Galerkin method using optimized wavelet-Gaussian mixed bases for electronic structure calculations in quantum chemistry

Dinh Huong Pham

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques Appliquées
Arrêté ministériel : 25 mai 2016

Présentée par

Dinh Huong PHAM

Thèse dirigée par **Valérie PERRIER**, Professeur des universités, Grenoble INP, co-dirigée par **Quang Huy TRAN**, Ingénieur de recherche, IFPEN, et co-encadrée par **Luigi GENOVESE**, Ingénieur de recherche, CEA, avec la participation de **Laurent DUVAL**, Ingénieur de recherche, IFPEN

préparée au sein d'**IFP Energies nouvelles** (Rueil-Malmaison) et du **Laboratoire Jean Kuntzmann** (Grenoble), dans l'**École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Bases mixtes ondelettes–gaussiennes pour le calcul de structures électroniques

Thèse soutenue publiquement le **30/06/2017**,
devant le jury composé de :

M. Stéphane LABBÉ
PROFESSEUR DES UNIVERSITÉS, UNIVERSITÉ GRENOBLE-ALPES, Président

Mme Silvia BERTOLUZZA
DIRECTEUR DE RECHERCHE, CNR-IMATI, PAVIA, Rapporteur
M. Yvon MADAY
PROFESSEUR DES UNIVERSITÉS, UPMC, LJLL, Rapporteur

M. Patrick FISCHER
MAÎTRE DE CONFÉRENCES, HDR, U. BORDEAUX 1, IMB, Examinateur
M. Christophe MORELL
PROFESSEUR DES UNIVERSITÉS, U. LYON 1, ISA, Examinateur

Mme Valérie PERRIER
PROFESSEUR DES UNIVERSITÉS, GRENOBLE INP, Directeur de thèse
M. Quang Huy TRAN
INGÉNIEUR DE RECHERCHE, HDR, IFPEN, Co-Directeur de thèse
M. Luigi GENOVESE
INGÉNIEUR DE RECHERCHE, CEA, INAC, Co-Encadrant

M. Thierry DEUTSCH
DIRECTEUR DE RECHERCHE, CEA, INAC, Invité
M. Laurent DUVAL
INGÉNIEUR DE RECHERCHE, IFPEN, Invité

Contents

Acknowledgements

First of all, I would like to thank my supervisors for their constant support. In particular, thanks to Valérie Perrier for being my thesis director and for the knowledge and counsels that she has brought. I want to express my sincere gratitude to Quang Huy Tran who was my co-director during these three years in IFP Energies nouvelles; I cannot count all the things I have learned with him and from him during this PhD. I am grateful to Luigi Genovese, without his guidance the thesis would not have been well finished.

A big thank-you goes to Laurent Duval for his timely support and encouragement. I also appreciate the help from Pascal Duchêne who provided us great assistance on the numerical front. I am deeply grateful to Zakia Benjelloun-Touimi for welcoming me into the Applied Mathematics Department.

I would like to thank Yvon Maday and Silvia Bertoluzza for reviewing this thesis and, with their comments, for helping me to improve it. Thanks also to Stéphane Labbé, Patrick Fischer, Christophe Morrel and Thierry Deutsch for participating in the jury.

I would take advantage of this forefront page to express my gratitude to all those of IFPEN with whom I shared these years. In particular, thanks to Riad, Nicolas and Adrien for all the helps and laughters. Although I wasted some precious time of yours, that time will be treasured in my mind; I wish the best for your remaining work in the department. My thank also goes to Thibaut, Frank and Aboubacar, to the doctoral students and internship students who have created a very good atmosphere for communication in IFPEN.

I am thankful to my friends who are always there for me when I need them. To Quynh Mai, Thu Trang, anh Chuong, Austin, and to Nga who is not there anymore, your friendships mean a lot to me.

Finally, a special thank to my parents and my brother who have been supporting me and advising me, especially in difficult times. Your love and understanding is the greatest companion in my journey of life.

Résumé

Cette thèse apporte une contribution aux méthodes numériques pour la simulation moléculaire *ab initio*, et plus spécifiquement pour le calcul de structures électroniques par l'équation de Schrödinger, ou par des formalismes comme la théorie de Hartree-Fock ou la théorie de la fonctionnelle de la densité. Elle propose une stratégie pour construire des bases mixtes ondelettes-gaussiennes dans l'approximation de Galerkin, combinant les qualités respectives de ces deux types de bases avec l'objectif de mieux capturer les points de rebroussement de la fonction d'onde.

Les nombreux logiciels actuellement disponibles à l'usage des chimistes dans ce domaine (VASP, Gaussian, ABINIT...) se différencient par divers choix méthodologiques, notamment celui des fonctions de base pour exprimer les orbitales atomiques. Nouvel arrivant sur le marché, le code massivement parallèle BigDFT a opté pour les bases d'ondelettes. Comme le nombre de niveaux de résolution y est limité pour des raisons de performance, ces fonctions ne peuvent déployer pleinement leur puissance. La question posée est alors de savoir comment accroître la précision des calculs *tous électrons* au voisinage des singularités de type *cusp* de la solution, sans augmenter excessivement la complexité de BigDFT.

La réponse que nous suggérons consiste à enrichir la base de fonctions d'échelles (niveau de résolution bas des bases d'ondelettes) par des fonctions gaussiennes centrées sur chaque position de noyau. La difficulté principale dans la construction d'une telle base mixte réside dans la détermination optimale du nombre de gaussiennes requises et de leurs écarts-types, de sorte que ces gaussiennes supplémentaires soient le mieux possible compatibles avec la base existante sous la contrainte d'un seuil d'erreur donné à l'avance. Nous proposons pour cela l'utilisation conjointe d'un estimateur *a posteriori* sur la diminution du niveau d'énergie et d'un algorithme glouton, ce qui aboutit à une suite incrémentale quasi-optimale de gaussiennes supplémentaires. Cette idée est directement inspirée des techniques de bases réduites.

Nous développons les fondements théoriques de cette stratégie sur deux modèles 1-D linéaires qui sont des simplifications de l'équation de Schrödinger pour un électron, posée en domaine infini ou domaine périodique. Ces modèles prototypes sont étudiés en profondeur dans la première partie. La définition de l'estimateur *a posteriori* de type norme duale du résidu, ainsi que la déclinaison de la philosophie glouton en différents algorithmes concrets, sont présentées en seconde partie, accompagnées de résultats numériques. Les algorithmes proposés vont dans le sens d'une économie croissante du temps de calcul. Ils sont aussi de plus en plus empiriques, au sens où ils reposent de plus en plus sur les intuitions avec lesquelles les chimistes sont familiers. En particulier, le dernier algorithme pour plusieurs noyaux s'appuie en partie sur la validité du transfert atome/molécule et rappelle dans une certaine mesure les bases d'orbitales atomiques.

Abstract

This thesis aims to be a contribution to numerical methods for *ab initio* molecular simulation, and more specifically for electronic structure calculations by means of the Schrödinger equation or formalisms such as the Hartree-Fock theory or the Density Functional Theory. It puts forward a strategy to build mixed wavelet-Gaussian bases for the Galerkin approximation, combining the respective advantages of these two types of bases in order to better capture the cusps of the wave function.

Numerous software programs are currently available to the chemists in this field (VASP, Gaussian, ABINIT...) and differ from each other by various methodological choices, notably that of the basis functions used for expressing atomic orbitals. As a newcomer to this market, the massively parallel BigDFT code has opted for a basis of wavelets. Due to performance considerations, the number of multiresolution levels has been limited and therefore users cannot benefit from the full potential of wavelets. The question is thus how to improve the accuracy of all-electron calculations in the neighborhood of the cusp-type singularities of the solution, without excessively increasing the complexity of BigDFT.

The answer we propose is to enrich the scaling function basis (low level of resolution of the wavelet basis) by Gaussian functions centered on each nucleus position. The main difficulty in constructing such a mixed basis lies in the optimal determination of the number of Gaussians required and their standard deviations, so that these additional Gaussians are compatible in the best possible way with the existing basis within the constraint of an error threshold given in advance. We advocate the conjunction of an *a posteriori* estimate on the diminution of the energy level and a greedy algorithm, which results in a quasi-optimal incremental sequence of additional Gaussians. This idea is directly inspired by the techniques of reduced bases.

We develop the theoretical foundations of this strategy on two 1-D linear models that are simplified versions of the Schrödinger equation for one electron in an infinite domain or a periodic domain. These prototype models are investigated in depth in the first part. The definition of the *a posteriori* estimate as a residual dual norm, as well as the implementation of the greedy philosophy into various concrete algorithms, are presented in the second part, along with extensive numerical results. These algorithms allow for more and more saving of CPU time and become more and more empirical, in the sense that they rely more and more on the intuitions with which chemists are familiar. In particular, the last proposed algorithm partly assumes the validity of the atom/molecule transfer and is somehow reminiscent of atomic orbitals bases.

List of principal notations

Chapter 1

In order to avoid conflict with chapters §§2–6, we have adopted symbols which could appear to be somewhat unconventional for chemists: u instead of ψ for the wave function, and φ instead of ϕ for the molecular orbitals.

u	wave function
m	number of electrons
i, j	indices of electrons
$\mathbf{x}_i$	position of the i-th electron in $\mathbb{R}^3$
M	number of nuclei
I	index of a nucleus
$\mathbf{X}_I$	position of the I-th nucleus in $\mathbb{R}^3$
Z_I	charge of the I-th nucleus
V	potential
$\mathcal{H}$	Hamiltonian operator
σ	spectrum of $\mathcal{H}$
$\mathfrak{E}$, E	energy functional and energy level
S	unit sphere in $\mathbb{R}^3$
ρ	density
$\mathbf{n}$	unit normal vector in $\mathbb{R}^3$
$\boldsymbol{\xi}$	wave vector in the Fourier space $\mathbb{R}^3$
i	imaginary number, $\mathrm{i}^2 = -1$
$\widehat{\psi}$	Fourier transform of ψ, with $\widehat{\psi}(\boldsymbol{\xi}) = \int_{\mathbb{R}^3} \psi(\mathbf{x}) \exp(-\mathrm{i}\boldsymbol{\xi} \cdot \mathbf{x}) \, \mathrm{d}\mathbf{x}$
φ_i, Φ	molecular orbital and admissible set of molecular orbitals
J	electrostatic energy of a distribution of charge
$\mathcal{K}$	Kohn-Sham operator
$\mathbf{r}$, r	relative position $\mathbf{x} - \mathbf{X}$ and its Euclidean norm
k, ℓ, m	indices of atomic orbitals for a single atom
ω, χ	generic element of a basis
$\mathbf{S}$, $\mathsf{S}_{\mu\nu}$	matrix of recovery and its entries
$\mathbf{C}$, $\mathsf{C}_{\mu i}$	matrix of decomposition and its entries
χ^{S}, χ^{G}, χ^{CG}	Slater, Gaussian and contracted Gaussian functions
ζ, α	multiplicative exponents of Slater and Gaussian functions
q, Q	index and number of primitives for a contracted Gaussian
$\boldsymbol{\tau}$, τ^q	set of exponents for a contracted Gaussian and its elements
$\mathbf{v}$, v^q	set of coefficients for a contracted Gaussian and its elements

Chapter 2

From now on, m becomes a dummy integer subscript, while i remains the imaginary unit ($\mathrm{i}^2 = -1$) and ξ denotes the wave number in the Fourier space $\mathbb{R}$.

ϕ	father wavelet or scaling function
n	index of shifting
$\mathbf{h}$, h_n	low-pass filter of the Multiresolution Analysis
J	index of multiresolution level
$\mathscr{V}_\mathsf{J}$	subspace of $L^2(\mathbb{R})$, generated by the $\phi_{\mathsf{J},n}$, $n \in \mathbb{Z}$
P_J	orthogonal projection on $\mathscr{V}_\mathsf{J}$
ψ	mother wavelet
$\mathbf{g}$, g_n	high-pass filter of the Multiresolution Analysis
$\mathscr{W}_\mathsf{J}$	subspace of $L^2(\mathbb{R})$, generated by the $\psi_{\mathsf{J},n}$, $n \in \mathbb{Z}$
M	order of the wavelet, number of vanishing moments for ψ
$\widehat{\phi}$	Fourier transform of ϕ, with $\widehat{\phi}(\xi) = \int_{\mathbb{R}} \exp(-\mathrm{i}\xi x)\phi(x)\,\mathrm{d}x$
$\widehat{h}$	transfer function of $\mathbf{h}$, $\widehat{h}(\xi) = 2^{-1/2}\sum_{n\in\mathbb{Z}} h_n \exp(-\mathrm{i}n\xi)$
$\widehat{\psi}$	Fourier transform of ψ
$\widehat{g}$	transfer function of $\mathbf{g}$, $\widehat{g}(\xi) = 2^{-1/2}\sum_{n\in\mathbb{Z}} g_n \exp(-\mathrm{i}n\xi)$
$\mathsf{db*}$	Daubechies minimal phase wavelet with specified order
$\mathsf{sy*}$	Daubechies least asymmetric wavelet with specified order (Symmlet)
γ_k	autocorrelation coefficient $\sum_{n\in\mathbb{Z}} h_n h_{n+k}$
α	Hölder regularity exponent
s	Sobolev regularity exponent
$a^{\mathsf{J}}_{i,j}$	connection coefficients $\int_{\mathbb{R}} \phi'_{\mathsf{J},i}\,\phi'_{\mathsf{J},j}$
$\mathbf{a}$, a_k	connection coefficients $\int_{\mathbb{R}} \phi'\,\phi'(\cdot + k)$
$\boldsymbol{\mu}$, μ_ℓ	second-order moments, $\mu_\ell = \int_{\mathbb{R}} x^2\phi(x-\ell)\,\mathrm{d}x$,
$\widetilde{\phi}$	1-periodization of ϕ, with $\widetilde{\phi}(x) = \sum_{n\in\mathbb{Z}}\phi(x+n)$
$\widetilde{\mathscr{V}}_\mathsf{J}$	subspace of $L^2(0,1)$, generated by the $\widetilde{\phi}_{\mathsf{J},n}$, $n \in \{0,\dots,2^{\mathsf{J}}-1\}$
$\widetilde{P}_\mathsf{J}$	orthogonal projection on $\widetilde{\mathscr{V}}_\mathsf{J}$
$\widetilde{\psi}$	1-periodization of ψ
$\widetilde{\mathscr{W}}_\mathsf{J}$	subspace of $L^2(0,1)$, generated by the $\widetilde{\psi}_{\mathsf{J},n}$, $n \in \{0,\dots,2^{\mathsf{J}}-1\}$
$\widetilde{a}^{\mathsf{J}}_{i,j}$	connection coefficients $\int_0^1 \widetilde{\phi}'_{\mathsf{J},i}\,\widetilde{\phi}'_{\mathsf{J},j}$
$\lvert i-j\rvert^{\sim}$	periodized distance, $\lvert i-j\rvert^{\sim} = \min\{\lvert i-j\rvert,\, 2^{\mathsf{J}} - \lvert i-j\rvert\}$

Chapter 3

From this chapter to the end of the manuscript, J becomes a dummy integer subscript, $\mathbf{C}$ refers to another type of matrix.

u	wave function
x	position of the electron in $\mathbb{R}$ or $[0, L]$
M	number of nuclei
I, J	indices of nuclei
X_I	position of the I-th nucleus in $\mathbb{R}$ or $[0, L]$

Z_I	charge of the I-th nucleus
Z	greatest charge, $Z = \max_{1\leq I\leq M} Z_I$
$\mathcal{Z}$	sum of charges, $\mathcal{Z} = \sum_{I=1}^{M} Z_I$
δ_X	Dirac delta located at X
V	potential
$\mathscr{V}$	space for u, $H^1(\mathbb{R})$ or $H^1_\#(0,L)$
$\mathscr{S}$	L^2-unit sphere in $\mathscr{V}$
$\mathfrak{a}$, $\mathfrak{b}$	Hamiltonian and mass bilinear forms
$\mathfrak{E}$, E	energy functional and energy level
c	continuity constant for the embedding $H^1 \subset C^0$ in 1-D
κ	continuity constant for $\mathfrak{a}$
(u_*, E_*) or $(u^{(1)}, E^{(1)})$	ground state and fundamental energy
ρ	density of probability $\lvert u\rvert^2$
$\widehat{u}$	Fourier transform of u, with $\widehat{u}(\xi) = \int_{\mathbb{R}} u(x)\exp(-\mathrm{i}\xi x)\,\mathrm{d}x$
ζ	auxiliary unknown, from which $E = -\zeta^2/2$ is deduced
$\mathrm{S}_{\zeta,X}$	Slater function, $\mathrm{S}_{\zeta,X} = \exp(-\zeta\lvert\cdot - X\rvert)$
$\mathbf{C}^\zeta$, C^ζ_{IJ}	matrix of compatibility and its entries
$\mathbf{u}$, u_J	vector whose entries are values of u at X_J
L	size of the domain in the periodic model
Λ_I	characteristic length associated with the I-th nucleus, $\Lambda_I = Z_I^{-1}$
$\lvert y\rvert^\sim$	distance from y to the closest integer multiple of L
$\widetilde{f}$	L-periodization of f, with $\widetilde{f} = \sum_{n\in\mathbb{Z}} f(\cdot + nL)$
$\widetilde{\mathrm{S}}_{\zeta,X}$	L-periodized Slater, $\widetilde{\mathrm{S}}_{\zeta,X} = \cosh(\zeta(\lvert\cdot - X\rvert^\sim - L/2))/\sinh(\zeta L/2)$
$\widetilde{z}$	L-alteration of $z > 0$, with $\widetilde{z} = z\coth(\widetilde{z}L/2)$
$\widehat{u}_k$	Fourier coefficients, $\widehat{u}_k = L^{-1}\int_0^L u(x)\exp(-\mathrm{i}2\pi kx/L)\,\mathrm{d}x$
R	internuclear distance, $\lvert X_2 - X_1\rvert$ or $\lvert X_2 - X_1\rvert^\sim$
W	Lambert function
$(u_\sharp, E_\sharp)$ or $(u^{(2)}, E^{(2)})$	first excited state (if exists) and second eigenvalue

Chapter 4

From this chapter to the end of the manuscript, i and j are subscripts denoting the nodes of a uniform mesh.

M	number of nuclei
I, J	indices of nuclei
$\mathscr{V}$	space for u, $H^1(\mathbb{R})$ or $H^1_\#(0,L)$
(u_*, E_*)	exact solution
$\mathscr{V}_b$	finite-dimensional subspace of $\mathscr{V}$
(u_b, E_b)	Galerkin approximation on $\mathscr{V}_b$
q, Q	index and number of elements in a Gaussian basis
$\boldsymbol{\sigma}$, σ_q	set of standard deviations and its elements
g_σ, $g_{\sigma,X}$	Gaussian centered at 0 or X, with standard deviation σ
$\mathscr{V}_{\boldsymbol{\sigma}}$	subspace spanned by a Gaussian basis
$(u_{\boldsymbol{\sigma}}, E_{\boldsymbol{\sigma}})$	Galerkin approximation on $\mathscr{V}_{\boldsymbol{\sigma}}$
$\mathbf{A}^{\boldsymbol{\sigma}}$, $\mathbf{B}^{\boldsymbol{\sigma}}$	Hamiltonian and mass matrices in the Gaussian basis

$\mathbf{u}^{\boldsymbol{\sigma}}$, $\mathsf{u}_q^{\boldsymbol{\sigma}}$	coefficients in the decomposition on the Gaussian basis
$\mathrm{CG}(\boldsymbol{\sigma}, \mathbf{v}, \cdot)$	contracted Gaussian with standard deviations $\boldsymbol{\sigma}$, coefficients $\mathbf{v}$
$\boldsymbol{\sigma}^*$, $\mathbf{v}^*$	optimal parameters for the contracted Gaussian
$\mathtt{err}$	relative error on energy
R_λ	scaling operator, $R_\lambda u(y) = \lambda^{1/2} u(\lambda y)$
L	size of the domain in the periodic model
h	mesh size
N	number of nodes, a power of 2, i.e., $N = 2^{\mathsf{J}}$
i, j	indices of nodes
$\widetilde{\chi}_i^h$	L-periodized basis scaling functions
$\mathscr{V}_h$	subspace spanned by the $\widetilde{\chi}_i^h$'s
(u_h, E_h)	Galerkin approximation on $\mathscr{V}_h$
$\mathbf{u}^h$, u_i^h	coefficients in the decomposition on the scaling function basis
$\mathbf{A}^h$, $\mathbf{B}^h$	Hamiltonian and mass matrices in the scaling function basis
e	difference $u_h - u_*$
K	continuity constant for $\mathfrak{a} - E_*\mathfrak{b}$
β	L^2-coercivity constant for $\mathfrak{a} - E_*\mathfrak{b}$ on $(u_*)^\perp$
γ	H^1-coercivity constant for $\mathfrak{a} - E_*\mathfrak{b}$ on $\mathbb{R}e$
$\mathsf{db*}$	Daubechies minimal phase wavelet with specified order
$\mathsf{sy*}$	Daubechies least asymmetric wavelet with specified order (Symmlet)
CG_I	optimized contracted Gaussian centered at X_I for charge Z_I
$\widetilde{\mathrm{CG}}_I$	L-periodization of CG_I
$\mathscr{V}_{h,g}$	subspace spanned by the χ_i^h's and the $\widetilde{\mathrm{CG}}_I$
$(u_{h,g}, E_{h,g})$	Galerkin approximation on $\mathscr{V}_{h,g}$
$\mathbf{u}^{h,g}$, $\mathsf{u}_i^{h,g}$	coefficients in the decomposition on the mixed basis
$\mathbf{A}^{h,g}$, $\mathbf{B}^{h,g}$	Hamiltonian and mass matrices in the mixed basis
$\langle f, \phi \rangle$	exact scalar product
$\langle\!\langle f, \phi \rangle\!\rangle$	approximate scalar product
ω_ℓ	weights of the quadrature rule
Q	degree of exactness for the quadrature rule
T_θ	translation operator, $T_\theta u = u(\cdot - \theta)$
S_ϑ	dilation operator, $S_\vartheta u = u(\vartheta \cdot)$
I, J, K, L	indices of multiresolution levels
$\langle\!\langle f, \phi \rangle\!\rangle_{\mathsf{L}}$	L-quadrature rule using two-scale relation
$\mathrm{e}(f)$, $\mathrm{e}_{\mathsf{L}}(f)$	quadrature errors
$\mathfrak{K}$, $\mathfrak{K}_{\mathsf{L}}$	Peano kernels
$\Xi_{\mathsf{Q},\mathsf{M}}$	apparent reduction factor

Chapter 5

From this chapter to the end of the manuscript, the subscript b stands for the current basis, while b, g stands for the augmented basis.

$\widetilde{g}_{\sigma,X}$	L-periodized Gaussian centered at X with standard deviation σ
$\breve{g}_\sigma$	additional Gaussian function, centered at some unprecised X_I

$\mathscr{V}_b$	subspace associated with the current basis
N_b	dimension of $\mathscr{V}_b$
(u_b, E_b)	Galerkin approximation on $\mathscr{V}_b$
$\mathscr{V}_{b,g}$	subspace associated with the augmented basis
$N_b + N_g$	dimension of $\mathscr{V}_{b,g}$
$(u_{b,g}, E_{b,g})$	Galerkin approximation on $\mathscr{V}_{b,g}$
e	difference $u_b - u_{b,g}$
K	continuity constant for $\mathfrak{a} - E_{b,g}\mathfrak{b}$
β_b	L^2-coercivity constant for $\mathfrak{a} - E_{b,g}\mathfrak{b}$ on $(u_{b,g})^\perp$
γ_b	H^1-coercivity constant for $\mathfrak{a} - E_{b,g}\mathfrak{b}$ on $\mathbb{R}e$
$\eta^2_{b,g}$	*a posteriori* estimate for $E_b - E_{b,g}$
$\mathbf{M}$	Gram matrix for a the (modified) H^1-norm
$\mathbf{r}$	residue vector
$(\!(\cdot,\cdot)\!)_\varepsilon$	modified H^1-scalar product
$\|\!\|\!\|\cdot\|\!\|\!\|_\varepsilon$	modified H^1-norm

Chapter 6

From this chapter to the end of the manuscript, the integer Q refers to the number of Gaussian functions in the basis, either pure Gaussian or mixed basis.

u	wave function
g_σ, $g_{\sigma,X}$	Gaussian centered at 0 or X, with standard deviation σ
L	size of the domain in the periodic model
$\widetilde{g}_{\sigma,X}$	L-periodized Gaussian from $g_{\sigma,X}$
q	index of Gaussians in the basis
σ_q, σ_{q*}	standard deviation of the qth Gaussian and its optimal value
$\boldsymbol{\sigma}$	set of standard deviations
$\mathfrak{E}$, E	energy functional and energy level
$\mathscr{V}$	space for u, $H^1(\mathbb{R})$ or $H^1_\#(0,L)$
(u_*, E_*)	exact solution
$\mathscr{V}_b$	subspace associated with the current basis
(u_b, E_b)	Galerkin approximation on $\mathscr{V}_b$
$\mathscr{V}_{b,g}$	subspace associated with the augmented basis
$(u_{b,g}, E_{b,g})$	Galerkin approximation on $\mathscr{V}_{b,g}$
$\eta^2_{b,g}$	*a posteriori* estimate for $E_b - E_{b,g}$
$\mathscr{V}_{\boldsymbol{\sigma}}, \mathscr{V}_2, \mathscr{V}_3$	subspaces spanned by Gaussian bases of dimension $Q, 2$ or 3
E_2, E_3	energy levels on $\mathscr{V}_2, \mathscr{V}_3$
(u_{1*}, E_{1*})	approximation on optimal Gaussian basis of dimension 1
(u_{2*}, E_{2*})	approximation on optimal Gaussian basis of dimension 2
$\mathbf{A}^{\boldsymbol{\sigma}}$, $\mathbf{B}^{\boldsymbol{\sigma}}$	Hamiltonian and mass matrices in the Gaussian basis
u_q	coefficients in the decomposition on the Gaussian basis
$\mathfrak{a}$, $\mathfrak{b}$	Hamiltonian and mass bilinear forms
$(\!(\cdot,\cdot)\!)$	modified H^1-scalar product
$\mathbf{M}$	Gram matrix for a the (modified) H^1-norm
r	dilation of a standard deviation

M	number of nuclei
I, J	indices of nuclei
X_I	position of the I-th nucleus in $\mathbb{R}$ or $[0, L]$
Z_I	charge of the I-th nucleus
δ_X	Dirac delta located at X
h	mesh size
N	number of nodes, a power of 2, i.e., $N = 2^J$
i, j	indices of nodes
$\widetilde{\chi}_i^h$	L-periodized basis scaling functions
$\mathscr{V}_h$	subspace spanned by the $\widetilde{\chi}_i^h$'s
(u_h, E_h)	Galerkin approximation on $\mathscr{V}_h$
$\mathscr{V}_{h,g_\sigma}$	subspace spanned by the mixed basis of the $\widetilde{\chi}_i^h$'s and g_σ
$\texttt{err}, \texttt{Err}$	relative differences between energy levels
err	relative difference of some quantity
$r_*^{\mathfrak{E}}, r_*^{\eta}$	optimal dilation by the energy criterion or the $\eta_{b,g}$ criterion
$E^{\mathfrak{E}}, E^{\eta}$	energy level obtained by using $r_*^{\mathfrak{E}}$ or r_*^{η}
$t^{\mathfrak{E}}, t^{\eta}$	time (in s) to obtain $r_*^{\mathfrak{E}}$ or r_*^{η}.

Introduction

Basis sets in quantum chemistry softwares

Ab initio molecular simulations aim at studying matter at subatomic scales by means of quantum models considered to be "fundamental," in contrast to those qualified as "empirical." The so-called fundamental models are derived from the Schrödinger equation via some additional formalisms, such as the Hartree-Fock or post-Hartree-Fock approximations [19, 22], the Density Functional Theory [25, 113]. Such calculations allow users to determine the electronic structure of the chemical system at hand, from which relevant macroscopic properties can be inferred. At IFPEN, *ab initio* computations are most widely used in catalysis [31, 46].

There are more than 70 software solutions in this area[1]; a few of the most common are pictorialized in Figure 1. A key difference between them lies in the basis functions selected to express the molecular orbitals, as the numerical discretization of the models relies on the Galerkin method. The choice of a basis set cannot be arbitrary. It has a direct impact on the validity range of the calculations, as well as on the computational cost.

VASP [84], ABINIT [67], CASTEP, Quantum ESPRESSO... use plane waves (PW), the idea of which will be explained in §1.4.1. PW basis sets are suitable for periodic and homogeneous systems that arise in solid state/crystal phase calculations. Augmented plane waves (APW) and Linearized augmented plane waves (LAPW) are methods that locally modify the PW basis functions in order to better manage the core regions (i.e., the neighborhood of a nucleus). LAPW basis sets are implemented for example in EXCITING, FLEUR, WIEN2k...

Gaussian [69], CP2K, MONSTERGAUSS, Q-Chem... opt for Gaussian Type Orbitals (GTO) [70], the construction of which will be sketched out in §1.3.2. Gaussian functions play a major role in quantum chemistry thanks to their efficiency for numerical approximation [14]. GTO basis sets are more natural for isolated or heterogeneous systems. Along with Contracted-Gaussian Type Orbitals (CGTO), their economical variants, they offer a possible solution for a better handling of the core regions. CGTO basis sets optimally designed for each atom are recorded in libraries and are part of the chemist's know-how.

CASINO, CONQUEST, Octopus, RMG... involve basis sets associated with a grid in real space, such as splines or sinc functions. The space localization of these functions gives one the hope that linear scaling algorithms could be designed for the problem [13, 63]. On the grounds of strong technical features elaborated on in §1.4.2 and §2.3, wavelets stand out as the most promising basis set among grid-based functions. They are implemented in BigDFT [60], DFT++, MADNESS...

[1] `http://en.wikipedia.org/wiki/List_of_quantum_chemistry_and_solid-state_physics_software`

Figure 1: A few quantum chemistry software packages.

BigDFT is a consortium[2] launched in 2005 with four partners: Laboratoire de Simulation Atomistique (CEA-INAC Grenoble), Universität Basel, Université catholique de Louvain, and Christian-Albrechts-Universität zu Kiel. Its initial mission was to develop an *ab initio* DFT code based on Daubechies wavelets to be integrated in ABINIT. The feasibility of Daubechies wavelets for DFT calculations had been demonstrated earlier in the doctoral work of Chauvin [27]. More than a DFT adventure, this code turned out to be the ideal case study for a number of scientific and computational questions. Among these, a special emphasis was laid on the optimal implementation of advanced DFT functionalities in High Performance Computing (HPC) frameworks. The BigDFT team was awarded the 2009 Bull-Joseph Fourier Prize for the parallelization of their software on hybrid CPU/GPU environments [61].

Accuracy of nuclear cusps in all-electron calculations

Wavelet analysis comes within an infinite multiresolution ladder, which allows one to add as many resolution levels as necessary. For the time being, though, only two resolution levels are implemented in BigDFT. This self-imposed limitation is intended to preserve its performance on large systems. As paradoxical as it may seem, two levels are indeed sufficient for most simulations in practice, where only *valence* electrons are of interest and where the effects of *core* electrons can be modeled by appropriate *pseudopotentials* in a sense that will be clarified in §1.4.1. For full-potential or all-electron calculations, which

[2] `http://bigdft.org`

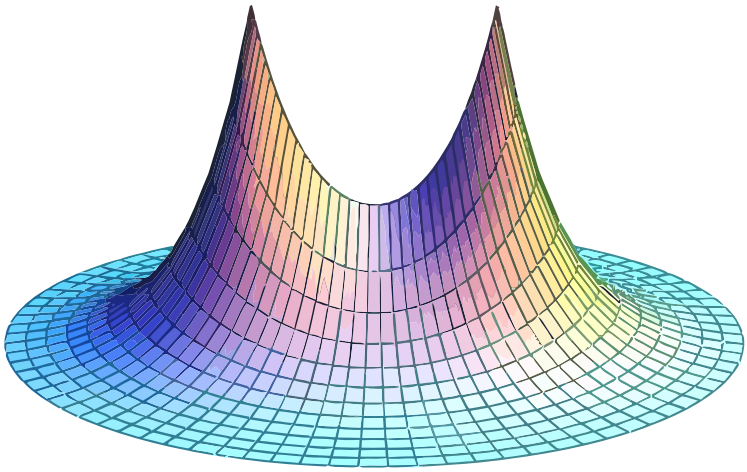

Figure 2: Cusp behavior exhibited by the ground state solution of the H_2^+ ion in 2-D.

require an increased accuracy in the neighbourhood of nuclei, two levels are obviously not sufficient. This thesis is concerned with a remedy for this shortcoming in the least expensive way possible.

It is established that at a nucleus position, the electronic wave function (for the Schrödinger setting) or each molecular orbital (for Hartree-Fock or DFT settings) exhibits a cusp singularity (as illustrated in Figure 2), at which the gradient is not well-defined but directional derivatives exist and satisfy the Kato condition [79]. As recalled in §1.1.2 and §1.2.2, this behavior drastically reduces the Sobolev regularity of the solution and clearly prevents convergence of numerical methods from being of high-order. To capture the singularity with enough accuracy, an effort must be expounded so as to insert a high-frequency content into the approximate solution.

There are many ways to do this. Perhaps the most natural one is to stick with the adaptivity of the multiresolution analysis while trying to further reduce the computational cost. This is the path followed by Harrison *et al.* [68] who take advantage of some non standard forms introduced by Beylkin and his co-authors [11,12] for separately representing operators. While the gain in accuracy is deemed satisfactory and the CPU time reasonable, this approach (see also Nagy and Pipek [109] for a review of similar ones) still has to cope with a large number of degrees of freedom associated with the extra wavelets.

When the grid is not subject to the constraint of uniform size, more sophisticated and presumably more efficient adaptivity strategies exist in the literature. To name a few: h adaptivity by Bao *et al.* [8], $h-P$ adaptivity by Maday [96], mesh redistribution adaptivity by Bao *et al.* [9], optimally convergent (with respect to the number of degrees of freedom) adaptivity for singular elliptic problems by Hunsicker *et al.* [75] and Li and Nistor [90], *a priori* error estimates-based adaptivity by Levine and Wilkins [89] and Motamarri et al. [107], *a posteriori* error estimates- based adaptivity by Dai *et al.* [41], Zhang *et al.* [137], Chen *et al.* [30] and Chen *et al.* [28]. Unfortunately, at least in the most common usage, the wavelet setting assumes that the mesh size is uniform, which rules out a direct application of all these advances. Despite this obstacle, let us keep in mind the idea of error estimates

for future use.

Another approach for improving the accuracy of the core region stems from a very simple intuition: take as a basis element some function that looks more or less like a cusp. Arguably, this already holds true for Atomic Orbital (AO) basis sets, of which GTO and CGTO are special instances, and is not in itself a novelty. But our hope here is that over a large spatial region, the combination of contracted Gaussians and wavelets would recover accuracy everywhere while being less expensive: wavelets would take care of smooth regions, while contracted Gaussians would deal with nuclear cusps.

The idea of mixed bases is not new either. In the thirties, the APW (augmented plane wave) method was proposed [126] in which the plane wave functions are hybridized with atomic orbitals in the core region (see §1.4.1 for more details). In the seventies and nineties, mixed basis sets with Gaussians around the nuclei and plane waves in the interstitial regions were successively investigated by Euwema [52], Lippert *et al.* [94] and VandeVondele *et al.* [131] in an attempt to correct the spatial delocalization of plane waves. More recently, Longo [95] suggested that a wavelet basis could be enriched by a small number of (contracted) Gaussians centered on each nucleus. It is mainly expected that, for a given accuracy threshold, the number of additional degrees of freedom required is much lower than in other approaches.

When mixed bases meet *a posteriori* estimates

In Longo's thesis [95], the determination of adequate or near-optimal parameters (namely, standard deviations) for Gaussians turned out to be difficult and could not be formulated in a systematic way, even though his results provide valuable insights into the nature of the difficulties. In the present thesis, our objective is to push one step further the mixed basis approach by providing a mathematically sound and computationally effective answer to the question of constructing the additional Gaussians.

The philosophy underlying our contribution, described with more details in §5, can be summarized as follows:

1. We work out a criterion that enables us to compare two approximate solutions without the knowledge of the exact solution. Since both approximate solutions are Galerkin approximations, they obey an energy minimizing principle, and therefore the natural criterion is the approximate energy level. When a new function is added into an existing basis, the approximate energy level can only decrease.

2. Given an approximate solution on a pure wavelet basis, the ideal parameters for the additional Gaussians are those that maximize the (positive) decay of the approximate energy level. Of course, it is out of reach for us to perform this maximization problem, insofar as it requires an eigenvalue problem to be solved in a mixed basis for each tentative value of the parameters.

3. But it is legitimate to replace the exact energy decay by an *a posteriori* estimate that remains to be devised. The advantage of such an estimate lies in the fact that it is explicitly computable without requiring the knowledge of the mixed basis approximate solution.

4. We can further save computational cost by means of a *greedy* algorithm, in which the original many-variable maximization problem is replaced by an incremental sequence

of one-variable maximization problems. The combination of *a posteriori* estimates and the greedy algorithm is reminiscent of reduced basis methods [71, 117], except for the fact that we are enlarging a basis instead of reducing it!

Error estimates have emerged as a major trend in various applications where a reliable tool is needed to monitor adaptive refinement or coarsening algorithms. The hard part is to design a good estimate for the PDE system at hand. In computational chemistry, *a priori* error estimates were first proposed by Zhou [139, 140] for the Gross-Pitaevskii and Thomas-Fermi models, followed by Chen *et al.* [29], Langwallner *et al.* [87], Cancès *et al.* [17, 18] for the Hartree-Fock and Kohn-Sham model. *A posteriori* error estimates were first proposed by Maday and Turinici [99] for the Hartree-Fock model, followed by Chen and his co-authors [28, 30] for the Kohn-Sham model and Dusson and Maday [50] for Gross-Pitaevskii model. An overview of this area is supplied in Maday's encyclopedia articles [97, 98].

The type of *a posteriori* estimate that we seek for the mixed basis problem is different from those quoted above. Indeed, step 3 of our battle plan requires an estimate for the energy decay between two (finite dimensional) approximate solutions[3], and not between a (finite dimensional) approximate solution and an (infinite dimensional) exact solution. As is well known in reduced basis techniques, the finite dimensionality of both solutions makes it possible to exactly compute the residual norm. It then only remains to establish that the residual norm is indeed related to the desired energy decay.

As a proof of concept for the above strategy, we have deliberately restricted ourselves to a physically significant but mathematically simplified Schrödinger model. Our toy model represents a single-electron (instead of many-electron) system evolving in a one-dimensional (instead of a three-dimensional) domain and subject to a multi-delta potential (instead of a multi-Coulomb potential). The simplicity of this toy model, from the theoretical and practical viewpoints, enables us to be more focused on the issue of mixed bases and to quickly test new paradigms for the choice of the additional Gaussians.

Outline of this thesis

Chapter 1

We begin, in chapter §1, by recalling some basic notions of *ab initio* simulations which are useful to contextualize this thesis. The first two sections of this chapter give an overview of two basic models in quantum chemistry, namely, the Schrödinger model in the Born-Oppenheimer approximation and the Kohn-Sham model in the Density Functional Theory. The objective of these two sections is to introduce the reader to the type of problems encountered in this area, with a particular emphasis on the questions of regularity and singularity for the wave function or the orbitals at the nuclear coalescence points.

The last two sections are devoted to the description of the standard categories of basis sets commonly used by chemists. In the first category, called atomic orbitals, we insist on the historical importance of Gaussian functions, as well as that of contracted Gaussians for the approximation of cusp points. The second category concerns planes waves, possibly "augmented" in the neighborhood of the nuclei. The third one includes bases that are

[3] In the definition of the estimate, the mixed basis solution will be considered as the "reference" solution, while the pure scaling functions basis solution will be considered as the "approximate" solution.

associated with a spatial mesh, in particular wavelets. The specific motivation for using wavelet bases is explained in §1.4.2.

