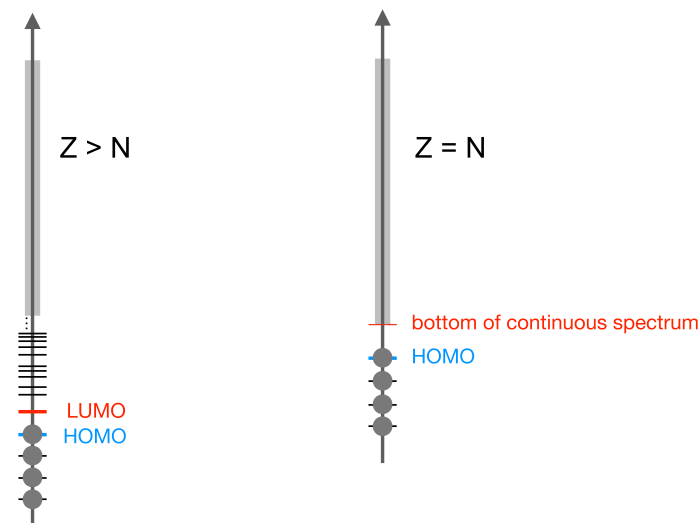
## 1. Introduction

Electronic excitations play an important role in the description of molecular properties such as absorption spectra, photoexcitation, state-to-state transition probabilities, reactivity, charge transfer processes, and reaction kinetics [5,6,11]. In numerical computations of these response properties, the standard model is Kohn–Sham density functional theory (KS-DFT), because of its good compromise between accuracy and feasibility for large systems (see [18] for a textbook account and [3] for a recent review). It is then of interest to investigate the mathematical status of excitations in KS-DFT.

In this paper we mathematically analyze the simplest such excitations, HOMO–LUMO transitions, in the setting of the local density approximation (LDA). For a systematic comparison of HOMO–LUMO excitations with experimental data see e.g. [2,22]. Even in this case we are not aware of previous rigorous results. Our findings are the following (see Fig. 1).

\* Corresponding author.

E-mail addresses: gf@ma.tum.de (G. Friesecke), benedikt.graswald@ma.tum.de (B. Graswald).



**Fig. 1.** Schematic picture of the spectrum of the KS Hamiltonian. Positively charged systems (left,  $Z > N$ ) have infinitely many excited states above the HOMO and below the continuous spectrum (see [Theorem 2](#)). For neutral systems (right,  $Z = N$ ), it can happen that there are no excited states, that is, the highest bound state eigenvalue is the HOMO (see [Theorem 4](#)).

For positively charged systems (i.e., total nuclear charge  $Z$  greater than the number  $N$  of electrons) such excitations – mathematically, higher eigenstates of the KS Hamiltonian, which is a certain elliptic partial integrodifferential operator – are rigorously proven to exist, under realistic assumptions on the exchange–correlation functional which we verify explicitly for the widely used PZ81 and PW92 functionals. See [Theorem 1](#) in [Section 3](#).

By contrast, the neutral case  $Z = N$  holds a surprise. In the case of the hydrogen and helium atoms, we prove that excited states do not exist when the self-consistent KS ground state density is replaced by a realistic but easier to analyze closed-form approximation (in case of hydrogen, the true Schrödinger ground state density). See [Theorem 4](#) in [Section 5](#).

Mathematically, the existence result relies on spectral properties of the corresponding hamiltonian combined with careful a priori estimates on the specific PDE system under study. The nonexistence result uses a not widely known method by Glaser, Martin, Grosse, and Thirring (GMGT) [10]. The latter method could, in principle, also be applied to numerical KS ground state densities; we expect that for *some* atoms and molecules, including hydrogen and helium, the GMGT nonexistence criterion (that a certain integral associated with the effective KS potential lies below a threshold value) would be satisfied.

The quadratic functional minimized by excitations, (2.13), can be viewed as an approximation to the Kohn–Sham energy functional (2.5). We remark that the latter is closely related to the Thomas–Fermi–Dirac–von Weizsäcker functional for which an interesting nonexistence result of minimizers was proved via completely different methods in [17].

Physically, these results indicate a significant artifact of KS-DFT. In the full  $N$ -electron Schrödinger equation, neutral systems (and even systems with  $Z > N - 1$ ) are known to possess infinitely many excited states below the bottom of the continuous spectrum. This is a celebrated result by Zhislin [23]; for a modern variational proof see [8]. The analogous result also holds in Hartree–Fock theory: for  $Z > N$  the Fock operator associated with the Hartree–Fock ground state density possess infinitely many bound states below the continuous spectrum [16, Lemma II.3], the latter being the interval  $[0, \infty)$ . Our results suggest that in KS-DFT, the threshold for existence of infinitely many excited states is shifted from  $Z > N - 1$  to  $Z > N$ . This is a previously unnoticed but important qualitative consequence of the (well known) incomplete cancellation of the self-interaction energy in KS-DFT.

It is interesting to interpret the nonexistence of excitations from the point of view of numerical computations in finite basis sets, or mathematical analysis (as in [9]) in bounded domains. Consider a neutral

system for which (exact) excitations do not exist. In a finite basis set, or a bounded domain, the spectrum of the KS Hamiltonian is purely discrete and therefore excited states exist. In the limit as the basis set approaches completeness, or the domain approaches the whole of  $\mathbb{R}^3$ ,

(i) the LUMO energy  $\varepsilon_L$  (i.e., the lowest unoccupied eigenvalue of the KS Hamiltonian) will remain well-defined, and approaches the bottom of the continuous spectrum (which equals 0, see [Theorems 2 and 4](#))

(ii) the LUMO (i.e., the lowest unoccupied eigenstate) will become more and more delocalized, failing to converge to a bound state.

Thus in contrast to common (explicit or implicit) belief, restriction to finite basis sets or bounded domains may be not just a negligible technicality, but significantly alters the physical nature of LUMO excitations, from stable bound state (i.e., invariant under the dynamics of the KS ground state Hamiltonian) to a delocalized, dispersing state associated with the continuous spectrum.

(ii) makes it very tempting to physically interpret the HOMO–LUMO excitation in the nonexistence case as an (approximation to an) ionization process. This interpretation together with (i) yields *ionization potential*  $\approx \varepsilon_L - \varepsilon_H = 0 - \varepsilon_H$  (where  $\varepsilon_H$  is the HOMO energy, i.e. the highest occupied eigenvalue of the KS Hamiltonian), lending new theoretical support to the famous semi-empirical formula

$$-\varepsilon_H \approx \textit{ionization potential}$$

which often agrees quite well with experimental data [\[2,22\]](#).

## 2. Mathematical setting

We start by recalling well-known mathematical facts about Kohn–Sham density functional theory (KS–DFT) [\[12,13,18\]](#). Readers familiar with these facts might want to skip this part. After that we give a variational definition of HOMO–LUMO excitations as introduced recently in [\[9\]](#), which works irrespective of degeneracies and is convenient for the mathematical analysis of excitations.

### 2.1. Kohn–Sham equations

We consider a system of  $N$  non-relativistic electrons in  $\mathbb{R}^3$  in the electrostatic potential generated by  $M$  nuclei of charges  $Z_1, \dots, Z_M$  located at positions  $R_1, \dots, R_M \in \mathbb{R}^3$ ,

$$v_{ext}(x) = - \sum_{\alpha=1}^M Z_\alpha \frac{1}{|x - R_\alpha|}. \quad (2.1)$$

In fact, for our analysis it is not essential that the nuclei are point particles. It suffices to assume the nuclear charge distribution is given by a nonnegative Radon measure  $\mu$  with total mass  $Z > 0$  supported on a compact set  $\Omega_{nuc} \subseteq \mathbb{R}^3$ , i.e. we consider any  $\mu$  belonging to

$$\mathcal{A}_{nuc} := \{ \mu \in \mathcal{M}(\Omega_{nuc}) : \mu \geq 0, \int_{\Omega_{nuc}} d\mu = Z \}, \quad (2.2)$$

where  $\mathcal{M}(\Omega_{nuc})$  denotes the space of signed Radon measures on  $\Omega_{nuc}$ , and

$$v_{ext}(x) := - \int_{\Omega_{nuc}} \frac{1}{|x - y|} d\mu(y). \quad (2.3)$$

For simplicity we look at a spin-unpolarized system, so the number  $N$  of electrons is even, i.e.  $N = 2n$  for some  $n \in \mathbb{N}$ . In this case Kohn–Sham DFT describes the electrons by  $n$  orbitals  $\varphi_1, \dots, \varphi_n : \mathbb{R}^3 \rightarrow \mathbb{C}$ , each occupied by two electrons of opposite spin. They are  $L^2$ -orthonormal, i.e.

$$\langle \varphi_i, \varphi_j \rangle_{L^2} = \int_{\mathbb{R}^3} \varphi_i(x) \overline{\varphi_j(x)} dx = \delta_{ij} \quad \forall i, j \in \{1, \dots, n\}, \quad (2.4)$$

and we denote  $\Phi := (\varphi_1, \dots, \varphi_n)$ . Note that in the following  $\langle \cdot, \cdot \rangle$  will always denote the  $L^2$  inner product. Then the corresponding Kohn–Sham energy functional is given by

$$\mathcal{E}_\mu[\Phi] = \underbrace{\sum_{k=1}^n 2 \int_{\mathbb{R}^3} \frac{1}{2} |\nabla \varphi_k|^2(x) \, dx}_{=:T[\Phi]} + \underbrace{\int_{\mathbb{R}^3} v_{ext}(x) \rho(x) \, dx}_{=:V[\rho]} + \underbrace{\frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|} \, dx \, dy}_{=:J[\rho]} + \underbrace{\int_{\mathbb{R}^3} e_{xc}(\rho(x)) \, dx}_{=:E_{xc}[\rho]}, \quad (2.5)$$

where  $e_{xc}$  gives the exchange–correlation energy per unit volume and  $\rho$  is the total electron density, that is,

$$\rho(x) := 2 \sum_{k=1}^n |\varphi_k(x)|^2. \quad (2.6)$$

The KS energy hence consists of the following terms:  $T$  is the kinetic energy of the electrons,  $V$  is the potential energy from the electron–nuclei interaction with  $v_{ext}$  being the electrostatic potential of the nuclei (2.3);  $J_H$  (the Hartree energy) describes the energy corresponding to the interelectron repulsion if the electrons were mutually independent;  $E_{xc}$  is the exchange–correlation energy which accounts for correlation effects correcting the simple independent ansatz of  $J_H$ .

Note here that the Coulomb potential over the whole  $\mathbb{R}^3$  is not in any  $L^p$ -space, but it is in  $L^2(\mathbb{R}^3) + L^\infty(\mathbb{R}^3)$  and we will be using the splitting  $\frac{1}{|\cdot|} = v_2 + v_\infty$ , where  $v_2, v_\infty$  lie in  $L^2, L^\infty$ , respectively.

Precise assumptions on  $e_{xc}$  which are sufficient for our mathematical results and cover standard local density approximation (LDA) exchange–correlation functionals used in practice are given in Section 3.1. A basic example derived from the homogeneous electron gas is the Dirac exchange energy

$$e_{xc}(\rho) = -\frac{3}{4} \left(\frac{3}{\pi}\right)^{\frac{1}{3}} \rho^{\frac{4}{3}}. \quad (2.7)$$

Ground states  $\Phi := (\varphi_1, \dots, \varphi_N)$  of the system are states satisfying

$$\Phi \in \operatorname{argmin} \mathcal{E}_\mu \text{ subject to the constraints (2.4),} \quad (2.8)$$

and are known to exist under suitable assumptions [1].

Any ground state satisfies the Euler–Lagrange equations of the system, the Kohn–Sham equations

$$h_{\mu,\rho} \varphi_i := \left( -\frac{1}{2} \Delta + v_{ext} + v_H + v_{xc} \right) \varphi_i = \sum_{j=1}^N \lambda_{ij} \varphi_j, \quad (2.9)$$

where the Lagrange multipliers  $\lambda_{ij}$  arise due to the orthonormality condition (2.4). The Hartree and exchange–correlation potentials are given by

$$v_H(x) = \int_{\mathbb{R}^3} \frac{1}{|x-y|} \, d\mu(y), \quad v_{xc} = \frac{d}{d\rho} e_{xc}. \quad (2.10)$$

Since the effective one-body operator  $h_{\mu,\rho}$  in (2.9) (the Kohn–Sham Hamiltonian) is invariant under unitary transformations, the KS equations can be brought into their canonical form

$$h_{\mu,\rho} \varphi_i := \left( -\frac{1}{2} \Delta + v_{ext} + v_H + v_{xc} \right) \varphi_i = \varepsilon_i \varphi_i. \quad (2.11)$$

## 2.2. Excitations

Following [9] we confine ourselves here to the simplest model for electronic excitations, the HOMO–LUMO transition. In this transition an electron pair migrates from the highest occupied molecular orbital (HOMO)

to the lowest unoccupied molecular orbital (LUMO). For the KS-orbitals  $\Phi = (\varphi_1, \dots, \varphi_n)$  ordered by the size of their eigenvalue in (2.11) this means

$$(\varphi_1, \dots, \varphi_{n-1}, \varphi_n) \longrightarrow (\varphi_1, \dots, \varphi_{n-1}, \varphi_{n+1}), \quad (2.12)$$

where  $\varphi_n$  is the HOMO and  $\varphi_{n+1}$  – the eigenstate corresponding to the next higher eigenvalue of  $h_{\mu,\rho}$  – is the LUMO.

To define HOMO and LUMO in a variational way, we consider the *excitation energy functional* [9] given by the quadratic form associated with KS Hamiltonian  $h_{\mu,\rho}$  (2.11),

$$\mathcal{E}_{\mu,\rho}[\chi] = \langle \chi, h_{\mu,\rho}\chi \rangle = \frac{1}{2} \int_{\mathbb{R}^3} |\nabla\chi|^2 + \int_{\mathbb{R}^3} (v_{ext} + v_H + v_{xc}(\rho))|\chi|^2. \quad (2.13)$$

Now define a HOMO  $\varphi_H$  by

$$\varphi_H \in \operatorname{argmax} \mathcal{E}_{\mu,\rho} \text{ subject to the constraints } \varphi_H \in \operatorname{Span}\{\varphi_1, \dots, \varphi_n\}, \langle \varphi_H, \varphi_H \rangle = 1 \quad (2.14)$$

and a LUMO  $\varphi_L$  by

$$\varphi_L \in \operatorname{argmin} \mathcal{E}_{\mu,\rho} \text{ subject to the constraints } \langle \varphi_i, \varphi_L \rangle = 0, \forall i \in \{1, \dots, n\}, \langle \varphi_L, \varphi_L \rangle = 1. \quad (2.15)$$

If they exist, HOMO and LUMO clearly satisfy the KS equations

$$h_{\mu,\rho}\varphi_H = \varepsilon_H\varphi_H, \quad h_{\mu,\rho}\varphi_L = \varepsilon_L\varphi_L, \quad (2.16)$$

for some eigenvalue  $\varepsilon_H$  (the HOMO energy) and  $\varepsilon_L$  (the LUMO energy).

### 3. Existence of HOMO–LUMO excitations

In this section we show that for positively charged systems ( $Z > N$ ) there always exist HOMO–LUMO excitations. This generalizes a corresponding result in [9] to unbounded domains, except that in bounded domains no restriction on  $Z$  is needed. In the latter case such a result is not straightforward due to the possibility of “mass escaping to infinity”, and requires concentration-compactness arguments [14,15]. The reader may wonder whether the assumption  $Z > N$ , which is essential in our proof, is really necessary. The authors of course asked themselves the same question. For counterexamples to existence in the case  $Z = N$  see Section 5.

#### 3.1. Assumptions

##### Assumptions on the exchange–correlation energy

We assume that  $e_{xc} : [0, \infty) \rightarrow \mathbb{R}$  is continuously differentiable with

$$e_{xc}(0) = 0 = v_{xc}(0), v_{xc} \leq 0, |v_{xc}(\rho)| \leq c_{xc}(1 + \rho^{p-1}) \quad (3.1)$$

for some exponent  $p$  with  $p \in [1, \frac{5}{3})$  and constant  $c_{xc} > 0$ .

Furthermore we need as in [1]:

$$\text{There exists } q \in [1, \frac{3}{2}) \text{ such that } \limsup_{\rho \rightarrow 0^+} \frac{e_{xc}(\rho)}{\rho^q} < 0. \quad (3.2)$$

**Remark 3.1.** These assumptions are trivially satisfied by the Dirac exchange (2.7) with  $p = q = \frac{4}{3}$ . In our appendix we check explicitly that these assumptions are also satisfied for the two most popular LDA exchange–correlation functionals: Perdew–Zunger (PZ81) [20] and Perdew–Wang (PW92) [19]

**Admissible sets of orbitals** In order to write these definitions in a more compact way we introduce the following sets: The KS admissible set

$$\mathcal{A} = \{(\varphi_1, \dots, \varphi_n) \in H^1(\mathbb{R}^3)^n : \langle \varphi_i, \varphi_j \rangle = \delta_{ij}\}$$

the HOMO admissible set

$$\mathcal{A}_\Phi^H = \{\varphi_H \in \text{Span}\{\varphi_1, \dots, \varphi_n\} : \|\varphi_H\|_2 = 1\}$$

and the LUMO admissible set

$$\mathcal{A}_\Phi^L = \{\varphi_L \in H^1(\mathbb{R}^3) : \langle \varphi_k, \varphi_L \rangle = 0 \text{ for } k \in \{1, \dots, n\}, \|\varphi_L\|_2 = 1\}.$$

The governing variational principles for the occupied KS orbitals, HOMO and LUMO can now be summarized as

$$\Phi \in \underset{\mathcal{A}}{\operatorname{argmin}} \mathcal{E}_\mu, \quad \varphi_H \in \underset{\mathcal{A}_\Phi^H}{\operatorname{argmax}} \mathcal{E}_{\mu, \rho}, \quad \varphi_L \in \underset{\mathcal{A}_\Phi^L}{\operatorname{argmin}} \mathcal{E}_{\mu, \rho}. \quad (3.3)$$

We start with some estimates for the KS energy functional (2.5).

**Lemma 1** (Lower bounds on KS energy functional). *The terms in the KS energy functionals have the following properties*

1.  $T[\Phi] \geq \frac{1}{2}T[\Phi] + \frac{1}{c}\|\rho\|_3$  and

$$\Phi \mapsto T[\Phi] \text{ is continuous and weakly lower semicontinuous on } H^1(\mathbb{R}^3)^n.$$

2.  $V[\rho] \geq -\|\mu\|_{\mathcal{M}}(\|v_\infty\|_\infty\|\rho\|_1 + \|v_2\|_2\|\rho\|_1^{1/4}\|\rho\|_3^{3/4})$  and

$$(\Phi, \mu) \mapsto V[\rho] \text{ is strong } \times \text{ weak}^* \text{ continuous on } (L^4 \cap L^2) \times \mathcal{M}.$$

3.  $J_H \geq 0$  and

$$\Phi \mapsto J_H[\rho] \text{ continuous on } (L^{12/5}(\mathbb{R}^3))^n.$$

4.  $E_{xc}[\rho] \geq -c_{xc}(\|\rho\|_1 + \frac{1}{p-1}\|\rho\|_1^{(3-p)/2}\|\rho\|_3^{(3(p-1)/2)})$ , where  $p \in (1, \frac{5}{3})$  is the exponent from assumption (3.1), and

$$\Phi \mapsto E_{xc}[\rho] \text{ is continuous on } (L^{2p}(\mathbb{R}^3))^n.$$

**Proof.** The first inequality is a standard result in DFT, but we include it for the sake of completeness. By a well-known result, see e.g. [4], we have  $T[\Phi] \geq \|\nabla \sqrt{\rho}\|_2^2$ , so the inequality follows by applying the Sobolev embedding  $H^1 \hookrightarrow L^6$  to the function  $u = \sqrt{\rho}$ .

Estimate 2 follows from the duality between  $\mathcal{M}(\Omega_{nuc})$  and  $C_b(\Omega_{nuc})$  and then Cauchy–Schwarz

$$V[\rho] = - \int_{\mathbb{R}^3} \left( \frac{1}{|\cdot|} * \rho \right) d\mu \geq -\|\mu\|_{\mathcal{M}} \|\frac{1}{|\cdot|} * \rho\|_\infty \geq -\|\mu\|_{\mathcal{M}} (\|v_2\|_2\|\rho\|_2 + \|v_\infty\|_\infty\|\rho\|_1)$$

and finally bounding the  $L^2$ -norm of  $\rho$  by the Hölder interpolation inequality,

$$\|\rho\|_p \leq \|\rho\|_q^\theta \|\rho\|_r^{1-\theta} \text{ with } q \leq p \leq r, \quad \frac{1}{p} = \frac{\theta}{q} + \frac{1-\theta}{r} \quad (3.4)$$

with  $p = 2, q = 1, r = 3$ .

The positivity of  $J_H$  is trivial.

Ad 4: By our assumption on  $v_{xc}$  we have

$$|e_{xc}(\rho)| = |e_{xc}(0) + \int_0^\rho v_{xc}(\xi) d\xi| \leq c_{xc}(\rho + \frac{1}{p-1}\rho^p)$$

The estimate again follows from Hölder interpolation (3.4) with  $q = 1, r = 3, \theta = \frac{(3-p)}{2p}$ . In all four cases the continuity results follow by pointwise continuity of the integrand and the proven bounds.  $\square$

The second Lemma considers the excitation functional (2.13). In the following we denote  $\rho_\chi = |\chi|^2$ .

**Lemma 2** (Lower Bounds on Excitation Functional). *The terms in the excitation functional have the following properties*

1.  $T[\psi] \geq \frac{1}{2}T[\psi] + \frac{1}{c}\|\rho_\psi\|_3$  and

$$\chi \mapsto T[\chi] \text{ is continuous and weakly lower semicontinuous on } H^1(\mathbb{R}^3).$$

2.  $\int_{\mathbb{R}^3} v_{ext}\rho_\psi \geq -\|\mu\|_{\mathcal{M}}(\|v_\infty\|_\infty\|\rho_\psi\|_1 + \|v_2\|_2\|\rho_\psi\|_1^{1/4}\|\rho_\psi\|_3^{3/4})$  and

$$(\chi, \mu) \mapsto \int_{\mathbb{R}^3} v_{ext}\rho_\chi \text{ is strong} \times \text{weak}^* \text{ continuous on } (L^4 \cap L^2) \times \mathcal{M}.$$

3.  $\int_{\mathbb{R}^3} (\frac{1}{|\cdot|} * \rho)\rho_\psi \geq 0$  and

$$(\Phi, \chi) \mapsto \int_{\mathbb{R}^3} (\frac{1}{|\cdot|} * \rho)\rho_\chi \text{ is continuous on } L^{12/5}(\mathbb{R}^3)^{n+1}.$$

4.  $\int_{\mathbb{R}^3} v_{xc}(\rho)\rho_\psi \geq -c_{xc}(\|\rho_\psi\|_1 + \|\rho\|_p^{p-1}\|\rho_\psi\|_1^{(3-p)/(2p)}\|\rho_\psi\|_3^{3(p-1)/(2p)})$  where  $p \in [1, \frac{5}{3}]$  is again the exponent from our assumptions on  $v_{xc}$  and

$$(\Phi, \chi) \mapsto \int_{\mathbb{R}^3} v_{xc}(\rho)\rho_\chi \text{ is continuous on } L^{2p}(\mathbb{R}^3)^n \times (L^2(\mathbb{R}^3) \cap L^{2p}(\mathbb{R}^3)).$$

In particular, the map  $(\Phi, \chi, \mu) \mapsto \mathcal{E}_{\mu, \rho}[\chi]$  is weak  $\times$  strong  $\times$  weak\* continuous and weak  $\times$  weak  $\times$  weak\* lower semicontinuous on  $H^1(\mathbb{R}^3)^n \times H^1(\mathbb{R}^3) \times \mathcal{M}$ .

**Proof.** Statements 1–3 follow by the same line of reasoning as in Lemma 2. The fourth assertion follows by the same argument given in [9], but we include it for the sake of completeness. By our assumption on the exchange–correlation potential  $v_{xc}$  we have

$$\int_{\mathbb{R}^3} v_{xc}(\rho)\rho_\psi \geq -c_{xc} \int_{\mathbb{R}^3} (1 + \rho^{p-1})\rho_\psi \geq -c_{xc} \left( \|\rho_\psi\|_1 + \underbrace{\|\rho^{p-1}\|_{p'}}_{=\|\rho\|_p^{p-1}} \cdot \|\rho_\psi\|_p^p \right) \quad \text{with } p' = \frac{p}{p-1}.$$

The asserted bound now follows from the Hölder interpolation inequality (3.4) with  $q = 1, r = 3, \theta = \frac{3-p}{2p}$ . The continuity follows from the pointwise continuity of  $\rho \mapsto v_{xc}(\rho)$  together with the bounds (3.1).  $\square$

**Remark.** Our existence results do not require  $\Phi$  to be the KS ground state, but only to satisfy a fast enough decay property, like

$$\exists \gamma > 0 \text{ s.t. } e^{\gamma|\cdot|}\varphi_j \in H^1(\mathbb{R}^3), \quad \forall j \in \{1, \dots, n\}. \tag{3.5}$$

By the results of [1], whenever  $Z \geq N = 2n$  and  $e_{xc}$  satisfies our assumptions (3.1), then there exists a KS ground state  $\Phi = (\varphi_1, \dots, \varphi_n)$  and any such ground state fulfills the additional property (3.5).

**Theorem 1** (Existence of HOMO–LUMO Excitations). *For any admissible nuclear charge distribution  $\mu \in \mathcal{A}_{nuc}$  and any set of orbitals  $\Phi = (\varphi_1, \dots, \varphi_n) \in \mathcal{A}$  the excitation functional possesses a maximizer  $\varphi_H$  on  $\mathcal{A}_\Phi^H$  (i.e., a HOMO).*

*If additionally we have  $Z > N$ , i.e. a positively charged system, and  $\Phi \in \mathcal{A}$  satisfies the decay property (3.5), then there exists also a minimizer  $\varphi_L$  on  $\mathcal{A}_\Phi^L$  (i.e., a LUMO).*



**Proof.** Existence of a HOMO is elementary since  $\chi \mapsto \mathcal{E}_{\mu,\rho}[\chi]$  is continuous on  $H^1(\mathbb{R}^3)$  (see Lemma 2) and the admissible set  $A_{\Phi}^H$  is compact since it is a closed and bounded subset of a finite dimensional space. The case of a LUMO is more difficult. Proving existence for more advanced excitations models would require concentration compactness arguments [14,15]. But since here we consider only a frozen core approximation with single orbital excitations, a simpler reasoning is possible. We start off with the following lemma:

**Lemma 3** (Negativity of the LUMO Energy). *Under the same assumptions in Theorem 1 the LUMO energy is strictly negative, i.e.*

$$-\infty < \varepsilon_L = \inf_{\chi \in \mathcal{A}_{\Phi}^L} \langle \chi, h_{\mu,\rho} \chi \rangle < 0. \quad (3.6)$$

**Proof.** The fact that  $\varepsilon_L > -\infty$  follows by the bounds in Lemma 2, so it only remains to prove the right inequality. For a given  $\Phi = (\varphi_1, \dots, \varphi_n)$ , define  $\gamma_{\Phi} := \sum_{k=1}^n |\varphi_k\rangle \langle \varphi_k|$ , the projection onto the span of the  $\varphi_k$ .

Take a radially symmetric function  $\psi \in C_c^{\infty}(\mathbb{R}^3)$  with  $\text{supp} \psi \subseteq B_K^c$  for some radius  $K > 0$  and  $\|\psi\| = 1$ . Here  $B_K^c$  denotes the complement of the ball  $B_K$  of radius  $K > 0$  around the origin. Then define

$$\psi_{\sigma}(x) = \sigma^{3/2} \psi\left(\sigma\left(x - \frac{1}{\sqrt{\sigma}} \hat{e}\right)\right), \quad \text{with } \hat{e} \text{ some unit vector in } \mathbb{R}^3 \text{ and } \sigma > 0.$$

We have that

$$\langle \varphi_k, \psi_{\sigma} \rangle = \int_{\mathbb{R}^3} \varphi_k(x) \sigma^{3/2} \psi\left(\sigma\left(x - \frac{1}{\sqrt{\sigma}} \hat{e}\right)\right) dx = \sigma^{-3/2} \int_{B_K^c} \psi(y) \varphi_k\left(\frac{y}{\sigma} + \frac{1}{\sqrt{\sigma}} \hat{e}\right) dy = O(\sigma^{-3/2} \exp(-\frac{c}{\sigma})),$$

where we have used  $\text{supp} \psi \subseteq B_K^c$ , the exponential decay of the KS orbitals, and the estimate

$$\left| \frac{y}{\sigma} + \frac{1}{\sqrt{\sigma}} \hat{e} \right| \geq \left| \left| \frac{y}{\sigma} \right| - \left| \frac{1}{\sqrt{\sigma}} \hat{e} \right| \right| \geq \frac{c}{\sigma}, \quad \text{for } y \in B_K^c, \sigma \text{ small enough, and some constant } c > 0.$$

Hence  $\langle \varphi_k, \psi_{\sigma} \rangle$  decays exponentially for  $\sigma \rightarrow 0$ , so it is negligible up to higher order terms, i.e.

$$|(Id - \gamma_{\Phi})\psi_{\sigma}|^2 = |\psi_{\sigma}|^2 + O(\exp(-\frac{c}{\sigma})) \quad \text{and} \quad |\nabla(Id - \gamma_{\Phi})\psi_{\sigma}|^2 = |\nabla\psi_{\sigma}|^2 + O(\exp(-\frac{c}{\sigma})).$$

We can now estimate the energy of  $(Id - \gamma_{\Phi})\psi_{\sigma}$  as follows:

$$\mathcal{E}_{\mu,\rho} \left[ \frac{(Id - \gamma_{\Phi})\psi_{\sigma}}{\|(Id - \gamma_{\Phi})\psi_{\sigma}\|} \right] = \frac{1}{2} \int_{\mathbb{R}^3} |\nabla\psi_{\sigma}(x)|^2 dx + \int_{\mathbb{R}^3} (v_{xc} + v_{ext} + v_H) |\psi_{\sigma}(x)|^2 dx + O(\exp(-\frac{c}{\sigma})).$$

Since we want to estimate the energy from above, we do not need to consider the exchange–correlation term since it only gives a negative contribution. The kinetic energy is easily estimated:

$$2T[\psi_{\sigma}] = \int_{\mathbb{R}^3} |\nabla\psi_{\sigma}|^2 dx = \sigma^2 \int_{\mathbb{R}^3} |\nabla\psi|^2 dx = O(\sigma^2).$$

The next task is to estimate the Hartree and external potential term. We have

$$\begin{aligned} \int_{\mathbb{R}^3} v_H |\psi_{\sigma}|^2 dx &= \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x) |\psi_{\sigma}(y)|^2}{|x - y|} dx dy = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x) |\psi(z)|^2}{|x - (\frac{z}{\sigma} + \frac{1}{\sqrt{\sigma}} \hat{e})|} dx dz \\ &= \sigma \int_{\mathbb{R}^3} \rho(x) \int_{\mathbb{R}^3} \frac{|\psi(z)|^2}{\max\{\sigma|x|, |z + \sqrt{\sigma} \hat{e}|\}} dz dx, \end{aligned}$$

where we have used the radial symmetry of  $\psi$  in the last step. The same steps transform the external potential term into

$$\int_{\mathbb{R}^3} v_{ext} |\psi_{\sigma}|^2 dx = -\sigma \int_{\mathbb{R}^3} d\mu(x) \int_{\mathbb{R}^3} \frac{|\psi(z)|^2}{\max\{\sigma|x|, |z + \sqrt{\sigma} \hat{e}|\}} dz.$$

Since the support of  $\mu$  is compact and  $\text{supp } \psi \subseteq B_K^c$ , for  $\sigma$  small enough the term in the last integral satisfies  $\max\{\sigma|x|, |z + \sqrt{\sigma}\hat{e}|\} = |z + \sqrt{\sigma}\hat{e}|$ .

Putting everything together, we obtain for  $\sigma$  small enough

$$\begin{aligned} \varepsilon_L &\leq \mathcal{E}_{\mu,\rho} \left[ \frac{(Id - \gamma_\Phi)\psi_\sigma}{\|(Id - \gamma_\Phi)\psi_\sigma\|} \right] \leq \int_{\mathbb{R}^3} (v_H + v_{ext})|\psi_\sigma|^2 dx + O(\sigma^2) \\ &= \sigma \left( -Z \int_{\mathbb{R}^3} \frac{|\psi(z)|^2}{|z + \sqrt{\sigma}\hat{e}|} dz + \int_{\mathbb{R}^3} \rho(x) \int_{\mathbb{R}^3} \frac{|\psi(z)|^2}{\max\{\sigma|x|, |z + \sqrt{\sigma}\hat{e}|\}} dz dx \right) + O(\sigma^2) \\ &\leq \sigma(2n - Z) \int_{\mathbb{R}^3} \frac{|\psi(z)|^2}{|z + \sqrt{\sigma}\hat{e}|} dz + O(\sigma^2) < 0. \end{aligned}$$

In this last inequality the assumption of a positively charged system  $Z > N$  is crucial.  $\square$

Now take a minimizing sequence  $(\chi_n)_n$ . Then by the bounds established in Lemma 2 we get for some  $\chi \in H^1(\mathbb{R}^3)$  that

$$\chi_n \rightharpoonup \chi \text{ weakly in } H^1(\mathbb{R}^3).$$

Since again by Lemma 2 the mapping  $\chi \mapsto \mathcal{E}_{\mu,\rho}[\chi]$  is weakly lower semicontinuous, the only issue which could arise is that some mass gets lost in the limit. So see that this cannot happen, note that  $\chi = 0$  implies by lower semicontinuity

$$\mathcal{E}_{\mu,\rho}[\chi] = 0 \leq \liminf_{n \rightarrow \infty} \mathcal{E}_{\mu,\rho}[\chi] = \varepsilon_L,$$

contradicting Lemma 3. And if  $\|\chi\| \in (0, 1)$ , we define  $\psi = \frac{\chi}{\|\chi\|}$  and obtain a trial function with strictly lower energy, since

$$\mathcal{E}_{\mu,\rho}[\psi] = \frac{1}{\|\chi\|^2} \mathcal{E}_{\mu,\rho}[\chi] < \mathcal{E}_{\mu,\rho}[\chi] \leq \varepsilon_L,$$

which cannot happen since  $\psi$  is an admissible trial function. Thus  $\|\chi\| = 1$  and hence  $\chi \in \mathcal{A}_\Phi^L$ , which finishes the proof.  $\square$

### 3.2. Higher excitations

We now complete the rigorous justification of the picture on the left-hand side of Fig. 1, i.e. we fully characterize the spectrum of the KS Hamiltonian in the positively charged case.

**Theorem 2** (*Spectrum of KS Hamiltonian in the Case  $Z > N$* ). *Consider  $Z > N$ , i.e. a positively charged system. Then the KS Hamiltonian  $h_{\mu,\rho}$  given in (2.11) has infinitely many negative eigenvalues of finite multiplicity below the bottom of its essential spectrum,*

$$\sigma(h_{\mu,\rho}) = \{\lambda_n\}_{n \geq 1} \cup [0, \infty), \quad \text{with } \lambda_n < 0 \text{ and } \lambda_n \xrightarrow{n \rightarrow \infty} 0. \tag{3.7}$$

**Proof.** Due to assumption (3.1) the potential  $v_{ext} + v_H + v_{xc}$  belongs to the space  $L^2(\mathbb{R}^3) + L_\varepsilon^\infty(\mathbb{R}^3)$ , consisting of potentials  $v$  which, for any given  $\varepsilon > 0$ , can be decomposed as  $v = v_1 + v_2$  with  $\|v_1\|_2 < \infty$ ,  $\|v_2\|_\infty < \varepsilon$ . Therefore the potential is relatively compact w.r.t. the Laplacian. Hence by Weyl's Theorem (see e.g. Chapter XIII.4 of [21])  $h_{\mu,\rho}$  is self-adjoint and we have  $\sigma_{ess}(h_{\mu,\rho}) = \sigma_{ess}(-\Delta) = [0, \infty)$ . Furthermore,  $h_{\mu,\rho}$  is bounded from below. To see this, write  $h_{\mu,\rho} = -\Delta + U_2 + U_\infty$  with  $U_2 \in L^2(\mathbb{R}^3)$  and  $U_\infty \in L_\varepsilon^\infty(\mathbb{R}^3)$ . Then for every  $\psi \in H^1(\mathbb{R}^3)$  with  $\|\psi\|_2 = 1$  we have

$$\langle \psi, h_{\mu,\rho} \psi \rangle \geq \frac{1}{2} \|\nabla \psi\|_2^2 - \|U_2\|_2 \|\psi\|_4^2 - \|U_\infty\|_\infty. \tag{3.8}$$

Now the Sobolev embedding  $\|\varphi\|_4^2 \leq C\|\nabla\varphi\|_2^{3/2}$  and Young's inequality give the asserted lower bound. Now note that

$$h_{\mu,\rho} \leq -\frac{1}{2}\Delta + v_{ext} + v_H \quad (3.9)$$

and since by assumption  $Z > N$  we can apply Lemma II.1 of [16] which gives us that the right-hand side operator of (3.9) is negative on an infinite-dimensional subspace. Hence so is  $h_{\mu,\rho}$ . The min–max principle now gives that it has infinitely many negative eigenvalues. This completes the proof.  $\square$

#### 4. Optimal HOMO–LUMO excitations

In [9], motivated by the design of photovoltaic materials, results are given for *optimal* HOMO–LUMO excitations with respect to some control goals. The control is the nuclear charge distribution (e.g., the doping profile or the heteroatom substitutions); typical control goals are the spatial electron–hole charge separation or the size of the HOMO–LUMO gap. The analysis in [9] relies on the simplifying assumption of bounded domains. Here we generalize this analysis to unbounded domains. The main difficulty again consists in handling loss of mass at infinity.

**Lemma 4** (*Analytic Properties of the Set of HOMO–LUMO Excitations*). *Consider a positively charged system, i.e.  $Z > N$ , then the joint solution set to the governing variational principles for occupied KS orbitals, HOMO and LUMO parametrized by the set of nuclear charge distributions  $\mu$ ,*

$$\mathcal{B} = \{(\Phi, \varphi_H, \varphi_L, \mu) : \mu \in \mathcal{A}_{nuc}, (\Phi, \varphi_H, \varphi_L) \text{ defined by (3.3)}\}$$

has the following properties:

- (a)  $\mathcal{B}$  is weak  $\times$  weak  $\times$  weak  $\times$  weak\*-closed in  $H^1(\mathbb{R}^3)^n \times H^1(\mathbb{R}^3) \times H^1(\mathbb{R}^3) \times \mathcal{M}$
- (b)  $\mathcal{B}$  is strong  $\times$  strong  $\times$  strong  $\times$  weak\*-compact in  $H^1(\mathbb{R}^3)^n \times H^1(\mathbb{R}^3) \times H^1(\mathbb{R}^3) \times \mathcal{M}$

**Proof. Part (a)** Let  $(\Phi^{(\nu)}, \varphi_H^{(\nu)}, \varphi_L^{(\nu)}) \rightharpoonup (\Phi, \varphi_H, \varphi_L)$  in  $H^1(\mathbb{R}^3)^{(n+2)}$  and  $\mu^{(\nu)} \rightharpoonup^* \mu$  in  $\mathcal{M}$ , then we need to prove

$$(i) \mu \in \mathcal{A}_{nuc} \quad (ii) \Phi \in \underset{\mathcal{A}}{\operatorname{argmin}} \mathcal{E}_\mu \quad (iii) \varphi_H \in \underset{\mathcal{A}_\Phi^H}{\operatorname{argmax}} \mathcal{E}_{\mu,\rho} \quad (iv) \varphi_L \in \underset{\mathcal{A}_\Phi^L}{\operatorname{argmin}} \mathcal{E}_{\mu,\rho}.$$

**Ad (i):** Since all measures  $\mu^{(\nu)}$  are supported on the compact set  $\Omega_{nuc}$ , the constant functions are in the predual of  $\mathcal{M}$ , that is the space of the continuous functions on  $\Omega_{nuc}$ , and hence  $Z = \mu^{(\nu)}(\Omega_{nuc}) \rightarrow \mu(\Omega_{nuc})$ , so  $\mu \in \mathcal{A}_{nuc}$ .

**Ad (ii):** For any admissible  $\Psi \in \mathcal{A}$  we have by the variational definition of  $\Phi^{(\nu)}$  and the weak\* continuity of  $\mu \mapsto \mathcal{E}_\mu[\Psi]$

$$\mathcal{E}_{\mu^{(\nu)}}[\Phi^{(\nu)}] \leq \mathcal{E}_{\mu^{(\nu)}}[\Psi] \xrightarrow{\nu \rightarrow \infty} \mathcal{E}_\mu[\Psi], \quad (4.1)$$

which implies

$$\limsup_{\nu \rightarrow \infty} \mathcal{E}_{\mu^{(\nu)}}[\Phi^{(\nu)}] \leq \inf_{\mathcal{A}} \mathcal{E}_\mu, \quad (4.2)$$

since  $\Psi \in \mathcal{A}$  was arbitrary.

Using the weak  $\times$  weak\* lower semicontinuity of  $(\Phi, \mu) \mapsto \mathcal{E}_\mu[\Phi]$  on  $H^1(\mathbb{R}^3)^n \times \mathcal{M}$  gives

$$\liminf_{\nu \rightarrow \infty} \mathcal{E}_{\mu^{(\nu)}}[\Phi^{(\nu)}] \geq \mathcal{E}_\mu[\Phi], \quad (4.3)$$

but unlike in the bounded domain case this does not directly give us the result since  $\mathcal{A}$  is not weakly closed.

Fortunately, the  $\Phi^{(\nu)}$  are not arbitrary elements of  $\mathcal{A}$ . Since the energy functional is invariant under unitary transformations, we can always assume the orbitals  $\Phi = (\varphi_1, \dots, \varphi_n)$  to be orthogonal. So the only thing that could go wrong is loss of mass due to weak convergence.

Assume for contradiction  $\alpha = \|\Phi\| < n$  so we lose mass in at least one orbital  $\varphi_k$ . In order to see that this is not possible, we will place some small mass at a large but finite distance, and obtain a state with lower energy.

We take  $\eta \in C_c^\infty(\mathbb{R}^3)$  with  $\|\eta\|_2 = 1$  and consider the test function  $\eta_{\lambda,\sigma} := \sigma^{1/2} \lambda^{3/2} \eta(\lambda \cdot)$ .

Then

$$\mathcal{E}_\mu[\eta_{\lambda,\sigma}] = \sigma \lambda^2 T[\eta] + V[|\eta_{\lambda,\sigma}|^2] + \sigma^2 \lambda J[|\eta|^2] + E_{xc}[|\eta_{\lambda,\sigma}|^2].$$

Using the assumption (3.2) and the fact that  $V[|\eta|^2] \leq 0$ , we obtain that for  $\sigma$  small enough there exists a constant  $c > 0$  such that

$$\mathcal{E}_\mu[\eta_{\lambda,\sigma}] \leq \sigma \lambda^2 T[\eta] + \sigma^2 \lambda J[|\eta|^2] - c \sigma^q \lambda^{3(q-1)} \int_{\mathbb{R}^3} |\eta|^{2q} dx.$$

Since  $q < \frac{3}{2}$ , the negative term dominates in the limit  $\sigma, \lambda \rightarrow 0$  (if we let both go to 0 at the same speed). Hence by choosing the parameters  $\lambda$  and  $\sigma$  small enough we ensure  $\mathcal{E}_\mu[\eta_{\lambda,\sigma}] < 0$ .

Now since we assume loss of mass in the  $k$ th orbital  $\varphi_k$ , let us consider  $\tilde{\varphi}_k^{(n)}(\cdot) = \varphi_k(\cdot) + \eta_{\lambda,\sigma}(\cdot - n\vec{e})$ , where  $\vec{e}$  is some unit vector in  $\mathbb{R}^3$ . Denoting the orbitals with  $\varphi_k$  replaced by  $\tilde{\varphi}_k$  by  $\tilde{\Phi}$ , we get

$$\mathcal{E}_\mu[\tilde{\Phi}] \leq \mathcal{E}_\mu[\Phi] + T[\eta_{\lambda,\sigma}] + J[|\eta_{\lambda,\sigma}|^2] + E_{xc}[|\eta_{\lambda,\sigma}|^2] + o(1) < \mathcal{E}_\mu[\Phi] + o(1).$$

So for  $n$  large enough we obtain  $\mathcal{E}_\mu[\tilde{\Phi}] < \mathcal{E}_\mu[\Phi]$  and  $\|\tilde{\Phi}\| > \|\Phi\|$ .

We can now repeat these steps until we have constructed a  $\Psi$  with  $\Psi \in \mathcal{A}$  (after a suitable unitary transformation), and arrive at the contradiction

$$\mathcal{E}_\mu[\Psi] < \mathcal{E}_\mu[\Phi] \stackrel{(4.2), (4.3)}{\leq} \inf_{\mathcal{A}} \mathcal{E}_\mu.$$

Hence there is no loss of mass,  $\Phi \in \mathcal{A}$ , and (ii) holds.

Before we move on to (iii) and (iv) we mention that due to the upper and lower bound above,  $\mathcal{E}_{\mu^{(\nu)}}[\Phi^{(\nu)}] \rightarrow \mathcal{E}_\mu[\Phi]$ . But this gives us  $T[\Phi^{(\nu)}] \rightarrow T[\Phi]$  since the other terms in the energy functional, namely  $V, J_h, E_{xc}$ , are continuous on  $L^2 \cap L^4$ . From this we infer strong convergence in  $L^p$  for  $p \in [2, 6)$ , since there is no loss of mass.

Hence the kinetic energy converges as well, which means  $\|\nabla \Phi^{(\nu)}\|_2 \rightarrow \|\nabla \Phi\|_2$ , giving us  $\nabla \Phi^{(\nu)} \rightarrow \nabla \Phi$  in  $L^2$ , so  $\Phi^{(\nu)} \rightarrow \Phi$  in  $H^1(\mathbb{R}^3)^n$ .

**Ad (iii):** As in [9] the statements (iii) and (iv) are more difficult since the HOMO and LUMO orbitals  $\varphi_H^{(\nu)}$  and  $\varphi_L^{(\nu)}$  are not defined via universal but  $\varphi^{(\nu)}$ -dependent sets and hence an admissible trial function  $\psi$  for the limiting HOMO and LUMO orbitals  $\varphi_H$  and  $\varphi_L$  may not be admissible for the variational principle for the approximating orbitals. In short, our argument in (4.1) is not valid anymore.

Hence we need to look at the  $L^2$ -projector of the sequence  $\Phi^{(\nu)} = (\varphi_1^{(\nu)}, \dots, \varphi_n^{(\nu)})$ , i.e.  $\gamma_{\Phi^{(\nu)}} \chi := \sum_{k=1}^n \langle \varphi_k^{(\nu)}, \chi \rangle \varphi_k^{(\nu)}$ . For any given  $\chi \in L^2(\mathbb{R}^3)$  the mapping  $\Phi \mapsto \gamma_\Phi \chi$  is strongly continuous from  $H^1(\mathbb{R}^3)^n$  to  $H^1(\mathbb{R}^3)$ , i.e.

$$\gamma_{\Phi^{(\nu)}} \chi \rightarrow \gamma_\Phi \chi \text{ in } H^1(\mathbb{R}^3), \quad \|\gamma_{\Phi^{(\nu)}} \chi\|_2 \rightarrow \|\gamma_\Phi \chi\|_2. \tag{4.4}$$

Furthermore by definition we have  $\gamma_\Phi \chi = \chi$  and  $\|\chi\|_2 = 1$  for any  $\chi \in \mathcal{A}_\Phi^H$ . So by (4.4) we have  $\|\gamma_{\Phi^{(\nu)}} \chi\|_2 > 0$  for all  $\nu$  large enough and therefore by the variational principle for the HOMO (3.3)

$$\mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}}[\varphi_H^{(\nu)}] \geq \mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}} \left[ \frac{\gamma_{\Phi^{(\nu)}} \chi}{\|\gamma_{\Phi^{(\nu)}} \chi\|_2} \right] = \frac{1}{\|\gamma_{\Phi^{(\nu)}} \chi\|_2^2} \mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}}[\gamma_{\Phi^{(\nu)}} \chi] \rightarrow 1 \cdot \mathcal{E}_{\mu, \Phi}[\chi]. \tag{4.5}$$

Here we used the strong convergence of (4.4) and the fact that the map  $(\Phi, \chi, \mu) \mapsto \mathcal{E}_{\mu, \rho}[\chi]$  is weak  $\times$  strong  $\times$  weak\* continuous on  $H^1(\mathbb{R}^3)^n \times H^1(\mathbb{R}^3) \times \mathcal{M}$  due to Lemma 2.

Since (4.5) holds for any  $\chi \in \mathcal{A}_\Phi^H$ , we have

$$\liminf_{\nu \rightarrow \infty} \mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}}[\varphi_H^{(\nu)}] \geq \sup_{\chi \in \mathcal{A}_\Phi^H} \mathcal{E}_{\mu, \Phi}[\chi]. \quad (4.6)$$

Next we prove that  $\varphi_H \in A_\Phi^H$ . Since by definition the  $\varphi_H^{(\nu)}$  lie in the span of  $\Phi^{(\nu)}$ , we obtain by the weak convergence of  $\varphi_H^{(\nu)}$  in  $H^1(\mathbb{R}^3)$  and the strong convergence of  $\Phi^{(\nu)}$  in  $H^1(\mathbb{R}^3)^n$  that

$$\varphi_H^{(\nu)} = \gamma_{\Phi^{(\nu)}} \varphi_H^{(\nu)} = \sum_{k=1}^n \langle \varphi_k^{(\nu)}, \varphi_H^{(\nu)} \rangle \varphi_k^{(\nu)} \longrightarrow \sum_{k=1}^n \langle \varphi_k, \varphi_H \rangle \varphi_k = \gamma_\Phi \varphi_H.$$

So taking the limit yields  $\varphi_H = \gamma_\Phi \varphi_H$  and hence  $\varphi_H^{(\nu)} \rightarrow \varphi_H$  strongly, so  $\varphi_H \in \mathcal{A}_\Phi^H$ .