Chapter 2

We then revisit, in chapter §2, those concepts of wavelets and multiresolution analysis that will be necessary for the discretization of PDEs in the next chapters. To the detriment of other frameworks such as semi-orthogonal, shift-orthogonal or biorthogonal, our exposition in the first section is restricted to the orthonormal one, since we wish to arrive quickly to the Daubechies wavelets that are implemented in BigDFT.

A special attention is paid in the second section to the properties and the algorithms related to the discretization of the partial differential equations of interest to us: the approximation property in the sense of the L^2 projection, the evaluation of scaling function values at a given point (dyadic or not) and the calculation of the "connection coefficients" which are the scalar products of two first derivatives of these functions. For the treatment of periodic boundary conditions, we also present the construction of periodic wavelets and its consequences on the previously mentioned algorithms. The last section briefly enumerates the implications of the Density Functional Theory's three-dimensional aspect on the use of wavelets that are one-dimensional by nature.

Chapter 3

The reader already familiar with the materials of chapters §1 or §2 may go directly to chapter §3, where we introduce two one-dimensional models that lie at the heart of our work: one in an infinite domain, the other in a periodic domain. These models faithfully reproduce the cusp behavior and result from simplifying the 3-D linear Schrödinger equation for one electron. Their advantage lies in the ease of implementation, which allows us to focus on the cusp issue.

The cusp is created by a Dirac delta potential, which is the 1-D counterpart of the 2-D and 3-D Coulomb potential. This idea dates back to Frost [57] for systems with one or two atoms and sporadically emerges in the literature, but to our knowledge has never been fully investigated from a mathematical perspective. This is why we undertake a thorough analysis of the models for an arbitrary number of nuclei: existence and uniqueness of the ground state, regularity of the wave function, bounds on the fundamental energy, exact solutions for single-delta and double-delta potentials... The knowledge of exact solutions by analytical or semi-analytic formulae makes it easier to study various approximation errors.

Chapter 4

As a preliminary exercice, we apply the Ritz-Galerkin method to the two 1-D models introduced in chapter §3. For the infinite-domain model, we propose to use the energy criterion to optimize the pure Gaussians basis associated with an isolated atom. This gives rise to a way of constructing contracted Gaussians that, at first sight, does not have good performance in practice because of the many-variable optimization problems involved. This approach will be reconsidered and improved in chapter §6 in conjunction with the greedy algorithm and the *a posteriori* estimate designed in chapter §5.

For the periodic-domain model, we derive the theoretical order of convergence of the method in pure scaling functions bases. The result is corroborated by numerical experiments for single-delta and double-delta potentials. Next, we show some simulations in mixed bases, in which the contracted Gaussians previously constructed are directly plugged into the basis without any adaptation to the existence scaling function basis. This procedure, called "pre-optimized" mixed basis, is of course not optimal but is a first attempt. Again, it will be reconsidered and improved in the forthcoming chapters.

Chapter 5

In chapter §5, we address the problem of enriching a scaling function basis by Gaussians. To optimize the construction of additional functions, we rely on a combination of *a posteriori* estimates and the greedy algorithm, in accordance with the general philosophy described earlier. From the discrete variational formulation we define a residue, the dual norm of which is shown to be effective and computable estimate for the energy decrease when the basis is augmented. To this end, we borrow and adapt some ideas from [50]. Furthermore, in order to increase the efficiency of the construction of the mixed basis, we recommend the greedy algorithm for building an incremental sequence of additional Gaussian functions. This makes our strategy look like the "dual" counterpart of reduced-basis techniques [116].

We finally come up with two algorithms, which coincide for a single nucleus but which differ from each other for a system with many nuclei. In the first one, the greedy algorithm dictates the order of the nuclei where some action must be taken. In the second one, we prescribe this sequence by visiting the nuclei in the decreasing order of charges.

Chapter 6

Extensive simulations are carried out in order to test the strategy proposed in chapter §5. As in chapter §4, the first section is devoted to the infinite-domain model with a single-delta potential in a pure Gaussians basis. This gives us the opportunity to resume the construction of contracted Gaussians left aside in §4.2.2 under a new vision.

In the second section, we analyze step by step the behavior of the two algorithms of chapter §5 for the periodic-domain model in a mixed basis. The outcome of this analysis is a third algorithm, which is a little more empirical but much more economical. This third algorithm has the flavor of atomic orbitals, since it relies on the transfer from an atom to a molecule of contracted Gaussians built with *a posteriori* estimates in the presence of an existing scaling functions basis. The limit of validity for this algorithm appears when two nuclei are too close from each other. In such a case, we switch back to the energy estimates in order to determine the additional Gaussians. At the end of the chapter, we give some thoughts to what remains to be done for three- or more-delta potentials.

Chapter 1

A brief survey of quantum chemistry

Contents

Dans cette synthèse bibliographique, nous présentons succinctement les notions essentielles en simulation moléculaire ab initio *qui sont utiles pour contextualiser cette thèse.*

Nous commençons par fournir un aperçu de deux modèles de base en chimie quantique : le modèle de Schrödinger dans l'approximation de Born-Oppenheimer et le modèle de Kohn-Sham dans la Théorie de la Fonctionnelle de Densité. L'objectif de ce survol est d'introduire le lecteur au type de problèmes rencontrés dans ce domaine, un accent particulier étant mis sur les questions de régularité et de singularité de la fonction d'onde ou des orbitales aux points de coalescence nucléaire.

Nous consacrons les deux dernières sections à la description des fonctions de base habituellement employées par les chimistes. Dans la première catégorie de bases, dite d'orbitales atomiques, nous insistons sur l'importance historique des gaussiennes, puis celle des gaussiennes contractées pour l'approximation des points de rebroussement. La deuxième catégorie concerne les ondes planes, éventuellement "augmentées" au voisinage des noyaux. La troisième regroupe les bases associées à un maillage spatial, en particulier les ondelettes.

1.1 The Schrödinger model

1.1.1 Physical backgrounds

The state of a system of $m \in \mathbb{N}^*$ non-relativistic electrons without spin is fully described by a time-independent wave function[1]

$$u(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) \in \mathbb{C}, \tag{1.1}$$

where the arguments $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) \in (\mathbb{R}^3)^m$ are the spatial positions likely to be occupied by the electrons. This wave function u is subjected to two conditions, namely

1. The property of *normalization*: as the modulus square $|u(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)|^2$ represents a probability density of presence, we must have

$$\int_{\mathbb{R}^{3m}} |u(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)|^2 \, \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \ldots \mathrm{d}\mathbf{x}_m = 1. \tag{1.2}$$

2. The property of *indistinguishability*: the electrons being identical *fermions*[2], it requires the antisymmetry relationship

$$u(\mathbf{x}_{\varsigma(1)}, \mathbf{x}_{\varsigma(2)}, \ldots, \mathbf{x}_{\varsigma(m)}) = \epsilon(\varsigma) u(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) \tag{1.3}$$

for every permutation ς of the subscripts $\{1, 2, \ldots, m\}$, of sign $\epsilon(\varsigma)$. It follows from this antisymmetry property the *Pauli exclusion principle*, which states that two electrons cannot occupy the same position.

The conditions (1.2)–(1.3) imply that u must belong to the antisymmetric tensor product

$$\mathscr{H}^0_{m,\mathbb{C}} = \bigwedge_{i=1}^{m} L^2(\mathbb{R}^3; \mathbb{C}), \tag{1.4}$$

which is the antisymmetric subspace of the full tensor product Hilbert space

$$\bigotimes_{i=1}^{m} L^2(\mathbb{R}^3; \mathbb{C}) = L^2(\mathbb{R}^{3m}, \mathbb{C}). \tag{1.5}$$

Among the wave functions satisfying (1.2)–(1.3), only those solutions of the eigenvalue problem

$$\mathcal{H}u = Eu, \tag{1.6}$$

called stationary ***Schrödinger equation***, are likely to characterize the state of the system. The operator $\mathcal{H}$ in the left-hand side of (1.6) is the ***electronic Hamiltonian***, defined by

$$\mathcal{H} = -\frac{1}{2} \sum_{i=1}^{m} \Delta_{\mathbf{x}_i} + \sum_{i=1}^{m} V(\mathbf{x}_i) + \sum_{1 \le i < j \le m} \frac{1}{|\mathbf{x}_j - \mathbf{x}_i|}, \tag{1.7}$$

where the potential

$$V(\mathbf{x}) = -\sum_{I=1}^{M} \frac{Z_I}{|\mathbf{x} - \mathbf{X}_I|} \tag{1.8}$$

[1] We use the symbol u instead of ψ, which denotes the mother wavelet in the subsequent chapters.

[2] As opposed to *bosons*, whose wave function is symmetric.

reflects the *attraction* pulled by M nuclei[3], of charges $(Z_1, Z_2, \dots, Z_M) \in (\mathbb{R}_+^*)^M$ and of assumed known positions $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M) \in (\mathbb{R}^3)^M$. The last term of (1.7) accounts for the *repulsion* between the electrons. The attraction and repulsion potentials act as multiplication operators. Meanwhile, the first term of the right-hand side of (1.7), called ***kinetic*** term, is up to a factor a sum of Laplacians with respect to each variable $\mathbf{x}_i$.

REMARK 1.1. In expressing the model (1.7)–(1.8), we have implicitly used the Hartree system of atomic units, in which the mass of the electron, the elementary charge, the Planck reduced constant and the Coulomb electric constant are all equal to 1.

REMARK 1.2. Strictly speaking, the electronic Hamiltonian (1.7) should be denoted

$$\mathcal{H}^{Z_1,Z_2,\dots,Z_M}_{\mathbf{X}_1,\mathbf{X}_2,\dots,\mathbf{X}_M}$$

to highlight its dependence on the parameters of the given *nuclear configuration*.

REMARK 1.3. Strictly speaking, what we have introduced so far is not the full Schrödinger model, but only the electronic problem that arises from the *Born-Oppenheimer approximation*. The idea of this approximation is to separate the effects of electrons and nuclei, which leads to omitting the nuclei repulsive interaction in the Hamiltonian. To determine the position of the nuclei, an outer minimization loop using classical mechanics is required [22, appendix A].

Spectral theory viewpoint

The Schrödinger equation (1.6) can be investigated from two points of view. First, it can be interpreted in light of the spectral properties of $\mathcal{H}$. The domain of this operator, i.e., the set of the u such that the image $\mathcal{H}u$ is well-defined and belongs to $\mathscr{H}^0_{m,\mathbb{C}}$ is the space

$$\mathscr{H}^2_{m,\mathbb{C}} = \bigwedge_{i=1}^{m} H^2(\mathbb{R}^3;\mathbb{C}). \tag{1.9}$$

This space coincides with $\mathscr{H}^0_{m,\mathbb{C}} \cap H^2(\mathbb{R}^{3m};\mathbb{C})$. Below is an overview of what is known about $\mathcal{H}$ from the standpoint of spectral theory, eluding many mathematical subtleties for which we refer the readers to [22, 23, 80, 120].

- It is proven in [78] that the operator $\mathcal{H}$ is self-adjoint on its domain. Consequently, its spectrum is contained in the real line: $\sigma(\mathcal{H}) \subset \mathbb{R}$. Recall that this spectrum can be divided into two separate parts [22, appendix B], which are
  - The *discrete spectrum* $\sigma_{\text{dis}}(\mathcal{H})$, containing the isolated eigenvalues of finite multiplicity ;
  - The *essential spectrum* $\sigma_{\text{ess}}(\mathcal{H})$, containing the non-isolated or of infinite multiplicity eigenvalues, as well as the *continuous spectrum*.
- It is demonstrated in [78] that the operator $\mathcal{H}$ is bounded from below, i.e. $\mathcal{H} \geq \varrho\mathcal{I}$ for some $\varrho \in \mathbb{R}$. Therefore, it makes sense to bring in the quantity

$$E_1 := \inf \sigma(\mathcal{H}) = \inf_{\substack{u \in \mathscr{H}^2_{m,\mathbb{C}} \\ \|u\|_{L^2(\mathbb{R}^{3m};\mathbb{C})}=1}} \langle \mathcal{H}u, u\rangle_{L^2}, \tag{1.10}$$

  called ***ground state energy*** of the system. Two cases must then be distinguished:

[3]In this modelling, a *nucleus* is a fictitious particle containing all protons and neutrons of an atom.

– If $E_1 \in \sigma_{\text{dis}}(\mathcal{H})$, then E_1 also appears to be the smallest eigenvalue of $\mathcal{H}$, and the eigenfunction u_1 corresponding to E_1 is called the ***ground state*** of the system. The eigenfunctions corresponding to other elements of the discrete spectrum are called ***excited state***. This favorable case occurs when there are "few" electrons, for example [138] when

$$m < \sum_{I=1}^{M} Z_I + 1. \tag{1.11}$$

– If $E_1 \in \sigma_{\text{ess}}(\mathcal{H})$, then there is no ground state and the system is *unstable*. This unfavorable case occurs when there are "too many" electrons, for example [124] when $m > m_{\text{c}}$ for some critical threshold m_{c}. An estimation of this threshold is provided by [93]

$$m_{\text{c}} \leq 2 \sum_{I=1}^{M} Z_I + M. \tag{1.12}$$

Energy viewpoint

The Schrödinger equation (1.6) can also be thought of as the characterization of critical points relative to some energy functional. This second perspective will be most relevant for us, insofar as it lays foundation to Galerkin approximation. Indeed, the product $\mathfrak{E}(u) := \langle \mathcal{H}u, u \rangle_{L^2}$ is rewritten in the symmetric form

$$\begin{aligned} \mathfrak{E}(u) = \frac{1}{2} \sum_{i=1}^{m} \int_{\mathbb{R}^{3m}} |\nabla_{\mathbf{x}_i} u(\mathbf{x}_1, \dots, \mathbf{x}_m)|^2 \, \mathrm{d}\mathbf{x}_1 \dots \mathrm{d}\mathbf{x}_m \\ + \sum_{i=1}^{m} \int_{\mathbb{R}^{3m}} V(\mathbf{x}_i) |u|^2 \, \mathrm{d}\mathbf{x}_1 \dots \mathrm{d}\mathbf{x}_m + \sum_{1 \leq i < j \leq m} \int_{\mathbb{R}^{3m}} \frac{|u(\mathbf{x}_1, \dots, \mathbf{x}_m)|^2}{|\mathbf{x}_j - \mathbf{x}_i|} \, \mathrm{d}\mathbf{x}_1 \dots \mathrm{d}\mathbf{x}_m. \end{aligned} \tag{1.13}$$

From Hardy's inequality

$$\int_{\mathbb{R}^3} \frac{|u(\mathbf{x})|^2}{|\mathbf{x}|^2} \, \mathrm{d}\mathbf{x} \leq 4 \int_{\mathbb{R}^3} |\nabla u(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x}$$

for all $u \in H^1(\mathbb{R}^3)$, it is proven that each term of the ***energy functional*** $\mathfrak{E}(u)$ is well-defined since u belongs to

$$\mathscr{H}^1_{m,\mathbb{C}} = \bigwedge_{i=1}^{m} H^1(\mathbb{R}^3; \mathbb{C}). \tag{1.14}$$

This space coincides with $\mathscr{H}^0_{m,\mathbb{C}} \cap H^1(\mathbb{R}^{3m}; \mathbb{C})$; the ground state can therefore be construed as a minimizer of $\mathfrak{E}(u)$ in the intermediate space $\mathscr{H}^1_{m,\mathbb{C}}$. In fact, the space of "good" wave functions can be further reduced to

$$\mathscr{H}^1_{m,\mathbb{R}} = \bigwedge_{i=1}^{m} H^1(\mathbb{R}^3; \mathbb{R}), \tag{1.15}$$

using the following argument. If there exists a ground state $u_1 \in \mathscr{H}^1_{m,\mathbb{C}}$, then it satisfies (in a weak sense) the Schrödinger equation

$$\mathcal{H}u_1 = E_1 u_1. \tag{1.16}$$

Because the operator $\mathcal{H}$ involves only real factors and $E_1 \in \mathbb{R}$, taking the complex conjugate of both sides yields

$$\mathcal{H}\overline{u}_1 = E_1\overline{u}_1. \tag{1.17}$$

This implies, in turn, that $\mathrm{Re}u_1$ et $\mathrm{Im}u_1$ are solutions of 1.16 as well, hence $E_1 = \mathfrak{E}(u_1) = \mathfrak{E}(\mathrm{Re}u_1) = \mathfrak{E}(\mathrm{Im}u_1)$ by scalar product. The last two functions $\mathrm{Re}u_1$ and $\mathrm{Im}u_1$ are real-valued.

From now on, we shall write $\mathscr{H}^1_m$ instead of $\mathscr{H}^1_{m,\mathbb{R}}$. The search for a ground state of the electronic problem is mathematically defined by

$$\inf_{\substack{u\in\mathscr{H}^1_m \\ \|u\|_{L^2(\mathbb{R}^{3m})}=1}} \mathfrak{E}(u)\,, \tag{1.18}$$

in which we "hope" that the infimum is reached at a state u_1 and that the minimum energy E_1 belongs to $\sigma_{\mathrm{dis}}(\mathcal{H})$. For the minimization problem (1.18), the Schrödinger equation (1.6) is formally the optimality condition, in which E acts as a Lagrange multiplier associated with the normalization constraint $\|u\|^2_{L^2(\mathbb{R}^{3m})} = 1$.

Besides the ground state (E_1, u_1), all other solutions (E, u) of the Schrödinger equation (1.6) may also be regarded as a critical point for the energy functional $\mathfrak{E}$ in $\mathscr{H}^1_m$ subject to the same normalization constraint.

1.1.2 Regularity and singularities

We now pay attention to the *a priori* regularity of the wave functions in the electronic problem. This question is of the utmost importance for the numerical discretization.

We call ***nuclear coalescence*** any point $(\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathbb{R}^{3m}$ where the attractive electron-nucleus potential becomes infinite. In other words, these are points for which $\mathbf{x}_i = \mathbf{X}_I$ for a pair of indices $(i, I) \in \{1, \ldots, m\} \times \{1, \ldots, M\}$ and $\mathbf{x}_i \neq \mathbf{x}_j$ for all $j \neq i$. According to Kato [79], every minimizer u of (1.18)

- is bounded and continuous on $\mathbb{R}^{3m}$, i.e., $u \in L^\infty(\mathbb{R}^{3m}) \cap C^0(\mathbb{R}^{3m})$;
- has a continuous gradient outside the coalescence points and discontinuous but bounded directional derivatives at these points, i.e., $|\nabla u| \in L^\infty_{\mathrm{loc}}(\mathbb{R}^{3m})$;
- exhibits, at a nuclear coalescence $\mathbf{x}_i = \mathbf{X}_I$, a cusp behavior satisfying the Kato condition

$$\lim_{\varepsilon\downarrow 0} \frac{1}{|S_i|} \oint_{S_i} \nabla_{\mathbf{x}_i} u(\mathbf{x}_1, \ldots, \mathbf{X}_I + \varepsilon\mathbf{n}_i, \ldots, \mathbf{x}_m) \cdot \mathbf{n}_i \,\mathrm{d}s_i = -Z_I u(\mathbf{x}_1, \ldots, \mathbf{X}_I, \ldots, \mathbf{x}_m), \tag{1.19}$$

  where S_i denotes the unit 2-sphere in $\mathbb{R}^3$ associated with the variable $\mathbf{x}_i$, $\mathbf{n}_i$ a 3D normal unit vector to S_i, $\mathrm{d}s_i$ the surface element on S_i at the neighbourhood of $\mathbf{n}_i$'s foot. This condition establishes a connection between the function value and the average of all first-order directional derivatives at a cusp singularity.

In the same spirit, let us note the contributions of Hoffmann-Ostenhof and their co-authors [72, 73] on the regularity and cusp conditions for the *density*

$$\rho(\mathbf{x}) = m \int_{\mathbb{R}^{3(m-1)}} |u(\mathbf{x}, \mathbf{x}_2, \ldots, \mathbf{x}_m)|^2 \,\mathrm{d}\mathbf{x}_2 \ldots \mathrm{d}\mathbf{x}_m. \tag{1.20}$$

This one-variable function will play a key role in approximate models such as Hartree-Fock's or Kohn-Sham's.

The above results are of local nature. It is also possible to measure the global regularity of the electronic wave functions in Sobolev norms. According to Yserentant [85, 136],

- if $m = 1$ (single electron), then $u \in H^{5/2-\epsilon}(\mathbb{R}^3)$ for all $\epsilon > 0$, where

$$H^s(\mathbb{R}^3) = \left\{ u \in L^2(\mathbb{R}^3) \mid \int_{\mathbb{R}^3} (1 + |\boldsymbol{\xi}|^2)^s \, |\widehat{u}(\boldsymbol{\xi})|^2 \, \mathrm{d}\boldsymbol{\xi} < \infty \right\}, \tag{1.21}$$

  is a fractional Sobolev space defined by means of $\widehat{u}$, the Fourier transform of u;

- if $m \geq 2$ (two or more electrons), then $u \in H^{1-\epsilon,1}_{\mathrm{mix}}(\mathbb{R}^{3m})$ for all $\epsilon > 0$, where

$$H^{\vartheta,1}_{\mathrm{mix}}(\mathbb{R}^{3m}) = \Bigg\{ u \in L^2(\mathbb{R}^{3m}) \mid \int_{\mathbb{R}^{3m}} \prod_{i=1}^{m} \left(1 + |\boldsymbol{\xi}_i|^2\right)^{\vartheta} \left(1 + \sum_{i=1}^{m} |\boldsymbol{\xi}_i|^2\right) |\widehat{u}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m)|^2 \, \mathrm{d}\boldsymbol{\xi}_1 \ldots \mathrm{d}\boldsymbol{\xi}_m < \infty \Bigg\}, \tag{1.22}$$

  is a mixed Sobolev space defined by means of $\widehat{u}$, the Fourier transform of u; the mixed component, quantified by the superscript ϑ, corresponds to partial derivatives involving at least two electrons.

These properties should be kept in mind as a guideline for numerical discretization: at the neighbourhood of a coalescence, one must refine the mesh or use suitable basis functions if one wishes to correctly capture the wave function's cusps.

1.2 The Kohn-Sham model

1.2.1 Mathematical derivation

It is unfortunately not a smart strategy to discretize the Schrödinger model directly, insofar as the calculations become unacceptably long for, say, more than ten electrons. There are basically two reasons for this fact.

1. The first reason is the nature of the unknown $u(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ which is a function of $3m$ scalar variables. Heuristically, if in discretization we use K points per spatial direction, this would give rise to K^{3m} discrete values in total. Thus the number of variables grows exponentially with m. This is known as the *curse of dimensionality*.

2. The second reason is the presence of integrals over $\mathbb{R}^{3m}$ in the energy functional $\mathfrak{E}$, especially the *bielectronic* integrals

$$\int_{\mathbb{R}^{3m}} \frac{|u(\mathbf{x}_1, \ldots, \mathbf{x}_m)|^2}{|\mathbf{x}_j - \mathbf{x}_i|} \, \mathrm{d}\mathbf{x}_1 \ldots \mathrm{d}\mathbf{x}_m \tag{1.23}$$

   for which there is no analytical formula, even with $m = 2$. Indeed, the approximate evaluation of these integrals is the bottleneck of the calculations.

Since the 1930s, many alternative models have been put forward that were supposedly better suited for numerical computations. One of the most prominent among these is the Kohn-Sham model. It is part of the *Density Functional Theory (DFT)*, the salient feature of which is to reformulate problem (1.18) in terms of the density ρ, defined by (1.20), in place of the wave function u. Such a formulation would obviously have the decisive advantage of working with a function of 3 variables instead of $3m$ variables. This possibility of such a formulation stems from the Hohenberg-Kohn theorem [74], the statement of which may be found in [19, p. 48].

A key role for the density

To expose the Density Functional Theory in an elementary way, we first observe that the attraction term

$$\sum_{i=1}^{m} \int_{\mathbb{R}^{3m}} V(\mathbf{x}_i)|u(\mathbf{x}_1,\dots,\mathbf{x}_m)|^2 \,\mathrm{d}\mathbf{x}_1 \dots \mathrm{d}\mathbf{x}_m$$

in the energy functional (1.13) easily becomes

$$\int_{\mathbb{R}^3} V(\mathbf{x})\rho(\mathbf{x}) \,\mathrm{d}\mathbf{x} \tag{1.24}$$

by antisymmetry of u. This is the only term that contains the potential V. Then, we break the infimum problem (1.18) into two steps, namely,

$$\inf_{\substack{u\in\mathscr{H}^1_m \\ \|u\|_{L^2(\mathbb{R}^{3m})}=1}} \mathfrak{E}(u) = \inf_{\rho\in\mathscr{I}_m} \left\{ \left(\inf_{u\to\rho} \mathfrak{E}_\flat(u) \right) + \int_{\mathbb{R}^3} V\rho \right\}, \tag{1.25a}$$

by defining

$$\mathfrak{E}_\flat(u) = \frac{1}{2}\sum_{i=1}^{m} \int_{\mathbb{R}^{3m}} |\nabla_{\mathbf{x}_i} u|^2 \,\mathrm{d}\mathbf{x}_1 \dots \mathrm{d}\mathbf{x}_m + \sum_{1\le i<j\le m} \int_{\mathbb{R}^{3m}} \frac{|u|^2}{|\mathbf{x}_j - \mathbf{x}_i|} \,\mathrm{d}\mathbf{x}_1 \dots \mathrm{d}\mathbf{x}_m \tag{1.25b}$$

and by letting $u \to \rho$ stand for the set of all $u \in \mathscr{H}^1_m$ whose L^2-norm is 1 and whose density is equal to ρ. By virtue of a non-trivial result obtained by Lieb [92], the set of such densities ρ is given by

$$\mathscr{I}_m = \left\{ \rho \ge 0, \quad \sqrt{\rho} \in H^1(\mathbb{R}^3), \quad \int_{\mathbb{R}^3} \rho = m \right\}. \tag{1.26}$$

At this point, the electronic problem has been reformulated as the search for

$$\inf_{\rho\in\mathscr{I}_m} \left\{ \mathfrak{F}_\flat(\rho) + \int_{\mathbb{R}^3} V\rho \right\}, \tag{1.27a}$$

where the *Levy-Lieb* functional

$$\mathfrak{F}_\flat(\rho) = \inf_{u\to\rho} \mathfrak{E}_\flat(u) \tag{1.27b}$$

appears as "universal," i.e., independent of any external potential V. This seems to be an overwhelming victory from the standpoint of complexity reduction, insofar as our unknown ρ is now a scalar function of 3 variables (instead of $3m$ variables). Unfortunately, there is no practical expression to calculate the functional $\mathfrak{F}_\flat$. At best, we know certain approximations that are based on accurate evaluations of $\mathfrak{F}_\flat$ on a "similar" system. For instance, if the reference system is homogeneous gas of electrons, we end up with the Thomas-Fermi models [91], which are grossly inaccurate from the standpoint of physics but excellent prototypes from the standpoint of mathematics.

An ansatz for the Levy-Lieb functional

The breakthrough was achieved by Kohn-Sham [83] by assuming that there is no interaction between the m electrons in the system[4] and by introducing a "modeled" correction term. This amounts to restricting ourselves to those densities $\rho \in \mathscr{I}_m$ that correspond to a wave function u given by the *Slater determinant*

$$u(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \frac{1}{m!} \begin{vmatrix} \varphi_1(\mathbf{x}_1) & \cdots & \varphi_1(\mathbf{x}_m) \\ \vdots & & \vdots \\ \varphi_m(\mathbf{x}_1) & \cdots & \varphi_m(\mathbf{x}_m) \end{vmatrix} \tag{1.28}$$

of m monoelectronic orthonormal functions[5] φ_i, called *molecular orbitals*. The idea of the Slater determinant is inspired from a previous model advocated by Hartree-Fock. Its purpose is twofold. First, it plainly complies with the antisymmetry condition (1.3). Second, it sets the unknowns as m functions of 3 variable, which is still more favorable than one function of $3m$ variables.

Plugging the determinant (1.28) into the energy functional (1.25b) and taking the infimum, we end up with

$$\mathfrak{F}_\flat(\rho) = T_{\mathrm{KS}}(\rho) + J(\rho) + E_{\mathrm{xc}}(\rho), \tag{1.29}$$

with

$$T_{\mathrm{KS}}(\rho) := \inf_{\Phi \in \mathscr{W}_m} \left\{ \frac{1}{2} \sum_{i=1}^{m} \int_{\mathbb{R}^3} |\nabla \varphi_i(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x}, \quad \rho_\Phi = \rho \right\}, \tag{1.30a}$$

$$J(\rho) := \frac{1}{2} \iint_{\mathbb{R}^3 \times \mathbb{R}^3} \frac{\rho(\mathbf{x})\rho(\mathbf{y})}{|\mathbf{y} - \mathbf{x}|} \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}, \tag{1.30b}$$

$$E_{\mathrm{xc}}(\rho) := \mathfrak{F}_\flat(\rho) - T_{\mathrm{KS}}(\rho) - J(\rho), \tag{1.30c}$$

where

$$\rho_\Phi(\mathbf{x}) := \sum_{i=1}^{m} |\varphi_i(\mathbf{x})|^2 \tag{1.31}$$

is the *density* of Φ and

$$\mathscr{W}_m = \left\{ \Phi = (\varphi_1, \ldots, \varphi_m) \in [H^1(\mathbb{R}^3)]^m, \ \ \langle \varphi_i, \varphi_j \rangle_{L^2(\mathbb{R}^3)} = \delta_{ij} \right\}. \tag{1.32}$$

is the set of admissible configurations. The Levy-Lieb functional thus appears as the sum of three terms. The first term $T_{\mathrm{KS}}(\rho)$ is the DFT approximation for the kinetic term of (1.13). The second term $J(\rho)$ is the DFT approximation for the repulsion term of (1.13), and is interpreted as the traditional electrostatic energy of a distribution of charge ρ. The third term $E_{\mathrm{xc}}(\rho)$, called *exchange-correlation energy*, must by definition compensate the errors caused by the first two terms. In practice, it must be of course modeled in an empirical way[6].

Finally, the electronic problem in the Kohn-Sham approximation is rewritten as the search for

$$\inf_{\Phi \in \mathscr{W}_m} \mathfrak{E}_{\mathrm{KS}}(\Phi), \tag{1.33}$$

[4] But that, in the same time, they still obey Fermi's statistics.

[5] We use the symbol φ instead of ϕ, which denotes the father wavelet in the subsequent chapters.

[6] Despite the empirical modeling of this term, the DFT is considered to be an *ab initio* method.

where

$$\mathfrak{E}_{\mathrm{KS}}(\Phi) = \frac{1}{2}\sum_{i=1}^{m}\int_{\mathbb{R}^3} |\nabla\varphi_i(\mathbf{x})|^2\,\mathrm{d}\mathbf{x} + \int_{\mathbb{R}^3} V(\mathbf{x})\rho_\Phi(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{1.34}$$
$$+\ \frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\rho_\Phi(\mathbf{x})\rho_\Phi(\mathbf{y})}{|\mathbf{y}-\mathbf{x}|}\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} + E_{\mathrm{xc}}(\rho_\Phi).$$

Note that all integrals in (1.34) are over $\mathbb{R}^3$ or $\mathbb{R}^6$ and not $\mathbb{R}^{3m}$, which is also favorable for practical computations. The optimality condition for (1.33) can be written as

$$-\frac{1}{2}\Delta\varphi_i + V\varphi_i + \left(\rho_\Phi \star \frac{1}{|\mathbf{x}|}\right)\varphi_i + v_{\mathrm{xc}}(\rho_\Phi)\varphi_i = \Upsilon_i\varphi_i, \tag{1.35}$$

with v_{xc} the derivative of E_{xc} with respect to ρ and $\Upsilon_i \in \mathbb{R}$ the i-th eigenvalue. The Kohn-Sham operator

$$\mathcal{F}_\Phi = -\frac{1}{2}\Delta + V + \rho_\Phi \star \frac{1}{|\mathbf{x}|} + v_{\mathrm{xc}}(\rho_\Phi) \tag{1.36}$$

acts locally and the last term of (1.36) is multiplicative. However, the definition of $v_{\mathrm{xc}}(\rho_\Phi)$ itself may not be local.

The existence of minimizers for neutral of positively charged systems has been proved in the framework of exchange-correlation energies E_{xc} of LDA and GGA[7] types [4, 88]. The uniqueness of minimizers is still an open problem.

Remark 1.4. We have not voluntarily ignored the issue of *representability* of ρ. Indeed, an arbitrary density in $\mathscr{I}_m$ is not always associated with a Slater determinant. This difficulty can be solved by extending the notions presented above to a larger context [19, p. 56], using reduced density matrices and mixed states.

Remark 1.5. The existence of minimizers for neutral or positively charged Kohn-Sham systems was proved by [4, 88] for functionals E_{xc} of the LDA type, which are the simplest possible. Uniqueness remains an open question.

1.2.2 Regularity and singularities

In the DFT setting, a nuclear coalescence is now any point $\mathbf{x} \in \mathbb{R}^3$ for which $\mathbf{x} = \mathbf{X}_I$ for some index $I \in \{1, \ldots, M\}$. The exchange-correlation potential v_{xc} is assumed to be smooth enough with respect to its argument. Then, by using the same proof techniques, it can be proven that the minimizers φ_i of the Kohn-Sham model behave in a similar fashion as the ground state for the Schrödinger model. More specifically,

- is bounded and continuous on $\mathbb{R}^3$, i.e., $\varphi_i \in L^\infty(\mathbb{R}^3) \cap C^0(\mathbb{R}^3)$;
- has a continuous gradient outside the coalescence points and discontinuous but bounded directional derivatives at these points, i.e., $|\nabla\varphi_i| \in L^\infty_{\mathrm{loc}}(\mathbb{R}^3)$;
- exhibits, at a nuclear coalescence $\mathbf{x} = \mathbf{X}_I$, a cusp behavior satisfying the Kato condition

$$\lim_{\varepsilon\downarrow 0}\frac{1}{|S|}\oint_S \nabla_{\mathbf{x}}\varphi_i(\mathbf{X}_I + \varepsilon\mathbf{n})\cdot\mathbf{n}\,\mathrm{d}s = -Z_I\varphi_i(\mathbf{X}_I), \tag{1.37}$$

  where S denotes the unit 2-sphere in $\mathbb{R}^3$, $\mathbf{n}$ the outgoing normal unit vector to S, $\mathrm{d}s$ the surface element on S in the neighbourhood of $\mathbf{n}$'s foot.

[7] Linear Density Approximation and Generalized Gradient Approximation, which are the simplest.

In terms of Sobolev regularity, we have $\varphi_i \in H^{5/2-\epsilon}(\mathbb{R}^3)$. The solution is thus no better than $H^{5/2}$ around a singularity.

1.2.3 Galerkin approximation

Let $\mathscr{V}_b \subset H^1(\mathbb{R}^3)$ be a subspace of finite dimension $N_b \geq 1$. Then,

$$\mathscr{W}_m(\mathscr{V}_b) = \Big\{\Phi = (\varphi_1, \ldots, \varphi_m) \in [\mathscr{V}_b]^m, \;\; \langle \varphi_i, \varphi_j \rangle_{L^2(\mathbb{R}^3)} = \delta_{ij}\Big\} \tag{1.38}$$

is a finite-dimensional subspace of $\mathscr{W}_m$, defined in (1.32). The basic idea of the Ritz-Galerkin approximation is to replace the minimization problem (1.33) by

$$\inf_{\Phi \in \mathscr{W}_m(\mathscr{V}_b)} \mathfrak{E}_{\mathrm{KS}}(\Phi). \tag{1.39}$$

For the moment, let us stay at the abstract level and not pay attention to how $\mathscr{V}_b$ is built. Let $\omega_\mu : \mathbf{x} \mapsto \omega_\mu(\mathbf{x})$, $1 \leq \mu \leq N_b$, be a generic basis of $\mathscr{V}_b$. In other words,

$$\mathscr{V}_b = \mathrm{Span}\{\omega_\mu, \;\; 1 \leq \mu \leq N_b\}. \tag{1.40}$$

We look for $\varphi_i \in \mathscr{V}_b$, $1 \leq i \leq m$, under the form

$$\varphi_i(\mathbf{x}) = \sum_{\mu=1}^{N_b} \mathsf{C}_{\mu i} \omega_\mu(\mathbf{x}) \tag{1.41}$$

The coefficients $\mathsf{C}_{\mu i}$ are encapsulated into a $N_b \times m$ matrix $\mathbf{C}$. The conditions of orthonormality $\langle \varphi_i, \varphi_j \rangle_{L^2(\mathbb{R}^3)} = \delta_{ij}$ then become

$$\mathbf{C}\mathbf{S}\mathbf{C}^T = \mathbf{I}_m, \tag{1.42a}$$

the entries of the overlap matrix $\mathbf{S}$ being

$$\mathsf{S}_{\mu\nu} = \langle \omega_\mu, \omega_\nu \rangle_{L^2(\mathbb{R}^3)} = \int_{\mathbb{R}^3} \omega_\mu(\mathbf{x}) \omega_\nu(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{1.42b}$$

Inserting the decomposition (1.41) into $\mathfrak{E}_{\mathrm{KS}}(\Phi)$, the first two terms become

$$\frac{1}{2} \sum_{i=1}^{m} \int_{\mathbb{R}^3} |\nabla \varphi_i(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x} + \int_{\mathbb{R}^3} V(\mathbf{x}) \rho_\Phi(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \sum_{\mu=1}^{N_b} \sum_{\nu=1}^{N_b} \sum_{i=1}^{m} \mathsf{A}_{\mu\nu} \mathsf{C}_{\nu i} \mathsf{C}_{\mu i} = \mathrm{Tr}(\mathbf{A}\mathbf{C}\mathbf{C}^T), \tag{1.43a}$$

the entries of the core Hamiltonian matrix $\mathbf{A}$ being

$$\mathsf{A}_{\mu\nu} = \frac{1}{2} \int_{\mathbb{R}^3} \nabla \omega_\mu(\mathbf{x}) \cdot \nabla \omega_\nu(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int_{\mathbb{R}^3} V(\mathbf{x}) \omega_\mu(\mathbf{x}) \omega_\nu(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{1.43b}$$

The third term is more complicated. It can be proven to be

$$\frac{1}{2} \iint_{\mathbb{R}^3 \times \mathbb{R}^3} \frac{\rho_\Phi(\mathbf{x}) \rho_\Phi(\mathbf{y})}{|\mathbf{y} - \mathbf{x}|} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} = \frac{1}{2} \sum_{\mu=1}^{N_b} \sum_{\nu=1}^{N_b} \sum_{\kappa=1}^{N_b} \sum_{\lambda=1}^{N_b} \sum_{i=1}^{m} \sum_{j=1}^{m} (\mu\nu|\kappa\lambda) \mathsf{C}_{\kappa j} \mathsf{C}_{\lambda j} \mathsf{C}_{\mu i} \mathsf{C}_{\nu i}, \tag{1.44a}$$

with the notation

$$(\mu\nu|\kappa\lambda) := \iint_{\mathbb{R}^3 \times \mathbb{R}^3} \frac{\omega_\mu(\mathbf{x}) \omega_\nu(\mathbf{x}) \omega_\kappa(\mathbf{y}) \omega_\lambda(\mathbf{y})}{|\mathbf{y} - \mathbf{x}|} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y}. \tag{1.44b}$$

For each $N_b \times N_b$ matrix $\mathbf{D}$, let $\mathbf{J}(\mathbf{D})$ be the $N_b \times N_b$ matrix whose entries are

$$\mathsf{J}(\mathsf{D})_{\mu\nu} = \sum_{\kappa=1}^{N_b} \sum_{\lambda=1}^{N_b} (\mu\nu|\kappa\lambda)\mathsf{D}_{\kappa\lambda}.$$

Then, the value of the repulsion term (1.44a) can be expressed as

$$\frac{1}{2} \iint_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\rho_\Phi(\mathbf{x})\rho_\Phi(\mathbf{y})}{|\mathbf{y}-\mathbf{x}|}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = \frac{1}{2}\mathrm{Tr}(\mathbf{J}(\mathbf{C}\mathbf{C}^T)\mathbf{C}\mathbf{C}^T). \tag{1.45}$$

Note that the evaluation of the integrals (1.42b), (1.43b) over $\mathbb{R}^3$ and (1.44b) over $\mathbb{R}^6$ and *a priori* undecomposable, appears as the bottleneck of the calculation. A rough assessment shows that there are N_b^2 products $\mathsf{S}_{\mu\nu}$, N_b^2 values of Hamiltonian $\mathsf{A}_{\mu\nu}$ and N_b^4 bielectronic integrals $(\mu\nu|\kappa\lambda)$ to calculate! Regarding the fourth term $E_{\mathrm{xc}}(\rho_\Phi)$, since

$$\rho_\Phi = \sum_{i=1}^{m} |\varphi_i|^2 = \sum_{i=1}^{m} \sum_{\mu=1}^{N_b} \sum_{\nu=1}^{N_b} \mathsf{C}_{\mu i}\mathsf{C}_{\nu i}\omega_\mu\omega_\nu = \sum_{\mu=1}^{N_b} \sum_{\nu=1}^{N_b} (\mathsf{C}\mathsf{C}^T)_{\mu\nu}\omega_\mu\omega_\nu,$$

it can be identified with a function of $\mathbf{C}\mathbf{C}^T$ for $\Phi \in \mathscr{W}_m(\mathscr{V}_b)$. Finally, the Kohn-Sham energy viewed as a function of $\mathbf{C}$ is equal to

$$\mathfrak{E}_{\mathrm{KS}}(\mathbf{C}) = \mathrm{Tr}(\mathbf{A}\mathbf{C}\mathbf{C}^T) + \frac{1}{2}\mathrm{Tr}(\mathbf{J}(\mathbf{C}\mathbf{C}^T)\mathbf{C}\mathbf{C}^T) + E_{\mathrm{xc}}(\mathbf{C}\mathbf{C}^T). \tag{1.46}$$

The first-order optimality conditions for minimizing (1.46) subject to the constraint (1.42a) lead to the nonlinear eigenvalue problem

$$\mathbf{F}(\mathbf{C}\mathbf{C}^T)\mathbf{C} = \mathbf{S}\mathbf{C}\boldsymbol{\Upsilon}, \tag{1.47a}$$

$$\boldsymbol{\Upsilon} = \mathrm{Diag}(\Upsilon_1, \ldots, \Upsilon_m), \tag{1.47b}$$

$$\mathbf{C}\mathbf{S}\mathbf{C}^T = \mathbf{I}_m, \tag{1.47c}$$

where $\mathbf{F} = \frac{1}{2}\nabla_{\mathbf{C}}\mathfrak{E}_{\mathrm{KS}}$ is the Fock matrix and $(\Upsilon_1, \ldots, \Upsilon_m)$ are the eigenvalues. Equation (1.47a) appears as the discrete counterpart of the continuous equation (1.35). Problem (1.47) is called a *Self-Consistent Field (SCF)* problem, for which appropriate algorithms should be devised [20, 21, 24] (see also [16] and [22, §6.2.5] for a review).

The question remains as to how the finite-dimensional subspace $\mathscr{V}_b$ should be built. The choice of the functions ω_μ, $1 \le \mu \le N_b$, which form a *basis set*, is a crucial step. It strongly depends on the physical effect that the user wishes to simulate and on the degree of accuracy he/she wants to achieve. It also heavily relies on his/her empirical know-how.

1.3 Atomic orbital basis sets

1.3.1 STO (Slater Type Orbitals)

A first natural intuition is that a molecule is an assembly of slightly distorted atoms. It is postulated, as an act of faith, that it is possible to combine the wave functions for isolated atoms (up to some slight alterations) to obtain a "good" basis set for the molecule. This is the general philosophy of *LCAO (Linear Combination of Atomic Orbitals)* bases.

Eigenstates of the hydrogen ion

To put the LCAO philosophy into practice, we have to deepen our knowledge about the wave functions of an isolated atom. In this respect, let us recall the properties of the hydrogen ion, that consists of a single electron and a single nucleus of charge Z located at $\mathbf{X} = \mathbf{0}$. The discrete spectrum of this isolated atom is determined by the Schrödinger equation

$$-\frac{1}{2}\Delta u - \frac{Z}{\mathrm{r}}u = Eu \tag{1.48}$$

with $\mathrm{r} = |\mathbf{x}|$. We have the following results:

- The eigenvalues of $\mathcal{H}$ form a sequence $\{E_k^Z\}_{k\geq 1}$ defined as

$$E_k^Z = -\frac{Z^2}{2k^2}. \tag{1.49}$$

- The eigenvalue E_k^Z is of multiplicity k^2. The corresponding eigenspace in $\mathscr{H}_m^1$ is generated by the k^2 eigenvectors

$$u_{k\ell m}^Z(\mathbf{x}) = \Xi_\ell^m(\theta,\varphi)Q_{k\ell}(Z\mathrm{r})\exp(-Z\mathrm{r}), \tag{1.50}$$

  which are called *atomic orbitals*, where

  – indices k (principal quantum number), ℓ (azimuthal quantum number) and m (magnetic quantum number) are integers satisfying

$$0 \leq \ell \leq k-1, \qquad -\ell \leq m \leq \ell\ ; \tag{1.51}$$

  – the real-valued functions $\Xi_\ell^m(\theta,\varphi)$ are equal to

$$\Xi_\ell^m = \begin{cases} \dfrac{\mathrm{i}}{\sqrt{2}}(Y_\ell^m - (-1)^m Y_\ell^{-m}) & \text{if } m \leq -1, \\ Y_\ell^0 & \text{if } m = 0, \\ \dfrac{1}{\sqrt{2}}(Y_\ell^{-m} + (-1)^m Y_\ell^m) & \text{if } m \geq 1, \end{cases} \tag{1.52}$$

    where $Y_\ell^m(\theta,\varphi)$ are the usual complex-valued spherical harmonics, with the angular variables θ, φ of the vector $\mathbf{x} = (x,y,z)$ having been defined after a random choice of the z-axis ;

  – the real-valued functions $Q_{k\ell}$ are equal to

$$Q_{k\ell}(\mathrm{R}) = C_{k\ell}\,\mathrm{R}^\ell\,\mathcal{L}_{k-\ell-1}^{(2\ell+1)}(\mathrm{R}), \tag{1.53}$$

    where $C_{k\ell}$ is a normalizing constant and $\mathcal{L}_{k-\ell-1}^{(2\ell+1)}$ is a generalized Laguerre polynomial of degree $k-\ell-1$, named *radial portion*.

- The ground state solution

$$u_{100}^Z(\mathbf{x}) = (Z^3/\pi)^{1/2}\exp(-Z\mathrm{r}). \tag{1.54}$$

  is known as the normalized *Slater function* and corresponds to the ground state energy $E_1^Z = -Z^2/2$. Among all orbitals $u_{k\ell m}^Z$, the fundamental state u_{100}^Z is the only one that does not vanish at $\mathbf{x} = \mathbf{0}$ (therefore, the only one that exhibits a non-trivial cusp). It is also the only one that keeps a constant sign over $\mathbb{R}^3$ (the remaining ones have to change sign in order to be orthogonal to the Slater function).

It is common usage in chemistry to classify the orbitals $u^Z_{k\ell m}$ according to another nomenclature, obtained from the previous system by the changes in notation recapitulated in Table 1.1. Each polynomial appearing as subscript in this table is the expression in Cartesian coordinates of the corresponding product $\mathrm{r}^\ell\,\Xi^m_\ell(\theta,\varphi)$. These polynomials are called *angular parts* of the wave function. In the new system, the Slater function (1.54) is referred to as $u^Z_{1\mathrm{s}}$.

	$m=-3$	$m=-2$	$m=-1$	$m=0$	$m=1$	$m=2$	$m=3$
$\ell=0$				s			
$\ell=1$			p_y	p_z	p_x		
$\ell=2$		d_{xy}	d_{yz}	$\mathrm{d}_{3z^2-r^2}$	d_{xz}	$\mathrm{d}_{x^2-y^2}$	
$\ell=3$	$\mathrm{f}_{y(3x^2-y^2)}$	f_{xyz}	f_{yz^2}	$\mathrm{f}_{z(5z^2-3r^2)}$	f_{xz^2}	$\mathrm{f}_{z(x^2-y^2)}$	$\mathrm{f}_{x(x^2-3y^2)}$

Table 1.1: Quantum chemical names for the hydrogen ion orbitals.

From an atom to a molecule

We now go back to a molecule of M nuclei and m electrons. If the I-th nucleus of charge Z_I and located at $\mathbf{X}_I$ were alone, it would have atomic orbitals

$$u^{Z_I}_{k\ell m}(\cdot-\mathbf{X}_I),$$

with $k\geq 1$, $0\leq\ell\leq k-1$ and $-\ell\leq m\leq\ell$ and $u^{Z_I}_{k\ell m}$ defined by (1.50). Starting with $k=\ell=m=0$ and going in increasing order of k, ℓ, $|m|$, we truncate this sequence after having visited $N^I_b\geq 1$ elements. This yields a set of functions $\chi^I_{\mu_I}$, $1\leq\mu_I\leq N^I_b$, all of which are localized and centered on the I-th nucleus. By assembling all these sets, we end up with the subspace

$$\mathscr{V}_b=\mathrm{Span}\{\chi^I_{\mu_I},\ \ 1\leq I\leq M,\ \ 1\leq\mu_I\leq N^I_b\},\tag{1.55}$$

whose dimension is

$$N_b=\sum_{I=1}^M N^I_b.\tag{1.56}$$

This abrupt way of building $\mathscr{V}_b$ does not give rise to the best possible basis set, to the extent that the functions at the atoms do not see each other. To account for the interaction between atoms, Slater [125] recommended to slightly distort the atomic orbitals by first considering $u^{\mathrm{S},I}_{k\ell m}(\cdot-\mathbf{X}_I)$ at the I-th nucleus, where

$$u^{\mathrm{S},I}_{k\ell m}(\mathbf{x})=\Xi^m_\ell(\theta,\varphi)\,c_{k\ell}\,\mathrm{r}^{k-1}\exp(-\zeta^I_{k\ell}\mathrm{r})\tag{1.57}$$

and $c_{k\ell}$ a normalization constant, then by applying the same procedure for (1.55). Comparing (1.57) with (1.50), we see two major differences:

1. The charge Z_I is replaced by the parameter $\zeta^I_{k\ell}$ which depends on (k,ℓ). Either we apply the formula
$$\zeta^I_{k\ell}=\frac{Z_I-s_{k\ell}}{k},\tag{1.58}$$
where $s_{k\ell}$ is a constant mesuring the *shielding effect*, or we directly adjust the $\zeta^I_{k\ell}$.

2. Since $\mathrm{r}^{k-1} = \mathrm{r}^{\ell}\mathrm{r}^{k-\ell-1}$, every thing happens as if the generalized Laguerre polynomial $\mathcal{L}_{k-\ell-1}^{(2\ell+1)}$ of (1.53) has been replaced by its higher degree monomial $\mathrm{r}^{k-\ell-1}$, up to a multiplicative constant. We have thus simplified the radial parts while preserving the angular parts.

To better understand these functions, let us explicit the first Slater orbitals. To alleviate notations, we omit the index I of the nucleus but it is implicitly understood that the various ζ's below depend on I, as well as $\mathbf{r} = \mathbf{x} - \mathbf{X}_I$ and $\mathrm{r} = |\mathbf{r}|$.

- For $k = 1$,
$$\chi_{1\mathrm{s}}^{\mathrm{S}}(\mathbf{r}) = (\zeta_{1\mathrm{s}}^3/\pi)^{1/2}\, \exp(-\zeta_{1\mathrm{s}}\mathrm{r}). \tag{1.59}$$
- For $k = 2$,
$$\chi_{2\mathrm{s}}^{\mathrm{S}}(\mathbf{r}) = (\zeta_{2\mathrm{s}}^5/3\pi)^{1/2}\, \mathrm{r}\exp(-\zeta_{2\mathrm{s}}\mathrm{r}), \tag{1.60a}$$
$$\chi_{2\mathrm{p}_x}^{\mathrm{S}}(\mathbf{r}) = (\zeta_{2\mathrm{p}}^5/\pi)^{1/2}\, x\exp(-\zeta_{2\mathrm{p}}\mathrm{r}), \tag{1.60b}$$
$$\chi_{2\mathrm{p}_y}^{\mathrm{S}}(\mathbf{r}) = (\zeta_{2\mathrm{p}}^5/\pi)^{1/2}\, y\exp(-\zeta_{2\mathrm{p}}\mathrm{r}), \tag{1.60c}$$
$$\chi_{2\mathrm{p}_z}^{\mathrm{S}}(\mathbf{r}) = (\zeta_{2\mathrm{p}}^5/\pi)^{1/2}\, z\exp(-\zeta_{2\mathrm{p}}\mathrm{r}). \tag{1.60d}$$

A basis of type (1.57), (1.55) is called *STO (Slater Type Orbitals)* basis. STO bases are particularly suitable to the capture of cusps, since the Slater function does have a genuine "peak" and is always a member of the basis set. For this reason, STO bases were very popular in the beginning of quantum chemistry, before being dethroned in the 1950s by GTO and CGTO (or STO-nG) bases that we will see in §1.3.2.

1.3.2 GTO (Gaussians) and CGTO (contracted Gaussians)

The drawback of STO bases is that there is no hope of finding an analytical formula for the $\mathsf{A}_{\mu\nu}$ and $(\mu\nu|\kappa\lambda)$. Numerical quadratures for those are extremely costly. Algebraically, the difficulty stems from $\exp(-\mathrm{r})$ in the radial parts of the functions (1.57): the distance r involves a square root, therefore is not easy to handle.

Gaussians and their "miraculous" efficiency

Things would be totally different if the radial parts were $\exp(-\mathrm{r}^2)$, in other words if the basis functions are Gaussian or Gaussian polynomials. This was discovered by Boys [14], who pointed out that with the Gaussian polynomials
$$u_{k\ell m}^{\mathrm{G},I}(\mathbf{r}) = c_{k\ell m}^I\, x^k y^\ell z^m \exp(-\alpha_{k+\ell+m}^I \mathrm{r}^2), \tag{1.61}$$
where the integers $(k,\ell,m) \in \mathbb{N}^3$ do not have the same meaning as earlier for STO bases, $\alpha_{k+\ell+m}^I > 0$ is a spreading parameter and $c_{k\ell m}^I$ is a normalizing constant, it is possible to

- calculate the quantities $\mathsf{S}_{\mu\nu}$ et $\int_{\mathbb{R}^3} \nabla\omega_\mu \cdot \nabla\omega_\nu$ introduced in (1.42b) and (1.43b) by means of explicit formulae;
- reduce the quantities $\int_{\mathbb{R}^3} V\omega_\mu\omega_\nu$ et $(\mu\nu|\kappa\lambda)$ introduced in (1.43b) and (1.44b) to 1-D integrals of the form
$$F(\gamma) = \int_0^1 \exp(-\gamma s^2)\, \mathrm{d}s, \tag{1.62}$$
which can be easily tabulated and interpolated.

Let us offer a glimpse of why the above claims are true for pure Gaussians, i.e., $k = \ell = m = 0$. The case of polynomial Gaussians $k + \ell + m \geq 1$ can be deduced by from that of pure Gaussians via integrations by parts (which lead to recursion formulae relating the integrals at issue to those corresponding to polynomial Gaussians with lower degrees). The details of the calculation of the integrals can be found in [19].

The primary reason why major simplifications occur is that the product of two Gaussians remains a Gaussian. Indeed, if

$$\omega_\mu(\mathbf{x}) = C_\mu \exp(-\alpha_\mu |\mathbf{x} - \mathbf{X}_\mu|^2), \qquad \omega_\nu(\mathbf{x}) = C_\nu \exp(-\alpha_\nu |\mathbf{x} - \mathbf{X}_\nu|^2),$$

then

$$\omega_\mu(\mathbf{x})\, \omega_\nu(\mathbf{x}) = C_{\mu\nu} \exp(-\beta_{\mu\nu} |\mathbf{x} - \mathbf{Y}_{\mu\nu}|^2), \tag{1.63}$$

with

$$\begin{aligned} \alpha_{\mu\nu} &= \frac{\alpha_\mu \alpha_\nu}{\alpha_\mu + \alpha_\nu}, \qquad & C_{\mu\nu} &= C_\mu C_\nu \exp(-\alpha_{\mu\nu} |\mathbf{X}_\mu - \mathbf{X}_\nu|^2), \\ \beta_{\mu\nu} &= \alpha_\mu + \alpha_\nu, \qquad & \mathbf{Y}_{\mu\nu} &= \frac{\alpha_\mu \mathbf{X}_\mu + \alpha_\nu \mathbf{X}_\nu}{\alpha_\mu + \alpha_\nu}. \end{aligned}$$

This entails the closed-form expression

$$\mathsf{S}_{\mu\nu} = \int_{\mathbb{R}^3} \omega_\mu \omega_\nu = C_{\mu\nu} \left(\frac{\pi}{\beta_{\mu\nu}} \right)^{3/2}$$

for the entries of the overlap matrix $\mathbf{S}$, as well as the explicit value

$$\frac{1}{2} \int_{\mathbb{R}^3} \nabla \omega_\mu \cdot \nabla \omega_\nu = \alpha_{\mu\nu} (3 - 2\alpha_{\mu\nu}) \mathsf{S}_{\mu\nu}$$

for the entries of the kinetic part of the core Hamiltonian matrix $\mathbf{A}$. The bielectronic integrals (1.44b) become

$$(\mu\nu|\kappa\lambda) = C_{\mu\nu} C_{\kappa\lambda} \iint_{\mathbb{R}^3 \times \mathbb{R}^3} \frac{\exp(-\beta_{\mu\nu} |\mathbf{x} - \mathbf{Y}_{\mu\nu}|^2) \exp(-\beta_{\kappa\lambda} |\mathbf{y} - \mathbf{Y}_{\kappa\lambda}|^2)}{|\mathbf{y} - \mathbf{x}|} \, \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y}. \tag{1.64}$$

The secondary reason why further simplifications occur at this stage comes from the Fourier decompositions

$$\begin{aligned} \exp(-\beta |\mathbf{x}|^2) &= \frac{1}{(4\pi\beta)^{3/2}} \int_{\mathbb{R}^3} \exp\left(-\frac{|\boldsymbol{\xi}|^2}{4\beta} \right) \exp(\mathrm{i}\boldsymbol{\xi} \cdot \mathbf{x}) \, \mathrm{d}\boldsymbol{\xi}, \\ \frac{1}{|\mathbf{x}|} &= \frac{1}{2\pi^2} \int_{\mathbb{R}^3} \frac{1}{|\boldsymbol{\xi}|^2} \exp(\mathrm{i}\boldsymbol{\xi} \cdot \mathbf{x}) \, \mathrm{d}\boldsymbol{\xi}, \end{aligned}$$

which allow us to transform (1.64) into

$$(\mu\nu|\kappa\lambda) = \frac{2\pi^2 C_{\mu\nu} C_{\kappa\lambda}}{(\beta_{\mu\nu}\beta_{\kappa\lambda})^{3/2} |\mathbf{Y}_{\mu\nu} - \mathbf{Y}_{\kappa\lambda}|} \int_0^{+\infty} \exp\left(-\frac{\xi^2}{4\gamma_{\mu\nu\kappa\lambda}} \right) \frac{\sin \xi}{\xi} \, \mathrm{d}\xi,$$

with

$$\gamma_{\mu\nu\kappa\lambda} = \frac{\beta_{\mu\nu}\beta_{\kappa\lambda}}{\beta_{\mu\nu} + \beta_{\kappa\lambda}} |\mathbf{Y}_{\mu\nu} - \mathbf{Y}_{\kappa\lambda}|^2.$$

The last ingredient is the equality

$$\int_0^{+\infty} \exp\left(-\frac{\xi^2}{4\gamma}\right)\frac{\sin\xi}{\xi}\,\mathrm{d}\xi = \sqrt{\pi\gamma}\int_0^1 \exp(-\gamma s^2)\,\mathrm{d}s,$$

which allows us to express $(\mu\nu|\kappa\lambda)$ using the auxiliairy function F of (1.62).

A basis of type (1.61), (1.55) is called *GTO (Gaussian Type Orbitals)* basis. The functions of a GTO basis are ordered by increasing index $k+\ell+m$. Let us explicit the first elements of a GTO basis associated with a nucleus of which the index I is omitted. Again, the various α's below depend on I, as well as the vector $\mathbf{r} = \mathbf{x} - \mathbf{X}_I$ and $\mathrm{r} = |\mathbf{r}|$.

- For $k+\ell+m = 0$,
$$\chi^{\mathrm{G}}_{1\mathrm{s}}(\mathbf{r}) = (2\alpha_{1\mathrm{s}}/\pi)^{3/4}\,\exp(-\alpha_{1\mathrm{s}}\mathrm{r}^2). \tag{1.65}$$
- For $k+\ell+m = 1$,
$$\chi^{\mathrm{G}}_{2\mathrm{p}_x}(\mathbf{r}) = (128\alpha^5_{2\mathrm{p}}/\pi^3)^{1/4}\,x\exp(-\alpha_{2\mathrm{p}}\mathrm{r}^2), \tag{1.66a}$$
$$\chi^{\mathrm{G}}_{2\mathrm{p}_y}(\mathbf{r}) = (128\alpha^5_{2\mathrm{p}}/\pi^3)^{1/4}\,y\exp(-\alpha_{2\mathrm{p}}\mathrm{r}^2), \tag{1.66b}$$
$$\chi^{\mathrm{G}}_{2\mathrm{p}_z}(\mathbf{r}) = (128\alpha^5_{2\mathrm{p}}/\pi^3)^{1/4}\,z\exp(-\alpha_{2\mathrm{p}}\mathrm{r}^2). \tag{1.66c}$$

It is worth noting that $\chi^{\mathrm{G}}_{2\mathrm{s}}$ does not appear in (1.66), while $\chi^{\mathrm{S}}_{2\mathrm{s}}$ did appear in (1.60). The rationale for this is that a GTO basis can only contain radial parts in $\mathrm{r}^{2L}\exp(-\alpha\mathrm{r}^2)$, because odd powers of r cannot be generated by linear combinations of functions (1.61).

Contracted Gaussians and the art of compromise

The outstanding efficiency of GTO bases for the calculation of elementary integrals has a downside: Gaussian polynomials are "flat" at nuclear singularities ($\mathbf{r} = \mathbf{0}$) and there is no chance to correctly capture the cusps. For certain types of calculation, where one is not interested in the core region, this disadvantage is not too bad since this region can be modeled by a *pseudo-potential*. For other types of calculation, called *all electrons*, where one seeks the point value of the density at singularities, GTO bases clearly suffer from a lack of accuracy.

Nevertheless, a compromise between STO and GTO, suggested by Hehre, Stewart and Pople [70], is to approach each function of the STO basis by a suitable linear combination of Gaussian polynomials. Such a combination, called ***contracted Gaussian***, is pre-calculated once for all using an error minimization process. This gives rise to the so-called *CGTO (Contracted Gaussian Type Orbitals)* basis, or more commonly *STO-nG*, where "n" denotes the number of *primitive* Gaussian polynomials involved.

Let us look at an example. Assume that, in the neighborhood of a nucleus, the ideal STO basis consists of 5 functions

$$\chi^{\mathrm{S}}_{1\mathrm{s}}(\zeta_{1\mathrm{s}},\mathbf{r}),\ \chi^{\mathrm{S}}_{2\mathrm{s}}(\zeta_{2\mathrm{s}},\mathbf{r}),\ \chi^{\mathrm{S}}_{2\mathrm{p}_x}(\zeta_{2\mathrm{p}},\mathbf{r}),\ \chi^{\mathrm{S}}_{2\mathrm{p}_y}(\zeta_{2\mathrm{p}},\mathbf{r}),\ \chi^{\mathrm{S}}_{2\mathrm{p}_z}(\zeta_{2\mathrm{p}},\mathbf{r}), \tag{1.67}$$

where the charge parameters $(\zeta_{1\mathrm{s}},\zeta_{2\mathrm{s}},\zeta_{2\mathrm{p}})$ are known by the practitioner. It is envisaged to replace these 5 functions by the contracted Gaussians

$$\chi^{\mathrm{CG}}_{1\mathrm{s}}(\boldsymbol{\tau}_1,\mathbf{v}_{1\mathrm{s}},\mathbf{r}) = \sum_{q=1}^{Q} \mathrm{v}^q_{1\mathrm{s}}\,\chi^{\mathrm{G}}_{1\mathrm{s}}(\tau^q_1,\mathbf{r}) \tag{1.68}$$

and

$$\chi^{\mathrm{CG}}_{2\mathrm{s}}(\boldsymbol{\tau}_2, \mathbf{v}_{2\mathrm{s}}, \mathbf{r}) = \sum_{q=1}^{Q} \mathsf{v}^q_{2\mathrm{s}}\, \chi^{\mathrm{G}}_{\underline{1\mathrm{s}}}(\tau_2^q, \mathbf{r}), \tag{1.69a}$$

$$\chi^{\mathrm{CG}}_{2\mathrm{p}_x}(\boldsymbol{\tau}_2, \mathbf{v}_{2\mathrm{p}}, \mathbf{r}) = \sum_{q=1}^{Q} \mathsf{v}^q_{2\mathrm{p}}\, \chi^{\mathrm{G}}_{2\mathrm{p}_x}(\tau_2^q, \mathbf{r}), \tag{1.69b}$$

$$\chi^{\mathrm{CG}}_{2\mathrm{p}_y}(\boldsymbol{\tau}_2, \mathbf{v}_{2\mathrm{p}}, \mathbf{r}) = \sum_{q=1}^{Q} \mathsf{v}^q_{2\mathrm{p}}\, \chi^{\mathrm{G}}_{2\mathrm{p}_y}(\tau_2^q, \mathbf{r}), \tag{1.69c}$$

$$\chi^{\mathrm{CG}}_{2\mathrm{p}_z}(\boldsymbol{\tau}_2, \mathbf{v}_{2\mathrm{p}}, \mathbf{r}) = \sum_{q=1}^{Q} \mathsf{v}^q_{2\mathrm{p}}\, \chi^{\mathrm{G}}_{2\mathrm{p}_z}(\tau_2^q, \mathbf{r}). \tag{1.69d}$$

In (1.68)–(1.69), $Q \in \mathbb{N}^*$ is the number of authorized primitives. The larger Q is, the better is the accuracy but the more expensive are the simulations. Usually, we take $2 \leq Q \leq 6$.

The exponents $\boldsymbol{\tau}_1 = (\tau_1^1, \ldots, \tau_1^Q)$ and the coefficients $\mathbf{v}_{1\mathrm{s}} = (\mathsf{v}^1_{1\mathrm{s}}, \ldots, \mathsf{v}^Q_{1\mathrm{s}})$ of the polynomial-gaussians (1.68) are determined as solution of the minimization problem

$$\min_{\substack{\boldsymbol{\tau}_1 \in (\mathbb{R}_+^*)^Q \\ \mathbf{v}_{1\mathrm{s}} \in \mathbb{R}^Q}} \|\chi^{\mathrm{CG}}_{1\mathrm{s}}(\boldsymbol{\tau}_1, \mathbf{v}_{1\mathrm{s}}, \cdot) - \chi^{\mathrm{S}}_{1\mathrm{s}}(\zeta_{1\mathrm{s}}, \cdot)\|^2_{L^2(\mathbb{R}^3)}.$$

There is only one set of exponents $\boldsymbol{\tau}_2 = (\tau_2^1, \ldots, \tau_2^Q)$ for the 4 contractions (1.69). This is justified by the desire to save algebraic operations by having the same exponential functions in the radial parts of level 2. The channel 2s has its own set of coefficients $\mathbf{v}_{2\mathrm{s}} = (\mathsf{v}^1_{2\mathrm{s}}, \ldots, \mathsf{v}^Q_{2\mathrm{s}})$, while the channels $2\mathrm{p}_x$, $2\mathrm{p}_y$, $2\mathrm{p}_z$ share the same set of coefficients $\mathbf{v}_{2\mathrm{p}} = (\mathsf{v}^1_{2\mathrm{p}}, \ldots, \mathsf{v}^Q_{2\mathrm{p}})$. Finally, the parameters $\boldsymbol{\tau}_2$, $\mathbf{v}_{2\mathrm{s}}$ and $\mathbf{v}_{2\mathrm{p}}$ are determined as solution of the minimization problem

$$\min_{\substack{\boldsymbol{\tau}_2 \in (\mathbb{R}_+^*)^Q \\ \mathbf{v}_{2\mathrm{s}} \in \mathbb{R}^Q \\ \mathbf{v}_{2\mathrm{p}} \in \mathbb{R}^Q}} \|\chi^{\mathrm{CG}}_{2\mathrm{s}}(\boldsymbol{\tau}_2, \mathbf{v}_{2\mathrm{s}}, \cdot) - \chi^{\mathrm{S}}_{2\mathrm{s}}(\zeta_{2\mathrm{s}}, \cdot)\|^2_{L^2(\mathbb{R}^3)} + \sum_{\natural \in \{x,y,z\}} \|\chi^{\mathrm{CG}}_{2\mathrm{p}_\natural}(\boldsymbol{\tau}_2, \mathbf{v}_{2\mathrm{p}}, \cdot) - \chi^{\mathrm{S}}_{2\mathrm{p}_\natural}(\zeta_{2\mathrm{p}}, \cdot)\|^2_{L^2(\mathbb{R}^3)}.$$

Under normalized forms, the contracted Gaussians resulting from the optimization process are stored in specialized libraries[8] and implemented in the code. There exist a huge variety of bases that are more sophisticated than STO, GTO and STO-nG: *double zeta*, *triple-zeta*, *split-valence*, *polarization*, *streaming*... to name only the most common [128].

1.4 Other types of basis sets

1.4.1 PW (plane waves) and APW (augmented plane waves)

Atomic orbital basis sets perfectly match the chemical picture. The electrons are well described, even with few orbitals. Core electrons, i.e., those in the vicinity of a nucleus, can be accurately described provided that the basis functions are sufficiently well tuned. In this respect, tunability is an advantage. However, tunability is also a disadvantage: from the standpoint of optimization, there are so many parameters!

[8] `http://bse.pnl.gov/bse/portal` for example, which covers a range of 490 published bases.

From periodic to non-periodic settings

In solid state/crystal phase calculations, the systems considered are periodic. There are an infinity of identically charged nuclei, equally distributed over a regular lattice. For this kind of problems, the *plane waves (PW)* are by far the preferred basis sets. By "plane waves" we mean complex-valued functions of the form

$$\omega_{\mathbf{G}}^{\mathrm{PW}}(\mathbf{x}) = \frac{1}{\sqrt{\Omega}} \exp(\mathrm{i}\mathbf{G} \cdot \mathbf{x}), \tag{1.70}$$

where Ω is the volume of the periodic cell and $\mathbf{G} \in \mathbb{R}^3$ a vector a the reciprocal lattice. This vector $\mathbf{G}$ acts as a 3-D discrete index of the Fourier series expansion of periodic functions defined on the initial lattice. The basis function (1.70) is a solution of the Schrödinger equation with zero potential

$$-\frac{1}{2}\Delta u = \frac{1}{2}|\mathbf{G}|^2 u,$$

which represents a free electron. This is why it is customarily said that, in the philosophy of plane waves, assemblies of atoms are slight distortions to free electrons.

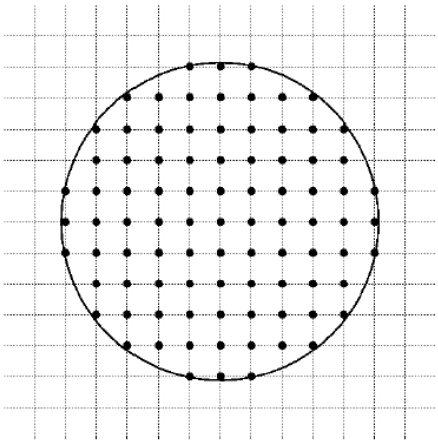

Figure 1.1: Wave vectors $\mathbf{G}$ in the cutoff region $|\mathbf{G}| \leq \sqrt{2E_c}$ of the reciprocal lattice.

Plane waves have many advantages. The most obvious ones are: they form an orthogonal set, they diagonalize the kinetic operator, they make other algebraic manipulations very simple and efficient (the Fast Fourier Transform is systematically used to switch between real and reciprocal spaces). The less obvious advantage is: they are easy to use, no basis set optimization is necessary, the size of the PW basis set is determined by one single parameter. Indeed, given a *cutoff energy* $E_c > 0$, all plane waves $\omega_{\mathbf{G}}^{\mathrm{PW}}$ such that

$$\frac{1}{2}|\mathbf{G}|^2 \leq E_c \tag{1.71}$$

must be selected to be members of the basis (this corresponds to a finite truncation of the Fourier series), as illustrated in Figure 1.1. It can then be established that the number of elements is roughly

$$N_b \simeq \Omega E_c^{3/2}. \tag{1.72}$$

As a consequence, it is also easy to improve the accuracy of the calculations: we just have to increase the cutoff energy. The delocalized nature of PW basis sets implies an all-or-nothing description for a given E_c, as no spot of the system is favored.

For an isolated non-periodic molecular system, it is possible to use PW basis sets at the price of a further approximation error. We first have to embed the system in a computational box, or *cell*, then to replicate this cell over and over by periodicity in order to obtain a lattice. The number of basis elements N_b determined by (1.71) is independent of the number of atoms in the initial molecule. However, a trade-off must be found between to incompatible requirements. On the one hand, if the cell is too small, the interaction between neighboring molecules are too strong and the periodic approximation is not good. On the other hand, if the cell is too big, the number N_b of plane waves required to enforce a given accuracy becomes prohibitively large, according to (1.72).

Managing the core regions

The inability of plane waves to handle strong oscillations of the orbitals near the nuclei without excessive cost can be addressed in two fashions, depending on whether or not one is interested in capturing their behavior in the vicinity of the nuclear singularities.

Generally speaking, core electrons have much less influence than valence electrons on the physical and chemical properties of the molecular system. After all, they do not participate in bonding, excitations, conductivity... The idea is then not to treat them explicitly any more. This will significantly reduce both the number of degrees of freedom and the size of the PW basis set to get the desired accuracy. To this end, the *pseudopotential* method replaces the Coulomb potential by a smeared out effective potential that takes into account the nuclei and the core electrons and that is to be "felt" by the valence electrons. Figure 1.2 provides an intuitive explanation of the pseudopotential method.

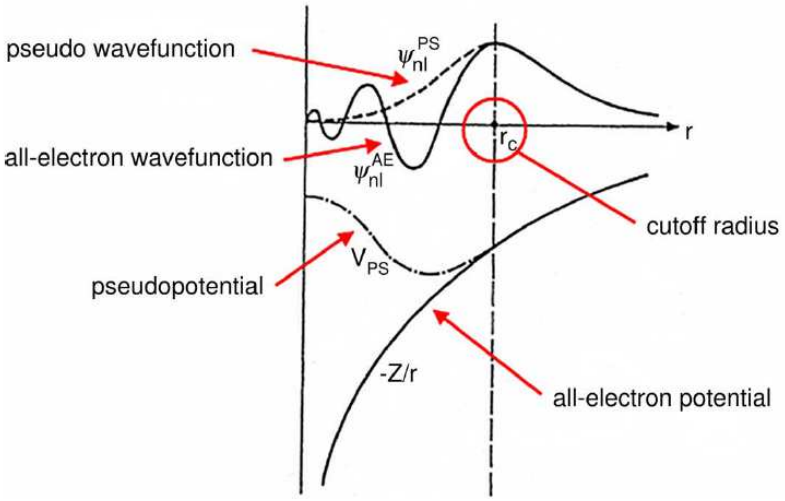

Figure 1.2: Pseudopotential and pseudo wave function.

A pseudopotential is subject to many constraints and its design is delicate. It should reproduce the necessary properties of the full problem in the reference state. Beyond a certain cutoff radius r_c, valence states should be identical to all-electron results. Occa-

sionally, it is also required to include relativistic effects. Last but not least, it should be *transferable*, which means that the same ionic potential can be used for different atomic arrangements. Numerous pseudopotentials have been proposed over the years and the user is faced with the abundance of choices. Anyhow, it has to kept in mind that pseudopotentials are nonlocal operators and they are often used in conjunction with plane waves in order to keep the size of the basis set manageable.

In some applications, for instance hyperfine fields and NMR (Nuclear Magnetic Resonance) computational spectroscopy, one is most interested in the piece of information contained in the core region. The pseudopotential approach is certainly not suitable. Let us describe the *APW (augmented plane waves)* method [126], one of the techniques traditionally used to help us cope with the situation. We start by dividing the lattice space into two regions:

1. The near-nuclei or *muffin-tin* region, that is the union of spheres $S_I = \{\mathbf{x} \in \mathbb{R}^3,\ |\mathbf{x} - \mathbf{X}_I| < \mathrm{R}\}$ centered about the atoms I and having radius R. Inside each sphere, the electrons behave almost as if they were in an isolated atom and are likely to exhibit atomic-like wave functions.

2. The far-away-from-nuclei or *interstitial* region, that is the remaining part of the space. Outside the spheres, the electrons are almost free and, as we saw earlier, can be described by plane waves.

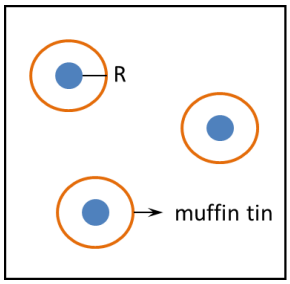

Figure 1.3: Partitioning the lattice space into muffin-tin and interstitial regions.

In other words, the basis functions considered are

$$\omega_{\mathbf{G}}^{\mathrm{APW}}(\mathbf{x}, E) = \begin{cases} \displaystyle\sum_{\ell \geq 0} \sum_{|m| \leq \ell} A_{\ell m}^{\mathbf{G},I} u_\ell^I(|\mathbf{x} - \mathbf{X}_I|, E) Y_\ell^m(\mathbf{x} - \mathbf{X}_I) & \text{if } \mathbf{x} \in S_I, \\ \dfrac{1}{\sqrt{\Omega}} \exp(\mathrm{i}\mathbf{G} \cdot \mathbf{x}) & \text{if } \mathbf{x} \in \text{inters.}, \end{cases} \tag{1.73}$$

where Y_ℓ^m are the spherical harmonics and $u_\ell^I(\mathrm{r}, E)$ is the solution of the radial equation of a free atom

$$-\frac{1}{\mathrm{r}} \frac{\partial^2}{\partial \mathrm{r}^2}(\mathrm{r} u_\ell^I) + \left[\frac{\ell(\ell+1)}{\mathrm{r}^2} - \frac{Z_I}{\mathrm{r}}\right] u_\ell^I = E u_\ell^I.$$

In 1.73, the expansion coefficients $A_{\ell m}^{\mathbf{G},I}$ and the energy level E are not degrees of freedom. Instead, they are have to be adjusted so as to satisfy some matching boundary conditions for the wave functions at $|\mathbf{x} - \mathbf{X}_I| = \mathrm{R}$. Thus, by taking advantage of the "best of each world," the APW method yields a very efficient basis set. In a sense, it is a remote ancestor of more sophisticated mixed-basis methods.

1.4.2 Real space grids and multiresolution

Atomic orbitals are localized in real space but centered on the nuclei positions. Usually, the corresponding Hamiltonian and overlap matrices are full, which makes computations expensive. Other spatially localized basis sets can be obtained by considering compacly supported functions centered on the nodes of a fixed grid. A typical instance of this is the family of finite element basis functions, which are piecewise polynomials. Spatial locality of the basis functions implies sparsity of the core Hamiltonian and overlap matrices. Sparsity in turn leads to computational efficiency. Naturally, the mesh can be refined in the neighborhood of a nucleus or coarsened in vacuum areas of the domain, which increases efficiency.