Furthermore by the continuity properties proven in Lemma 2 we obtain

$$\lim_{\nu \rightarrow \infty} \mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}}[\varphi_H^{(\nu)}] = \mathcal{E}_{\mu, \Phi}[\varphi_H]. \quad (4.7)$$

Combining (4.6) and (4.7) yields (iii).

**Ad (iv):**

The corresponding proof for the LUMO starts similarly. Take any  $\chi \in \mathcal{A}_\Phi^L$ , i.e.  $\gamma_\Phi \chi = 0$  and  $\|\chi\|_2 = 1$ . Then again by (4.4) and the variational principle for the LUMO (3.3) we obtain

$$\mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}}[\varphi_L^{(\nu)}] \leq \mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}} \left[ \frac{(I - \gamma_\Phi^{(\nu)})\chi}{\|(I - \gamma_\Phi^{(\nu)})\chi\|_2} \right] = \frac{1}{\|(I - \gamma_\Phi^{(\nu)})\chi\|_2^2} \mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}}[\gamma_\Phi^{(\nu)} \chi] \rightarrow 1 \cdot \mathcal{E}_{\mu, \Phi}[\chi]. \quad (4.8)$$

Minimizing over  $\chi \in A_\Phi^L$  gives

$$\limsup_{\nu \rightarrow \infty} \mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}}[\varphi_L^{(\nu)}] \leq \inf_{\chi \in \mathcal{A}_\Phi^L} \mathcal{E}_{\mu, \Phi}[\chi]. \quad (4.9)$$

For the lower bound we use the weak  $\times$  weak  $\times$  weak\* lower semicontinuity of the map  $(\Phi, \mu, \chi) \mapsto \mathcal{E}_{\mu, \Phi}[\chi]$  by Lemma 2

$$\mathcal{E}_{\mu, \Phi}[\varphi_L] \leq \liminf_{\nu \rightarrow \infty} \mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}}[\varphi_L^{(\nu)}]. \quad (4.10)$$

Unfortunately we are not done since the  $\varphi_L^{(\nu)}$  are not known to converge strongly but just weakly, so we still need to prove that  $\varphi_L$  is admissible. The limit  $\varphi_L$  lies in the orthogonal complement of the  $(\varphi_k)_{k=1}^n$ , since  $\Phi^{(\nu)}$  converges strongly and therefore

$$0 = \langle \varphi_L^{(\nu)}, \varphi_k^{(\nu)} \rangle \xrightarrow{\nu \rightarrow \infty} \langle \varphi_L, \varphi_k \rangle \quad \forall k \in \{1, \dots, n\}.$$

So we only need to prove  $\|\varphi_L\|_2 = 1$ . Assume  $\|\varphi_L\|_2 < 1$ . Then by (4.9) and (4.10)

$$I_1 = \inf_{\chi \in \mathcal{A}_\Phi^L} \mathcal{E}_{\mu, \Phi}[\chi] \geq \mathcal{E}_{\mu, \Phi}[\varphi_L] = \mathcal{E}[\varphi_L] \geq I_{\|\varphi_L\|}. \quad (4.11)$$

But we proved in Lemma 3 that the mass-to-LUMO-energy map  $\lambda \mapsto I_\lambda$  is strictly decreasing. Hence  $\|\varphi_L\|_2 = 1$ , so  $\varphi_L$  is admissible and together with (4.9) and (4.10) this establishes (iv).

Before we prove part (b) let us make some remarks. Since there is no loss of mass,  $\varphi_L^{(\nu)}$  converges strongly in  $L^2(\mathbb{R}^3)$  and hence strongly in  $L^p(\mathbb{R}^3)$  for  $p \in [2, 6)$ . But since  $\mathcal{E}_{\mu^{(\nu)}, \Phi^{(\nu)}}[\varphi_L^{(\nu)}] \rightarrow \mathcal{E}_{\mu, \Phi}[\varphi_L]$  and all terms except the kinetic energy converge due to the continuity results in Lemma 2, we must also have  $T[\varphi_L^{(\nu)}] \rightarrow T[\varphi_L]$ , i.e.  $\varphi_L^{(\nu)}$  converges strongly in  $H^1(\mathbb{R}^3)$ .

In conclusion, we know that  $(\Phi^{(\nu)}, \varphi_H^{(\nu)}, \varphi_L^{(\nu)})$  converges strongly in  $H^1(\mathbb{R}^3)^{(n+2)}$ .

**Part (b)**

To prove sequential compactness of the set  $\mathcal{B}$  is now quite easy since by part (a) and Banach–Alaoglu we just need to prove that any sequence  $(\Phi^{(\nu)}, \varphi_H^{(\nu)}, \varphi_L^{(\nu)}, \mu^{(\nu)}) \in \mathcal{B}$  is bounded in  $H^1(\mathbb{R}^3)^{(n+2)} \times \mathcal{M}$ . For  $\Phi^{(\nu)}$  and  $\varphi_L^{(\nu)}$  this follows from the bounds in Lemmas 1 and 2, respectively, noting that the exponent  $p \in [1, \frac{5}{3})$  in the assumption on  $v_{xc}$  has the property that the exponents  $\frac{3(p-1)}{2}$  and  $\frac{3(p-1)}{2p}$  of  $\|\rho\|_3$  are strictly less than 1. Since the  $\Phi^{(\nu)}$  stay bounded, so do the  $\varphi_H^{(\nu)}$ . Lastly, since  $\mu^{(\nu)} \geq 0$  we have  $\|\mu^{(\nu)}\|_{\mathcal{M}} = \int_{\mathbb{R}^3} d\mu^{(\nu)} = Z$ , which concludes the proof.  $\square$

As an example of a control goal we consider bandgap tuning as introduced in [9]. Here the quantity which one wants to influence by a suitable choice of the nuclear charge distribution  $\mu$  is the HOMO–LUMO bandgap  $\varepsilon_H - \varepsilon_L$ , where  $\varepsilon_H$  and  $\varepsilon_L$  stand for the HOMO and LUMO eigenvalues of the KS Hamiltonian (2.11). Any bandgap tuning functional promoting a desired target value  $\varepsilon_*$  has to reach its minimum when  $\varepsilon_H - \varepsilon_L = \varepsilon_*$ . A simple choice suggested in [9] is

$$J[\Phi, \varphi_H, \varphi_L, \mu] = |\varepsilon_L - \varepsilon_H - \varepsilon_*|^2 = |\mathcal{E}_{\mu, \rho}[\varphi_L] - \mathcal{E}_{\mu, \rho}[\varphi_H] - \varepsilon_*|^2. \tag{4.12}$$

**Theorem 3.** For any  $\varepsilon_* > 0$  and for  $Z > N$ , there exists a nuclear charge distribution  $\mu \in \mathcal{A}_{nuc}$  which minimizes the bandgap tuning functional (4.12) over  $\mathcal{A}_{nuc}$  subject to the constraints (3.3).

**Proof.** The bandgap functional  $J$  in (4.12) is, due to Lemma 2, weak  $\times$  strong  $\times$  strong  $\times$  weak\* continuous on  $(H^1)^n \times H^1 \times H^1 \times \mathcal{M}$ . Hence by the compactness of the set  $\mathcal{B}$  proven in Lemma 4 it attains its minimum over this set.  $\square$

**5. Nonexistence of HOMO–LUMO excitations in the neutral case  $Z = N$**

We now introduce carefully chosen and realistic model densities  $\rho$  and prove that the excitation functional  $\mathcal{E}_{\mu, \rho}$  admits no excited states, i.e. no bound states other than the ground state. See the picture on the right in Fig. 1.

This finding suggests that also for the true KS ground state density  $\rho$ , it may happen that there are no exact HOMO–LUMO excitations. Of course, the Hamiltonian possesses continuous spectrum above the ground state energy and therefore “metastable” excitations (suitable square-integrable superpositions of continuous eigenstates) still exist.

From a mathematical point of view, the results in this section show that the assumption  $Z > N$  in our existence result (Theorem 1) was in fact sharp and cannot be weakened to  $Z \geq N$ . Note that the model densities considered here satisfy the assumptions of Theorem 1, in particular (3.5).

*H-atom ground state density*

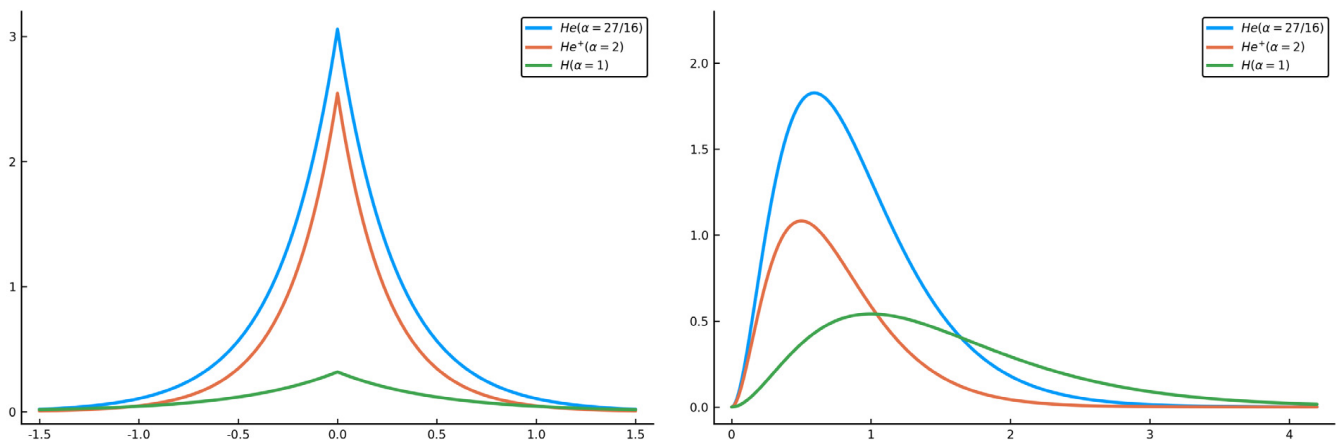
Let  $\mu = \delta_0$  and let  $\varphi_H(x) = \frac{1}{\sqrt{\pi}}e^{-|x|}$  be the hydrogen atom ground state for the Schrödinger equation, i.e. the lowest eigenfunction of  $-\frac{1}{2}\Delta - \frac{1}{|\cdot|}$ . Its density is

$$\rho_H(x) = |\varphi_H|^2(x) = \frac{1}{\pi}e^{-2|x|}. \tag{5.1}$$

We expect this to be a good approximation for the KS density, hence the KS operator  $h_{\rho_H}$  should be a good approximation to the self-consistent hydrogen KS operator.

In this case the Hartree-potential  $v_H$  can be explicitly computed. The well known result is

$$v_H = \int_{\mathbb{R}^3} \frac{1}{|x-y|} \rho(y) dy = \frac{1}{|x|} - e^{-2|x|} \left( 1 + \frac{1}{|x|} \right).$$



**Fig. 2.** Comparison of the density (left) and radial density (right) of hydrogen, of the Helium ion  $\text{He}^+$ , and of the model (5.2) for Helium.

The first term cancels the external potential  $v_{ext}$ , hence the excitation functional becomes

$$\mathcal{E}[\chi] = \frac{1}{2} \int_{\mathbb{R}^3} |\nabla \chi|^2 + \int_{\mathbb{R}^3} |\chi|^2 \left( -e^{-2|x|} \left( 1 + \frac{1}{|x|} \right) + v_{xc} \left( \frac{1}{\pi} e^{-2|x|} \right) \right)$$

with corresponding Hamiltonian

$$h_{\rho_H} = -\frac{1}{2} \Delta + V(x) = -\frac{1}{2} \Delta - e^{-2|x|} \left( 1 + \frac{1}{|x|} \right) + v_{xc} \left( \frac{1}{\pi} e^{-2|x|} \right).$$

*Model ground state density for the He-atom*

In order to construct a model density for Helium, we make the following ansatz.

We take a dilated version of the hydrogen orbital, i.e.

$$\varphi_\alpha(x) = \frac{\alpha^{3/2}}{\sqrt{\pi}} e^{-\alpha|x|} \text{ and } \rho_\alpha = 2|\varphi_\alpha|^2, \tag{5.2}$$

and – following Hans Bethe – determine the parameter  $\alpha$  by

$$\alpha = \underset{\beta > 0}{\operatorname{argmin}} E_\beta, \quad E_\beta = 2T[\varphi_\beta] + 2V_{ne}^{He}[\varphi_\beta] + \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{|\varphi_\beta|^2(x)|\varphi_\beta|^2(y)}{|x-y|} dx dy.$$

The last term describing the electron–electron interaction comes from

$$V_{ee}[[\vec{\psi}_1 \vec{\psi}_2]] = \sum_{i < j} \left( \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{|\vec{\psi}_i|^2(x)|\vec{\psi}_j|^2(y)}{|x-y|} - \frac{(\vec{\psi}_i \cdot \vec{\psi}_j^*)(x)(\vec{\psi}_i^* \cdot \vec{\psi}_j)(y)}{|x-y|} \right),$$

with the spinors  $\vec{\psi}_1 = \begin{pmatrix} \varphi \\ 0 \end{pmatrix}$  and  $\vec{\psi}_2 = \begin{pmatrix} 0 \\ \varphi \end{pmatrix}$ .

The energy is easily computed as

$$E_\beta = \beta^2 - 4\beta + \frac{5}{8}\beta,$$

which implies

$$\alpha = \frac{27}{16} = 1.6875.$$

In Fig. 2 one sees that this value corresponds to the fact that the second electron does not see the full Coulomb potential of the nucleus but a screened one.

With this density the Hartree-potential can again be computed explicitly and plugging this density into our excitation functional gives us the Hamiltonian

$$h_{\rho_\alpha} = -\frac{1}{2}\Delta - 2e^{-2\alpha|x|} \left( \frac{1}{|x|} + \frac{1}{\alpha} \right) + v_{xc}(\rho_\alpha(x)).$$

With these two densities at hand we can state the main result of this section and complete the picture given in Fig. 1.

**Theorem 4** (Spectrum of the KS Hamiltonian in the Case  $Z = N$ ). Consider either the hydrogen atom ( $N = 1, \mu = \delta_0$ ) with the density  $\rho = \rho_H$  given by (5.1) or the helium atom ( $N = 2, \mu = 2\delta_0$ ) with the density  $\rho = \rho_\alpha$  given by (5.2). Furthermore let the exchange–correlation energy be given by either Dirac exchange, PW81 or PZ92. Then the spectrum of the KS Hamiltonian has the form

$$\sigma(h_{\mu,\rho}) = \{\varepsilon_0\} \cup [0, \infty), \quad \text{for some } \varepsilon_0 < 0. \tag{5.3}$$

In particular, the Hamiltonian possesses exactly one bound state (up to spin in the hydrogen case) and no excited states, i.e. no bound states above the ground state.

The result of Theorem 4 is quite significant from a computational point of view. In numerical methods one, of course, obtains excited states, but in the limit of complete basis sets in infinite volume these might dissolve into metastable states associated with the continuous spectrum.

**Proof.** As in the proof of Theorem 2 the potential of the KS Hamiltonian is in  $L^2 + L^\infty$ , hence in both cases  $\sigma_{ess}(h_{\mu,\rho}) = [0, \infty)$ .

Next, we prove that there is at least one bound state with eigenvalue  $\varepsilon_0 < 0$ . By the Rayleigh–Ritz method it suffices to find a  $\psi \in D(h_{\mu,\rho}) = H^1(\mathbb{R}^3)$  with  $\|\psi\|_2 = 1$  and  $\langle \psi, h_{\mu,\rho}\psi \rangle < 0$ . Since for any LDA functional we have  $e_{xc} \leq e_x$ , it suffices to prove the inequality for Dirac exchange. As a test function we choose the corresponding orbitals we used in the construction of our densities. These terms are easily computed to give for the hydrogen case

$$\varepsilon_0^H \leq \langle \varphi_H, h_{\delta_0,\rho_H}\psi_H \rangle = \frac{1}{2} - \frac{3}{8} - \left( \frac{3}{\pi^2} \right)^{1/3} \frac{27}{64} = -0.1587 < 0, \tag{5.4}$$

and for the helium atom with  $\alpha = \frac{27}{16}$

$$\varepsilon_0^\alpha \leq \langle \varphi_\alpha, h_{2\delta_0,\rho_\alpha}\psi_\alpha \rangle = \frac{\alpha^2}{2} - \frac{2\alpha^2 + 1}{4\alpha} - \left( \frac{6}{\pi^2} \right)^{1/3} \frac{27}{64}\alpha = -0.1711 < 0. \tag{5.5}$$

So for both atoms we have at least one bound state.

Now we use the upper bound on the number of bound states given in [10] which says that the number  $N_\ell$  of bound states with angular momentum  $\ell$  satisfies

$$N_\ell \leq (2\ell + 1)^{1-2p} I_p(V), \tag{5.6}$$

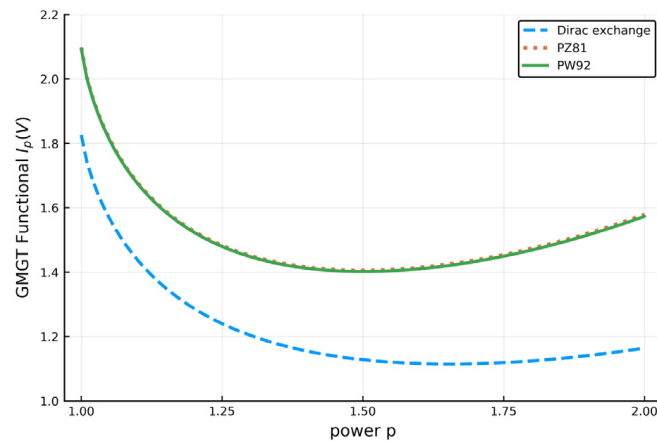
where the functional  $I_p(V)$  is given by

$$I_p(V) = C_p \int_{\mathbb{R}^3} \frac{1}{4\pi} |x|^{2p-3} (V_-(x))^p dx, \quad \text{with } C_p = \frac{(p-1)^{p-1} \Gamma(2p)}{p^p \Gamma(p)}. \tag{5.7}$$

For a radially symmetric potential this reduces to

$$I_p(V) = C_p \int_0^\infty \frac{1}{x} (x^2 V_-(x))^p dx. \tag{5.8}$$





**Fig. 3.** The Glaser–Martin–Grosse–Thirring functional (5.8) of the effective Kohn–Sham potential for different choices of the exchange–correlation functional and the approximate ground state density  $\rho_\alpha$  (5.2). Values strictly less than 2 mean that there is at most one bound state.

We call  $I_p$  the Glaser–Martin–Grosse–Thirring (GMGT) functional. Here  $V_-$  denotes the negative part of the potential, and the parameter  $p \geq 1$  for the radially symmetric case while for the general case we have the restriction  $p \geq \frac{3}{2}$ .

If one calculates this integral numerically for our hydrogen density (5.1) with Dirac exchange, one obtains that the minimum value is attained at  $p = 1.4$  with  $I_p(V) \approx 1.61587$ . This gives us the upper bound

$$N_\ell < \frac{1}{(2\ell + 1)^{1.8}} 1.61587,$$

which means  $N_0 \leq 1$  and all other  $N_\ell = 0$  for  $\ell \geq 1$ . So up to the degeneracy with respect to spin there is only the ground state and there are no excited states.

For the helium density (5.2) we computed the GMGT functional  $I_p(V)$  for the Dirac [7], the Perdew–Zunger [20] and the Perdew–Wang [19] exchange–correlation functional and obtained

$$\begin{aligned} I_p(V^D) &= 1.11465, \text{ at the value } p = 1.65, \\ I_p(V^{PZ}) &= 1.40558, \text{ at the value } p = 1.5, \\ I_p(V^{PW}) &= 1.40184, \text{ at the value } p = 1.5. \end{aligned}$$

Here we denote by  $V_-$  the potential  $V_-(x) = 2e^{-2\alpha|x|} \left( \frac{1}{|x|} + \frac{1}{\alpha} \right) + v_{xc}(\rho_\alpha(x))$ , so in e.g. the Dirac case we have

$$V_-(x) = 2e^{-2\alpha|x|} \left( \frac{1}{|x|} + \frac{1}{\alpha} \right) + \left( \frac{3}{\pi} \right)^{\frac{1}{3}} \frac{2\alpha}{\pi} e^{-2\alpha|x|}.$$

As before, it now follows from (5.6) and (5.8) that no bound state other than the ground state orbital exists.

Hence for both hydrogen and helium we have exactly one bound state – the ground state itself – which corresponds to an eigenfunction with eigenvalue  $\varepsilon_0 < 0$ .  $\square$

In Fig. 3 the values of  $I_p(V)$  as a function of  $p$  are given in the case of helium for the three LDAs mentioned above. The figure shows that the minimum value of the GMGT-functional is always attained at some  $p \geq 1.5$ , i.e. in the interval  $p \in [\frac{3}{2}, \infty)$ , where the upper bound is also valid for non-symmetric potentials. Hence, as long as the real KS-density is only a small perturbation of our model density  $\rho_\alpha$ , even if it were not symmetric, our results would hold.

The overall conclusion of this section is that there are no LUMO excitation for the model densities (5.1) and (5.2).

Our findings raise the following interesting questions which lie beyond the scope of the present paper. First, is the GMGT nonexistence criterion satisfied for numerically obtained ground state densities of hydrogen and helium, or more complex atoms and molecules? A second question is whether the absence of exact excitations persists for more advanced excitation models like the Casida ansatz.

## Acknowledgments

Support from the International Research Training Group IGDK Munich - Graz funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 188264188/GRK1754 is gratefully acknowledged.

## Appendix. Analytical properties of PZ81 and PW92

In the following we verify explicitly that the correlation functionals PZ81 and PW92 from the physics literature satisfy the mathematical assumptions (3.1) and (3.2) of this paper. In the physics literature the exchange–correlation functional is usually specified by the energy per particle at the density  $\rho$ , denoted  $\varepsilon_c(\rho)$ . We work with the mathematically convenient energy per unit volume  $e_{xc}(\rho) = \rho\varepsilon_c(\rho)$ . The exchange part was already discussed in Remark 3.1, so we only need to deal with the correlation part.

Furthermore, recall the Wigner–Seitz radius

$$r_s(x) := \left( \frac{3}{4\pi\rho(x)} \right)^{\frac{1}{3}}, \quad (\text{A.1})$$

which is a standard parameter in physics to describe the local electron density of a system. Lastly, we remark that in the following  $C$  will describe a generic constant, which may have different values at each appearance, but is independent of  $\rho$  and  $r_s$ .

### Perdew–Wang (PW92)

In this paper we consider only spin-unpolarized systems, so we have  $\zeta = \frac{n_\uparrow - n_\downarrow}{n_\uparrow + n_\downarrow} = 0$ . So (in the notation of the original paper) we only need to check the assumptions (3.1) and (3.2) for  $e_c(r_s, 0)$ . The PW92 correlation functional is given by

$$\varepsilon_c(r_s) = -2A(1 + \alpha_1 r_s) \log \left( 1 + \frac{1}{2A(\beta_1 r_s^{1/2} + \beta_2 r_s + \beta_3 r_s^{3/2} + \beta_4 r_s^2)} \right).$$

In order to improve readability of the arguments below, we define the following functions

$$f(r_s) := -2A(1 + \alpha_1 r_s), \quad g(r_s) := 2A(\beta_1 r_s^{1/2} + \beta_2 r_s + \beta_3 r_s^{3/2} + \beta_4 r_s^2).$$

So returning to the notation of our paper we need to check (3.1) and (3.2) for the function

$$e_c(\rho) = \rho \cdot \varepsilon_c(r_s(\rho)) = \rho f(r_s(\rho)) \log \left( 1 + \frac{1}{g(r_s(\rho))} \right).$$

This function is clearly continuously differentiable for  $\rho > 0$ , so we only check the limit  $\rho \rightarrow 0$ . Since  $f(r_s(\rho)) = O(r_s) = O(\rho^{-1/3})$  and  $g(r_s) = \Omega(\sqrt{r_s})$  for  $\rho \rightarrow 0$ , respectively  $r_s \rightarrow \infty$ , we get

$$\limsup_{\rho \rightarrow 0} |e_c(\rho)| \leq \limsup_{\substack{\rho \rightarrow 0 \\ r_s \rightarrow \infty}} C \rho^{2/3} \log(1 + C r_s^{-1/2}) = 0,$$

hence with  $e_c(0) = 0$ ,  $e_c$  is continuous.

Next we calculate the derivative

$$v_c(\rho) = \frac{d}{d\rho} e_c(\rho) = f(r_s) \log\left(1 + \frac{1}{g(r_s)}\right) + \rho \left(\frac{d}{d\rho} r_s\right) \log\left(1 + \frac{1}{g(r_s)}\right) \frac{d}{dr_s} f(r_s) + \rho f(r_s) \left(\frac{d}{d\rho} r_s\right) \frac{d}{dr_s} \log\left(1 + \frac{1}{g(r_s)}\right)$$

using  $\frac{d}{d\rho} r_s = -\frac{1}{3} \frac{r_s}{\rho}$  we obtain

$$v_c(\rho) = \log\left(1 + \frac{1}{g(r_s)}\right) \left[f(r_s) + \frac{2A\alpha_1}{3} r_s\right] + \frac{1}{3} r_s f(r_s) \frac{\frac{d}{dr_s} g(r_s)}{g(r_s) + g(r_s)^2}.$$

Since  $g(r_s)$  consists only of powers of  $r_s$  its derivative is  $\frac{d}{dr_s} g(r_s) = \frac{1}{r_s} \Theta(g(r_s))$  and we get

$$v_c(\rho) = \log\left(1 + \frac{1}{g(r_s)}\right) \left[f(r_s) + \frac{2A\alpha_1}{3} r_s\right] + f(r_s) \Theta\left(\frac{1}{g(r_s)+1}\right). \tag{A.2}$$

Applying the inequality  $\log(1+x) \leq x$  for  $x > -1$ , using  $\frac{1}{g(r_s)} = O(r_s^{-2})$  and taking the limit  $\rho \rightarrow 0, r_s \rightarrow \infty$  gives now

$$\limsup_{\rho \rightarrow 0} |e_c(\rho)| \leq \limsup_{\substack{\rho \rightarrow 0 \\ r_s \rightarrow \infty}} C \frac{1}{r_s^2} (1 + r_s) = 0.$$

So also  $v_c$  is continuous and hence it suffices to prove  $v_c \leq c_{xc}(1 + \rho^{p-1})$  for  $\rho \rightarrow \infty$ , i.e.  $r_s \rightarrow 0$ . Using  $f(r_s) \rightarrow C, g(r_s) \rightarrow 0$  and  $\frac{1}{g(r_s)} = O(r_s^{-1/2})$  we obtain

$$|v_c(\rho)| \leq C \left(1 + \log\left(1 + \frac{1}{g(r_s)}\right)\right) \leq C(1 + r_s^{-1/2}) = C(1 + \rho^{1/6}),$$

so (3.1) holds with  $p = \frac{7}{6}$ .

For (3.2) we can choose  $q = \frac{17}{12}$  (any value between  $\frac{4}{3}$  and  $\frac{3}{2}$  will do). Writing the condition in terms of  $r_s$  then gives us

$$\limsup_{r_s \rightarrow \infty} (r_s^3)^{q-1} f(r_s) \log\left(1 + \frac{1}{g(r_s)}\right) < 0.$$

Realizing that  $f(r_s) \sim -r_s$  for  $r_s \rightarrow \infty$  and that  $\frac{x}{1+x} \leq \log(1+x) \leq x$  implies for the log-term  $\Theta\left(\frac{1}{1+g(r_s)}\right) = \log\left(1 + \frac{1}{g(r_s)}\right)$ , transforms this into

$$\limsup_{r_s \rightarrow \infty} (r_s^3)^{q-1} (-r_s + 1) \Theta\left(\frac{1}{r_s^2}\right) < 0,$$

which is true if  $q > \frac{4}{3}$ .

*Perdew-Zunger (PZ81)*

Here we again consider spin-unpolarized systems, i.e.  $\zeta = 0$ , as in the case of PW92. The precise value of the PZ81 constants is important for continuity and continuous differentiability; they were chosen in such a way that  $e_c$  is continuously differentiable.

The PZ81 correlation functional is given by

$$\varepsilon_c(r_s) := \begin{cases} \frac{\gamma}{1+\beta_1\sqrt{r_s}+\beta_2r_s} & \text{for } r_s > 1, \\ A \log(r_s) + B + Cr_s \log(r_s) + Dr_s & \text{for } r_s \leq 1. \end{cases} \tag{A.3}$$

Here we used the notation of the original paper and hence  $\gamma, B, D$  are negative. For our analysis we now need to consider  $e_c = \rho\varepsilon(r_s(\rho))$ . Hence, we calculate  $v_c$  to be

$$\begin{aligned} v_c(\rho) &= \begin{cases} \frac{\gamma}{1+\beta_1\sqrt{r_s}+\beta_2r_s} + \rho \left(\frac{d}{d\rho} r_s\right) \frac{d}{dr_s} \left(\frac{\gamma}{1+\beta_1\sqrt{r_s}+\beta_2r_s}\right) & \text{for } r_s > 1, \\ \left(A \log(r_s) + B + Cr_s \log(r_s) + Dr_s\right) + \rho \left(\frac{d}{d\rho} r_s\right) \frac{d}{dr_s} \left(A \log(r_s) + b + Cr_s \log(r_s) + Dr_s\right) & \text{for } r_s \leq 1. \end{cases} \\ &= \begin{cases} \frac{\gamma}{1+\beta_1\sqrt{r_s}+\beta_2r_s} + \frac{\gamma}{3} \frac{\frac{\beta_1}{2}\sqrt{r_s}+\beta_2r_s}{\left(1+\beta_1\sqrt{r_s}+\beta_2r_s\right)^2} & \text{for } r_s > 1, \\ A \log(r_s) + B + Cr_s \log(r_s) + Dr_s + -\frac{1}{3} \left(A + Cr_s + Cr_s \log(r_s) + Dr_s\right) & \text{for } r_s \leq 1. \end{cases} \end{aligned}$$

The continuity for  $\rho \rightarrow 0$  is now checked easily:

$$\lim_{\rho \rightarrow 0} e_c(\rho) = \lim_{r_s \rightarrow \infty} \frac{3}{4\pi} r_s^{-3} \frac{\gamma}{1 + \beta_1 \sqrt{r_s} + \beta_2 r_s} = 0$$

and

$$\lim_{\rho \rightarrow 0} v_c(\rho) = \lim_{r_s \rightarrow \infty} \frac{\gamma}{1 + \beta_1 \sqrt{r_s} + \beta_2 r_s} + \frac{\frac{\gamma}{3} \frac{\beta_1}{2} \sqrt{r_s} + \beta_2 r_s}{(1 + \beta_1 \sqrt{r_s} + \beta_2 r_s)^2} = 0.$$

The continuity of  $e_c$  and  $v_c$  at the value  $r_s = 1$  follows from the choice of constants in the original paper [20], since

$$\lim_{r_s \rightarrow 1^+} \varepsilon_c(\rho(r_s)) = \frac{\gamma}{1 + \beta_1 + \beta_2} = B + D = \lim_{r_s \rightarrow 1^-} \varepsilon_c(\rho(r_s))$$

and

$$\lim_{r_s \rightarrow 1^+} v_c(\rho(r_s)) = \frac{\gamma}{1 + \beta_1 + \beta_2} + \frac{\gamma}{3} \frac{\frac{\beta_1}{2} + \beta_2}{(1 + \beta_1 + \beta_2)^2} = B + D - \frac{1}{3}(D + A + C) = \lim_{r_s \rightarrow 1^-} v_c(\rho(r_s)).$$

Since  $v_c$  is continuous, we only need to check the bound (3.1) for  $\rho \rightarrow \infty$ , i.e.  $r_s \rightarrow 0$ . But here the only term which is not bounded is  $A \log(r_s)$  for which we again use a standard log-bound,  $0 \geq \log(x) \geq 1 - \frac{1}{x}$  for  $0 < x \leq 1$ , so  $|v_c(r_s)| \leq C(1 + \frac{1}{r_s})$  for  $r_s$  small enough. In terms of  $\rho$  this means  $|v_c(\rho)| = C(1 + \rho^{1/3})$ , so (3.1) holds with  $p = \frac{4}{3}$ .

Also, (3.2) holds with  $q = \frac{4}{3}$ , since then plugging in the relation (A.1) between  $r_s$  and  $\rho$  (A.1) yields

$$\limsup_{\rho \rightarrow 0} \frac{e_c(\rho)}{\rho^q} = \limsup_{r_s \rightarrow \infty} \frac{\gamma r_s (\frac{4\pi}{3})^{1/3}}{1 + \beta_1 \sqrt{r_s} + \beta_2 r_s} = (\frac{4\pi}{3})^{1/3} \frac{\gamma}{\beta_2} < 0.$$

## References

- [1] A. Anantharaman, E. Cancès, Existence of minimizers for Kohn–Sham models in quantum chemistry, *Ann. Inst. H. Poincaré Anal. Non Linéaire* 26 (6) (2009) 2425–2455.
- [2] E.J. Baerends, O.V. Gritsenko, R. Van Meer, The Kohn–Sham gap, the fundamental gap and the optical gap: The physical meaning of occupied and virtual Kohn–Sham orbital energies, *Phys. Chem. Chem. Phys.* 15 (39) (2013) 16408–16425.
- [3] A.D. Becke, Perspective: Fifty years of density-functional theory in chemical physics, *J. Chem. Phys.* 140 (18) (2014).
- [4] I. Catto, C. Le Bris, P.-L. Lions, On the thermodynamic limit for Hartree–Fock type models, *Ann. Inst. H. Poincaré Anal. Non Linéaire* 18 (6) (2001) 687–760.
- [5] C.J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, Wiley, 2002.
- [6] H.-L. Dai, W. Ho, *Laser Spectroscopy and Photochemistry on Metal Surfaces*, World Scientific Publishing Company, 1995.
- [7] P.A.M. Dirac, Note on exchange phenomena in the Thomas atom, *Math. Proc. Camb. Phil. Soc.* 26 (3) (1930) 376–385.
- [8] G. Friesecke, The multiconfiguration equations for atoms and molecules: Charge quantization and existence of solutions, *Arch. Ration. Mech. Anal.* 169 (1) (2003) 35–71.
- [9] G. Friesecke, M. Kniely, New optimal control problems in density functional theory motivated by photovoltaics, 2018.
- [10] V. Glaser, A. Martin, H. Grosse, W. Thirring, A family of optimal conditions for the absence of bound states in a potential, *Les Rencontres Phys.-Math. Strasbourg-RCP25* 23 (1976) 0–21.
- [11] W. Ho, Reactions at metal surfaces induced by femtosecond lasers, Tunneling electrons, and heating, *J. Phys. Chem.* 100 (31) (1996) 13050–13060.
- [12] P. Hohenberg, W. Kohn, Inhomogeneous electron gas, *Phys. Rev. (2)* 136 (1964) B864–B871.
- [13] W. Kohn, L.J. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev. (2)* 140 (1965) A1133–A1138.
- [14] P.-L. Lions, The concentration-compactness principle in the calculus of variations. The locally compact case. I, *Ann. Inst. H. Poincaré Anal. Non Linéaire* 1 (2) (1984) 109–145.
- [15] P.-L. Lions, The concentration-compactness principle in the calculus of variations. The locally compact case. II, *Ann. Inst. H. Poincaré Anal. Non Linéaire* 1 (4) (1984) 223–283.
- [16] P.-L. Lions, Solutions of Hartree–Fock equations for Coulomb systems, *Comm. Math. Phys.* 109 (1) (1987) 33–97.
- [17] J. Lu, F. Otto, Nonexistence of minimizer for Thomas–Fermi–Dirac–von Weizsäcker model, *Comm. Pure Appl. Math.* 67 (10) (2014) 1605–1617.
- [18] R.G. Parr, W. Yang, *Density-Functional Theory of Atoms and Molecules*, in: *International Series of Monographs on Chemistry*, Oxford University Press, USA, 1994.

- [19] J.P. Perdew, Y. Wang, Accurate and simple analytic representation of the electron-gas correlation energy, *Phys. Rev. B* 45 (1992) 13244–13249.
- [20] J.P. Perdew, A. Zunger, Self-interaction correction to density-functional approximations for many-electron systems, *Phys. Rev. B* 23 (1981) 5048–5079.
- [21] M. Reed, B. Simon, *Methods of Modern Mathematical Physics, Vol. IV. Analysis of Operators*. Academic, Harcourt Brace Jovanovich Publishers, New York, 1978.
- [22] G. Zhang, C.B. Musgrave, Comparison of DFT methods for molecular orbital eigenvalue calculations, *J. Phys. Chem. A* 111 (8) (2007) 1554–1561.
- [23] G.M. Zhislin, Discussion of the spectrum of Schrödinger operators for systems of many particles, *Trudy Moskovskogo Matematicheskogo Obščestva* 9 (1960) 81–120.

## **A.2 Electronic wavefunction with maximally entangled MPS representation**

# Electronic wavefunction with maximally entangled MPS representation

Benedikt R. Graswald and Gero Friesecke

---

It has long been recognized that matrix product states (MPS) yield accurate representations of quantum chemical wavefunctions. These representations – even though they originated in spin-physics – now lie at the heart of Quantum-Chemistry Density Renormalization Group (QC-DMRG), a state-of-art method for strongly correlated systems. While in theory it is possible to capture the wavefunction  $\Psi$  completely, this exactness requires exponentially large matrices with respect to the system size. Additionally, it is known that the quality of the approximation is governed by the size of the discarded singular values of the corresponding unfoldings  $\psi_{\mu_k+1, \dots, \mu_L}^{\mu_1, \dots, \mu_k}$ . Furthermore unlike in spin chains with identical sites, where the required matrix sizes are connected to the entanglement between subsystems which is in turn governed by area laws, the situation in quantum chemistry is more complicated. Here the role of the sites is taken by the system’s molecular orbitals, and the matrix ranks, the singular values, and the overall approximation quality is strongly influenced by the ordering of this one-body basis. Optimizing this underlying basis set to reduce the size of the involved matrices is an important task in QC-DMRG.

Since the molecular orbitals used in practice are carefully constructed by physicist and chemists, one usually only considers a reordering of this basis instead of arbitrary fermionic mode transformations, i.e. unitary basis changes. Reordering the orbitals corresponds to changing the topology of the tensor network underlying the MPS.

Therefore in this article we consider the problem whether re-orderings can always decrease the bond dimension of a given state. We start by defining a fermionic analogon of the prototypical examples of strong entanglement from spin physics and quantum information theory, which we call fermionic Bell states. To demonstrate how powerful reorderings can be, we show in Section III. that these are in fact only weakly entangled in the MPS sense if re-ordering of the “sites” is allowed. To be precise, they have the feature that the largest matrix rank for  $L$  molecular orbitals occupied by  $N = L/2$  electrons drops from maximal,  $2^{L/2}$ , to just 2 independently of  $L$ , under optimal re-ordering.

In contrast, in Section IV. we provide example states which can be constructed with any basis and for any number of electrons  $N$  and orbitals  $L$  whose bond dimension can not be lower at all by any reordering, which we thus call maximally entangled in MPS sense. This result is a consequence of an old theorem about prime numbers by Besicovitch.

In Section V. we investigate numerically the singular value distribution by providing the singular values of the corresponding unfolding  $\psi_{\mu_k+1, \dots, \mu_L}^{\mu_1, \dots, \mu_k}$ . The singular values are seen to decay extremely slowly, and exhibit a remarkable almost-invariance under re-ordering. The numerics were done using the code tensor-train-julia by Mi-Song Dupuy.

*Own contribution.* I was significantly involved in finding the ideas and carrying out the scientific work of all parts of this article. In particular, I discovered these states, found the old result by Besicovitch and handled the numerics. Furthermore, I was in charge of writing the first draft of this article and involved in writing all major parts of the final version.

# Permission to include:

Benedikt R. Graswald and Gero Friesecke.

Electronic wavefunction with maximally entangled MPS representation.

Eur. Phys. J. D 75, 176 (2021).

<https://doi.org/10.1140/epjd/s10053-021-00189-2>



---

Journal Name:	The European Physical Journal D	(the 'Journal')
Manuscript Number:	EPJD-D-21-00173R1	
Proposed Title of Article:	Electronic wavefunction with maximally entangled MPS representation	
Author(s) [Please list all named Authors]:	Benedikt Graswald, Gero Friesecke	(the 'Author')
Corresponding Author Name:	Benedikt Graswald	

## Licence Applicable to the Article:

Creative Commons licence CC BY: This licence allows readers to copy, distribute and transmit the Article as long as it is attributed back to the author. Readers are permitted to alter, transform or build upon the Article, and to use the Article for commercial purposes. Please read the full licence for further details at - <http://creativecommons.org/licenses/by/4.0/>

## 1 Publication

EDP Sciences, Società Italiana di Fisica and Springer-Verlag GmbH Germany, part of Springer Nature (the 'Licensee') will consider publishing this article, including any supplementary information and graphic elements therein (e.g. illustrations, charts, moving images) (the 'Article'), including granting readers rights to use the Article on an open access basis under the terms of the stated Creative Commons licence.

Headings are for convenience only.

## 2 Grant of Rights

Subject to editorial acceptance of the Article, it will be published under the Creative Commons licence shown above.

In consideration of the Licensee evaluating the Article for publication, the Author grants the Licensee the non exclusive, irrevocable and sub-licensable right, unlimited in time and territory, to copy-edit, reproduce, publish, distribute, transmit, make available and store the Article, including abstracts thereof, in all forms of media of expression now known or developed in the future, including pre- and reprints, translations, photographic reproductions and extensions.

Furthermore, to enable additional publishing services, such as promotion of the Article, the Author grants the Licensee the right to use the Article (including the use of any graphic elements on a stand-alone basis) in whole or in part in electronic form, such as for display in databases or data networks (e.g. the Internet), or for print or download to stationary or portable devices. This includes interactive and multimedia use as well as posting the Article in full or in part or its abstract on social media, and the right to alter the Article to the extent necessary for such use. Author grants to Licensee the right to re-license Article metadata without restriction (including but not limited to author name, title, abstract, citation, references, keywords and any additional information as determined by Licensee).

If the Article is rejected by the Licensee and not published, all rights under this agreement shall revert to the Author.

## 3 Copyright

Ownership of copyright in the Article shall vest in the Author. When reproducing the Article or extracts from it, the Author shall acknowledge and reference first publication in the Journal.

## 4 Self Archiving

Author is permitted to self-archive a preprint and the accepted manuscript version of their Article.

The rights and licensing terms applicable to the version of the Article as published by the Licensee are set out in sections 2 and 3 above. The following applies to versions of the Article preceding publication by the Licensee and/or copyediting and typesetting by the Licensee. Author is permitted to self-archive a preprint and an Author's accepted manuscript version of their Article.

a) A preprint is the version of the Article before peer-review has taken place ("Preprint"). Prior to



# Electronic wavefunction with maximally entangled MPS representation

Benedikt R. Graswald<sup>a</sup> and Gero Friesecke<sup>b</sup>

Department of Mathematics, Technical University of Munich, Munich, Germany

Received 7 April 2021 / Accepted 28 May 2021  
© The Author(s) 2021

**Abstract.** We present an example of an electronic wavefunction with maximally entangled MPS representation, in the sense that the bond dimension is maximal and cannot be lowered by any re-ordering of the underlying one-body basis. Our construction works for any number of electrons and orbitals.

## 1 Introduction

It has long been recognized that matrix product states (MPS) yield accurate representations of quantum chemical wavefunctions. Such representations lie at the heart of the QC-DMRG method, a state-of-the-art method for strongly correlated systems [1–5]. However, exactness requires exponentially large matrices with respect to the system size and the quality of the approximation is governed by the size of the discarded singular values of the corresponding unfoldings  $\psi_{\mu_k+1, \dots, \mu_L}^{\mu_1, \dots, \mu_k}$  [5, 6].

Unlike in spin chains with identical sites, where the required matrix sizes are connected to the entanglement between subsystems which is in turn governed by area laws [7–10], the situation in quantum chemistry is more complicated. The role of the sites is then taken by the system's molecular orbitals, and the matrix ranks, the singular values, and the overall approximation quality is strongly influenced by the ordering of the orbitals [3, 11–14]. Reordering the orbitals corresponds to changing the topology of the tensor network underlying the MPS; see Fig. 1. As turns out, standard examples with maximal entanglement such as the fermionic Bell states (see below) have the feature that the largest matrix rank (or bond dimension) for  $L$  molecular orbitals occupied by  $N = L/2$  electrons drops from maximal,  $2^{L/2}$ , to just 2 independently of  $L$ , under optimal re-ordering.

Here we present an explicit, rather more intricately correlated state whose bond dimension stays at the maximal value  $2^{L/2}$ , regardless of any re-ordering.

## 2 MPS representation

Given a suitable orthonormal set  $\{\varphi_1, \dots, \varphi_L\}$  of molecular spin orbitals, typically consisting of occupied and

<sup>a</sup>e-mail: [benedikt.graswald@ma.tum.de](mailto:benedikt.graswald@ma.tum.de) (corresponding author)

<sup>b</sup>e-mail: [gf@ma.tum.de](mailto:gf@ma.tum.de)

unoccupied Hartree–Fock orbitals, recall the exponential-sized full-CI expansion of an electronic wavefunction which reads, in  $N$ -particle, respectively, Fock space,

$$\begin{aligned}\Psi &= \sum_{i_1 < \dots < i_N} \lambda_{i_1, \dots, i_N} |\varphi_{i_1}, \dots, \varphi_{i_N}\rangle \\ &= \sum_{\mu_1, \dots, \mu_L=0}^1 \psi_{\mu_1, \dots, \mu_L} \Phi_{\mu_1, \dots, \mu_L}.\end{aligned}\quad (1)$$

Here the  $|\varphi_{i_1}, \dots, \varphi_{i_N}\rangle$  are Slater determinants and

$$\begin{aligned}\psi_{\mu_1, \dots, \mu_L} &= \begin{cases} \lambda_{i_1, \dots, i_N}, & \text{if } \mu_{i_1} = \dots = \mu_{i_N} = 1, \sum_j \mu_j = N \\ 0, & \text{else,} \end{cases} \\ \Phi_{\mu_1, \dots, \mu_L} &= \begin{cases} |\varphi_{i_1}, \dots, \varphi_{i_N}\rangle, & \text{if } \mu_{i_1} = \dots = \mu_{i_N} = 1 \\ 0, & \text{else.} \end{cases}\end{aligned}\quad (2)$$

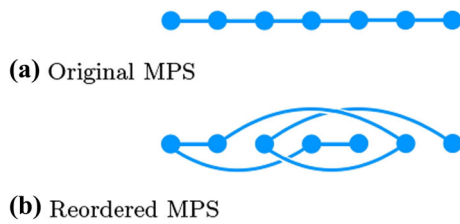
The MPS approximation consists in the ansatz

$$\psi_{\mu_1, \dots, \mu_L} = A_1[\mu_1] \cdots A_L[\mu_L], \quad (3)$$

where the  $A_k[\mu_k]$  are matrices of size  $1 \times M$  (for  $k = 1$ ),  $M \times M$  (for  $k = 2, \dots, L - 1$ ), and  $M \times 1$  (for  $k = L$ ) for some moderate value of  $M$ .

## 3 Fermionic Bell states

Next, we argue that prototype examples of strong entanglement from spin physics and QIT—like Bell states—are in fact only weakly entangled in the MPS sense if re-ordering of the “sites” is allowed. Of course re-ordering only makes sense for molecular orbitals, not sites in 1D spin chains.



**Fig. 1** Schematic picture of a MPS before and after reordering the orbitals (vertices). Bonds represent virtual variables, i.e., summation indices in the matrix product; see Eq. (3)

For  $N$  electrons occupying  $L = 2N$  orbitals  $\{\varphi_1, \dots, \varphi_L\}$ , one can easily write down a fermionic analogon to the standard Bell states.

Set  $\psi_k := (\varphi_k + \varphi_{k+N})/\sqrt{2}$  for  $k = 1, \dots, N$  and consider the Slater determinant  $\Psi := |\psi_1, \dots, \psi_N\rangle$ . It is then not hard to see (e.g., [15]) that its minimal MPS representation in the basis  $(\varphi_k)_k$  has bond dimension  $2^N$ .

Now apply a re-ordering which puts paired-up orbitals next to each other,

$$(\tilde{\varphi}_1, \tilde{\varphi}_2, \dots, \tilde{\varphi}_{L-1}, \tilde{\varphi}_L) = (\varphi_1, \varphi_{N+1}, \dots, \varphi_N, \varphi_L).$$

We claim that in the new basis  $(\tilde{\varphi}_k)_k$ ,  $\Psi$  has an MPS representation with bond dimension just 2. Indeed

$$\Psi = \sum_{\mu_1, \dots, \mu_L=0}^1 A_1[\mu_1] \cdots A_L[\mu_L] \tilde{\Phi}_{\mu_1, \dots, \mu_L} \quad (4)$$

where  $\tilde{\Phi}$  is specified as in (2) and the matrices  $A_k$  are

$$A_1[\mu_1] = (\delta_{\mu_1}^0 \ \delta_{\mu_1}^1), \quad A_L[\mu_L] = (\delta_{\mu_L}^1, \delta_{\mu_L}^0)^T, \\ A_{2\ell}[\mu_{2\ell}] = \begin{pmatrix} \delta_{\mu_{2\ell}}^1 & 0 \\ 0 & \delta_{\mu_{2\ell}}^0 \end{pmatrix}, \quad A_{2\ell+1}[\mu_{2\ell+1}] = \begin{pmatrix} \delta_{\mu_{2\ell+1}}^0 & \delta_{\mu_{2\ell+1}}^1 \\ \delta_{\mu_{2\ell+1}}^0 & \delta_{\mu_{2\ell+1}}^1 \end{pmatrix}.$$

Here  $\ell = 1, \dots, N - 1$  and  $\delta_{\nu}^k$  denotes the Kronecker delta.

### 4 Maximally entangled state

To construct a state whose bond dimension cannot be reduced by any re-ordering, we start off by recalling an old result by Besicovitch [16]; let  $p_1, \dots, p_s$ , be different primes. Then

**Theorem 1** (Corollary 1 in [16]) *A polynomial  $P(\sqrt{p_1}, \dots, \sqrt{p_s})$  with rational coefficients and degree w.r.t. each entry less than or equal to 1, not all equal to zero, cannot vanish.*

Now we consider the set  $\mathcal{P} := \{\sqrt{p_j} : p_j \text{ prime}\}$ . Then every matrix  $A$  whose elements belong to  $\mathcal{P}$  and are pairwise different has maximal rank, since—for every

square submatrix  $B$ — $\det(B)$  is exactly a polynomial of the above form.