Among real space basis sets, wavelets and more specifically *Daubechies wavelets* [43] have the most attractive properties for the simulation of inhomogeneous chemical systems:

1. They are localized in the real space, which enables orbitals to be represented by a small number of significiant coefficients. The operators and operations involved in DFT can be expressed using efficient algorithms whose complexity is linear with respect to the number of basis functions. In quantum chemistry, those methods whose computational and memory requirements scale linearly with the number of atoms are called *linear scaling* or $O(N)$. A comprehensive review of $O(N)$ methods can be found in [13, 63].

2. They are localized in the Fourier space, which is helpful from the standpoint of preconditioning. Indeed, as elaborated on in [60], the convergence rate of the energy minimization process depends on the highest eigenvalue of the Hamiltonian operator. The high frequency spectrum of the latter is essentially dominated by that of the kinetic energy operator. A Fourier-localized function is an approximate eigenfunction of the kinetic energy operator. Therefore, using such functions amounts to preconditioning the high energy spectrum of the Hamiltonian operator.

3. They are orthonormal, at least for Daubechies wavelets, which not only makes the overlap matrix equal to identity but also simplifies other steps, such as matrix inversions and eigenvalue problems. Orthogonality also improves the condition number of the overlap matrix.

4. They offer a high degree of adaptivity. Although at a fixed level of resolution they are associated with a uniform mesh size, the multiresolution formalism empowers us with the possibility of successively refining and coarsening wherever it is necessary.

Since the pioneering work of Cho *et al.* [32] and Fischer and Defrancheschi [54, 55] on simple test cases, there has been tremendous development in this field with the works of Brewster *et al.* [15], Arias and his co-authors [5, 45, 51], Markvoort *et al.* [102], Niklasson *et al.* [112], Harrison *et al.* [68], Yanai *et al.* [134, 135], Chauvin [27]... The linear scaling property has been carefully investigated by Goedecker and Ivanov [64, 65]. The BigDFT team has greatly contributed to the parallel implementation of wavelets for DFT in GPU environments [60, 61]. In chapter §2, we will give some preliminary notions on wavelets in 1-D. In section §2.3, some details will be provided on wavelets in a 3-D DFT context.

Chapter 2

Prerequisites on wavelets for PDEs discretization

Contents

De la théorie des ondelettes, nous passons en revue les concepts les plus basiques qui sont indispensables pour la suite. En privilégiant d'emblée le cadre orthonormal — au détriment d'autres comme le semi-orthogonal [34], le shift-orthogonal [130] ou le biorthogonal [38] — nous souhaitons arriver le plus rapidement à la famille des ondelettes de Daubechies [42] qui sont implémentées dans BigDFT.

Nous insistons plus particulièrement sur les propriétés et les algorithmes en rapport avec la discrétisation des équations aux dérivées partielles qui nous intéressent : la propriété d'approximation au sens de la projection L^2, l'évaluation de la valeur de la fonction d'échelle en un point donné (dyadique ou non) et le calcul des "coefficients de connexion" qui sont les produits scalaires de deux dérivées premières de ces fonctions. En vue du traitement des conditions aux limites périodiques, nous examinons également la construction des ondelettes périodiques ainsi que ses conséquences sur les algorithmes précédemment évoqués.

Enfin, même si cela n'est pas directement utile pour le reste de ce mémoire, nous consacrons la dernière section du chapitre aux implications de l'aspect tridimensionnel de la Théorie de la Fonctionnelle de Densité sur l'utilisation des ondelettes qui sont la plupart du temps mises en œuvre dans leur version unidimensionnelle.

2.1 From multiresolution analyses to Daubechies wavelets

2.1.1 Multiresolution analyses, scaling functions, wavelets

Multiresolution analyses (MRA)

The theory of wavelet bases is most conveniently explained using the concept of Multiresolution Analysis (MRA) introduced by S. Mallat [100] and Y. Meyer [104]. In essence, a multiresolution analysis computes the approximation of signals at various resolutions by means of the orthogonal projection on a sequence of embedded spaces $\{\mathscr{V}_\jmath\}_{\jmath\in\mathbb{Z}}$. The most remarkable feature of this ladder of spaces is that it can be generated by the images under integer translations and dilations of a single special function, called scaling function or father wavelet.

Definition 2.1. A ***multiresolution analysis*** (MRA) of $L^2(\mathbb{R})$ consists of a sequence of nested closed subspaces

$$\{0\} \subset \cdots \subset \mathscr{V}_\jmath \subset \mathscr{V}_{\jmath+1} \cdots \subset L^2(\mathbb{R}), \quad \text{for all } \jmath \in \mathbb{Z}, \tag{2.1}$$

that satisfy certain self-similarity relations in scale/frequency and time/space, as well as the following completeness and regularity relations.

- *Self-similarity in scale*: all subspaces $\mathscr{V}_\jmath$ are dyadic-scaled versions of each other, i.e.,

$$u \in \mathscr{V}_\jmath \Longleftrightarrow u(2\cdot) \in \mathscr{V}_{\jmath+1} \Longleftrightarrow u(2^{-\jmath}\cdot) \in \mathscr{V}_0, \quad \text{for all } \jmath \in \mathbb{Z}. \tag{2.2}$$

- *Self-similarity in time*: the model subspace $\mathscr{V}_0$ is invariant under integer translations, i.e.,

$$u \in \mathscr{V}_0 \Longrightarrow u(\cdot - n) \in \mathscr{V}_0, \quad \text{for all } n \in \mathbb{Z}. \tag{2.3}$$

- *Regularity*: there exists a function $\phi \in \mathscr{V}_0$ such that the family

$$\{\phi_n := \phi(\cdot - n), \ \ n \in \mathbb{Z}\} \tag{2.4}$$

  is a Riesz basis of $\mathscr{V}_0$. Such a function ϕ is called ***scaling function*** or a ***father wavelet*** of the multiresolution analysis.

- *Completeness*: the union of these subspaces is dense in $L^2(\mathbb{R})$, i.e.,

$$\overline{\bigcup_{\jmath\in\mathbb{Z}} \mathscr{V}_\jmath} = L^2(\mathbb{R}) \tag{2.5}$$

  and their intersection should only contain the zero element, i.e.,

$$\bigcap_{\jmath\in\mathbb{Z}} \mathscr{V}_\jmath = \{0\}. \tag{2.6}$$

To clarify the regularity condition (2.4), we recall that the family $\{\phi_n\}_{n\in\mathbb{Z}}$ is said to be a *Riesz basis* of the Hilbert space $\mathscr{V}_0$, equipped with the $L^2(\mathbb{R})$-inner product, if the set of all finite linear combinations of the ϕ_n's is dense in $\mathscr{V}_0$, and if there exist two constants $0 < C_1, C_2 < \infty$ such that for all sequences $\{u_n\}_{n\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$, we have

$$C_1 \sum_{n\in\mathbb{Z}} |u_n|^2 \le \Big\| \sum_{n\in\mathbb{Z}} u_n \phi_n \Big\|^2_{L^2(\mathbb{R})} \le C_2 \sum_{n\in\mathbb{Z}} |u_n|^2.$$

This amounts to saying that the mapping

$$T : \{u_n\}_{n\in\mathbb{Z}} \mapsto \sum_{n\in\mathbb{Z}} u_n \phi_n$$

is an isomorphism from $\ell^2(\mathbb{Z})$ to $\mathscr{V}_0$. In such a case, we shall be writing

$$\mathscr{V}_0 = \mathrm{Span}\{\phi_n = \phi(\cdot - n), \;\; n \in \mathbb{Z}\}. \tag{2.7a}$$

By the self-similarity in scale condition (2.2), at all levels $\jmath \in \mathbb{Z}$ we have

$$\mathscr{V}_\jmath = \mathrm{Span}\{\phi_{\jmath,n} = 2^{\jmath/2}\phi(2^\jmath \cdot - n), \;\; n \in \mathbb{Z}\}. \tag{2.7b}$$

Of tremendous interest is a MRA associated with an orthonormal scaling function in the following sense.

Definition 2.2. A scaling function $\phi \in L^2(\mathbb{R})$ is said to be *orthonormal* if it satisfies

$$\langle \phi(\cdot - n), \phi(\cdot - m)\rangle_{L^2(\mathbb{R})} = \delta_{n-m} \tag{2.8}$$

for all $(m, n) \in \mathbb{Z}^2$, where δ is the Kronecker symbol.

Indeed, besides making some calculations simpler and some matrices better conditioned, orthonormality also enables one to design fast algorithms. It can be characterized in the Fourier space as a single algebraic identity.

Proposition 2.1. *A scaling function $\phi \in L^2(\mathbb{R})$ is orthonormal if and only if*

$$\sum_{k\in\mathbb{Z}} |\widehat{\phi}(\xi + 2\pi k)|^2 = 1 \tag{2.9}$$

for all $\xi \in \mathbb{R}$, where

$$\widehat{\phi}(\xi) = \int_{\mathbb{R}} \phi(x) \exp(-\mathrm{i}\xi x)\, \mathrm{d}x \tag{2.10}$$

denotes the Fourier transform of ϕ.

Proof. See [43, p. 132] or [101, p. 267]. □

Starting from an arbitrary scaling function ϕ, it is always possible to work out an orthonormal scaling function by the "orthonormalization trick" that is described below.

Corollary 2.1. *Let $\{\mathscr{V}_\jmath\}_{\jmath\in\mathbb{Z}}$ be a MRA associated with the scaling function ϕ. Let $\varphi \in L^2(\mathbb{R})$ be the function whose Fourier transform is*

$$\widehat{\varphi}(\xi) = \frac{\widehat{\phi}(\xi)}{\left\{\displaystyle\sum_{k\in\mathbb{Z}} |\widehat{\phi}(\xi + 2\pi k)|^2\right\}^{1/2}}$$

and consider

$$\varphi_{\jmath,n} = 2^{\jmath/2}\varphi(2^\jmath \cdot - n).$$

Then, $\{\varphi_{\jmath,n}\}_{n\in\mathbb{Z}}$ is an orthonormal basis of $\mathscr{V}_\jmath$ at all levels $\jmath \in \mathbb{Z}$.

Proof. See [101, Theorem 7.1, p. 267]. □

Let us now review two elementary examples of MRA and orthonormal scaling function.

1. *The Haar basis.* Let ϕ be the indicator function of $[0,1)$. Then, $\mathscr{V}_0$ is the space of piecewise-constant functions on intervals $[n, n+1)$, $n \in \mathbb{Z}$. Obviously, ϕ is an orthonormal scaling function.

2. *The Battle-Lemarié family.* Let ϕ be a B-spline of some fixed degree $\mathtt{D} \geq 1$, with knots at the integers. This compactly supported scaling function is not orthonormal. Application of Corollary 2.1 results in a Battle-Lemarié scaling function φ. This function φ decays exponentially but has an infinite support. This is the price to be paid for orthonormality. See [43, §5.4, p. 146] or [101, §7.1.2, p. 269] for more details.

Two-scale relation, low-pass filter

The self-similarity condition (2.2) and the regularity condition (2.4) impose stringent restrictions on the set of admissible scaling functions. Since

$$\mathscr{V}_0 \subset \mathscr{V}_1 = \mathrm{Span}\{\sqrt{2}\,\phi(2\cdot - n),\ \ n \in \mathbb{Z}\}, \tag{2.11}$$

it should be possible to express $\phi \in \mathscr{V}_0$ as

$$\phi = \sqrt{2}\sum_{n\in\mathbb{Z}} h_n \phi(2\cdot - n), \tag{2.12}$$

with $\sum_{n\in\mathbb{Z}} |h_n|^2 < \infty$. This equation is called "refinement equation" or ***two-scale relation***, in which the real sequence $\mathbf{h} = \{h_n\}_{n\in\mathbb{Z}}$ is referred to as the ***low-pass filter*** of the MRA. Within the orthonormal framework of Definition 2.2, that we shall be assuming throughout the rest of this chapter, we have

$$h_n = \langle \phi,\, \sqrt{2}\,\phi(2\cdot - n)\rangle_{L^2(\mathbb{R})}.$$

Our purpose here is to show that it is possible to recover the scaling function ϕ from the filter $\mathbf{h}$. To this end, we apply the Fourier transform (2.10) to turn (2.12) into

$$\widehat{\phi}(\xi) = \widehat{h}\left(\frac{\xi}{2}\right)\widehat{\phi}\left(\frac{\xi}{2}\right), \tag{2.13}$$

for all $\xi \in \mathbb{R}$, where

$$\widehat{h}(\xi) := \frac{1}{\sqrt{2}}\sum_{n\in\mathbb{Z}} h_n \exp(-\mathrm{i}n\xi) \tag{2.14}$$

is the transfer function of the low-pass filter. Applying $\mathtt{K}$ times (2.13), we obtain

$$\widehat{\phi}(\xi) = \left[\prod_{k=1}^{\mathtt{K}} \widehat{h}(2^{-k}\xi)\right]\widehat{\phi}(2^{-\mathtt{K}}\xi).$$

If $\widehat{\phi}$ were continuous at $\xi = 0$, then by letting $\mathtt{K} \to +\infty$ we would have

$$\widehat{\phi}(\xi) = \left[\prod_{k=1}^{\infty} \widehat{h}(2^{-k}\xi)\right]\widehat{\phi}(0). \tag{2.15}$$

The upcoming Theorem gives necessary and sufficient conditions on $\widehat{h}$ to guarantee that the above infinite product is the Fourier transform of an orthonormal scaling function.

Theorem 2.1. *If $\phi \in L^2(\mathbb{R})$ is an orthonormal scaling function, then the transfer function* (2.14) *of its low-pass filter satisfies*

$$|\widehat{h}(\xi)|^2 + |\widehat{h}(\xi+\pi)|^2 = 1, \quad \forall \xi \in \mathbb{R}, \tag{2.16a}$$

$$\widehat{h}(0) = 1. \tag{2.16b}$$

Conversely, given a 2π-periodic function $\widehat{h}$ that is continuously differentiable in a neighborhood of $\xi = 0$, if it satisfies (2.16) *and if*

$$\inf_{|\xi| \le \pi/2} |\widehat{h}(\xi)| > 0, \tag{2.17}$$

then the infinite product

$$\widehat{\phi}(\xi) = \prod_{k=1}^{\infty} \widehat{h}(2^{-k}\xi) \tag{2.18}$$

exists and is the Fourier transform of an orthonormal scaling function $\phi \in L^2(\mathbb{R})$.

Proof. See [101, Theorem 7.2, p. 271]. □

Thus, an orthonormal scaling function is entirely specified by a discrete filter whose transfer function satisfies (2.16) and Mallat's sufficient condition (2.17). The latter is not necessary, but turns out to be always satisfied in practice. As a matter of fact, it can be replaced by a weaker but more sophisticated necessary and sufficient condition discovered by Cohen [35].

In view of the infinite product (2.15), the choice (2.18) for $\widehat{\phi}$ is associated with the normalization $\widehat{\phi}(0) = 1$. This implies

$$\int_{\mathbb{R}} \phi = 1. \tag{2.19}$$

Integrating both sides of the two-scale relation (2.12) over $\mathbb{R}$ yields

$$\sum_{n\in\mathbb{Z}} h_n = \sqrt{2}. \tag{2.20}$$

The low-pass filter coefficients also satisfy

$$\sum_{n\in\mathbb{Z}} h_n h_{n+2k} = \delta_k, \tag{2.21}$$

which follows from (2.16a) and from writing out the explicit Fourier series for $|\widehat{h}(\xi)|^2 + |\widehat{h}(\xi+\pi)|^2$. In other words, their auto-correlation coefficients of even order vanish, except for the 0-th order coefficient $\sum_{n\in\mathbb{Z}} |h_n|^2 = 1$. The following Proposition points out another connection between the scaling function and the low-pass filter, anticipating the construction of compactly supported scaling functions.

Proposition 2.2. *The scaling function ϕ is has a compact support if and only if the low-pass filter $\mathbf{h} = \{h_n\}_{n\in\mathbb{Z}}$ has a compact support. In such a case, their supports are equal.*

Proof. See [42, §4, p. 965] or [101, Theorem 7.5, p. 286]. □

Wavelets, high-pass filter

As said earlier, we restrict ourselves to the orthonormal framework of Definition 2.2. We recall that $\mathscr{V}_0$ is the subspace spanned by the orthonormal scaling function ϕ and its integer translates. Let $\mathscr{W}_0$ be the orthogonal complementary subspace of $\mathscr{V}_0$ in $\mathscr{V}_1$, i.e.,

$$\mathscr{V}_1 = \mathscr{V}_0 \oplus \mathscr{W}_0. \tag{2.22}$$

This orthogonal complement is called ***detail space*** at level 0. As before, the question of interest is whether or not there exists some function ψ such that the family

$$\{\psi(\cdot - n), \quad n \in \mathbb{Z}\} \tag{2.23}$$

is an orthonormal basis of $\mathscr{W}_0$, in the sense that $\langle \psi(\cdot - n), \psi(\cdot - m)\rangle_{L^2(\mathbb{R})} = \delta_{n-m}$. If such a function ψ exists, it is called ***mother wavelet*** or simply ***wavelet*** of the MRA. In such a case, we shall be writing

$$\mathscr{W}_0 = \mathrm{Span}\{\psi_n = \psi(\cdot - n), \;\; n \in \mathbb{Z}\}. \tag{2.24a}$$

By the self-similarity in scale condition (2.2), at all level $\jmath \in \mathbb{Z}$ we have

$$\mathscr{W}_\jmath = \mathrm{Span}\{\psi_{\jmath,n} = 2^{\jmath/2}\psi(2^\jmath \cdot - n), \;\; n \in \mathbb{Z}\}, \tag{2.24b}$$

where the detail subspace $\mathscr{W}_\jmath$ is the orthogonal complement of $\mathscr{V}_\jmath$ in $\mathscr{V}_{\jmath+1}$, i.e.,

$$\mathscr{V}_{\jmath+1} = \mathscr{V}_\jmath \oplus \mathscr{W}_\jmath. \tag{2.25}$$

Again, the direct-sum decomposition (2.22), the spanning property of ψ and the self-similar property (2.2) impose stringent restrictions on the set of admissible wavelets. As $\mathscr{W}_0 \subset \mathscr{V}_1 = \mathrm{Span}\{\phi(2\cdot - n),\ n \in \mathbb{Z}\}$, it should be possible to express ψ as

$$\psi = \sqrt{2}\sum_{n\in\mathbb{Z}} g_n \phi(2\cdot - n), \tag{2.26}$$

with $\sum_{n\in\mathbb{Z}} |g_n|^2 < \infty$. The real sequence $\mathbf{g} = \{g_n\}_{n\in\mathbb{Z}}$ is referred to as the ***high-pass filter*** of the MRA. Within the orthonormal framework, we naturally have

$$g_n = \langle \psi,\, \sqrt{2}\,\phi(2\cdot - n)\rangle_{L^2(\mathbb{R})}.$$

Our purpose here is to show that it is possible to recover the wavelet ψ and the high-pass filter $\mathbf{g}$ from the low-pass filter $\mathbf{h}$. To this end, we apply the Fourier transform to turn (2.26) into

$$\widehat{\psi}(\xi) = \widehat{g}\left(\frac{\xi}{2}\right)\widehat{\phi}\left(\frac{\xi}{2}\right), \tag{2.27}$$

for all $\xi \in \mathbb{R}$, where

$$\widehat{g}(\xi) := \frac{1}{\sqrt{2}}\sum_{n\in\mathbb{Z}} g_n \exp(-\mathrm{i}n\xi) \tag{2.28}$$

is the transfer function of the high-pass filter. The following Theorem provides a recipe to construct an orthonormal wavelet ψ and the high-pass filter $\mathbf{g}$ from the low-pass filter $\mathbf{h}$.

Theorem 2.2. *Let ϕ be an orthonormal scaling function and $\mathbf{h} = \{h_n\}_{n\in\mathbb{Z}}$ the associated low-pass filter. Let*

$$\widehat{g}(\xi) = \exp(\mathrm{i}\xi)\,\overline{\widehat{h}(\xi+\pi)}. \tag{2.29}$$

Then, the function

$$\widehat{\psi}(\xi) = \widehat{g}\left(\frac{\xi}{2}\right)\widehat{\phi}\left(\frac{\xi}{2}\right) \tag{2.30}$$

is well-defined and is the Fourier transform of an orthonormal wavelet $\psi \in L^2(\mathbb{R})$.

Proof. See [43, Theorem 5.1.1, p. 135] or [101, Theorem 7.3, p. 278]. □

Note that other choices for $\widehat{g}$ are possible. These differ from the standard choice (2.29) by a multiplicative phase function. Identifying (2.28) with (2.29) and (2.14), we end up with

$$g_n = (-1)^n h_{1-n} \tag{2.31}$$

for the high-pass filter coefficients. The following Proposition investigates a consequence of (2.31), anticipating once again the construction of compactly supported wavelets.

Proposition 2.3. *If the scaling function ϕ and the low-pass filter $\mathbf{h}$ have compact support*

$$\operatorname{supp}\phi = [n_1, n_2], \qquad \operatorname{supp}\mathbf{h} = \{n_1, \dots, n_2\}$$

where $(n_1, n_2) \in \mathbb{Z}^2$, then the wavelet ψ and the high-pass filter $\mathbf{g}$ have compact supports

$$\operatorname{supp}\psi = \left[\frac{n_1 - n_2 + 1}{2}, \frac{n_2 - n_1 + 1}{2}\right], \qquad \operatorname{supp}\mathbf{g} = \{1 - n_2, \dots, 1 - n_1\}.$$

Proof. See [101, Theorem 7.5, p. 286]. □

Let us go back to the two previous examples.

1. *The Haar basis.* The scaling function is the indicator of $[0, 1)$, with the low-pass filter $h_0 = h_1 = 1/\sqrt{2}$. The construction of Theorem 2.2 yields the high-pass filter $g_0 = -g_1 = 1/\sqrt{2}$, which generates the wavelet

$$\psi(x) = \begin{cases} 1 & \text{if } 0 \le x < 1/2 \\ -1 & \text{if } 1/2 \le x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

2. *The Battle-Lemarié family.* For each degree $\mathrm{D} \ge 1$ of the B-spline, the (orthonormalized) scaling function φ has infinite support. It is symmetric around $x = 1/2$ for even D, symmetric around $x = 0$ for odd D. The construction of Theorem 2.2 yields a wavelet ψ which is antisymmetric around $x = 1/2$ for even D, symmetric around $x = 1/2$ for odd D. This wavelet has an infinite support but still decays exponentially, see [43, §5.4, p. 146].

Orthogonal decomposition of $L^2(\mathbb{R})$, Fast Wavelet Transform (FWT)

Applying recursively the direct-sum decompositions (2.25) and using (2.5), we obtain

$$L^2(\mathbb{R}) = \overline{\mathscr{V}_0 \oplus \mathscr{W}_0 \oplus \mathscr{W}_1 \oplus \mathscr{W}_2 \oplus \dots} \tag{2.32}$$

This means that any function in $L^2(\mathbb{R})$ can be uniquely expanded into the sum of an approximation in $\mathscr{V}_0$ and an aggregation of more and more refined details in $\mathscr{W}_\jmath$, $\jmath \geq 0$. The set

$$\{\phi_n\}_{n\in\mathbb{Z}} \cup \{\psi_{\jmath,n}\}_{\jmath\in\mathbb{N},\ n\in\mathbb{Z}}$$

appears to be an orthonormal basis of $L^2(\mathbb{R})$. In this basis, any function $u \in L^2(\mathbb{R})$ can be decomposed into

$$u = \sum_{n\in\mathbb{Z}} \langle u,\, \phi_n\rangle \phi_n + \sum_{\jmath\geq 0} \sum_{n\in\mathbb{Z}} \langle u,\, \psi_{\jmath,n}\rangle \psi_{\jmath,n}, \tag{2.33}$$

where the shortened notation $\langle\cdot,\cdot\rangle$ stands for the scalar product $\langle\cdot,\cdot\rangle_{L^2(\mathbb{R})}$, as long as no ambiguity occurs. The first term

$$P_0 u = \sum_{n\in\mathbb{Z}} \langle u,\, \phi_n\rangle \phi_n$$

represents the orthogonal projection of u onto $\mathscr{V}_0$. More generally, for a fixed level $\jmath \in \mathbb{Z}$, the orthogonal projection $P_\jmath : L^2(\mathbb{R}) \to \mathscr{V}_\jmath$ is defined as

$$P_\jmath u = \sum_{n\in\mathbb{Z}} \langle u,\, \phi_{\jmath,n}\rangle \phi_{\jmath,n}. \tag{2.34}$$

This projection represents a natural approximation of u at level $\jmath$. The quality of this linear approximation will be studied in the next section §2.1.2.

For $\jmath \geq 0$, we have

$$P_\jmath u = P_0 u + \sum_{\imath=0}^{\jmath-1} \sum_{n\in\mathbb{Z}} \langle u,\, \psi_{\imath,n}\rangle \psi_{\imath,n}. \tag{2.35}$$

Given the left-hand side $P_\jmath u$ of (2.35), regarded as the "finest" level, there is a fast algorithm, known as the ***Fast Wavelet Transform*** (FWT), to efficiently compute the decomposition coefficients $\langle u, \phi_{\imath,n}\rangle$ and $\langle u, \psi_{\imath,n}\rangle$ at all coarser levels $\imath \leq \jmath$. The same algorithm can be used the other way around, that is, to efficiently reconstruct the coefficients $\langle u, \phi_{\jmath,n}\rangle$ given the decomposition coefficients $\langle u, \phi_{\imath,n}\rangle$ and $\langle u, \psi_{\imath,n}\rangle$ at levels $\imath \leq \jmath - 1$. To present this algorithm, let us introduce the shorthand notations

$$c_{\jmath,n} = \langle u, \phi_{\jmath,n}\rangle, \qquad d_{\jmath,n} = \langle u, \psi_{\jmath,n}\rangle, \tag{2.36}$$

and encapsulate these coefficients into the countably infinite vectors

$$\mathbf{c}_\jmath = \{c_{\jmath,n}\}_{n\in\mathbb{Z}}, \qquad \mathbf{d}_\jmath = \{d_{\jmath,n}\}_{n\in\mathbb{Z}}.$$

For any filter or infinite vector $\mathbf{v} = \{v_n\}_{n\in\mathbb{Z}}$, we consider

- the reversion $\check{\mathbf{v}}$, whose elements are

$$\check{v}_n = v_{-n};$$

- the 2-decimation or subsampling $\downarrow_2 \mathbf{v}$, whose elements are

$$\downarrow_2 v_n = v_{2n};$$

- the 2-insertion or oversampling $\uparrow_2 \mathbf{v}$, whose elements are

$$\uparrow_2 v_n = \begin{cases} v_m & \text{if } n = 2m; \\ 0 & \text{if } n \text{ odd.} \end{cases}$$

The following Theorem reveals that the decomposition (also called *analysis*) and the reconstruction (also called *synthesis*) can be achieved by means of a cascade of discrete convolutions (denoted by $\star$), reversions, decimations and insertions.

Theorem 2.3. *The decomposition can be performed as*

$$\mathbf{c}_{\jmath-1} = \downarrow_2 (\check{\mathbf{h}} \star \mathbf{c}_\jmath), \tag{2.37a}$$

$$\mathbf{d}_{\jmath-1} = \downarrow_2 (\check{\mathbf{g}} \star \mathbf{c}_\jmath), \tag{2.37b}$$

which means that

$$c_{\jmath-1,n} = \sum_{m\in\mathbb{Z}} h_{m-2n} c_{\jmath,m},$$

$$d_{\jmath-1,n} = \sum_{m\in\mathbb{Z}} g_{m-2n} c_{\jmath,m}.$$

The reconstruction can be performed as

$$\mathbf{c}_\jmath = \mathbf{h} \star (\uparrow_2 \mathbf{c}_\jmath) + \mathbf{g} \star (\uparrow_2 \mathbf{d}_\jmath), \tag{2.38}$$

which means that

$$c_{\jmath,n} = \sum_{m\in\mathbb{Z}} h_{n-2m} c_{\jmath-1,m} + \sum_{m\in\mathbb{Z}} g_{n-2m} d_{\jmath-1,m}.$$

Proof. See [101, Theorem 7.10, p. 298]. □

The complexity of these operations is linear with respect to the number of nonzero coefficients $c_{\jmath,n}$ on the higher level $\jmath$. More specifically, suppose that the filters $\mathbf{h}$ and $\mathbf{g}$ have K nonzero coefficients and that the signal is of size $N = 2^\jmath$ on level $\jmath$. Then, a careful inspection of (2.37) and (2.38) shows that the wavelet representation (2.35) can be calculated with at most $2KN$ additions and multiplications [101, p. 302]. This is also the main motivation for designing compactly supported wavelets.

As explained in §4.4.2, the decomposition part of the Fast Wavelet Transform will be at the heart of the accuracy enhancement procedure for the numerical quadrature of scaling function-Gaussian inner products.

2.1.2 Approximation property of MRA

Vanishing moments, Strang-Fix condition

Depending on the type of applications that users have in mind, the wavelet ψ must be designed so as to be the most efficient possible, that is, to produce a maximum number of coefficients $\langle u, \psi_{\jmath,n}\rangle$ that are close to zero. This can be usually ensured if u is regular and if ψ has enough "vanishing moments" in the following sense.

Definition 2.3. The wavelet ψ is said to have M *vanishing moments*, or to be *of order* M, if

$$\int_{\mathbb{R}} x^m \psi(x)\,\mathrm{d}x = 0 \quad \text{for } m = 0, 1, \ldots, \mathsf{M}-1. \tag{2.39}$$

The integer M is called the *order of polynomial reproduction* of the multiresolution analysis (or of the wavelet).

The intuition behind this notion is that, if u is smooth, say, locally C^m, then it can be approximated by its Taylor series truncated at degree m. If $m < \mathsf{M}$, then the wavelets are orthogonal to this Taylor polynomial, thus gives rise to small-amplitude coefficients at fine scales. The property (2.39) of vanishing moments can be understood from a great many complementary viewpoints.

Theorem 2.4. *Let ϕ and ψ be a scaling function and a wavelet that generate an orthonormal basis. Assume that*

$$|\phi(x)| = O((1+|x|^2)^{-\mathsf{M}/2-1}), \quad |\psi(x)| = O((1+|x|^2)^{-\mathsf{M}/2-1}), \quad \mathsf{M} \in \mathbb{N}^*. \tag{2.40}$$

Then, the following statements are equivalent:

1. *ψ has M vanishing moments in the sense of Definition 2.3.*
2. *$\widehat{\psi}$ and its first $\mathsf{M}-1$ derivatives are zero at $\xi = 0$.*
3. *$\widehat{g}$ and its first $\mathsf{M}-1$ derivatives are zero at $\xi = 0$.*
4. *$\widehat{h}$ and its first $\mathsf{M}-1$ derivatives are zero at $\xi = \pi$.*
5. *$\widehat{\phi}$ and its first $\mathsf{M}-1$ derivatives are zero at $\xi = 2k\pi$, $k \neq 0$.*
6. *$\wp_m = \sum_{n\in\mathbb{Z}} n^m \phi(\cdot - n)$ is a polynomial of degree m, for any $0 \leq m \leq \mathsf{M}-1$.*

Proof. See [101, Theorem 7.4, p. 284] for statements 1, 2, 4, 6. See [127, §1.2, p. 1242] for statements 3, 5. □

Statement 6, commonly referred to as the ***Strang-Fix condition***, expresses the fact that the polynomials of degree less than or equal to $\mathsf{M}-1$ can be reproduced as a linear expansion of the functions $\{\phi_n\}_{n\in\mathbb{Z}}$. Note, however, that the decomposition coefficients of $\wp_m$ and of $x \mapsto x^m$ in this basis do not have finite energy, because polynomials do not belong to $L^2(\mathbb{R})$.

Corollary 2.2. *If ψ is of order M, i.e., has M vanishing moments, then*

$$\sum_{n\in\mathbb{Z}} (x-n)^m \phi(x-n) = \int_{\mathbb{R}} y^m \phi(y)\,\mathrm{d}y, \quad \forall x \in \mathbb{R}, \tag{2.41a}$$

$$\sum_{n\in\mathbb{Z}} (-1)^n n^m h_n = 0, \tag{2.41b}$$

for all $0 \leq m \leq \mathsf{M}-1$.

Proof. See [127, §1.2, p. 1242] for (2.41a). As for (2.41b), this is none other than Statement 4 of Theorem 2.4. □

The equality (2.41a) will be invoked in §4.4.2 to prove that the trapezoidal quadrature rule for the dot product $\langle u, \phi\rangle$ has degree of exactness $\mathsf{Q} = \mathsf{M}-1$. The special case $m = 0$ reads

$$\sum_{n\in\mathbb{Z}} \phi(\cdot - n) = 1 \tag{2.42}$$

and is sometimes dubbed "partition of unity." It will be helpful in §2.2.1 for the evaluation of the values of ϕ at dyadic rational points. Combined with the orthonormality constraint, the requirement of having M vanishing moments leads to a somewhat unfavorable consequence on the support size of ϕ and ψ.

Proposition 2.4. *If ψ is of order M, then the low-pass filter $\mathbf{h} = \{h_n\}_{n\in\mathbb{Z}}$ has at least $2\mathsf{M}$ nonzero coefficients, and the support of ϕ is at least of length $2\mathsf{M}-1$.*

PROOF. See [101, Theorem 7.7, p. 293]. □

In this light, Daubechies wavelets will be seen to be "optimal" insofar as they have a minimum-length support for a given number of vanishing moments.

Error estimate for linear approximation

We said earlier that $P_\mathsf{J} u$ is a natural approximation of u. In order to assess the quality of this linear approximation, let us introduce the Sobolev space defined for $s \geq 0$ as

$$H^s(\mathbb{R}) = \left\{ u \in L^2(\mathbb{R}) \mid \int_{\mathbb{R}} (1+|\xi|^2)^s |\widehat{u}(\xi)|^2 \,\mathrm{d}\xi < \infty \right\},$$

where $\widehat{u}$ denotes the Fourier transform of u. This space is equipped with the norm

$$\|u\|^2_{H^s(\mathbb{R})} = \int_{\mathbb{R}} (1+|\xi|^2)^s |\widehat{u}(\xi)|^2 \,\mathrm{d}\xi. \tag{2.43}$$

Theorem 2.5. *Assume that the MRA is of order $\mathsf{M} \geq 1$ and that*

$$\phi \in H^s(\mathbb{R}), \quad \text{for some } s \in [0, \mathsf{M}).$$

Moreover, assume that we are given a function

$$u \in H^t(\mathbb{R}), \quad \text{for some } t \in (s, \mathsf{M}).$$

Then, there exists a constant $\Gamma_{\mathsf{M},s,t}$ (dependent on M, s and t but not on J) such that

$$\|u - P_\mathsf{J} u\|_{H^s(\mathbb{R})} \leq \Gamma_{\mathsf{M},s,t} \|u\|_{H^t(\mathbb{R})} 2^{-\mathsf{J}(t-s)} \tag{2.44}$$

for all J large enough.

PROOF. See [36, §3.3, p. 165]. □

2.1.3 Orthonormal compactly supported wavelets

General construction

To design orthornormal compactly supported wavelets with a prescribed order M, the idea of Daubechies [42] was to start afresh from Statement 4 of Theorem 2.4 ($\widehat{h}$ and its first $\mathsf{M}-1$ derivatives are zero at $\xi = \pi$) and to seek $\widehat{h}$ under the form

$$\widehat{h}(\xi) = \left[\frac{1+\exp(-\mathrm{i}\xi)}{2}\right]^{\mathsf{M}} \mathcal{L}(\xi), \tag{2.45}$$

where $\mathcal{L}$ is a trigonometric polynomial: the finite length of this trigonometric polynomial would then ensure that of the low-pass filter. By squaring the modulus of both sides, we have

$$|\widehat{h}(\xi)|^2 = \left(\cos\frac{\xi}{2}\right)^{2\mathsf{M}} P\left(\sin^2\frac{\xi}{2}\right),$$

where

$$P\left(\sin^2\frac{\xi}{2}\right) = |\mathcal{L}(\xi)|^2 \tag{2.46}$$

is a polynomial with respect to its argument. $P(y)$ must take positive values for $y \in [0,1]$. Besides, according to conditions (2.16) of Theorem 2.1, P must be subject to

$$(1-y)^{\mathsf{M}}P(y) + y^{\mathsf{M}}P(1-y) = 1, \quad \forall y \in [0,1], \tag{2.47a}$$

$$P(0) = 1. \tag{2.47b}$$

Proposition 2.5. *The polynomial of degree* $\mathsf{M}-1$

$$P(y) = \sum_{k=0}^{\mathsf{M}-1} \binom{\mathsf{M}-1+k}{k} y^k \tag{2.48}$$

is positive on $[0,1]$ *and solves* (2.47).

Proof. See [43, Proposition 6.1.2, p. 171]. In fact, the solutions of (2.47) are

$$P(y) = \sum_{k=0}^{\mathsf{M}-1} \binom{\mathsf{M}-1+k}{k} y^k + y^{\mathsf{M}} R(1/2-y),$$

where R is an odd polynomial, chosen such that $P(y) \geq 0$ for $y \in [0,1]$. Nevertheless, for the actual construction of wavelets, $R \equiv 0$ is always selected. □

Formula (2.48) completely determines $|\widehat{h}(\xi)|^2$. However, what we ultimately want is $\widehat{h}(\xi)$ and not $|\widehat{h}(\xi)|^2$; therefore, we have to carry out a *spectral factorization* to extract $\widehat{h}(\xi)$. The following Lemma, due to Riesz, asserts the existence of such a factorization.

Lemma 2.1. *There exists a (non-unique) trigonometric polynomial of order* $\mathsf{M}-1$

$$\mathcal{L}(\xi) = \sum_{m=0}^{\mathsf{M}-1} \lambda_m \exp(-\mathrm{i}m\xi), \qquad \textit{with } \lambda_m \in \mathbb{R}, \tag{2.49}$$

that satisfies (2.46).

Proof. See [43, Lemma 6.1.3, p. 172].

Apart from having the form (2.49), the function $\mathcal{L}$ should also be adjusted in such a way that the transfer function $\widehat{h}$, given by (2.45), satisfies the technical condition (2.17), so that once $\widehat{h}$ is known, ϕ can be recovered by Theorem 2.1. Before discussing about the choice of the "square root" in more details, let us mention a very important consequence of (2.49) on the supports of various functions and filters.

Corollary 2.3. *The general construction* (2.45), (2.49) *implies*

$$\operatorname{supp}\phi = [0, 2\mathsf{M}-1], \qquad \operatorname{supp}\mathbf{h} = \{0,\dots,2\mathsf{M}-1\},$$
$$\operatorname{supp}\psi = [-\mathsf{M}+1, \mathsf{M}], \qquad \operatorname{supp}\mathbf{g} = \{-\mathsf{M}+1,\dots,\mathsf{M}\},$$

regardless of the actual choice for $\mathcal{L}$.

PROOF. Plugging (2.49) into (2.45), we see that $\widehat{h}(\xi)$ is a linear combination of $\exp(-in\xi)$, with n ranging from 0 to $2m-1$. Therefore, $\operatorname{supp}\mathbf{h} = \{0,\dots,2\mathsf{M}-1\}$. From Proposition 2.2 and Proposition 2.3, we infer the three remaining supports. □

Thus, the supports of ϕ, ψ and $\mathbf{h}$, $\mathbf{g}$ do not depend on the specific choice for $\mathcal{L}$. Also independent of the actual choice for $\mathcal{L}$ are the autocorrelation coefficients

$$\gamma_k = \sum_{n\in\mathbb{Z}} h_n h_{n+k} \tag{2.50}$$

of the low-pass filter. By equation (2.21), we know that even-rank autocorrelation coefficients γ_{2k} are zero, except for $\gamma_0 = 1$. The odd-rank ones can be explicitly computed as indicated in the following Proposition, and this will in turn make the connection coefficients defined in §2.2.2 independent of $\mathcal{L}$.