Now define the state  $\Psi_{\mathcal{P}}$  by

$$\Psi_{\mathcal{P}} = \sum_{i_1 < \dots < i_N} \lambda_{i_1, \dots, i_N} |\varphi_{i_1}, \dots, \varphi_{i_N}\rangle \\ = \sum_{\mu_1, \dots, \mu_L=0}^1 \psi_{\mu_1, \dots, \mu_L} \Phi_{\mu_1, \dots, \mu_L}, \quad (5)$$

where the coefficients  $\lambda_{i_1, \dots, i_N}$  are mutually different elements of  $\mathcal{P}$  and the second equation gives the occupation representation with  $\psi_{\mu_1, \dots, \mu_L}$  corresponding to  $\lambda_{i_1, \dots, i_N}$  as in (2). Then every unfolding  $\psi_{\mu_{k+1}, \dots, \mu_L}^{\mu_1, \dots, \mu_k}$  is a matrix of the above form and thus has maximal rank. In particular [17],  $\Psi$  has maximal bond dimension.

Furthermore if we consider any new ordering, that is, we change our orbitals according to  $(\varphi_1, \dots, \varphi_L) = Q(\tilde{\varphi}_1, \dots, \tilde{\varphi}_L)$  with  $Q \in \mathbb{R}^{L \times L}$  a permutation matrix, then we cannot decrease the rank of any unfolding. Indeed, it is easy to see that we then obtain the following representation:

$$\Psi_{\mathcal{P}} = \sum_{j_1 < \dots < j_N} \tilde{\lambda}_{j_1, \dots, j_N} |\tilde{\varphi}_{j_1}, \dots, \tilde{\varphi}_{j_N}\rangle \\ = \sum_{\mu_1, \dots, \mu_L=0}^1 \tilde{\psi}_{\mu_1, \dots, \mu_L} \tilde{\Phi}_{\mu_1, \dots, \mu_L},$$

with

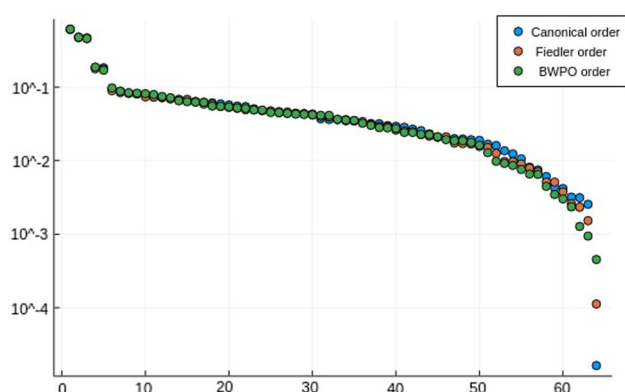
$$\tilde{\lambda}_{j_1, \dots, j_N} = \sum_{i_1 < \dots < i_N} \lambda_{i_1, \dots, i_N} \begin{vmatrix} q_{i_1, j_1} & \cdots & q_{i_1, j_N} \\ \vdots & & \vdots \\ q_{i_N, j_1} & \cdots & q_{i_N, j_N} \end{vmatrix},$$

where  $q_{ij}$  denotes the elements of  $Q$ . Since  $Q$  is a permutation, exactly one determinant will be non-zero. Thus every unfolding still contains the same elements but only their positions change. But by construction of the set  $\mathcal{P}$ , the position within the unfolding  $\psi_{\mu_{k+1}, \dots, \mu_L}^{\mu_1, \dots, \mu_k}$  is irrelevant as long as all entries are different elements of  $\mathcal{P}$ . Hence the unfolding still has full rank. Therefore  $\Psi_{\mathcal{P}}$  still has maximal bond dimension.

We remark that in contrast to orderings, arbitrary fermionic mode transformations, i.e., choosing the transformation  $Q$  above as a unitary, can always somewhat decrease the bond dimension. In the two-particle case ( $N = 2$ ), this can even achieve the optimal bond dimension of 3, for an arbitrary number of orbitals  $L$  [18].

### 5 Singular value distribution

We have also numerically calculated the singular value distribution of our example states for different values of  $N$  and  $L$  and different orderings (such as the widely



**Fig. 2** Singular value distribution of the matricization  $\psi_{\mu_1^1, \dots, \mu_{12}^6}$  of the state  $\Psi_{\mathcal{P}}$  [Eq. (5)] with  $N = 6$  electrons and  $L = 12$  orbitals, for different orderings

used Fiedler order [11]) using the code tensor-train-julia [19].

Figure 2 corresponds to  $N = 6$ ,  $L = 12$ , and a random choice of  $\binom{L}{N}$  primes of size less than  $2^{N+L}$ . The different orderings shown are the original (canonical) order, the Fiedler order [11], and the more recent best weighted prefactor order [15]. In particular all  $2^{L/2}$  singular values are non-zero, as predicted.

The singular values are seen to decay extremely slowly, and exhibit a remarkable almost-invariance under re-ordering. A less extreme but related observation, that the bond dimension cannot be lowered much by re-ordering, was made in an interesting numerical study of strongly correlated states in the 1D Hubbard model [20]. By contrast, for weakly correlated states re-ordering typically reduces the tail by several orders of magnitude [15].

Physically, the slow decay in Fig. 2 means that for the state  $\Psi_{\mathcal{P}}$ , any two subsystems obtained by partitioning the molecular orbitals into two equal-size parts are strongly entangled.

**Acknowledgements** The authors thank M.-S. Dupuy for helpful discussions. Support from the International Research Training Group IGDK Munich—Graz funded by DFG, Project Number 188264188/GRK1754, is gratefully acknowledged.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Data Availability Statement** This manuscript has no associated data or the data will not be deposited. [Authors' comment: In Figure 2 we took a random choice of  $L$  choose  $N$  random primes of size less than  $2^{(N+L)}$ . The shape of the plot and all the implications drawn in the paper do not depend on this particular choice. We always obtained the same overall pattern for any random selection of primes. Even averaging over thousands of different samples did not change the profile of the plot. So, since Fig. 2 just provides a generic example and is in no way special, there is no scientific data to provide.]

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. S.R. White, R.L. Martin, Ab initio quantum chemistry using the density matrix renormalization group. *J. Chem. Phys.* **110**, 4127 (1999)
2. H.-G. Luo, M.-P. Qin, T. Xiang, Optimizing hartree-fock orbitals by the density-matrix renormalization group. *Phys. Rev. B* **81**, 235129 (2010)
3. O. Legeza, J. Röder, B.A. Hess, Controlling the accuracy of the density-matrix renormalization-group method: the dynamical block state selection approach. *Phys. Rev. B* **67**, 125114 (2003a)
4. S. Szalay, M. Pfeffer, V. Murg, G. Barcza, F. Verstraete, R. Schneider, O. Legeza, Tensor product methods and entanglement optimization for ab initio quantum chemistry. *Int. J. Quantum Chem.* **115**, 1342 (2015)
5. U. Schollwöck, The density-matrix renormalization group. *Rev. Mod. Phys.* **77**, 259 (2005)
6. W. Hackbusch, *Tensor Spaces and Numerical Tensor Calculus*, vol. 42 (Springer, Berlin, 2012)
7. M.B. Plenio, J. Eisert, J. Dreißig, M. Cramer, Entropy, entanglement, and area: analytical results for harmonic lattice systems. *Phys. Rev. Lett.* **94**, 060503 (2005)
8. J. Eisert, M. Cramer, M.B. Plenio, Colloquium: area laws for the entanglement entropy. *Rev. Mod. Phys.* **82**, 277 (2010)
9. K. Van Acoleyen, M. Mariën, F. Verstraete, Entanglement rates and area laws. *Phys. Rev. Lett.* **111**, 170501 (2013)
10. M.B. Hastings, An area law for one-dimensional quantum systems. *J. Stat. Mech Theory Exp.* **2007**, P08024 (2007)
11. G. Barcza, O. Legeza, K.H. Marti, M. Reiher, Quantum-information analysis of electronic states of different molecular structures. *Phys. Rev. A* **83**, 012508 (2011)
12. K. Boguslawski, P. Tecmer, O. Legeza, M. Reiher, Entanglement measures for single- and multireference correlation effects. *J. Phys. Chem. Lett.* **3**, 3129 (2012)
13. O. Legeza, J. Sólyom, Optimizing the density-matrix renormalization group method using quantum information entropy. *Phys. Rev. B* **68**, 195116 (2003)
14. O. Legeza, J. Röder, B.A. Hess, QC-DMRG study of the ionic-neutral curve crossing of LiF. *Mol. Phys.* **101**, 2019 (2003b)

15. M.-S. Dupuy, G. Friesecke, Inversion symmetry of singular values and a new orbital ordering method in tensor train approximations for quantum chemistry. *SIAM J. Sci. Comput.* **43**, B108 (2021)
16. A.S. Besicovitch, On the linear independence of fractional powers of integers. *J. Lond. Math. Soc.* **1**, 3 (1940)
17. S. Holtz, T. Rohwedder, R. Schneider, On manifolds of tensors of fixed TT-rank. *Numer. Math.* **120**, 701 (2012)
18. G. Friesecke, B. Grawald, Two-particle correlation in QC-DMRG, in preparation (2021)
19. M.-S. Dupuy, Tensor-train-julia (2021). <https://github.com/msdupuy/Tensor-Train-Julia>. Accessed 3 Apr 2021
20. Ö. Legeza, F. Gebhard, J. Rissler, Entanglement production by independent quantum channels. *Phys. Rev. B* **74**, 195112 (2006)

### A.3 Dissociation limit in Kohn-Sham density functional theory

# Dissociation limit in Kohn-Sham density functional theory

Sören Behr and Benedikt R. Graswald

---

Our main goal in this paper is to analyze the dissociation limit of any symmetric diatomic molecule, i.e. any molecule of the form  $X_2$ , in Kohn-Sham density functional theory (KS-DFT). Simply put we ask the question, what happens to the energy of the system, when the distance between the two atoms is artificially increased further and further until they are torn infinitely far apart? Our main result takes the following form

**Theorem** (Main Theorem – Informal Version). Let  $I_{2N,R}^{X_2}$  and  $I_\lambda^X$  be the energy of the  $X_2$ -molecule with distance  $R$  between the atoms and the  $X$ -atom with  $\lambda$  electrons, respectively. Then we have

$$\lim_{R \rightarrow \infty} I_{2N,R}^{X_2} = \min_{\alpha \in [0, N]} (I_\alpha^X + I_{2N-\alpha}^X). \quad (\text{A.1})$$

This result is proven by applying concentration-compactness theorem to a nonstandard object which is the sequence of minimizers for distances between the two atoms going to infinity. This quite technical proof is carried out in Section 4 and was completely done by myself.

In the long range limit, the ground-state energy of the  $X_2$ -molecule is identical to the energy of two non-interacting atoms - one with electron mass  $\alpha$  and one with electron mass  $2N - \alpha$ . The physical expectation here is, that it is optimal to split the electrons evenly (i.e. the minimum is attained for  $\alpha = N$ ).

The question if or rather for which  $\lambda$  one has symmetric splitting, i.e. given a family of infima  $I_\lambda$  with mass  $\lambda$  if

$$2I_\lambda < I_{\lambda+\varepsilon} + I_{\lambda-\varepsilon} \quad \text{for all } 0 < \varepsilon < \lambda,$$

already plays an important role in Thomas-Fermi and related theories.

To our knowledge the fact that the lowest energy splitting is always given by two neutral atoms is not even proven in full quantum mechanics, rather only in Thomas-Fermi theory and perturbations thereof, where the behaviour of the energy with respect to the particle number is completely understood.

In Section 3 we discuss that if the exchange becomes too strong, we observe symmetry breaking, i.e. the right hand side of (A.1) does not equal  $2I_N^X$ . This is done in two different setups. In Subsection 3.1 we introduce a one-dimensional model which we are able to solve analytically and thus characterize the splitting completely. The full three-dimensional case is discussed in Subsection 3.2, where we prove the symmetric splitting in the positively charged case for sufficiently small exchange relying on an old result by Le Bris in TFDW theory. This is done in Proposition 3 and 4 which we proved by me. Furthermore we analyze the neutrally charged case numerically. These numerical plots were the main part of my co-author Sören Behr and were done using the OCTOPUS package.

*Own contribution.* I was the one developing the main concepts and ideas of this manuscript as well as carrying out the scientific work of all parts of this article with the exception of the plots in Figure 2. Furthermore, I was in charge of writing all parts of the article.

# Permission to include:

Sören Behr and Benedikt R. Graswald (2021).

Dissociation limit in Kohn-Sham density functional theory.

*Nonlinear Analysis* 215, 112633.

<https://doi.org/10.1016/j.na.2021.112633>





Home (<https://www.elsevier.com>) > About (<https://www.elsevier.com/about>)  
> Policies (<https://www.elsevier.com/about/policies>) > Copyright (<https://www.elsevier.com/about/policies/copyright>)

## Copyright

[Overview](#)   [Author rights](#)   [Institution rights](#)   [Government rights](#)   [Find out more](#)

### Overview

In order for Elsevier to publish and disseminate research articles, we need certain publishing rights from authors, which are determined by a publishing agreement between the author and Elsevier.

For articles published open access, the authors license exclusive rights in their article to Elsevier.

For articles published under the subscription model, the authors transfer copyright to Elsevier.

Regardless of whether they choose to publish open access or subscription with Elsevier, authors have many of the same rights under our publishing agreement, which support their need to share, disseminate and maximize the impact of their research.

For open access articles, authors will also have additional rights, depending on the Creative Commons end user license that they select. This Creative Commons license sets out the rights that readers (as well as the authors) have to re-use and share the article: please see here (<https://www.elsevier.com/about/policies/open-access-licenses>) for more information on how articles can be re-used and shared under these licenses.

This page aims to summarise authors' rights when publishing with Elsevier; these are explained in more detail in the [↓ publishing agreement between the author and Elsevier](#).

Irrespective of how an article is published, Elsevier is committed to protect and defend authors' works and their reputation. We take allegations of infringement, plagiarism, ethical disputes, and fraud very seriously.

### Author rights

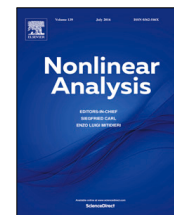
The below table explains the rights that authors have when they publish with Elsevier, for authors who choose to publish either open access or subscription. These apply to the corresponding author and all co-authors.

Author rights in Elsevier's proprietary journals	Published open access	Published subscription
Retain patent and trademark rights	√	√
Retain the rights to use their research data freely without any restriction	√	√
Receive proper attribution and credit for their published work	√	√

Author rights in Elsevier's proprietary journals	Published open access	Published subscription
Re-use their own material in new works without permission or payment (with full acknowledgement of the original article): <ol style="list-style-type: none"> <li>1. Extend an article to book length</li> <li>2. Include an article in a subsequent compilation of their own work</li> <li>3. Re-use portions, excerpts, and their own figures or tables in other works.</li> </ol>	√	√
Use and share their works for scholarly purposes (with full acknowledgement of the original article): <ol style="list-style-type: none"> <li>1. In their own classroom teaching. Electronic and physical distribution of copies is permitted</li> <li>2. If an author is speaking at a conference, they can present the article and distribute copies to the attendees</li> <li>3. Distribute the article, including by email, to their students and to research colleagues who they know for their personal use</li> <li>4. Share and publicize the article via Share Links, which offers 50 days' free access for anyone, without signup or registration</li> <li>5. Include in a thesis or dissertation (provided this is not published commercially)</li> <li>6. Share copies of their article privately as part of an invitation-only work group on commercial sites with which the publisher has a hosting agreement</li> </ol>	√	√
Publicly share the preprint on any website or repository at any time.	√	√
Publicly share the accepted manuscript on non-commercial sites	√	√ using a CC BY-NC-ND license and usually only after an embargo period (see Sharing Policy ( <a href="https://www.elsevier.com/about/policies/sharing">https://www.elsevier.com/about/policies/sharing</a> ) for more information)
Publicly share the final published article	√ in line with the author's choice of end user license	×
Retain copyright	√	×

## Institution rights

Regardless of how the author chooses to publish with Elsevier, their institution has the right to use articles for classroom teaching and internal training. Articles can be used for these purposes throughout the author's institution, not just by the author:



## Dissociation limit in Kohn–Sham density functional theory

Sören Behr, Benedikt R. Graswald\*

Department of Mathematics, Technische Universität München, Germany



## ARTICLE INFO

## Article history:

Received 25 November 2020

Accepted 28 September 2021

Communicated by Enrico Valdinoci

## Keywords:

Density functional theory

Analysis of PDE

Nonlinear analysis

Mathematical physics

## ABSTRACT

We consider the dissociation limit for molecules of the type  $X_2$  in the Kohn–Sham density functional theory setting, where  $X$  can be any element with  $N$  electrons. We prove that when the two atoms in the system are torn infinitely far apart, the energy of the system converges to  $\min_{\alpha \in [0, N]} (I_\alpha^X + I_{2N-\alpha}^X)$ , where  $I_\alpha^X$  denotes the energy of the atom with  $\alpha$  electrons surrounding it. Depending on the “strength” of the exchange this minimum might not be equal to the symmetric splitting  $2I_N^X$ . We show numerically that for the  $H_2$ -molecule with Dirac exchange this gives the expected result of twice the energy of a H-atom  $2I_1^H$ .

© 2021 Elsevier Ltd. All rights reserved.

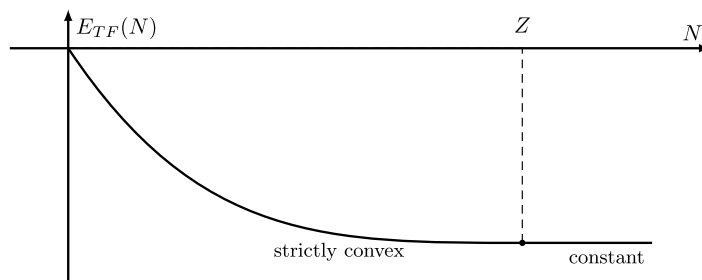
## 1. Introduction

Density functional theory (DFT) was developed by Hohenberg, Kohn and Sham [16,19] in the 1960s and is to this day one of the most widely spread electronic structure models in quantum chemistry, biology and materials science because of its good compromise between accuracy and computational cost. The idea behind DFT is to transform the high-dimensional Schrödinger equation into a low-dimensional and thus computationally manageable problem.

The trade-off in this approach is the introduction of the so-called exchange–correlation functional, which is in theory exact but in practice unknown. Therefore a lot of effort [3,34,36,37,40] has gone into building good approximations to this functional. In this paper we consider the simplest form of these models, the local density approximation (LDA) first proposed in [19], some standard references are [8,33]. Even here the resulting mathematical properties are still far from being well understood. Furthermore as observed in Ref. [32] starting in the early 2000s newer approximations actually become worse in predicting the electron densities. This is due to only focusing on the energies and in the process sacrificing mathematical rigor in favor of the flexibility of fitting to empirical data. Thus in the present article we want to focus on fundamental properties that the exchange–correlation functional should fulfill.

\* Corresponding author.

E-mail addresses: behr@ma.tum.de (S. Behr), graswabe@ma.tum.de (B.R. Graswald).



**Fig. 1.** The Thomas–Fermi energy  $E_{TF}(N)$  with respect to the particle number  $N$ . For positively charged systems  $N < Z$  it is strictly convex, for  $N > Z$  it remains constant.

Our main goal is to analyze the dissociation limit of any symmetric diatomic molecule, i.e. any molecule of the form  $X_2$ , in Kohn–Sham (KS) DFT. Simply put we ask the question, what happens to the energy of the system, when the distance between the two atoms is artificially increased further and further until they are torn infinitely far apart? Our main result takes the following form

**Theorem (Theorem 1 – Informal Version).** Let  $I_{2N,R}^{X_2}$  and  $I_\lambda^X$  be the energy of the  $X_2$ -molecule with distance  $R$  between the atoms and the  $X$ -atom with  $\lambda$  electrons, respectively, defined by (30). Then we have

$$\lim_{R \rightarrow \infty} I_{2N,R}^{X_2} = \min_{\alpha \in [0,N]} (I_\alpha^X + I_{2N-\alpha}^X). \tag{1}$$

In the long range limit, the ground-state energy of the  $X_2$ -molecule is identical to the energy of two non-interacting atoms — one with electron mass  $\alpha$  and one with electron mass  $2N - \alpha$ . The physical expectation here is, that it is optimal to split the electrons evenly (i.e. the minimum is attained for  $\alpha = N$ ).

The question if or rather for which  $\lambda$  one has symmetric splitting, i.e. given a family of infima  $I_\lambda$  with mass  $\lambda$  if

$$2I_\lambda < I_{\lambda+\varepsilon} + I_{\lambda-\varepsilon} \quad \text{for all } 0 < \varepsilon < \lambda,$$

already plays an important role in Thomas–Fermi and related theories, see e.g. [24].

To our knowledge the fact that the lowest energy splitting is always given by two neutral atoms is not even proven in full quantum mechanics, rather only in Thomas–Fermi theory and perturbations thereof, where the behavior of the energy with respect to the particle number is completely understood. A simple sketch of this is presented in Fig. 1.

Note furthermore that in full quantum mechanics and thus for exact HK-DFT charge quantization occurs, i.e.  $\alpha$  in (1) can be restricted to integer values, as proven in [10].

As will be discussed in detail in Section 3 if the exchange becomes too strong, we observe symmetry breaking, i.e. the right hand side of (1) does not equal  $2I_N^X$ .

In the physics literature, this is a well-known challenge: While spin-restricted Kohn–Sham calculations yield qualitatively correct results (i.e. by nature preserve spin-symmetry) they only give reasonable energies close to the actual bond length. Spin unrestricted schemes on the other hand yield better energies but may prefer ionic solutions at long ranges [11,35].

This dilemma has recently attracted mathematical interest. In case of the  $H_2$  molecule at fixed bond-length (see [17]) and for periodic systems (see [14]), symmetry breaking occurs for sufficiently strong exchange contributions. These issues in LDA-DFT and related theories like Thomas–Fermi–Dirac–von Weizsäcker is caused by the Dirac term  $-\int \rho^{4/3}$ , which to some extent makes the functional concave and can thus lead also to nonattainment, see e.g. [30].

The rest of the paper is structured as follows: The next section sets the stage by defining and motivating all the energy functionals needed, giving our main result in Theorem 1 the necessary details. In Section 3 we

put it into context by considering first a one-dimensional DFT model where we can always determine the right hand side of (1). Then we consider the full three dimensional case and fill the gap in our theoretical results by numerical evidence.

The last section contains all the proofs, with the most interesting point being that we apply the concentration–compactness lemma not to a minimizing sequence but to a sequence of minimizers. Fig. 3 summarizes the structure of the proof to help not get lost in technical details.

## 2. Setting the stage

### 2.1. Density functional theory

To put our result into perspective we recall here shortly the basic fundamentals of DFT. A standard reference would be [33]. Readers familiar with the topic might want to skip this section. As a quick reference guide for the notation in use, we created a list of symbols at the end of this paper, which the reader might want to consult from time to time.

The starting point is a system in Born–Oppenheimer approximation [4,15], i.e. a system of  $N$  non-relativistic electrons under influence of an external potential  $V(x)$  and with a repulsive interaction potential  $v_{ee}(x - y)$ .

For a molecule with  $M$  atomic nuclei at positions  $R_1, \dots, R_M \in \mathbb{R}^3$ , with individual charges  $Z_1, \dots, Z_M \in \mathbb{N}$  and total atomic charge  $Z = \sum_{i=1}^M Z_i$ , and with  $N$  electrons the potential  $v(x)$  is just the ensuing Coulomb potential of their positions and charges

$$V(x) := - \sum_{i=1}^M \frac{Z_i}{|x - R_i|}, \quad x \in \mathbb{R}^3.$$

The class of admissible functions  $\mathcal{A}_N$  – the so-called  $N$ -electron wave functions – is given by

$$\mathcal{A}_N := \{ \Psi \in L^2((\mathbb{R}^3 \times \Sigma)^N; \mathbb{C}) : \nabla \Psi \in L^2, \Psi \text{ antisymmetric, } \|\Psi\|_{L^2} = 1 \},$$

where  $\Sigma := \{|\uparrow\rangle, |\downarrow\rangle\}$  denotes the set of spin-states.

In the following we denote the space of position and spin by  $\mathbb{R}_\Sigma^3 = \mathbb{R}^3 \times \Sigma$  and write  $z_i = (x_i, s_i) \in \mathbb{R}_\Sigma^3$  for the pair of position and spin of the  $i$ th particle. Now we can finally define the quantum mechanical energy functional  $\mathcal{E}^{QM}$ ,

$$\mathcal{E}^{QM}[\Psi] := T[\Psi] + V_{ne}[\Psi] + V_{ee}[\Psi], \tag{2}$$

where

$$T[\Psi] := \frac{1}{2} \int_{(\mathbb{R}_\Sigma^3)^N} \sum_{i=1}^N |\nabla_{x_i} \Psi(x_1, s_1, \dots, x_N, s_N)|^2 dz_1 \dots dz_N$$

describes the kinetic energy,

$$V_{ne}[\Psi] := \int_{(\mathbb{R}_\Sigma^3)^N} \sum_{i=1}^N V(x_i) |\Psi(x_1, s_1, \dots, x_N, s_N)|^2 dz_1 \dots dz_N$$

gives the electron–nuclei interaction energy, and

$$V_{ee}[\Psi] := \int_{(\mathbb{R}_\Sigma^3)^N} \sum_{1 \leq i < j \leq N} v_{ee}(x_i - x_j) |\Psi(x_1, s_1, \dots, x_N, s_N)|^2 dz_1 \dots dz_N$$

is the electron–electron interaction energy. Here we used the notation  $\int_{\mathbb{R}_\Sigma^3} f(z) dz = \sum_{s \in \Sigma} \int_{\mathbb{R}^3} f(x, s) dx$ . The exact quantum mechanical ground state energy is now defined as

$$E_0^{QM} := \inf_{\Psi \in \mathcal{A}_N} \mathcal{E}^{QM}[\Psi]. \tag{3}$$

Unfortunately due to the curse of dimensionality there is no hope of ever solving (3) for interesting molecular systems. This is where a central result going back to Hohenberg and Kohn [16] comes into play. We state it in the more modern formulation due to Levy and Lieb [22,25], see also [5]: The quantum mechanical ground state energy  $E_0^{QM}$  (3) only depends on the one-body density  $\rho$  given by

$$\rho(x) = \sum_{s \in \Sigma} \int_{\mathbb{R}^{3(N-1)}} |\Psi(x, s, z_2, \dots, z_N)|^2 dz_2 \dots dz_N. \tag{4}$$

Furthermore it can be recovered exactly by the following minimization

$$E_0^{QM} = \inf_{\rho \in \mathcal{R}_N} \left( F_{LL}[\rho] + \int_{\mathbb{R}^3} v(x)\rho(x) dx \right), \tag{5}$$

where the functional  $F_{LL}$  is given by

$$F_{LL}[\rho] = \min_{\Psi \in \mathcal{A}_N, \Psi \mapsto \rho} (T[\Psi] + V_{ee}[\Psi]). \tag{6}$$

Here the map  $\Psi \mapsto \rho$  describes the relationship in (4), i.e.  $\Psi$  has one-body density  $\rho$  and  $\mathcal{R}_N$  denotes the set of admissible densities  $\rho$  arising via (4) from the set of admissible wavefunctions  $\mathcal{A}_N$ . Note that due to [25]  $\mathcal{R}_N$  has an explicit form

$$\mathcal{R}_N = \left\{ \rho : \mathbb{R}^3 \rightarrow \mathbb{R} \mid \rho \geq 0, \sqrt{\rho} \in H^1(\mathbb{R}^3), \int_{\mathbb{R}^3} \rho(x) dx = N \right\} \tag{7}$$

and provided  $\rho \in \mathcal{R}_N$  the minimum in (6) is attained.

The problem one now faces is that there is no tractable expression of  $F_{LL}$  which could be used in practice. In physics one usually takes the starting point of splitting  $F_{LL}[\rho]$  into three parts

$$F_{LL}[\rho] = T[\rho] + V_{ee}[\rho] + E_{xc}[\rho],$$

where  $T$  describes a kinetic part and  $V_{ee}$  an interaction part and the exchange–correlation  $E_{xc}$  contains all the other terms ensuring that equality holds. There are of course different choice for the individual functionals, but a particularly successful one has been proposed by Kohn and Sham [19]. They came up with the idea to construct the kinetic term  $T$  by considering a non-interacting reference system with the same density  $\rho$  described by single-particle orbitals  $(\varphi_i)_{i=1}^N$  given by

$$T[\rho] = T_{KS}[\rho] = \min \left\{ \frac{1}{2} \sum_{i=1}^N \int_{\mathbb{R}^3_\Sigma} |\nabla \varphi_i|^2(z) dz \mid \varphi \in H^1(\mathbb{R}^3_\Sigma), \int_{\mathbb{R}^3_\Sigma} \overline{\varphi_i(z)} \varphi_j(z) dz = \delta_{ij}, \sum_{i=1}^N \sum_{s \in \Sigma} |\varphi_i(x, s)|^2 = \rho(x) \right\}. \tag{8}$$

Note that these orbitals coming from the fictitious non-interacting system, are only connected to the real system by having the same density, a direct interpretation while sometimes loosely done in practice is not theoretically justified. The interaction term is modeled by an independence ansatz

$$V_{ee}[\rho] = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|} dx dy.$$

Thus the challenge becomes finding an accurate approximation for  $E_{xc}[\rho]$ . There is a huge variety of different exchange–correlation functionals (see e.g [34,36,37]), each with its advantages and disadvantages. In the following we will be working with the so-called local density approximation (LDA), meaning that the exchange–correlation functional is assumed to be of the following form

$$E_{xc}[\rho] = \int_{\mathbb{R}^3} e_{xc}(\rho(x)) dx, \tag{9}$$

where the function  $e_{xc} : \mathbb{R} \rightarrow \mathbb{R}$  has to fulfill certain properties. In our case they will be specified in Section 2.3 under Assumption 1.

The prototypical example for an  $E_{xc}[\rho]$ -approximation stems from considering the homogeneous electron gas. It goes back to Dirac [7] (for a mathematical derivation see [9]) and is given by

$$E_{xc}[\rho] = \int_{\mathbb{R}^3} e_{xc}(\rho(x)) \, dx, \quad e_{xc}(\rho) = -c_{xc}\rho^{4/3}. \tag{10}$$

Employing the above ansatz for the different parts of  $F_{LL}$ , in particular using the orbitals  $\Phi = (\varphi_1, \dots, \varphi_N)$  of the KS-ansatz, Eq. (5) takes the form

$$\begin{aligned} E_0^{QM} \approx E_0^{LDA} = \inf \left\{ \frac{1}{2} \sum_{i=1}^N \int_{\mathbb{R}^3_\Sigma} |\nabla \varphi_i|^2(z) \, dz + \int_{\mathbb{R}^3} V(x)\rho(x) \, dx \right. \\ \left. + \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|} \, dx \, dy + \int_{\mathbb{R}^3} e_{xc}(\rho(x)) \, dx \right\} \\ \varphi_i \in H^1(\mathbb{R}^3_\Sigma), \quad \int_{\mathbb{R}^3_\Sigma} \overline{\varphi_i(z)}\varphi_j(z) \, dz = \delta_{ij}, \quad \rho(x) = \sum_{i=1}^N \sum_{s \in \Sigma} |\varphi_i(x, s)|^2 \end{aligned}$$

### 2.2. Mixed-states

This section shortly recalls the description of the above problem using mixed states, i.e. density matrices. For more details see e.g. [1].

Let  $\mathfrak{S}_1$  denote the vector space of trace class operators on  $L^2(\mathbb{R}^3)$  and introduce the subspace  $\mathcal{H} := \{\gamma \in \mathfrak{S}_1 : |\nabla|\gamma|\nabla| \in \mathfrak{S}_1\}$  endowed with the norm  $\|\cdot\|_{\mathcal{H}} := \text{tr}(|\cdot|) + \text{tr}(|\nabla|\cdot|\nabla|)$  and the convex set

$$K := \{\gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) : 0 \leq \gamma \leq 1, \text{tr}(\gamma) < \infty, \text{tr}(|\nabla|\gamma|\nabla|) < \infty\}, \tag{11}$$

where  $\mathcal{S}(L^2(\mathbb{R}^3))$  denotes the space of bounded self-adjoint operators on  $L^2(\mathbb{R}^3)$ . Next, let us remark that

$$E_0^{QM} = \inf \left\{ \langle \Psi | H_N^V | \Psi \rangle \quad : \quad \Psi \in \mathcal{A}_N \right\} \tag{12}$$

$$= \inf \left\{ \text{tr}(H_N^V \Gamma) \quad : \quad \Gamma \in D_N \right\}, \tag{13}$$

where  $H_N^V$  is the electronic hamiltonian

$$H_N^V := -\frac{1}{2} \sum_{i=1}^N \Delta_{x_i} - \sum_{i=1}^N V(x_i) + \sum_{1 \leq i < j \leq N} \frac{1}{|x_i - x_j|}, \tag{14}$$

and  $D_N$  is the set of  $N$ -body density matrices defined by

$$D_N = \left\{ \Gamma \in \mathcal{S}(\mathcal{H}_N) : 0 \leq \Gamma \leq 1, \text{tr}(\Gamma) = 1, \text{tr}(-\Delta\Gamma) < \infty \right\}. \tag{15}$$

In the above expression,  $\mathcal{S}(\mathcal{H}_N)$  denotes the vector space of bounded self-adjoint operators on the Hilbert space  $\mathcal{H}_N$ , where

$$\mathcal{H}_N = \bigwedge_{i=1}^N L^2(\mathbb{R}^3_\Sigma),$$

endowed with the standard inner product

$$\langle \Psi | \Psi' \rangle_{\mathcal{H}_N} = \int_{(R^3_\Sigma)^N} \overline{\Psi(z_1, \dots, z_N)} \Psi'(z_1, \dots, z_N) \, dz_1 \dots \, dz_N.$$

Furthermore the condition  $0 \leq \Gamma \leq 1$  stands for  $0 \leq \langle \Psi | \Gamma | \Psi \rangle \leq \|\Psi\|_{\mathcal{H}_N}^2$  for all  $\Psi \in \mathcal{H}_N$ .

From a physical point of view, (12) and (13) mean that the ground state energy can be computed either by minimizing over pure states – characterized by wave functions  $\Psi$  – or by minimizing over mixed states – characterized by density operators  $\Gamma$ .

As before we define the electronic density for any  $N$ -electron density operator  $\Gamma \in D_N$

$$\rho_\Gamma(x) := N \sum_{\sigma \in \Sigma} \int_{(\mathbb{R}_\Sigma^3)^{(N-1)}} \Gamma(x, \sigma, z_2, \dots, z_N; x, \sigma, z_2, \dots, z_N) dz_2 \dots dz_N. \tag{16}$$

Note that here and below we use the same notation for an operator and its Green kernel.

Then we get for the electron densities

$$\left\{ \rho : \mathbb{R}^3 \rightarrow \mathbb{R} : \exists \Gamma \in D_N, \rho_\Gamma = \rho \right\} = \mathcal{R}_N = \left\{ \rho : \mathbb{R}^3 \rightarrow \mathbb{R} : \rho \geq 0, \sqrt{\rho} \in H^1(\mathbb{R}^3), \int_{\mathbb{R}^3} \rho dx = N \right\}.$$

Let  $\Gamma \in D_N$  be in the set of  $N$ -body density matrices, then the one-electron reduced density operator  $\Upsilon_\Gamma$  associated with  $\Gamma$  which is the self-adjoint operator on  $L^2(\mathbb{R}_\Sigma^3)$  with kernel

$$\Upsilon_\Gamma(x, s; y, t) = N \int_{(\mathbb{R}_\Sigma^3)^{N-1}} \Gamma(x, s, z_2, \dots, z_N; y, t, z_2, \dots, z_N) dz_2 \dots dz_N.$$

Furthermore it is known, see e.g. [6], that

$$\left\{ \Upsilon : \exists \Gamma \in D_N, \rho_\Gamma = \rho \right\} = \left\{ \Upsilon \in \mathcal{R}D_N : \rho_\Upsilon = \rho \right\}, \tag{17}$$

where

$$\mathcal{R}D_N = \left\{ \Upsilon \in S(L^2(\mathbb{R}_\Sigma^3)) : 0 \leq \Upsilon \leq 1, \text{tr}(\Upsilon) = N, \text{tr}(-\Delta_x \Upsilon) < \infty \right\} \quad \text{and} \tag{18}$$

$$\rho_\Upsilon(x) := \sum_{\sigma \in \Sigma} \Upsilon(x, \sigma; x, \sigma). \tag{19}$$

This leads to the so-called extended Kohn–Sham models

$$I_N^{EKS}[V] := \inf \left\{ \text{tr}(-\frac{1}{2} \Delta_x \Upsilon) + \int_{\mathbb{R}^3} \rho_\Upsilon V dx + J[\rho_\Upsilon] + E_{ex}[\rho_\Upsilon] : \Upsilon \in \mathcal{R}D_N \right\}. \tag{20}$$

Note, up to now no approximation has been made, such that for the exact exchange–correlation functional  $E_0^{QM} = I_N^{EKS}$  for any molecular system containing  $N$  electrons. Unfortunately as mentioned in Section 2.1, there is no tractable expression of  $E_{xc}[\rho]$  that can be used in numerical simulations.

Before proceeding further, and for the sake of simplicity, we will restrict ourselves to closed-shell, spin-unpolarized systems. This means that we will only consider molecular systems with an even number of electrons  $N = 2N_p$ , where  $N_p$  is the number of electron pairs in the system, and we will assume that electrons “go by pairs”.

Hence, the constraints on the one-electron reduced density operator originating from the closed-shell approximation read:

$$\Upsilon(x, |\uparrow\rangle, y, |\uparrow\rangle) = \Upsilon(x, |\downarrow\rangle, y, |\downarrow\rangle) \quad \text{and} \quad \Upsilon(x, |\uparrow\rangle, y, |\downarrow\rangle) = \Upsilon(x, |\downarrow\rangle, y, |\uparrow\rangle) = 0. \tag{21}$$

Introducing  $\gamma(x, y) = \Upsilon(x, |\uparrow\rangle, y, |\uparrow\rangle)$  and denoting  $\rho_\gamma(x) = 2\gamma(x, x)$ , we obtain the spin-unpolarized (or closed-shell or restricted) extended Kohn–Sham model

$$I_N^{REKS}(V) = \inf \left\{ \mathcal{E}(\gamma) : \gamma \in K_{N_p} \right\}, \tag{22}$$



where the energy functional  $\mathcal{E}$  is given by

$$\mathcal{E}(\gamma) = \text{tr}(-\Delta\gamma) + \int_{\mathbb{R}^3} \rho_\gamma V \, dx + J[\rho_\gamma] + E_{xc}[\rho_\gamma], \tag{23}$$

and the admissible set looks like

$$K_{N_p} = \left\{ \gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) : 0 \leq \gamma \leq 1, \text{tr}(\gamma) = N_p, \text{tr}(-\Delta\gamma) < \infty \right\}. \tag{24}$$

Note that the factor  $\frac{1}{2}$  in front of the kinetic part of  $H_N^V$  from (14) vanishes here in front of the trace due to the definition of  $\gamma$  and accounts for the spin.

Furthermore, by spectral theory we have for any  $\gamma \in K_{N_p}$

$$\gamma = \sum_{i \geq 1} \lambda_i |\varphi_i\rangle \langle \varphi_i| \tag{25}$$

with

$$\varphi_i \in H^1(\mathbb{R}^3), \quad \int_{\mathbb{R}^3} \varphi_i \varphi_j \, dx = \delta_{ij}, \quad \lambda_i \in [0, 1], \quad \sum_{i=1}^\infty \lambda_i = N_p, \quad \sum_{i=1}^\infty \lambda_i \|\nabla \varphi_i\|_{L^2}^2 < \infty. \tag{26}$$

### 2.3. Dissociation

In this section we shortly introduce the energy functionals we will be using in this paper. The Kohn–Sham energy functional is given by

$$\mathcal{E}^V[\gamma] := \text{tr}[-\Delta\gamma] + \int_{\mathbb{R}^3} V\rho \, dx + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|} \, dx \, dy + \int_{\mathbb{R}^3} e_{xc}(\rho(x)) \, dx, \tag{27}$$

where  $\rho(x) = 2\gamma(x, x)$  and  $V$  denotes the external potential. Note that the factor 2 is used since we are considering a spin-unpolarized system. Let  $X$  be any atom with  $Z$  number of protons. Then for the  $X_2$  molecule we have

$$V_R^{X_2} = -\frac{Z}{|\cdot|} - \frac{Z}{|\cdot - R|}, \quad \mathcal{E}_R^{X_2}[\gamma] := \mathcal{E}^{V_R^{X_2}}[\gamma], \tag{28}$$

and similar for the  $X$ -atom

$$V^X = -\frac{Z}{|\cdot|}, \quad \mathcal{E}^X[\gamma] := \mathcal{E}^{V^X}[\gamma]. \tag{29}$$

Here and in the following to keep notation a bit simpler we will denote by  $R$  the position of the second nucleus and also its distance to the origin, as long as it is clear from context which one we are referring to.

We then define the ground state energies

$$I_{\lambda,R}^{X_2} := \inf_{\gamma \in K_\lambda} \mathcal{E}_R^{X_2}[\gamma], \quad I_\lambda^X := \inf_{\gamma \in K_\lambda} \mathcal{E}^X[\gamma], \tag{30}$$

where the admissible set is given by

$$K_\lambda := \left\{ \gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) : 0 \leq \gamma \leq 1, \text{tr}(\gamma) = \lambda, \text{tr}(-\Delta\gamma) < \infty \right\}. \tag{31}$$

Furthermore we introduce the problem at infinity, corresponding to a system without nuclei

$$I_\lambda^\infty := \inf_{\gamma \in K_\lambda} \mathcal{E}^\infty[\gamma], \quad \mathcal{E}^\infty[\gamma] := \text{tr}[-\Delta\gamma] + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|} \, dx \, dy + \int_{\mathbb{R}^3} e_{xc}(\rho) \, dx. \tag{32}$$

To shorten notation we will denote by  $T[\gamma] = \text{tr}[-\Delta\gamma]$  the kinetic energy,  $V[\rho] = \int V(x)\rho(x)$  describes the electron–nuclei interaction and the exchange–correlation term is given by  $E_{xc}[\rho] = \int e_{xc}(\rho) \, dx$ . Furthermore the Hartree energy is given by

$$J[\rho] = \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|} \, dx \, dy$$

with its corresponding bilinear form  $D[f, g]$  being

$$D[f, g] = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{f(x)g(y)}{|x - y|} dx dy.$$

Furthermore if a certain statement holds true for all of the three infima, we will sometimes simply write it holds for the map  $\lambda \mapsto I_\lambda$ .

Next let us give the assumption on the exchange–correlation term. Note that we can use the same setting as [1] for the local-density approximation (LDA).

**Assumption 1** (*LDA-exchange–correlation*). Let  $e_{xc} : \mathbb{R}_+ \rightarrow \mathbb{R}$  be a  $C^1$ -function such that

1.  $e_{xc}(0) = 0$ ,
2.  $e'_{xc} \leq 0$ ,
3.  $\exists 0 < \beta_- \leq \beta_+ < \frac{2}{3}$  such that  $|e'_{xc}(\rho)| \leq C(\rho^{\beta_-} + \rho^{\beta_+})$ ,
4.  $\exists 1 \leq \alpha < \frac{3}{2}$  such that  $\limsup_{\rho \rightarrow 0} \frac{e_{xc}(\rho)}{\rho^\alpha} < 0$ .

Note that the prototypical exchange–correlation functional in the LDA-setting (10) coming from the uniform electron gas satisfies these assumptions with  $\alpha = \frac{4}{3}$  and  $\beta_- = \beta_+ = \frac{1}{3}$ .

**Remark.** The existence of minimizers to these functionals for neutral or positively charged systems is due to [1]. We will also be using the following standard results proven there, which we summarize in the Lemmata 1–4.

First some properties of the electron mass to ground state energy map  $\lambda \mapsto I_\lambda$ .

**Lemma 1** (*Properties of the Infimum [1]*). Let  $I_{\lambda,R}^{X_2}, I_\lambda^X$  and  $I_\lambda^\infty$  be as defined above. For the molecular energy assume  $R$  is fixed, but arbitrary. Then the following holds

1. All three maps  $\lambda \mapsto I_{\lambda,R}^{X_2}$ ,  $\lambda \mapsto I_\lambda^X$  and  $\lambda \mapsto I_\lambda^\infty$  are continuous and strictly decreasing for any  $\lambda$  in the domain  $\lambda \in [0, \infty)$ .
2. We always have  $I_{0,R}^{X_2} = I_0^X = I_0^\infty = 0$  and  $-\infty < I_{R,\lambda}^{X_2} < I_\lambda^X < I_\lambda^\infty < 0$  for  $\lambda > 0$ .
3. Furthermore all three maps satisfy the subadditivity condition, i.e. for any of the maps denoted by  $\lambda \mapsto I_\lambda$  we have

$$I_\lambda \leq I_\alpha + I_{\lambda-\alpha}^\infty \quad \forall \alpha \in [0, \lambda] \tag{33}$$

Furthermore the next lemma says that minimizing sequences of our problems cannot vanish in the sense of [27].

**Lemma 2** (*Non-Vanishing [1]*). Let  $\lambda > 0$  and  $(\gamma_n)_n$  a minimizing sequence for any of the problems (30) or (32). Then the sequence  $(\rho_{\gamma_n})_n$  cannot vanish in the sense of [27], which means that

$$\exists M > 0 : \text{ such that } \lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}^3} \int_{B_M(x)} \rho_{\gamma_n}(x) dx > 0.$$

Additionally we remark the classical continuity properties.

**Lemma 3** (*Continuity [1]*). The three functionals  $\mathcal{E}^{X_2}, \mathcal{E}^X, \mathcal{E}^\infty$  are all continuous on the space  $\mathcal{H} = \{\gamma \in \mathfrak{S}_1 : |\nabla|\gamma|\nabla| \in \mathfrak{S}_1\}$ .

The next lemma summarizes the standard bounds on the energy functional. We note that in the following  $C$  will describe a generic constant, which may have different values at each appearance, indicating some finite positive constant independent of the surrounding variables.

**Lemma 4** (Bounds on the Energy Functional [1]). *For all  $\gamma \in K$ , where  $K$  is the convex set defined in (11), we get  $\sqrt{\rho_\gamma} \in H^1(\mathbb{R}^3)$  and the following inequalities:*

(i) Lower bound on the kinetic energy:

$$\frac{1}{2} \|\nabla \sqrt{\rho_\gamma}\|_{L^2}^2 \leq \text{tr}[-\Delta \gamma] \tag{34}$$

(ii) Upper bound on the Coulomb energy:

$$0 \leq J[\rho_\gamma] \leq C \text{tr}[\gamma]^{\frac{3}{2}} \text{tr}[-\Delta \gamma]^{\frac{1}{2}} \tag{35}$$

(iii) Bounds on the interaction energy between nuclei and electrons:

$$-4Z \text{tr}[\gamma]^{\frac{1}{2}} \text{tr}[-\Delta \gamma]^{\frac{1}{2}} \leq \int_{\mathbb{R}^3} \rho_\gamma(x) V(x) dx \leq 0 \tag{36}$$

(iv) Bounds on the exchange–correlation energy:

$$-C \left( \text{tr}[\gamma]^{1-\frac{\beta_-}{2}} \text{tr}[-\Delta \gamma]^{\frac{3\beta_-}{2}} + \text{tr}[\gamma]^{1-\frac{\beta_+}{2}} \text{tr}[-\Delta \gamma]^{\frac{3\beta_+}{2}} \right) \leq E_{xc}[\rho_\gamma] \leq 0 \tag{37}$$

(v) Lower bound on the energy:

$$\mathcal{E}[\gamma] \geq \frac{1}{2} \left( \text{tr}[-\Delta \gamma]^{\frac{1}{2}} - 4Z \text{tr}[\gamma]^{\frac{1}{2}} \right)^2 - 8Z^2 \text{tr}[\gamma] - C \left( \text{tr}[\gamma]^{\frac{2-\beta_-}{2-3\beta_-}} + \text{tr}[\gamma]^{\frac{2-\beta_+}{2-3\beta_+}} \right) \tag{38}$$

(vi) Lower bound on the energy at infinity:

$$\mathcal{E}^\infty[\gamma] \geq \frac{1}{2} \text{tr}[-\Delta \gamma] - C \left( \text{tr}[\gamma]^{\frac{2-\beta_-}{2-3\beta_-}} + \text{tr}[\gamma]^{\frac{2-\beta_+}{2-3\beta_+}} \right). \tag{39}$$

In particular, minimizing sequences of  $I_\lambda$  (30) and  $I_\lambda^\infty$  (32) are bounded in  $\mathcal{H}$ .

Lemma 4 is a central point for the existence of minimizer in the fixed nuclei setting but more importantly for us it bounds the minimizers independently of the position of the nuclei.

Let us now restate the main result of this paper.

**Theorem 1** (Dissociation Limit). *Let  $I_{\lambda,R}^{X_2}$  and  $I_\lambda^X$  be defined by (30), then we have for positively and neutrally charged molecules, i.e. for  $\lambda \leq 2Z$ ,*

$$\lim_{R \rightarrow \infty} I_{\lambda,R}^{X_2} = \min_{\alpha \in [0,\lambda]} (I_\alpha^X + I_{\lambda-\alpha}^X). \tag{40}$$

**Proof.** The proof of Theorem 1 is quite technical and is split into several parts. Since we first want to concentrate on its implications and in order to help with the reading flow of this paper we moved it into its own Section 4.  $\square$

**Theorem 1** says the energy of the  $X_2$ -molecule converges – as the nuclei are pulled infinitely far apart – to the minimum over distributing the amounts of electrons  $\lambda$  on two separated  $X$ -atoms. For linear problems this directly gives  $2I_{\lambda/2}^X$ , i.e. a symmetric splitting, but for nonlinear problems  $\alpha \mapsto I_\alpha^X + I_{\lambda-\alpha}^X$  might take its minimum at another value. Whether the right hand side gives the expected symmetric minimum or not, will be discussed on the basis of the  $H_2$  molecule in the next section.

We want to stress again that we consider the spin-restricted setting also for the two individual atoms. Hence applying it to an  $H$ -Atom with a single electron has to be taken with a grain of salt.

Before we move on, let us make the following remarks regarding the modeling perspective of **Theorem 1**.

**Remark (Possible Generalizations).** We remark that the result of **Theorem 1** can be extended to a large class of finite systems, i.e. arbitrary molecules, provided we stay in a neutral or positively charged setting and the minimum distance between the nuclei of each sub-system goes to zero.

To ease presentation and focus on the main implications, we concentrate on the diatomic case, most prevalent in the related physics literature (see e.g. [17,18,42]). Moreover **Theorem 1** does not guarantee a symmetric splitting even in the case of hydrogen, so a simple extension to larger systems seems to be of limited interest.

**Remark (Spin-Restricted Vs. Spin-Unrestricted).** For the question of existence of minimizers the extension to spin-polarized systems was of independent interest [1,13]. Conversely, we expect that the equivalent of **Theorem 1** would be a quantitative result about distribution of mass of the eigenfunctions of the density matrix, which does not favor a straightforward interpretation in terms of individual orbitals or electrons. This also seems to be the reason why the spin-restricted case is more prevalent in practical computations.

Finally, staying in the spin-restricted setting allows for a comparison with models outside the Kohn–Sham framework as done in Sections 3.1 and 3.2.

So unless new techniques are developed to improve on our result an extension in this direction also seems to be of limited interest.

### 3. Symmetric dissociation or not?

The question which arises now is of course: Does

$$\min_{\alpha \in [0,1]} (I_\alpha^H + I_{2-\alpha}^H) \stackrel{?}{=} 2I_1^H, \tag{41}$$

hold or not, i.e. do we have the right dissociation limit which we expect from physical intuition or which holds also for the Schrödinger equation. The answer is it depends on the “strength” of the exchange–correlation functional. To discuss this further we consider in the following only the Dirac exchange  $e_{xc}(\rho) = -c_{xc}\rho^{4/3}$  – the prototypical example arising from the homogeneous electron gas – with the constant  $c_{xc}$  determining the strength of the exchange term.

To get a better feeling for what determines if the splitting is symmetric or not, i.e. if  $\alpha = 1$  is the minimizer in (41), we consider first a one-dimensional model.

#### 3.1. One-dimensional model

As we will see in the following section, it is quite hard to determine when

$$\min_{\alpha \in [0,1]} (I_\alpha^H + I_{2-\alpha}^H) = 2I_1^H.$$

To understand the problem better we study in this section the one-dimensional problem. Since the Coulomb potential is not well suited for the one dimensional case, we consider  $v(x) = \delta_0(x)$ , i.e. a simple contact potential [31]. The corresponding full Schrödinger system for the  $H_2$ -molecule looks like

$$I_R^{H_2} = \inf_{\substack{\psi \in H^1(\mathbb{R}_\Sigma^2), \\ \|\psi\|_{L^2} = 1, \\ \psi \text{ antisymm.}}} \langle \psi, H(x, y)\psi \rangle, \quad H(x, y) = \sum_{z \in \{x, y\}} -\frac{1}{2} \frac{d^2}{dz^2} - \delta_0(z) - \delta_R(z) + \delta_{|x-y|}(z) \quad (42)$$

and the energy for the  $H$ -atom becomes

$$I^H = \inf_{\substack{\varphi \in H^1(\mathbb{R}), \\ \|\varphi\|_{L^2} = 1}} \langle \varphi, h(x)\varphi \rangle, \quad h(x) = -\frac{1}{2} \frac{d^2}{dx^2} - \delta_0(x). \quad (43)$$

Note that  $\delta_{|x-y|}(x)$  in (42) denotes the delta-distribution, i.e.

$$\int_{\mathbb{R}} \delta_{|x-y|}(x) f(x, z) dx = f(y, z)$$

As for the standard Schrödinger system also here we have the right dissociation limit.

**Proposition 1** (*Dissociation Limit for the Schrödinger Setting*). *For the full Schrödinger setting we always have*

$$\lim_{R \rightarrow \infty} I_R^{H_2} = 2I^H, \quad (44)$$

i.e. the right dissociation limit.

**Proof.** See Section 4.3.  $\square$

Note that Theorem 1 gives exactly the same result in the nonlinear case, but in the linear case every pair  $(\alpha, 2 - \alpha)$  gives the same result, so we always have symmetric dissociation.

Now we consider the DFT version of this system. Note that in this case the Hartree term takes the form

$$J[\rho] = \frac{1}{2} \iint \rho(x)v(x-y)\rho(y) dx dy = \frac{1}{2} \int \rho^2(x) dx.$$

Furthermore the exchange energy per volume looks like  $e_{xc}(\rho) = -c_{xc}\rho^2$ , where the exponent is  $2 = 1 + \frac{1}{d}$  and  $c_{xc} = \frac{1}{4}$  see [20,31].

In total our energy functional for the  $H$ -atom takes the form

$$\begin{aligned} \mathcal{E}^H[\rho] &= \frac{1}{2} \int (\sqrt{\rho}')^2 dx - \int v\rho dx + \frac{1}{2} \iint \rho(x)\rho(y)v(|x-y|) dx dy + E_{xc}[\rho] \\ &= \frac{1}{2} \int (\sqrt{\rho}')^2 dx - \rho(0) + \left(\frac{1}{2} - c_{xc}\right) \int \rho^2 dx. \end{aligned} \quad (45)$$

And analogously for the  $H_2$ -molecule

$$\mathcal{E}^{H_2}[\rho] = \int (\sqrt{\rho}')^2 dx - \rho(0) - \rho(R) + \left(\frac{1}{2} - c_{xc}\right) \int |\rho|^2 dx.$$

In the same way as in Section 4 we can show the dissociation limit

$$\lim_{R \rightarrow \infty} I_R^{H_2} = \min_{\alpha \in [0,1]} \left( I_\alpha^H + I_{2-\alpha}^H \right).$$

Due to replacing the Coulomb potential by a contact potential we simplify the problem because the Hartree and the exchange energy take the same form. Hence the energy functional  $\rho \mapsto \mathcal{E}^H[\rho]$  is clearly convex for  $c_{xc} \leq \frac{1}{2}$ , since the von-Weizsäcker kinetic energy is.

This property is inherited by the infimum. Take any  $\rho_\alpha, \rho_\beta$  non-negative and with  $L^1$ -norm  $\alpha, \beta$ , respectively. Then,

$$I_{\lambda\alpha+(1-\lambda)\beta}^H \leq \mathcal{E}^H[\lambda\rho_\alpha + (1-\lambda)\rho_\beta] \leq \lambda\mathcal{E}^H[\rho_\alpha] + (1-\lambda)\mathcal{E}^H[\rho_\beta],$$

taking the infimum over  $\rho_\alpha, \rho_\beta$  gives the convexity of  $\alpha \mapsto I_\alpha$ .

Therefore we have for  $c_{xc} \leq \frac{1}{2}$

$$2I_1^H = 2I_{\frac{1}{2}\alpha+\frac{1}{2}(2-\alpha)}^H \leq I_\alpha^H + I_{2-\alpha}^H,$$

so symmetric splitting occurs.

For  $c_{xc} > \frac{1}{2}$  there is no symmetric splitting anymore. In order to see this, note that taking the test-functions  $(1 \pm \eta)\rho_1$  with  $\rho_1$  the minimizer to  $I_1^H$  yields

$$I_{1+\eta}^H + I_{1-\eta}^H \leq 2I_1^H + 2\eta^2(\frac{1}{2} - c_{xc}) \int \rho_1^2 dx < 2I_1^H,$$

i.e.  $2I_1^H$  is the strict global maximum. Furthermore in this setting, i.e. the one-dimensional DFT system with contact potential given by the energy functional (45), the ground state density can be found explicitly, see e.g. [41]:

$$\rho = \alpha|\psi|^2, \quad \text{with} \quad \psi(x) = a \cdot \text{sech}(b|x| + x_0), \tag{46}$$

where the parameters  $a, b, x_0$  only depend on  $\alpha$  and  $c_{xc}$  and are given by

$$x_0 = \text{arctanh}\left(\frac{1}{b}\right), \quad a = \sqrt{\frac{b^2}{2(b-1)}}, \quad b = 1 - \alpha \frac{1 - 2c_{xc}}{2}.$$

With this we obtain

$$I_\alpha^H + I_{2-\alpha}^H = \frac{1}{12}(\alpha^2(3 - 12c_{xc}^2) + 6\alpha(4c_{xc}^2 - 1) - 4(1 + 2c_{xc} + 4c_{xc}^2)), \tag{47}$$

where the exact integrals are carried out in the appendix.

Eq. (47) directly implies

$$\min_{\alpha \in [0,1]} \left( I_\alpha^H + I_{2-\alpha}^H \right) = I_2^H.$$

Therefore for  $c_{xc} > \frac{1}{2}$ , we always have both electrons bound at one nucleus.

The fact that the minimum is attained at an integer, is also something we observe numerically in the three-dimensional case. From the view point of physics this make sense since we cannot split an electron in half, but it is non obvious why this drops out of the mathematics.

### 3.2. The three-dimensional case

Now we go back to the physically more interesting case of three dimensions. Since we are just considering the  $H$ -atom the kinetic energy is the same as the von Weizäcker kinetic energy, i.e.

$$\mathcal{E}[\rho] = \frac{1}{2} \int_{\mathbb{R}^3} |\nabla\sqrt{\rho}|^2 dx + \int_{\mathbb{R}^3} V\rho dx + J[\rho] + \int_{\mathbb{R}^3} e_{xc}(\rho) dx \tag{48}$$

and with the energy as before

$$E_\alpha = \inf_{\rho \in \mathcal{A}_\alpha} \mathcal{E}[\rho], \quad \mathcal{A}_\alpha := \{ \rho \in L^1 : \sqrt{\rho} \in H^1(\mathbb{R}^3), \int_{\mathbb{R}^3} \rho dx = \alpha \}.$$

In this section we assume the exchange functional in (48) is given by  $e_{xc}(\rho) = -c_{xc}\rho^{4/3}$  (Dirac-exchange).

**Remark (TFDW).** In the remainder of this section we will be considering the Thomas–Fermi–Dirac–von Weizsäcker (TFDW) energy functional given by (48). As mentioned this coincides with our DFT energy functional in the case of the H-atom, i.e.  $N = 1$ . So this case is the most interesting for our results. Nevertheless we will be considering an arbitrary real number  $N$  of electrons in the following statements whenever possible.

Then for  $c_{xc} \gg 1$  we observe symmetry breaking as in the one-dimensional case.

**Proposition 2 (Neutrally Charged Case).** For  $c_{xc} = 0$  we have the correct splitting, i.e.  $\alpha = 1$  is the unique global minimizer to  $\alpha \mapsto E_\alpha + E_{2N-\alpha}$ . On the other hand there exists a  $c(N) > 0$  such that if  $c_{xc} > c(N)$  we obtain

$$2E_N > (E_\alpha + E_{2N-\alpha}) \quad \forall \alpha \neq N$$

i.e. symmetry breaking occurs.

**Proof.**

We start with the extreme case with  $c_{xc} = 0$ , then the functional  $\rho \mapsto \mathcal{E}[\rho]$  is strictly convex and hence we obtain for any admissible densities  $\rho_\alpha, \rho_{2N-\alpha}$  with mass  $\alpha$  and  $2N - \alpha$ , respectively

$$2E_N \leq 2\mathcal{E}[\frac{1}{2}\rho_\alpha + \frac{1}{2}\rho_{2N-\alpha}] < \mathcal{E}[\rho_\alpha] + \mathcal{E}[\rho_{2N-\alpha}]. \tag{49}$$

Taking now the infimum over  $\rho_\alpha$  and  $\rho_{2N-\alpha}$  gives

$$2E_N = \min_{\alpha \in [0, N]} (E_\alpha + E_{2N-\alpha}).$$

So here the minimum is really attained at the symmetric splitting. Furthermore we also have that  $\alpha = N$  is always the strict global minimizer. Indeed this can be seen by a case distinguishment:

Case 1 : Assume minimizers exist also for slightly negatively charged systems, i.e. there is some  $\varepsilon > 0$  such that minimizer exist for all  $\alpha \in [0, N + \varepsilon]$ . Note we do not assume anything about uniqueness of minimizers just existence. Then we directly get a strong inequality  $2E_N < E_\alpha + E_{2N-\alpha}$  by taking the corresponding minimizers in (49) for  $\alpha \neq N$ . Thus, in this case  $\alpha = N$  would be a strict local minimum and by convexity the unique global minimum.

Case 2 : If we do not have a minimizer for slightly negatively charged systems, then this can only happen if  $\alpha \mapsto E_\alpha$  is not strictly decreasing anymore for  $\alpha > N$ . Otherwise we would have a strict subadditivity inequality because here the problem at infinity  $E_\beta^\infty = 0$  is trivial and the strict subadditivity condition (compare Lemma 1) would give us existence directly [27,28].

Additionally due to convexity and the fact that  $\alpha \mapsto E_\alpha$  is always non-increasing, we must have  $E_\alpha = E_N$  for every  $\alpha \in [N, 2N]$ . But in this case we have

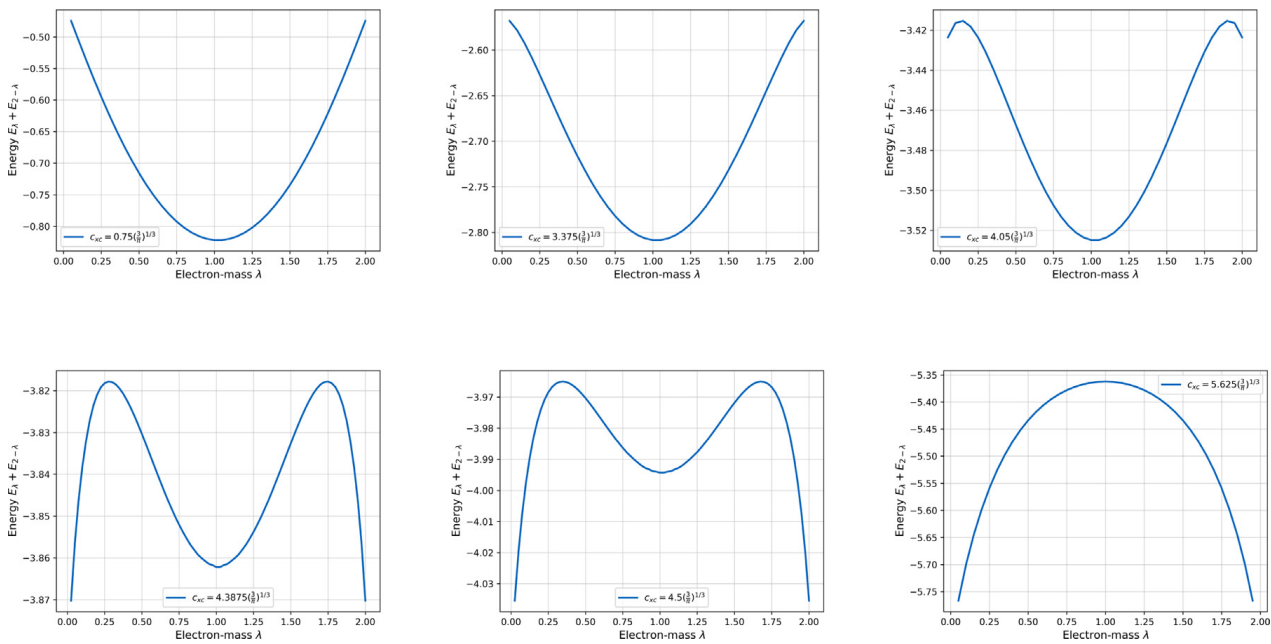
$$E_\alpha + E_{2N-\alpha} = E_\alpha + E_N > 2E_N \quad \forall \alpha \in [0, N).$$

This is what happens in Thomas–Fermi theory, compare Fig. 1. Therefore also in the second case the symmetric splitting is the minimum if we set the exchange constant  $c_{xc} = 0$ .

Now we consider the second statement of our proposition, i.e. we take  $c_{xc}$  to be large:

Let  $\rho_N$  be a minimizer of  $E_N$  (which is known to exist [24]) and  $\eta \in (0, N)$ . Then

$$\begin{aligned} E_{(N+\eta)} + E_{(N-\eta)} &\leq \mathcal{E}[(1 + \frac{\eta}{N})\rho_N] + \mathcal{E}[(1 - \frac{\eta}{N})\rho_N] \\ &= 2(T[\rho_N] + V[\rho_N]) + ((1 + \frac{\eta}{N})^2 + (1 - \frac{\eta}{N})^2)J[\rho_N] + ((1 + \frac{\eta}{N})^{4/3} + (1 - \frac{\eta}{N})^{4/3})E_{xc}[\rho_N] \\ &= 2E_N + \frac{\eta^2}{N^2} \left( 2J[\rho_N] + \frac{4}{9}E_{xc}[\rho_N] \right) + o\left(\frac{\eta^2}{N^2}\right), \end{aligned}$$



**Fig. 2.** The function  $\lambda \mapsto E_\lambda^H + E_{2-\lambda}^H$  for increasing values of  $c_{xc}$ . Note that the plot in the top left corner corresponds to the physically interesting case of  $c_{xc} = \frac{3}{4} \left(\frac{3}{\pi}\right)^{1/3}$ ; here we get numerically a symmetric splitting.

where we used the Taylor-expansion for  $(1 \pm \eta)^{4/3}$ . Now we can use Hardy–Littlewood–Sobolev and then Hölder interpolation to bound  $J[\rho_N]$ .

$$2J[\rho_N] = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_N(x)\rho_N(y)}{|x-y|} dx dy \leq C_{HLS} \|\rho_N\|_{L^{\frac{6}{5}}}^2 \leq C_{HLS} \|\rho_N\|_{L^1}^{2/3} \|\rho_N\|_{L^{4/3}}^{4/3}.$$

So we get using  $\|\rho_N\|_{L^1} = N$

$$2J[\rho_N] + \frac{4}{9}E_{xc}[\rho_N] \leq \left(C_{HLS}N^{2/3} - \frac{4}{9}c_{xc}\right) \int_{\mathbb{R}^3} \rho_1^{4/3} dx < 0,$$

for  $c_{xc} > \frac{9}{4}C_{HLS}N^{2/3}$ . Putting in the numbers, i.e. using the optimal  $C_{HLS}$  given in [23] we see that

$$c_{xc} > \frac{9}{4}\sqrt{\pi} \frac{\Gamma(1)}{\Gamma(\frac{5}{2})} \left(\frac{\Gamma(3)}{\Gamma(\frac{3}{2})}\right)^{2/3} N^{2/3} \approx 5.1615 N^{2/3}$$

suffices. In this case the symmetric splitting of the mass is not the minimum, in fact it is the maximum since the remaining terms in the Taylor expansion all have negative sign.

Note again that for our result of [Theorem 1](#) concerning KS-DFT only the case  $N = 1$ , i.e. the H-atom, is covered by this argument.  $\square$

While [Proposition 2](#) deals with the extreme cases  $c_{xc} = 0$  and  $c_{xc} \gg 1$ , we were not able to prove symmetric splitting for the physically most interest case  $c_{xc} = \frac{3}{4} \left(\frac{3}{\pi}\right)^{1/3}$ . Therefore, we studied the behavior numerically. As in the one-dimensional setting [3.1](#), the minimum seems to be always attained at an integer pair. But the transition from symmetric to asymmetric seems to be more interesting since the function  $\lambda \mapsto I_\lambda^H + I_{2-\lambda}^H$  does not simply switch from convex to concave.

The computations for [Fig. 2](#) were done using the OCTOPUS package [2]. We remark that after rescaling to an  $L^2$  normalized orbital,  $I_\lambda^H$  can be computed from a more standard DFT problem with modified electron–electron interaction potential (fractional charge) and modified exchange constant (cp. Slater X- $\alpha$  exchange [39]).



If we already start with a positively charged molecule in the beginning and thus wonder about the minimum of  $\alpha \mapsto E_{2\lambda-\alpha} + E_\alpha$  for  $\lambda < N$  we can get a stronger result.

**Proposition 3 (Positively Charged Case).** *Let  $\lambda < N$ , then there exists a constant  $c(\lambda) > 0$  such that for all  $c_{xc} < c(\lambda)$  we have*

$$\min_{\alpha \in [0, \lambda]} (E_{2\lambda-\alpha} + E_\alpha) = 2E_\lambda.$$

**Proof.** As in the proof of Proposition 2 we know that for  $c_{xc} = 0$  the symmetric splitting is the strict global minimum. By continuity it thus suffices to show that it stays a local one for all  $c_{xc}$  small enough.

This follows by a result of Le Bris [21]. Indeed in Theorem 4 of [21] he proved that the mapping  $\alpha \mapsto E_\alpha$  is strictly convex for  $\alpha \leq Z$  and  $c_{xc} > 0$  small enough. Note that Le Bris originally proved his convexity result for the Thomas–Fermi–Dirac–von Weizsäcker model. But since he considered an arbitrary non-negative constant in front of the Thomas–Fermi term, this reduces to our model in the hydrogen case, if we set this constant to zero.

This directly implies

$$2E_\lambda = 2E_{\frac{1}{2}(\lambda-\alpha) + \frac{1}{2}(\lambda+\alpha)} < E_{\lambda-\alpha} + E_{\lambda+\alpha}, \quad \forall \alpha \in (0, Z - \lambda).$$

So the symmetric splitting is a local minimum, as mentioned above since it is a strict global minimum for  $c_{xc} = 0$  and thus by continuity it remains a global minimum for  $c_{xc}$  small enough.  $\square$

#### 4. Dissociation limit — the proof

This section contains the proof to Theorem 1, it is split into two parts containing the upper bound and lower bound, respectively.

##### 4.1. Upper bound

We begin by proving the upper bound to Theorem 1, i.e.

$$\limsup_{R \rightarrow \infty} I_{\lambda, R}^{X_2} \leq \min_{\alpha \in [0, \frac{\lambda}{2}]} (I_\alpha^X + I_{\lambda-\alpha}^X). \tag{50}$$

For this purpose, given  $\varepsilon > 0$  take  $\gamma_\alpha \in K_\alpha$  and  $\gamma_{\lambda-\alpha} \in K_{\lambda-\alpha}$ , s.t.

$$\mathcal{E}[\gamma_\alpha] \leq I_\alpha^X + \frac{\varepsilon}{2} \quad \text{and} \quad \mathcal{E}[\gamma_{\lambda-\alpha}] \leq I_{\lambda-\alpha}^X + \frac{\varepsilon}{2}.$$

Thanks to the continuity of the energy functionals established in Lemma 3 and the fact that the finite rank operator and the functions  $C_c^\infty(\mathbb{R}^3)$  are dense in  $\mathcal{H}$  and  $L^2(\mathbb{R}^3)$ , respectively, we may assume that both  $\gamma_\alpha$  and  $\gamma_{\lambda-\alpha}$  have finite rank with range in  $C_c^\infty(\mathbb{R}^3)$ .

Then define the operator  $\gamma_R := \gamma_\alpha + \tau_R \gamma_{\lambda-\alpha} \tau_{-R}$ , where  $\tau_R$  is the unitary operator on  $L^2(\mathbb{R}^3)$  defined by

$$(\tau_R f)(x) := f(x - R).$$

For  $R$  large enough we have  $\gamma_R \in K_\lambda$  and thus

$$\begin{aligned} I_{\lambda, R}^{X_2} &\leq \mathcal{E}_R^{X_2}[\gamma_R] \\ &\leq \mathcal{E}^X[\gamma_\alpha] + \mathcal{E}^X[\gamma_{\lambda-\alpha}] + \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_{\gamma_{\lambda-\alpha}}(x - R) \rho_{\gamma_\alpha}(y)}{|x - y|} \, dx \, dy \\ &\leq I_\alpha^X + I_{\lambda-\alpha}^X + \varepsilon + \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_{\gamma_{\lambda-\alpha}}(x - R) \rho_{\gamma_\alpha}(y)}{|x - y|} \, dx \, dy \xrightarrow{R \rightarrow \infty} I_\alpha^X + I_{\lambda-\alpha}^X + \varepsilon, \end{aligned}$$

where we used that  $\rho_{\gamma_{\lambda-\alpha}}$  and  $\rho_{\gamma_\alpha}$  have compact support. Taking the limsup yields

$$\limsup_{R \rightarrow \infty} I_{\lambda,R}^{X_2} \leq I_\alpha^X + I_{\lambda-\alpha}^X + \varepsilon.$$

Since  $\varepsilon > 0$  and also  $\alpha \in [0, \lambda]$  were arbitrary, we get the desired assertion.

#### 4.2. Lower bound

The lower bound is more difficult, we want to prove

$$\liminf_{R \rightarrow \infty} I_{\lambda,R}^{X_2} \geq \min_{\alpha \in [0, \lambda]} (I_\alpha^X + I_{\lambda-\alpha}^X). \tag{51}$$

Our proof idea is to use the concentration–compactness lemma, which is usually applied to a minimizing sequence, but this time act on a sequence of minimizers  $(\gamma_{R_n})_n$  of  $I_{\lambda,R_n}^{X_2}$  for a sequence  $(R_n)_n$  tending to infinity.

In the following we will denote the arising subsequence also with  $(\gamma_{R_n})_n$  to keep notation clearer. Furthermore in the following  $C > 0$  will denote a generic constant, which may have different values at each appearance, indicating some finite positive constant independent of the surrounding variables.

**Lemma 5** (Lemma I.1., Lions [27]). *Let  $(\rho_n)_{n \geq 1} \subseteq L^1(\mathbb{R}^D)$  be a sequence of non-negative functions such that*

$$\int_{\mathbb{R}^D} \rho_n \, dx = \lambda, \quad \lambda > 0 \text{ fixed.}$$

*Then there exists a subsequence  $(\rho_{n_k})_{k \geq 1}$  satisfying one (and only one) of the following properties:*

(1) *Concentration There is  $(y_{n_k})_{k \geq 1} \subseteq \mathbb{R}^D$  such that  $\rho(\cdot + y_{n_k})$  is tight, i.e.*

$$\forall \varepsilon > 0 \quad \exists R < \infty : \int_{B_R(y_k)} \rho_{n_k} \, dx \geq \lambda - \varepsilon$$

(2) *Vanishing For any  $R < \infty$ , we have*

$$\lim_{k \rightarrow \infty} \sup_{y \in \mathbb{R}^D} \int_{B_R(y)} \rho_{n_k} \, dx = 0,$$

(3) *Dichotomy There is  $\alpha \in (0, \lambda)$  and  $\rho_k^1, \rho_k^2 \in L^1(\mathbb{R}^D)$  non-negative such that*

$$\begin{aligned} \int_{\mathbb{R}^D} |\rho_{n_k} - \rho_k^1 - \rho_k^2| &\longrightarrow 0, \\ \int_{\mathbb{R}^D} \rho_k^1 &\longrightarrow \alpha \text{ and } \int_{\mathbb{R}^D} \rho_k^2 &\longrightarrow \lambda - \alpha \\ \text{dist}(\text{supp}(\rho_k^1), \text{supp}(\rho_k^2)) &\longrightarrow \infty. \end{aligned}$$

For the dichotomy case we will actually use the stronger statement given in [28] see below.

Hence we have to distinguish three cases. Note that the concentration case is the extreme case where the entire mass stays at one nucleus. Dichotomy corresponds to the electron mass being distributed in some way over the two nuclei. Finally vanishing means that the electron mass separates from both nuclei completely, which does not fit into our result.

Therefore let us start with the vanishing case.

Case 1 : Vanishing :

We apply the bounds for the energy functional  $\mathcal{E}^{X_2}$  established in (34), which yields

$$\left\| \nabla \sqrt{\rho_{R_n}^{X_2}} \right\|_{L^2}^2 \leq C + \mathcal{E}^{X_2}[\rho_{R_n}^{X_2}] = C + I_{\lambda, R_n}^{X_2} \leq C, \tag{52}$$

which implies that the sequence  $(\rho_{R_n}^{X_2})_n$  is bounded in  $H^1(\mathbb{R}^3)$ . Hence we can apply the following lemma by Lions.

**Lemma 6** (Lemma I.1., Lions [28]). *Let  $1 \leq p \leq \infty$ ,  $1 \leq q < \infty$  with  $q \neq \frac{Dp}{D-p} =: p^*$  if  $p < D$ . Assume that  $(u_n)_{n \geq 1}$  and  $(\nabla u_n)_n$  are bounded in  $L^q(\mathbb{R}^D)$  and  $L^p(\mathbb{R}^D)$ , respectively. If*

$$\sup_{y \in \mathbb{R}^D} \int_{B_R(y)} |u_n|^q \, dx \xrightarrow{n \rightarrow \infty} 0, \quad \text{for some } R > 0,$$

then  $u_n \rightarrow 0$  in  $L^\alpha(\mathbb{R}^D)$  for  $\alpha$  between  $q$  and  $p^*$  (if  $p \geq D$  set  $p^* = \infty$ ).

With  $p = q = 2$  we obtain that

$$\sqrt{\rho_{R_n}^{X_2}} \xrightarrow{n \rightarrow \infty} 0, \quad \text{in } L^\alpha(\mathbb{R}^3), \quad \alpha \in (2, 6).$$

So clearly we get by applying 1 and 3 from Assumption 1

$$0 \leq -E_{xc}[\rho_{R_n}^{X_2}] \leq C \int_{\mathbb{R}^3} (\rho_{R_n}^{X_2})^{1+\beta_+} + (\rho_{R_n}^{X_2})^{1+\beta_-} \, dx \xrightarrow{n \rightarrow \infty} 0,$$

where  $1 + \beta_\pm \in (1, \frac{5}{3})$  by assumption.

Furthermore we can split the Coulomb potential  $V = v_1 + v_2$  with  $v_1 \in L^q(\mathbb{R}^3)$  and  $v_2 \in L^r(\mathbb{R}^3)$  with  $q < 3$  and  $r > 3$ . Hence by taking e.g.  $q = 2, r = 4$  and applying Hölder inequality with we obtain

$$\left| \int_{\mathbb{R}^3} \frac{1}{|x|} \rho_{R_n}^{X_2} \, dx \right| \leq \|v_1\|_{L^2} \left\| \rho_{R_n}^{X_2} \right\|_{L^2} + \|v_2\|_{L^4} \left\| \rho_{R_n}^{X_2} \right\|_{L^{4/3}} \xrightarrow{n \rightarrow \infty} 0.$$

Analogously for the second nucleus with Coulomb potential  $\frac{1}{|\cdot - R_n|}$ . So combining those two results yields

$$\liminf_{n \rightarrow \infty} \mathcal{E}_{R_n}^{X_2}[\gamma_{R_n}^{X_2}] \geq 0,$$

but this contradicts the upper bound (50) we established

$$0 \leq \liminf_{n \rightarrow \infty} \mathcal{E}_{R_n}^{X_2}[\rho_{R_n}^{X_2}] \leq \limsup_{n \rightarrow \infty} \mathcal{E}_{R_n}^{X_2}[\rho_{R_n}^{X_2}] \leq \min_{\alpha \in [0, \frac{\lambda}{2}]} (I_\alpha^X + I_{\lambda-\alpha}^X) < 0 \quad \text{!}.$$

The last inequality comes from the fact that  $I_\lambda < 0$  for  $\lambda > 0$ , which is the second statement of Lemma 1. Therefore vanishing cannot occur.

Case 2 : Concentration :

Assume concentration occurs, i.e.

$$\forall \varepsilon > 0 \exists (y_n)_n, M < \infty : \int_{B_M(y_n)} \rho_{R_n}^{X_2}(x) \, dx \geq \lambda - \varepsilon \quad \forall n.$$

Intuitively this corresponds to

$$\mathcal{E}_{R_n}^{X_2}[\rho_{R_n}^{X_2}] \xrightarrow{n \rightarrow \infty} I_\lambda^X + I_0^X = I_\lambda^X.$$

We start off with a small lemma.

**Lemma 7.** *The sequence  $(y_n)_n$  stays bounded around 0 or  $(R_n)_n$ , to be more precise (up to a subsequence)*

$$\exists L < \infty \quad \forall n \geq 0: \quad |y_n| \leq L \text{ or } |y_n - R_n| \leq L.$$

**Proof.** Assume  $|y_n| > L$  and  $|y_n - R_n| > L$  for any  $L > 0$  (in particular  $L \gg M$ ). We estimate the Coulomb interaction by applying Cauchy–Schwarz and then Hardy’s inequality to obtain

$$\begin{aligned} \int \frac{\rho_{R_n}^{X_2}}{|x|} dx &= \int \frac{\sqrt{\rho_{R_n}^{X_2}} \sqrt{\rho_{R_n}^{X_2}}}{|x|} dx \leq \left( \int \rho_{R_n}^{X_2} dx \right)^{1/2} \left( \int \frac{\rho_{R_n}^{X_2}}{|x|^2} dx \right)^{1/2} \\ &\leq 2 \left( \int \rho_{R_n}^{X_2} dx \right)^{1/2} \left( \int |\nabla \sqrt{\rho_{R_n}^{X_2}}|^2 dx \right)^{1/2} \leq C \left( \int \rho_{R_n}^{X_2} dx \right)^{1/2}, \end{aligned}$$

where we used that the  $H^1$ -seminorm of  $(\sqrt{\rho_{R_n}^{X_2}})_n$  stays bounded (52). Now if  $|y_n| > L$  we obtain

$$\left( \int_{B_{L-M}(0)} \frac{\rho_{R_n}^{X_2}}{|x|} dx \right)^2 \leq C \int_{B_{L-M}(0)} \rho_{R_n}^{X_2} dx \leq C \left( \lambda - \int_{B_M(y_n)} \rho_{R_n}^{X_2} dx \right) \leq C\varepsilon$$

and the analogous result for the Coulomb interaction with the other nucleus.

Then,

$$\begin{aligned} -\frac{1}{Z} V_{R_n}^{X_2}[\rho_{R_n}^{X_2}] &= \int_{\mathbb{R}^3} \rho_{R_n}^{X_2} \left( \frac{1}{|x|} + \frac{1}{|x - R|} \right) dx \\ &= \int_{B_{L-M}(0)} \frac{\rho_{R_n}^{X_2}}{|x|} dx + \int_{B_{L-M}(R_n)} \frac{\rho_{R_n}^{X_2}}{|x - R_n|} dx + \int_{B_{L-M}^c(0)} \frac{\rho_{R_n}^{X_2}}{|x|} dx + \int_{B_{L-M}^c(R_n)} \frac{\rho_{R_n}^{X_2}}{|x - R_n|} dx \\ &\leq \frac{2\lambda}{L - M} + 2C\varepsilon^{1/2}. \end{aligned}$$

Since this inequality holds for any  $L > M$  we can take  $L \rightarrow \infty$  and then  $\varepsilon \rightarrow 0$ , which gives

$$V_{R_n}^{X_2}[\rho_{R_n}^{X_2}] \xrightarrow{n \rightarrow \infty} 0.$$

But this would imply

$$I_\lambda^X \geq \min_{\alpha \in [0, \frac{\lambda}{2}]} (I_\alpha^X + I_{\lambda-\alpha}^X) \geq \limsup_{n \rightarrow \infty} \mathcal{E}_{R_n}^{X_2}[\rho_{R_n}^{X_2}] \geq \liminf_{n \rightarrow \infty} \mathcal{E}_{R_n}^{X_2}[\rho_{R_n}^{X_2}] = \liminf_{n \rightarrow \infty} \mathcal{E}^\infty[\rho_{R_n}^{X_2}] \geq I_\lambda^\infty, \quad (53)$$

where the first inequality simply comes from the fact that  $I_\lambda^X$  is the same as setting  $\alpha = 0$ , thus it is for sure bigger than the minimum. The second inequality is exactly our upper bound (50) proven in the previous subsection. But Eq. (53) is a contradiction to the strict inequality in Lemma 1 (ii). This finishes the proof.  $\square$

So Lemma 7 gives us either  $|y_n| \leq L$  or  $|y_n - R| \leq L$  for some  $L > 0$ . W.l.o.g. we can in the following assume that  $|y_n| \leq L$  (otherwise transform the coordinate system by a reflection s.t. 0 gets mapped to  $R_n$ . This leaves the energy functional unchanged.)

The last step consists now in a cut-off argument.

By Lemma 4 the sequence  $(\gamma_{R_n}^{X_2})_n$  stays uniformly bounded in  $\mathcal{H}$  and hence we have (up to subsequence)

$$\gamma_{R_n}^{X_2} \xrightarrow{*} \gamma^* \text{ in } \mathcal{H}, \quad \sqrt{\rho_{R_n}^{X_2}} \rightharpoonup \sqrt{\rho^*} \text{ in } H^1(\mathbb{R}^3).$$

Since we are in the concentration case and the  $(y_n)_n$  stays bounded, we can choose for any  $\varepsilon > 0$  a compact set  $Q \subseteq \mathbb{R}^3$  such that

$$\int_Q \rho_{R_n}^{X_2} dx \geq \int_{B_M(y_n)} \rho_{R_n}^{X_2} dx \geq \lambda - \varepsilon.$$

Therefore we get for the limit  $\rho^*$  using convergence in  $L^1_{loc}$

$$\|\rho^*\|_{L^1} \geq \int_Q \rho^* dx = \lim_{n \rightarrow \infty} \int_Q \rho_{R_n}^{X_2} dx \geq \lambda - \varepsilon,$$

since  $\varepsilon > 0$  was arbitrary we get  $\|\rho^*\|_{L^1} = \lambda$ . Therefore  $\sqrt{\rho_{R_n}^{X_2}}$  converges also strongly in  $L^2$  and due to the weak convergence in  $H^1$  also strongly in  $L^p(\mathbb{R}^3)$  for  $p \in [2, 6)$ .

Therefore we get by using the same line of argument as above with Hardy’s inequality

$$\int_{\mathbb{R}^3} \frac{1}{|x - R_n|} \rho_{R_n}^{X_2} dx \xrightarrow{n \rightarrow \infty} 0.$$

Using now the sequential weak lower semi-continuity of the kinetic energy functional  $T$  we obtain

$$\liminf_{n \rightarrow \infty} \mathcal{E}^{X_2}[\gamma_{R_n}^{X_2}] \geq \mathcal{E}^\infty[\gamma^*] - \int_{\mathbb{R}^3} \frac{1}{|x|} \rho^* dx = \mathcal{E}^X[\gamma^*] \geq I_\lambda^X = I_\lambda^X + I_0^X \geq \min_{\alpha \in [0, \frac{\lambda}{2}]} (I_\alpha^X + I_{\lambda-\alpha}^X).$$

This establishes our desired lower bound from (51) in the concentration case.

Case 3 : Dichotomy :

Take a smooth partition of unity  $\xi^2 + \zeta^2 = 1$  such that

$$0 \leq \xi, \zeta \leq 1, \xi(x) = 1, \text{ if } |x| \leq 1, \xi(x) = 0 \text{ if } |x| \geq 2 \text{ and } \zeta(x) = 0, \text{ for } |x| \leq 1, \zeta(x) = 1 \text{ for } |x| \geq 2.$$

Furthermore assume

$$\|\nabla \xi\|_\infty \leq 2 \text{ and } \|\nabla \zeta\|_\infty \leq 2,$$

and consider the dilated functions  $\xi_K(x) = \xi(\frac{x}{K})$  and  $\zeta_K(x) = \zeta(\frac{x}{K})$ . Now if we use the detailed construction of the dichotomy case given in [28] (compare also [1]), we can assume that (up to a subsequence), there exists

- $\alpha \in (0, \lambda)$
- a sequence of points  $(y_n)_n \in \mathbb{R}^3$
- two increasing sequences of positive real numbers  $(K_n^{(1)})_n$  and  $(K_n^{(2)})_n$  such that

$$\lim_{n \rightarrow \infty} K_n^{(1)} = \infty \text{ and } \lim_{n \rightarrow \infty} \frac{K_n^{(2)}}{2} - K_n^{(1)} = \infty \tag{54}$$

such that the sequences  $\gamma_n^{(1)} := \xi_{K_n^{(1)}} \gamma_{R_n}^{X_2} \xi_{K_n^{(1)}}$  and  $\gamma_n^{(2)} := \zeta_{K_n^{(2)}} \gamma_{R_n}^{X_2} \zeta_{K_n^{(2)}}$  satisfy

$$\left\{ \begin{array}{ll} \rho_{\gamma_{R_n}^{X_2}} = \rho_{\gamma_n^{(1)}} \text{ on } B_{K_n^{(1)}}(y_n), \quad \rho_{\gamma_{R_n}^{X_2}} = \rho_{\gamma_n^{(2)}} \text{ on } B_{K_n^{(2)}}^c(y_n), & \text{(a)} \\ \lim_{n \rightarrow \infty} \text{tr } \gamma_n^{(1)} = \alpha, & \text{(b)} \\ \lim_{n \rightarrow \infty} \text{tr } \gamma_n^{(2)} = \lambda - \alpha, & \text{(c)} \\ \rho_{\gamma_n^{(1)}} + \rho_{\gamma_n^{(2)}} - \rho_{\gamma_{R_n}^{X_2}} \xrightarrow{n \rightarrow \infty} 0 \text{ in } L^p \text{ for all } p \in [1, 3), & \text{(d)} \\ \left\| \rho_{\gamma_{R_n}^{X_2}} \right\|_{L^p(B_{K_n^{(2)}}(y_n) \setminus \bar{B}_{K_n^{(1)}}(y_n))} \xrightarrow{n \rightarrow \infty} 0 \text{ in } L^p \text{ for all } p \in [1, 3), & \text{(e)} \\ \lim_{n \rightarrow \infty} \text{dist}(\text{supp}(\rho_{\gamma_n^{(1)}}), \text{supp}(\rho_{\gamma_n^{(2)}})) = \infty, & \text{(f)} \\ \liminf_{n \rightarrow \infty} \text{tr}[-\Delta(\gamma_{R_n}^{X_2} - \gamma_n^{(1)} - \gamma_n^{(2)})] \geq 0. & \text{(g)} \end{array} \right. \tag{55}$$

In terms of the energy functional this splitting gives

$$\begin{aligned} \mathcal{E}^{X_2}[\gamma_{R_n}^{X_2}] &= \mathcal{E}^\infty[\gamma_n^{(1)}] + \mathcal{E}^\infty[\gamma_n^{(2)}] + \int_{\mathbb{R}^3} \rho_{\gamma_n^{(1)}} V^{X_2} + \int_{\mathbb{R}^3} \rho_{\gamma_n^{(2)}} V^{X_2} + \int_{\mathbb{R}^3} \tilde{\rho}_n V^{X_2} \\ &+ \operatorname{tr}[-\Delta(\gamma_{R_n}^{X_2} - \gamma_n^{(1)} - \gamma_n^{(2)})] \\ &+ D[\rho_{\gamma_n^{(1)}}, \rho_{\gamma_n^{(2)}}] + D[\tilde{\rho}_n, \rho_{\gamma_n^{(1)}} + \rho_{\gamma_n^{(2)}}] + J[\tilde{\rho}_n] \\ &+ \int_{\mathbb{R}^3} e_{xc}(\rho_{R_n}^{X_2}) - e_{xc}(\rho_{\gamma_n^{(1)}}) - e_{xc}(\rho_{\gamma_n^{(2)}}) \, dx, \end{aligned}$$

where we have denoted  $\tilde{\rho}_n = \rho_{R_n}^{X_2} - \rho_{\gamma_n^{(1)}} - \rho_{\gamma_n^{(2)}}$ . Since (55d) we know  $\tilde{\rho}_n$  converges to zero in  $L^p(\mathbb{R}^3)$  for all  $p \in [1, 3)$ , so we obtain

$$\int_{\mathbb{R}^3} \tilde{\rho}_n V^{X_2} + D[\tilde{\rho}_n, \rho_{\gamma_n^{(1)}} + \rho_{\gamma_n^{(2)}}] + J[\tilde{\rho}_n] \xrightarrow{n \rightarrow \infty} 0.$$

Indeed, the first term can be handled by again splitting up the Coulomb potential and the second and third term are dealt with using Hardy–Littlewood–Sobolev

$$|D[f, g]| = \left| \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{f(x)g(y)}{|x - y|} \, dx \, dy \right| \leq C \|f\|_{L^{6/5}} \|g\|_{L^{6/5}}.$$

Furthermore for the Coulomb-interaction between  $\rho_{\gamma_n^{(1)}}$  and  $\rho_{\gamma_n^{(2)}}$  we have

$$D[\rho_{\gamma_n^{(1)}}, \rho_{\gamma_n^{(2)}}] \leq \operatorname{dist}(\operatorname{supp}(\rho_{\gamma_n^{(1)}}, \rho_{\gamma_n^{(2)}}))^{-1} \|\rho_{\gamma_n^{(1)}}\|_{L^1} \|\rho_{\gamma_n^{(2)}}\|_{L^1} \xrightarrow{n \rightarrow \infty} 0,$$

where we used (55f). Also the difference in the exchange terms vanishes

$$\begin{aligned} &\left| \int_{\mathbb{R}^3} e_{xc}(\rho_{R_n}^{X_2}) - e_{xc}(\rho_{\gamma_n^{(1)}}) - e_{xc}(\rho_{\gamma_n^{(2)}}) \right| \\ &\leq \int_{B_{K_n^{(2)}}(y_n) \setminus \overline{B_{K_n^{(1)}}}(y_n)} |e_{xc}(\rho_{R_n}^{X_2})| + |e_{xc}(\rho_{\gamma_n^{(1)}}^H)| + |e_{xc}(\rho_{\gamma_n^{(2)}}^H)| \\ &\leq 3C \left( \|\rho_{R_n}^{X_2}\|_{L^{p_-}(B_{K_n^{(2)}}(y_n) \setminus \overline{B_{K_n^{(1)}}}(y_n))}^{p_-} + \|\rho_{R_n}^{X_2}\|_{L^{p_+}(B_{K_n^{(2)}}(y_n) \setminus \overline{B_{K_n^{(1)}}}(y_n))}^{p_+} \right) \xrightarrow{n \rightarrow \infty} 0, \end{aligned}$$

where the exponents  $p_\pm$  are given by  $p_\pm = 1 + \beta_\pm$ . Indeed here we employed Assumption 1 part 3 on  $e'_{xc}(\rho)$  together with part 2  $e_{xc}(0) = 0$  in order to get

$$|e_{xc}(\rho)| \leq C(\rho^{p_+} + \rho^{p_-})$$

and Eq. (55e), i.e. that the density vanishes on the annulus  $B_{K_{2,n}}(y_n) \setminus \overline{B_{K_{1,n}}}(y_n)$ .

Using the lim inf estimate for the kinetic energy from (55g), we obtain

$$\mathcal{E}^{X_2}[\gamma_{R_n}^{X_2}] \geq \mathcal{E}^\infty[\gamma_n^{(1)}] + \mathcal{E}^\infty[\gamma_n^{(2)}] + \int_{\mathbb{R}^3} \rho_{\gamma_n^{(1)}} V^{X_2} + \int_{\mathbb{R}^3} \rho_{\gamma_n^{(2)}} V^{X_2} + \mathcal{R}(n) \tag{56}$$

with a remainder  $\mathcal{R}(n) \xrightarrow{n \rightarrow \infty} 0$ . The last step is to deal with the nuclei part and to go from  $V^{X_2}$  to  $V^X$ ; here we again have to distinguish three cases.

Case 1:  $\rho_{\gamma_n^{(1)}}$  stays close to exactly one nucleus (w.l.o.g. the one at the origin), i.e.

$$\operatorname{dist}(0, B_{K_n^{(1)}}(y_n)) \text{ stays bounded and } \operatorname{dist}(R_n, B_{K_n^{(1)}}(y_n)) \xrightarrow{n \rightarrow \infty} \infty.$$

Note that this necessarily implies that  $\text{dist}(0, \text{supp}(\rho_{\gamma_n^{(2)}})) \rightarrow \infty$  due to triangle inequality. Hence,

$$\int_{\mathbb{R}^3} \rho_{\gamma_n^{(1)}} \frac{1}{|x - R_n|} dx, \int_{\mathbb{R}^3} \rho_{\gamma_n^{(2)}} \frac{1}{|x|} dx \xrightarrow{n \rightarrow \infty} 0$$

and thus taking the limit in (56) and using the continuity of  $\lambda \mapsto I_\lambda^X$  gives

$$\liminf_{n \rightarrow \infty} I_{\lambda, R_n}^{X_2} = \liminf_{n \rightarrow \infty} \mathcal{E}^{X_2}[\gamma_{R_n}^{X_2}] \geq \liminf_{n \rightarrow \infty} \mathcal{E}^X[\gamma_n^{(1)}] + \mathcal{E}^X[\gamma_n^{(2)}] \geq I_\alpha^X + I_{\lambda-\alpha}^X \geq \min_{\alpha \in [0, \frac{\lambda}{2}]} (I_\alpha^X + I_{\lambda-\alpha}^X).$$

Case 2:  $\rho_{\gamma_n^{(1)}}$  does not stay close to any of the two nuclei, i.e.

$$\text{dist}(\{0, R_n\}, \text{supp}(\rho_{\gamma_n^{(1)}})) \xrightarrow{n \rightarrow \infty} \infty.$$

Then by using again our upper bound (50) from Section 4.1 Eq. (56) becomes

$$\begin{aligned} \min_{\alpha \in [0, \frac{\lambda}{2}]} (I_\alpha^X + I_{\lambda-\alpha}^X) &\geq \liminf_{n \rightarrow \infty} \mathcal{E}^{X_2}[\gamma_{R_n}^{X_2}] \geq I_\alpha^\infty + \liminf_{n \rightarrow \infty} \left( \mathcal{E}^\infty[\gamma_n^{(2)}] + \int_{\mathbb{R}^3} \rho_{\gamma_n^{(2)}} V^{X_2} \right) \\ &= I_\alpha^\infty + \liminf_{n \rightarrow \infty} \mathcal{E}^{X_2}[\gamma_n^{(2)}] \\ &\geq I_\alpha^\infty + \underbrace{\liminf_{n \rightarrow \infty} \mathcal{E}^{X_2}[\tilde{\gamma}_n]}_{=: J_{\lambda-\alpha}}, \end{aligned}$$

where  $\tilde{\gamma}_n$  is a minimizer of the problem  $I_{\lambda-\alpha, R_n}^{X_2}$  for each  $n$ .