Proposition 2.6. *The general construction* (2.45), (2.46), (2.48), (2.49) *implies*

$$\gamma_{2k-1} = \frac{(-1)^{k-1}}{2(\mathsf{M}-k)!(\mathsf{M}+k-1)!(2k-1)}\left[\frac{(2\mathsf{M}-1)!}{(\mathsf{M}-1)!4^{\mathsf{M}-1}}\right]^2$$

for all $1 \le k \le \mathsf{M}$*, regardless of the actual choice for* $\mathcal{L}$*.*

PROOF. See [10, (3.52), p. 1725]. The main argument relies on the formula

$$|\widehat{h}(\xi)|^2 = \frac{1}{2} + \sum_{k=1}^{\mathsf{M}} \gamma_{2k-1}\cos(2k-1)\xi,$$

which stems from squaring the modulus of (2.14) and which demonstrates that the γ_k's are Fourier series coefficients of $|\widehat{h}(\xi)|^2$. □

Specific choices of phase

We now go back to the extraction of $\mathcal{L}$ from P so as to satisfy (2.46). The spectral factorization procedure is explained in [43, p. 172], to which the reader is referred for full details. Here, following the presentation of Mallat [101, p. 293], we just sketch out the principle in order to lay emphasis on the origin of various choices.

We look for $\mathcal{L}(\xi) = L(\exp(-\mathrm{i}\xi))$, where the polynomial L and its factorized form

$$L(z) = \sum_{m=0}^{\mathsf{M}-1} \lambda_m z^m = \lambda_0 \prod_{m=1}^{\mathsf{M}-1}(1-\nu_m z)$$

are considered as function of the complex variable $z \in \mathbb{C}$. As $\lambda_m \in \mathbb{R}$, condition (2.46) reads

$$|\mathcal{L}(\xi)|^2 = L(\exp(-\mathrm{i}\xi))L(\exp(\mathrm{i}\xi)) = P\Big(\frac{2-\exp(-\mathrm{i}\xi)-\exp(\mathrm{i}\xi)}{4}\Big).$$

Extending this equality to the whole complex plane, we obtain

$$L(z)L(z^{-1}) = \lambda_0^2 \prod_{m=1}^{\mathsf{M}-1}(1-\nu_m z)(1-\nu_m z^{-1}) = P\Big(\frac{2-z-z^{-1}}{4}\Big). \tag{2.51}$$

The $2(\mathsf{M}-1)$ roots of the right-hand side of (2.51) come in reciprocal pairs and their conjugates. Put another way, if $\zeta_k \in \mathbb{C}$ is a root of the right-hand side, its inverse $1/\zeta_k$ is

also a root and their conjugates $\overline{\zeta}_k$, $1/\overline{\zeta}_k$ are also roots. To cook up L, we pick each root μ_m of L among a pair $(\zeta_k, 1/\zeta_k)$ as well as $\overline{\zeta}_k$ if $\zeta_k \notin \mathbb{R}$ to get real coefficients. Thus, the multiplicity of choice lies in the selection of the roots ζ_k to become ν_m.

In this fashion, there might be up to $2^{\mathsf{M}-1}$ different solutions for $\mathcal{L}$. Only two of these are of common use among practitioners.

- *Minimal phase.* Daubechies systematically retained the roots ζ_k that are inside the unit circle $|z| \leq 1$. The resulting wavelets are called ***Daubechies wavelets*** and denoted by dbM. As shown by the examples in Figure 2.1, they are highly asymmetric. The reason for this is that their filters have their energy maximally concentrated near the starting point of their support. They are not very smooth either, at least for small M. The regularity of ϕ and ψ increases with M and be quantified accurately by means of advanced techniques proposed by Daubechies and Lagarias [44] (for the Hölder regularity) and Cohen and Daubechies [37] (for the Sobolev regularity). The results are reported in Table 2.1 for M from 2 to 8.

- *Least asymmetric.* Daubechies proved that the Haar wavelet is the only real compactly supported, orthonormal and symmetric wavelet. There is thus no hope to achieve complete symmetry for $\mathsf{M} \geq 2$ within this general construction framework. However, it is possible to optimize the choice of the square root to obtain an "almost linear phase". The resulting wavelets are called ***symmlets*** and denoted by syM. As shown by the examples in Figure 2.2, they are less markedly asymmetric. In reality, symmlets coincide with Daubechies wavelets for $\mathsf{M} \leq 3$ and are therefore relevant only for $\mathsf{M} \geq 4$.

M	α	s
2	0.5500	0.9998
3	1.0878	1.4149
4	1.6179	1.7753
5	1.9690	2.0965
6	2.1891	2.3880
7	2.4604	2.6585
8	2.7608	2.9146

Table 2.1: Hölder (C^α) and Sobolev (H^s) exponents for the minimal phase Daubechies scaling functions and wavelets.

Note that for $\mathsf{M} = 1$, any choice of $\mathcal{L}$ degenerates to the Haar system, in which ϕ and ψ are both discontinuous. For $\mathsf{M} = 2$, the minimal phase (which coincides with the least asymmetric) scaling function ϕ is associated with the low-pass filter

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}}, \quad h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}}, \quad h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}}, \quad h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}.$$

It is continuous but not differentiable, since the Hölder exponent is 0.5500. It does not even belong to $H^1(\mathbb{R})$, since the Sobolev exponent is 0.9998. Consequently, it cannot be an appropriate basis function for the discretization of PDEs involving a second derivative. This is why we shall be considering $\mathsf{M} \geq 3$ from now on.

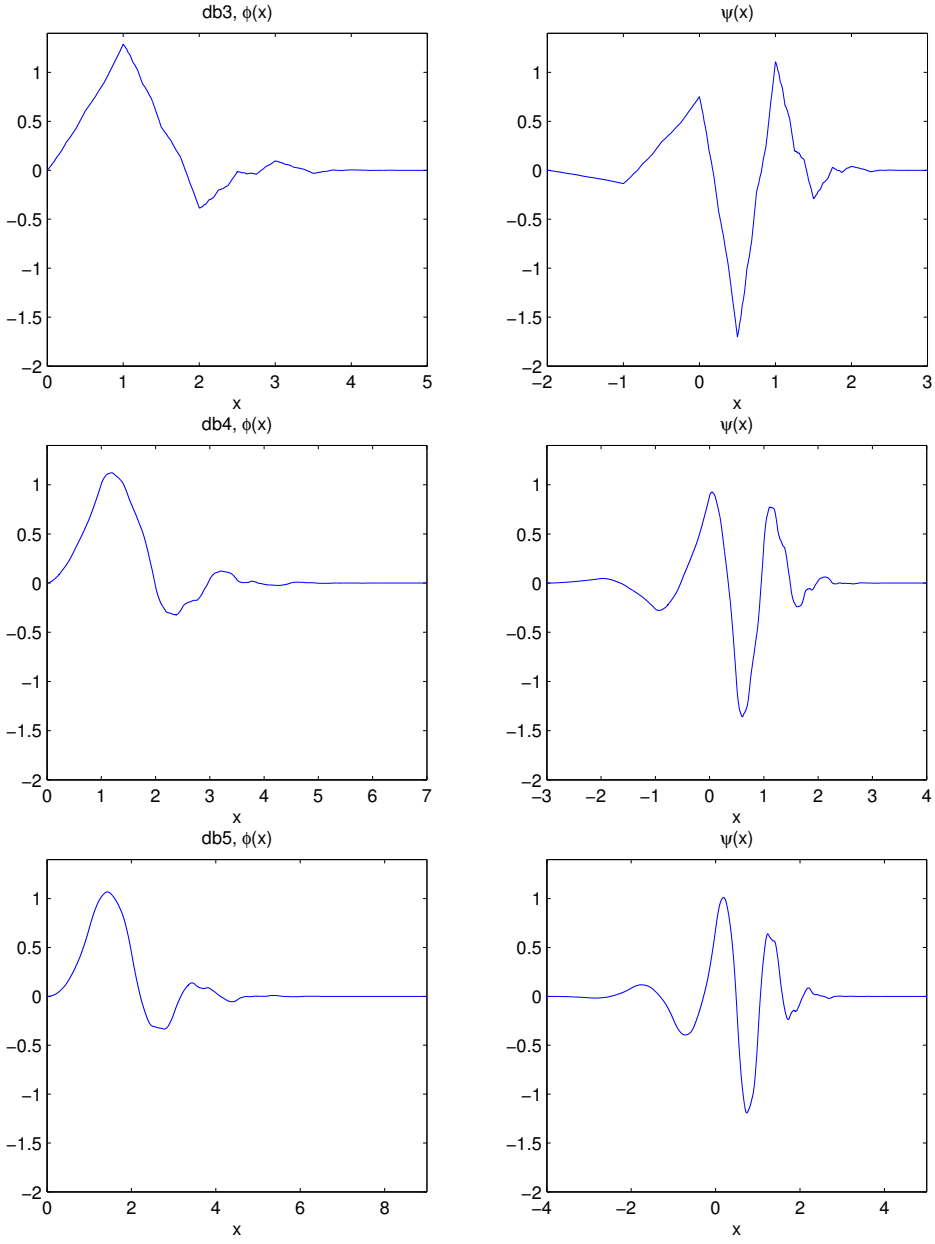

Figure 2.1: Minimal phase (db) scaling functions ϕ and wavelets ψ for $3 \leq \mathsf{M} \leq 5$.

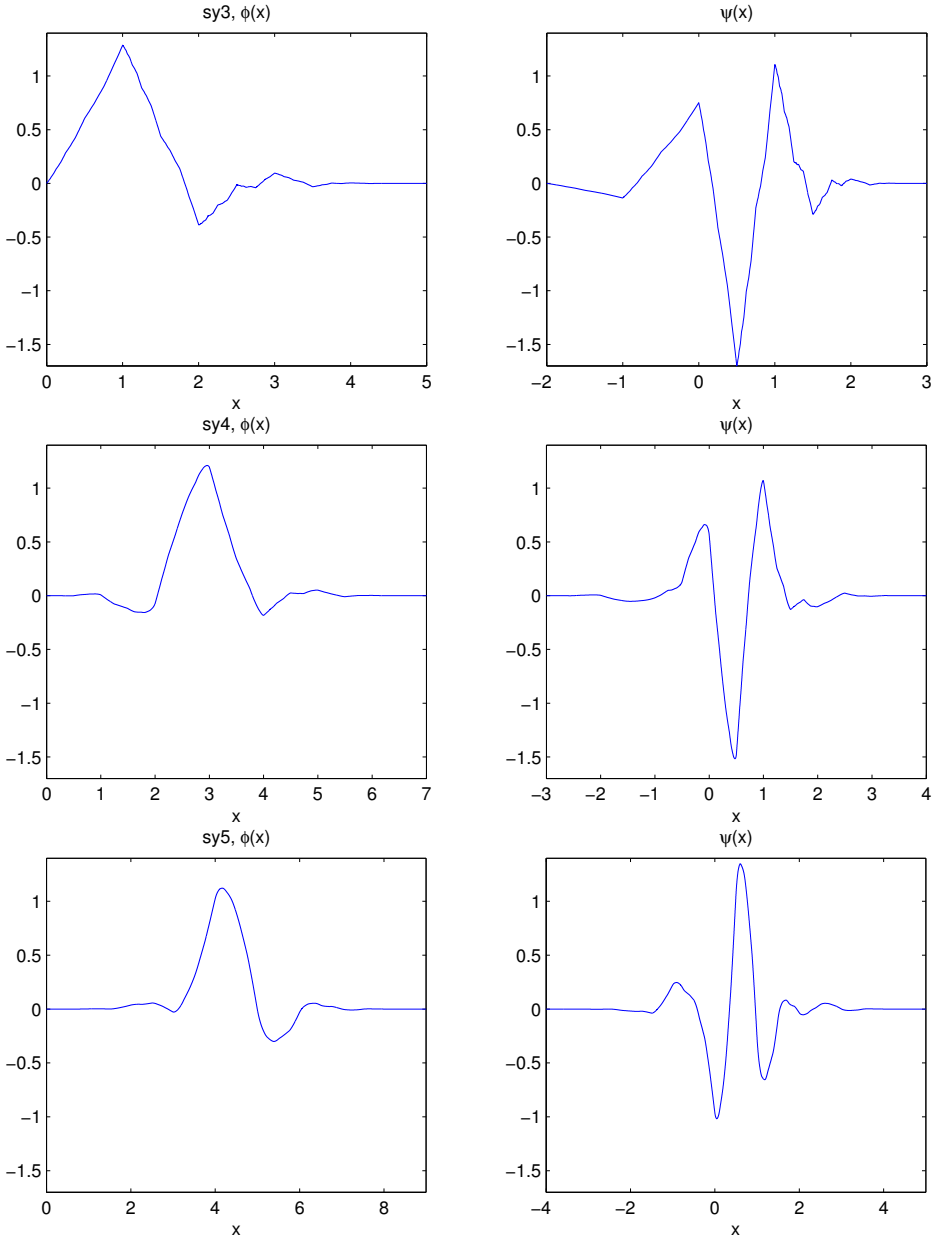

Figure 2.2: Least asymmetric (sy) scaling functions ϕ and wavelets ψ for $3 \leq \mathsf{M} \leq 5$.

2.2 Technical issues for PDE discretization

In preparation of chapter §4, this section is devoted to several technical issues related to the numerical resolution of PDEs.

2.2.1 Evaluation of function values at a given point

There is no closed-form analytic formula for the Daubechies father and mother wavelets, except for the Haar case. Instead, these are given by their filter coefficients. There are situations in which we need to know the values of ϕ, ψ at some specific abscissa x. This can be achieved exactly (for integers and dyadic rationals) or approximately (for reals) by various algorithms.

At integers

Because $\operatorname{supp}\phi = [0, 2\mathsf{M}-1]$, it is plain that $\phi(k) = 0$ for integers $k \le -1$ and $k \ge 2\mathsf{M}$. By virtue of continuity of ϕ for $\mathsf{M} \ge 2$, we necessarily have $\phi(0) = \phi(2\mathsf{M}-1) = 0$. Thus,

$$\phi(k) = 0, \qquad \text{for } \; k \le 0 \;\text{ or }\; k \ge 2\mathsf{M}-1. \tag{2.52}$$

The integers at which ϕ can be nonzero are $\{1, \ldots, 2\mathsf{M}-2\}$. These values can be computed simultaneously via a suitably normalized eigenvector problem.

Proposition 2.7. *The vector of values of ϕ at integers*

$$\boldsymbol{\phi} = (\phi(1), \phi(2), \ldots, \phi(2\mathsf{M}-3), \phi(2\mathsf{M}-2))^T \in \mathbb{R}^{2\mathsf{M}-2}$$

solves the eigenvector problem

$$\mathbf{H}\boldsymbol{\phi} = \boldsymbol{\phi}, \tag{2.53a}$$

$$\mathbf{e}^T\boldsymbol{\phi} = 1, \tag{2.53b}$$

for the eigenvalue 1, where $\mathbf{H}$ is the $(2\mathsf{M}-2)\times(2\mathsf{M}-2)$ matrix whose entries are

$$\mathsf{H}_{k\ell} = \sqrt{2}\, h_{2k-\ell}, \qquad \text{for } \; (k,\ell) \in \{1, \ldots, 2\mathsf{M}-2\}^2, \tag{2.54}$$

and $\mathbf{e}$ is the vector of size $2\mathsf{M}-2$ with components

$$\mathsf{e}_\ell = 1, \qquad \text{for } \; \ell \in \{1, \ldots, 2\mathsf{M}-2\}. \tag{2.55}$$

Proof. The eigenvector problem (2.53a) comes from writing out the two-scale relation (2.12) at each $x = k \in \{1, \ldots, 2\mathsf{M}-2\}$, i.e.,

$$\phi(k) = \sqrt{2}\sum_{n\in\mathbb{Z}} h_n \phi(2k-n) = \sqrt{2}\sum_{\ell\in\mathbb{Z}} h_{2k-\ell}\phi(\ell) = \sqrt{2}\sum_{\ell=1}^{2\mathsf{M}-2} h_{2k-\ell}\phi(\ell),$$

the last equality being due to (2.52). The normalization (2.53b) comes from the "partition of unity" property (2.42). □

For instance, when $\mathsf{M} = 3$, the eigenvector problem with eigenvalue 1 reads

$$\sqrt{2}\begin{bmatrix} h_1 & h_0 & & \\ h_3 & h_2 & h_1 & h_0 \\ h_5 & h_4 & h_3 & h_2 \\ & & h_5 & h_4 \end{bmatrix}\begin{bmatrix}\phi(1)\\ \phi(2)\\ \phi(3)\\ \phi(4)\end{bmatrix} = \begin{bmatrix}\phi(1)\\ \phi(2)\\ \phi(3)\\ \phi(4)\end{bmatrix}.$$

In practical implementations, it could be sometimes judicious [103] to reintroduce $\phi(0)$ and to artificially increase the system as

$$\mathbf{H}_0\boldsymbol{\phi}^{(0)} = \sqrt{2}\begin{bmatrix} h_0 & & & & \\ h_2 & h_1 & h_0 & & \\ h_4 & h_3 & h_2 & h_1 & h_0 \\ & h_5 & h_4 & h_3 & h_2 \\ & & & h_5 & h_4 \end{bmatrix}\begin{bmatrix}\phi(0)\\ \phi(1)\\ \phi(2)\\ \phi(3)\\ \phi(4)\end{bmatrix} = \begin{bmatrix}\phi(0)\\ \phi(1)\\ \phi(2)\\ \phi(3)\\ \phi(4)\end{bmatrix} = \boldsymbol{\phi}^{(0)}, \tag{2.56}$$

so as to reuse the matrix $\mathbf{H}_0$ later.

At dyadic rationals

From the values $\phi(k)$ for $1 \le k \le 2\mathsf{M}-2$, successive applications of the scaling relation (2.12) enable one to deduce the values of ϕ at dyadic rationals, namely,

$$\phi(2^{-\mathsf{J}}k), \qquad \text{for } \mathsf{J} \ge 1 \text{ and } 1 \le k \le 2^{\mathsf{J}}(2\mathsf{M}-1)-1.$$

The recursion from J to $\mathsf{J}+1$ is totally explicit from the refinement equation (2.12), as

$$\phi\left(\frac{k}{2^{\mathsf{J}+1}}\right) = \sqrt{2}\sum_{n\in\mathbb{Z}} h_n \phi\left(\frac{k}{2^{\mathsf{J}}} - n\right) = \sqrt{2}\sum_{n\in\mathbb{Z}} h_n \phi\left(\frac{k-2^{\mathsf{J}}n}{2^{\mathsf{J}}}\right).$$

This *dyadic cascade* algorithm involves a finite number of matrix-vector products and no eigenvector problem. To benefit from a really efficient implementation of this cascade algorithm, we have incorporated in our code a collection of Matlab routines developed by Mehra and Goyal[1] [103], in which several tricks have been deployed in order to optimally reuse memory data.

For instance, when $\mathsf{M} = 3$ and $\mathsf{J} = 1$, we have

$$\boldsymbol{\phi}^{(1)} = \begin{bmatrix}\phi(1/2)\\ \phi(3/2)\\ \phi(5/2)\\ \phi(7/2)\\ \phi(9/2)\end{bmatrix} = \sqrt{2}\begin{bmatrix} h_1 & h_0 & & & \\ h_3 & h_2 & h_1 & h_0 & \\ h_5 & h_4 & h_3 & h_2 & h_1 \\ & & h_5 & h_4 & h_3 \\ & & & & h_5 \end{bmatrix}\begin{bmatrix}\phi(0)\\ \phi(1)\\ \phi(2)\\ \phi(3)\\ \phi(4)\end{bmatrix} = \mathbf{H}_1\boldsymbol{\phi}^{(0)}.$$

Although $\phi(0) = 0$, it is advised to leave it in the system, since the matrix $\mathbf{H}_1$ will be reused later. At the next level $\mathsf{J} = 2$, it is recommended to split the unknown vector $\boldsymbol{\phi}^{(2)}$ into two smaller vectors, that is,

$$\boldsymbol{\phi}^{(2)}_{1/4} = (\phi(1/4), \phi(5/4), \phi(9/4), \phi(13/4), \phi(17/4))^T,$$
$$\boldsymbol{\phi}^{(2)}_{3/4} = (\phi(3/4), \phi(7/4), \phi(11/4), \phi(15/4), \phi(19/4))^T,$$

[1] available at `http://www.netlib.org/toms/929.zip`

and to implement the products

$$\boldsymbol{\phi}^{(2)}_{1/4} = \mathbf{H}_0 \boldsymbol{\phi}^{(1)}, \qquad \boldsymbol{\phi}^{(2)}_{3/4} = \mathbf{H}_1 \boldsymbol{\phi}^{(1)}.$$

As $\jmath$ increases, the unknown vector $\boldsymbol{\phi}^{(\jmath)}$ keeps being split into subvectors of length 5, which gives rise to matrix-vector products involving $\mathbf{H}_0$ and $\mathbf{H}_1$ [103].

At reals

For a fixed $\jmath \in \mathbb{N}$, we define $\varphi_\jmath : \mathbb{R} \to \mathbb{R}$ as the function that is piecewise-linear on each interval $[2^{-\jmath}k, 2^{-\jmath}(k+1)]$, $k \in \mathbb{Z}$, and that takes values

$$\varphi_\jmath(2^{-\jmath}k) = \phi(2^{-\jmath}k).$$

For all $x \in \mathbb{R}$, $\varphi_\jmath(x)$ appears to be an approximation of $\phi(x)$. The quality of this approximation is given by the following Proposition.

Proposition 2.8. *If ϕ is Hölder continuous with exponent α, then there exists $C > 0$ and $\jmath_0 \in \mathbb{N}$ such that, for all $\jmath \geq \jmath_0$,*

$$\|\phi - \varphi_\jmath\|_{L^\infty(\mathbb{R})} \leq C 2^{-\alpha\jmath}.$$

Proof. See [43, Proposition 6.5.2, p. 205]. □

2.2.2 Connection coefficients

Another prerequisite for the discretization of the PDE models introduced in §3 is the computation of the so-called connection coefficients. Let ϕ be a scaling function of regularity $H^1(\mathbb{R})$. For $\jmath \in \mathbb{N}$ and $(i,j) \in \mathbb{Z}^2$, we define the *connection coefficients* at level $\jmath$ as

$$a^\jmath_{i,j} := \int_{\mathbb{R}} \phi'_{\jmath,i} \phi'_{\jmath,j}. \tag{2.57}$$

It is easy to see that the connection coefficients are symmetric with respect to the space subscripts, i.e.,

$$a^\jmath_{i,j} = a^\jmath_{j,i}$$

for all $(i,j) \in \mathbb{Z}^2$. From the knowledge of the connection coefficients at level 0

$$a^0_{i,j} := \int_{\mathbb{R}} \phi'_{0,i} \phi'_{0,j} = \int_{\mathbb{R}} \phi'(\cdot - i)\phi'(\cdot - j), \tag{2.58}$$

all connection coefficients at level $\jmath$ can be deduced by

$$a^\jmath_{i,j} = 2^{2\jmath} a^0_{i,j}.$$

The 0-th level connection coefficients (2.58) are easily seen to depend only on the difference $i - j$. This motivates the introduction of the quantities

$$a_k = \int_{\mathbb{R}} \phi' \phi'(\cdot - k) \tag{2.59}$$

for $k \in \mathbb{Z}$. The numbers (2.59) are even with respect to k, i.e.,

$$a_{-k} = a_k. \tag{2.60}$$

From this knowledge, we can infer all other connection coefficients via

$$a^\jmath_{i,j} = 2^{2\jmath} a_{|i-j|}. \tag{2.61}$$

Practical computation

If the scaling function ϕ has compact support, the numbers a_k are zero when $|k|$ is greater than the support size of ϕ. In particular, if ϕ is a Daubechies scaling function of order M, then $a_k = 0$ for all $|k| \geq 2\mathsf{M} - 1$. As for the nonzero coefficients corresponding to $|k| \leq 2\mathsf{M} - 2$, they can be computed all at once by means of an eigenvalue problem based on the two-scale relation (2.12). The idea is described in [10, 40] for a general scaling function. Below, we give a statement that is restricted to a Daubechies scaling function of order $\mathsf{M} \geq 3$ but that includes a suitable normalization for the eigenvector.

Proposition 2.9. *If ϕ is a Daubechies scaling function of order $\mathsf{M} \geq 3$, then the vector*

$$\mathbf{a} = (a_{-2\mathsf{M}+2}, a_{-2\mathsf{M}+3}, \ldots, a_{2\mathsf{M}-3}, a_{2\mathsf{M}-2})^T \in \mathbb{R}^{4\mathsf{M}-3}$$

solves the eigenvector problem

$$\mathbf{G}\mathbf{a} = \frac{1}{4}\mathbf{a}, \tag{2.62a}$$

$$\boldsymbol{\mu}^T\mathbf{a} = -2, \tag{2.62b}$$

for the eigenvalue $1/4$, where $\mathbf{G}$ is the $(4m-3) \times (4m-3)$ matrix whose entries are

$$\mathsf{G}_{k\ell} = \sum_{n\in\mathbb{Z}} h_n h_{2k+n-\ell}, \qquad \text{for } (k,\ell) \in \{-2\mathsf{M}+2, \ldots, 2\mathsf{M}-2\}^2, \tag{2.63}$$

and $\boldsymbol{\mu}$ is the vector of second-order moments with components

$$\mu_\ell = \int_{\mathbb{R}} x^2 \phi(x-\ell)\,\mathrm{d}x, \qquad \text{for } \ell \in \{-2\mathsf{M}+2, \ldots, 2\mathsf{M}-2\}. \tag{2.64}$$

Proof. Taking the derivative of the two-scale relation (2.12), we obtain

$$\phi'(x) = 2\sqrt{2} \sum_{n\in\mathbb{Z}} h_n \phi'(2x-n). \tag{2.65}$$

Substitution of this into definition (2.59) yields

$$\begin{aligned}
a_k &= \int_{\mathbb{R}} \phi'(x)\phi'(x-k)\,\mathrm{d}x \\
&= 8 \sum_{m\in\mathbb{Z}} h_m \sum_{n\in\mathbb{Z}} h_n \int_{\mathbb{R}} \phi'(2x-m)\phi'(2x-2k-n)\,\mathrm{d}x \\
&= 4 \sum_{m\in\mathbb{Z}} h_m \sum_{n\in\mathbb{Z}} h_n \int_{\mathbb{R}} \phi'(y)\phi'(y-2k-n+m)\,\mathrm{d}x \\
&= 4 \sum_{m\in\mathbb{Z}} h_m \sum_{n\in\mathbb{Z}} h_n a_{2k+n-m} = 4 \sum_{\ell\in\mathbb{Z}} \Big(\sum_{n\in\mathbb{Z}} h_n h_{2k+n-\ell} \Big) a_\ell,
\end{aligned}$$

which is a finite sum. This proves (2.62a). To derive (2.62b), we notice that by orthonormality and by polynomial exactness up to degree $\mathsf{M} - 1 \geq 2$,

$$x^2 = \sum_{\ell\in\mathbb{Z}} \mu_\ell \phi(x-\ell).$$

Differentiating the above identity with respect to x, we get

$$2x = \sum_{\ell\in\mathbb{Z}} \mu_\ell \phi'(x-\ell),$$

the pointwise existence of ϕ' resulting from the fact that $\phi \in C^1(\mathbb{R})$ as soon as $\mathsf{M} \geq 3$. Multiplying both sides by $\phi'(x)$ and integrating over $\mathbb{R}$, we end up with

$$2\int_{\mathbb{R}} x\phi'(x)\,\mathrm{d}x = \sum_{\ell\in\mathbb{Z}} \mu_\ell a_\ell = \boldsymbol{\mu}^T \mathbf{a}.$$

An integration by parts shows that the left-hand side is equal to $-2\int_{\mathbb{R}}\phi = -2$, which completes the proof of (2.62b). □

The entries of the matrix $\mathbf{G}$, defined in (2.63), coincide with some of the autocorrelation coefficients introduced in (2.50). More specifically,

$$\mathsf{G}_{k\ell} = \gamma_{2k-\ell}.$$

After Proposition 2.6, the autocorrelation coefficients do not depend on the specific choice of phase for $\mathcal{L}$ within Daubechies general construction. It follows that the matrix $\mathbf{G}$ is the same for the minimal phase or least asymmetric or any other choice of $\mathcal{L}$.

We now turn to the normalization (2.62b). The second-order moments $\boldsymbol{\mu}$ defined in (2.64) do depend on the choice of phase for $\mathcal{L}$. However, we will show that the condition (2.62b) is equivalent to another normalization that does not depend on $\mathcal{L}$.

Proposition 2.10. *The normalization condition* $\boldsymbol{\mu}^T\mathbf{a} = -2$ *is equivalent to*

$$\sum_{\ell=-2\mathsf{M}+2}^{2\mathsf{M}-2} \ell^2 a_\ell = -2. \tag{2.66}$$

Proof. By the change of variable $y = x + \ell$, formula (2.64) becomes

$$\begin{aligned}\mu_\ell = \int_{\mathbb{R}} (y+\ell)^2\phi(y)\,\mathrm{d}y &= \int_{\mathbb{R}} y^2\phi(y)\,\mathrm{d}y + 2\ell\int_{\mathbb{R}} y\phi(y)\,\mathrm{d}y + \ell^2\int_{\mathbb{R}}\phi(y)\,\mathrm{d}y \\ &= \mu_0 + 2\lambda_0\ell + \ell^2,\end{aligned}$$

with $\lambda_0 = \int_{\mathbb{R}} y\phi(y)\,\mathrm{d}y$. Hence,

$$\boldsymbol{\mu}^T\mathbf{a} = \mu_0\sum_{\ell\in\mathbb{Z}} a_\ell + 2\lambda_0\sum_{\ell\in\mathbb{Z}} \ell a_\ell + \sum_{\ell\in\mathbb{Z}} \ell^2 a_\ell. \tag{2.67}$$

The first term of the right-hand side is zero because

$$\sum_{\ell\in\mathbb{Z}} a_\ell = \sum_{\ell\in\mathbb{Z}}\int_{\mathbb{R}} \phi'\phi'(\cdot-\ell) = \int_{\mathbb{R}} \phi'\left(\sum_{\ell\in\mathbb{Z}}\phi(\cdot-\ell)\right)' \tag{2.68}$$

and thanks to the partition of unity property (2.42). The second term of the right-hand side of (2.67) is also zero because of the symmetry $a_{-\ell} = a_\ell$. The last sum in (2.67) is the left-hand side of (2.66). □

To summarize, the connection coefficients a_k are solutions of an eigenvector problem that does not depend on the specific choice of roots for $\mathcal{L}$. Thus, they depend only on M within Daubechies general construction of orthonormal compactly supported wavelets. With the help of Maple for the eigenvector problem (2.62a)–(2.66), we have computed "by hands" a few of these coefficients and they all turn out to be rational numbers, as reported in Tables 2.2–2.4. Our findings are confirmed by Goedecker's calculations [62, §31.3], of which we were not initially aware.

k	Exact value	Decimal approximation
0	295/56	$5.267857142857 \cdot 10^{0}$
1	−356/105	$-3.390476190476 \cdot 10^{0}$
2	92/105	$8.761904761904 \cdot 10^{-1}$
3	−4/35	$-1.142857142857 \cdot 10^{-1}$
4	−3/560	$-5.357142857143 \cdot 10^{-3}$

Table 2.2: Connection coefficients a_k for the Daubechies family $\mathsf{M} = 3$.

k	Exact value	Decimal approximation
0	342643/82248	$4.165973640696 \cdot 10^{0}$
1	−2852128/1079505	$-2.642070208105 \cdot 10^{0}$
2	12053651/17272080	$6.978691043580 \cdot 10^{-1}$
3	−162976/1079505	$-1.509728996160 \cdot 10^{-1}$
4	−60871/5757360	$-1.057272777801 \cdot 10^{-2}$
5	352/215901	$1.630376885702 \cdot 10^{-3}$
6	−55/3454416	$-1.592164927444 \cdot 10^{-5}$

Table 2.3: Connection coefficients a_k for the Daubechies family $\mathsf{M} = 4$.

k	Exact value	Decimal approximation
0	2370618501415/618154371936	$3.834994313783 \cdot 10^{0}$
1	−1632655076608/676106344305	$-2.414790351193 \cdot 10^{0}$
2	439132551286/676106344305	$6.495021899808 \cdot 10^{-1}$
3	−367031529728/2028319032915	$-1.809535500934 \cdot 10^{-1}$
4	80883901277/2704425377220	$2.990798043766 \cdot 10^{-2}$
5	−107449600/135221268861	$-7.946205571436 \cdot 10^{-4}$
6	−148937594/405663806583	$-3.671453838944 \cdot 10^{-4}$
7	−32000/19317324123	$-1.656544136043 \cdot 10^{-6}$
8	−4375/1236308743872	$-3.538760056244 \cdot 10^{-9}$

Table 2.4: Connection coefficients a_k for the Daubechies family $\mathsf{M} = 5$.

Approximation of the 1-D Laplacian operator

Let $u : \mathbb{R} \to \mathbb{R}$ be a smooth function. On the real line, consider the grid points $x_i = 2^{-\mathsf{J}} i$ for $i \in \mathbb{Z}$ at some fixed $\mathsf{J} \geq 0$, which correspond to a regular mesh of size $2^{-\mathsf{J}}$. Denoting by u_i the value of u at x_i, we introduce the operator $-\Delta_{\mathsf{M},\mathsf{J}}$ defined as

$$(-\Delta_{\mathsf{M},\mathsf{J}} u)_i = 2^{2\mathsf{J}} \sum_{k=-2\mathsf{M}+2}^{2\mathsf{M}-2} a_k u_{i+k}. \tag{2.69}$$

where the a_k's are the connection coefficients of the Daubechies family of order M.

Proposition 2.11. *The discrete operator $-\Delta_{\mathsf{M},\mathsf{J}}$ defined by (2.69) is a $(2\mathsf{M}-2)$-th order approximation of the 1-D continuous operator $-\Delta$ in the finite difference sense, that is,*

$$(-\Delta_{\mathsf{M},\mathsf{J}} u)_i = -u''(x_i) + O((2^{-\mathsf{J}})^{2\mathsf{M}-2}) \tag{2.70}$$

as $\mathsf{J} \to +\infty$ and for $u \in C^{2\mathsf{M}}(\mathbb{R})$.

Proof. From the Taylor expansion

$$u_{i\pm k} = u_i + \sum_{\ell=1}^{2\mathsf{M}-1} \frac{(\pm 2^{-\mathsf{J}} k)^\ell}{\ell!} u^{(\ell)}(x_i) + O((2^{-\mathsf{J}} k)^{2\mathsf{M}}),$$

we easily get

$$u_{i+k} - 2u_i + u_{i-k} = 2 \sum_{m=1}^{\mathsf{M}-1} \frac{(2^{-\mathsf{J}} k)^{2m}}{(2m)!} u^{(2m)}(x_i) + O((2^{-\mathsf{J}} k)^{2\mathsf{M}}).$$

Multiplying by a_k and summing over $k \in \{1, \ldots, 2\mathsf{M}-2\}$, we end up with

$$\sum_{k=1}^{2\mathsf{M}-2} a_k (u_{i+k} - 2u_i + u_{i-k}) = 2 \sum_{m=1}^{\mathsf{M}-1} \left(\sum_{k=1}^{2\mathsf{M}-2} k^{2m} a_k \right) \frac{(2^{-\mathsf{J}})^{2m}}{(2m)!} u^{(2m)}(x_i) + O((2^{-\mathsf{J}})^{2\mathsf{M}}). \tag{2.71}$$

We saw in (2.68) that

$$a_0 + 2 \sum_{k=1}^{2\mathsf{M}-2} a_k = \sum_{\ell=-2+2\mathsf{M}}^{2\mathsf{M}-2} a_\ell = 0. \tag{2.72a}$$

By Proposition 2.10, we have

$$\sum_{k=1}^{2\mathsf{M}-2} k^2 a_k = \frac{1}{2} \sum_{\ell=-2\mathsf{M}+2}^{2\mathsf{M}-2} \ell^2 a_\ell = -1. \tag{2.72b}$$

Furthermore, it is proven (see, e.g., Beylkin [10]) that the connection coefficients a_k satisfy

$$\sum_{k=1}^{2\mathsf{M}-2} k^4 a_k = \sum_{k=1}^{2\mathsf{M}-2} k^6 a_k = \ldots = \sum_{k=1}^{2\mathsf{M}-2} k^{2\mathsf{M}-2} a_k = 0. \tag{2.72c}$$

The equalities (2.72), along with the symmetry $a_{-k} = a_k$, allow us to rewrite (2.71) as

$$\sum_{k=-2\mathsf{M}+2}^{2\mathsf{M}-2} a_k u_{i+k} = -2^{-2\mathsf{J}} u''(x_i) + O((2^{-\mathsf{J}})^{2\mathsf{M}}).$$

Multiplication by $2^{2\mathsf{J}}$ finally yields the desired order (2.70). □

2.2.3 Periodic wavelets

So far, the functions u, $\phi(\cdot - n)$, $\psi(\cdot - n)$ have been defined on the entire real line and belong to $L^2(\mathbb{R})$. In chapters §3 and §4, however, we shall be considering bounded computational domains equipped with periodic boundary conditions. This can be dealt with by resorting to periodized scaling functions and wavelets introduced below. For convenience, let us assume that the domain of interest is the finite interval $[0, 1]$.

Definition 2.4. Let ϕ and ψ be the compactly supported and orthonormal scaling function and wavelet from a MRA. For any $(\jmath, n) \in \mathbb{Z}^2$, we define

- the 1-periodic scaling function

$$\widetilde{\phi}_{\jmath,n} = \sum_{k\in\mathbb{Z}} \phi_{\jmath,n}(\cdot + k) = 2^{\jmath/2} \sum_{k\in\mathbb{Z}} \phi(2^\jmath(\cdot + k) - n); \tag{2.73a}$$

- the 1-periodic wavelet

$$\widetilde{\psi}_{\jmath,n} = \sum_{k\in\mathbb{Z}} \psi_{\jmath,n}(\cdot + k) = 2^{\jmath/2} \sum_{k\in\mathbb{Z}} \psi(2^\jmath(\cdot + k) - n). \tag{2.73b}$$

The sums in the right-hand sides of (2.73) are well defined thanks to the compact supports: for each $x \in \mathbb{R}$ at which the sums are evaluated, only a finite number of terms are nonzeros.

The newly defined objects $\widetilde{\phi}_{\jmath,n}$ and $\widetilde{\psi}_{\jmath,n}$ can be seen as 1-periodic functions over $\mathbb{R}$ or functions over $[0, 1]$ subject to the periodic boundary conditions $\phi_{\jmath,n}(0) = \phi_{\jmath,n}(1)$, $\psi_{\jmath,n}(0) = \psi_{\jmath,n}(1)$. The second viewpoint makes sense provided that ϕ and ψ are continuous, which is the case if they are Daubechies scaling function and wavelet with order m large enough. The following Proposition recapitulates some basic properties of $\widetilde{\phi}_{\jmath,n}$ and $\widetilde{\psi}_{\jmath,n}$.

Proposition 2.12. *Assume that ϕ and ψ are Daubechies scaling function and wavelet of order $\mathsf{m} \geq 3$. Then,*

1. *For all $\jmath \leq 0$ and $n \in \mathbb{Z}$, $\widetilde{\phi}_{\jmath,n}$ is a constant function, namely,*

$$\widetilde{\phi}_{\jmath,n}(\cdot) = 2^{\jmath/2}.$$

2. *For all $\jmath \leq -1$ and $n \in \mathbb{Z}$, $\widetilde{\psi}_{\jmath,n}$ is a constant function, namely,*

$$\widetilde{\psi}_{\jmath,n}(\cdot) = 0.$$

3. *For all $\jmath \geq 1$ and $n \in \mathbb{Z}$, $\widetilde{\phi}_{\jmath,n}$ and $\widetilde{\psi}_{\jmath,n}$ are $2^\jmath$-periodic with respect to the shift parameter, namely,*

$$\widetilde{\phi}_{\jmath,n+2^\jmath k} = \widetilde{\phi}_{\jmath,n}, \qquad \widetilde{\psi}_{\jmath,n+2^\jmath k} = \widetilde{\psi}_{\jmath,n},$$

   for all $k \in \mathbb{Z}$.

4. *For all $\jmath$ such that $2^\jmath \geq 2\mathsf{m} - 1$ and $n \in \mathbb{Z}$, we have*

$$\widetilde{\phi}_{\jmath,n}(x) = \begin{cases} \phi_{\jmath,n}(x) & \text{if } x \in [0,1] \cap \operatorname{supp} \phi_{\jmath,n} \\ \phi_{\jmath,n}(x+1) & \text{if } x \in [0,1] \setminus \operatorname{supp} \phi_{\jmath,n}, \end{cases} \tag{2.74a}$$

   and

$$\widetilde{\psi}_{\jmath,n}(x) = \begin{cases} \psi_{\jmath,n}(x) & \text{if } x \in [0,1] \cap \operatorname{supp} \psi_{\jmath,n} \\ \psi_{\jmath,n}(x+1) & \text{if } x \in [0,1] \setminus \operatorname{supp} \psi_{\jmath,n}. \end{cases} \tag{2.74b}$$

PROOF. See Nielsen [111, §3.3]. □

Statements 1 and 2 tell us that at coarse scales ($\jmath \leq 0$ or $\jmath \leq -1$), the periodized scaling function and wavelet do not behave similarly as their counterparts in an infinite domain. The sums (2.73) are saturated to constant values because $\sum_{k\in\mathbb{Z}} \phi(\cdot + k) = 1$. Statement 3 mean that at a fixed level $\jmath \geq 1$, there are only $2^\jmath$ distinct periodized scaling functions and wavelets, which can be described by the subscript range $n \in \{0, \ldots, 2^\jmath - 1\}$. Statement 4 expresses the fact that when $2^\jmath$ is large enough relatively to support length $2\mathsf{M}-1$, then the functions involved in the sums (2.73) have disjoint supports. Consequently, the periodized scaling functions $\widetilde{\phi}_{\jmath,n}$ and wavelets $\widetilde{\psi}_{\jmath,n}$ coincide with their infinite counterparts $\phi_{\jmath,n}$ and $\psi_{\jmath,n}$ on $[0,1] \cap \operatorname{supp} \phi_{\jmath,n}$ and $[0,1] \cap \operatorname{supp} \psi_{\jmath,n}$. In other words, the restriction to $[0,1]$ of the periodized scaling function and wavelets is different from the original ones only when the latter have a support containing $x = 0$ or $x = 1$. In such a case, the periodized scaling function and wavelets are obtained by "wrapping around" the original ones. As a wrapped scaling function or wavelet cannot overlap itself when $2^\jmath \geq 2\mathsf{M} - 1$, this procedure gives rise to two disconnected components as illustrated in Figure 2.3.

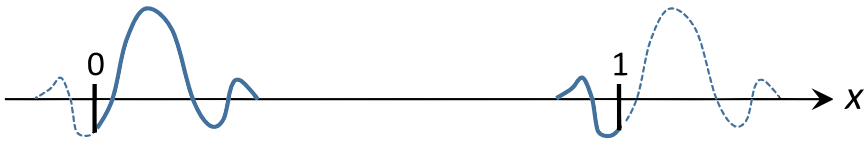

Figure 2.3: Restriction to $[0,1]$ of a periodized function without self-overlapping.

The periodized scaling functions and wavelets, restricted to $[0,1]$, clearly belong to $L^2(0,1)$. Furthermore, they generate an MRA of $L^2(0,1)$ analogous to that of $L^2(\mathbb{R})$, as clarified by the following Theorem.

Theorem 2.6. *For $\jmath \geq 0$, consider the $2^\jmath$-dimensional subspaces*

$$\widetilde{\mathscr{V}_\jmath} = \operatorname{Span}\{\widetilde{\phi}_{\jmath,n},\ 0 \leq n \leq 2^\jmath - 1\},$$
$$\widetilde{\mathscr{W}_\jmath} = \operatorname{Span}\{\widetilde{\psi}_{\jmath,n},\ 0 \leq n \leq 2^\jmath - 1\},$$

of $L^2(0,1)$. Then,

1. *The $\widetilde{\phi}_{\jmath,n}$, $0 \leq n \leq 2^\jmath - 1$, form an orthonormal basis of $\widetilde{\mathscr{V}_\jmath}$, that is,*

$$\int_0^1 \widetilde{\phi}_{\jmath,m}\widetilde{\phi}_{\jmath,n} = \delta_{m,n}. \tag{2.75a}$$

   for $(m,n) \in \{0, \ldots, 2^\jmath - 1\}^2$.

2. *The $\widetilde{\psi}_{\jmath,n}$, $0 \leq n \leq 2^\jmath - 1$, form an orthonormal basis of $\widetilde{\mathscr{W}_\jmath}$, that is,*

$$\int_0^1 \widetilde{\psi}_{\jmath,m}\widetilde{\psi}_{\jmath,n} = \delta_{m,n}. \tag{2.75b}$$

   for $(m,n) \in \{0, \ldots, 2^\jmath - 1\}^2$.

3. $\widetilde{\mathscr{W}_\jmath}$ is the orthogonal complement of $\widetilde{\mathscr{V}_\jmath}$ in $\widetilde{\mathscr{V}_{\jmath+1}}$, that is,

$$\widetilde{\mathscr{V}_\jmath} \oplus \widetilde{\mathscr{W}_\jmath} = \widetilde{\mathscr{V}_{\jmath+1}}.$$

4. The nested sequence

$$\{0\} \subset \widetilde{\mathscr{V}_0} \subset \widetilde{\mathscr{V}_1} \subset \cdots \widetilde{\mathscr{V}_\jmath} \subset \widetilde{\mathscr{V}_{\jmath+1}} \subset \cdots \subset L^2(0,1)$$

is a multiresolution analysis of $L^2(0,1)$ in a sense similar to that of Definition 2.1.

Proof. See Perrier and Basdevant [114] and Restrepo *et al.* [119]. □

The orthogonality property of Statement 3 implies that

$$\int_0^1 \widetilde{\phi}_{\jmath,n}\widetilde{\psi}_{\imath,m} = 0$$

for all $\imath \geq \jmath \geq 0$, $n \in \{0, \ldots, 2^\jmath - 1\}$, $m \in \{0, \ldots, 2^\imath - 1\}$. The MRA property of Statement 4 contains the completeness condition

$$L^2(0,1) = \overline{\bigcup_{\jmath \geq 0} \widetilde{\mathscr{V}_\jmath}},$$

which reads in turn

$$L^2(0,1) = \overline{\widetilde{V}_0 \oplus \widetilde{\mathscr{W}_0} \oplus \widetilde{\mathscr{W}_1} \oplus \widetilde{\mathscr{W}_2} \oplus \ldots}$$

This implies that the system

$$\{1\} \cup \{\widetilde{\psi}_{\jmath,n}\}_{\jmath \geq 0,\, 0 \leq n \leq 2^\jmath - 1}$$

is an orthonormal basis for $L^2(0,1)$. In this basis, any function $u \in L^2(0,1)$ can be decomposed into

$$u = \langle u, 1\rangle_{L^2(0,1)} 1 + \sum_{\jmath \geq 0} \sum_{n=0}^{2^\jmath - 1} \langle u, \widetilde{\psi}_{\jmath,n}\rangle_{L^2(0,1)} \widetilde{\psi}_{\jmath,n},$$

where the first term $\widetilde{P}_0 u = \langle u, 1\rangle_{L^2(0,1)} 1$ represents the orthogonal projection of u onto $\widetilde{V}_0$.

Error estimate for linear projection

More generally, for $\jmath \geq 0$, the orthogonal projection $\widetilde{P}_\jmath : L^2(0,1) \to \widetilde{V}_\jmath$ is defined as

$$\widetilde{P}_\jmath u = \sum_{n=0}^{2^\jmath - 1} \langle u, \widetilde{\phi}_{\jmath,n}\rangle_{L^2(0,1)} \widetilde{\phi}_{\jmath,n} = \widetilde{P}_0 u + \sum_{\imath=0}^{\jmath-1} \sum_{m=0}^{2^\imath - 1} \langle u, \widetilde{\psi}_{\imath,m}\rangle_{L^2(0,1)} \widetilde{\psi}_{\imath,m}.$$

As in the non-periodic case, $\widetilde{P}_\jmath$ is a natural approximation of u. In order to assess the quality of this linear approximation, let us introduce the periodic Sobolev space for $s \geq 0$ as

$$H^s_\#(0,1) = \left\{ u \in L^2(0,1) \,|\, \sum_{k \in \mathbb{Z}} (1 + |2\pi k|^2)^s |\widehat{u}_k|^2 < \infty \right\},$$

where $\widehat{u}_k = \int_0^1 u(x) \exp(-\mathrm{i}2\pi k x)\, \mathrm{d}x$ the k-th Fourier coefficient. This space is equipped with the norm

$$\|u\|^2_{H^s_\#(0,1)} = \sum_{k \in \mathbb{Z}} (1 + |2\pi k|^2)^s |\widehat{u}_k|^2.$$

Theorem 2.7. *Assume that the scaling function ϕ of order $\mathsf{M} \geq 1$ is such that*

$$\phi \in H^s(\mathbb{R}), \qquad \text{for some } s \in [0, \mathsf{M}).$$

Moreover, assume that we are given a function

$$u \in H^t_{\#}(0,1), \qquad \text{for some } t \in (s, \mathsf{M}).$$

Then, there exists a constant $\widetilde{\Gamma}_{\mathsf{M},s,t}$ (dependent on M, s and t but not on J) such that

$$\|u - \widetilde{P}_{\mathsf{J}} u\|_{H^s_{\#}(0,1)} \leq \widetilde{\Gamma}_{\mathsf{M},s,t} \|u\|_{H^s_{\#}(0,1)} 2^{-\mathsf{J}(t-s)} \tag{2.76}$$

for all $\mathsf{J} \geq 0$ large enough.

Proof. See Restrepo *et al.* [119], Restrepo and Leaf [118] or Walter and Cai [133]. Note that, as a consequence of the Poisson summation formula, $\phi \in H^s(\mathbb{R})$ implies $\widetilde{\phi}_{\mathsf{J},n} \in H^s_{\#}(0,1)$ for all (J, n), so that $\widetilde{P}_{\mathsf{J}} u \in H^s_{\#}(0,1)$ and the left-hand side of (2.76) is well-defined. □

This approximation result is to be compared with its infinite domain counterpart (2.44) in Theorem 2.5. It will be used in the *a priori* error estimate of §4.3.2.

Evaluation of function values at a given point

In some numerical simulations of §4 and §6, we need to be able to evaluate $\widetilde{\phi}_{\mathsf{J},n}(x)$ for $n \in \{0, \ldots, 2^{\mathsf{J}} - 1\}$ at some given location $x \in [0,1]$. To make this evaluation simpler, we shall always assume that $2^{\mathsf{J}} \geq 2\mathsf{M} - 1$. Indeed, thanks to property (2.74a) of Proposition 2.12, $\widetilde{\phi}_{\mathsf{J},n}(x)$ is equal either to $\phi_{\mathsf{J},n}(x)$ or $\phi_{\mathsf{J},n}(x+1)$. This brings us back to the cascade algorithm presented in §2.2.1.

Connection coefficients

Another prerequisite for the discretization of the PDE models introduced in §3 is the computation of the connection coefficients

$$\widetilde{a}^{\mathsf{J}}_{i,j} := \int_0^1 \widetilde{\phi}'_{\mathsf{J},i}\, \widetilde{\phi}'_{\mathsf{J},j} \tag{2.77}$$

for $(i,j) \in \{0, \ldots, 2^{\mathsf{J}} - 1\}^2$. In order to write down a closed-form expression for $\widetilde{a}^{\mathsf{J}}_{i,j}$, we need the following notation.

Definition 2.5. Given $(i,j) \in \{0, \ldots, 2^{\mathsf{J}} - 1\}^2$, the *periodized distance* between i and j is the non-negative integer $|i-j|^{\sim} \in \{0, \ldots, 2^{\mathsf{J}-1}\}$ defined as

$$|i-j|^{\sim} = \min\left\{|i-j|,\, 2^{\mathsf{J}} - |i-j|\right\}. \tag{2.78}$$

The periodized distance $|i-j|^{\sim}$ is the shortest distance between i and j, seen as points on the ordered discrete set $\{0, 1, \ldots, 2^{\mathsf{J}} - 1, 2^{\mathsf{J}}\}$ whose ends 0 and 2^{J} have been identified. This distance is always less than $2^{\mathsf{J}-1}$. The concept of periodized distance enables us to state the following result, that will be used in §4.3 and §4.4.

Proposition 2.13. *If $2^{\mathsf{J}} \geq 4\mathsf{M}-2$, then for $(i,j) \in \{0,\dots,2^{\mathsf{J}}-1\}^2$,*

$$\widetilde{a}^{\mathsf{J}}_{i,j} = 2^{2\mathsf{J}}\, a_{|i-j|^{\sim}} \tag{2.79}$$

where the a_k's are the infinite domain connection coefficients defined in (2.59) *and $|i-j|^{\sim}$ the periodized distance of Definition 2.5.*

Proof. Since periodization and derivation commute when ϕ is of compact support, we have

$$\widetilde{a}^{\mathsf{J}}_{i,j} = \int_0^1 \widetilde{\phi}'_{\mathsf{J},i}\,\widetilde{\phi}'_{\mathsf{J},j} = \int_0^1 \widetilde{\phi}'_{\mathsf{J},i} \sum_{\ell\in\mathbb{Z}} \phi'_{\mathsf{J},j}(\cdot+\ell) = \sum_{\ell\in\mathbb{Z}} \int_0^1 \widetilde{\phi}'_{\mathsf{J},i}\,\phi'_{\mathsf{J},j}(\cdot+\ell).$$

By the change of variable $y = x+\ell$ and thanks to the periodicity of $\widetilde{\phi}'_{\mathsf{J},i}$, the above expression can be turned into

$$\widetilde{a}^{\mathsf{J}}_{i,j} = \sum_{\ell\in\mathbb{Z}} \int_{\ell}^{\ell+1} \widetilde{\phi}'_{\mathsf{J},i}\,\phi'_{\mathsf{J},j} = \int_{\mathbb{R}} \widetilde{\phi}'_{\mathsf{J},i}\,\phi'_{\mathsf{J},j}.$$

It follows that

$$\widetilde{a}^{\mathsf{J}}_{i,j} = \int_{\mathbb{R}} \Big[\sum_{k\in\mathbb{Z}} \phi'_{\mathsf{J},i}(\cdot+k)\Big]\phi'_{\mathsf{J},j} = \sum_{k\in\mathbb{Z}} \int_{\mathbb{R}} \phi'_{\mathsf{J},i-2^{\mathsf{J}}k}\,\phi'_{\mathsf{J},j} = \sum_{k\in\mathbb{Z}} a^{\mathsf{J}}_{i-2^{\mathsf{J}}k,j},$$

where the equality $\phi'_{\mathsf{J},i}(\cdot+k) = \phi'_{\mathsf{J},i-2^{\mathsf{J}}k}$ comes from differentiating $\phi_{\mathsf{J},i}(\cdot+k) = \phi_{\mathsf{J},i-2^{\mathsf{J}}k}$, the latter being easy to establish. By virtue of (2.60)–(2.61), we obtain

$$\widetilde{a}^{\mathsf{J}}_{i,j} = 2^{2\mathsf{J}} \sum_{k\in\mathbb{Z}} a_{i-j-2^{\mathsf{J}}k} = 2^{2\mathsf{J}} \sum_{k\in\mathbb{Z}} a_{|i-j-2^{\mathsf{J}}k|}. \tag{2.80}$$

For a_n to be nonzero, it is necessary that $-2\mathsf{M}+2 \leq n \leq 2\mathsf{M}-2$, which makes for $4\mathsf{M}-3$ consecutive slots of feasible subscript n. Since $2^{\mathsf{J}} \geq 4\mathsf{M}-2$ by assumption, when k is shifted by ± 1, $2^{\mathsf{J}}k$ is shifted strictly more than $\pm(4\mathsf{M}-3)$. Therefore, the first infinite sum in (2.80) boilds down to at most a single term, the one that would possibly correspond to $|i-j-2^{\mathsf{J}}k| \leq 2\mathsf{M}-2$. Anyhow, we can assert that

$$\widetilde{a}^{\mathsf{J}}_{i,j} = 2^{2\mathsf{J}} a_{|i-j-2^{\mathsf{J}}k(i,j)|},$$

where $k(i,j) \in \mathbb{Z}$ is the unique relative integer such that

$$-2^{\mathsf{J}-1}+1 \leq i-j-2^{\mathsf{J}}k(i,j) \leq 2^{\mathsf{J}-1},$$

because $2^{\mathsf{J}-1} \geq 2\mathsf{M}-1$ and so $2^{\mathsf{J}-1}-1 \geq 2\mathsf{M}-2$. The existence and uniqueness of such a $k(i,j)$ can be proven by carrying out the Euclidean division of $i-j+2^{\mathsf{J}-1}-1$ by 2^{J}, i.e.,

$$i-j+2^{\mathsf{J}-1}-1 = 2^{\mathsf{J}}k(i,j) + r(i,j)$$

in which $r(i,j) \in \{0,\dots,2^{\mathsf{J}}-1\}$ is the remainder. Then,

$$i-j-2^{\mathsf{J}}k(i,j) = r(i,j) - 2^{\mathsf{J}-1} + 1 \in \{-2^{\mathsf{J}-1}+1,\dots,2^{\mathsf{J}-1}\}.$$

A more careful scrutiny, taking into account the fact that $(i,j) \in \{0,\dots 2^{\mathsf{J}}-1\}^2$, reveals that $k(i,j) \in \{-1,0,1\}$ and that in all cases

$$|i-j-2^{\mathsf{J}}k(i,j)| = |i-j|^{\sim},$$

which is the desired result. □

2.3 Wavelets for DFT

For the sake of completeness, we shortly present how wavelets have been applied in the context of DFT calculations. More specifically, we will describe the main features of BigDFT, a GPL software for DFT computations that makes use of Daubechies wavelets to express the Kohn-Sham orbitals [95].

2.3.1 Wavelets in 3-D

Consider a uniform grid of $\mathbb{R}^3$ with grid spacing 1 in every single dimension. We define a multidimensional [77, 108] multiresolution analysis in $\mathbb{R}^3$ by taking the tensor product

$$\begin{aligned}\mathscr{V}_{\jmath}^{(3\mathrm{D})} &= \mathscr{V}_{\jmath} \otimes \mathscr{V}_{\jmath} \otimes \mathscr{V}_{\jmath} \\ &= \overline{\mathrm{Span}}\{\, \Phi_{\jmath,i,j,k}(x,y,z) = \phi_{\jmath,i}(x)\,\phi_{\jmath,j}(y)\,\phi_{\jmath,k}(z),\ \ (i,j,k) \in \mathbb{Z}^3 \,\},\end{aligned}$$

with the scaling function $\phi_{\jmath,n}$ previously defined. We would like to introduce the orthogonal complement $\mathscr{W}_{\jmath}^{(3\mathrm{D})}$ of $\mathscr{V}_{\jmath}^{(3\mathrm{D})}$ in $\mathscr{V}_{\jmath+1}^{(3\mathrm{D})}$. In order to simplify the notations, let us define

$$\mathscr{U}_{\jmath}^{\ell} := \begin{cases} \mathscr{V}_{\jmath} & \text{if } \ell = 0, \\ \mathscr{W}_{\jmath} & \text{if } \ell = 1. \end{cases}$$

Then, the detail space in dimension 3 is defined as

$$\mathscr{W}_{\jmath}^{(3\mathrm{D})} = \bigoplus_{(\ell_1,\ell_2,\ell_3)\in\{0,1\}^3\setminus(0,0,0)} \mathscr{U}_{\jmath}^{\ell_1} \otimes \mathscr{U}_{\jmath}^{\ell_2} \otimes \mathscr{U}_{\jmath}^{\ell_3}. \tag{2.81}$$

This means that it is spanned by the $2^3 - 1 = 7$ types of wavelets $\{\Psi^q_{\jmath,i,j,k},\ 1 \le q \le 7\}$, where

$$\begin{aligned}
\Psi^1_{\jmath,i,j,k} &= \psi_{\jmath,i}(x)\,\phi_{\jmath,j}(y)\,\phi_{\jmath,k}(z), \\
\Psi^2_{\jmath,i,j,k} &= \phi_{\jmath,i}(x)\,\psi_{\jmath,j}(y)\,\phi_{\jmath,k}(z), \\
\Psi^3_{\jmath,i,j,k} &= \phi_{\jmath,i}(x)\,\phi_{\jmath,j}(y)\,\psi_{\jmath,k}(z), \\
\Psi^4_{\jmath,i,j,k} &= \phi_{\jmath,i}(x)\,\psi_{\jmath,j}(y)\,\psi_{\jmath,k}(z), \\
\Psi^5_{\jmath,i,j,k} &= \psi_{\jmath,i}(x)\,\phi_{\jmath,j}(y)\,\psi_{\jmath,k}(z), \\
\Psi^6_{\jmath,i,j,k} &= \psi_{\jmath,i}(x)\,\psi_{\jmath,j}(y)\,\phi_{\jmath,k}(z), \\
\Psi^7_{\jmath,i,j,k} &= \psi_{\jmath,i}(x)\,\psi_{\jmath,j}(y)\,\psi_{\jmath,k}(z).
\end{aligned}$$

The definition (2.81) is therefore equivalent to

$$\mathscr{W}_{\jmath}^{(3\mathrm{D})} = \overline{\mathrm{Span}}\{\, \Psi^q_{\jmath,i,j,k},\ \ (i,j,k) \in \mathbb{Z}^3,\ \ 1 \le q \le 7 \,\}. \tag{2.82}$$

2.3.2 DFT in a wavelet basis

As discussed at length in §1.4.2 and [60, 95], Daubechies wavelets have four advantages for DFT calculations: (1) localization in real space, which allows for efficient algorithms, in particular linear scaling ones [64, 65]; (2) localization in Fourier space, which helps improving preconditioning; (3) orthonormality, which saves computational time and also

improves the condition number; (4) adaptivity, which is desirable for accuracy and efficiency. We refer the reader to the end of §1.4.2 for a review of previous works on wavelets in quantum chemistry.

Details of the implementation for DFT methods using such a basis is described in [60]. In particular, BigDFTuses two levels of basis functions: a level $\jmath$ of scaling functions and a level $\jmath$ of wavelets defined via the formalism of §2.3.1. The nodes of the discretization are divided into 3 different types, as depicted in Figure 2.4. The nodes very far from the atoms will have zero charge density and therefore will not be associated to any basis function. The remaining grid points are either in the high resolution region $\mathcal{R}_2$ which contains the chemical bonds or in the low resolution region $\mathcal{R}_1$ which contains the exponentially decaying tails of the wavefunctions. In the low resolution region $\mathcal{R}_1$, we use only one scaling funtion $\Phi_{\jmath,i,j,k}$ per coarse grid point, whereas in the high resolution region $\mathcal{R}_2$, we use both the scaling function and the 7 wavelets $\Psi^q_{\jmath,i,j,k}$. Hence, in the high resolution region, the resolution is doubled in each spatial dimension compared to the low resolution region.

A molecular orbital φ, such as defined in (1.28), can be expanded in this basis as

$$\varphi(x,y,z) = \sum_{(i,j,k)\in\mathcal{R}_1} c_{i,j,k}\Phi_{\jmath,i,j,k}(x,y,z) + \sum_{(i,j,k)\in\mathcal{R}_2}\sum_{q=1}^{7} d^q_{i,j,k}\Psi^q_{\jmath,i,j,k}(x,y,z) \tag{2.83}$$

after rescaling of $\Phi_{\jmath,i,j,k}$ and $\Psi^q_{\jmath,i,j,k}$ to match an imposed grid size h. The decomposition of scaling functions into coarser scaling functions and wavelets can be continued recursively to obtain more than 2 resolution levels. However, a high degree of adaptivity is not of essential importance in pseudopotential calculations. In fact, the pseudopotentials smooth the wavefunctions so that two levels of resolution are enough in most cases to achieve good computational accuracy. In addition, more than two resolution levels lead to more complicated algorithms.

The molecular orbitals are stored in a compressed form where only the nonzero scaling function and wavelet coefficients are kept. The basis being orthogonal, several operations such as scalar products among different orbitals and the projector of the nonlocal pseudopotential can directly be done in this compressed form. The number N_{basis} of degrees of freedom "per wave function" scales linearly with the number N_{atom} of atoms. Since the number N_{orb} of molecular orbitals have the same scaling and considering that the Hamiltonian and overlap matrices among molecular orbitals have to be calculated by scalar product, the overall complexity scales roughly in $O(N^2_{orb}N_{basis}) = O(N^3_{atom})$ with respect to the number of atoms. In the first versions of BigDFT, this cubic scaling could not be improved. Recently, a breakthrough was made by Mohr *et al.* [105] using an intermediate basis consisting of "support functions" (generalizing Wannier functions). Linear scaling $O(N_{atom})$ has thus been achieved and is now available in BigDFT.

In this thesis, we will not consider a basis of scaling functions plus one or more levels of wavelets, but instead, a basis of scaling functions plus some contracted or optimal Gaussians. In order to test and define a strategy for this new method, we will introduce simplified 1-D models in chapter §3.

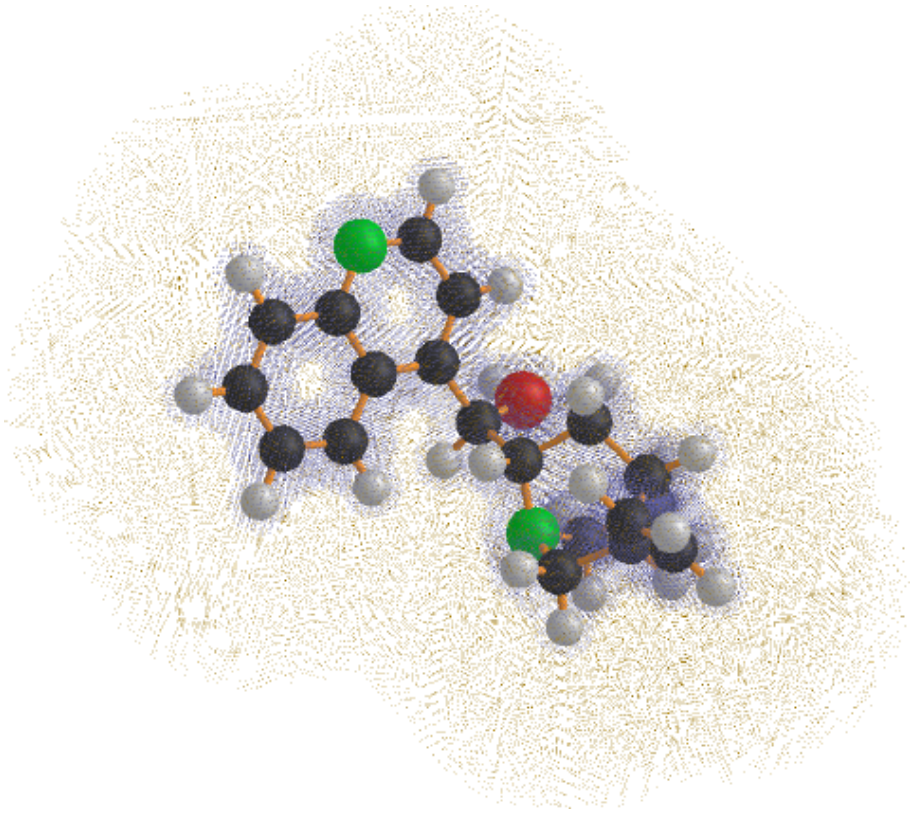

Figure 2.4: Computational domain in BigDFT: high-resolution region with 2 levels (blue dots), low-resolution region with 1 level (yellow dots), far away region with no grid point (white).

Chapter 3

One-dimensional models with cusp behaviors

Contents

Nous présentons deux modèles linéaires 1-D pour la simulation des systèmes moléculaires exhibant des points de rebroussement dans la fonction d'onde. Le premier modèle, posé en domaine infini, est une simplification de l'équation de Schrödinger 3-D pour un électron. Le deuxième modèle est la transposition du premier en domaine périodique. L'avantage de ces modèles réside dans la facilité d'implémentation, ce qui permet de nous concentrer sur la capture des singularités.

Les points de rebroussement sont crées par un potentiel de type Dirac, qui est l'équivalent 1-D du potentiel coulombien en 2-D et 3-D. Cette idée remonte à Frost [57] pour des systèmes à un ou deux atomes et revient sporadiquement dans la litérature. N'ayant connaissance d'aucun travail théorique antérieur sur de tels systèmes, nous entreprenons ici leur analyse mathématique pour un nombre arbitraire de noyaux : existence et unicité de l'état fondamental, régularité des fonctions d'onde et bornes sur le niveau fondamental, expression analytique ou semi-analytique des solutions. Ce dernier aspect facilite d'ailleurs l'étude des erreurs d'approximation.

3.1 Multi-delta model in an infinite domain

Starting from the 3-D linear Schrödinger equation (1.6)–(1.8), we apply the following series of simplifications: (i) go from 3-D to 1-D; (ii) consider just one electron; (iii) replace Coulomb potentials by Dirac delta potentials. The resulting toy model, set on an infinite domain, is shown to be a good one from the standpoint of cusp behavior and to enjoy many other favorable mathematical properties.

3.1.1 Physical ideas

We consider a one-dimensional system consisting of one electron and $M \geq 1$ nuclei of known charges $(Z_1, Z_2, \ldots, Z_M) \in (\mathbb{R}_+^*)^M$ located at known positions $(X_1, X_2, \ldots, X_M) \in \mathbb{R}^M$ such that

$$X_1 < X_2 < \ldots < X_M.$$

The state of this electron is described by a wave function $u : x \in \mathbb{R} \mapsto u(x) \in \mathbb{R}$. Only those wave functions satisfying

$$-\frac{1}{2}u'' + \left(-\sum_{I=1}^{M} Z_I \delta_{X_I}\right) u = Eu, \tag{3.1a}$$

$$\int_{\mathbb{R}} |u|^2 = 1, \tag{3.1b}$$

are relevant to characterize the state of the system. In (3.1), the unknowns are the wave function u and the energy E. The second line (3.1b) is the normalization condition for u, since $|u|^2$ represents a density of probability of presence. The first line is a Schrödinger equation, whose potential

$$V(x) = -\sum_{I=1}^{M} Z_I \delta_{X_I}(x)$$

represents the attraction generated by the nuclei on the single electron. This potential is a linear combination of M Dirac masses $\delta_{X_I} = \delta_0(\cdot - X_I)$, located at the X_I's.

There might be several solutions (u, E) to problem (3.1). While we shall be primarily interested in the "ground state" or "fundamental state" solution (u_*, E_*) that corresponds to the lowest possible energy level E, it is informative to keep an eye on all other solutions, called "excited states."

Delta potential and cusp generation

The delta potential $-\delta_{X_I}(\cdot)$ is the 1-D counterpart of the 2-D or 3-D Coulomb potential $-1/|\cdot - \mathbf{X}_I|$. Indeed, the 1-D Coulomb potential is not "stiff" enough to cause a cusp to appear at a nucleus position. The idea of using delta potentials in 1-D toy models dates back to Frost [57] for systems with one or two atoms. Thus, equation (3.1a) is best thought of as the 1-D counterpart of the 3-D one-electron models

$$-\frac{1}{2}\Delta u + \left(-\sum_{I=1}^{M} \frac{Z_I}{|\cdot - \mathbf{X}_I|}\right) u = Eu,$$

where $u \in H^1(\mathbb{R}^3)$. It should not be regarded as the 1-D equivalent of the 3-D equations

$$-\frac{1}{2}\Delta u - \sum_{I=1}^{M} Z_I \delta_{\mathbf{X}_I} u = Eu.$$

Indeed, the latter are ill-behaved and do not have finite binding energy, even though they are often used in some models of inter-particle interactions for condensed matter [47, 58].

The 1-D model (3.1) is capable of reproducing all the cusp properties that we already know for the 3-D Schrödinger equation with Coulomb potentials. In §1.1.2, we mentioned that at a nuclear coalescence, the Kato condition [79] must hold. For a one-electron system, the Kato condition (1.19) reads

$$\lim_{\epsilon \downarrow 0} \frac{1}{|S|} \oint_S \nabla u(\mathbf{X}_I + \epsilon \mathbf{n}) \cdot \mathbf{n} = -Z_I u(\mathbf{X}_I),$$

where the limits in the left-hand sides denote the average of all directional derivatives of u at $\mathbf{X}_I$, with $S = \{\mathbf{n} \in \mathbb{R}^3 \text{ such that } |\mathbf{n}| = 1\}$ being the unit sphere. In Theorem 3.2, it will be proven that any solution of the 1-D model (3.1) satisfies

$$\frac{u'(X_I^+) - u'(X_I^-)}{2} = -Z_I u(X_I),$$

whose left-hand side is clearly the average of the two possible directional derivatives of u at X_I. In §1.3.1 and §1.1.2, we saw that the ground state of the Schrödinger equation

$$-\frac{1}{2}\Delta u - \frac{Z}{|\cdot - \mathbf{X}|} u = Eu$$

for the single-electron single-nucleus case is

$$\begin{aligned} u_*(\mathbf{x}) &= \pi^{-1/2} Z^{3/2} \exp(-Z|\mathbf{x} - \mathbf{X}|) \quad \in H^{5/2-\epsilon}(\mathbb{R}^3), \\ E_* &= -\frac{1}{2} Z^2, \end{aligned}$$

which highlights the role of the Slater function

$$\mathrm{S}_{Z,\mathbf{X}}(\cdot) = \exp(-Z|\cdot - \mathbf{X}|).$$

In Theorem 3.4 and Corollary 3.2, we will show that the ground state for the single-delta model (3.1) is

$$\begin{aligned} u_*(x) &= Z^{1/2} \exp(-Z|x - X|) \quad \in H^{3/2-\epsilon}(\mathbb{R}), \\ E_* &= -\frac{1}{2} Z^2, \end{aligned}$$

and again we meet the Slater function $\mathrm{S}_{Z,X}$. As a matter of fact, the latter will play a major role in our 1-D model, to the extent that every solution of (3.1) for all $M \geq 1$ will be shown (Theorem 3.3) to be a superposition of M Slater functions S_{ζ,X_I}, where $\zeta > 0$ is a zero of some nonlinear equation. This extremely constraining feature of the solutions is specific to the 1-D nature of the model and does not arise in 3-D.

So far, we have put forward the advantages of the 1-D model (3.1), the most prominent of which is the ease of implementation. This model also has a few shortcomings. As can be seen from the above formulae, the first price to be paid for simplicity is the low regularity of the solution we want to capture: the Sobolev regularity of u_* is $3/2 - \epsilon$ in 1-D, instead of $5/2 - \epsilon$ in 3-D. Although this will certainly worsen the quality of the numerical approximation, this does not have too much a negative impact on the study of mixed bases.

Energy viewpoint

Perhaps the most troubling flaw of the 1-D model (3.1) comes from the observation that, due to the singularity of the Dirac delta, we can no longer define the "Hamiltonian operator" in the customary way. As introduced in §1.1.1, the Hamiltonian operator

$$\mathcal{H} = -\frac{1}{2}\Delta + V$$

for a one-electron system in 3-D is traditionally defined on the domain $H^2(\mathbb{R}^3)$, where both Δu and Vu belong to $L^2(\mathbb{R}^3)$. In our 1-D problem, taking $u \in H^2(\mathbb{R})$ does not guarantee that $\delta_{X_I} u$ belongs to $L^2(\mathbb{R})$. The spectral theory of operators is ineffective here, since the very notion of operator has become dubious. This is the reason why we avoided talking about the "eigenvector" u or the "eigenvalue" E at the beginning of this section.

The correct approach to define the "eigenvalues" E is the variational one, by means of an extended Courant-Fischer principle. For instance, from now on, the fundamental energy E_* is declared to be

$$E_* = \inf_{u \in \mathscr{V}} \frac{\mathfrak{a}(u,u)}{\mathfrak{b}(u,u)} \tag{3.2}$$

with the space

$$\mathscr{V} = H^1(\mathbb{R}) = \{u \in L^2(\mathbb{R}) \,|\, u' \in L^2(\mathbb{R})\}$$

and the bilinear forms

$$\mathfrak{a}(u,v) = \frac{1}{2}\int_{\mathbb{R}} u'v' - \sum_{I=1}^{M} Z_I u(X_I) v(X_I) \tag{3.3a}$$

$$\mathfrak{b}(u,v) = \int_{\mathbb{R}} uv \tag{3.3b}$$

for $(u,v) \in \mathscr{V}^2$. Since $\mathfrak{a}$ and $\mathfrak{b}$ are homogeneous of degree 2, a straightforward reformulation of (3.2) is

$$E_* = \inf_{\substack{u \in \mathscr{V} \\ \mathfrak{b}(u,u)=1}} \mathfrak{E}(u), \tag{3.4}$$

where

$$\mathfrak{E}(u) = \frac{1}{2}\int_{\mathbb{R}} |u'|^2 - \sum_{I=1}^{M} Z_I |u(X_I)|^2 = \mathfrak{a}(u,u) \tag{3.5}$$

is the energy functional. We remind [2] that $H^1(\mathbb{R}) \subset C^0(\mathbb{R})$, i.e., every function in $H^1(\mathbb{R})$ is continuous[1]. Therefore, $u(X_I)$ and $v(X_I)$ are well-defined, and consequently, $\mathfrak{a}$, $\mathfrak{b}$ and $\mathfrak{E}$ are well-defined as well. However, there remain a few difficulties with the energy viewpoint (3.4), namely:

1. We do not know whether or not the fundamental energy E_* exists. For the infimum problem (3.4) to make sense, the energy functional $\mathfrak{E}$ must be bounded from below on the unit sphere $\mathfrak{b}(u,u) = 1$.

2. Even if the fundamental energy E_* exists, we do not know whether or not there exists some $u_* \in \mathscr{V}$ such $E_* = \mathfrak{E}(u_*)$. From now on, such a minimizer u_* is declared to be a ground state, should it exist.

[1]This property is no longer true in higher dimensions.

3. Even if the infimum E_* is achieved at some minimizer $u_* \in \mathscr{V}$, we do not know whether or not such a ground state u_* is unique.

Questions 1 and 2 will be addressed in §3.1.2 using the techniques explained in the book by Cancès, Le Bris and Maday [22]. The reader who is not interested in the issue of existence and uniqueness is advised to go directly to §3.1.3, where we derive the properties shared by all solutions of (3.1), with a special focus on single-delta and double-delta potentials. These properties will, in turn, help us tackle question 3 in §3.1.5.

Variational formulation

Before embarking on this journey, let us broaden the scope of (3.4) and consider the problem of seeking all critical points of the energy functional $\mathfrak{E}$ on the manifold

$$\mathfrak{I}(u) := \|u\|^2_{L^2(\mathbb{R})} - 1 = 0.$$

The following Proposition characterizes such a critical point —should it exist at all— as a solution in the variational sense of the 1-D model (3.1).

Proposition 3.1. *Every critical point $u \in \mathscr{V}$ of the energy functional $\mathfrak{E}$, subject to the constraint $\mathfrak{I}(u) = 0$, is necessarily a solution of* (3.1) *in the variational sense:*

$$\mathfrak{a}(u, v) = E\, \mathfrak{b}(u, v), \tag{3.6a}$$

$$\mathfrak{b}(u, u) = 1, \tag{3.6b}$$

for all $v \in \mathscr{V}$. Furthermore,

$$E = \mathfrak{E}(u). \tag{3.7}$$

Proof. By the Euler-Lagrange optimality condition, a critical point u for $\mathfrak{E}$ subject to the constraint $\mathfrak{I}(u) = 0$ is as a point at which the gradient $\nabla_u \mathfrak{E}$ is collinear to that of $\nabla_u \mathfrak{I}$. In our case, it is easy to see that

$$\langle \nabla_u \mathfrak{E}, v\rangle_{\mathscr{V}'\times\mathscr{V}} = 2\mathfrak{a}(u, v),$$
$$\langle \nabla_u \mathfrak{I}, v\rangle_{\mathscr{V}'\times\mathscr{V}} = 2\mathfrak{b}(u, v),$$

for all $v \in \mathscr{V}$. The required collinearity implies that there exists a Lagrange multiplier $E \in \mathbb{R}$ such that $\nabla_u \mathfrak{E} = E\, \nabla_u \mathfrak{I}$ in $\mathscr{V}' = H^{-1}(\mathbb{R})$. Expressing this for all $v \in \mathscr{V}$, we end up with (3.6a). Setting $v = u$ in (3.6a) and invoking (3.6b), we obtain (3.7). □

The equality (3.7) means that the critical levels of $\mathfrak{E}$ are also "eigenvalues" of the Schrödinger problem (3.1), which is a feature of quadratic functionals. Of course, all of this is subject to the hypothetical existence of a critical point. Independently of any concern about energy, the variational formulation (3.6) is the correct sense in which problem (3.1) must be understood. It is also the starting point for building up a Galerkin approximation.