Case 3: Here  $\rho_{\gamma_n^{(1)}}$  stays close to both of the nuclei and hence  $\rho_{\gamma_n^{(2)}}$  does not stay close to any of the two. Thus we get the same result as in case 2, but with  $\alpha$  and  $\lambda - \alpha$  exchanged.

So we obtain

$$J_\lambda \geq I_\alpha^\infty + J_{\lambda-\alpha} \text{ (case 2) } \quad \text{or} \quad J_\lambda \geq I_{\lambda-\alpha}^\infty + J_\alpha \text{ (case 3)}.$$

Note furthermore that the opposite inequality always holds. As in the proof of the upper bound (Section 4.1) we can for any  $\varepsilon > 0$  take finite rank approximations with range in  $C_c^\infty(\mathbb{R}^3)$  of the minimizers  $\gamma_\alpha^\infty$  to  $I_\alpha^\infty$  and  $\tilde{\gamma}_n$  to  $J_{\lambda-\alpha}$ , respectively, to obtain

$$J_\lambda \leq \mathcal{E}^{X_2}[\tilde{\gamma}_n + \tau_R \gamma_\alpha^\infty \tau_{-R}] \leq J_{\lambda-\alpha} + I_\alpha^\infty + \varepsilon + O\left(\frac{1}{R}\right) \xrightarrow{\varepsilon, \frac{1}{R} \rightarrow 0} J_{\lambda-\alpha} + I_\alpha^\infty,$$

so we arrive at

$$J_\lambda = I_\alpha^\infty + J_{\lambda-\alpha} \text{ (case 2) } \quad \text{or} \quad J_\lambda = I_{\lambda-\alpha}^\infty + J_\alpha \text{ (case 3)}. \tag{57}$$

This furthermore implies that the part splitting off to infinity (i.e.  $\gamma_n^{(1)}$  in case 2 and  $\gamma_n^{(2)}$  in case 3) is almost a minimizing sequence for the problem at infinity ( $I_\alpha^\infty$  in case 2 and  $I_{\lambda-\alpha}^\infty$  in case 3) in the sense that

$$\lim_{n \rightarrow \infty} \mathcal{E}^\infty[\gamma_n^{(1)}] = I_\alpha^\infty \quad \text{and} \quad \lim_{n \rightarrow \infty} \text{tr}[\gamma_n^{(1)}] = \alpha. \tag{58}$$

Now we are in the same position as in the beginning of the proof: We have a sequence  $\tilde{\gamma}_n$  of minimizers to the functional  $\mathcal{E}_{R_n}^{X_2}$  but now with the mass constraint

$$\text{tr}[\tilde{\gamma}_n] = \lambda - \alpha < \lambda \leq N. \tag{59}$$

Going through the entire procedure of the proof again, it either ends after a finite amount of steps (i) or we always end up into the dichotomy case and there case 2 or 3 (ii) compare Fig. 3.

Case (i): After a finite amount of steps we get

$$J_\lambda \geq \sum_{l=1}^T I_{\alpha_l}^\infty + J_{\lambda - \sum_{l=1}^T \alpha_l} \geq \sum_{l=1}^T I_{\alpha_l}^\infty + \min_{\beta} (I_\beta^X + I_{\lambda-\beta - \sum_{l=1}^T \alpha_l}^X). \tag{60}$$

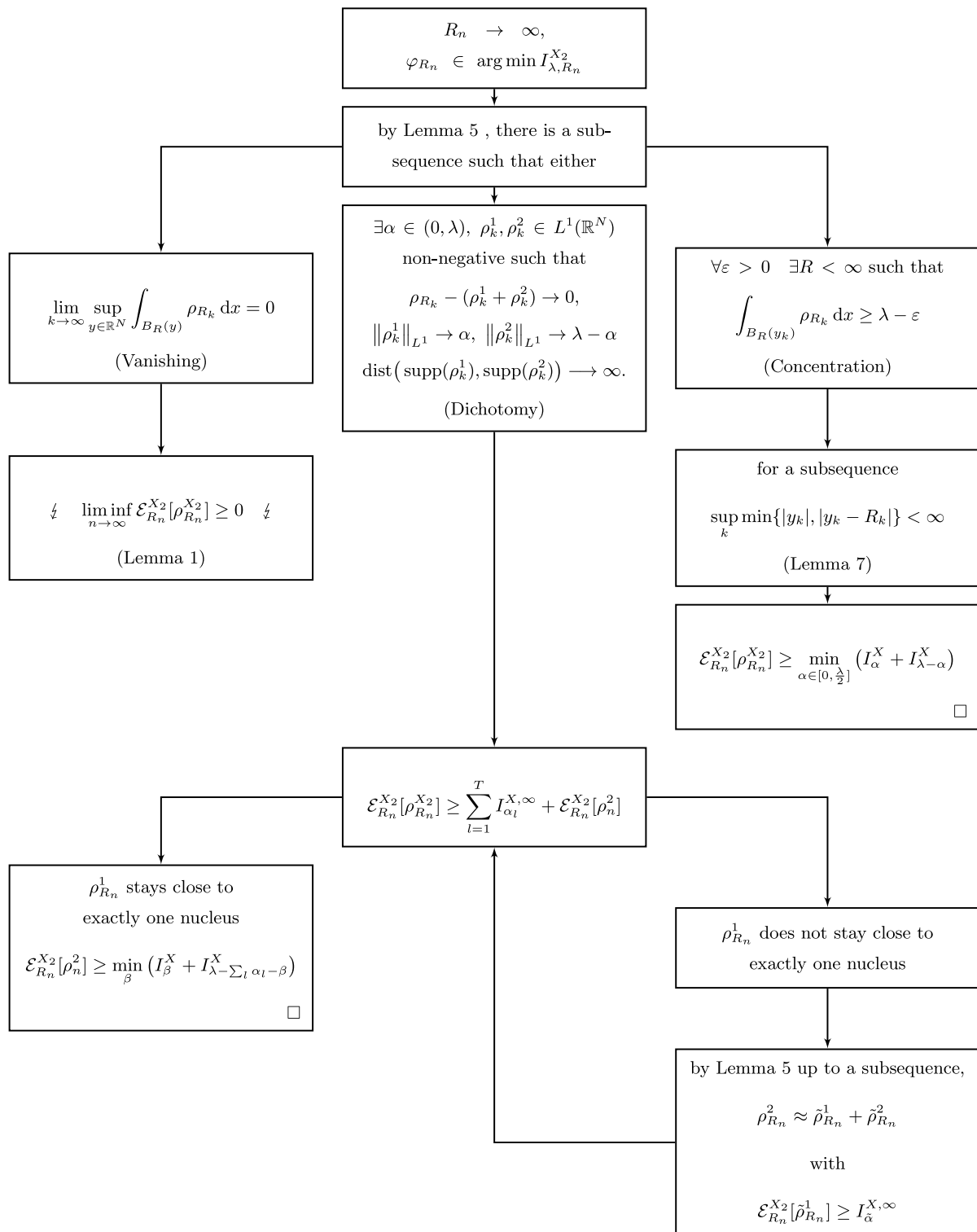


Fig. 3. Structure of the proof for the lower bound. The loop on the bottom-right can only be visited a finite number of times.

Let the minimum at the right hand side be attained at  $\tilde{\beta}$ , then we can just apply the weak subadditivity inequality from Lemma 1 to obtain the desired assertion

$$(60) \geq \sum_{l=1}^T I_{\alpha_l}^\infty + I_{\tilde{\beta}}^X + I_{\lambda-\tilde{\beta}-\sum_{l=1}^T \alpha_l}^X \geq I_{\tilde{\beta}}^X + I_{\lambda-\tilde{\beta}}^X \geq \min_{\alpha \in [0, \frac{\lambda}{2}]} (I_\alpha^X + I_{\lambda-\alpha}^X).$$



Case (ii) : This case is the more intricate one.

After the first splitting we have for the sequence  $\tilde{\gamma}_n$  that  $\|\rho_{\tilde{\gamma}_n}\|_{L^1} = \lambda - \alpha < \lambda$ . In order to show that such a splitting cannot occur infinitely many times we start with considering the Euler–Lagrange equations of the system.

**Lemma 8 (Euler–Lagrange Equations).** *Let  $\gamma_R$  be a minimizer to the energy functional  $\mathcal{E}_R^{X^2}$  to the mass constraint  $\text{tr}[\gamma_R] = \tilde{\lambda} < \lambda$  then it satisfies the Euler–Lagrange equations*

$$\gamma_R = \mathbf{1}_{(-\infty, \varepsilon_F)}(h_{\gamma_R}) + \delta, \quad \text{with } 0 \leq \delta \subset \text{Ker}(h_{\gamma_R} - \varepsilon_F) \tag{61}$$

for some  $\varepsilon_F < 0$  called the Fermi energy, and with the Hamiltonian

$$h_{\gamma_R} = \left(-\frac{1}{2} \Delta + \rho_{\gamma_R} * \frac{1}{|x|} + V_R^{X^2} + e'_{xc}(\rho_{\gamma_R})\right).$$

Furthermore, we have

$$\gamma_R \in \arg \min\{\text{tr}[h_{\gamma_R}\gamma] : \gamma \in K_\lambda\}. \tag{62}$$

**Proof.** This is a standard result, but let us shortly prove it (for a more detailed version see [13]). If  $\gamma_R$  is a minimizer for  $\mathcal{E}_R^{X^2}$  with  $\text{tr}[\gamma_R] = \tilde{\lambda}$  we have for any  $\gamma \in K_{\tilde{\lambda}}$  and for any  $0 \leq t \leq 1$  that the convex combination  $t\gamma + (1-t)\gamma_R$  is admissible, i.e  $t\gamma + (1-t)\gamma_R \in K_{\tilde{\lambda}}$ . Therefore as  $\gamma_R$  is a minimizer on  $K_{\tilde{\lambda}}$  this yields the inequality  $\mathcal{E}_R^{X^2}[t\gamma + (1-t)\gamma_R] \geq \mathcal{E}_R^{X^2}[\gamma_R]$ .

In particular

$$\frac{\mathcal{E}_R^{X^2}[t\gamma + (1-t)\gamma_R] - \mathcal{E}_R^{X^2}[\gamma_R]}{t} \geq 0 \quad \text{for all } 0 < t \leq 1.$$

This implies

$$\lim_{t \rightarrow 0^+} \frac{\mathcal{E}_R^{X^2}[t\gamma + (1-t)\gamma_R] - \mathcal{E}_R^{X^2}[\gamma_R]}{t} = \frac{\partial}{\partial t} \mathcal{E}_R^{X^2}[t\gamma + (1-t)\gamma_R] \Big|_{t=0} \geq 0. \tag{63}$$

A direct calculation leads to

$$\frac{\partial}{\partial t} \mathcal{E}_R^{X^2}[t\gamma + (1-t)\gamma_R] \Big|_{t=0} = \text{tr}[h_{\gamma_R}(\gamma - \gamma_R)]$$

with  $h_{\gamma_R}$  as above. Due to (63) we must have  $\gamma_R \in \arg \min\{\text{tr}[h_{\gamma_R}\gamma] : \gamma \in K_{\tilde{\lambda}}\}$ .

The representation of  $\gamma_R$  then follows if we can show  $\varepsilon_F < 0$ . But as  $\rho_{\gamma_R} * \frac{1}{|x|} + V_R^{X^2} + e'_{xc}(\rho_{\gamma_R})$  is  $\Delta$ -compact, since it is in  $L^\infty_\varepsilon + L^2$ , by Weyl theorem [38] the essential spectrum is that of the Laplacian  $\sigma_{ess}(h_{\gamma_R}) = [0, \infty)$ . Furthermore  $h_{\gamma_R}$  is bounded from below and due to  $e'_{xc}(x) \leq 0$  we have the bound

$$h_{\gamma_R} \leq \left(-\frac{1}{2} \Delta + \rho_{\gamma_R} * \frac{1}{|x|} + V_R^{X^2}\right). \tag{64}$$

For the operator on the right hand side we know by Lemma 19 from [29] that as long as the nuclear charge  $2Z$  is larger than  $\tilde{\lambda}$ , which is satisfied, since  $2Z \geq \lambda > \tilde{\lambda}$ , it has infinitely many negative eigenvalues of finite multiplicity. Therefore the same holds true for  $h_{\gamma_R}$ , which gives us  $\varepsilon_F < 0$ .  $\square$

Note that (62) implies that in fact only finitely many orbitals are occupied, i.e.

$$\tilde{\gamma}_n = \sum_{l=1}^k |\varphi_n^l\rangle \langle \varphi_n^l| + \sum_{l=k}^m \lambda_l |\varphi_n^l\rangle \langle \varphi_n^l|,$$

with  $\lambda_l \in (0, 1)$ . Here the first  $n$  orbitals are fully occupied, while the rest might be fractionally occupied.

Furthermore every occupied orbital  $\varphi_n^l$  is an eigenstate of the corresponding hamiltonian  $h_{\tilde{\gamma}_n}$ , i.e. satisfies

$$\left(-\frac{1}{2} \Delta + \rho_{\tilde{\gamma}_n} * \frac{1}{|x|} + V_{R_n}^{X^2} + e'_{xc}(\rho_{\tilde{\gamma}_n})\right) \varphi_n^l + \theta_n^l \varphi_n^l = 0, \tag{65}$$

where  $-\theta_n^1 < -\theta_n^2 \leq \dots$  denotes the ordered eigenvalues. Our first step consists in proving that for fixed  $l$  the sequence  $(\theta_n^l)_n$  stays bounded away from 0.

**Lemma 9.** Denote by  $(\theta_n^l)_n$  the sequence of smallest eigenvalues in (65), then we have

$$\liminf_{n \rightarrow \infty} \theta_n^l > 0. \tag{66}$$

**Proof.** To see this note

$$h_{\rho_{\tilde{\gamma}_n}} \leq -\frac{1}{2} \Delta + \rho_{\tilde{\gamma}_n} * \frac{1}{|x|} + V_{R_n}^{X_2} = \tilde{h}_n,$$

so it is enough to consider the latter operator  $\tilde{h}_n$ . As in [29] consider a radially symmetric function  $\psi \in C_c^\infty$  with  $\|\psi\|_{L^2} = 1$  and set  $\psi_\sigma = \sigma^{3/2}\psi(\sigma \cdot)$ . Then we get

$$\langle \psi_\sigma, \tilde{h}_n \psi_\sigma \rangle = \sigma^2 \frac{1}{2} \int_{\mathbb{R}^3} |\nabla \psi|^2 dx + \sigma \int_{\mathbb{R}^3} V_\sigma(x) |\psi|^2 dx + \sigma \int_{\mathbb{R}^3} \left( \rho_{\sigma, \tilde{\gamma}_n} * \frac{1}{|x|} \right) |\psi|^2 dx,$$

where  $V_\sigma(x) = -\frac{Z}{|x|} - \frac{Z}{|x - R_n \sigma|}$  and  $\rho_{\sigma, \tilde{\gamma}_n} = \sigma^{-3} \rho_{\tilde{\gamma}_n}(\frac{1}{\sigma} \cdot)$ . Note that the  $\sigma^2$  in front of the kinetic energy comes from the chain rule while the prefactors of the two remaining terms come from substitution. Due to radial symmetry of  $\psi$  we have

$$\begin{aligned} \int_{\mathbb{R}^3} \left( \rho_{\sigma, \tilde{\gamma}_n} * \frac{1}{|x|} \right) |\psi|^2 dx &= \int_{\mathbb{R}^3} \left( |\psi|^2 * \frac{1}{|x|} \right) \rho_{\sigma, \tilde{\gamma}_n}(x) dx \\ &= \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{|\psi|^2(y)}{\max\{|x|, |y|\}} dy \rho_{\sigma, \tilde{\gamma}_n}(x) dx \\ &\leq \underbrace{\|\rho_{\sigma, \tilde{\gamma}_n}\|_{L^1}}_{\lambda - \alpha} \int_{\mathbb{R}^3} \frac{|\psi|^2(y)}{|y|} dy. \end{aligned}$$

In the second equality we have used Newton’s theorem for radial functions  $f(x) = f(|x|)$  with  $f = |\psi|^2$  and that is

$$\int_{\mathbb{R}^3} \frac{f(|y|)}{|x - y|} dy = \int_{\mathbb{R}^3} \frac{f(|y|)}{\max\{|x|, |y|\}} dy.$$

In our case of three dimensions this follows directly by integration in spherical coordinates, for the general case see Theorem 9.7 in [26].

By Rayleigh–Ritz we thus have for every fixed  $n$

$$-\theta_n^1 = \inf \langle \psi_\sigma, h_{\rho_{\tilde{\gamma}_n}} \psi_\sigma \rangle \leq \inf \langle \psi_\sigma, \tilde{h}_n \psi_\sigma \rangle \leq \sigma \underbrace{(\lambda - \alpha - 2Z)}_{< 0} \int_{\mathbb{R}^3} \frac{|\psi|^2(y)}{|y|} dy + \sigma^2 \frac{1}{2} \int_{\mathbb{R}^3} |\nabla \psi|^2 dx.$$

Taking now  $\sigma \rightarrow 0$  the linear term will eventually dominate and since the right hand side is independent of  $n$ , we obtain a lower bound

$$\theta_n^1 > c > 0 \quad \forall n \in \mathbb{N} \text{ and some constant } c > 0 \text{ independent of } n.$$

For  $l > 1$  simply take a family of orthogonal functions  $(\psi_j)_{j=1}^k$  with the same properties as  $\psi$  above, the min–max principle [38] then gives the result.  $\square$

Since also for  $\tilde{\gamma}_n$  the dichotomy case occurs we get  $\tilde{\gamma}_n^{(1)}$  and  $\tilde{\gamma}_n^{(2)}$  with the same properties as listed in (55)–(55g). Define now

$$\omega_n^l := (1 - \xi_{K_n^{(1)}} - \zeta_{K_n^{(2)}}) \varphi_n^l = \varepsilon_n \varphi_n^l \quad \text{and} \quad \varphi_{1,n}^l = \xi_{K_n^{(1)}} \varphi_n^l, \quad \varphi_{2,n}^l = \zeta_{K_n^{(2)}} \varphi_n^l,$$

where  $\varphi_n^l$  are the orbitals corresponding to  $\tilde{\gamma}_n$  and  $\xi, \zeta$  are the smooth partitions of unity given by the dichotomy case with  $K_n^{(j)}$  as in (54).

Note that  $0 \leq \varepsilon_n \leq 1$  and  $\|\nabla \varepsilon_n\|_\infty \rightarrow 0$ . Furthermore we have

$$\rho_{\tilde{\gamma}_n^{(i)}} = \sum_l \lambda_l^{(n)} |\varphi_{i,n}^l|^2,$$

where  $0 < \lambda_l^{(n)} \leq 1$  is the occupation number of the  $l$ th orbital. By multiplying (65) with  $\omega_n^l$ , we obtain

$$\int_{\mathbb{R}^3} \nabla \omega_n^l \cdot \nabla \varphi_n^l \, dx \xrightarrow{n \rightarrow \infty} 0.$$

Since  $\nabla \omega_n^l = \varepsilon_n \nabla \varphi_n^l + \varphi_n^l \nabla \varepsilon_n$  and  $\varepsilon_n^2 \leq \varepsilon_n$  we also get

$$\int_{\mathbb{R}^3} \varepsilon_n^2 |\nabla \varphi_n^l|^2 \, dx \xrightarrow{n \rightarrow \infty} 0,$$

which finally implies  $\nabla \omega_n^l \rightarrow 0$  in  $L^2(\mathbb{R}^3)$ . Combining this with the fact that the supports of  $\varphi_{1,n}^l$  and  $\varphi_{2,n}^l$  go infinitely far apart for  $n \rightarrow \infty$  (65) becomes

$$\left(-\frac{1}{2} \Delta + \rho_{\tilde{\gamma}_n^{(1)}} * \frac{1}{|x|} + V_{R_n}^{X_2} + e'_{xc}(\rho_{\tilde{\gamma}_n^{(1)}})\right) \varphi_{1,n}^l + \theta_n^l \varphi_{1,n}^l \xrightarrow{H^{-1}} 0 \tag{67}$$

$$\left(-\frac{1}{2} \Delta + \rho_{\tilde{\gamma}_n^{(2)}} * \frac{1}{|x|} + V_{R_n}^{X_2} + e'_{xc}(\rho_{\tilde{\gamma}_n^{(2)}})\right) \varphi_{2,n}^l + \theta_n^l \varphi_{2,n}^l \xrightarrow{H^{-1}} 0 \tag{68}$$

Note here that the eigenvalues  $\theta_n^l$  are the ones from  $h_{\tilde{\gamma}_n}$  and that the support of one of the two sequences drifts infinitely far way of both nuclei. W.l.o.g. let it be  $\tilde{\gamma}_n^{(1)}$ , then

$$\text{dist}(\{0, R_n\}, \rho_{\tilde{\gamma}_n^{(1)}}) \xrightarrow{n \rightarrow \infty} \infty$$

and since  $\tilde{\gamma}_n^{(1)}$  is almost a minimizing sequence to  $I_\alpha^\infty$  in the sense of (58), it cannot vanish. Therefore there exist  $\kappa, M > 0$  and a sequence  $(y_n)_n$  of points in  $\mathbb{R}^3$  such that

$$\int_{B_M(y_n)} \rho_{\tilde{\gamma}_n^{(1)}}(x) \, dx \geq \kappa > 0. \tag{69}$$

Furthermore we necessarily have

$$\text{dist}(\{0, R_n\}, (y_n)_n) \xrightarrow{n \rightarrow \infty} \infty$$

and thus (67) becomes for the translated density matrix  $\bar{\gamma}_n^{(1)} := \tau_{y_n} \tilde{\gamma}_n^{(1)} \tau_{-y_n}$  with orbitals  $\bar{\varphi}_{1,n}^l$

$$\left(-\frac{1}{2} \Delta + \rho_{\bar{\gamma}_n^{(1)}} * \frac{1}{|x|} + e'_{xc}(\rho_{\bar{\gamma}_n^{(1)}})\right) \bar{\varphi}_{1,n}^l + \theta_n^l \bar{\varphi}_{1,n}^l \xrightarrow{H^{-1}} 0.$$

Finally note that  $\sqrt{\rho_{\bar{\gamma}_n^{(1)}}} \rightharpoonup \sqrt{\rho} \neq 0$  in  $H^1(\mathbb{R}^3)$  due to (69).

Now if our procedure never stops we can as in [1] construct an infinity of sequences of orbitals  $(\varphi_{k,n}^l)_{k,n \in \mathbb{N}}$  with  $\|\varphi_{k,n}^l\|_{L^2} = 1$  such that for every  $k, n \in \mathbb{N}$

$$\left\{ \begin{array}{l} \psi_{l,k,n} := (\sqrt{\lambda_l^{(n)}} \varphi_{k,n}^l), \sqrt{\rho_{\gamma_n^{(k)}}} \text{ bounded in } H^1(\mathbb{R}^3), \int_{\mathbb{R}^3} \rho_{\gamma_n^{(k)}} = \alpha_k, \rho_{\gamma_n^{(k)}} = \sum_l |\psi_{l,k,n}|^2 \quad \text{(a)} \\ \left(-\frac{1}{2} \Delta + \rho_{\gamma_n^{(k)}} * \frac{1}{|x|} + e'_{xc}(\rho_{\gamma_n^{(k)}})\right) \varphi_{k,n}^l + \theta_n^l \varphi_{k,n}^l = \eta_{n,k} \xrightarrow{n \rightarrow \infty} 0 \quad \text{(b)} \\ \psi_{l,k,n} \text{ converges to } \psi_{l,k} \text{ weakly in } H^1, \text{ strongly in } L_{loc}^p \text{ for } 2 \leq p < 6 \text{ and a.e. on } \mathbb{R}^3, \quad \text{(c)} \\ \sqrt{\rho_{\gamma_n^{(k)}}} \text{ converges to } \sqrt{\rho_k} \neq 0 \text{ weakly in } H^1, \text{ str. in } L_{loc}^p \text{ for } 2 \leq p < 6 \text{ and a.e. on } \mathbb{R}^3, \quad \text{(d)} \end{array} \right. \tag{70}$$

where

$$\sum_{k \in \mathbb{N}} \alpha_k \leq \lambda - \alpha. \tag{71}$$

Note furthermore that

$$\sum_l \|\psi_{l,k,n}\|_{H^1}^2$$

stays bounded independent of  $k$  or  $n$ . For the  $L^2$ -norm this is clear from (70), since  $\alpha_k$  stays bounded (71). For the  $H^1$ -norm, this can be seen by applying (70) to  $\lambda_l^{(n)} \varphi_{k,n}^l$  and summing over  $l$ . Then the term

$$\left( \frac{1}{2} \sum_l \|\nabla \psi_{l,k,n}\|_{L^2}^2 + J[\rho_{\gamma_n^{(k)}}] + \int_{\mathbb{R}^3} e'_{xc}(\rho_{\gamma_n^{(k)}}) \rho_{\gamma_n^{(k)}} \, dx + \sum_l \theta_n^l \|\psi_{l,k,n}\|_{L^2}^2 \right).$$

stays bounded independently of  $k$  and  $n$ . Indeed by (70)  $\sqrt{\rho_{\gamma_n^{(k)}}}$  is bounded in  $H^1$ , thus in  $L^p$  for  $2 \leq p \leq 6$ . So each of the last three terms stays bounded, because  $J$  is bounded by Hardy–Littlewood–Sobolev

$$J[\rho] \leq C \|\rho\|_{L^{\frac{6}{5}}}^2 < \infty,$$

the exchange–correlation term by

$$\int_{\mathbb{R}^3} e'_{xc}(\rho) \rho \, dx \leq C \int_{\mathbb{R}^3} \sqrt{\rho}^{2+2\beta_-} + \sqrt{\rho}^{2+2\beta_+}$$

with  $2 + 2\beta_{\pm} \in [2, 6]$  and the last term by the boundedness of  $\theta_n^l$  and the already discussed  $L^2$ -norm bound. Thus taking the limit  $n \rightarrow \infty$  we get

$$\left(-\frac{1}{2} \Delta + \rho_k * \frac{1}{|x|} + e'_{xc}(\rho_k)\right) \psi_{l,k} + \theta^l \psi_{l,k} = 0, \tag{72}$$

where  $\theta^l = \liminf_{n \rightarrow \infty} \theta_n^l > 0$ . Furthermore we have

$$\rho_k = \sum_l |\psi_{l,k}|^2. \tag{73}$$

Since the mass of the  $\rho_{\gamma_n^{(k)}}$  does not depend on  $n$  we obtain from (71)

$$\lim_{k \rightarrow \infty} \|\rho_k\|_{L^1} = 0. \tag{74}$$

By multiplying (72) with  $\psi_{l,k}$ , integrating and summing over  $l$  we obtain

$$\begin{aligned} 0 &\geq - \sum_l \theta^l \|\psi_{l,k}\|_{L^2}^2 - \int_{\mathbb{R}^3} \left(\rho_k * \frac{1}{|x|}\right) |\psi_{l,k}|^2 \, dx \\ &= \frac{1}{2} \sum_l \|\nabla \psi_{l,k}\|_{L^2}^2 + \sum_l \int_{\mathbb{R}^3} |\psi_{l,k}|^2 e'_{xc}(\rho_k) \, dx. \end{aligned}$$

Now we can again use Assumption 1 stating  $e'_{xc}(\rho) \leq C(\rho^{\beta_-} + \rho^{\beta_+})$  on the second term and then apply Hölder’s inequality with  $\beta_+$  and  $\beta_-$ , respectively, to obtain the following bound

$$\begin{aligned} 0 &\geq \frac{1}{2} \sum_l \|\nabla \psi_{l,k}\|_{L^2}^2 - C \sum_l \left( \|\psi_{l,k}^2\|_{L^{\frac{1}{1-\beta_-}}} + \|\psi_{l,k}^2\|_{L^{\frac{1}{1-\beta_+}}} \right) \|\rho_k\|_{L^1} \\ &\geq \frac{1}{2} \sum_l \|\nabla \psi_{l,k}\|_{L^2}^2 - C \|\rho_k\|_{L^1} \underbrace{\sum_l \|\psi_{l,k}\|_{H^1}^2}_{< \infty}, \end{aligned}$$

where we applied the Sobolev embedding in the last inequality on the two terms  $\|\psi^2\|_{L^{\frac{1}{1-\beta}}} = \|\psi\|_{L^{\frac{2}{1-\beta}}}^2$ . Indeed, by Assumption 1 we have  $0 < \beta < \frac{2}{3}$ , implying  $\frac{2}{1-\beta} \in (2, 6)$ . Thus by taking the limit  $k \rightarrow \infty$  and using (74) we obtain

$$\sum_l \|\nabla \psi_{l,k}\|_{L^2}^2 \xrightarrow{k \rightarrow \infty} 0.$$

Applying standard elliptic regularity results (see e.g. [12]) to (72) now give us the inequality

$$\|\psi_{l,k}\|_{L^\infty} \leq C \|\psi_{l,k}\|_{H^1},$$

where the constant  $C > 0$  does not depend on  $k$  and thus

$$\lim_{k \rightarrow \infty} \sum_l \|\psi_{l,k}\|_{L^\infty}^2 = 0.$$

Thus by (73) we also obtain

$$\lim_{k \rightarrow \infty} \|\rho_k\|_{L^\infty} = 0.$$

Again from (72) and from Assumption 1 we deduce

$$\theta^l \|\psi_{l,k}\|_{L^2}^2 \leq C (\|\rho_k\|_{L^\infty}^{2\beta_-} + \|\rho_k\|_{L^\infty}^{2\beta_+}) \|\psi_{l,k}\|_{L^2}^2. \tag{75}$$

Now note that due to (62) at most  $N$  different energy levels are occupied. Thus

$$\|\psi_{l,k,n}\|_{L^2}^2 = \lambda_l^n = 0,$$

for all  $l$  corresponding to the  $(N + 1)$ th or higher eigenvalues without counting multiplicity. Note that due to degeneracies this might not be the same as  $l > N$ . Therefore we directly get for those  $l$

$$\|\psi_{l,k}\|_{L^2} = 0.$$

Thus the mass of  $\rho_k$  is distributed among only finitely many energy levels  $l$ . Therefore for at least one fixed level  $l$  we can find up to a subsequence in  $k$   $\psi_{l,k}$  such that

$$\|\psi_{l,k}\|_{L^2} \neq 0, \quad \forall k,$$

because otherwise we would have  $\|\rho_k\|_{L^1} = 0$ . Hence (75) becomes

$$\theta^l \leq C (\|\rho_k\|_{L^\infty}^{2\beta_-} + \|\rho_k\|_{L^\infty}^{2\beta_+}) \xrightarrow{k \rightarrow \infty} 0,$$

which is a contradiction to Lemma 9.

Thus case (ii) cannot happen and the proof is hence complete.

### 4.3. Proof of Proposition 1

First we will show that we can get rid of the antisymmetry condition, i.e.

$$I_R^{H_2} = \inf_{\substack{\psi \in H^1(\mathbb{R}_\Sigma^2), \\ \|\psi\|_{L^2} = 1, \psi \text{ antisymm.}}} \langle \psi, H(x, y)\psi \rangle = \inf_{\substack{\psi \in H^1(\mathbb{R}^2), \\ \|\psi\|_{L^2} = 1}} \langle \psi, H(x, y)\psi \rangle =: \tilde{I}_R^{H_2}$$

Take any  $\psi \in H^1(\mathbb{R}_\Sigma^2), \|\psi\|_{L^2} = 1$ , then

$$\begin{aligned} \langle \psi, H(x, y)\psi \rangle &= \sum_{s,t \in \Sigma} \langle \psi(\cdot, s, \cdot, t), H(x, y)\psi(\cdot, s, \cdot, t) \rangle \\ &\geq \sum_{s,t \in \Sigma} \tilde{I}_R^{H_2} \|\psi(\cdot, s, \cdot, t)\|_{L^2(\mathbb{R}^2)}^2 \\ &= \tilde{I}_R^{H_2} \|\psi\|_{L^2((\mathbb{R}_\Sigma)^2)}^2 = \tilde{I}_R^{H_2}. \end{aligned}$$

So we have  $I_R^{H_2} \geq \tilde{I}_R^{H_2}$ . For the other direction define for any given  $\psi \in H^1(\mathbb{R}^2)$  with  $\|\psi\|_{L^2}^2$

$$\tilde{\psi}(x, s, y, t) := \frac{1}{\sqrt{2}} \left( \psi(x, y) \delta_{|\uparrow\rangle}(s) \delta_{|\downarrow\rangle}(t) - \psi(y, x) \delta_{|\uparrow\rangle}(t) \delta_{|\downarrow\rangle}(s) \right),$$

where  $\delta_{|\uparrow\rangle}(s)$  denotes the Kronecker delta for the spin component. By definition this  $\tilde{\psi}$  satisfies the antisymmetry condition and is normalized. Furthermore we have

$$\langle \tilde{\psi}, H(x, y) \tilde{\psi} \rangle_{L^2(\mathbb{R}_\Sigma^2)} = \langle \psi, H(x, y) \psi \rangle_{L^2(\mathbb{R}^2)},$$

taking the infimum over normalized  $\psi$  gives

$$I_R^{H_2} \leq \inf_{\tilde{\psi}} \langle \tilde{\psi}, H(x, y) \tilde{\psi} \rangle = \inf_{\psi} \langle \psi, H(x, y) \psi \rangle = \tilde{I}_R^{H_2}.$$

So from here on we will consider the system without the antisymmetry condition.

### 4.3.1. Upper bound

For the upper bound we can take a  $\varphi \in C_c^\infty(\mathbb{R})$  with  $\|\varphi\|_{L^2} = 1$  and consider as a testfunction for the  $H_2$  Hamiltonian just the tensor product  $\psi = \varphi \otimes \varphi(\cdot - R)$ , i.e.  $\psi(x, y) = \varphi(x)\varphi(y - R)$ . Then we directly get

$$\begin{aligned} \lim_{R \rightarrow \infty} E_R^{H_2} &\leq \lim_{R \rightarrow \infty} \langle \psi, H_R(x, y) \psi \rangle \\ &= \lim_{R \rightarrow \infty} 2\langle \varphi, h(x) \varphi \rangle + |\varphi|^2(R) + |\varphi|^2(-R) + 2 \int \varphi(\pm y) \varphi(y - R) dy \\ &= 2\langle \varphi, h(x) \varphi \rangle, \end{aligned}$$

where we used that the last three terms vanish as soon as  $R > \text{diam}(\text{supp } \varphi)$ . Taking now the infimum w.r.t.  $\varphi$  and noting that the Hamiltonian  $h$  is continuous on  $H^1(\mathbb{R})$  we get the result.

### 4.3.2. Lower bound

Since the electron–electron interaction is positive we directly get

$$H(x, y) \geq \tilde{h}(x) + \tilde{h}(y), \tag{76}$$

where  $\tilde{h}(x) = -\frac{1}{2} \frac{d^2}{dx^2} - \delta_0(x) - \delta_R(x)$ . To determine the infimum over the right hand side, we can just consider tensor-products of functions due to the additive structure. Hence we only need to consider

$$\langle \varphi, \tilde{h}(x) \varphi \rangle, \quad \varphi \in L^2(\mathbb{R}).$$

Therefore consider any arbitrary  $\varphi \in H^1(\mathbb{R})$ , and two cut-off functions  $\xi_1$  and  $\xi_2$  with

$$\xi_1^2 + \xi_2^2 = 1, \quad \xi_1(x) = 1 \text{ for } x \leq \frac{1}{3}, \quad \xi_1(x) = 0 \text{ for } x \geq \frac{2}{3}.$$

Defining then

$$\varphi_i = \xi_i \left( \frac{\cdot}{R} \right) \varphi,$$

gives with a straightforward calculation

$$\langle \varphi, \tilde{h}(x) \varphi \rangle \geq \langle \varphi, h_0(x) \varphi \rangle + \langle \varphi, h_R(x) \varphi \rangle + o(1). \tag{77}$$

Here  $h_0(x), h_R(x)$  denote the hamiltonian  $h(x)$  with the nucleus sitting at the origin  $x = 0$  and  $x = R$ , respectively. This now directly gives

$$(77) \geq \varepsilon \left( \|\varphi_1\|_{L^2}^2 + \|\varphi_2\|_{L^2}^2 \right) + o(1) = \varepsilon \|\varphi\|_{L^2}^2 + o(1), \tag{78}$$

where we used that the lowest eigenvalue  $\varepsilon = I^H$  of  $h(x)$  does not depend on the position of the single nucleus in the system.

Combining this lower bound with (76) directly gives the desired assertion.

## Symbols

$\mathcal{A}_N$	set of admissible $N$ -electron wavefunctions
$\beta_{\pm}$	exponents in part 3 of <a href="#">Assumption 1</a> for controlling the derivative of $e_{xc}$
$\gamma$	closed-shell one-electron reduced density operator
$D[\cdot, \cdot]$	bilinear form associated with the Hartree energy
$D_N$	set of $N$ -body density matrices
$\mathcal{E}^X[\cdot]$	energy functional for the $X$ -atom
$\mathcal{E}^{X_2}[\cdot]$	energy functional for the $X_2$ -molecule
$E_{\alpha}$	TFDW energy of the system with mass $\alpha$
$E_0^{QM}$	quantum mechanical ground state energy
$E_{ex}[\cdot]$	exchange–correlation energy
$e_{ex}$	LDA exchange–correlation function
$F_{LL}[\cdot]$	Levy–Lieb energy functional
$\mathcal{H}_N$	Hilbert space given by $\bigwedge_{i=1}^N L^2(\mathbb{R}^3)$
$\mathcal{H}$	subspace of $\mathfrak{S}_1$ with finite kinetic energy
$h_{\gamma}$	hamiltonian coming from the Euler–Lagrange equation
$H_N^V$	$N$ -electron hamiltonian with potential $V$
$I_{\lambda}^X$	energy of the $X$ -atom with $\lambda$ electrons surrounding it
$I_{\lambda,R}^{X_2}$	energy of the $X_2$ -molecules with $\lambda$ electrons surrounding it and distance $R$ between the nuclei
$I_{\lambda}^{\infty}$	energy of the problem at infinity, i.e. without potential, with $\lambda$ electrons surrounding it
$J[\cdot]$	Hartree energy of a density
$\Psi$	$N$ -body wavefunction
$K_{\lambda}$	set of admissible density matrices $\gamma$ with mass $\text{tr}[\gamma] = \lambda$
$\rho$	one-body reduced electron density
$\mathcal{R}_N$	set of admissible one-body electron densities arising from the wavefunctions in $\mathcal{A}_N$
$\mathcal{RD}_N$	set of admissible one-body electron densities arising from the density matrices in $D_N$
$\mathbb{R}_{\Sigma}^3$	space of position and spin; elements are denoted by $z = (x, y)$
$\mathfrak{S}_1$	set of trace class operators on $L^2(\mathbb{R}^3)$
$\Sigma$	set of spin-states $\{ \uparrow\rangle,  \downarrow\rangle\}$
$T[\cdot]$	kinetic energy of the system, depending on the setting of $\rho, \gamma$ or $\Psi$
$\tau_R$	unitary translation operator
$\theta_i$	eigenvalues corresponding to $h_{\gamma}$
$V$	Coulomb potential generated by the clamped nuclei
$V_{ee}[\cdot]$	electron–electron interaction energy
$V_{ne}[\cdot]$	electron–nuclei interaction energy
$\xi$	partition of unity in the dichotomy case
$\zeta$	partition of unity in the dichotomy case

## Acknowledgments

The authors thank Gero Friesecke for helpful discussions and the anonymous reviewers for their helpful comments which helped improve and clarify this manuscript.

Both B.R.G. and S.B. gratefully acknowledge support from the International Research Training Group IGDK Munich - Graz funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 188264188/GRK1754.

**Appendix. Calculating the exact one-dimensional DFT energy**

For the readers' convenience we sketch the calculations for the exact ground state energy of the one-dimensional DFT model (see Section 3.1) with the ground-state first reported in [41]. We consider the corresponding energy functional from (45)

$$\mathcal{E}^H[\rho] = \frac{1}{2} \int (\sqrt{\rho'})^2 dx - \rho(0) + \left(\frac{1}{2} - c_{xc}\right) \int \rho^2 dx$$

with ground-state given by (46)

$$\rho = \alpha|\psi|^2, \quad \text{with} \quad \psi(x) = a \cdot \text{sech}(b|x| + x_0), \tag{79}$$

where the parameters  $a, b, x_0$  only depend on  $\alpha$  and  $c_{xc}$  and are given by

$$x_0 = \text{arctanh}\left(\frac{1}{b}\right), \quad a = \sqrt{\frac{b^2}{2(b-1)}}, \quad b = 1 - \alpha \frac{1 - 2c_{xc}}{2}.$$

Note that we are interested in the case  $c_{xc} > \frac{1}{2}$ , thus  $b > 0$ . We start off by recalling some basic properties of the sech-Function

$$\frac{d}{dx} \text{sech}(x) = -\text{sech}(x) \cdot \tanh(x)$$

and

$$\int \text{sech}(x)^4 dx = \frac{2}{3} \tanh(x) + \frac{1}{3} \tanh(x)\text{sech}(x)^2, \quad \int \text{sech}(x)^2 \tanh(x)^2 dx = \frac{1}{3} \tanh(x)^3.$$

This implies for the last term in the energy functional

$$\begin{aligned} \int \rho^2 dx &= \alpha^2 a^4 \int_{-\infty}^{\infty} \text{sech}(b|x| + x_0)^4 dx \\ &= \alpha^2 2a^4 \frac{1}{b} \int_{x_0}^{\infty} \text{sech}(y)^4 dy \\ &= \alpha^2 \frac{2a^4}{b} \left( \frac{2}{3} - \frac{2}{3} \tanh(x_0) - \frac{1}{3} \tanh(x_0)\text{sech}(x_0)^2 \right) \\ &= \alpha^2 \frac{2a^4}{b} \left( \frac{2}{3} - \frac{2}{3b} - \frac{1}{3b} \left(1 - \frac{1}{b^2}\right) \right) \\ &= \frac{\alpha^2}{6} (2b + 1). \end{aligned}$$

For the kinetic energy we obtain

$$\begin{aligned} \int (\sqrt{\rho'})^2 dx &= \alpha a^2 b^2 \int_{-\infty}^{\infty} \text{sech}(b|x| + x_0)^2 \tanh(b|x| + x_0)^2 dx \\ &= 2\alpha a^2 b \int_{x_0}^{\infty} \text{sech}(y)^2 \tanh(y)^2 dy \\ &= 2\alpha a^2 b \frac{1}{3} \left( 1 - \tanh(x_0)^3 \right) = 2\alpha a^2 b \frac{1}{3} \left( 1 - \frac{1}{b^3} \right) \\ &= \frac{\alpha}{3} (b^2 + b + 1). \end{aligned}$$



For the term  $\rho(0)$  we have

$$\rho(0) = \alpha a^2 \operatorname{sech}(x_0)^2 = \alpha a^2 \left(1 - \frac{1}{b^2}\right) = \alpha \frac{b+1}{2}.$$

Therefore we obtain for the total energy

$$I_\alpha^H = \mathcal{E}^H[\rho_\alpha] = \frac{\alpha}{6}(b^2 + b + 1) - \alpha \frac{b+1}{2} + \left(\frac{1}{2} - c_{xc}\right) \frac{\alpha^2}{6}(2b + 1)$$

Plugging now  $b = 1 - \alpha\left(\frac{1}{2} - c_{xc}\right)$  into this equation gives now the desired result

$$I_\alpha^H + I_{2-\alpha}^H = \frac{1}{12}(\alpha^2(3 - 12c_{xc}^2) + 6\alpha(4c_{xc}^2 - 1) - 4(1 + 2c_{xc} + 4c_{xc}^2)).$$

## References

- [1] A. Anantharaman, E. Cancès, Existence of minimizers for Kohn–Sham models in quantum chemistry, *Ann. Inst. Henri Poincaré C* 26 (6) (2009) 2425–2455.
- [2] X. Andrade, D. Strubbe, U. De Giovannini, A. Larsen, M.J.T. Oliveira, J. Alberdi-Rodriguez, A. Varas, I. Theophilou, N. Helbig, M.J. Verstraete, L. Stella, F. Nogueira, A. Aspuru-Guzik, A. Castro, M.A.L. Marques, A. Rubio, Real-space grids and the Octopus code as tools for the development of new simulation approaches for electronic systems, *Phys. Chem. Chem. Phys.* 17 (2015) 31371–31396.
- [3] A.D. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior, *Phys. Rev. A* 38 (1988) 3098–3100.
- [4] M. Born, R. Oppenheimer, Zur quantentheorie der molekeln, *Ann. Phys.* 389 (20) (1927) 457–484.
- [5] H. Chen, G. Friesecke, Pair densities in density functional theory, *Multiscale Model. Simul.* 13 (4) (2015) 1259–1289.
- [6] E. Davidson, *Reduced Density Matrices in Quantum Chemistry*, Academic Press, New York, 1976.
- [7] P.A.M. Dirac, Note on exchange phenomena in the Thomas atom, in: *Math. Proc. Cambridge Philos. Soc.*, 26 (3) (1930) 376–385.
- [8] E. Engel, R.M. Dreizler, *Density Functional Theory*, Springer, 2013.
- [9] G. Friesecke, Pair correlations and exchange phenomena in the free electron gas, *Comm. Math. Phys.* 184 (1) (1997) 143–171.
- [10] G. Friesecke, The multiconfiguration equations for atoms and molecules: charge quantization and existence of solutions, *Arch. Ration. Mech. Anal.* 169 (1) (2003) 35–71.
- [11] M. Fuchs, Y.-M. Niquet, X. Gonze, K. Burke, Describing static correlation in bond dissociation by Kohn–Sham density functional theory, *J. Chem. Phys.* 122 (9) (2005) 094116.
- [12] D. Gilbarg, N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, 2015.
- [13] D. Gontier, Existence of minimizers for Kohn–Sham within the local spin density approximation, *Nonlinearity* 28 (1) (2014) 57–76.
- [14] D. Gontier, M. Lewin, F.Q. Nazar, The nonlinear Schrödinger equation for orthonormal functions: I. Existence of ground states, 2020, arXiv preprint [arXiv:2002.04963](https://arxiv.org/abs/2002.04963).
- [15] S.J. Gustafson, I.M. Sigal, *Mathematical Concepts of Quantum Mechanics*, vol. 33, Springer, 2003.
- [16] P. Hohenberg, W. Kohn, Inhomogeneous electron gas, *Phys. Rev. (2)* 136 (1964) B864–B871.
- [17] M. Holst, H. Hu, J. Lu, J.L. Marzuola, D. Song, J. Weare, Symmetry breaking in density functional theory due to Dirac exchange for a hydrogen molecule, 2019, arXiv preprint [arXiv:1902.03497](https://arxiv.org/abs/1902.03497).
- [18] B. Jackson, H. Metiu, The dynamics of H2 dissociation on Ni (100): A quantum mechanical study of a restricted two-dimensional model, *J. Chem. Phys.* 86 (2) (1987) 1026–1035.
- [19] W. Kohn, L.J. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev. (2)* 140 (1965) A1133–A1138.
- [20] A. Laestadius, One-dimensional Lieb–Oxford bounds, 2019, [arXiv:1910.01925](https://arxiv.org/abs/1910.01925).
- [21] C. Le Bris, Some results on the Thomas–Fermi–Dirac–von Weizsäcker model, *Differential Integral Equations* 6 (2) (1993) 337–353.
- [22] M. Levy, Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem, *Proc. Natl. Acad. Sci.* 76 (12) (1979) 6062–6065.
- [23] E.H. Lieb, Sharp constants in the Hardy–Littlewood–Sobolev and related inequalities, *Ann. of Math.* 118 (2) (1983) 349–374.
- [24] E.H. Lieb, Thomas–Fermi And related theories of atoms and molecules, in: *The Stability of Matter: From Atoms To Stars*, Springer, 1997, pp. 259–297.
- [25] E.H. Lieb, Density functionals for Coulomb systems, in: *Inequalities*, Springer, 2002, pp. 269–303.
- [26] E.H. Lieb, M. Loss, *Analysis*, Graduate Series in Mathematics, vol. 14, Amer. Math. Soc., Providence, 2001.

- [27] P.-L. Lions, The concentration-compactness principle in the calculus of variations. The locally compact case, part 1, *Ann. Inst. Henri Poincaré Anal. Nonlinéaire* 1 (2) (1984) 109–145.
- [28] P.-L. Lions, The concentration-compactness principle in the calculus of variations. The locally compact case, part 2, *Ann. Inst. Henri Poincaré Anal. Nonlinéaire* 1 (4) (1984) 223–283.
- [29] P.-L. Lions, Solutions of Hartree-Fock equations for Coulomb systems, *Comm. Math. Phys.* 109 (1) (1987) 33–97.
- [30] J. Lu, F. Otto, Nonexistence of a minimizer for Thomas–Fermi–Dirac–von Weizsäcker model, *Comm. Pure Appl. Math.* 67 (2014).
- [31] R.J. Magyar, K. Burke, Density-functional theory in one dimension for contact-interacting fermions, *Phys. Rev. A* 70 (3) (2004) 032508.
- [32] M.G. Medvedev, I.S. Bushmarinov, J. Sun, J.P. Perdew, K.A. Lyssenko, Density functional theory is straying from the path toward the exact functional, *Science* 355 (6320) (2017) 49–52.
- [33] R.G. Parr, W. Young, *Density Functional Theory of Atoms and Molecules*, Oxford University Press, 1989.
- [34] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.* 77 (1996) 3865–3868.
- [35] J.P. Perdew, A. Savin, K. Burke, Escaping the symmetry dilemma through a pair-density interpretation of spin-density functional theory, *Phys. Rev. A* 51 (6) (1995) 4531.
- [36] J.P. Perdew, Y. Wang, Accurate and simple analytic representation of the electron-gas correlation energy, *Phys. Rev. B* 45 (1992) 13244–13249.
- [37] J.P. Perdew, A. Zunger, Self-interaction correction to density-functional approximations for many-electron systems, *Phys. Rev. B* 23 (1981) 5048–5079.
- [38] M. Reed, B. Simon, *Methods of Modern Mathematical Physics: Analysis of Operators*, in: *Methods of Modern Mathematical Physics*, (no. Bd. 4) Academic Press, 1978.
- [39] J.C. Slater, A simplification of the Hartree-Fock method, *Phys. Rev.* 81 (1951) 385–390.
- [40] P.J. Stephens, F.J. Devlin, C.F. Chabalowski, M.J. Frisch, Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields, *J. Phys. Chem.* 98 (45) (1994) 11623–11627.
- [41] D. Witthaut, S. Mossmann, H.J. Korsch, Bound and resonance states of the nonlinear Schrödinger equation in simple model systems, *J. Phys. A: Math. Gen.* 38 (8) (2005) 1777.
- [42] Y. Zhang, C.H. Cheng, J. Kim, J. Stanojevic, E. Eyler, Dissociation energies of molecular hydrogen and the hydrogen molecular ion, *Phys. Rev. Lett.* 92 20 (2004) 203003.