3.1.2 Existence of a ground state

Taking the infimum problem (3.4) as our point of departure, we wish to prove that it is well-defined and does have a minimizer u_*. We first recall a technical result.

Lemma 3.1. *Every function $u \in H^1(\mathbb{R})$ goes to zero at infinity, i.e.,*

$$\lim_{x\to\pm\infty} u(x) = 0. \tag{3.8}$$

PROOF. Let $\widehat{u}$ be the Fourier transform of u, which exists for $u \in L^2(\mathbb{R})$. The assumption $u \in H^1(\mathbb{R})$ is equivalent to

$$\int_{\mathbb{R}} (1+\xi^2)|\widehat{u}(\xi)|^2 \,\mathrm{d}\xi < \infty.$$

From the Cauchy-Schwarz inequality

$$\int_{\mathbb{R}} |\widehat{u}(\xi)| \,\mathrm{d}\xi \leq \Big(\int_{\mathbb{R}} \frac{\mathrm{d}\xi}{1+\xi^2} \Big)^{1/2} \Big(\int_{\mathbb{R}} (1+\xi^2)|\widehat{u}(\xi)|^2 \,\mathrm{d}\xi \Big)^{1/2}$$

whose right-hand side is well defined, we infer that $\widehat{u} \in L^1(\mathbb{R})$. It remains to apply the Riemann-Lebesgue lemma to $\widehat{u}$ to conclude about the limits (3.8). □

We said earlier that $\mathscr{V} = H^1(\mathbb{R}) \subset C^0(\mathbb{R})$. Lemma 3.1 gives us legitimacy to define the norm

$$\|v\|_{L^\infty(\mathbb{R})} = \sup_{x \in \mathbb{R}} |v(x)|$$

for all $v \in \mathscr{V}$. Throughout the remainder of section §3.1, we shall be writing

$$\|\cdot\|_{L^\infty},\ \|\cdot\|_{L^2},\ \|\cdot\|_{H^1} \quad \text{instead of} \quad \|\cdot\|_{L^\infty(\mathbb{R})},\ \|\cdot\|_{L^2(\mathbb{R})},\ \|\cdot\|_{H^1(\mathbb{R})}.$$

Thanks to the continuous embedding [2]

$$H^1(\mathbb{R}) \subset C^{0,1/2}(\mathbb{R}),$$

where $C^{0,1/2}$ denotes the Hölder space with exponent $1/2$, there exists *a fortiori* a constant $c > 0$ such that, for all $v \in \mathscr{V}$,

$$\|v\|_{L^\infty} \leq c\|v\|_{H^1}. \tag{3.9}$$

Neverthless, let us work out some finer control of the L^∞-norm for functions in $\mathscr{V}$.

Lemma 3.2. *For all $v \in \mathscr{V}$ and for all $\theta > 0$,*

$$\|v\|_{L^\infty}^2 \leq \theta\|v'\|_{L^2}^2 + \frac{1}{\theta}\|v\|_{L^2}^2. \tag{3.10}$$

PROOF. For any $v \in \mathscr{V}$ and $x, y \in \mathbb{R}$ we have

$$|v(x)|^2 - |v(y)|^2 = \int_y^x 2v'(t)v(t) \,\mathrm{d}t.$$

Letting $y \to -\infty$ and recalling that $v(y) \to 0$ by Lemma 3.1, we obtain

$$|v(x)|^2 = 2\int_{-\infty}^x v'(t)v(t) \,\mathrm{d}t \leq 2\|v'\|_{L^2}\|v\|_{L^2}$$

by the Cauchy-Schwarz inequality. Taking the supremum in $x \in \mathbb{R}$, we end up with $\|v\|_{L^\infty}^2 \leq 2\|v'\|_{L^2}\|v\|_{L^2}$. Application of Young's inequality

$$\|v'\|_{L^2}\|v\|_{L^2} \leq \frac{\theta}{2}\|v'\|_{L^2}^2 + \frac{1}{2\theta}\|v\|_{L^2}^2$$

for $\theta > 0$ results in (3.10). □

Inequalities (3.9) and (3.10) have tremendous consequences on the properties of the bilinear form $\mathfrak{a}$ and the energy functional $\mathfrak{E}$. In the following Proposition, the first two statements are helpful for error estimates while the last one is exactly what we need for the well-posedness of the infimum problem (3.4).

Proposition 3.2. *The following properties hold true:*

1. *There exists $\kappa > 0$ such that for all $(v, w) \in \mathscr{V}^2$,*
$$|\mathfrak{a}(v, w)| \leq \kappa \|v\|_{H^1} \|w\|_{H^1}. \tag{3.11}$$
2. *There exists $\Theta > 0$ such that for all $v \in \mathscr{V}$,*
$$\mathfrak{a}(v, v) \geq \frac{1}{4}\|v\|^2_{H^1} - \Theta \|v\|^2_{L^2}. \tag{3.12}$$
3. *Over the L^2-unit sphere*
$$\mathscr{S} = \{v \in \mathscr{V} \,|\, \|v\|_{L^2} = 1\}, \tag{3.13}$$
the energy functional $\mathfrak{E}$ is bounded from below.

Proof. By the triangle inequality and from the definition of $\mathfrak{a}(\cdot, \cdot)$, we have

$$\begin{aligned} |\mathfrak{a}(v, w)| &\leq \frac{1}{2}\int_{\mathbb{R}} |v'||w'| + \sum_{I=1}^{M} Z_I |v(X_I)||w(X_I)| \\ &\leq \left(\frac{1}{2}\|v'\|^2_{L^2} + \sum_{I=1}^{M} Z_I |v(X_I)|^2\right)^{1/2} \left(\frac{1}{2}\|w'\|^2_{L^2} + \sum_{I=1}^{M} Z_I |w(X_I)|^2\right)^{1/2}, \end{aligned}$$

the last line being due to the Cauchy-Schwarz inequality. From $|v(X_I)| \leq \|v\|_{L^\infty}$ and $\|v\|_{L^\infty} \leq c\|v\|_{H^1}$ after (3.9), we infer that

$$|\mathfrak{a}(v, w)| \leq \left(\frac{1}{2}\|v'\|^2_{L^2} + \mathcal{Z}c\|v\|^2_{H^1}\right)^{1/2} \left(\frac{1}{2}\|w'\|^2_{L^2} + \mathcal{Z}c\|w\|^2_{H^1}\right)^{1/2},$$

where $\mathcal{Z} = \sum_{I=1}^{M} Z_I$ is the total charge. Taking $\kappa = 1/2 + \mathcal{Z}c$, we easily get (3.11).

To derive (3.12), we first notice that

$$\mathfrak{a}(v, v) \geq \frac{1}{2}\|v'\|^2_{L^2} - \mathcal{Z}\|v\|^2_{L^\infty} \geq \left(\frac{1}{2} - \mathcal{Z}\theta\right)\|v'\|^2_{L^2} - \frac{\mathcal{Z}}{\theta}\|v\|^2_{L^2}, \tag{3.14}$$

the last inequality being due to (3.10) of Lemma 3.2. Selecting $\theta > 0$ such that $\mathcal{Z}\theta = 1/4$ and writing $\|v'\|^2_{L^2} = \|v\|^2_{H^1} - \|v\|^2_{L^2}$, we obtain

$$\mathfrak{a}(v, v) \geq \frac{1}{4}\|v\|^2_{H^1} - \left(\frac{1}{4} + 4\mathcal{Z}^2\right)\|v\|^2_{L^2}$$

which proves (3.12) with $\Theta = 1/4 + 4\mathcal{Z}^2$.

In (3.14), we now select $\theta > 0$ such that $\mathcal{Z}\theta = 1/2$. This cancels out the first term in the lower bound and leaves us with

$$\mathfrak{E}(v) = \mathfrak{a}(v, v) \geq -2\mathcal{Z}^2, \tag{3.15}$$

for $\|v\|^2_{L^2} = 1$ when $v \in \mathscr{S}$. □

The facts that the energy functional $\mathfrak{E}$ is well-defined on $\mathscr{V}$ and that it is bounded from below on the unit sphere $\mathscr{S}$ testify to the well-posedness of the infimum problem (3.4). The following remark on the value of $E_* = \inf_{u\in\mathscr{S}} \mathfrak{E}(u)$ will help us shorten the proof of existence for a minimizer.

Lemma 3.3. $E_* < 0$.

Proof. Let $v_1(x) = Z_1^{1/2}\exp(-Z_1|x - X_1|)$. One easily computes

$$\frac{1}{2}\|v_1'\|_{L^2}^2 - Z_1|v_1(X_1)|^2 = -\frac{1}{2}Z_1^2,$$

so that

$$\mathfrak{E}(v_1) = -\frac{1}{2}Z_1^2 - \sum_{I=2}^{M} Z_I|v_1(X_I)|^2 < 0.$$

In addition, $\|v_1\|_{L^2}^2 = 1$. Therefore, $E_* \leq \mathfrak{E}(v_1) < 0$. □

The function v_1 introduced in the proof of Lemma 3.3 is, apart from a multiplicative constant, a Slater function about which more will be said in §3.1.3 and §3.1.4. More stringent bounds of E_* will be provided in Theorem 3.7.

Theorem 3.1. *Let $\mathscr{S}$ be the L^2-unit sphere defined in* (3.13)*. There exists a minimizer $u_* \in \mathscr{S} \subset \mathscr{V}$ such that*

$$E_* = \mathfrak{E}(u_*) = \min_{u\in\mathscr{S}} \mathfrak{E}(u).$$

Proof. In accordance with the strategy developed in [22, §3], we divide the proof into several steps. We go quickly over the easy ones, detailing only those requiring specific properties of the model.

1. Let $\{u_n\}_{n\geq 1}$ be a minimizing sequence, i.e., such that $u_n \in \mathscr{S}$ and $\lim_{n\to\infty} \mathfrak{E}(u_n) = E_*$. Then, there exists $C \in \mathbb{R}$ such that $\mathfrak{E}(u_n) \leq C$ for all $n \geq 1$. According to (3.12),
$$C \geq \mathfrak{E}(u_n) = \mathfrak{a}(u_n, u_n) \geq \frac{1}{4}\|u_n\|_{H^1}^2 - \Theta\|u_n\|_{L^2}^2.$$
From $\|u_n\|_{L^2} = 1$, we infer that $\|u_n\|_{H^1}^2 \leq 4(C + \Theta)$. Thus, the minimizing sequence $\{u_n\}_{n\geq 1}$ is bounded in $H^1(\mathbb{R})$.

2. We can therefore extract a subsequence, also denoted by $\{u_n\}_{n\geq 1}$, that converges weakly toward some element $u_* \in \mathscr{V}$. The weak convergence
$$u_n \overset{H^1}{\rightharpoonup} u_*$$
and the convexity of the functionals $v \mapsto \int_{\mathbb{R}} |v|^2$ and $v \mapsto \int_{\mathbb{R}} |v'|^2$ implies that
$$\|u_*\|_{L^2}^2 \leq \liminf_{n\to+\infty} \|u_n\|_{L^2}^2 = 1, \tag{3.16a}$$
$$\|u_*'\|_{L^2}^2 \leq \liminf_{n\to+\infty} \|u_n'\|_{L^2}^2. \tag{3.16b}$$

3. By the Rellich-Kondrashov theorem [2], we have the compact embedding
$$H^1(\Omega) \subset\subset C^0(\overline{\Omega})$$

for all bounded open set $\Omega \subset \mathbb{R}$. Taking Ω to be an open set containing $[X_1, X_M]$, we can extract from $\{u_n\}_{n\geq 1}$ a subsequence, again denoted by $\{u_n\}_{n\geq 1}$, that converges strongly toward u_* in $C^0(\overline{\Omega})$. This entails, in particular, that

$$u_*(X_I) = \lim_{n\to+\infty} u_n(X_I) \tag{3.17}$$

for all $1 \leq I \leq M$. Combining (3.16b) and (3.17), we end up with

$$\mathfrak{E}(u_*) \leq \liminf_{n\to+\infty} \mathfrak{E}(u_n) = E_*. \tag{3.18}$$

4. From the definition of the infimum, $E_* \leq \mathfrak{E}(u_*/\|u_*\|_{L^2})$. By homogeneity, we deduce that $E_*\|u_*\|_{L^2}^2 \leq \mathfrak{E}(u_*)$. The combination of this with (3.18) and Lemma 3.3 yields

$$E_*\|u_*\|_{L^2}^2 \leq \mathfrak{E}(u_*) \leq E_* < 0,$$

from which a division by $E_* < 0$ gives $\|u_*\|_{L^2}^2 \geq 1$. But by (3.16a), $\|u_*\|_{L^2}^2 \leq 1$. Hence, $\|u_*\|_{L^2}^2 = 1$ and $u_* \in \mathscr{S}$. This results in $E_* \leq \mathfrak{E}(u_*)$, and in view of (3.18), implies $E_* = \mathfrak{E}(u_*)$. □

3.1.3 Properties of all eigenstates

For the moment, we know that there is at least a minimizer u_* and such a ground state u_* is characterized by (3.1). In order to prove uniqueness —which will be done in §3.1.5— we have to know more about the properties of any possible ground state. We take this opportunity to investigate the properties of all solutions of (3.1), referred to as "eigenstates." Assuming that eigenstates exist, we derive necessary conditions to be satisfied by them.

Kato's condition

As mentioned in §3.1.1, the 1-D model (3.1) is able to mimick the cusp properties known for the 3-D Schrödinger equation with Coulomb potentials. Let us see the first of these.

Theorem 3.2. *The wave function u of every solution $(u, E) \in H^1(\mathbb{R}) \times \mathbb{R}$ of* (3.1)

1. *is infinitely differentiable on each open interval*

$$(-\infty, X_1), (X_1, X_2), \ldots, (X_{M-1}, X_M), (X_M, +\infty).$$

2. *has a jump in derivative at each nucleus location X_I, $1 \leq I \leq M$, that satisfies the Kato condition*

$$\frac{u'(X_I^+) - u'(X_I^-)}{2} = -Z_I u(X_I). \tag{3.19}$$

Proof. As said earlier, the sense to be given to (3.1) is the variational formulation (3.6). Plugging into (3.6) a test function $v \in C_0^\infty(\mathbb{R})$ whose support does not contain any X_I, we see that u is the solution in the sense of distributions of

$$u'' = -2Eu \tag{3.20}$$

on each interval $(-\infty, X_1)$, (X_1, X_2), … , (X_{M-1}, X_M), $(X_M, +\infty)$. However, we know that weak solutions of (3.20) coincide with strong solutions on an open interval. The

classical solution of (3.20) on each of these open intervals is the sum of at most two exponentials and is obviously infinitely differentiable.

Fix $I \in \{1, \dots, M\}$ and take a smooth test function v such that

$$v(X_I) \neq 0 \quad \text{and} \quad \operatorname{supp} v \subset (X_I - \epsilon, X_I + \epsilon),$$

with $\epsilon > 0$ small enough so that the interval $(X_I - \epsilon, X_I + \epsilon)$ does not contain any other nucleus. Integration by parts gives us

$$\begin{aligned} \int_{X_I-\epsilon}^{X_I} u'v' &= \quad u'(X_I^-)v(X_I) - \int_{X_I-\epsilon}^{X_I} u''v, \\ \int_{X_I}^{X_I+\epsilon} u'v' &= -u'(X_I^+)v(X_I) - \int_{X_I}^{X_I+\epsilon} u''v. \end{aligned}$$

Summing these two equalities and arguing that (3.20) holds on each open interval $(X_I - \epsilon, X_I)$ and $(X_I, X_I + \epsilon)$, we end up with

$$\int_{\mathbb{R}} u'v' = \int_{X_I-\epsilon}^{X_I} u'v' + \int_{X_I}^{X_I+\epsilon} u'v' = -[u'(X_I^+) - u'(X_I^-)]v(X_I) + 2E \int_{\mathbb{R}} uv.$$

After division by 2 and subtraction to the variational formulation (3.6a), the above equation leads to

$$-\frac{u'(X_I^+) - u'(X_I^-)}{2} v(X_I) = Z_I u(X_I) v(X_I).$$

A further simplification by $-v(X_I) \neq 0$ yields the Kato condition (3.19). □

Corollary 3.1. *The energy E of every solution $(u, E) \in H^1(\mathbb{R}) \times \mathbb{R}$ of (3.1) is necessarily negative, i.e.,*

$$E < 0. \tag{3.21}$$

Proof. If $E > 0$, the solutions of (3.1) are of oscillatory type. To fix ideas, on $(-\infty, X_1)$, a solution satisfies $u'' = -2Eu$ and must be of the form

$$u(x) = a_1 \cos(\sqrt{2E}x) + b_1 \sin(\sqrt{2E}x).$$

Such a function cannot be a bound state (which means $\|u\|_{L^2} < \infty$), unless $a_1 = b_1 = 0$. But then it vanishes identically on $(-\infty, X_1)$ and gives rise by continuity to $u(X_1) = u'(X_1^-) = 0$. From the Kato condition at X_1 we deduce that

$$u'(X_1^+) = u'(X_1^-) - 2Z_1 u(X_1) = 0.$$

Starting from $u(X_1) = u'(X_1^+) = 0$ and solving the differential equation (3.20) "eastward," we easily show that $u \equiv 0$ on (X_1, X_2). Repeating this procedure on and on, we show that u is identically zero on $\mathbb{R}$, which contradicts $\|u\|_{L^2} = 1$.

By a similar argument, we also succeed in excluding the hypothetical case $E = 0$, which completes the proof. □

Multi-Slater form

The second analogy between model (3.1) and the 3-D Schrödinger equation with Coulomb potentials is the role played by the Slater function. We shall be using the symbol

$$\mathrm{S}_{\zeta,X}(x) = \exp(-\zeta|x - X|) \tag{3.22}$$

for the Slater function centered at X and having $\zeta > 0$ as orbital exponent. For $M = 1$, the analogy is perfect, as will be elaborated on in Theorem 3.4. For $M \geq 2$, the Slater function keeps a very strong influence in the shape of a solution of problem (3.1). This feature differs from the situation in 3-D and can be attributed to the 1-D setting.

Theorem 3.3. *Every solution* $(u, E) \in H^1(\mathbb{R}) \times \mathbb{R}$ *of* (3.1) *is necessarily of the form*

$$u = \sum_{J=1}^{M} \frac{Z_J}{\zeta} u(X_J)\, \mathrm{S}_{\zeta,X_J}, \tag{3.23a}$$

$$E = -\frac{1}{2}\zeta^2, \tag{3.23b}$$

where

- $\zeta > 0$ *is a zero of the equation*

$$\det(\mathsf{C}^\zeta - \zeta \mathsf{I}) = 0, \tag{3.24a}$$

with I *the* $M \times M$ *identity matrix and* C^ζ *the* $M \times M$ matrix of compatibility *whose entries are*

$$\mathsf{C}^\zeta_{IJ} = Z_J \exp(-\zeta|X_I - X_J|); \tag{3.24b}$$

- *the vector* $\mathsf{u} \in \mathbb{R}^M$ *of components* $\mathsf{u}_J = u(X_J)$ *is a non-trivial solution of the* relations of compatibility

$$\mathsf{C}^\zeta \mathsf{u} = \zeta \mathsf{u}. \tag{3.25}$$

Furthermore, there are at most a finite number of distinct zeros ζ *for* (3.24a).

Proof. Since non-negative energies $E \geq 0$ have been ruled out by Corollary 3.1, we are restricted to looking for

$$E = -\frac{1}{2}\zeta^2, \qquad \text{for some } \zeta > 0.$$

The first equation of (3.1) can be put under the form

$$-\frac{1}{2}u'' - \sum_{J=1}^{M} Z_J u(X_J)\delta_{X_J} = -\frac{1}{2}\zeta^2 u.$$

Applying the Fourier transform

$$\widehat{u}(\xi) = \int_{\mathbb{R}} u(x) \exp(-\mathrm{i}\xi x)\, \mathrm{d}x \tag{3.26}$$

to both sides and using the properties $\widehat{u''}(\xi) = -\xi^2 \widehat{u}(\xi)$ and $\widehat{\delta_{X_J}}(\xi) = \exp(-\mathrm{i}X_J\xi)$, we end up with

$$\frac{\xi^2}{2}\widehat{u}(\xi) - \sum_{J=1}^{M} Z_J u(X_J) \exp(-\mathrm{i}X_J\xi) = -\frac{\zeta^2}{2}\widehat{u}(\xi).$$

From this, we can extract

$$\begin{aligned}\widehat{u}(\xi) &= \frac{2}{\zeta^2+\xi^2}\sum_{J=1}^{M} Z_J u(X_J)\exp(-\mathrm{i}X_J\xi)\\ &= \sum_{J=1}^{M} Z_J u(X_J)\frac{2\exp(-\mathrm{i}X_J\xi)}{\zeta^2+\xi^2}\\ &= \sum_{J=1}^{M} Z_J \frac{u(X_J)}{\zeta}\left\{\frac{1}{\zeta}\cdot\frac{2\exp(-\mathrm{i}X_J\xi))}{1+(\xi/\zeta)^2}\right\}. \end{aligned} \tag{3.27}$$

Let us recognize the expression in the brackets as the Fourier transform of some elementary function. Starting from the classical Fourier transform pair

$$\widehat{\mathrm{S}_{1,0}}(\xi) = \widehat{\exp(-|\cdot|)}(\xi) = \frac{2}{1+\xi^2},$$

we have by the dilation formula

$$\widehat{\mathrm{S}_{\zeta,0}}(\xi) = \widehat{\exp(-\zeta|\cdot|)}(\xi) = \frac{1}{\zeta}\cdot\frac{2}{1+(\xi/\zeta)^2},$$

and by the translation formula

$$\widehat{\mathrm{S}_{\zeta,X_J}}(\xi) = \widehat{\exp(-\zeta|\cdot - X_J|)}(\xi) = \frac{1}{\zeta}\cdot\frac{2\exp(-\mathrm{i}X_J\xi)}{1+(\xi/\zeta)^2}. \tag{3.28}$$

Taking the inverse Fourier transform of (3.27), we obtain the desired form (3.23a), that is,

$$u(x) = \sum_{J=1}^{M}\frac{Z_J}{\zeta}u(X_J)\,\exp(-\zeta|x-X_J|). \tag{3.29}$$

The values $u(X_J)$ cannot be prescribed freely. They are subject to the relations of compatibility

$$u(X_I) = \sum_{J=1}^{M}\frac{Z_J}{\zeta}\exp(-\zeta|X_I-X_J|)\,u(X_J),$$

that result from specifying $x = X_I$ in (3.29). Gathering all of these conditions for $I \in \{1,\dots,M\}$ and multiplying by ζ, we have the matrix-vector relation (3.25), where $\mathbf{C}^\zeta$ is defined in (3.24b). For a solution $\mathbf{u} \neq \mathbf{0}$ to exist, ζ must be an eigenvalue of the matrix $\mathbf{C}^\zeta$ (depending itself on ζ), whence the characterization (3.24a) for ζ.

Expanding the determinant (3.24a) using (3.24b) and regrouping terms multiplied by the same exponential, we see that it takes the form

$$f(\zeta) := \det(\mathbf{C}^\zeta - \zeta\mathbf{I}) = \sum_{\ell=1}^{\mathcal{M}}\exp(q_\ell\zeta)P_\ell(\zeta), \tag{3.30}$$

in which $\mathcal{M}$ is some finite integer, P_ℓ is a non-zero polynomial and the exponents $q_\ell \in \mathbb{R}$ are distinct from each other. Assume that f has an infinite number of distinct zeros. From this infinity, we can extract a countable sequence and we can assert that

$$\exp(-q_1\zeta)f(\zeta) = P_1(\zeta) + \sum_{\ell=2}^{\mathcal{M}}\exp((q_\ell-q_1)\zeta)P_\ell(\zeta)$$

has a countable infinity of distinct zeros. Applying Rolle's theorem $m_1 + 1$ times, where $m_1 = \deg P_1$, we can say that

$$f^{(1)}(\zeta) := \frac{\mathrm{d}^{m_1+1}}{\mathrm{d}\zeta^{m_1+1}} \exp(-q_1\zeta) f(\zeta) = \sum_{\ell=2}^{\mathcal{M}} \exp((q_\ell - q_1)\zeta) P_\ell^{(1)}(\zeta)$$

has a countable infinity of distinct zeros. Note that, here, $f^{(1)}$ and $P_\ell^{(1)}$ are just notations and do not mean first derivatives. Since $q_\ell - q_1 \neq 0$ for $2 \leq \ell \leq \mathcal{M}$, we are sure that $\deg P_\ell^{(1)} = \deg P_\ell =: m_\ell$. Multiplying by $f^{(1)}$ by $\exp((q_1 - q_2)\zeta)$ so as to clear the factor of $P_2^{(1)}$ and applying Rolle's theorem $m_2 + 1$ times, we can claim that

$$f^{(2)}(\zeta) := \frac{\mathrm{d}^{m_2+1}}{\mathrm{d}\zeta^{m_2+1}} \exp((q_1 - q_2)\zeta) f^{(1)}(\zeta) = \sum_{\ell=3}^{\mathcal{M}} \exp((q_\ell - q_2)\zeta) P_\ell^{(2)}(\zeta)$$

has a countable infinity of distinct zeros, with $\deg P_\ell^{(2)} = \deg P_\ell^{(1)} = m_\ell$. Again, note that $f^{(2)}$ and $P_\ell^{(2)}$ are just notations and do not mean second derivatives. Going on with this process, we end up with a countable infinity of distinct zeros for some non-zero polynomial $P_{\mathcal{M}}^{(\mathcal{M}-1)}$, which is impossible. From this contradiction, we conclude that f has at most a finite number of zeros. □

Should a solution of (3.1) exist, it is the superposition of M Slater functions S_{ζ,X_J} centered at each nucleus location and having the same orbital exponent ζ. The latter must be a zero of the "characteristic" equation (3.24a), whose number of solutions depend on the parameters (X_I, Z_I) of the problem. Even for a given "eigenvalue" $\zeta > 0$ of (3.24a), there might be several "eigenvectors" $\mathbf{u}$ that satisfy the compatibility system (3.25). The existence of a minimizer for $\mathfrak{E}$ does guarantee that there exists at least a solution ζ_* to (3.24a), the largest one (corresponding to the smallest E_*).

Sobolev regularity

As a consequence of the multi-Slater form (3.23a), we are going to determine the Sobolev regularity of any eigenstate u of (3.1). Let us recall that, for a given $s \in \mathbb{R}$,

$$u \in H^s(\mathbb{R}) \Longleftrightarrow \|u\|^2_{H^s(\mathbb{R})} := \int_{\mathbb{R}} (1 + |\xi|^2)^s |\widehat{u}(\xi)|^2 \,\mathrm{d}\xi < \infty, \tag{3.31}$$

where $\widehat{u}$ is the Fourier transform of u defined in (3.26).

Corollary 3.2. *If $(u, E) \in H^1(\mathbb{R}) \times \mathbb{R}$ is a solution of (3.1), then*

$$u \in H^{3/2-\epsilon}(\mathbb{R}) \qquad \textit{for all } \epsilon > 0.$$

Proof. We first prove that any Slater function $\mathrm{S}_{\zeta,X}$ belongs to $H^{3/2-\epsilon}(\mathbb{R})$ for any $\epsilon > 0$. Indeed, taking the square of the modulus in (3.28),

$$|\widehat{\mathrm{S}_{Z,X}}(\xi)|^2 = \frac{4\zeta^2}{(\zeta^2 + |\xi|^2)^2}.$$

Consequently, when $|\xi| \to \infty$,

$$(1+|\xi|^2)^s |\widehat{\mathrm{S}_{Z,X}}(\xi)|^2 \sim |\xi|^{2s} \cdot \frac{4\zeta^2}{|\xi|^4} = \frac{4\zeta^2}{|\xi|^{4-2s}}$$

assuming that $s \geq 0$. This assumption is justified since $\mathrm{S}_{\zeta,X} \in L^2(\mathbb{R}) = H^0(\mathbb{R})$. Therefore,

$$\int_{\mathbb{R}} (1+|\xi|^2)^s |\widehat{\mathrm{S}_{Z,X}}(\xi)|^2 \,\mathrm{d}\xi < \infty \iff \int_1^{+\infty} \frac{\mathrm{d}\xi}{|\xi|^{4-2s}} < \infty.$$

The convergence of the latter integral is equivalent to $4 - 2s > 1$, that is, $s < 3/2$. As a finite linear combination (3.23a) of Slater functions, u has at least the same Sobolev regularity $3/2 - \epsilon$. □

3.1.4 Single- and double-delta potentials

From the general form in Theorem 3.3, we can write down explicit formulae for the exact solutions in the special and important cases $M = 1$ and $M = 2$. The single-delta case is the prototype of an isolated atom, while the double-delta case is the prototype of a molecule.

Single-delta

When $M = 1$, the system consists of only one nucleus and one electron, which is a model of a hydrogenoid atom. The equations to be solved are

$$-\frac{1}{2}u'' - Z\delta_X u = Eu, \tag{3.32a}$$

$$\|u\|_{L^2(\mathbb{R})} = 1, \tag{3.32b}$$

for a given charge $Z > 0$ and a given nucleus location $X \in \mathbb{R}$.

Theorem 3.4. *The only solution of problem* (3.32) *is, up to sign of the wave function,*

$$u_* = Z^{1/2}\,\mathrm{S}_{Z,X}, \tag{3.33a}$$

$$E_* = -\frac{1}{2}Z^2. \tag{3.33b}$$

It is thus also the minimizer of $\mathfrak{E}(u) = \frac{1}{2}\|u'\|_{L^2}^2 - Z|u(X)|^2$ *on* $\mathscr{S}$.

Proof. The general form (3.23a) of Theorem (3.3) boils down in our case to

$$u(x) = \frac{Z}{\zeta} u(X) \exp(-Z|x - X|),$$

while $u(X)$ is subject to the relation of compatibility (3.25), which reads

$$Zu(X) = \zeta u(X).$$

If $u(X) = 0$, then u vanishes identically. Therefore, $u(X) \neq 0$ and $\zeta = Z$. The normalization (3.32b) gives $u(X) = \sqrt{Z}$. By Theorem 3.1, there exists a minimizer u_*. We also know that any minimizer must satisfy (3.32). Therefore, formulae (3.33) supply us with the only minimizer possible. □

The knowledge of the solution for a single-delta potential is helpful for finding bounds on the ground state energy in the case of a multi-delta potential. In §3.1.5, we shall be using Theorem 3.4 under the following form: for all $\zeta > 0$ and all $X \in \mathbb{R}$,

$$-\frac{1}{2}\zeta^2 = \min_{\substack{v\in H^1(\mathbb{R})\\ \|v\|_{L^2}=1}} \frac{1}{2}\|v'\|_{L^2}^2 - \zeta|v(X)|^2 = \frac{1}{2}\|\mathrm{S}'_{\zeta,X}\|_{L^2}^2 - \zeta|\mathrm{S}_{\zeta,X}(X)|^2. \tag{3.34}$$

Double-delta

When $M = 2$, the system consists of one electron orbiting two nuclei. The equations to be solved are

$$-\frac{1}{2}u'' - (Z_1\delta_{X_1} + Z_2\delta_{X_2})u = Eu, \tag{3.35a}$$

$$\|u\|_{L^2(\mathbb{R})} = 1. \tag{3.35b}$$

We are able to determine all eigenstates of (3.35). It is convenient to introduce the characteristic lengths

$$\Lambda_1 = \frac{1}{Z_1}, \qquad \Lambda_2 = \frac{1}{Z_2},$$

as well as the internuclear distance

$$R = X_2 - X_1 > 0.$$

Theorem 3.5. *The solutions of problem* (3.35) *are given by*

$$u = \frac{Z_1}{\zeta}u(X_1)\,\mathrm{S}_{\zeta,X_1} + \frac{Z_2}{\zeta}u(X_2)\,\mathrm{S}_{\zeta,X_2}, \tag{3.36a}$$

$$E = -\frac{1}{2}\zeta^2 \tag{3.36b}$$

where ζ is a zero of the equation

$$(Z_1 - \zeta)(Z_2 - \zeta) - Z_1Z_2\exp(-2R\zeta) = 0, \tag{3.37a}$$

and where $(u(X_1), u(X_2))^T$ is a suitably normalized non-trivial vector satisfying

$$\begin{pmatrix} Z_1 & Z_2\exp(-R\zeta) \\ Z_1\exp(-R\zeta) & Z_2 \end{pmatrix}\begin{pmatrix} u(X_1) \\ u(X_2) \end{pmatrix} = \zeta\begin{pmatrix} u(X_1) \\ u(X_2) \end{pmatrix}. \tag{3.37b}$$

These solutions consist of

1. *a fundamental state (u_*, E_*), that corresponds to the unique zero ζ_* of* (3.37a) *such that*

$$\max\{Z_1, Z_2\} < \zeta_* < Z_1 + Z_2. \tag{3.38}$$

2. *an excited state, that exists if and only if*

$$R > \frac{\Lambda_1 + \Lambda_2}{2} \tag{3.39}$$

and that corresponds to the unique zero $\zeta_\sharp$ of (3.37a) *such that*

$$0 < \zeta_\sharp < \min\{Z_1, Z_2\}. \tag{3.40}$$

PROOF. The general form (3.23a) reduces in our case to (3.36a). The relations of compatiblity (3.25) for $(u(X_1), u(X_2))$ become (3.37b). Let

$$f(\zeta) := \det(\mathbf{C}^\zeta - \zeta \mathbf{I}) = (Z_1 - \zeta)(Z_2 - \zeta) - Z_1 Z_2 \exp(-2R\zeta) \tag{3.41}$$

be the left-hand side of the characteristic equation.

The intuition for the number of roots of f is depicted in Figure 3.1, where we have plotted: the parabola $\mathfrak{p}$ (in blue) representing $\zeta \mapsto (Z_1 - \zeta)(Z_2 - \zeta)$ and the exponential curve $\mathfrak{e}$ (in red) representing $\zeta \mapsto Z_1 Z_2 \exp(-2R\zeta)$. $\mathfrak{p}$ is convex and cuts the axis $y = 0$ at $x = \min\{Z_1, Z_2\}$ and $x = \max\{Z_1, Z_2\}$. $\mathfrak{e}$ is decreasing and lies above the axis $y = 0$. If $\mathfrak{e}$ starts with a small slope (in absolute value) at $\zeta = 0$, it stays above $\mathfrak{p}$ for a while and will intersect $\mathfrak{p}$ at only one point farther that $\max\{Z_1, Z_2\}$. If $\mathfrak{e}$ starts with a slope large enough (in absolute value) at $\zeta = 0$, it dives below $\mathfrak{p}$ from the beginning and will intersect $\mathfrak{p}$ at two points.

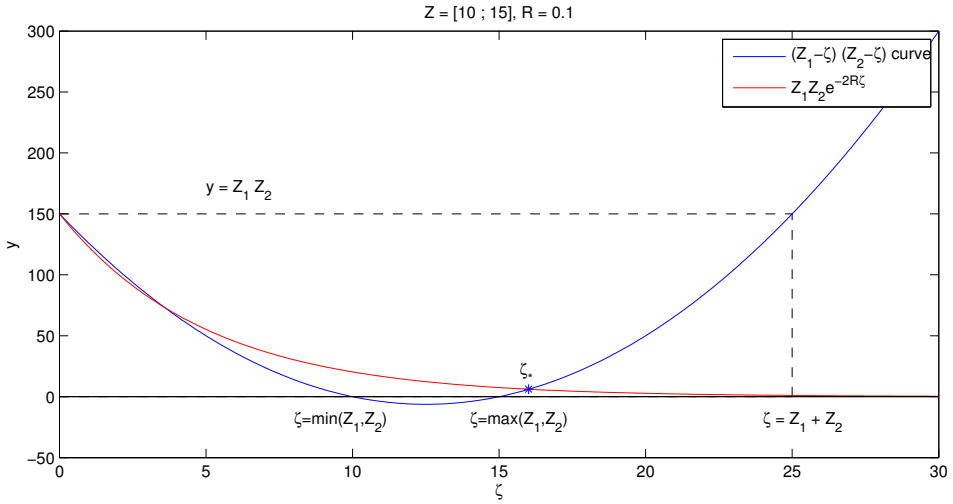

Figure 3.1: Intersection of the parabola $\mathfrak{p} : \zeta \mapsto (Z_1 - \zeta)(Z_2 - \zeta)$ and the exponential $\mathfrak{e} : \zeta \mapsto Z_1 Z_2 \exp(-2R\zeta)$.

To have a rigorous proof, let us introduce

$$Z = \max\{Z_1, Z_2\}, \qquad \mathcal{Z} = Z_1 + Z_2, \tag{3.42}$$

and study f defined in (3.41).

— *Existence and uniqueness of* ζ_*. It is plain that

$$f(Z) = 0 - Z_1 Z_2 \exp(-2RZ) < 0,$$
$$f(\mathcal{Z}) = Z_1 Z_2 (1 - \exp(-2R\mathcal{Z})) > 0.$$

By virtue of the intermediate value theorem, there exists $\zeta_* \in (Z, \mathcal{Z})$ such that $f(\zeta_*) = 0$. To show that such a ζ_* is unique in $(Z, +\infty)$, let us study the derivative of f, which is

$$f'(\zeta) = 2\left(\zeta - \frac{Z_1 + Z_2}{2}\right) + 2RZ_1 Z_2 \exp(-2R\zeta).$$

Clearly $f'(\zeta) > 0$ for all $\zeta \geq Z$ since $Z \geq (Z_1 + Z_2)/2$. Thus, f is strictly increasing on $[Z, +\infty)$ and there is a unique zero ζ_* of f not only in $(Z, \mathcal{Z})$ but also in $(Z, +\infty)$. Because $\zeta_* > \max\{Z_1, Z_2\}$, the matrix

$$\mathbf{C}^{\zeta_*} - \zeta_* \mathbf{I} = \begin{pmatrix} Z_1 - \zeta_* & Z_2 \exp(-R\zeta) \\ Z_1 \exp(-R\zeta) & Z_2 - \zeta_* \end{pmatrix}$$

is not identically zero, which implies that there is just one "eigenvector" $\mathbf{u}_*$ such that $\mathbf{C}^{\zeta_*}\mathbf{u}_* = \zeta_* \mathbf{u}_*$, up to a normalization constant.