## Appendix B

### Further Articles under Review

- B.1 Two-electron wavefunctions are matrix product states with bond dimension Three

# Two-electron wavefunctions are matrix product states with bond dimension Three

Gero Friesecke and Benedikt R. Graswald

---

In contrast to Core Article II, this article considers not only re-orderings of the underlying one-body basis, but general unitary transformations, also known as fermionic mode transformations. Our work is motivated by the following empirical observation by Krumnow, Veis, Legeza, and Eisert [89, 90]: Going beyond ordering and *optimizing over fermionic mode transformations* can reduce the approximation error a great deal further in systems of interest. After introducing the mathematical framework in Section 2 and 3, we describe how QC-DMRG together with this optimization over the single-particle basis can be viewed as generalization of the classical Hartree-Fock (HF) method. In particular, utilizing the size of the bond dimensions as parameters it interpolates between HF (bond dimension = 1) and the full configuration-interaction method (FCI) (bond dimension  $2^{L/2}$ , where  $L$  is the number of basis functions).

The main results are given in Section 4 and consist of three theorems. First, we prove an upper bound by finding a good basis consisting of the so-called natural orbitals, i.e., the eigenstates of the single-particle reduced density matrix, and providing explicitly the MPS decomposition.

Even though discovering this representation was joint work, the final version presented in the paper which starts with a pair states point of view was developed by my co-author Gero Friesecke, in particular Lemma 2 is by him. He was also the one suggesting to consider the natural orbitals to simplify the structure of the coefficient tensor of the wavefunction.

Theorem 2 is the technical most challenging part and was proven by myself. It provides the corresponding lower bound. Here, the crucial point consists in analyzing the algebraic structure of the unfoldings of the coefficient tensor of the wavefunction independently of the underlying one-particle basis.

We find a dramatic effect, namely a *reduction of the bond dimension needed for exactness of the method from  $2 + \frac{L}{2}$  to 3*, where  $L$  is the number of single-particle basis functions. Furthermore, as the bounds in Theorem 1 and 2 coincide, we obtain a full characterization of the optimal bound dimension in the two-electron setting, i.e., we find that 3 is in fact optimal. Previous exact representations of quantum states in the form of low-bond-dimension MPS were, to our knowledge, limited to very special states, the prototype example being the AKLT state from spin physics [1] which arises as the ground state of a particular translation invariant Hamiltonian. On the other hand, the present result – unlike that in [1] – is limited to  $N = 2$ .

Theorem 3 then summarizes our results and its application to the QC-DMRG method and is mostly due to my co-author Gero Friesecke. Finally, we remark that the exact bond-dimension-three representation of two-fermion wavefunctions carries over to the infinite-dimensional single-particle Hilbert space  $L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$  of full two-electron quantum mechanics.

*Own contribution.* I was significantly involved in finding the ideas and carrying out the scientific work of all parts of this article. In particular, I discovered the explicit matrices for the upper bound and found the arguments for and carried out the lower bounds. Furthermore, I took an active part in writing the first draft of the article as well as all major parts of the final version.

# Permission to include:

Gero Friesecke and Benedikt R. Graswald (2021).

Two-electron wavefunctions are matrix product states with bond dimension Three  
*arXiv preprint* arXiv:2109.10091.

Submitted to *Journal of Mathematical Physics*.

# Permission to Reuse Content

## REUSING AIP PUBLISHING CONTENT

Permission from AIP Publishing is required to:

- republish content (e.g., excerpts, figures, tables) if you are not the author
- modify, adapt, or redraw materials for another publication
- systematically reproduce content
- store or distribute content electronically
- copy content for promotional purposes

To request permission to reuse AIP Publishing content, use RightsLink® for the fastest response or contact AIP Publishing directly at [rights@aip.org](mailto:rights@aip.org) (<mailto:rights@aip.org>) and we will respond within one week:

For RightsLink, use Scitation to access the article you wish to license, and click on the Reprints and Permissions link under the TOOLS tab. (For assistance click the “Help” button in the top right corner of the RightsLink page.)

To send a permission request to [rights@aip.org](mailto:rights@aip.org) (<mailto:rights@aip.org>), please include the following:

- Citation information for the article containing the material you wish to reuse
- A description of the material you wish to reuse, including figure and/or table numbers
- The title, authors, name of the publisher, and expected publication date of the new work
- The format(s) the new work will appear in (e.g., print, electronic, CD-ROM)
- How the new work will be distributed and whether it will be offered for sale

Authors do **not** need permission from AIP Publishing to:

- quote from a publication (please include the material in quotation marks and provide the customary acknowledgment of the source)
- reuse any materials that are licensed under a Creative Commons CC BY license (please format your credit line: “Author names, Journal Titles, Vol.#, Article ID#, Year of Publication; licensed under a Creative Commons Attribution (CC BY) license.”)
- reuse your own AIP Publishing article in your thesis or dissertation (please format your credit line: “Reproduced from [FULL CITATION], with the permission of AIP Publishing”)
- make multiple copies of articles—although you must contact the Copyright Clearance Center (CCC) at [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) to do this

## REUSING CONTENT PUBLISHED BY OTHERS

To request another publisher’s permission to reuse material in AIP Publishing articles, please use our

# Two-electron wavefunctions are matrix product states with bond dimension Three

Gero Friesecke and Benedikt R. Graswald

Department of Mathematics, Technical University of Munich, Germany

gf@ma.tum.de, graswabe@ma.tum.de

September 27, 2021

## Abstract

We prove the statement in the title, for a suitable (wavefunction-dependent) choice of the underlying orbitals, and show that Three is optimal. Thus for two-electron systems, the QC-DMRG method with bond dimension Three combined with fermionic mode optimization exactly recovers the FCI energy.

## 1 Introduction

The  $N$ -electron Schrödinger equation is a partial differential equation in  $\mathbb{R}^{3N}$  and its direct numerical solution is prohibited for large  $N$  by the curse of dimension. As a consequence, a large variety of approximate methods have been developed since the early days of quantum mechanics, starting with the work of Thomas, Fermi, Dirac, Hartree, and Fock. In the past decade, the Quantum Chemistry Density Matrix Renormalization Group (QC-DMRG) method [19, 16, 4, 15] has become the state-of-the-art choice for systems with up to a few dozen electrons; see [18] for a recent review.

In QC-DMRG, one chooses a suitable finite single-particle basis, makes a matrix product state (MPS) alias tensor train ansatz for the coefficient tensor of the many-particle wavefunction in Fock space, and optimizes the Rayleigh quotient over the matrices (see Sections 3 and 4 for a detailed description). The key parameter in the method is the maximal allowed size of the matrices, called bond dimension. For bond dimension 1 the MPS ansatz reduces to a single Slater determinant built from the basis functions. For large bond dimension the ansatz recovers all wavefunctions in the Fock space, but large means impractically large (more precisely:  $2^{L/2}$ , where  $L$  is the number of single-particle basis functions [17]).

It has long been known that the accuracy strongly depends on the choice of basis, and can typically be improved by re-ordering the basis (see [2, 6, 18]; also, see [9] for extreme examples where ordering does not yield an improvement).

This paper is motivated by an empirical phenomenon observed by Krumnow, Veis, Legeza, and Eisert [14, 13]: going beyond ordering and *optimizing over fermionic mode transformations* (i.e., general unitary transformations of the single-particle basis) can reduce the approximation error a great deal further in systems of interest. QC-DMRG together with optimization over the single-particle basis as introduced in [14, 13] can be viewed as a generalization of the classical Hartree-Fock method, to which it reduces for bond dimension 1 (see Section 4).

In the absence of previous mathematical results on the influence of mode transformations on the approximation error, we investigate here the simplest case  $N = 2$ . We find a dramatic effect, namely a *reduction of the bond dimension needed for exactness of the method from  $2 + \frac{L}{2}$  to 3*, where  $L$  is the number of single-particle basis functions (Theorem 3 in Section 4). This is proved by showing that general two-particle wavefunctions can be represented exactly with bond dimension 3 after a (wavefunction-dependent) optimal mode transformation, with 3 being optimal. See Theorems 1 and 2 in Section 4.

Previous exact representations of quantum states in the form of low-bond-dimension MPS were, to our knowledge, limited to very special states, the prototype example being the AKLT state from spin physics [1] which arises as the ground state of a particular translation invariant Hamiltonian. On the other hand, the present result – unlike that in [1] – is limited to  $N = 2$  (see the Conclusions for further discussion of this point).

Finally, we remark that the exact bond-dimension-three representation of two-fermion wavefunctions carries over to the infinite-dimensional single-particle Hilbert space  $L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$  of full two-electron quantum mechanics, as shown in the last part of this paper.

## 2 Fock space and occupation representation

*Fermionic Fock space.* We first consider a finite dimensional single-particle Hilbert space  $\mathcal{H}_L$ , whose dimension we denote by  $L$ . The associated state space for a system of  $N$  fermions is the  $N$ -fold antisymmetric product  $\mathcal{V}_{N,L} := \bigwedge_{i=1}^N \mathcal{H}_L$ , and the resulting Fock space is defined as the direct sum of the  $N$ -particle spaces,

$$\mathcal{F}_L := \bigoplus_{N=0}^L \mathcal{V}_{N,L}, \quad (1)$$

where  $V_{0,L} \cong \mathbb{C}$  is spanned by the vacuum state  $\Omega$ . When the particles are electrons,  $\mathcal{H}_L$  would correspond to a subspace of  $L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$  spanned by  $L$  spin orbitals. If the orbitals are the occupied and lowest unoccupied eigenstates of the Hartree-Fock Hamiltonian associated with the electronic Schrödinger equation,  $\mathcal{V}_{N,L}$  is known in physics as the full configuration interaction (full CI) space (see e.g. [11]).



Now given an orthonormal basis  $\{\varphi_1, \dots, \varphi_L\}$  of the single-particle Hilbert space  $\mathcal{H}_L$ , we can write any element  $\Psi \in \mathcal{F}_L$  in the form

$$\Psi = c_0 \Omega + \sum_{i=1}^L c_i \varphi_i + \sum_{1 \leq i < j \leq L} c_{ij} |\varphi_i \varphi_j\rangle + \sum_{1 \leq i < j < k \leq L} c_{ijk} |\varphi_i \varphi_j \varphi_k\rangle + \dots, \quad (2)$$

with  $|\varphi_{i_1} \dots \varphi_{i_N}\rangle$  denoting the antisymmetric tensor product alias Slater determinant

$$|\varphi_{i_1} \dots \varphi_{i_N}\rangle = \varphi_{i_1} \wedge \dots \wedge \varphi_{i_N} \in \mathcal{V}_{N,L}. \quad (3)$$

*Occupation representation.* Instead of the above 'first quantized' representation, in QC-DMRG one considers a 'second quantized' representation by occupation numbers of orbitals in Fock space. A Slater determinant  $|\varphi_{i_1} \dots \varphi_{i_N}\rangle \in \mathcal{V}_{N,L}$  is represented by a binary string  $(\mu_1, \dots, \mu_L) \in \{0, 1\}^L$ , with  $\mu_i$  indicating whether or not the orbital  $\varphi_i$  is present (occupied) or absent (unoccupied). An example with  $N = 4$  and  $L = 8$  is

$$|\varphi_2 \varphi_3 \varphi_6 \varphi_8\rangle \longleftrightarrow (0, 1, 1, 0, 0, 1, 0, 1), \quad \begin{array}{cccccccc} \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\ | & | & | & | & | & | & | & | \\ \circ & \bullet & \bullet & \circ & \circ & \bullet & \circ & \bullet \\ | & | & | & | & | & | & | & | \\ \varphi_1 & \varphi_2 & & & & & & \varphi_L \end{array}$$

since  $\varphi_1$  is unoccupied,  $\varphi_2$  is occupied,  $\varphi_3$  is occupied, and so on. The Slater determinant (3) indexed by its binary label is in the following denoted  $\Phi_{\mu_1 \dots \mu_L}$ , that is to say

$$\Phi_{\mu_1 \dots \mu_L} := |\varphi_{i_1} \dots \varphi_{i_N}\rangle \text{ if } \mu_i = 1 \text{ exactly when } i \in \{i_1, \dots, i_N\}, \quad i_1 < \dots < i_N. \quad (4)$$

The coefficients in the expansion (2) indexed by the corresponding binary label are called  $C_{\mu_1 \dots \mu_L}$ , that is to say

$$C_{\mu_1 \dots \mu_L} = c_{i_1 \dots i_N} \text{ if } \mu_i = 1 \text{ precisely when } i \in \{i_1, \dots, i_N\}, \quad i_1 < \dots < i_N, \quad (5)$$

yielding the occupation representation

$$\Psi = \sum_{\mu_1, \dots, \mu_L=0}^1 C_{\mu_1 \dots \mu_L} \Phi_{\mu_1 \dots \mu_L}. \quad (6)$$

### 3 Matrix product states

A matrix product state (MPS) or tensor train (TT) with respect to the basis  $\{\varphi_i\}_{i=1}^L$  with size parameters ('bond dimensions')  $r_i$  ( $i = 1, \dots, L - 1$ ) is a state of the form

$$\Psi = \sum_{\mu_1, \dots, \mu_L=0}^1 A_1[\mu_1] A_2[\mu_2] \dots A_L[\mu_L] \Phi_{\mu_1 \dots \mu_L} \in \mathcal{F}_L \quad (7)$$

where for every  $(\mu_1, \dots, \mu_L)$ ,  $A_i[\mu_i]$  is a  $r_{i-1} \times r_i$  matrix, with the convention  $r_0 = r_L = 1$ . Writing out the above matrix multiplications,

$$A_1[\mu_1]A_2[\mu_2] \dots A_L[\mu_L] = \sum_{\alpha_1=1}^{r_1} \sum_{\alpha_2=1}^{r_2} \dots \sum_{\alpha_{L-1}=1}^{r_{L-1}} (A_1[\mu_1])_{\alpha_1} (A_2[\mu_2])_{\alpha_1\alpha_2} \dots (A_L[\mu_L])_{\alpha_{L-1}}.$$

Hence the  $A_i$  can be viewed as tensors of order 3 (depending on three indices  $\alpha_{i-1}$ ,  $\mu_i$ ,  $\alpha_i$ ) in  $\mathbb{C}^{r_{i-1} \times 2 \times r_i}$ . The name 'bond dimensions' for the  $r_i$  has nothing to do with chemical bonds, but is related to the standard graphical representation of MPS in Figure 1, in which each contraction index  $\alpha_i$  is represented by a horizontal 'bond'. The minimal bond dimensions with which a given state can be represented have a well known meaning as ranks of matricizations of the coefficient tensor  $C$ , as recalled in Lemma 3. The set of tensor trains (TT) or matrix product states (MPS) with respect to the basis  $\{\varphi_i\}_{i=1}^L$  with bond dimensions  $r_i$  ( $i = 1, \dots, L-1$ ) is denoted by

$$\text{MPS}(L, \{r_i\}_i, \{\varphi_i\}_i) \subseteq \mathcal{F}_L. \quad (8)$$

For bond dimension One, i.e.  $r_i = 1$  for all  $i$ , the MPS set (8) reduces to the set of Slater determinants  $\Phi_{\mu_1, \dots, \mu_L}$  built from the basis functions. Representing arbitrary states in  $\mathcal{F}_L$  as MPS is possible, but requires bond dimensions  $2^{L/2}$ , i.e. bond dimensions growing exponentially with  $L$  [17]. (Here we have assumed that  $L$  is even.) A simple example where this exponential bound is saturated is the Slater determinant with orbitals  $\psi_i := (\varphi_i + \varphi_{i+L/2})/\sqrt{2}$  for  $i = 1, \dots, L/2$ , see e.g. [6, 9]. (In this example the bond dimension could be lowered to 2 by re-ordering the basis; an example where the exponential bound is saturated *regardless* of the ordering of the basis is given in [9].) Here we are interested in the best bond dimensions achievable by choosing the basis optimally, i.e. performing an optimal fermionic mode transformation.

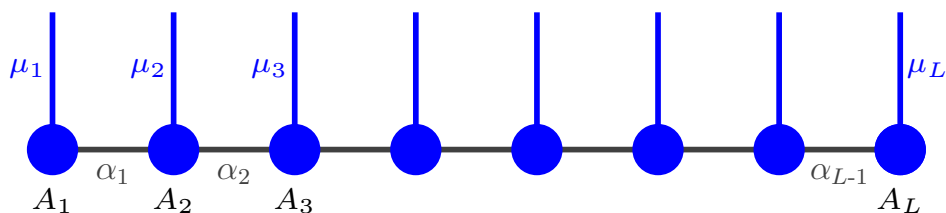


Figure 1: Graphical representation of a matrix product state in the finite-dimensional case; the virtual indices  $\alpha_j$  are contracted over.

## 4 Low-rank representation of two-electron wavefunctions and exactness of QC-DMRG with mode optimization

We now show that for representing two-electron wavefunctions in the MPS format, bond dimension Three always suffices independently of  $L$ , provided the basis of the single-particle space is chosen optimally.

In the following, the  $N$ -particle Hilbert space  $\mathcal{V}_{N,L} = \bigwedge_{i=1}^N \mathcal{H}_L$  will be identified with the  $N$ -particle sector

$$\{\Psi \in \mathcal{F}_L : \mathcal{N}\Psi = N\Psi\}$$

of Fock space, to which it is canonically isomorphic. Here  $\mathcal{N} = \sum_{i=1}^L a^\dagger(\varphi_i)a(\varphi_i)$  is the number operator, with  $a^\dagger(\varphi)$  and  $a(\varphi)$  denoting the usual creation and annihilation operators associated with an orbital  $\varphi \in \mathcal{H}_L$ . Also, we will make use of the single-particle reduced density matrix  $\gamma_\Psi : \mathcal{H}_L \rightarrow \mathcal{H}_L$  defined by

$$\langle \Psi, a^\dagger(\varphi_i)a(\varphi_j)\Psi \rangle = \langle \varphi_j, \gamma_\Psi \varphi_i \rangle \quad \text{for all } i, j.$$

Our precise result on two-particle states is as follows.

**Theorem 1** (Upper bound on the bond dimensions). *For any two-particle state  $\Psi \in \mathcal{V}_{2,L}$  with  $L \geq 4$ , there exists a basis  $\{\varphi_1, \dots, \varphi_L\}$  of the single-particle Hilbert space  $\mathcal{H}_L$  for which  $\Psi$  is an MPS with bond dimensions*

$$(r_1, \dots, r_{L-1}) = (2, \underbrace{2, 3, \dots, 2, 3, 2, 2}_{L-4 \text{ times}}). \quad (9)$$

If  $L \leq 3$ , we can simply achieve  $(r_1, \dots, r_{L-1}) = (1, \dots, 1)$ .

The somewhat counterintuitive looking bond dimension vector in (9) is in fact optimal for generic two-particle wavefunctions.

**Theorem 2** (Lower bound on the bond dimensions). *Suppose that  $L \geq 4$  is even,  $\Psi \in \mathcal{V}_{2,L}$ , and  $\gamma_\Psi$  has maximal rank (i.e., its rank equals  $L$ ). Then the bond dimensions given in Theorem 1 are optimal, that is to say for any basis  $\{\varphi_1, \dots, \varphi_L\}$  of the single-particle Hilbert space  $\mathcal{H}_L$  and any MPS-representation with bond dimensions  $(r_1, \dots, r_{L-1})$  we have*

- $r_j \geq 2$  for every  $j \in \{1, \dots, L-1\}$
- At least one of two consecutive elements  $(r_j, r_{j+1})$  for  $j \in \{2, \dots, L-2\}$  is at least 3.

Furthermore the bond dimension vector  $(r_1, \dots, r_{L-1})$  with lowest  $\ell^1$ -norm  $r_1 + \dots + r_{L-1}$  is unique and given by (9).

As will become clear in the proof of Theorem 1, the optimal representation is achieved for a basis consisting of natural orbitals, i.e. eigenstates of  $\gamma_\Psi$ .

These results have an important implication for the QC-DMRG method for computing the electronic structure of molecules.

## 4.1 QC-DMRG method

This method approximates, for a given  $N$ -electron system, a given self-adjoint and particle-number-conserving Hamiltonian  $H : \mathcal{F}_L \rightarrow \mathcal{F}_L$ , and a given  $L$ -dimensional one-particle Hilbert space  $\mathcal{H}_L$ , the ground and excited energy levels and eigenstates of the system as follows: for a given basis  $\{\varphi_1, \dots, \varphi_L\}$  of  $\mathcal{H}_L$ ,

$$E_0^{\text{QC-DMRG}}(\varphi_1, \dots, \varphi_L) = \min_{\substack{\Psi \in \text{MPS}(L, \{r_i\}_i, \{\varphi_i\}_i) \\ \Psi \neq 0, \mathcal{N}\Psi = N\Psi}} \frac{\langle \Psi, H\Psi \rangle}{\langle \Psi, \Psi \rangle} \quad (10)$$

and

$$E_j^{\text{QC-DMRG}}(\varphi_1, \dots, \varphi_L) = \min_{\substack{\Psi \in \text{MPS}(L, \{r_i\}_i, \{\varphi_i\}_i) \\ \Psi \neq 0, \mathcal{N}\Psi = N\Psi, \\ \langle \Psi_k^{\text{QC-DMRG}}, \Psi \rangle = 0 \forall k=0, \dots, j-1}} \frac{\langle \Psi, H\Psi \rangle}{\langle \Psi, \Psi \rangle} \quad (j \geq 1), \quad (11)$$

with the  $\Psi_j^{\text{QC-DMRG}}$  being corresponding optimizers. Our notation emphasizes that these quantities depend on the chosen single-particle basis. The exact (full configuration-interaction or FCI) eigenvalues  $E_j$  and eigenstates  $\Psi_j$  in the finite one-body basis are given, thanks to the Rayleigh-Ritz variational principle, by the analogous formulae with the MPS set  $\text{MPS}(L, \{r_i\}_i, \{\varphi_i\}_i)$  replaced by the full Fock space  $\mathcal{F}_L$ .

## 4.2 Mode transformations

In recent simulations [14] it has been found to be beneficial to also optimize over the underlying one-body basis, i.e. the ‘modes’  $\varphi_1, \dots, \varphi_L$ . Mathematically this corresponds to the following improved approximation to the eigenvalues and eigenstates:

$$E_0^{\text{QC-DMRG-MO}} = \min_{\substack{(\varphi_1, \dots, \varphi_L) \in \mathcal{H}_L \times \dots \times \mathcal{H}_L : \\ \langle \varphi_i, \varphi_j \rangle = \delta_{ij} \forall i, j}} E_0^{\text{QC-DMRG}}(\varphi_1, \dots, \varphi_L) \quad (12)$$

and

$$E_j^{\text{QC-DMRG-MO}} = \min_{\substack{(\varphi_1, \dots, \varphi_L) \in \mathcal{H}_L \times \dots \times \mathcal{H}_L : \\ \langle \varphi_i, \varphi_j \rangle = \delta_{ij} \forall i, j}} E_j^{\text{QC-DMRG}}(\varphi_1, \dots, \varphi_L) \quad (j \geq 1), \quad (13)$$

where the superscript MO stands for mode-optimized. Corresponding optimizers in (10), (11) with optimal  $\varphi_i$ ’s are denoted  $\Psi_0^{\text{QC-DMRG-MO}}$  respectively  $\Psi_j^{\text{QC-DMRG-MO}}$ . Note that such optimizers exist, since sets of normalized MPS states with given bond dimensions are closed [12, 3] and bounded, and hence compact.

Obviously, we have the inequalities

$$E_j \leq E_j^{\text{QC-DMRG-MO}} \leq E_j^{\text{QC-DMRG}}(\varphi_1, \dots, \varphi_N) \quad \forall j \geq 0.$$

Note also that for bond dimension 1, i.e.  $r_i = 1$  for all  $i$ , the QC-DMRG-MO ground state energy reduces precisely to the famous Hartree-Fock energy defined by

$$E_0^{HF} = \min_{\substack{(\varphi_1, \dots, \varphi_N) \in \mathcal{H}_L \times \dots \times \mathcal{H}_L \\ \langle \varphi_i, \varphi_j \rangle = \delta_{ij}}} \frac{\langle \varphi_1 \wedge \dots \wedge \varphi_N, H \varphi_1 \wedge \dots \wedge \varphi_N \rangle}{\langle \varphi_1 \wedge \dots \wedge \varphi_N, \varphi_1 \wedge \dots \wedge \varphi_N \rangle},$$

that is to say

$$E_0^{\text{QC-DMRG-MO}} \Big|_{r_1 = \dots = r_{L-1} = 1} = E_0^{HF}.$$

The following interesting result is an immediate consequence of Theorem 1.

**Theorem 3** (Low-rank exactness of QC-DMRG). *For  $N = 2$  electrons, any particle-number-conserving self-adjoint Hamiltonian  $H$ , and any finite-dimensional single-particle Hilbert space  $\mathcal{H}_L$ , the QC-DMRG method with fermionic mode optimization is exact for bond dimension Three. That is to say,*

$$E_j^{\text{QC-DMRG-MO}} \Big|_{r_1 = \dots = r_{L-1} = 3} = E_j, \quad \forall j \geq 0,$$

and any corresponding optimizers  $\Psi_j^{\text{QC-DMRG-MO}}$  are exact eigenstates.

### 4.3 Necessity of mode optimization

Mode optimization is essential for Theorems 1 and 3, as the following example demonstrates.

**Example 1.** *Consider an arbitrary fixed underlying basis  $\{\varphi_i\}_{i=1}^L$ . As recalled in Lemma 3 below, the minimal bond dimensions of a state  $\Psi$  correspond to the ranks of the unfolding of its coefficient tensor  $C$ . These take the form*

$$C_{\substack{\mu_1, \dots, \mu_k \\ \mu_{k+1}, \dots, \mu_L}}^{\mu_1, \dots, \mu_k} = \left( \begin{array}{c|c|c} & & v_{1,k} \\ \hline & D_k & \\ \hline v_{2,k} & & \end{array} \right), \quad (14)$$

where  $D_k \in \mathbb{C}^{k \times L-k}$  contains the coefficients corresponding to one  $\mu_i = 1$  in the  $k$  upper indices and one  $\mu_j = 1$  in the lower  $L - k$  indices, analogously for  $v_{1,k} \in \mathbb{C}^{1 \times \binom{k}{2}}$  and  $v_{2,k} \in \mathbb{C}^{\binom{L-k}{2} \times 1}$ .

Thus a generic state  $\Psi = \sum C_{\mu_1 \dots \mu_L} \Phi_{\mu_1 \dots \mu_L} \in \mathcal{V}_{2,L}$ , resulting e.g. from its coefficients being drawn independently from a continuous probability distribution – like a standard Gaussian – will have minimal bond dimensions

$$r_k = 2 + \min\{k, L - k\},$$

see [7].

In this example, the overall bond dimension necessary,  $\max_k r_k = 2 + \frac{L}{2}$ , grows with the number of orbitals  $L$ . Note also that the state  $\Psi$  above arises as the ground state of the parent Hamiltonian given by minus the orthogonal projector onto the state, that is,  $H = -|\Psi\rangle\langle\Psi|$ . Further, it follows from the results in [9] that there always exist states in  $\mathcal{V}_{2,L}$  for which the overall bond dimension  $2 + \frac{L}{2}$  cannot be reduced by re-ordering the basis.

#### 4.4 Upper bounds on the ranks

In this subsection we prove Theorem 1. We begin by recalling the following well known result [5].

**Lemma 1** (Two-particle wave-functions). *For any two-particle wavefunction  $\Psi \in \mathcal{V}_{2,L} = \mathcal{H}_L \wedge \mathcal{H}_L$  there exists a basis  $\{\varphi_i\}_{i=1}^L$  of  $\mathcal{H}_L$  and coefficients  $\lambda_i$  ( $i = 1, \dots, k$ ),  $k \leq L/2$ , such that*

$$\Psi = \sum_{\ell=1}^k \lambda_\ell |\varphi_{2\ell-1}, \varphi_{2\ell}\rangle, \quad (15)$$

*i.e. each basis function appears only in one Slater determinant.*

This can be proved by using the antisymmetry of  $\Psi$  to write it in the form

$$\Psi = \sum_{1 \leq i \neq j \leq L} c_{ij} |\varphi_i \varphi_j\rangle \quad (16)$$

with  $c_{ij} = -c_{ji}$ , and applying spectral theory to the coefficient matrix. We note that the orbitals appearing in (15) are automatically of  $\Psi$ , i.e. *natural orbitals* or *norbs*.

In the following we will always assume the basis to be chosen such that  $\Psi$  is of the form (15).

The coefficient tensor in the occupation representation then takes the following form

$$C_{\mu_1, \dots, \mu_L} = \sum_{\ell=1}^k \lambda_\ell \delta_{11}^\ell \prod_{\substack{i=1 \\ i \neq \ell}}^k \delta_{00}^i, \quad (17)$$

where we introduced the short-hand notation

$$\delta_{00}^n := \delta_0(\mu_{2n-1})\delta_0(\mu_{2n}), \quad \delta_{11}^n := \delta_1(\mu_{2n-1})\delta_1(\mu_{2n}).$$

Due to this special structure it makes sense to first seek a pair states decomposition, i.e. an MPS factorization of  $C_{\mu_1, \dots, \mu_L}$  into tensors  $B_\ell$  associated with pairs  $(\mu_{2\ell-1}, \mu_{2\ell})$  of occupation numbers, i.e.  $B_\ell \in \mathbb{C}^{r_{2\ell-2} \times 4 \times r_{2\ell}}$ .

With respect to pair states, (17) looks like a non-translation-invariant version of the W-state  $|\uparrow\downarrow\downarrow \dots \downarrow\rangle + |\downarrow\uparrow\downarrow \dots \downarrow\rangle + \dots + |\downarrow\downarrow\downarrow \dots \uparrow\rangle$  from spin physics, which is known to have bond dimension 2. The following lemma gives a corresponding low-bond-dimension factorization in the non-translation-invariant case.

**Lemma 2** (Matrix lemma). *For any two sequences  $(a_n)_{n \geq 2}$  and  $(b_n)_{n \geq 2}$  of complex numbers,*

$$\begin{pmatrix} a_2 & b_2 \\ & a_2 \end{pmatrix} \begin{pmatrix} a_3 & b_3 \\ & a_3 \end{pmatrix} \cdots \begin{pmatrix} a_{n-1} & b_{n-1} \\ & a_{n-1} \end{pmatrix} = \begin{pmatrix} \prod_{i=2}^{n-1} a_i & \sum_{i=2}^{n-1} b_i \prod_{j \neq i} a_j \\ & \prod_{i=2}^{n-1} a_i \end{pmatrix}. \quad (18)$$

*Proof.* We can write the left hand side of (18) as

$$\left[ a_2 \text{Id} + b_2 \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{=:S} \right] \left[ a_3 \text{Id} + b_3 \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \right] \cdots \left[ a_{n-1} \text{Id} + b_{n-1} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \right].$$

By the nilpotence of the matrix  $S$ , this becomes

$$\prod_{i=2}^{n-1} a_i \text{Id} + \sum_{i=2}^{n-1} b_i \prod_{j \neq i} a_j S,$$

which is our assertion. □

Consequently, letting

$$\begin{aligned} B_1[\mu_1, \mu_2] &:= \begin{pmatrix} \delta_{00}^1 & \lambda_1 \delta_{11}^1 \end{pmatrix}, \\ B_\ell[\mu_{2\ell-1}, \mu_{2\ell}] &:= \begin{pmatrix} \delta_{00}^{2\ell} & \lambda_\ell \delta_{11}^{2\ell} \\ & \delta_{00}^{2\ell} \end{pmatrix} \quad (1 < \ell < k), \\ B_k[\mu_{2k-1}, \mu_{2k}] &:= \begin{pmatrix} \lambda_k \delta_{11}^{2k} \\ & \delta_{00}^{2k} \end{pmatrix} \end{aligned} \quad (19)$$

we obtain the following MPS representation:

$$\begin{aligned} C_{\mu_1, \dots, \mu_L} &= B_1[\mu_1, \mu_2] \cdots B_k[\mu_{2k-1}, \mu_{2k}] \\ &= \begin{pmatrix} \delta_{00}^1 & \lambda_1 \delta_{11}^1 \end{pmatrix} \begin{pmatrix} \prod_{\ell=2}^{k-1} \delta_{00}^\ell & \sum_{\ell=2}^{k-1} \lambda_\ell \delta_{11}^\ell \prod_{\substack{i=2 \\ i \neq \ell}}^{k-1} \delta_{00}^i \\ & \prod_{\ell=2}^{k-1} \delta_{00}^\ell \end{pmatrix} \begin{pmatrix} \lambda_k \delta_{11}^k \\ \delta_{00}^k \end{pmatrix} \\ &= \begin{pmatrix} \prod_{\ell=1}^{k-1} \delta_{00}^\ell & \sum_{\ell=1}^{k-1} \lambda_\ell \delta_{11}^\ell \prod_{\substack{i=1 \\ i \neq \ell}}^{k-1} \delta_{00}^i \end{pmatrix} \begin{pmatrix} \lambda_k \delta_{11}^k \\ \delta_{00}^k \end{pmatrix} \\ &= \sum_{\ell=1}^k \lambda_\ell \delta_{11}^\ell \prod_{\substack{i=1 \\ i \neq \ell}}^k \delta_{00}^i. \end{aligned}$$

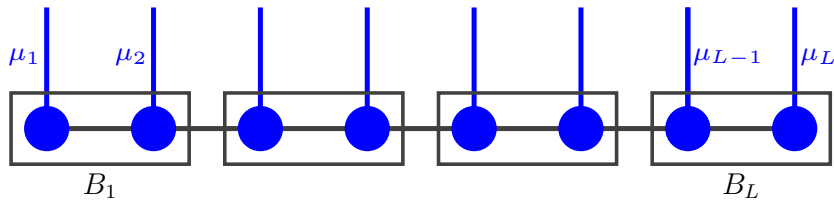


Figure 2: Graphical representation of the MPS decomposition associated with orbital pairs.

The last step consists now in passing from the pair states to the original states, i.e. decomposing the tensors  $B_\ell[\mu_{2\ell-1}, \mu_{2\ell}]$  into two tensors depending only on one of the  $\mu_i$ 's. This can be either guessed directly or obtained via reshaping and carrying out a singular value decomposition as in the derivation of the MPS representation of a general state (see e.g. [17]).

The result is

$$\begin{aligned}
 A_1[\mu_1] &:= \begin{pmatrix} \delta_0(\mu_1) & \delta_1(\mu_1) \end{pmatrix}, & A_2[\mu_2] &:= \begin{pmatrix} \delta_0(\mu_2) & \lambda_1 \delta_1(\mu_2) \end{pmatrix}, \\
 A_{2\ell-1}[\mu_{2\ell-1}] &:= \begin{pmatrix} \delta_0(\mu_{2\ell-1}) & \lambda_\ell \delta_1(\mu_{2\ell-1}) \\ & \delta_0(\mu_{2\ell-1}) \end{pmatrix}, & A_{2\ell}[\mu_{2\ell}] &:= \begin{pmatrix} \delta_0(\mu_{2\ell}) & \delta_1(\mu_{2\ell}) \\ & \delta_0(\mu_{2\ell}) \end{pmatrix} \quad (1 < \ell < k), \\
 A_{2k-1}[\mu_{2k-1}] &:= \begin{pmatrix} \lambda_k \delta_1(\mu_{2k-1}) & \\ & \delta_0(\mu_{2k-1}) \end{pmatrix}, & A_{2k}[\mu_{2k}] &:= \begin{pmatrix} \delta_1(\mu_{2k}) \\ \delta_0(\mu_{2k}) \end{pmatrix}.
 \end{aligned}$$

Therefore we have found an MPS representation for  $\Psi$  with bond dimensions  $r = (2, 2, 3, 2, \dots, 2, 3, 2, 2)$ . Note that in the case  $L = 4$ , i.e.  $k = 2$ , this reduces to  $r = (2, 2, 2)$ . Finally, in the case of just one Slater-determinant, i.e.  $L = 2$ , we can use  $a_1[\mu_1] := \lambda_1 \delta_1(\mu_1)$ ,  $a_2[\mu_2] := \delta_1(\mu_2)$ . This completes the proof of Theorem 1.

#### 4.5 Lower bounds on the ranks

In the previous subsection we saw that we can choose a basis such that all states can be represented with bond dimensions  $r = (2, 2, 3, \dots, 2, 3, 2, 2)$ . But as the  $\underbrace{(2, 2, 3, \dots, 2, 3, 2, 2)}_{L-4 \text{ times}}$  product of the matrices of two orbitals can be written as a  $2 \times 2$ -matrix (see the  $B_\ell$  above) one might wonder if the maximal bond dimension can be brought down to 2. This turns out not to be the case and the above size vector is optimal as stated in Theorem 2.

Our starting point to prove Theorem 2 is the following well known fact.



**Lemma 3** (TT-rank equals separation rank [12], [10]). *Let  $C \in \mathbb{C}^{n_1 \times \dots \times n_d}$  be an arbitrary tensor (representing the coefficients of a quantum state with respect to a fixed basis). For each bond between the  $i$ th and the  $i + 1$ st matrix, there exists a minimal  $r_i$  such that  $C$  admits a TT-decomposition with  $A_i$  of size  $n \times r_i$  and  $A_{i+1}$  of size  $r_i \times m$ , and this  $r_i$  is given by the Schmidt rank of the unfolding  $C_{\mu_{i+1} \dots \mu_L}^{\mu_1 \dots \mu_i}$ . Also, there exists a TT-decomposition with all  $r_i$  being simultaneously minimal.*

*Proof of Theorem 2.* We start with the case  $L = 4$  to convey the proof idea. Due to Lemma 3 it is enough to consider the unfoldings  $C_{\mu_1 \mu_2 \mu_3}^{\mu_1}$  and  $C_{\mu_3 \mu_4}^{\mu_1 \mu_2}$  of our tensor  $C_{\mu_1 \mu_2 \mu_3 \mu_4}$ . Note that here the coefficients  $c_{ij}$  in (16) are with respect to some arbitrary underlying basis  $\{\varphi_i\}_{i=1}^L$ . We begin with

$$C_{\mu_2, \mu_3, \mu_4}^{\mu_1} = \begin{matrix} & 110 & 101 & 101 & 100 & 010 & 001 \\ 0 & \left( \begin{array}{ccc|ccc} c_{23} & c_{24} & c_{34} & & & \\ & & & c_{12} & c_{13} & c_{14} \end{array} \right) \\ 1 & & & & & & \end{matrix}$$

The second row cannot vanish since then the state  $\varphi_1$  would not appear at all, meaning  $\gamma_\Psi$  has rank  $< 4$ , a contradiction.

And if the first row vanishes we can define  $\tilde{\varphi}_2 := \sum_{k=2}^4 c_{1,k} \varphi_k$  and thus get

$$\Psi = |\varphi_1, \tilde{\varphi}_2\rangle,$$

so again rank  $\gamma_\Psi < 4$ . So the matrix  $C_{\mu_2, \mu_3, \mu_4}^{\mu_1}$  must have rank 2.

Next let us consider

$$C_{\mu_3, \mu_4}^{\mu_1, \mu_2} = \begin{matrix} & & 00 & 10 & 01 & 11 \\ 00 & \left( \begin{array}{cc|cc} & & & c_{34} \\ & c_{13} & c_{14} & \\ 01 & c_{23} & c_{24} & \\ 11 & c_{12} & & \end{array} \right) \\ 10 & & & & & \\ 01 & & & & & \\ 11 & & & & & \end{matrix}$$

We want to show that rank  $C_{\mu_3, \mu_4}^{\mu_1, \mu_2} \geq 2$ .

If the submatrix in the middle vanishes, then both  $c_{12} \neq 0$  and  $c_{34} \neq 0$ , otherwise  $L \leq 2$ . But then rank  $C_{\mu_3, \mu_4}^{\mu_1, \mu_2} = 2$ . So we can assume that the submatrix in the middle does not vanish. If it has rank 2 we are already done. Thus assume that the rank equals 1 and  $c_{12} = c_{34} = 0$ . Then we know that there is a  $\lambda \in \mathbb{C}$  such that

$$\begin{pmatrix} c_{14} \\ c_{24} \end{pmatrix} = \lambda \begin{pmatrix} c_{13} \\ c_{23} \end{pmatrix}$$

but then we can write  $\Psi$  as

$$\Psi = c_{13} |\varphi_1, \varphi_3 + \lambda \varphi_4\rangle + c_{23} |\varphi_2, \varphi_3 + \lambda \varphi_4\rangle,$$

so  $L < 4$ . The unfolding  $C_{\mu_4}^{\mu_1, \mu_2, \mu_3}$  is dealt with in the same way as  $C_{\mu_2, \mu_3, \mu_4}^{\mu_1}$ . Therefore if rank  $\gamma_\Psi = L = 4$ , then the lowest possible rank vector is  $r = (2, 2, 2)$ .

Let us now turn to the general case  $L \geq 6$ . We start by noting that the first unfolding  $C_{\mu_2, \dots, \mu_L}^{\mu_1}$  and the last unfolding  $C_{\mu_L}^{\mu_1, \dots, \mu_{L-1}}$  both always have rank 2. The argument is exactly the same as in the  $L = 4$  case. Thus we already know  $r_1 = r_{L-1} = 2$ . Consider now the unfoldings

$$M_n := C_{\mu_{n+1}, \dots, \mu_L}^{\mu_1, \dots, \mu_n} = \begin{pmatrix} 0\dots0 & 10\dots0 & \dots & \dots & 0\dots01 & 110\dots0 & \dots & \dots & 0\dots011 \\ 0\dots0 & & & & & c_{n+1, n+2} & \dots & \dots & c_{L-1, L} \\ 10\dots0 & c_{1, n+1} & \dots & \dots & c_{1L} & & & & \\ \vdots & \vdots & & & \vdots & & & & \\ \vdots & \vdots & & & \vdots & & & & \\ 0\dots01 & & c_{n, n+1} & \dots & \dots & c_{nL} & & & \\ 110\dots0 & c_{12} & & & & & & & \\ \vdots & \vdots & & & & & & & \\ \vdots & \vdots & & & & & & & \\ 0\dots011 & c_{n-1, n} & & & & & & & \end{pmatrix}.$$

We start by proving that these matrices  $(M_n)_{n=2, \dots, L-2}$  always have rank  $\geq 2$ .

Assume the first row vanishes. If the submatrix corresponding to one  $\mu_i$  being 1 in the upper indices and one  $\mu_j$  being 1 in the lower indices – denoted by  $M_n^{11}$  – has rank  $\geq 2$ , there is nothing to show. So assume that  $\text{rank } M_n^{11} \leq 1$ . Then we can recombine the states  $\varphi_{n+1}, \dots, \varphi_L$  to see that  $\text{rank } \gamma_\Psi < L$ , as follows. Since  $\text{rank } M_n^{11} \leq 1$ , all columns are multiples of a single column, i.e. w.l.o.g.

$$\exists \alpha_j : \begin{pmatrix} c_{1j} \\ \vdots \\ c_{nj} \end{pmatrix} = \alpha_j \begin{pmatrix} c_{1, n+1} \\ \vdots \\ c_{n, n+1} \end{pmatrix} \quad \forall j \in \{n+1, \dots, L\}.$$

Then we can write  $\Psi$  as

$$\begin{aligned} \Psi &= \sum_{1 \leq r < s \leq n} c_{rs} |\varphi_r, \varphi_s\rangle + \sum_{\substack{1 \leq r \leq n \\ n+1 \leq s \leq L}} c_{rs} |\varphi_r, \varphi_s\rangle \\ &= \sum_{1 \leq r < s \leq n} c_{rs} |\varphi_r, \varphi_s\rangle + \sum_{1 \leq r \leq n} c_{r, n+1} |\varphi_r, \underbrace{\sum_{n+1 \leq s \leq L} \alpha_s \varphi_s}_{\tilde{\varphi}_{n+1}}\rangle. \end{aligned}$$

So  $\Psi$  can be represented with only at most  $n+1$  basis functions, i.e.  $\text{rank } \gamma_\Psi \leq n+1 < L$ , a contradiction.

In the same way, assuming that the first column vanishes and that  $\text{rank } M_n^{11} \leq 1$ , we obtain w.l.o.g.

$$\exists \beta_j : (c_{j, n+1}, \dots, c_{jL}) = \beta_j (c_{1, n+1}, \dots, c_{1L}) \quad \forall j \in \{1, \dots, n\}.$$

Then we can write  $\Psi$  as

$$\begin{aligned}\Psi &= \sum_{n+1 \leq r < s \leq L} c_{rs} |\varphi_r, \varphi_s\rangle + \sum_{\substack{1 \leq r \leq n \\ n+1 \leq s \leq L}} c_{rs} |\varphi_r, \varphi_s\rangle \\ &= \sum_{n+1 \leq r < s \leq L} c_{rs} |\varphi_r, \varphi_s\rangle + \sum_{n+1 \leq s \leq L} c_{1s} \underbrace{\left| \sum_{1 \leq r \leq n} \beta_r \varphi_r, \varphi_s \right\rangle}_{\tilde{\varphi}_n}.\end{aligned}$$

Consequently  $\Psi$  can be represented with at most  $L + 1 - n$  basis functions, i.e.  $\text{rank } \gamma_\Psi \leq L + 1 - n < L$ , so we again obtain a contradiction. Hence we have proven that if the first row or the first column of  $M_n$  vanish, the submatrix  $M_n^{11}$  has rank  $\geq 2$ . Since we have dealt with  $M_1$  and  $M_{L-1}$  separately, we thus have shown

$$\text{rank } M_n \geq 2 \quad \forall n \in \{1, \dots, L-1\}.$$

Our next steps now consists in considering two unfoldings at the same time and prove that at least one of them has rank  $\geq 3$ .

Therefore, consider for every  $\ell \in \{1, \dots, k-2\}$  the following matrices  $M_{2\ell}$  and  $M_{2\ell+1}$ , where  $M_i := C_{\mu_{i+1}, \dots, \mu_L}^{\mu_1, \dots, \mu_i}$  and  $L = 2k$ .

$$M_{2\ell} := \begin{array}{c} \begin{array}{c} 0\dots0 \\ 10\dots0 \\ \vdots \\ \vdots \\ 0\dots01 \\ 110\dots0 \\ \vdots \\ \vdots \\ 0\dots011 \end{array} \left( \begin{array}{cccccccc} 0\dots0 & 10\dots0 & \dots & \dots & 0\dots01 & 110\dots0 & \dots & \dots & 0\dots011 \\ & & & & & c_{2\ell+1, 2\ell+2} & \dots & \dots & c_{L-1, L} \\ & & & & & c_{1, 2\ell+1} & \dots & \dots & c_{1L} \\ & & & & & \vdots & & & \vdots \\ & & & & & \vdots & & & \vdots \\ & & & & & c_{2\ell, 2\ell+1} & \dots & \dots & c_{2\ell, L} \\ & & & & & c_{12} & & & \\ & & & & & \vdots & & & \\ & & & & & \vdots & & & \\ & & & & & c_{2\ell-1, 2\ell} & & & \end{array} \right) \end{array}$$



which yields the lowest possible rank vector

$$(r_1, \dots, r_{L-1}) = (2, 2, 3, \underbrace{\dots, 2, 3}_{L-4 \text{ times}}, 2, 2).$$

The proof of Theorem 1 is complete.  $\square$

## 5 Matrix product states – Infinite dimensions

We now deal with infinite-dimensional single-particle Hilbert spaces  $\mathcal{H}$ . As we will see, this calls for half-infinite matrix product states which we will introduce in a rigorous manner below. Graphically this corresponds to a half-infinite chain, see Figure 3.

So let  $\mathcal{H}$  be an infinite-dimensional separable Hilbert space spanned by orthonormal orbitals  $\{\varphi_i\}_{i=1}^\infty$ , let  $\mathcal{V}_N$  be the  $N$ -fold antisymmetric product  $\bigwedge_{i=1}^N \mathcal{H}$ , and let  $\mathcal{F}$  be the ensuing Fock space,

$$\mathcal{F} := \bigoplus_{N=0}^{\infty} \mathcal{V}_N.$$

Analogously to (7), we define a matrix product state (MPS) or tensor train (TT) with respect to the basis  $\{\varphi_i\}_{i=1}^\infty$  with size parameters ('bond dimensions')  $\{r_i\}_{i=1}^\infty$  to be a state of the form

$$\Psi = \lim_{L \rightarrow \infty} \sum_{\mu_1, \dots, \mu_L=0}^1 A_1[\mu_1] A_2[\mu_2] \dots A_L[\mu_L] \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \Phi_{\mu_1 \dots \mu_L} \in \mathcal{F}, \quad (20)$$

where the  $A_i[\mu_i]$  ( $i = 1, 2, \dots$ ) are  $r_{i-1} \times r_i$  matrices,  $r_0 = 1$ , the column vector above has length  $r_L$  (so as to make the coefficient of  $\Phi_{\mu_1 \dots \mu_L}$  scalar), and the  $A_i$  are such that the above limit exists as a strong limit in the Fock space  $\mathcal{F}$ . The key point about the representation (20) is that the  $A_i$  are *fixed* matrices which only depend on the *exact* infinite-dimensional quantum state  $\Psi$  and encode its true entanglement structure, whereas first truncating the one-body Hilbert space to dimension  $L$  and then MPS-factorizing the ensuing approximation to  $\Psi$  would lead to  $L$ -dependent  $A_i$ 's.

The vector  $(0, \dots, 0, 1)$  appearing in (20) may look arbitrary at first, but as we show in a companion paper [8] every normalized state  $\Psi$  in the Fock space  $\mathcal{F}$  can be represented in the form (20) with left-normalized  $A_i$  (i.e.  $\sum_{\mu_i} A_i(\mu_i)^\dagger A_i(\mu_i) = I$ ) if the  $r_i$  are allowed to grow exponentially (i.e.  $r_i = 2^i$ ).

The set of tensor trains (TT) or matrix product states (MPS) with respect to the basis  $\{\varphi_i\}_{i=1}^\infty$  with bond dimensions  $\{r_i\}_{i=1}^\infty$  is denoted by

$$\text{MPS}(\infty, \{r_i\}_i, \{\varphi_i\}_i) \subseteq \mathcal{F}. \quad (21)$$

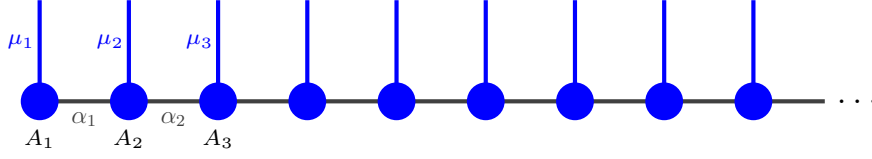


Figure 3: Graphical representation of a matrix product state in the infinite-dimensional case.

## 6 Two-particle systems – Infinite dimensions

We now extend our results from section 4 to infinite dimensions.

**Theorem 4.** *Let  $\mathcal{H}$  be an infinite-dimensional separable Hilbert space. For any two-particle state  $\Psi \in \mathcal{H} \wedge \mathcal{H}$ , there exists an orthonormal basis  $\{\varphi_i\}_{i=1}^{\infty}$  of the single-particle Hilbert space  $\mathcal{H}$  for which  $\Psi$  is an MPS with bond dimensions  $r_i = 2$  for  $i$  even,  $r_i = 3$  for  $i$  odd and  $> 1$ , and  $r_1 = 2$ . In particular, there is an MPS representation with maximal bond dimension 3.*

*Moreover the value 3 is minimal, that is, not all two-particle states can be represented by an MPS with bond dimension 2.*

*Proof.* Let  $\Psi$  be in  $\mathcal{H} \wedge \mathcal{H}$ , and let  $\{\varphi_i\}_{i=1}^{\infty}$  be an ONB of  $\mathcal{H}$  consisting of eigenstates of  $\gamma_{\Psi}$  (i.e., of natural orbitals), ordered by size of the eigenvalue of  $\gamma_{\Psi}$ . After a unitary transformation in each eigenspace,  $\Psi$  has the normal form

$$\Psi = \sum_{\ell=1}^{\infty} \lambda_{\ell} |\varphi_{2\ell-1} \varphi_{2\ell}\rangle,$$

with  $\sum_{\ell=1}^{\infty} |\lambda_{\ell}|^2 = \|\Psi\|^2 < \infty$ . For  $L$  even, define

$$\Psi_L = \sum_{\ell=1}^L \lambda_{\ell} |\varphi_{2\ell-1} \varphi_{2\ell}\rangle.$$

Applying now our analysis from Section 4 gives that  $\Psi_L$  has a representation of the form (8) with  $L$ -dependent tensors  $A_1^{(L)}[\mu_1], \dots, A_L^{(L)}[\mu_L]$ .

By inspection these  $A_i^{(L)}$  only depend on the coefficients  $\lambda_{\ell}$  up to  $\lceil \frac{i}{2} \rceil$ , so they are independent of  $L$  for  $L \gg i$ ; thus denote these by  $A_i$ . As in Section 4, let  $B_{\ell}[\mu_{2\ell-1}, \mu_{2\ell}] = A_{2k-1}[\mu_{2\ell-1}] A_{2\ell}[\mu_{2\ell}]$ , whence  $B_{\ell}$  is given by eq. (19). Now let us compute the expression inside the limit in eq. (20). When  $L$  is even, that is,  $L = 2k$  for some integer  $k$ , we have

$$B_k[\mu_{2k-1}, \mu_{2k}] \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda_k \delta_{11}^{2k}(\mu_{2k-1}, \mu_{2k}) \\ \delta_{00}^{2k}(\mu_{2k-1}, \mu_{2k}) \end{pmatrix}$$

and therefore

$$\left(\prod_{\ell=1}^k B_{\ell}[\mu_{2\ell-1}, \mu_{2\ell}]\right) \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \sum_{\ell=1}^k \lambda_{\ell} \delta_{11}^{\ell} \prod_{\substack{i=1, \\ i \neq \ell}}^k \delta_{00}^i.$$

It follows that

$$\sum_{\mu_1, \dots, \mu_L=0}^1 A_1[\mu_1] A_2[\mu_2] \dots A_L[\mu_L] \begin{pmatrix} 0 \\ 1 \end{pmatrix} \Phi_{\mu_1 \dots \mu_L} = \Psi_{2k}. \quad (22)$$

For  $L$  odd, that is,  $L = 2k + 1$  for integer  $k$ , one finds analogously that

$$\left(\prod_{i=1}^{2k+1} A_i\right) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \left(\prod_{i=1}^k B_i\right) A_{2k+1} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \sum_{\ell=1}^k \lambda_{\ell} \delta_{11}^{\ell} \prod_{\substack{i=1, \\ i \neq \ell}}^k \delta_{00}^i \delta_0(\mu_{2k+1})$$

and therefore the left hand side in eq. (22) is again given by  $\Psi_{2k}$ . Since  $\Psi_L$  converges by construction to  $\Psi$ , we obtain

$$\Psi = \lim_{L \rightarrow \infty} \Psi_L = \lim_{L \rightarrow \infty} \sum_{\mu_1, \dots, \mu_L=0}^1 A_1[\mu_1] A_2[\mu_2] \dots A_L[\mu_L] \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \Phi_{\mu_1 \dots \mu_L}.$$

Thus  $\Psi$  has an MPS representation with the asserted bond dimensions with respect to our natural orbital basis.

The fact that  $\Psi$  does not in general belong to the set of MPS with bond dimension 2 regardless of the choice of basis follows directly from Theorem 2.  $\square$

As a corollary, the exactness of the QC-DMRG method combined with fermionic mode transformations for two-electron systems (Theorem 3) generalizes in a straightforward manner to the full infinite-dimensional single-particle Hilbert space  $\mathcal{H} = L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$  for electrons. The resulting versions of (10)–(13) constitute an exact reformulation of (time-independent) two-electron quantum mechanics.

## 7 Conclusions and Outlook

We have shown that the QC-DMRG method combined with fermionic mode optimization is exact for two-electron systems with the (extremely low) bond dimension  $M = 3$ . This can be viewed as a theoretical contribution towards explaining the remarkable success of the QC-DMRG method in practical computations, and as a theoretical argument in favour of including mode optimization. The numerical favourability of the latter was emphasized in [14].

An interesting theoretical question beyond the scope of the present paper is whether any analogs of our findings hold for larger particle numbers provided the Hamiltonian is of two-body form. In the two-electron case investigated here, this form was satisfied automatically; in electronic structure it continues to be satisfied for arbitrary particle numbers.

## Acknowledgements

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 188264188/GRK1754 within the International Research Training Group IGDK 1754 is gratefully acknowledged.

## Data availability statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## References

- [1] I. Affleck, T. Kennedy, E. H. Lieb, and H. Tasaki. Rigorous results on valence-bond ground states in antiferromagnets. *Phys. Rev. Lett.*, 59:799–802, Aug 1987.
- [2] G. Barcza, O. Legeza, K. H. Marti, and M. Reiher. Quantum-information analysis of electronic states of different molecular structures. *Phys. Rev. A*, 83:012508, Jan 2011.
- [3] T. Barthel, J. Lu, and G. Friesecke. On the closedness and geometry of tensor network state sets. *arXiv preprint arXiv:2108.00031*, 2021.
- [4] G. K.-L. Chan and M. Head-Gordon. Highly correlated calculations with a polynomial cost algorithm: A study of the density matrix renormalization group. *J. Chem. Phys.*, 116(11):4462–4476, 2002.
- [5] A. J. Coleman and V. I. Yukalov. *Reduced density matrices: Coulson’s challenge*, Lecture Notes in Chemistry volume 72. Springer Science & Business Media, 2000.
- [6] M.-S. Dupuy and G. Friesecke. Inversion symmetry of singular values and a new orbital ordering method in tensor train approximations for quantum chemistry. *SIAM J. Sci. Comput.*, 43(1):B108–B131, 2021.
- [7] X. Feng and Z. Zhang. The rank of a random matrix. *Appl. Math. Comput.*, 185(1):689–694, 2007.



- [8] G. Friesecke and B. R. Graswald. *In preperation.*
- [9] B. R. Graswald and G. Friesecke. Electronic wavefunction with maximally entangled MPS representation. *Eur. Phys. J. D*, 75(6):1–4, 2021.
- [10] W. Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42. Springer, 2012.
- [11] T. Helgaker, P. Jørgensen, and J. Olsen. *Configuration-Interaction Theory*, chapter 11 of *Molecular Electronic-Structure Theory*, pages 523–597. John Wiley & Sons, Ltd, 2000.
- [12] S. Holtz, T. Rohwedder, and R. Schneider. On manifolds of tensors of fixed TT-rank. *Numer. Math.*, 120(4):701–731, 2012.
- [13] C. Krumnow, L. Veis, J. Eisert, and O. Legeza. Effective dimension reduction with mode transformations: Simulating two-dimensional fermionic condensed matter systems with matrix-product states. *Phys. Rev. B*, 104:075137, 2021.
- [14] C. Krumnow, L. Veis, O. Legeza, and J. Eisert. Fermionic orbital optimization in tensor network states. *Phys. Rev. Lett.*, 117:210402, Nov 2016.
- [15] O. Legeza, J. Röder, and B. Hess. Controlling the accuracy of the density-matrix renormalization-group method: The dynamical block state selection approach. *Phys. Rev. B*, 67(12):125114, 2003.
- [16] A. O. Mitrushenkov, G. Fano, F. Ortolani, R. Linguerri, and P. Palmieri. Quantum chemistry using the density matrix renormalization group. *J. Chem. Phys.*, 115(15):6815–6821, 2001.
- [17] U. Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Ann. Physics*, 326(1):96–192, 2011.
- [18] S. Szalay, M. Pfeffer, V. Murg, G. Barcza, F. Verstraete, R. Schneider, and Ö. Legeza. Tensor product methods and entanglement optimization for ab initio quantum chemistry. *Int. J. Quantum Chem.*, 115(19):1342–1391, 2015.
- [19] S. R. White and R. L. Martin. Ab initio quantum chemistry using the density matrix renormalization group. *J. Chem. Phys.*, 110(9):4127–4130, 1999.

FURTHER ARTICLES UNDER REVIEW

## Appendix C

### Articles as Co-author

#### C.1 Necessary Criteria for Markovian Divisibility of Linear Maps

# Necessary Criteria for Markovian Divisibility of Linear Maps

Matthias C. Caro and Benedikt R. Graswald

---

In 1976 Lindblad as well as Gorini, Kossakowski, and Sudarshan characterized the generators which give rise to semigroups of quantum channels via the corresponding (time-independent) master equation. This constituted an important step towards understanding the connection between master equations and the framework of quantum channels for describing quantum evolutions. In this paper, we consider the converse question, i.e., the open problem of characterizing those quantum channels that can arise from the solution of a (possibly time-dependent) Lindblad master equation. Endeavours towards a resolution of this problem have given rise to different notions of Markovianity for quantum evolutions. We concentrate on the definition which is based on connecting Markovianity to certain divisibility properties of quantum evolutions, in particular to the possibility of dividing the evolution into infinitesimal pieces.

While this gives an intuitively plausible notion of time-dependent quantum Markovianity and some structural properties can be established on its basis, it has so far not given rise to easily verifiable criteria for Markovianity. Only the trivial necessary criterion on non-negativity of the determinant was known. In contrast to higher dimensions, in the qubit case, this notion is well understood and completely characterized, which we recall in Section III.

In the main part of this work, Section IV, we go beyond this characterization for the 2-dimensional case. We start off in Part A with describing the setup of (infinitesimal) Markovian divisibility of general linear maps w.r.t. a closed and convex set of generators. Next, we provide a general proof strategy which can be applied to obtain necessary criteria for (infinitesimal) Markovian divisibility, if the generators satisfy certain spectral properties. In Part B we establish that for the specific case of quantum channels, where the set of generators are the Lindblad generators, these spectral properties are fulfilled.

Thus, we obtain necessary criteria for a quantum channel to be divisible into infinitesimal Markovian pieces, which take the form of an upper bound on the determinant in terms of a  $\Theta(d)$ -power of the smallest singular value, and in terms of a product of  $\Theta(d)$  smallest singular values.

In Part C of Section IV we also discuss the classical counterpart of this scenario, i.e., stochastic matrices with the generators given by transition rate matrices. Here, we show that no necessary criteria for infinitesimal Markovian divisibility of the form proved for quantum channels can hold in general.

The project's idea was motivated by discussions between Matthias C. Caro and myself. I want to emphasize again that Matthias C. Caro is the main author of this contribution. He wrote the majority of the text for the first draft and it was his idea to use the (sub-)multiplicativity properties of the determinant and of products of largest singular values as well as Trotterization to reduce the problem to a question about spectral property of the generators. Furthermore, he had the idea to consider general sets of generators, in particular proving Lemma IV.1 and Lemma IV.4, in addition to Theorem IV.5 and Corollary IV.6.

I proved Proposition IV.13. as well as Lemma A.1. in the appendix. Additionally, I was involved in all aspects of this work, with exception of the parts mentioned above.

# Permission to include:

Matthias C. Caro and Benedikt R. Graswald (2020).  
Necessary Criteria for Markovian Divisibility of Linear Maps.  
*Journal of Mathematical Physics* 62, 042203 (2021).  
<https://doi.org/10.1063/5.0031760>

# Permission to Reuse Content

## REUSING AIP PUBLISHING CONTENT

Permission from AIP Publishing is required to:

- republish content (e.g., excerpts, figures, tables) if you are not the author
- modify, adapt, or redraw materials for another publication
- systematically reproduce content
- store or distribute content electronically
- copy content for promotional purposes

To request permission to reuse AIP Publishing content, use RightsLink® for the fastest response or contact AIP Publishing directly at [rights@aip.org](mailto:rights@aip.org) (<mailto:rights@aip.org>) and we will respond within one week:

For RightsLink, use Scitation to access the article you wish to license, and click on the Reprints and Permissions link under the TOOLS tab. (For assistance click the “Help” button in the top right corner of the RightsLink page.)

To send a permission request to [rights@aip.org](mailto:rights@aip.org) (<mailto:rights@aip.org>), please include the following:

- Citation information for the article containing the material you wish to reuse
- A description of the material you wish to reuse, including figure and/or table numbers
- The title, authors, name of the publisher, and expected publication date of the new work
- The format(s) the new work will appear in (e.g., print, electronic, CD-ROM)
- How the new work will be distributed and whether it will be offered for sale

Authors do **not** need permission from AIP Publishing to:

- quote from a publication (please include the material in quotation marks and provide the customary acknowledgment of the source)
- reuse any materials that are licensed under a Creative Commons CC BY license (please format your credit line: “Author names, Journal Titles, Vol.#, Article ID#, Year of Publication; licensed under a Creative Commons Attribution (CC BY) license.”)
- reuse your own AIP Publishing article in your thesis or dissertation (please format your credit line: “Reproduced from [FULL CITATION], with the permission of AIP Publishing”)
- make multiple copies of articles—although you must contact the Copyright Clearance Center (CCC) at [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) to do this

## REUSING CONTENT PUBLISHED BY OTHERS

To request another publisher’s permission to reuse material in AIP Publishing articles, please use our

# Necessary criteria for Markovian divisibility of linear maps

Cite as: J. Math. Phys. 62, 042203 (2021); doi: 10.1063/5.0031760

Submitted: 3 October 2020 • Accepted: 24 March 2021 •

Published Online: 13 April 2021



Matthias C. Caro<sup>1,2,a)</sup>  and Benedikt R. Graswald<sup>1,b)</sup> 

## AFFILIATIONS

<sup>1</sup>Technical University of Munich, Department of Mathematics, Boltzmannstraße 3, 85748 Garching bei München, Germany

<sup>2</sup>Munich Center for Quantum Science and Technology (MCQST), Munich, Germany

<sup>a)</sup> Author to whom correspondence should be addressed: [caro@ma.tum.de](mailto:caro@ma.tum.de). URL: <https://sites.google.com/view/matthiasccaro>

<sup>b)</sup> [graswabe@ma.tum.de](mailto:graswabe@ma.tum.de). URL: <https://www-m7.ma.tum.de/bin/view/Analysis/BenediktGraswald>

## ABSTRACT

We describe how to extend the notion of infinitesimal Markovian divisibility from quantum channels to general linear maps and compact and convex sets of generators. We give a general approach toward proving necessary criteria for (infinitesimal) Markovian divisibility. With it, we prove two necessary criteria for infinitesimal divisibility of quantum channels in any finite dimension  $d$ : an upper bound on the determinant in terms of a  $\Theta(d)$ -power of the smallest singular value and in terms of a product of  $\Theta(d)$  smallest singular values. These allow us to analytically construct, in any given dimension, a set of channels that contains provably non-infinitesimal Markovian divisible ones. Moreover, we show that, in general, no such non-trivial criteria can be derived for the classical counterpart of this scenario.

Published under license by AIP Publishing. <https://doi.org/10.1063/5.0031760>

## I. INTRODUCTION

References 1 and 2 made an important step toward understanding the connection between master equations and the framework of quantum channels for describing quantum evolutions by characterizing the generators, which give rise to semigroups of quantum channels via the corresponding (time-independent) master equation. The converse question, i.e., the problem of characterizing those quantum channels that can arise from the solution of a (possibly time-dependent) Lindblad master equation, is, however, still awaiting an answer.

Endeavors toward a resolution of this problem have given rise to different notions of (non-) Markovianity for quantum evolutions. One line of research is based on connecting Markovianity to certain divisibility properties of quantum evolutions, particularly to the possibility of dividing the evolution into infinitesimal pieces. While this gives an intuitively plausible notion of time-dependent quantum Markovianity and some structural properties can be established on its basis, it has so far not given rise to easily verifiable criteria for Markovianity (with a simple exception). Only for evolutions of qubit systems is this notion completely understood. We go beyond this characterization for the two-dimensional case and establish necessary criteria for a quantum channel—or a linear map in general—to be divisible into infinitesimal Markovian pieces. Our criteria take the form of an upper bound on the determinant in terms of the power of a product of smallest singular values.

Our proof strategy is not specific to quantum channels but can be applied to obtain necessary criteria for (infinitesimal) Markovian divisibility of general linear maps with respect to a closed and convex set of generators if the generators satisfy certain spectral properties.

### A. Overview of our results

In this work, we study the following question: Given a linear map  $T$  and a set of linear maps  $\mathcal{G}$ , acting on  $\mathbb{C}^d$ , can  $T$  be approximated arbitrarily well by linear maps of the form  $\prod_i e^{G_i}$ , where  $G_i \in \mathcal{G}$ ? If that is the case, we say that  $T$  is *Markovian divisible with respect to the set of generators*  $\mathcal{G}$ .

We aim toward establishing necessary criteria for Markovian divisibility of the form

$$|\det(T)| \leq \left( \prod_{i=1}^k s_i^\uparrow(T) \right)^p,$$

where  $k = k(d)$  and  $p = p(d)$  depend on the underlying dimension. Proving such criteria becomes tractable by combining multiplicativity of the determinant and sub-/super-multiplicativity of products of largest/smallest singular values with Trotterization.

In Sec. IV A, we describe how to use these properties to reduce the problem of establishing necessary criteria of the above form to a spectral property of the generators. We can summarize our reduction as follows:

**Theorem** (Theorem IV.5—informal version). *Let  $\mathcal{G} \subseteq \mathcal{M}_d$  be a set of generators. Let  $T$  be Markovian divisible with respect to  $\mathcal{G}$ , and suppose that every  $G \in \mathcal{G}$  satisfies  $\text{Tr}[G + G^*] - p \sum_{i=1}^k \lambda_i^\uparrow(G + G^*) \leq 0$ . Then,  $|\det(T)| \leq \left( \prod_{i=1}^k s_i^\uparrow(T) \right)^p$ .*

We employ our proof strategy for the physically motivated scenario of *infinitesimal Markovian divisibility*. Here, the objects of interest are linear maps  $T$  that, for any  $\varepsilon > 0$ , can be arbitrarily well approximated by linear maps of the form  $\prod_i e^{G_i}$ , where  $G_i \in \mathcal{G}$  are such that  $\|e^{G_i} - \mathbb{1}_d\| \leq \varepsilon$ .

We first study the case in which  $\mathcal{G}$  is the set of Lindblad generators seen as linear maps on  $d \times d$ -matrices, i.e., we consider those generators that give rise to semigroups of quantum channels. With this choice, the notion of infinitesimal Markovian divisibility of a linear map  $T$  on  $d \times d$ -matrices becomes that of infinitesimal Markovian divisibility of quantum channels introduced in Ref. 3.

We prove necessary criteria for infinitesimal Markovian divisibility of quantum channels in any finite dimension. Specifically, for an infinitesimal Markovian divisible quantum channel  $T$  on  $d \times d$ -matrices, we show in Corollaries IV.9 and IV.16 that

$$|\det(T)| \leq \left( s_1^\uparrow(T) \right)^{\frac{d}{2}} \text{ and } |\det(T)| \leq \prod_{i=1}^{\lfloor 2d-2\sqrt{2d+1} \rfloor} s_i^\uparrow(T).$$

Moreover, we give explicit examples (Examples IV.12 and IV.17) of infinitesimal divisible channels from which we can conclude that the  $d$ -dependence of the exponent (in the first bound) and of the number of singular value factors (in the second bound) is close to optimal, respectively.

We also describe how to interpolate between these bounds in Corollary IV.21 and obtain that for an infinitesimal divisible quantum channel  $T$  acting on  $d \times d$ -matrices,

$$|\det(T)| \leq \left( \prod_{i=1}^k s_i^\uparrow(T) \right)^{\frac{2d}{k+2\sqrt{k+1}}} \text{ for } 1 \leq k \leq d^2.$$

These criteria allow us to give new examples of provably non-infinitesimal divisible channels in dimensions strictly bigger than 2, which were not recognizable as such previously (Example IV.11).

As a second application of our proof strategy, we take  $\mathcal{G}$  to be the set of transition rate matrices of dimension  $d$  and thereby study the question of (infinitesimal) Markovian divisibility of stochastic matrices. We first show via an explicit example (Example IV.24) that no necessary criterion of the above form can hold in this scenario when we allow all transition rate matrices as generators. Combined with our results for infinitesimal Markovian divisible quantum channels, this implies that stochastic matrices cannot be embedded into quantum channels while preserving both the singular values and the property of infinitesimal Markovian divisibility at the same time.

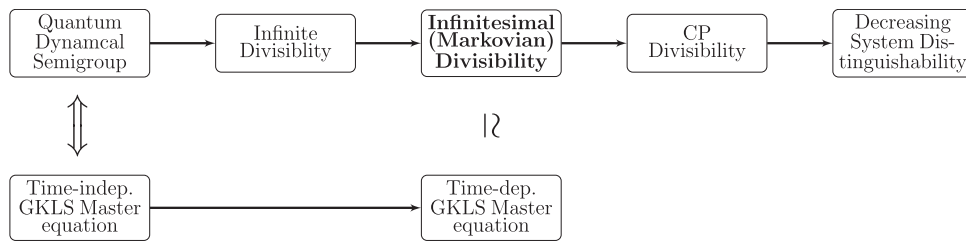
If, however, we restrict our set of generators to transition rate matrices whose diagonal elements differ by at most a constant factor, our proof strategy can be applied and yields an upper bound on the determinant in terms of a power of the smallest singular value (Corollary IV.27).

## B. Related work

The quantum Markovianity problem, the question of deciding whether a given quantum channel is a member of a quantum dynamical semigroup, was considered from a complexity-theoretic perspective in Ref. 4. Therein, it was shown to be NP-hard and the same is true for the classical counterpart of this problem, with stochastic matrices instead of quantum channels and transition rate matrices instead of Lindblad generators. The computational complexity of a related divisibility problem for stochastic matrices, namely, that of finite divisibility, was studied in Ref. 5. In addition, this divisibility problem turns out to be NP-hard, even NP-complete.

When fixing the system dimension, however, deciding whether a quantum channel is an exponential of a Lindblad generator, in which case it can be called time-independent Markovian because it solves a time-independent Lindblad master equation, becomes feasible. Corresponding necessary and sufficient criteria and an efficient (in the desired precision) algorithmic procedure for this case with a fixed dimension were given in Refs. 4 and 6. These results pertain to time-independent (quantum) Markovianity and cannot be directly applied to the time-dependent case.





**FIG. 1.** A depiction of the relations between different notions of divisibility and Markovianity of quantum channels and quantum dynamical maps. A simple arrow indicates that a channel or dynamical map satisfying the condition at the tail also satisfies that at the head.  $\updownarrow$  indicates the equivalence of two notions.  $\approx$  is used to indicate a correspondence that, to the best of our knowledge, has been rigorously proven only for the qubit case.

Our focus is on infinitesimal Markovian divisible quantum channels. These were introduced and studied in detail for qubit channels by Ref. 3. Therein, it is also observed that every infinitely divisible quantum channel, i.e., every channel that can be written as an  $n$ th power of a quantum channel for every  $n \in \mathbb{N}$ , is infinitesimal divisible. The notion of infinitesimal Markovian divisibility can be seen as corresponding to time-dependent Markovianity, i.e., to solutions of time-dependent Lindblad master equations. Thereby, it offers a route to studying a time-dependent version of the Markovianity problem.

A plethora of different notions of Markovianity for quantum evolutions and relations between them are discussed in several review papers.<sup>7–10</sup> On the one hand, one considers notions of quantum Markovianity based on the divisibility of the evolution, either for quantum channels or for quantum dynamical maps with corresponding propagators. This line of research was initiated by Ref. 3, and Refs. 11 and 12 constitute recent additions to it. In relation to this approach, Ref. 13 proposed a measure of non-Markovianity on the basis of infinitesimal deviations from complete positivity. On the other hand, there are notions and measures of non-Markovianity based on (quantum) information backflow, often formalized in terms of distinguishability measures that are known to be non-increasing under completely positive and trace-preserving maps. This idea was introduced in Ref. 14, and Ref. 9 recently proposed a variant of it.

In Fig. 1, we present only a selected few of these notions and the connections between them.

### C. Structure of the paper

Section II introduces basic notions from quantum information that provide our overall framework. In Sec. III, we introduce the core definition of infinitesimal Markovian divisibility in a general setting and discuss prior work in the quantum scenario. Section IV contains our main results: We describe the general proof approach in Subsection IV A and apply it to derive necessary criteria for infinitesimal Markovian divisibility of quantum channels in Subsection IV B. The same type of criterion does not, in general, hold for infinitesimal divisibility of stochastic matrices, only for suitable subsets, as we argue in Subsection IV C. We conclude with some open questions and the references.

## II. PRELIMINARIES

We introduce some of the basic notions of quantum information with focus on quantum channels and the corresponding semigroups. The interested reader is referred to Ref. 15 for more details.

Throughout this paper, we denote the set of  $d \times d$  complex matrices as  $\mathcal{M}_d$  for a dimension  $d \in \mathbb{N}$ . The identity matrix in  $\mathcal{M}_d$  is written as  $\mathbb{1}_d$ , whereas  $\text{id} = \text{id}_{\mathcal{M}_d}$  denotes the identity map on  $\mathcal{M}_d$ . For  $A \in \mathcal{M}_d$ , we use  $\lambda_i = \lambda_i(A)$  to denote its eigenvalues. If  $A \in \mathcal{M}_d$  is Hermitian, we use  $\lambda_i^\downarrow$  ( $\lambda_i^\uparrow$ ) to denote the eigenvalues in decreasing (increasing) order. Similarly, we use the notation  $s_i^\downarrow$  and  $s_i^\uparrow$  for singular values. Finally,  $\text{Tr}[A]$  will denote the trace of  $A$ .

### A. Quantum states and channels

A  $d$ -level quantum system (for  $d \in \mathbb{N}$ ) is described by a  $d \times d$  density matrix, i.e., an element of

$$\mathcal{S}(\mathbb{C}^d) := \{\rho \in \mathcal{M}_d \mid \rho \geq 0, \text{Tr}[\rho] = 1\},$$

where  $\rho \geq 0$  means that the matrix  $\rho$  is positive semidefinite.

Physically admissible transformations of quantum systems are described by *quantum channels* (in the Schrödinger picture), i.e., by elements of

$$\mathcal{T}(\mathbb{C}^d, \mathbb{C}^{d'}) := \{T : \mathcal{M}_d \rightarrow \mathcal{M}_{d'} \mid T \text{ is linear, completely positive, and trace - preserving}\}.$$

Here, we call  $T$  *completely positive* iff  $T \otimes \text{id}_{\mathcal{M}_n}$  is positivity-preserving for every  $n \in \mathbb{N}$ . This definition guarantees that a quantum channel maps states to states and that this is still the case when embedding the quantum system of interest into a larger system with trivial evolution on the environmental subsystem.

We will also use the shorthand  $\mathcal{T}_d := \mathcal{T}(\mathbb{C}^d, \mathbb{C}^d)$  for channels with equal input and output dimension.

### B. Quantum dynamical semigroups

It is a foundational postulate in quantum theory that the dynamics of a closed quantum system can be described in terms of a Schrödinger equation, which gives rise to a one-parameter group of unitaries. For open quantum systems, we will work with one-parameter semigroups.

*Definition II.1 (Continuous dynamical semigroups).* A family of linear maps  $T_t : \mathcal{M}_d \rightarrow \mathcal{M}_d$  with time parameter  $t \in \mathbb{R}_+$  is called a dynamical semigroup if  $\forall t, s \in \mathbb{R}_+ = [0, \infty) : T_t T_s = T_{t+s}$  and  $T_0 = Id$ . If in addition, the map  $t \mapsto T_t$  is continuous (we are working on finite dimensional spaces, so there is no need to specify the type of continuity here), then the family is called a continuous dynamical semigroup.

It is well-known that such continuous dynamical semigroups can be represented via a generator, i.e., if  $\{T_t\}_{t \geq 0}$  is a continuous dynamical semigroup, then there exists a linear map  $L : \mathcal{M}_d \rightarrow \mathcal{M}_d$  such that  $T_t = e^{tL}$  for all  $t \geq 0$ .

When requiring such a semigroup to consist of physically admissible evolutions of a quantum system, i.e., of quantum channels, the question arises of what the corresponding generators are. This was answered in the following.

**Theorem II.2 (Generators of quantum dynamical semigroups—GKLS, Refs. 1 and 2).** A linear map  $L : \mathcal{M}_d \rightarrow \mathcal{M}_d$  is the generator of a continuous dynamical semigroup of quantum channels if and only if it can be written as

$$L(\rho) = i[\rho, H] + \sum_j \mathcal{L}_j \rho \mathcal{L}_j^\dagger - \frac{1}{2} \{ \mathcal{L}_j^\dagger \mathcal{L}_j, \rho \}, \tag{1}$$

where  $H = H^\dagger \in \mathcal{M}_d$  is self-adjoint and  $\{\mathcal{L}_j\}_j$  is a set of matrices in  $\mathcal{M}_d$ . Here,  $\{\cdot, \cdot\}$  denotes the anti-commutator.

For such generators, often called GKLS or Lindblad generators, we refer to the term  $i[\cdot, H]$  as the Hamiltonian part and to  $\sum_j \mathcal{L}_j \cdot \mathcal{L}_j^\dagger - \frac{1}{2} \{ \mathcal{L}_j^\dagger \mathcal{L}_j, \cdot \}$  as the dissipative part with Lindbladians  $\{\mathcal{L}_j\}_j$ .

We will call a quantum channel *Markovian* if it is an element of a quantum dynamical semigroup.

### III. MARKOVIAN DIVISIBILITY

The main motivation for our work is the following problem: Given a quantum channel, decide whether it comes from a (possibly time-dependent) Lindblad master equation. We take two different perspectives on this task to motivate our definitions.

The first perspective is that of differential equations. Specifically, we want to understand which quantum channels can arise as a solution of a time-dependent master equation of the form  $\frac{d}{dt} T_t = L(t) T_t$ , where  $L(t)$  is a time-dependent Lindblad generator. More generally, we want to study the possible solutions of a linear ordinary differential equation  $\frac{d}{dt} T_t = G(t) T_t$ , where  $t \mapsto G(t) \in \mathcal{G}$ , with  $\mathcal{G} \subset \mathcal{M}_d$  being a fixed set of generators.

Our second perspective on the problem comes from the semigroup structure of the solutions to time-independent master equations. Specifically, each such equation corresponds to a quantum dynamical semigroup. If we now also want to take into account a possible time-dependence of the generator while still preserving the semigroup structure, we can consider the semigroup generated by all elements of quantum dynamical semigroups. On an intuitive level, the question about solutions of master equations that we asked above now becomes the question of whether a given quantum channel is an element of this semigroup, i.e., we are dealing with the membership problem for this semigroup. Again, we can generalize the question by going from Lindblad generators to general generators.

#### A. Markovian divisibility with respect to general sets of generators

The two perspectives given above lead us to two slightly different definitions. In the first, we focus on the semigroup structure.

*Definition III.1 (Markovian divisibility).* Let  $\mathcal{G} \subset \mathcal{M}_d$  be a set of matrices, whose elements we call generators. We define the set

$$\mathcal{D}_{\mathcal{G}} := \left\{ T \in \mathcal{M}_d \mid \exists n \in \mathbb{N}, \text{ generators } \{G_i\}_{1 \leq i \leq n} \subset \mathcal{G} \text{ so that } \prod_{i=1}^n e^{G_i} = T \right\}.$$

We call the closure  $\overline{\mathcal{D}_{\mathcal{G}}}$  the set of linear maps that are Markovian divisible with respect to  $\mathcal{G}$ .

When translating the mathematical motivation of semigroups to a more physical motivation, Definition III.1 can be seen as an approach to the question of which linear maps can be arbitrarily well approximated using alternating exponentials of a fixed set of (control) generators.

Now, we give a definition based much on the perspective of differential equations determining the overall evolution on infinitesimal time intervals while keeping the semigroup structure in mind.

*Definition III.2 (Infinitesimal Markovian divisibility).* Let  $\mathcal{G} \subset \mathcal{M}_d$  be a compact and convex set of matrices containing  $0 \in \mathcal{M}_d$ . We will again refer to the elements of  $\mathcal{G}$  as generators. We define the set

$$\mathcal{I}_{\mathcal{G}} := \left\{ T \in \mathcal{M}_d \mid \forall \varepsilon > 0 \exists n \in \mathbb{N}, \text{ generators } \{G_j\}_{1 \leq j \leq n} \subset \mathcal{G} \right. \\ \left. \text{so that (i) } \|e^{G_j} - \mathbb{1}_d\| \leq \varepsilon \forall j \text{ and (ii) } \prod_{j=1}^n e^{G_j} = T \right\}.$$

We call the closure  $\overline{\mathcal{I}_{\mathcal{G}}}$  the set of linear maps that are infinitesimal Markovian divisible with respect to  $\mathcal{G}$ .

*Remark III.3.* In the definition, we require  $\mathcal{G}$  to be compact. This can be assumed without loss of generality. First, closedness can be assumed without loss of generality since for non-closed  $\mathcal{G}_0$ , we have  $\overline{\mathcal{I}_{\mathcal{G}_0}} = \overline{\mathcal{I}_{\overline{\mathcal{G}_0}}}$ . Second, boundedness can also be assumed without loss of generality. Specifically, suppose that  $\tilde{\mathcal{G}} \subset \mathcal{M}_d$  is an unbounded closed and convex set with  $0 \in \tilde{\mathcal{G}}$  and  $T \in \mathcal{I}_{\tilde{\mathcal{G}}}$ . Then, by definition,  $\forall \varepsilon > 0 \exists n \in \mathbb{N}$  and  $\{G_j\}_{1 \leq j \leq n} \subset \tilde{\mathcal{G}}$  such that  $\|e^{G_j} - \mathbb{1}_d\| \leq \varepsilon$  and  $\prod_{j=1}^n e^{G_j} = T$ . By convexity, also  $\frac{1}{N}G_j \in \tilde{\mathcal{G}} \forall 1 \leq j \leq n$  for every  $N > 1$ . By continuity of the matrix exponential, there exists  $N_0 \in \mathbb{N}$  such that  $\|e^{\frac{1}{N}G_j} - \mathbb{1}_d\| \leq \varepsilon$  for all  $N \geq N_0$ . Clearly, we can write  $T = \prod_{j=1}^n e^{G_j} = \prod_{j=1}^n \left(e^{\frac{1}{N}G_j}\right)^N$ . Thus, as  $\|e^{\frac{1}{N}G_j}\| \rightarrow 0$  as  $N \rightarrow \infty$ , we conclude that for every  $B > 0$ , we have  $T \in \mathcal{I}_{\tilde{\mathcal{G}}_{\leq B}}$ , where  $\tilde{\mathcal{G}}_{\leq B} := \{G \in \tilde{\mathcal{G}} \mid \|G\| \leq B\}$ . Hence, we can impose an arbitrary (non-zero) norm bound on our generators without changing the set of infinitesimal Markovian divisible channels.

Therefore, we are justified in using Definition III.2 also for non-compact  $\mathcal{G}$  (in particular, Lindblad generators and transition rate matrices).

*Remark III.4.* By continuity of the matrix exponential, it is easy to see that, if  $G \in \mathcal{G}$  implies  $\frac{1}{n}G \in \mathcal{G}$  for all  $n \in \mathbb{N}$ , then  $\mathcal{D}_{\mathcal{G}} = \mathcal{I}_{\mathcal{G}}$ . This is particularly the case if  $\mathcal{G}$  satisfies the assumptions of Definition III.2.

If, however,  $\mathcal{G}$  does not have this property, then (i) in the definition of  $\mathcal{I}_{\mathcal{G}}$  will, in general, lead to  $\mathcal{I}_{\mathcal{G}} \neq \mathcal{D}_{\mathcal{G}}$  (e.g.,  $\mathcal{I}_{\mathcal{G}}$  could be empty even if  $\mathcal{D}_{\mathcal{G}}$  is not).

When specifying  $\mathcal{G}$  to be the set of Lindblad generators and thus the linear maps of interest to be quantum channels, Definitions III.1 and III.2 become connected to quantum channels arising from master equations. Studying such channels via a notion of Markovian divisibility into infinitesimal pieces was first proposed in Ref. 3. Next, we discuss some results of that work.

## B. Infinitesimal Markovian divisibility of quantum channels

For ease of notation, we will denote by  $\mathcal{I}_d$  the set  $\mathcal{I}_{\mathcal{G}}$  for the specific choice of  $\mathcal{G}$  being the set of Lindblad generators acting on  $d \times d$ -matrices. Then, the set  $\overline{\mathcal{I}_d}$  is the set of *infinitesimal Markovian divisible quantum channels*, as defined in Ref. 3.

When referring to these channels, we will sometimes drop the ‘‘Markovian’’ for convenience. This can also be justified in a rigorous sense (see Theorem 16 in Ref. 3).

While some insight into the structure of infinitesimal Markovian divisible quantum channels has been obtained in Ref. 3, so far, there are no simple-to-check criteria for infinitesimal divisibility for a general dimension  $d$ . Such criteria are the main focus of this work.

A straightforward necessary criterion for infinitesimal divisibility is already observed in Ref. 3, namely, we have the following as a direct consequence of multiplicativity and continuity of the determinant:

*Proposition III.5.* An infinitesimal divisible quantum channel  $T$  satisfies  $\det(T) \geq 0$ .

This is, to our knowledge, the only necessary criterion for infinitesimal divisibility known so far that holds in any finite dimension.

For the special case of qubit channels, the set of infinitesimal divisible channels can be explicitly characterized by making use of the Lorentz normal form (the latter is discussed in Ref. 16).

**Theorem III.6 (Infinitesimal divisible qubit channels<sup>3</sup>—informal).** Let  $T : \mathcal{M}_2 \rightarrow \mathcal{M}_2$  be a generic qubit channel with the Lorentz normal form  $\begin{pmatrix} 1 & 0 \\ 0 & \Delta \end{pmatrix}$ .

$T$  is infinitesimal Markovian divisible if and only if  $0 \leq \det(\Delta) \leq s_{\min}^2$ , where  $s_{\min}$  is the smallest singular value of  $\Delta$ .

This characterization serves as one motivation for our results in higher dimensions, which we derive in Subsection IV B.

## IV. NECESSARY CRITERIA FOR MARKOVIAN DIVISIBILITY

We now develop necessary criteria for a linear map to be (infinitesimal) Markovian divisible. More precisely, our discussion aims toward establishing inequalities of the form

$$|\det(T)| \leq \left( \prod_{i=1}^k s_i^\uparrow(T) \right)^P. \quad (2)$$

We first present some results for the case of general linear maps and generators and later combine these observations with a more detailed analysis for quantum channels and Lindblad generators and stochastic matrices and transition rate matrices, respectively.

### A. General sets of generators

We first observe that if each of two matrices satisfies the desired inequality (2), then so does the product of the matrices.

*Lemma IV.1.* Let  $T_1, T_2 \in \mathcal{M}_d$ . Suppose that  $1 \leq k \leq d$  and  $p > 0$  such that

$$|\det(T_j)| \leq \left( \prod_{i=1}^k s_i^\uparrow(T_j) \right)^p$$

holds for  $j = 1, 2$ . Then, also

$$|\det(T_1 T_2)| \leq \left( \prod_{i=1}^k s_i^\uparrow(T_1 T_2) \right)^p.$$

*Proof.* A well-known majorization inequality for singular values states that

$$\prod_{i=1}^k s_i^\downarrow(AB) \leq \prod_{i=1}^k s_i^\downarrow(A) s_i^\downarrow(B) \tag{3}$$

for any  $1 \leq k \leq n$  for  $n \times n$ -matrices  $A, B$  (see Ref. 17, Theorem 3.3.4). With this, we obtain

$$\begin{aligned} |\det(T_1 T_2)| &= |\det(T_1)| |\det(T_2)| \\ &\leq \left( \prod_{i=1}^k s_i^\uparrow(T_1) \right)^p \left( \prod_{i=1}^k s_i^\uparrow(T_2) \right)^p \\ &= \left( \frac{|\det(T_1)| |\det(T_2)|}{\prod_{i=1}^{d-k} s_i^\downarrow(T_1) s_i^\downarrow(T_2)} \right)^p \\ &\leq \left( \frac{|\det(T_1 T_2)|}{\prod_{i=1}^{d-k} s_i^\downarrow(T_1 T_2)} \right)^p \\ &= \left( \prod_{i=1}^k s_i^\uparrow(T_1 T_2) \right)^p \end{aligned}$$

as claimed. Here, the first inequality is that, by assumption, the following step uses  $|\det(T_i)| = \prod_{j=1}^d s_j^\downarrow(T_i)$ , the second inequality is due to Eq. (3),

and the last step uses  $|\det(T_1 T_2)| = \prod_{j=1}^d s_j^\downarrow(T_1 T_2)$ . □

This means that, when trying to establish an inequality of the form (2), if  $T$  is a finite product, it suffices to consider the single factors separately.

Now we show that, once we have our desired inequality (2) for non-negative multiples of two separate generators, the exponential of the sum of these two generators also satisfies the inequality. This observation will be particularly useful in our analysis of Lindblad generators.

*Lemma IV.2.* Let  $G_1, G_2 \in \mathcal{M}_d$ . Suppose that  $1 \leq k \leq d$  and  $p > 0$  are such that

$$|\det\left(e^{\frac{G_j}{n}}\right)| \leq \left( \prod_{i=1}^k s_i^\uparrow\left(e^{\frac{G_j}{n}}\right) \right)^p$$

holds for all  $n \in \mathbb{N}$  and  $j = 1, 2$ . Then, also

$$|\det(e^{G_1+G_2})| \leq \left( \prod_{i=1}^k s_i^\uparrow(e^{G_1+G_2}) \right)^p.$$

*Proof.* By the Lie–Trotter formula,  $e^{A+B} = \lim_{n \rightarrow \infty} (e^{\frac{A}{n}} e^{\frac{B}{n}})^n$ . As both the determinant and the singular values depend continuously on the matrix, we can combine this with (an iterative application of) Lemma IV.1 to see whether it suffices to have  $|\det(e^{\frac{G_i}{n}})| \leq \left( \prod_{i=1}^k s_i^\uparrow(e^{\frac{G_i}{n}}) \right)^p$  for arbitrary  $n \in \mathbb{N}$ . We can summarize this reasoning as follows:

$$\begin{aligned} |\det(e^{G_1+G_2})| &= \lim_{n \rightarrow \infty} \left| \det \left( \begin{pmatrix} e^{\frac{G_1}{n}} & \\ & e^{\frac{G_2}{n}} \end{pmatrix} \right)^n \right| \\ &\leq \lim_{n \rightarrow \infty} \left( \prod_{i=1}^k s_i^\uparrow \left( \begin{pmatrix} e^{\frac{G_1}{n}} & \\ & e^{\frac{G_2}{n}} \end{pmatrix} \right)^n \right)^p \\ &= \left( \prod_{i=1}^k s_i^\uparrow(e^{G_1+G_2}) \right)^p, \end{aligned}$$

where the inequality follows by combining the assumption with Lemma IV.1. □

*Remark IV.3.* If  $G_j$  in Lemma IV.2 are normal matrices, then it is easy to see that the assumed inequality for  $n = 1$  already implies the corresponding inequality for any  $n \in \mathbb{N}$ . In general, however, this implication is not true. This can be seen as considering  $L$  and  $\frac{1}{2}L$ , with  $L$  given in Example IV.12. Therefore, we make the assumption for all  $n \in \mathbb{N}$ . This is also why we formulate Definition III.2 for convex sets of generators that contain the zero-matrix.

Next, we discuss how to reduce an inequality of the form (2) for a single matrix exponential to an inequality of eigenvalues of the exponent.

*Lemma IV.4.* Suppose that  $G \in \mathcal{M}_d$  satisfies  $\text{Tr}[G + G^*] - p \sum_{i=1}^k \lambda_i^\uparrow(G + G^*) \leq 0$ , then

$$|\det(e^G)| \leq \left( \prod_{i=1}^k s_i^\uparrow(e^G) \right)^p.$$

*Proof.* We observe that

$$\prod_{i=1}^k s_i^\uparrow(e^G) = \frac{|\det(e^G)|}{\prod_{i=1}^{d-k} s_i^\downarrow(e^G)} \geq \frac{|\det(e^G)|}{\prod_{i=1}^{d-k} s_i^\downarrow(e^{12(G+G^*)})} = \frac{\det(e^{\frac{1}{2}(G+G^*)})}{\prod_{i=1}^{d-k} e^{12\lambda_i^\downarrow(G+G^*)}} = \prod_{i=1}^k e^{\frac{1}{2}\lambda_i^\uparrow(G+G^*)},$$

where we used  $\prod_{i=1}^{d-k} s_i^\downarrow(e^G) \leq \prod_{i=1}^{d-k} s_i^\downarrow(e^{\Re(G)})$  (see p. 259 of Ref. 18) as well as  $|\det(e^G)| = \det(e^{\frac{1}{2}(G+G^*)})$ , which can be seen via Lie–Trotter. With this, we now obtain

$$|\det(e^G)|^2 = e^{\text{Tr}[G+G^*]} \leq \left( e^{\sum_{i=1}^k \lambda_i^\uparrow(G+G^*)} \right)^p = \left( \prod_{i=1}^k e^{\frac{1}{2}\lambda_i^\uparrow(G+G^*)} \right)^{2p} \leq \left( \prod_{i=1}^k s_i^\uparrow(e^G) \right)^{2p},$$

where the first inequality is exactly our assumption. Now we take the square root and obtain the claimed inequality. □

We summarize the results of the foregoing discussion for Markovian divisibility in the following.

**Theorem IV.5.** Let  $\mathcal{G} \subseteq \mathcal{M}_d$  be a set of generators. Let  $T \in \overline{\mathcal{D}}_{\mathcal{G}}$  and suppose that every  $G \in \mathcal{G}$  satisfies  $\text{Tr}[G + G^*] - p \sum_{i=1}^k \lambda_i^\uparrow(G + G^*) \leq 0$ .

Then,  $|\det(T)| \leq \left( \prod_{i=1}^k s_i^\uparrow(T) \right)^p$ .

*Proof.* By continuity of the determinant and the singular values, we can restrict our attention to  $T \in \mathcal{D}_{\mathcal{G}}$ . In that case, there exist  $n \in \mathbb{N}$  and generators  $\{G_i\}_{1 \leq i \leq n} \subset \mathcal{G}$  such that  $\prod_{i=1}^n e^{G_i} = T$ . By Lemma IV.1, it suffices to have the desired inequality for each factor  $e^{G_i}$ . These now satisfy the inequality by Lemma IV.4.

We obtain an analogous result for infinitesimal Markovian divisibility:

*Corollary IV.6.* Let  $\mathcal{G} \subset \mathcal{M}_d$  be a compact and convex set of matrices containing  $0 \in \mathcal{M}_d$ . Let  $\tilde{\mathcal{G}} := \{\lambda G \mid \lambda \in [0,1]\}$ ,  $G$  an extreme point of  $\mathcal{G} \subset \mathcal{G}$ . Assume that every  $\tilde{G} \in \tilde{\mathcal{G}}$  satisfies  $\text{Tr}[\tilde{G} + \tilde{G}^*] - p \sum_{i=1}^k \lambda_i^\uparrow(\tilde{G} + \tilde{G}^*) \leq 0$ . Let  $T \in \overline{\mathcal{I}}_{\mathcal{G}}$ . Then,  $0 \leq \det(T) \leq \left( \prod_{i=1}^k s_i^\uparrow(T) \right)^p$ .

*Proof.*  $\det(T) \geq 0$  follows in the same way as in Proposition III.5. By continuity, it suffices to prove the desired upper bound for  $T \in \mathcal{I}_G$ . By the definition of the set  $\mathcal{I}_G$  and Lemma IV.1, it then suffices to consider single factors of the form  $e^G$ ,  $G \in \mathcal{G}$ . By definition of  $\tilde{\mathcal{G}}$ ,  $\tilde{G} \in \tilde{\mathcal{G}}$ , in particular, implies that  $\frac{1}{n}\tilde{G} \in \tilde{\mathcal{G}}$  for all  $n \in \mathbb{N}$ . In addition, every element of  $\mathcal{G}$  can be expressed as a finite sum of elements of  $\tilde{\mathcal{G}}$  (by Krein–Milman). Therefore, we can apply Lemma IV.2 to conclude that it suffices to consider single factors of the form  $e^{\tilde{G}}$ ,  $\tilde{G} \in \tilde{\mathcal{G}}$ . Now we apply Lemma IV.4 to finish the proof.  $\square$

The assumption in Corollary IV.6 is about (truncated) rays through extreme points of the convex set of interest. In light of Remark IV.3, we expect that this can, in general, not be further simplified to an assumption only about the extreme points themselves (without multiples).

## B. Quantum channels

We now want to apply the reasoning from Subsection IV A to the more specific question of infinitesimal (Markovian) divisibility of quantum channels.

To avoid confusion about notation, in this subsection, we will denote the eigenvalues of a matrix  $\mathcal{L}$  as  $\lambda_i = \lambda_i(\mathcal{L})$ , whereas the eigenvalues of a linear map  $L$  on matrices are written as  $\Lambda_K = \Lambda_K(L)$ . For real eigenvalues of such linear superoperators, we use  $\Lambda_K^\downarrow$  ( $\Lambda_K^\uparrow$ ) to denote the eigenvalues in decreasing (increasing) order.

### 1. Determinant vs power of the smallest singular value

We first show that purely dissipative Lindblad generators with one Lindbladian satisfy an inequality, as assumed in Lemma IV.4 with only one summand:

*Lemma IV.7.* Let  $L : \mathcal{M}_d \rightarrow \mathcal{M}_d$  and  $L(\rho) = \mathcal{L}\rho\mathcal{L}^\dagger - \frac{1}{2}\{\mathcal{L}^\dagger\mathcal{L}, \rho\}$  be a purely dissipative Lindblad generator with one Lindbladian  $\mathcal{L} \in \mathcal{M}_d$ . Then,

$$\text{Tr}[L + L^*] - \frac{d}{2}\Lambda_1^\uparrow(L + L^*) \leq 0. \quad (4)$$

*Proof.* We adopt the following convention for vectorization of matrices: If  $A$  is an  $n \times n$ -matrix with column vectors  $a_i$ , then  $\text{vec}(A) = (a_1^T, \dots, a_n^T)^T$  is the column vector obtained by stacking the columns of  $A$  on top of one another. When using  $\text{vec}(ABC) = (C^T \otimes A)\text{vec}(B)$  to rewrite  $L + L^*$  as a  $d^2 \times d^2$ -matrix, we obtain

$$\text{vec}(L + L^*) = \overline{\mathcal{L}} \otimes \mathcal{L} + \overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger - \mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} - \overline{\mathcal{L}^\dagger} \mathcal{L} \otimes \mathbb{1}_d.$$

From this, it is easy to see that

$$\text{Tr}[L + L^*] = |\text{Tr}[\mathcal{L}]|^2 - 2d\|\mathcal{L}\|_F^2.$$

We observe that the Lindbladians  $\mathcal{L}$  and  $\lambda\mathbb{1}_d + \mathcal{L}$  give rise to the same superoperator  $L + L^*$  for every  $\lambda \in \mathbb{C}$ . Hence, we can, without loss of generality, assume that  $\text{Tr}[\mathcal{L}] = 0$  and therefore  $\text{Tr}[L + L^*] = -2d\|\mathcal{L}\|_F^2$ . Thus, we obtain

$$\begin{aligned} \text{Tr}[L + L^*] - \frac{d}{2}\Lambda_1^\uparrow(L + L^*) &\leq -2d\|\mathcal{L}\|_F^2 + \frac{d}{2}\|L + L^*\|_\infty \\ &\leq -2d\|\mathcal{L}\|_F^2 + \frac{d}{2}\left(\|\overline{\mathcal{L}} \otimes \mathcal{L}\|_\infty + \|\overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger\|_\infty + \|\mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L}\|_\infty + \|\overline{\mathcal{L}^\dagger} \mathcal{L} \otimes \mathbb{1}_d\|_\infty\right) \\ &= -2d\|\mathcal{L}\|_F^2 + \frac{d}{2} \cdot 4\|\mathcal{L}\|_\infty^2 \\ &\leq 0, \end{aligned}$$

which finishes the proof.  $\square$

*Remark IV.8.* In our Proof of Lemma IV.7, one step might strike the reader as particularly simplistic. Specifically, we estimate

$$\frac{d}{2}\|L + L^*\|_\infty \leq \frac{d}{2}\left(\|\overline{\mathcal{L}} \otimes \mathcal{L}\|_\infty + \|\overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger\|_\infty + \|\mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L}\|_\infty + \|\overline{\mathcal{L}^\dagger} \mathcal{L} \otimes \mathbb{1}_d\|_\infty\right) \leq \frac{d}{2} \cdot 4\|\mathcal{L}\|_\infty^2.$$

With a more thorough analysis, we can slightly improve this upper bound and thereby increase the prefactor in the statement of Lemma IV.7 from  $\frac{d}{2}$  to  $\approx 0.610733d$ . (We then get the same improvement in Corollary IV.9.) We derive this improvement in the [Appendix](#).

We can now apply the reasoning from Subsection IV A (for  $k = 1$  and  $p = \frac{d}{2}$ ) to obtain the following corollary:

*Corollary IV.9.* Let  $T \in \overline{\mathcal{I}}_d$ . Then,  $0 \leq \det(T) \leq (s_1^\uparrow(T))^{\frac{d}{2}}$ .

*Proof.* By combining the form of Lindblad generators from Theorem II.2 with Corollary IV.6, it suffices to consider Lindblad generators with a single summand, i.e., of the form

$$L(\rho) = \begin{cases} i[\rho, H] \text{ with } H = H^\dagger \\ \mathcal{L}\rho\mathcal{L}^\dagger - \frac{1}{2}\{\mathcal{L}^\dagger\mathcal{L}, \rho\}. \end{cases}$$

$[\cdot, H] : \mathcal{M}_d \rightarrow \mathcal{M}_d$  is a self-adjoint map if  $H = H^\dagger$ , and therefore,  $e^{i[\cdot, H]}$  has 1 as only singular value. The desired singular value inequality (2) is thus trivially satisfied for factors of this form. For factors of the form  $e^L$  with  $L(\rho) = \mathcal{L}\rho\mathcal{L}^\dagger - \frac{1}{2}\{\mathcal{L}^\dagger\mathcal{L}, \rho\}$ , the desired eigenvalue inequality is exactly shown in Lemma IV.7.  $\square$

This necessary criterion can be used to find channels that are not infinitesimal divisible and are given by convex combinations of a rank-deficient channel with the identity channel.

*Corollary IV.10.* Let  $T : \mathcal{M}_d \rightarrow \mathcal{M}_d$  be a quantum channel that has singular value of 0 of multiplicity  $1 \leq k < \frac{d}{2}$ . Then, every neighborhood of  $T$  contains a non-infinitesimal divisible channel.

*Proof.* Given such a quantum channel  $T$ , we can explicitly write down non-infinitesimal divisible channels via convex combination with the identity,  $T_\epsilon = (1 - \epsilon)T + \epsilon \text{Id}$ . By assumption,  $T_\epsilon$  has exactly  $k$  singular values, which go to 0 as  $\epsilon \rightarrow 0$ . Thus, either  $\det(T_\epsilon) < 0$  or we have

$$\det(T_\epsilon) = \prod_{j=1}^{d^2} s_j^\uparrow(T_\epsilon) \geq \left(s_1^\uparrow(T_\epsilon)\right)^k \prod_{j=k+1}^{d^2} s_j^\uparrow(T_\epsilon) > \left(s_1^\uparrow(T_\epsilon)\right)^{d/2} \text{ for } \epsilon \text{ small enough,}$$

where we just used that the  $d^2 - k$  largest singular values do not go to 0 for  $\epsilon \rightarrow 0$ . Hence, for  $\epsilon > 0$  small enough,  $T_\epsilon$  does not satisfy the criterion given in Corollary IV.9 and is therefore not infinitesimal divisible.  $\square$

*Example IV.11.* We can use the above Corollary to find infinitesimal divisible channels near the channel  $T : \mathcal{M}_d \rightarrow \mathcal{M}_d$ ,  $T(\rho) = \frac{\text{Tr}[\rho]}{d} \mathbb{1}_d$ .  $T$  is diagonal with respect to the generalized Gell–Mann basis of  $\mathcal{M}_d$  with the corresponding matrix given by  $\hat{T} = \text{diag}[1, 0, 0, \dots, 0]$ . The Choi matrix  $\tau$  of  $T$  has full rank and is thus particularly strictly positive definite (because complete positivity of  $T$  translates to positive semidefiniteness of its Choi matrix  $\tau$ ; see Ref. 15).

Hence, we can pick  $\epsilon > 0$  small enough such that  $\hat{T}_\epsilon = \text{diag}[1, \epsilon, \dots, \epsilon, 0]$  is the matrix representation of a completely positive map in the generalized Gell–Mann basis. As such a matrix  $\hat{T}_\epsilon$  describes by its very form a trace-preserving map, it corresponds to a quantum channel  $T_\epsilon$ , which now has an eigenvalue of 0 with a multiplicity of 1. Hence, we can apply Corollary IV.10 to  $T_\epsilon$  and thus find channels arbitrarily close to  $T$  that are not infinitesimal divisible.

Naturally, the question arises whether the power  $\frac{d}{2}$  in Corollary IV.9 is optimal. Our next example shows that the dependence on  $d$  cannot be better than linear and that the factor of  $\frac{1}{2}$  cannot be improved by much.

*Example IV.12.* When considering the pathological case of a matrix of the form

$$\mathcal{L} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

we can easily compute that

$$L + L^* = \begin{pmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & D_d \end{pmatrix}$$

with  $D_i = \text{diag}(0, \dots, 0, -1) \in \mathbb{R}^{d \times d}$  for  $1 \leq i \leq d-1$  and  $D_d = \text{diag}(-1, \dots, -1, -2) \in \mathbb{R}^{d \times d}$ . Hence,  $L + L^*$  has eigenvalues  $-1$  of multiplicity  $2(d-1)$ ,  $0$  of multiplicity  $d^2 - 2d$ , and  $-1 \pm \sqrt{2}$ , each of multiplicity  $1$ . In particular,  $\text{Tr}[L + L^*] - p\Lambda_1^\dagger(L + L^*) = -2d + (1 + \sqrt{2})p \leq 0$  iff  $p \leq \frac{2}{1+\sqrt{2}}d$ .

This example also shows that in Theorem IV.9, nothing better than  $\det(T) \leq (s_1^\dagger(T))^p$  with  $p = \mathcal{O}(d)$  can be achieved. Specifically, with the above choice of  $\mathcal{L}$ , we get

$$L = \begin{pmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & D_d \end{pmatrix}.$$

This can now be exponentiated to obtain

$$T := e^L = \begin{pmatrix} 0 & 0 & \cdots & 1 - e^{-1} \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} e^{\frac{1}{2}D_1} & 0 & \cdots & 0 \\ 0 & e^{\frac{1}{2}D_2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\frac{1}{2}D_d} \end{pmatrix},$$

where  $e^{1/2D_i} = \text{diag}(1, \dots, 1, e^{-1/2})$  for  $1 \leq i \leq d-1$  and  $e^{1/2D_d} = \text{diag}(e^{-1/2}, \dots, e^{-1/2}, e^{-1})$ .

We can now compute

$$T^*T = \begin{pmatrix} 0 & 0 & \cdots & 1 - e^{-1} \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 1 - e^{-1} & 0 & \cdots & (1 - e^{-1})^2 \end{pmatrix} + \begin{pmatrix} e^{D_1} & 0 & \cdots & 0 \\ 0 & e^{D_2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & e^{D_d} \end{pmatrix},$$

from which we see that  $T$  has singular values of  $1$  of multiplicity  $(d-1)^2 - 1$ ,  $e^{-\frac{1}{2}}$  of multiplicity  $2(d-1)$ ,  $\frac{\sqrt{1-e+e^2+(e-1)\sqrt{1+e^2}}}{e} \approx 1.200$  of multiplicity  $1$ , and  $\frac{\sqrt{1-e+e^2-(e-1)\sqrt{1+e^2}}}{e} \approx 0.306$  of multiplicity  $1$ . In particular, we have

$$\det(T) \leq (s_1^\dagger(T))^{\frac{d}{2}},$$

but

$$\det(T) > (s_1^\dagger(T))^d.$$

More precisely, we see that  $\det(T) \leq (s_1^\dagger(T))^p$  requires, as  $d \rightarrow \infty$ ,

$$p \leq \frac{\ln(s_1^\dagger(T)) + \ln(s_1^\dagger(T)) - (d-1)}{\ln(s_1^\dagger(T))} \approx \frac{\ln(1.200) + \ln(0.306) - (d-1)}{\ln(0.306)} \sim \frac{1}{-\ln(0.306)} d \approx 0.845 d.$$

If we do the same computation for  $\frac{1}{n}L$  instead of  $L$ , we obtain, in the limit of large  $n$ , the upper bound,

$$p \leq \frac{2}{1+\sqrt{2}} d + 1 + \frac{\sqrt{2}}{1+\sqrt{2}},$$

which coincides up to an additive constant with the bound obtained above on the level of eigenvalues.

This concludes our discussion of the example.



The result of Theorem IV.9 applied to the qubit case does not reproduce the criterion from Theorem III.6. In particular, we do not obtain  $s_{\min}^2$  but merely  $s_{\min}$ . For normal Lindbladians and thus products of unital channels, our reasoning can, however, be improved.

*Proposition IV.13.* For normal Lindbladians, the prefactor in Lemma IV.7 (thus the exponent in Corollary IV.9) can be improved to  $d$ . Furthermore, this estimate is sharp, i.e., cannot be improved for general normal  $\mathcal{L}$ .

*Proof.* For normal  $\mathcal{L}$ , we know all the eigenvalues of  $L + L^*$ , and they are given by  $\{-|\lambda_i - \lambda_j|^2\}_{i,j}$ , where  $\lambda_i$  are the eigenvalues of  $\mathcal{L}$  (see Remark IV.15 for a detailed derivation). Now choose two indices  $i^*, j^*$  such that

$$|\lambda_{i^*} - \lambda_{j^*}|^2 = \max_{i,j} |\lambda_i - \lambda_j|^2.$$

Then, (4) for exponent  $d$  becomes

$$\text{Tr}[L + L^*] - d\Lambda_1^\uparrow(L + L^*) = -\sum_{i,j} |\lambda_i - \lambda_j|^2 + d|\lambda_{i^*} - \lambda_{j^*}|^2. \tag{5}$$

Now using  $|a + b|^2 \leq 2(|a|^2 + |b|^2)$  and denoting the indices  $\{1, \dots, d\} \setminus \{i^*, j^*\} = \{n_1, \dots, n_{d-2}\}$ , we obtain

$$(5) \leq -\sum_{i,j} |\lambda_i - \lambda_j|^2 + 2|\lambda_{i^*} - \lambda_{j^*}|^2 + 2\sum_{k=1}^{d-2} (|\lambda_{i^*} - \lambda_{n_k}|^2 + |\lambda_{j^*} - \lambda_{n_k}|^2) \leq 0.$$

In the last step, we used that every difference  $|\lambda_{i^*/j^*} - \lambda_{n_k}|^2$  appears twice in the first sum.

In order to see that  $d$  is also optimal, consider the example  $\mathcal{L} = \text{diag}[1, -1, 0, \dots, 0]$ . Here, a straightforward calculation shows that  $L + L^*$  has eigenvalues  $-4$  of multiplicity 2,  $-1$  of multiplicity  $4(d - 2)$ , and 0 of multiplicity  $2 + (d - 2)^2$ . Thus,

$$\text{Tr}[L + L^*] = -4d = -d|\lambda_1 - \lambda_2|^2 = d\Lambda_1^\uparrow(L + L^*),$$

so  $d$  is optimal. □

Note that the example used in the previous proof can also be used to show that for normal  $\mathcal{L}$ , the exponent in  $\det(e^L) \leq (s_1^\uparrow(e^L))^d$  cannot be improved.

## 2. Determinant vs product of smallest singular values

So far, we have used the ideas from Subsection IV A to derive an upper bound on the determinant of infinitesimal divisible quantum channels in terms of the power of its smallest singular value. Now we focus on the other aspect of Lemma IV.4 and bound the determinant via a product of smallest singular values.

*Lemma IV.14.* Let  $L : \mathcal{M}_d \rightarrow \mathcal{M}_d$  and  $L(\rho) = \mathcal{L}\rho\mathcal{L}^\dagger - \frac{1}{2}\{\mathcal{L}^\dagger\mathcal{L}, \rho\}$  be a purely dissipative Lindblad generator with one Lindbladian  $\mathcal{L} \in \mathcal{M}_d$ . Then, for  $f(d) = 2d - 2\sqrt{2d} + 1$ , we have

$$\text{Tr}[L + L^*] - \sum_{K=1}^{\lfloor f(d) \rfloor} \Lambda_K^\uparrow(L + L^*) \leq 0. \tag{6}$$

*Proof.* As in the Proof of Lemma IV.7, we can, without loss of generality, assume that  $\text{Tr}[\mathcal{L}] = 0$ , and therefore,  $\text{Tr}[L + L^*] = -2d\|\mathcal{L}\|_F^2$ . We can now bound

$$\begin{aligned} -\sum_{K=1}^{\lfloor f(d) \rfloor} \Lambda_K^\uparrow(L + L^*) &\leq \sum_{K=1}^{\lfloor f(d) \rfloor} |\Lambda_K^\uparrow(L + L^*)| \\ &\leq \sum_{K=1}^{\lfloor f(d) \rfloor} s_K^\downarrow(L + L^*) \\ &= \|L + L^*\|_{(\lfloor f(d) \rfloor)} \\ &= \|\overline{\mathcal{L}} \otimes \mathcal{L} + \overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger - \mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} - \overline{\mathcal{L}^\dagger} \overline{\mathcal{L}} \otimes \mathbb{1}_d\|_{(\lfloor f(d) \rfloor)} \\ &\leq 2\|\overline{\mathcal{L}} \otimes \mathcal{L}\|_{(\lfloor f(d) \rfloor)} + \|\mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} + \overline{\mathcal{L}^\dagger} \overline{\mathcal{L}} \otimes \mathbb{1}_d\|_{(\lfloor f(d) \rfloor)}, \end{aligned}$$

where we used the  $k$ th Ky Fan norm,

$$\|A\|_{(k)} := \sum_{i=1}^k s_i^\downarrow(A).$$

We bound those two norms separately: For the first term,

$$\begin{aligned} \|\overline{\mathcal{L}} \otimes \mathcal{L}\|_{(\lfloor f(d) \rfloor)} &= \sum_{K=1}^{\lfloor f(d) \rfloor} s_K^\downarrow(\overline{\mathcal{L}} \otimes \mathcal{L}) \\ &\leq \sqrt{\lfloor f(d) \rfloor} \left( \sum_{K=1}^{\lfloor f(d) \rfloor} \left( s_K^\downarrow(\overline{\mathcal{L}} \otimes \mathcal{L}) \right)^2 \right)^{\frac{1}{2}} \\ &\leq \sqrt{\lfloor f(d) \rfloor} \|\overline{\mathcal{L}} \otimes \mathcal{L}\|_F \\ &= \sqrt{\lfloor f(d) \rfloor} \|\mathcal{L}\|_F^2, \end{aligned}$$

where the first inequality is an application of Cauchy–Schwarz.

For the second term, we choose an ONB with respect to which  $\mathcal{L}^\dagger \mathcal{L}$  is diagonal with the squares of the singular values  $s_i$  of  $\mathcal{L}$  on the diagonal (which is possible by unitary invariance of the Ky Fan norms) and then compute

$$\begin{aligned} \|\mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} + \overline{\mathcal{L}^\dagger \mathcal{L}} \otimes \mathbb{1}_d\|_{(\lfloor f(d) \rfloor)} &= \|\text{diag}[2s_1^2, s_1^2 + s_2^2, \dots, s_1^2 + s_d^2, s_1^2 + s_2^2, \dots, 2s_d^2]\|_{(\lfloor f(d) \rfloor)} \\ &\leq (\lfloor f(d) \rfloor + 1) \sum_{i=1}^d s_i^2 \\ &\leq (\lfloor f(d) \rfloor + 1) \|\mathcal{L}\|_F^2. \end{aligned}$$

Plugging this into the above, we obtain

$$\text{Tr}[L + L^*] - \sum_{K=1}^{\lfloor f(d) \rfloor} \Lambda_K^\dagger(L + L^*) \leq -2d \|\mathcal{L}\|_F^2 + (1 + 2\sqrt{\lfloor f(d) \rfloor} + \lfloor f(d) \rfloor) \|\mathcal{L}\|_F^2.$$

This is  $\leq 0$  if  $1 + 2\sqrt{\lfloor f(d) \rfloor} + \lfloor f(d) \rfloor - 2d \leq 0$ , which is guaranteed by the choice  $f(d) = 2d - 2\sqrt{2d} + 1$ . □

*Remark IV.15.* The reasoning in the Proof of Lemma IV.14 becomes particularly simple if the Lindbladian  $\mathcal{L}$  is normal. In that case, let  $\{v_j\}_j$  be an orthonormal basis for  $\mathbb{R}^d$  consisting of eigenvectors of  $\mathcal{L}$  corresponding to eigenvalues  $\{\lambda_j\}_j$ . By normality, the  $\{v_j\}_j$  are also eigenvectors of  $\mathcal{L}^\dagger$  to eigenvalues  $\{\overline{\lambda_j}\}_j$ . Recalling that in the matrix representation, we can write  $L + L^* = \overline{\mathcal{L}} \otimes \mathcal{L} + \overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger - \mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} - \overline{\mathcal{L}^\dagger \mathcal{L}} \otimes \mathbb{1}_d$ , it is now easy to see that  $\{\tilde{v}_i \otimes v_j\}_{i,j}$  is an orthonormal basis of  $\mathbb{C}^{d^2}$  consisting of eigenvectors of  $L + L^*$  to eigenvalues  $\{-|\lambda_i - \lambda_j|^2\}_{i,j}$ . Hence, all eigenvalues of  $L + L^*$  are  $\leq 0$ , and the inequality of Lemma IV.14 is trivially satisfied.

We can now apply our reasoning from Subsection IV A (with  $k = \lfloor 2d - 2\sqrt{2d} + 1 \rfloor$  and  $p = 1$ ) to obtain the following corollary:

*Corollary IV.16.* Let  $T \in \overline{\mathcal{T}}_d$ . Then, with  $f(d) = \lfloor 2d - 2\sqrt{2d} + 1 \rfloor$ , we have

$$0 \leq \det(T) \leq \prod_{i=1}^{f(d)} s_i^\dagger(T).$$

*Example IV.17.* Consider again the Lindblad generator  $L$  from Example IV.12 and the corresponding channel  $T$ . With the eigenvalues and singular values computed in Example IV.12, we see that in this case,  $\sum_{i=1}^{d^2-k} \Lambda_i^\dagger(L + L^*) > 0$  for all  $k \geq 2d - 1$ , and we have

$$\det(T) \leq \prod_{i=1}^{2d-2} s_i^\dagger(T),$$

but

$$\det(T) > \prod_{i=1}^k s_i^\uparrow(T)$$

for every  $d^2 > k > 2d - 2$ . This shows that in Corollary IV.16, nothing better than  $\det(T) \leq \prod_{i=1}^k s_i^\uparrow(T)$  with  $k = 2d - 2$  can be achieved.

*Remark IV.18.* After establishing the optimality of picking the smallest  $2d - C$  singular values in Corollary IV.16, the question naturally arises whether this bound can, in principle, be achieved with our proof strategy. In other words, what is the optimal choice for  $k$  such that

$$\left\| \overline{\mathcal{L}} \otimes \mathcal{L} + \overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger - \mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} - \overline{\mathcal{L}^\dagger} \mathcal{L} \otimes \mathbb{1}_d \right\|_{(k)} \leq 2d \|\mathcal{L}\|_F^2?$$

We clearly have

$$\left\| \overline{\mathcal{L}} \otimes \mathcal{L} + \overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger - \mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} - \overline{\mathcal{L}^\dagger} \mathcal{L} \otimes \mathbb{1}_d \right\|_{(k)} \leq 2 \left\| \overline{\mathcal{L}} \otimes \mathcal{L} \right\|_{(k)} + \left\| \mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} + \overline{\mathcal{L}^\dagger} \mathcal{L} \otimes \mathbb{1}_d \right\|_{(k)}.$$

The first term has the singular values  $s_i(\mathcal{L})s_j(\mathcal{L})$ , and the second one has singular values  $s_i^2(\mathcal{L}) + s_j^2(\mathcal{L})$ . Thus, if we normalize the Frobenius norm of  $\mathcal{L}$  to 1 and write  $p_i = s_i^2(\mathcal{L})$ , we can reduce the desired bound to the following conjecture:

*Conjecture IV.19.* Let  $p \in \mathbb{R}_{\geq 0}^d$  with  $\sum_{i=1}^d p_i = 1$ . Define the matrices  $a, g \in \mathbb{R}^{d \times d}$  via

$$a_{ij} = \frac{p_i + p_j}{2}, \quad g_{ij} = \sqrt{p_i p_j}.$$

Denote by  $a_k^\downarrow$  and  $g_k^\downarrow$  the  $k$ th largest entry of  $a$  and  $g$ , respectively. Define

$$A = \sum_{k=1}^{h(d)} a_k^\downarrow, \quad G = \sum_{k=1}^{h(d)} g_k^\downarrow.$$

We conjecture that the maximal integer  $h(d)$  such that  $A + G \leq d$  holds for any probability vector  $p$  is given by  $h(d) = 2d - 5$ .

We have tested this conjecture numerically for a wide range of dimensions. Theoretically, it stems from the fact that we know the optimal values and corresponding probability vectors for the arithmetic [ $h(d) = 2d - 2$ ] and geometric mean [ $h(d) = d^2$ ], respectively. Hence,  $A$  is by far more decisive and  $G$  can only worsen the maximal number of summands by a bit. If we were able to prove this conjecture, we could choose  $f(d) = h(d) = 2d - 5$  in Corollary IV.16, which would bring us closer to the optimum of  $2d - 2$  up to an additive constant.

*Remark IV.20.* In contrast to Subsection IV B 1, here, we cannot provide an example of a quantum channel that violates the criterion from Corollary IV.16. As any channel having only singular values  $\leq 1$  trivially satisfies the criterion, no unital channel will provide a violation, which makes analytically constructing an example more difficult. We have also tried to find an example of a non-infinitesimal divisible channel that is recognized as such by the conjectured optimal version of our criterion (which we cannot prove yet) numerically via minimizing the fraction  $\prod_{i=1}^{2d-2} s_i^\uparrow(T) / \det(T)$  over channels. This has, however, not been successful. We would be interested in any comments as to how such an example can be found or why finding one is a challenging task.

So far in our treatment of infinitesimal divisible quantum channels, we considered two extreme cases, namely, estimating the determinant by the highest possible power of the smallest singular value and by the product of the largest possible number of the lowest singular values all with exponent 1. The next proposition corresponds to an interpolation between those two results.

*Proposition IV.21.* Let  $T \in \overline{\mathcal{I}}_d$ . Then, for any  $1 \leq k \leq d^2$  with  $g(d) = \frac{2d}{k+2\sqrt{k+1}}$ , we have

$$0 \leq \det(T) \leq \left( \prod_{i=1}^k s_i^\uparrow(T) \right)^{g(d)}.$$

*Proof.* As shown in Subsection IV A, it suffices to show that any Lindblad generator  $L$  satisfies

$$\text{Tr}[L + L^*] - g(d) \sum_{\ell=1}^k \Lambda_\ell^\uparrow(L + L^*) \leq -2d \|\mathcal{L}\|_F^2 + g(d) \|L + L^*\|_{(k)} \leq 0.$$

Again, we only need to consider purely dissipative Lindblad generators with a single Lindbladian. For such generators, the desired assertion follows from the bound on the Ky Fan norm provided in the Proof of Lemma IV.14,

$$\|L + L^*\|_{(k)} \leq (k + 2\sqrt{k} + 1) \|\mathcal{L}\|_F^2.$$

□

*Remark IV.22.* In our numerical tests, we observe the result of Corollary IV.9 to be the strongest in generic cases in higher dimensions, since generically, the smallest singular value seems to be of some orders of magnitude smaller than the others. However, the result in Proposition IV.21 might give useful improvements for small dimensions, especially if some of the lowest singular values are all of the same order of magnitude. Take the case  $d = 3, k = 2$ , and then, we get the three results,

$$0 \leq \det(T) \leq \begin{cases} s_1^\uparrow(T)^{3/2} & \text{(Corollary IV.9)} \\ s_1^\uparrow(T)s_2^\uparrow(T) & \text{(Corollary IV.16)} \\ (s_1^\uparrow(T)s_2^\uparrow(T))^{\frac{6}{3+2\sqrt{2}}} & \text{(Proposition IV.21)}. \end{cases}$$

Hence, if  $s_1^\uparrow(T)$  is a lot smaller than  $s_2^\uparrow(T)$ , the first result is the strongest. However, if  $s_1^\uparrow(T) \approx s_2^\uparrow(T)$ , then the last result becomes the strongest criterion out of the three.

### C. Stochastic matrices

The classical counterparts of quantum channels and Lindblad generators are stochastic matrices and transition rate matrices, respectively. In particular, when choosing the set of generators to be the set of all transition rate matrices, we obtain a notion of (infinitesimal) Markovian divisibility for stochastic matrices.

Motivated by the results of Subsections IV A and IV B, we now study whether similar criteria for infinitesimal divisibility of stochastic matrices can be established. More precisely, we define the following:

*Definition IV.23 (Markovian divisible stochastic matrices).* We define the set of  $d \times d$  stochastic matrices to be

$$\mathcal{S}_d := \left\{ S \in \mathbb{R}^{d \times d} \mid S_{ij} \geq 0 \ \forall i, j \ \text{and} \ \sum_{j=1}^d S_{ij} = 1 \ \forall i \right\}$$

and the set of  $d \times d$  transition rate matrices to be

$$\mathcal{Q}_d := \left\{ Q \in \mathbb{R}^{d \times d} \mid Q_{ij} \geq 0 \ \forall i \neq j \ \text{and} \ \sum_{j=1}^d Q_{ij} = 0 \ \forall i \right\}.$$

We call a stochastic matrix  $S \in \mathcal{S}_d$  Markovian divisible if it is Markovian divisible with respect to the set of generators  $\mathcal{Q}_d$  in the sense of Definition III.1.

Note that, as discussed in Remark III.4, the “infinitesimal” requirement is automatically contained in this definition due to the structure of the set  $\mathcal{Q}_d$ , which is why we do not write it out explicitly.

Our first observation is that, in contrast to the case of Lindblad generators studied in Subsection IV B, when allowing all transition rate matrices as generators, no non-trivial necessary criteria of our desired form (2) can hold.

*Example IV.24.* Take the transition rate matrix

$$Q = \begin{pmatrix} -1 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \in \mathcal{Q}_d, \quad \text{and then, } e^Q = \begin{pmatrix} \frac{1}{e} & 0 & \cdots & 0 & 1 - \frac{1}{e} \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

which has singular values  $\frac{\sqrt{1-e+e^2+(e-1)\sqrt{1+e^2}}}{e} \approx 1.200$  of multiplicity 1, 1 of multiplicity  $d - 2$ , and  $\frac{\sqrt{1-e+e^2-(e-1)\sqrt{1+e^2}}}{e} \approx 0.306$  of multiplicity 1. In particular, we see that for every  $1 \leq k < d$ ,

$$\det(e^Q) > \prod_{i=1}^k s_i^\uparrow(e^Q).$$

Hence, for Markovian divisible stochastic matrices, there cannot be a non-trivial necessary criterion of the form of Corollary IV.16. Similarly, no non-trivial necessary criterion as in Corollary IV.9 with an exponent growing with some positive power of  $d$  can hold when we take the set  $\mathcal{G}$  of generators to be all transition rate matrices.

This example, together with Corollaries IV.16 and IV.9, implies the following:

*Corollary IV.25.* *There cannot be a mapping from  $d^2 \times d^2$  stochastic matrices to  $\mathcal{T}_d$  that both preserves infinitesimal Markovian divisibility and leaves singular values invariant.*

We can, however, restrict our attention to strict subsets of all transition rate matrices and derive analogous criteria there.

*Lemma IV.26.* *Let  $c \in (0, 1]$ . Consider the set of generators*

$$\mathcal{G}_c := \{Q \in \mathbb{R}^{d \times d} \mid Q \text{ is a transition rate matrix and } Q_{kk} \leq c \min_{1 \leq l \leq d} Q_{ll} \forall 1 \leq k \leq d\}.$$

Then,  $\text{Tr}[Q + Q^T] - \frac{1+c(d-1)}{2} \lambda_1^\uparrow(Q + Q^T) \leq 0$ .

*Proof.* Clearly, for  $Q \in \mathcal{G}_c$ , we have  $\text{Tr}[Q + Q^T] = 2 \sum_{i=1}^d Q_{ii} \leq 2(1 + c(d - 1)) \min_{1 \leq l \leq d} Q_{ll}$ . As  $\sum_{j=1}^d Q_{ij} = 0$  for all  $1 \leq i \leq d$ , we can use Gerschgorin discs to obtain  $\lambda_1^\uparrow(Q + Q^T) \geq 4 \min_{1 \leq l \leq d} Q_{ll}$ . In particular, we have that

$$\text{Tr}[Q + Q^T] - \frac{1 + c(d - 1)}{2} \lambda_1^\uparrow(Q + Q^T) \leq 2(1 + c(d - 1)) \min_{1 \leq l \leq d} Q_{ll} - 2(1 + c(d - 1)) \min_{1 \leq l \leq d} Q_{ll} = 0,$$

as claimed. □

According to our reasoning from Subsection IV A, this directly implies the following corollary:

*Corollary IV.27.* *Let  $c \in (0, 1]$ . Suppose that  $S \in [0, 1]^{d \times d}$  is a stochastic matrix that is Markovian divisible with respect to  $\mathcal{G}_c$ . Then,  $\det(S) \leq (s_1^\uparrow(S))^{\frac{1+c(d-1)}{2}}$ .*

If we set  $c = 1$ , then  $\mathcal{G}_1$  describes the set of transition rate matrices with constant diagonal. For Markovian divisibility of a stochastic matrix  $S$  with respect to this restricted set of generators, we obtain again the criterion  $\det(S) \leq (s_1^\uparrow(S))^{\frac{d}{2}}$ .

## V. CONCLUSION

In this work, we described how the notion of infinitesimal Markovian divisibility introduced in Ref. 3 as a notion of Markovianity for quantum channels with the generators in Lindblad form can be extended to a notion applicable to general linear maps and a (closed and convex) set of generators.

Our main contribution toward an understanding of this notion is a general proof strategy based on (sub-) multiplicativity properties of the determinant and products of largest singular values as well as Trotterization, with which we can establish necessary criteria for infinitesimal Markovian divisibility from a spectral property of the generators.

We showed that all Lindblad generators satisfy such a property, and therefore, our approach yields necessary criteria for infinitesimal Markovian divisibility of quantum channels in any (finite) dimension. These are the first such criteria beyond dimension 2 aside from non-negativity of the determinant. Using these criteria, we gave new examples of provably non-infinitesimal Markovian divisible quantum channels that can be found in any neighborhood of any rank-deficient quantum channel.

However, when studying the classical counterpart—stochastic matrices as maps of interest and transition rate matrices as generators—we found that in the general scenario in which all possible transition rate matrices are allowed as generators, no necessary criterion of our desired form can hold. We could apply our proof strategy only after imposing an additional restriction on the allowed transition rate matrices, which

can be interpreted as requiring that the time scales for remaining in any of the states of the Markov chain are comparable. (In particular, we have to assume that there are no absorbing states.)

Several follow-up questions arise naturally from our work. The first such question is for improvements of our results of Corollaries IV.9 and IV.16. In Examples IV.12 and IV.17, we have shown that our results are close to optimal with respect to the dimension dependence of the exponent in Corollary IV.9 and optimal in the leading order with respect to the number of factors in Corollary IV.16. Nevertheless, there remains a gap to be closed. One possible step for improving Corollary IV.16 might lie in a better understanding of Conjecture IV.19. One might also wonder whether there is a subclass of Lindblad operators for which our proof strategy yields stronger bounds.

More generally, we are hoping for a better understanding of the result of Corollary IV.16. A crucial first step would be to find—either analytically or numerically—examples of not infinitesimal Markovian divisible quantum channels that violate the inequality in Corollary IV.16 (or, for that matter, our conjectured improvement of it). As our proof of this inequality makes extensive use of the assumed divisibility structure, we would consider it surprising if no such examples could be found, which would make it trivial as a necessary criterion.

We mention one more natural question concerning the case of infinitesimal Markovian divisible quantum channels. Specifically, now that we have established necessary criteria for this property, can these be complemented by sufficient criteria of a similar form? The results of Ref. 3 show that for generic qubit channels, an inequality between the determinant of a channel and the square of its smallest singular value is indeed a necessary and sufficient criterion for infinitesimal Markovian divisibility. However, it is not at all clear whether this generalizes to higher dimensions.

Finally, here, we have applied our general proof strategy to two scenarios: that of Lindblad generators and that of transition rate matrices as generators. It would be interesting to find other sets of matrix semigroups whose generators satisfy a spectral property as required in Theorem IV.5.

#### ACKNOWLEDGMENTS

M.C.C. and B.R.G. thank Michael M. Wolf for suggesting this problem and for many insightful discussions. We are also grateful for the suggestions made by the anonymous reviewer at the Journal of Mathematical Physics.

M.C.C. gratefully acknowledges support from the TopMath Graduate Center of the TUM Graduate School at the Technical University of Munich, Germany, and the TopMath Program at the Elite Network of Bavaria. M.C.C. was supported by a doctoral scholarship of the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes).

B.R.G. gratefully acknowledges support from the International Research Training Group (IGDK Munich—Graz) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project No. 188264188/GRK1754.

#### APPENDIX: PROOF OF AN IMPROVEMENT TO COROLLARY IV.9

As mentioned in Remark IV.8, we are able to improve the exponent in Corollary IV.9 from  $\frac{d}{2}$  to  $\frac{2}{2+\sqrt{\frac{13}{8}}} d \approx 0.610733 d$ .

The idea behind the improvement is to estimate more carefully the smallest (“most negative”) eigenvalue  $\Lambda_1^\dagger(L + L^*)$ . In the proof of Corollary IV.9, we simply estimate  $\Lambda_1^\dagger(L + L^*)$  from below by  $-4\|\mathcal{L}\|_F^2$ , which yields the exponent  $\frac{d}{2}$  when comparing it to the  $-2d\|L\|_F^2$  from the trace of  $L + L^*$ . To obtain our improved version, we prove the following lemma.

*Lemma A.1.* Let  $\mathcal{L} \in \mathcal{M}_d$  and  $L(\rho) = \mathcal{L}\rho\mathcal{L}^\dagger - \frac{1}{2}\{\mathcal{L}^\dagger\mathcal{L}, \rho\}$ . Then,

$$\Lambda_1^\dagger(L + L^*) \geq -\left(2 + \sqrt{\frac{13}{8}}\right)\|\mathcal{L}\|_F^2.$$

*Proof.* The starting point for our reasoning is the  $l^2$ -version of the Gerschgorin disc Theorem (see Ref. 18), which states that for a Hermitian matrix  $A = (a_{ij})_{ij}$ , each interval  $[a_{ii} - r_i, a_{ii} + r_i]$  contains at least one eigenvalue of  $A$ , where

$$r_i = \left(\sum_{j \neq i} |a_{ij}|^2\right)^{1/2}.$$

Next, note that due to the tensor-structure of  $L + L^*$ , we can write its entries in a matrix representation as

$$(L + L^*)_{kl} = \bar{\mathcal{L}}_{(q+1)(p+1)} \mathcal{L}_{rs} + \mathcal{L}_{(p+1)(q+1)} \bar{\mathcal{L}}_{sr} - \delta_{qp}(\mathcal{L}^\dagger \mathcal{L})_{rs} - (\bar{\mathcal{L}}^\dagger \bar{\mathcal{L}})_{(q+1)(p+1)} \delta_{rs},$$

where  $k = qd + r, l = pd + s$  with  $q \in \{0, \dots, d-1\}, r \in \{1, \dots, d\}$ . If we now choose an orthonormal basis such that  $\mathcal{L}^\dagger \mathcal{L} = \text{diag}[\sigma_1^2, \dots, \sigma_d^2]$ , we obtain, for the diagonal entries,

$$(L + L^*)_{kk} = \bar{\mathcal{L}}_{(q+1)(q+1)} \mathcal{L}_{rr} + \mathcal{L}_{(q+1)(q+1)} \bar{\mathcal{L}}_{rr} - \sigma_r^2 - \sigma_{(q+1)}^2.$$

For the off-diagonal entries, we need to consider only the first two terms in  $L + L^*$  due to the choice of our basis, i.e., we get, for  $k \neq l$ ,

$$(L + L^*)_{kl} = \bar{\mathcal{L}}_{(q+1)(p+1)} \mathcal{L}_{rs} + \mathcal{L}_{(p+1)(q+1)} \bar{\mathcal{L}}_{sr}.$$

We need to distinguish two cases.

Case  $k = 1$ : Here, we have

$$(L + L^*)_{11} = 2|\mathcal{L}_{11}|^2 - 2\sigma_1^2$$

and

$$\begin{aligned} \sum_{k \neq 1} |(L + L^*)_{1k}|^2 &= \sum_{q,r} |\bar{\mathcal{L}}_{1(q+1)} \mathcal{L}_{1r} + \mathcal{L}_{(q+1)1} \bar{\mathcal{L}}_{r1}|^2 \\ &\leq \sum_q |\bar{\mathcal{L}}_{1(q+1)}|^2 \sum_r |\bar{\mathcal{L}}_{1r}|^2 + \sum_q |\bar{\mathcal{L}}_{(q+1)1}|^2 \sum_r |\bar{\mathcal{L}}_{r1}|^2 + 2 \left( \sum_r \underbrace{|\mathcal{L}_{1r} \bar{\mathcal{L}}_{r1}|}_{\leq \frac{1}{2}(|\mathcal{L}_{1r}|^2 + |\bar{\mathcal{L}}_{r1}|^2)} \right)^2 \\ &\leq \|\mathcal{L}\|_F^2 (\|\mathcal{L}\|_F^2 + |\mathcal{L}_{11}|^2) + \frac{1}{2} (\|\mathcal{L}\|_F^2 + |\mathcal{L}_{11}|^2)^2, \end{aligned}$$

where in the last step, we used that, since we are summing up the first row and column, only the diagonal entry  $|\mathcal{L}_{11}|^2$  appears twice and the sum of the remaining squares can be bounded by one Frobenius norm.

Before we proceed, let us note that without loss of generality, we can normalize  $\|\mathcal{L}\|_F^2 = 1$  to make the following computations more readable. Then, we obtain, by completing the square,

$$\sum_{k \neq 1} |(L + L^*)_{1k}|^2 \leq 1 + |\mathcal{L}_{11}|^2 + \frac{1}{2} (1 + |\mathcal{L}_{11}|^2)^2 = \left( \sqrt{\frac{3}{2}} + \sqrt{\frac{2}{3}} |\mathcal{L}_{11}|^2 \right)^2 - \frac{1}{6} |\mathcal{L}_{11}|^4.$$

Thus,

$$(L + L^*)_{11} - \left( \sum_{k \neq 1} |(L + L^*)_{1k}|^2 \right)^{1/2} \geq 2|\mathcal{L}_{11}|^2 - 2\sigma_1^2 - \sqrt{\frac{3}{2}} - \sqrt{\frac{2}{3}} |\mathcal{L}_{11}|^2 \geq - \left( 2 + \sqrt{\frac{3}{2}} \right).$$

Hence, in this case, we are even able to bound  $a_{ii} - r_i$  from below by  $-(2 + \sqrt{\frac{3}{2}}) \|\mathcal{L}\|_F^2$ .

Case  $k \neq 1$ : Here, we obtain, for the diagonal entries using Young's inequality,

$$(L + L^*)_{kk} = \bar{\mathcal{L}}_{(q+1)(q+1)} \mathcal{L}_{rr} + \mathcal{L}_{(q+1)(q+1)} \bar{\mathcal{L}}_{rr} - \sigma_r^2 - \sigma_{(q+1)}^2 \geq -2|\bar{\mathcal{L}}_{(q+1)(q+1)} \mathcal{L}_{rr}| - \|\mathcal{L}\|_F^2.$$

Note that the two singular values might be the same but can, nevertheless, be bounded by just one Frobenius norm, which is the important difference to the case  $k = 1$ .

For the off-diagonal entries, we start off in the same way as above,

$$\begin{aligned} \sum_{l \neq k} |(L + L^*)_{kl}|^2 &\leq \sum_{(p,s) \neq (q,r)} |\bar{\mathcal{L}}_{(q+1)(p+1)} \mathcal{L}_{rs}|^2 + |\mathcal{L}_{(p+1)(q+1)} \bar{\mathcal{L}}_{sr}|^2 + 2|\bar{\mathcal{L}}_{(q+1)(p+1)} \mathcal{L}_{rs} \mathcal{L}_{(p+1)(q+1)} \bar{\mathcal{L}}_{sr}| \\ &= \left( \sum_p |\mathcal{L}_{(q+1)(p+1)}|^2 \right) \left( \sum_s |\mathcal{L}_{rs}|^2 \right) + \left( \sum_p |\mathcal{L}_{(p+1)(q+1)}|^2 \right) \left( \sum_s |\bar{\mathcal{L}}_{sr}|^2 \right) \\ &\quad + 2 \left( \sum_p |\bar{\mathcal{L}}_{(q+1)(p+1)} \mathcal{L}_{(p+1)(q+1)}| \right) \left( \sum_s |\mathcal{L}_{rs} \bar{\mathcal{L}}_{sr}| \right) - 4|\bar{\mathcal{L}}_{(q+1)(q+1)} \mathcal{L}_{rr}|^2 \\ &\leq \|\mathcal{L}\|_F^2 (\|\mathcal{L}\|_F^2 + \min\{|\mathcal{L}_{rr}|^2, |\mathcal{L}_{(q+1)(q+1)}|^2\}) - 4|\bar{\mathcal{L}}_{(q+1)(q+1)} \mathcal{L}_{rr}|^2 \\ &\quad + \frac{1}{2} (\|\mathcal{L}\|_F^2 + |\mathcal{L}_{rr}|^2) (\|\mathcal{L}\|_F^2 + |\mathcal{L}_{(q+1)(q+1)}|^2). \end{aligned}$$

Again normalizing  $\|\mathcal{L}\|_F^2 = 1$  and denoting  $x = |\mathcal{L}_{(q+1)(q+1)}|, y = |\mathcal{L}_{rr}|$  give us

$$(L + L^*)_{kk} - \left( \sum_{l \neq k} |(L + L^*)_{kl}|^2 \right)^{1/2} \geq -2xy - 1 - \left( (1 + \min\{x^2, y^2\}) + \frac{1}{2}(1 + x^2)(1 + y^2) - 4x^2y^2 \right)^{1/2} \\ =: g(x, y).$$

Taking the minimum of the function on the right-hand side over (the upper half of) the unit disk  $x^2 + y^2 \leq 1$  gives us

$$(L + L^*)_{kk} - \left( \sum_{l \neq k} |(L + L^*)_{kl}|^2 \right)^{1/2} \geq \min_{B_1(0)} g(x, y) = g\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = -2 - \sqrt{\frac{13}{8}}.$$

As the second case  $k \neq 1$  gives us the worse bound, our final estimate is precisely the statement from Lemma A.1.  $\square$

Again, this has to be compared to  $-2d\|\mathcal{L}\|_F^2$  in the reasoning of the proof of Corollary IV.9, whereby we obtain the claimed exponent  $\frac{2}{2 + \sqrt{\frac{13}{8}}} d$  (instead of the previous  $\frac{d}{2}$ ).

#### DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

#### REFERENCES

- <sup>1</sup>V. Gorini, A. Kossakowski, and E. C. G. Sudarshan, *J. Math. Phys.* **17**, 821 (1976).
- <sup>2</sup>G. Lindblad, *Commun. Math. Phys.* **48**, 119 (1976).
- <sup>3</sup>M. M. Wolf and J. I. Cirac, *Commun. Math. Phys.* **279**, 147 (2008).
- <sup>4</sup>T. S. Cubitt, J. Eisert, and M. M. Wolf, *Commun. Math. Phys.* **310**, 383 (2012).
- <sup>5</sup>J. Bausch and T. Cubitt, *Linear Algebra Appl.* **504**, 64 (2016).
- <sup>6</sup>M. M. Wolf, J. Eisert, T. S. Cubitt, and J. I. Cirac, *Phys. Rev. Lett.* **101**, 150402 (2008).
- <sup>7</sup>Á. Rivas, S. F. Huelga, and M. B. Plenio, *Rep. Prog. Phys.* **77**, 094001 (2014).
- <sup>8</sup>H.-P. Breuer, E.-M. Laine, J. Piilo, and B. Vacchini, *Rev. Mod. Phys.* **88**, 021002 (2016).
- <sup>9</sup>L. Li, M. J. W. Hall, and H. M. Wiseman, *Phys. Rep.* **759**, 1 (2018).
- <sup>10</sup>C.-F. Li, G.-C. Guo, and J. Piilo, *Europhys. Lett.* **127**, 50001 (2019).
- <sup>11</sup>D. Davalos, M. Ziman, and C. Pineda, *Quantum* **3**, 144 (2019).
- <sup>12</sup>D. Chruscinski and U. Chakraborty, *New J. Phys.* **23**, 013009 (2021).
- <sup>13</sup>A. Rivas, S. F. Huelga, and M. B. Plenio, *Phys. Rev. Lett.* **105**, 050403 (2010).
- <sup>14</sup>H.-P. Breuer, E.-M. Laine, and J. Piilo, *Phys. Rev. Lett.* **103**, 210401 (2009).
- <sup>15</sup>M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, 10th ed. (Cambridge University Press, Cambridge, 2010).
- <sup>16</sup>F. Verstraete and H. Verschelde, "On quantum channels," eprint [arXiv:quant-ph/0202124](https://arxiv.org/abs/quant-ph/0202124) (2002).
- <sup>17</sup>R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis* (Cambridge University Press, Cambridge, 1991).
- <sup>18</sup>R. Bhatia, *Matrix Analysis*, Graduate Texts in Mathematics Vol. 169 (Springer, New York, NY, 1997).