— *Non-existence of $\zeta_\sharp$ when $R \leq (\Lambda_1 + \Lambda_2)/2$.* At $\zeta = 0$, we have

$$f(0) = 0, \qquad f'(0) = -(Z_1 + Z_2) + 2RZ_1Z_2, \qquad f''(0) = 2 - 4Z_1Z_2R^2.$$

If $f'(0) \leq 0$, that is, if $R \leq (Z_1 + Z_2)/2Z_1Z_2 = (\Lambda_1 + \Lambda_2)/2$, then there exists $\zeta_0 > 0$ in the vicinity of 0 such that $f(\zeta) < 0$ for all $\zeta \in (0, \zeta_0]$. This assertion is obvious when $f'(0) < 0$, but when $f'(0) = 0$ it is due to

$$f''(0) = 2 - 4Z_1Z_2\left(\frac{Z_1 + Z_2}{2Z_1Z_2}\right)^2 < 0.$$

In this case, we will prove that f does not have a zero in (ζ_0, Z), which leads to the unicity of ζ_*. By contradiction, should there exist a zero for f in (ζ_0, Z), there must be at least 2 zeros ζ_1 and ζ_2 in (ζ_0, Z), since we have $f(\zeta_0) < 0$ and $f(Z) < 0$. Now, we have globally 4 distinct zeros for f: 0, ζ_1, ζ_2 and ζ_*. By successive applications of Rolle's theorem, there exist 3 distinct zeros for f', 2 distinct zeros for f'' and 1 zero for f'''. But

$$f'''(x) = 8R^3Z_1Z_2\exp(-2R\zeta) > 0$$

cannot vanish. In summary, when $R \leq (\Lambda_1 + \Lambda_2)/2$, there is no zero of f other than ζ_*.

— *Existence and uniqueness of $\zeta_\sharp$ when $R > (\Lambda_1 + \Lambda_2)/2$.* If $f'(0) > 0$, that is, if $R > (Z_1 + Z_2)/2Z_1Z_2 = (\Lambda_1 + \Lambda_2)/2$, then there is a $\zeta_0 > 0$ in the vicinity of 0 such that $f(\zeta) > 0$ for all $\zeta \in (0, \zeta_0]$. Since $f(\zeta_0) > 0$ and $f(Z) < 0$, there exists a zero $\zeta_\sharp \in (\zeta_0, Z)$ for f. Should there appear two or more zeros in (ζ_0, Z), let us call these ζ_1 and ζ_2 and repeat the argument above using Rolle's theorem three times to arrive at a contradiction. In summary, when $R > (\Lambda_1 + \Lambda_2)/2$, there is exactly one second zero $\zeta_\sharp$ of f. Because $f < 0$ on $[\min\{Z_1, Z_2\}, Z]$, we must have $\zeta_\sharp < \min\{Z_1, Z_2\}$ and the matrix

$$\mathbf{C}^{\zeta_\sharp} - \zeta_\sharp \mathbf{I} = \begin{pmatrix} Z_1 - \zeta_\sharp & Z_2 \exp(-R\zeta) \\ Z_1 \exp(-R\zeta) & Z_2 - \zeta_\sharp \end{pmatrix}$$

cannot be identically zero, which implies that there is just one "eigenvector" $\mathbf{u}_\sharp$ such that $\mathbf{C}^{\zeta_\sharp}\mathbf{u}_\sharp = \zeta_\sharp \mathbf{u}_\sharp$, up to a normalization constant. □

Remark 3.1. The fact that the solution for a double-delta potential is the superposition of two Slater functions is very specific to the 1-D setting. In 3-D, the ground state of a diatomic monoelectronic system (such as H_2^+) is not a sum of two Slater functions, but is a prolate-spheroïdal orbital, for the calculation of which numerical algorithms are available [6] but explicit formulae are not.

For unequal charges $Z_1 \neq Z_2$, the exact solutions can be given a closed-form expression by means of a generalized Lambert function [122]. In practice, it is more efficient to numerically solve equation (3.37a) by the Newton method. For identical charges $Z_1 = Z_2 = Z$ (a one-dimensional model of homonuclear diatomic molecules), explicit formulae for the solutions are available [121] using the standard Lambert function W. This function is defined by the implicit equation

$$\mathrm{W}(z)\exp(\mathrm{W}(z)) = z \tag{3.43}$$

for $z \in [-1/e, +\infty)$ and $\mathrm{W}(z) \in [-1, +\infty)$. Put another way, W is the reciprocal function of $w \mapsto w\exp(w)$, whose domain is $[-1, +\infty)$ and whose range is $[-1/e, +\infty)$.

Corollary 3.3. *When $Z_1 = Z_2 = Z$, the fundamental state of* (3.35) *is given by*

$$u_* = \Gamma_*(\mathrm{S}_{\zeta_*,X_1} + \mathrm{S}_{\zeta_*,X_2}), \tag{3.44a}$$

$$E_* = -\frac{1}{2}\zeta_*^2, \tag{3.44b}$$

with

$$\zeta_* = Z + \frac{\mathrm{W}(RZ\exp(-RZ))}{R}, \qquad \Gamma_*^2 = \frac{Z}{2\,(1 + \mathrm{W}(RZ\exp(-RZ)))}. \tag{3.45}$$

It is called gerade *for its symmetry with respect to the mid-point $(X_1 + X_2)/2$.*

For $R > \Lambda = 1/Z$, the excited state is given by

$$u_\sharp = \Gamma_\sharp(\mathrm{S}_{\zeta_\sharp,X_1} - \mathrm{S}_{\zeta_\sharp,X_2}), \tag{3.46a}$$

$$E_\sharp = -\frac{1}{2}\zeta_\sharp^2, \tag{3.46b}$$

with

$$\zeta_\sharp = Z + \frac{\mathrm{W}(-RZ\exp(-RZ))}{R}, \qquad \Gamma_\sharp^2 = \frac{Z}{2\,(1 + \mathrm{W}(-RZ\exp(-RZ)))}. \tag{3.47}$$

It is called ungerade *for its anti-symmetry with respect to the mid-point $(X_1 + X_2)/2$.*

Proof. When $Z_1 = Z_2 = Z$, the characteristic equation (3.37a) becomes

$$(\zeta - Z)^2 = Z^2\exp(-2R\zeta),$$

and it is possible to extract the square roots as

$$\begin{aligned}\zeta - Z &= \pm Z\exp(-R\zeta)\\ &= \pm Z\exp(-RZ)\exp(-R(\zeta - Z)).\end{aligned}$$

Multiplying by R and setting $y = R(\zeta - Z)$, we get $y = \pm RZ\exp(-RZ)\exp(-y)$, hence $y\exp(y) = \pm RZ\exp(-RZ)$. The first branch corresponds to

$$y\exp(y) = RZ\exp(-RZ). \tag{3.48}$$

The right-hand side of (3.48) belongs to $[0, +\infty)$ and its image by the Lambert function is well defined. From the definition (3.43), we infer that

$$y\exp(y) = z \quad \Leftrightarrow \quad y = \mathrm{W}(z). \tag{3.49}$$

With $RZ\exp(-RZ)$ in place of z, we have

$$R(\zeta - Z) = y = \mathrm{W}(RZ\exp(-RZ))$$

and ζ is given by the gerade formula (3.45). We cannot yet conclude that $\zeta = \zeta_*$ before examining the second branch

$$y\exp(y) = -RZ\exp(-RZ). \tag{3.50}$$

We can apply the Lambert function to (3.50) since its right-hand side is always greater or equal to $-1/e$.

- If $RZ \leq 1$, then
$$y = \mathrm{W}(-RZ\exp(-RZ)) = -RZ \geq -1,$$
which implies $\zeta = 0$. Such a value is not acceptable and the second branch does not produce any new solution.

- If $RZ > 1$, then
$$y = \mathrm{W}(-RZ\exp(-RZ)) \geq -1 > -RZ,$$
which implies $\zeta > 0$. This value for ζ, corresponding to the *ungerade* formula (3.47), is acceptable and can be checked to be less than that of the first branch.

In summary, the first branch always matches with the fundamental solution ζ_*, while the second branch matches with the excited solution $\zeta_\sharp$ when $RZ > 1$. Normalizing u_* and $u_\sharp$ to $\|u\|_{L^2(\mathbb{R})} = 1$ gives us the values of Γ_* and $\Gamma_\sharp$ in (3.45), (3.47). □

Figure 3.2 displays the gerade solution for a double-delta model with $X_1 = -R/2$ and $X_2 = R/2$, as well as the two individual Slater functions that contribute to this solution.

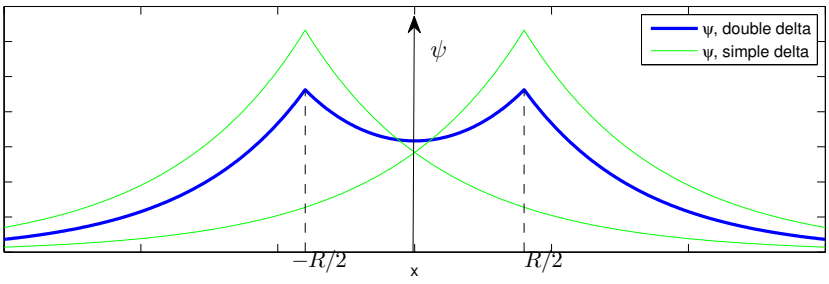

Figure 3.2: Fundamental state of H_2^+ ion in 1-D.

Returning to the general double-delta model (3.35), let us concentrate on a remarkable feature that it exhibits, which is the "sensitivity" of the solution with respect to any slight difference in the charges. Before explaining what we mean by "sensitivity," let us first state the following linear perturbation result.

Proposition 3.3. *In model* (3.35)*, consider the almost identical charges*

$$Z_1 = Z + \Delta Z, \qquad Z_2 = Z,$$

with ΔZ a small variation, i.e., $|\Delta Z| \ll Z$. Then, the values of the fundamental state u at the two cusps are in the ratio

$$\frac{u(X_1)}{u(X_2)} = 1 + [\exp(R\zeta_*) - 1]\frac{\Delta Z}{2Z} + O(\Delta Z^2), \tag{3.51}$$

where ζ_* is the gerade *solution* (3.45) *of the equal charges problem.*

Proof. Let $\zeta_* + \Delta\zeta_*$ be the gerade solution of the double-delta problem with almost identical charges. Then, $\zeta_* + \Delta\zeta_*$ is a root of

$$(\zeta_* + \Delta\zeta_* - Z - \Delta Z)(\zeta_* + \Delta\zeta_* - Z) - Z(Z + \Delta Z)\exp(-2R(\zeta_* + \Delta\zeta_*)) = 0. \tag{3.52}$$

Carrying out the first-order Taylor expansion and dropping the zeroth-order terms (which cancel out each other), taking into account the property $\zeta_* - Z = Z\exp(-R\zeta_*)$ for the gerade solution (cf. proof of Corollary 3.3), we end up with

$$\Delta\zeta_* = \frac{1 + \exp(-R\zeta_*)}{2(1 + RZ\exp(-R\zeta_*))}\Delta Z + O(\Delta Z^2) \tag{3.53}$$

after some algebra. In view of the first line of (3.37b), the ratio of amplitudes $u(X_1)/u(X_2)$ is equal to

$$\begin{aligned}\frac{u(X_1)}{u(X_2)} &= \frac{Z\exp(-R(\zeta_* + \Delta\zeta_*))}{\zeta_* + \Delta\zeta_* - Z - \Delta Z} \\ &= \frac{Z\exp(-R\zeta_*)}{\zeta_* - Z}\frac{1 - R\Delta\zeta_*}{1 + (\zeta_* - Z)^{-1}(\Delta\zeta_* - \Delta Z)} + O(\Delta Z^2).\end{aligned}$$

Again, because $\zeta_* - Z = Z\exp(-R\zeta_*)$, the above fraction is reduced to

$$\begin{aligned}\frac{u(X_1)}{u(X_2)} &= \frac{1 - R\Delta\zeta_*}{1 + Z^{-1}\exp(R\zeta_*)(\Delta\zeta_* - \Delta Z)} + O(\Delta Z^2) \\ &= (1 - R\Delta\zeta_*)(1 - Z^{-1}\exp(R\zeta_*)(\Delta\zeta_* - \Delta Z)) + O(\Delta Z^2) \\ &= 1 - (R + Z^{-1}\exp(R\zeta_*))\Delta\zeta_* + Z^{-1}\exp(R\zeta_*)\Delta Z + O(\Delta Z^2).\end{aligned}$$

Inserting the value (3.53) for $\Delta\zeta_*$ into the previous equation, it becomes

$$\begin{aligned}\frac{u(X_1)}{u(X_2)} &= 1 - \frac{\Delta Z}{Z}(RZ + \exp(R\zeta_*))\frac{1 + \exp(-R\zeta_*)}{2(1 + RZ\exp(-R\zeta_*))} \\ &\quad + \frac{\Delta Z}{Z}\exp(R\zeta_*) + O(\Delta Z^2) \\ &= 1 + \frac{\Delta Z}{Z}\frac{(\exp(R\zeta_*) - 1)(1 + RZ\exp(-R\zeta_*))}{2(1 + RZ\exp(-R\zeta_*))} + O(\Delta Z^2).\end{aligned}$$

After simplification, we finally obtain the ratio (3.51). □

By "sensitivity," we mean that the amplification factor $\frac{1}{2}[\exp(R\zeta_*) - 1]$ in the ratio $u(X_1)/u(X_2)$ can be extremely large, even for "reasonable" values of $R\zeta_*$. For instance, if we take

$$Z = 20, \quad R = \frac{1}{2},$$

as in Figure 3.3, then $RZ = 10$ and by (3.45),

$$R\zeta_* = 10 + \mathrm{W}(10\exp(-10)) \approx 10.$$

This entails

$$\frac{u(X_1)}{u(X_2)} \approx 1 + \frac{\exp(10) - 1}{2}\frac{\Delta Z}{Z} \approx 1 + 11012\frac{\Delta Z}{Z}.$$

In other words, when the distance R between the nuclei is large enough compared to $\Lambda_* = 1/\zeta_*$, the characteristic length associated with the ground state solution, a slight perturbation in one of the charge is likely to cause a dramatic distortion in the shape of the solution. As displayed in Figure 3.3, the cusp with the smaller charge becomes very quickly insignificant in comparison with the other one, and may even seem to have vanished. From the viewpoint of physics, this sensitivity phenomenon bears some similarity with the *ionic bonding*, where the electron is transferred from one ion to another by electrovalence. Here, as soon as there is a difference in the charges of the nuclei and provided that these nuclei are far enough, the electron "decides" to choose the stronger nucleus to be associated with.

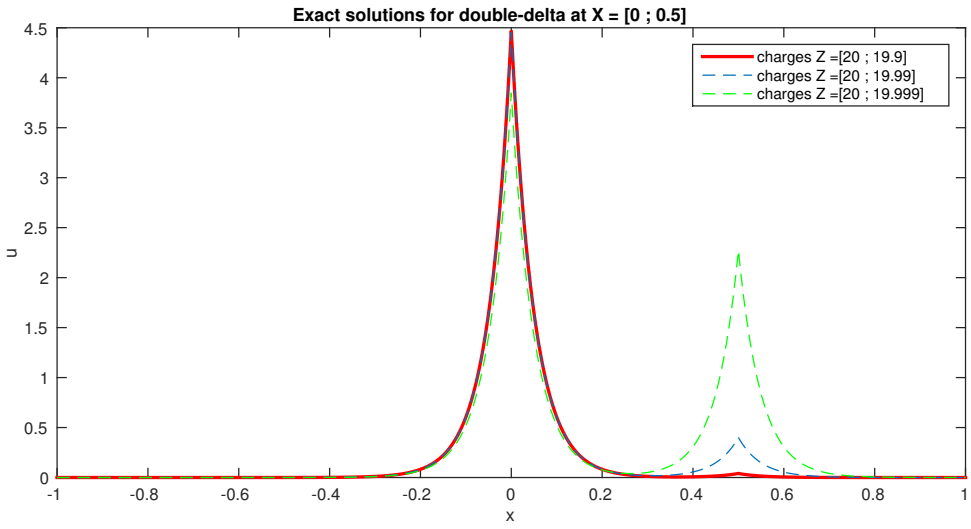

Figure 3.3: Wave function in the double-delta model.

3.1.5 Uniqueness and other properties of the ground state

After having explored the properties of all eigenstates, defined to be solutions of (3.1), we return to the energy viewpoint (3.4).

Theorem 3.6. *The minimizer u_* of* (3.4) *is unique, up to a sign.*

Proof. Following once again the strategy developed in [22, §3], we divide the proof into several steps.

1. If u_* is a minimizer, then $v := |u_*| \geq 0$ is also a minimizer, as $|u_*| \in \mathscr{S}$ and $\mathfrak{E}(|u_*|) = \mathfrak{E}(u_*)$. As such, v is a weak solution of 3.1.

Suppose that v vanishes at some point $x_0 \in \mathbb{R}$. If x_0 is not one of the nuclei positions X_I, we know by Theorem 3.2 that v is (infinitely) differentiable in the neighborhood of x_0. In other words, $v'(x_0)$ exists. Necessarily, $v'(x_0) = 0$ since otherwise v would take strictly negative values in some half-neighborhood of x_0. Starting from x_0 with $v(x_0) = v'(x_0) = 0$, we solve the differential equation $v'' = -2E_* v$ to find that $v = v' \equiv 0$ until we reach a nucleus location X_I. By continuity, $v(X_I) = 0$ and $v'(X_I^-) = 0$ or $v'(X_I^+) = 0$. The Kato condition (3.19) allows us to "jump" to the other side of X_I, where the first derivative also vanishes. We can then go on integrating the differential equation until the next nucleus. At the end of the process, we will find $v \equiv 0$ on $\mathbb{R}$, which contradicts $\|v\|_{L^2} = 1$.

If $x_0 = X_I$ for some $I \in \{1, \dots, M\}$, the Kato condition (3.19) yields $v'(X_I^-) = v'(X_I^+)$. This means that v is differentiable at X_I. The common value $v'(X_I)$ must then be zero, otherwise v would take negative value. The rest of the argument is identical to the above. In conclusion, if u_* is a minimizer, then u_* cannot vanish anywhere. In other words, u_* must keep a constant sign on $\mathbb{R}$.

2. If u_* minimizes $\mathfrak{E}$ over the unit sphere $\mathscr{S}$, then the probability density $\rho_* = |u_*|^2$ remains strictly positive (according to the previous step) and minimizes the functional

$$\rho \mapsto \mathfrak{E}(\sqrt{\rho}) = \frac{1}{2}\int_{\mathbb{R}} |(\sqrt{\rho})'|^2 - \sum_{I=1}^{M} Z_I \rho(X_I) \tag{3.54}$$

over

$$\mathscr{C} = \left\{ \rho > 0, \quad \sqrt{\rho} \in H^1(\mathbb{R}), \quad \int_{\mathbb{R}} \rho = 1 \right\}.$$

The functional (3.54) can be shown to be strictly convex on $\mathscr{C}$. Besides, $\mathscr{C}$ is a convex set. Therefore, the minimizer ρ_* is unique. We are thus left with only two choices, namely, $u_* = \sqrt{\rho_*} > 0$ or $u_* = -\sqrt{\rho_*} < 0$. This completes the proof of uniqueness. □

Let $E^{(1)}$ be another name for E_*, since it represents the "first eigenvalue." We define the "second eigenvalue" as

$$E^{(2)} = \inf_{\substack{v \in (u_*)^\perp \subset \mathscr{V} \\ \|v\|_{L^2}=1}} \mathfrak{E}(v), \tag{3.55}$$

the infimum problem being well-defined on the orthogonal complement $(u_*)^\perp$ in $\mathscr{V}$. Naturally, $E^{(2)} \geq E^{(1)}$. But $E^{(2)}$ might not be attained and nothing could prevent it from being equal to $E^{(1)}$. The uniqueness result forbids this scenario to take place.

Corollary 3.4. *$E^{(2)} > E^{(1)}$.*

Proof. Assume $E^{(2)} = E_*$. By Lemma 3.3, $E^{(2)} < 0$. Then, the whole proof machinery of Theorem 3.1 can be repeated, in which $\mathscr{V}$ is replaced by $(u_*)^\perp$, so as to establish the existence of a minimizer $u^{(2)} \in (u_*)^\perp$ such that $\|u^{(2)}\|_{L^2} = 1$ and $E^{(2)} = \mathfrak{E}(u^{(2)})$. Indeed, the elements of the minimizing (sub)sequence $\{u_n\}_{n\geq 1}$ belong to $(u_*)^\perp$. Regarding the weak limit $u^{(2)}$ in $H^1(\mathbb{R})$, it also belongs to $(u_*)^\perp$, since

$$H^1\text{-boundedness} \Longrightarrow L^2\text{-boundedness} \Longrightarrow L^2\text{-weak convergence}$$

and the latter preserves L^2-orthogonality. The value E_* would then be reached at two minimizers u_* and $u^{(2)}$, which are linearly independent because of orthogonality. This obviously contradicts Theorem 3.6. □

The fact that $E^{(2)} > E^{(1)}$ ensures that the bilinear form $\mathfrak{a}(\cdot,\cdot) - E_*\mathfrak{b}(\cdot,\cdot)$ is L^2-coercive on $(u_*)^\perp$, which turns out to be a cornerstone requirement for error estimates. To finish with multi-delta models on infinite domains, let us give a sharpened bounds for the fundamental energy level. This can be useful for the numerical computation of E_*.

Theorem 3.7. *The fundamental energy E_* is bounded by*

$$-\frac{1}{2}\mathcal{Z}^2 \leq E_* \leq -\frac{1}{2}Z^2, \tag{3.56}$$

where $\mathcal{Z} = \sum_{J=1}^M Z_J$ is the total charge and $Z = \max\{Z_1, \dots, Z_M\}$ is the greatest charge.

Proof. For a fixed $I \in \{1, \dots, M\}$, introduce the Slater function $v_I = Z_I^{1/2}\,\mathrm{S}_{Z_I,X_I}$. From Theorem 3.4, we know that $\|v_I\|_{L^2} = 1$ and that from the reformulation (3.34), we have

$$\frac{1}{2}\|v_I'\|_{L^2}^2 - Z_I|v_I(X_I)|^2 = -\frac{1}{2}Z_I^2.$$

Going back to the multi-delta potential problem at hand, we have

$$E_* = \min_{\substack{v\in H^1(\mathbb{R})\\ \|v\|_{L^2}=1}} \mathfrak{E}(v) \leq \mathfrak{E}(v_I).$$

But

$$\mathfrak{E}(v_I) = \frac{1}{2}\|v_I'\|_{L^2}^2 - Z_I|v_I(X_I)|^2 - \sum_{J\neq I} Z_J|v_I(X_J)|^2 = -\frac{1}{2}Z_I^2 - \sum_{J\neq I} Z_J|v_I(X_J)|^2$$

and thus $E_* \leq -Z_I^2/2$. The minimum of the last quantity over $I \in \{1, \dots, M\}$ yields the upper bound $E_* \leq -Z^2/2$, with $Z = \max\{Z_1, \dots, Z_M\}$.

To derive the lower bound, let $\mathcal{Z} = \sum_{J=1}^M Z_J$ stand for the total charge and X some abscissa to be specified later. Again, by (3.34), we know that

$$-\frac{1}{2}\mathcal{Z}^2 = \min_{\substack{v\in H^1(\mathbb{R})\\ \|v\|_{L^2}=1}} \frac{1}{2}\|v'\|_{L^2}^2 - \mathcal{Z}|v(X)|^2 \leq \frac{1}{2}\|u_*'\|_{L^2}^2 - \mathcal{Z}|u_*(X)|^2.$$

But

$$E_* = \mathfrak{E}(u_*) = \frac{1}{2}\|u_*'\|_{L^2}^2 - \sum_{J=1}^M Z_J|u_*(X_J)|^2,$$

so that

$$-\frac{1}{2}\mathcal{Z}^2 \leq E_* + \sum_{J=1}^M Z_J|u_*(X_J)|^2 - \mathcal{Z}|u_*(X)|^2 = E_* + \sum_{J=1}^M Z_J(|u_*(X_J)|^2 - |u_*(X)|^2).$$

By choosing $X = X_I$ such that $|u(X_I)| = \max_{1\leq J\leq M}|u(X_J)|$, we can make sure that every summand of the second term is non-positive. As a result, $-\mathcal{Z}^2/2 \leq E_*$. □

The bounds in (3.56) correspond to two physically meaningful extreme cases. The lower bound $-\mathcal{Z}^2/2$ is the fundamental energy of an fictitious system whose M nuclei, having the same charges, are concentrated at the same point. The upper bound $-Z^2/2$ is the fundamental energy of an fictitious system whose M nuclei, having the same charges, are scattered to infinity and act as if each were alone.

3.2 Multi-delta model in a periodic domain

The infinite model of §3.1 can be approximated by a Galerkin method on a meshless basis, e.g., consisting of infinite-supported Gaussians as in §4.2. It is unfortunately not suitable to numerical approximation on any basis associated with a mesh. To overcome this difficulty, we opted for the periodic model that is obtained from the infinite one by imposing periodic boundary conditions. The periodic model is shown to preserve almost all of the desirable properties of the infinite model.

3.2.1 Physical ideas

We consider a system evolving in an interval $[0, L] \subset \mathbb{R}$ with $L > 0$, consisting of one electron and $M \geq 1$ nuclei of known charges $(Z_1, Z_2, \ldots, Z_M) \in (\mathbb{R}_+^*)^M$ located at known positions $(X_1, X_2, \ldots, X_M) \in (0, L)^M$ such that

$$0 < X_1 < X_2 < \ldots < X_M < L.$$

The state of this electron is described by a wave function $u : x \in [0, L] \mapsto u(x) \in \mathbb{R}$ subject to the L-periodicity condition

$$u(0) = u(L). \tag{3.57}$$

Only those wave functions satisfying

$$-\frac{1}{2}u'' + \Big(- \sum_{J=1}^{M} Z_J \delta_{X_J} \Big) u = Eu, \tag{3.58a}$$

$$\int_0^L |u|^2 = 1, \tag{3.58b}$$

in some variational sense to be precised later, are relevant to characterize the state of the system. In (3.58), the unknowns are the wave function u and the energy E. The meanings of these two equations and the various terms contained in them are the same as in (3.1). As in §3.1, our primary interest lies in the ground state solution (u_*, E_*) that corresponds to the lowest possible energy level E, but we shall also be concerned with some limited aspects of excited states.

The conditions $0 < X_1$ and $X_M < L$ have been imposed for convenience. A delta potential located at the boundary is not a difficulty in itself, but makes formulae a little awkward to write down. Regarding the periodicity of the wave function, it is sufficient to prescribe equality (3.57) between its two end values. As will be shown in Theorem 3.9, it turns out that any solution of (3.57)–(3.58) automatically satisfies the additional periodicity $u'(0^+) = u'(L^-)$ between its derivatives at the boundary.

Approximating the infinite domain

Another reason for prohibiting a nucleus at the boundary is that, in fact, the purpose of (3.57)–(3.58) is to approximate the infinite domain problem (3.1) on a bounded computational domain. It is expected that when the domain is large enough ($L \to +\infty$), the ground state solution of (3.57)–(3.58) degenerates in some sense to that of (3.1). The notion of "large enough" can be further quantified by the requirement

$$L \gg \max_{1 \leq I \leq M} \Lambda_I, \qquad \text{where } \Lambda_I = \frac{1}{Z_I}$$

is the characteristic length associated with the I-th nucleus.

At this point, the question arises as to whether or not transparent boundary conditions would have been preferable. After all, these can be worked out for our problem, since the overall potential vanishes outside $[X_1, X_M]$. In a previous work [115], we considered the boundary conditions

$$u'(0) = \sqrt{-2E}\, u(0), \qquad u'(L) = -\sqrt{-2E}\, u(L), \tag{3.59}$$

that are exactly satisfied by the ground state of the infinite domain model (3.1). While conditions (3.59) allow for a perfect imitation of the infinite domain solution in a bounded computational domain, they suffer from two drawbacks:

1. The variational formulation that takes (3.59) into account reads

$$\frac{1}{2}\int_0^L u'v' + \sqrt{-2E}\,\frac{u(0)v(0)+u(L)v(L)}{2} - \sum_{I=1}^{M} Z_I u(X_I)v(X_I) = E\int_0^L uv,$$
$$\int_0^L |u|^2 + \frac{1}{\sqrt{-2E}}\,\frac{|u(0)|^2+|u(L)|^2}{2} = 1.$$

   It is nonlinear with respect to the eigenvalue E and requires an iterative numerical procedure, which may incur extra cost and convergence issue.

2. Wavelets whose supports intersect the boundary cannot be easily "truncated" for the calculation of the integrals $\int_0^L$. A framework was proposed by Monasse and Perrier [106] for adapting wavelets to an interval with homogeneous boundary conditions, but to our knowledge there is no such framework for (3.59). Anyhow, the implementation would be much more technical. In [115], we stayed away from this problem by working with P_1 finite elements. But here, wavelets are part of our objectives.

In this context, periodic boundary conditions appear to be a good compromise. For one, their implementation is pretty easy using the periodized scaling functions and wavelets introduced in §2. For another, the solutions of the periodic model (3.57)–(3.58) enjoy many properties similar to those of the infinite model (3.1). In particular, we still have a cusp at each X_I, where the Kato conditions remain satisfied (Theorem 3.9). The superposition principle for the general solution is preserved (Theorem 3.10), modulo the fact that the elementary component has now changed from the Slater function $\mathrm{S}_{\zeta,X}$ to its L-periodized version $\widetilde{\mathrm{S}}_{\zeta,X}$ defined in (3.79). Generally speaking, every formula for the infinite domain has a counterpart in the periodic domain, which looks more heavy and more difficult to apprehend but which tends to the correct limit as $L \to +\infty$.

There is one feature of the periodic model that we should be aware of and that distinguishes it from the infinite model. In Theorem 3.3, we saw that the infinite model has at most a finite number of excited states, all of which have a negative energy level. For the periodic model, it can be shown that there exists countable sequence of excited states, whose energy levels eventually become positive. This is due to the boundedness of the domain. We shall not give the proof of this. Neither shall we investigate excited states with positive energies, since their complexity goes well beyond the scope of this thesis.

Energy viewpoint

As pointed out in §3.1.1, the delta potential is an obstruction to defining the Hamiltonian operator in the customary way. The correct approach to envisage "eigenvalues" is the variational one. From now on, the fundamental energy E_* is declared to be

$$E_* = \inf_{u\in\mathscr{V}} \frac{\mathfrak{a}(u,u)}{\mathfrak{b}(u,u)} \tag{3.60}$$

with the space

$$\mathscr{V} = H^1_{\#}(0,L) = \{u \in L^2(0,L) \,|\, u' \in L^2(0,L),\ u(0) = u(L)\}$$

and the bilinear forms

$$\mathfrak{a}(u,v) = \frac{1}{2}\int_0^L u'v' - \sum_{I=1}^M Z_I u(X_I)v(X_I) \tag{3.61a}$$

$$\mathfrak{b}(u,v) = \int_0^L uv \tag{3.61b}$$

for $(u,v) \in \mathscr{V}^2$. Since $\mathfrak{a}$ and $\mathfrak{b}$ are homogeneous of degree 2, a straightforward reformulation of (3.60) is

$$E_* = \inf_{\substack{u\in\mathscr{V}\\ \mathfrak{b}(u,u)=1}} \mathfrak{E}(u), \tag{3.62}$$

where

$$\mathfrak{E}(u) = \frac{1}{2}\int_0^L |u'|^2 - \sum_{I=1}^M Z_I|u(X_I)|^2 = \mathfrak{a}(u,u) \tag{3.63}$$

is the energy functional. Thanks to the embedding $H^1_{\#}(0,L) \subset H^1(0,L) \subset C^0([0,L])$, the bilinear forms $\mathfrak{a}$, $\mathfrak{b}$ and the functional $\mathfrak{E}$ are well-defined. The questions that remain to be elucidated are the same as in the infinite model, namely:

1. Existence of the fundamental energy E_*, which amounts to saying that the infimum problem (3.62) is well-defined.
2. Existence of a minimizer $u_* \in \mathscr{V}$ such $E_* = \mathfrak{E}(u_*)$, from now on referred to as a ground state.
3. Uniqueness of the ground state u_*.

Questions 1 and 2 will be addressed in §3.2.2. In §3.2.3, we derive the properties shared by all solutions of (3.58), with an emphasis laid on single-delta and double-delta potentials. These properties will, in turn, help us tackle question 3 in §3.2.5.

Variational formulation

Broadening the scope of (3.62), we consider the problem of seeking all critical points of the energy functional $\mathfrak{E}$ on the manifold

$$\mathfrak{I}(u) := \|u\|^2_{L^2(0,L)} - 1 = 0.$$

The following Proposition characterizes such a critical point —should it exist at all— as a solution in the variational sense of the 1-D model (3.57)–(3.58).

Proposition 3.4. *Every critical point $u \in \mathscr{V}$ of the energy functional $\mathfrak{E}$, subject to the constraint $\mathfrak{I}(u) = 0$, is necessarily a solution of (3.57)–(3.58) in the variational sense:*

$$\mathfrak{a}(u, v) = E\, \mathfrak{b}(u, v), \tag{3.64a}$$
$$\mathfrak{b}(u, u) = 1, \tag{3.64b}$$

for all $v \in \mathscr{V}$. Furthermore,

$$E = \mathfrak{E}(u).$$

Proof. Similar to the proof of Proposition 3.1. □

The variational formulation (3.64) is the correct sense in which problem (3.57)–(3.58) must be understood. It is also the starting point for building up a Galerkin approximation.

3.2.2 Existence of a ground state

The space $\mathscr{V} = H^1_{\#}(0, L)$ is equipped with the norm

$$\|u\|^2_{H^1(0,L)} = \|u\|^2_{L^2(0,L)} + \|u'\|^2_{L^2(0,L)}.$$

We said earlier that $\mathscr{V} \subset H^1(0, L) \subset C^0([0, L])$. This gives us legitimacy to define the norm

$$\|v\|_{L^\infty(0,L)} = \sup_{x\in[0,L]} |v(x)|$$

for all $v \in \mathscr{V}$. Throughout the remainder of section §3.2, we shall be writing

$$\|\cdot\|_{L^\infty},\ \|\cdot\|_{L^2},\ \|\cdot\|_{H^1} \quad \text{instead of} \quad \|\cdot\|_{L^\infty(0,L)},\ \|\cdot\|_{L^2(0,L)},\ \|\cdot\|_{H^1(0,L)}.$$

Thanks to the continuous embedding [2]

$$H^1(0, L) \subset C^{0,1/2}([0, L]),$$

where $C^{0,1/2}$ denotes the Hölder space with exponent $1/2$, there exists *a fortiori* a constant $c > 0$ such that, for all $v \in \mathscr{V}$,

$$\|v\|_{L^\infty} \le c\|v\|_{H^1}. \tag{3.65}$$

Neverthless, let us work out some finer control of the L^∞-norm for functions in $\mathscr{V}$.

Lemma 3.4. *For all $v \in \mathscr{V}$ and for all $\theta > 0$,*

$$\|v\|^2_{L^\infty} \le \theta\|v'\|^2_{L^2} + \left(\frac{1}{\theta} + \frac{1}{L}\right)\|v\|^2_{L^2}. \tag{3.66}$$

Proof. For any $v \in \mathscr{V}$ and $x, y \in [0, L]$ we have

$$|v(x)|^2 - |v(y)|^2 = \int_y^x 2v'(t)v(t)\,\mathrm{d}t \le 2\|v'\|_{L^2}\|v\|_{L^2}$$

by the Cauchy-Schwarz inequality. It follows that $|v(x)|^2 \le |v(y)|^2 + 2\|v\|_{L^2}\|v'\|_{L^2}$, and by taking the supremum in $x \in [0, L]$, we find that

$$\|v\|^2_{L^\infty} \le |v(y)|^2 + 2\|v'\|_{L^2}\|v\|_{L^2}.$$

Integrating this inequality over $[0, L]$, we obtain

$$\begin{aligned} L\|v\|_{L^\infty}^2 &\leq \|v\|_{L^2}^2 + 2L\|v\|_{L^2}\|v'\|_{L^2} \\ &\leq \|v\|_{L^2}^2 + L\left(\theta\|v'\|_{L^2}^2 + \frac{1}{\theta}\|v\|_{L^2}^2\right) \end{aligned}$$

by virtue of Young's inequality. Division by L yields (3.66). □

When $L \to +\infty$, the upper bound (3.66) degenerates to its infinite domain counterpart (3.10). Inequalities (3.65) and (3.66) have tremendous consequences on the properties of the bilinear form $\mathfrak{a}$ and the energy functional $\mathfrak{E}$. In the following Proposition, the first two statements will be used for various error estimates in §4.3.2 and §5.3, while the last one is exactly what we need for the well-posedness of the infimum problem (3.4).

Proposition 3.5. *The following properties hold true:*

1. *There exists $\kappa > 0$ such that for all $(v, w) \in \mathscr{V}^2$,*

$$|\mathfrak{a}(v, w)| \leq \kappa\|v\|_{H^1}\|w\|_{H^1}. \tag{3.67}$$

2. *There exists $\Theta > 0$ such that for all $v \in \mathscr{V}$,*

$$\mathfrak{a}(v, v) \geq \frac{1}{4}\|v\|_{H^1}^2 - \Theta\|v\|_{L^2}^2. \tag{3.68}$$

3. *Over the L^2-unit sphere*

$$\mathscr{S} = \{v \in \mathscr{V},\ \|v\|_{L^2} = 1\}, \tag{3.69}$$

the energy functional $\mathfrak{E}$ is bounded from below.

Proof. By the triangle inequality and from the definition of $\mathfrak{a}(\cdot, \cdot)$, we have

$$\begin{aligned} |\mathfrak{a}(v, w)| &\leq \frac{1}{2}\int_0^L |v'||w'| + \sum_{I=1}^M Z_I|v(X_I)||w(X_I)| \\ &\leq \left(\frac{1}{2}\|v'\|_{L^2}^2 + \sum_{I=1}^M Z_I|v(X_I)|^2\right)^{1/2}\left(\frac{1}{2}\|w'\|_{L^2}^2 + \sum_{I=1}^M Z_I|w(X_I)|^2\right)^{1/2}, \end{aligned}$$

the last line being due to the Cauchy-Schwarz inequality. From $|v(X_I)| \leq \|v\|_{L^\infty}$ and $\|v\|_{L^\infty} \leq c\|v\|_{H^1}$ after (3.65), we infer that

$$|\mathfrak{a}(v, w)| \leq \left(\frac{1}{2}\|v'\|_{L^2}^2 + \mathcal{Z}c\|v\|_{H^1}^2\right)^{1/2}\left(\frac{1}{2}\|w'\|_{L^2}^2 + \mathcal{Z}c\|w\|_{H^1}^2\right)^{1/2},$$

where we remind that $\mathcal{Z} = \sum_{I=1}^M Z_I$ is the total charge. Taking $\kappa = 1/2 + \mathcal{Z}c$, we easily get (3.67). To derive (3.68), we first notice that

$$\mathfrak{a}(v, v) \geq \frac{1}{2}\|v'\|_{L^2}^2 - Z\|v\|_{L^\infty}^2 \geq \left(\frac{1}{2} - \mathcal{Z}\theta\right)\|v'\|_{L^2}^2 - \mathcal{Z}\left(\frac{1}{\theta} + \frac{1}{L}\right)\|v\|_{L^2}^2, \tag{3.70}$$

the last inequality being due to (3.66) of Lemma 3.4. Selecting $\theta > 0$ such that $\mathcal{Z}\theta = 1/4$ and writing $\|v'\|_{L^2}^2 = \|v\|_{H^1}^2 - \|v\|_{L^2}^2$, we obtain

$$\mathfrak{a}(v, v) \geq \frac{1}{4}\|v\|_{H^1}^2 - \left(\frac{1}{4} + 4\mathcal{Z}^2 + \frac{\mathcal{Z}}{L}\right)\|v\|_{L^2}^2$$

which proves (3.68) with $\Theta = 1/4 + 4\mathcal{Z}^2 + \mathcal{Z}/L$.

In (3.70), we select $\theta > 0$ such that $\mathcal{Z}\theta = 1/2$. This cancels out the first term in the lower bound and leaves us with

$$\mathfrak{E}(v) = \mathfrak{a}(v, v) \geq -2\mathcal{Z}^2 - \frac{\mathcal{Z}}{L}, \tag{3.71}$$

for $\|v\|_{L^2}^2 = 1$ for all $v \in \mathscr{S}$. □

When $L \to +\infty$, the lower bound (3.71) degenerates to its infinite domain counterpart (3.15). The facts that the energy functional $\mathfrak{E}$ is well-defined on $\mathscr{V}$ and that it is bounded from below on the unit sphere $\mathscr{S}$ testify to the well-posedness of the infimum problem (3.62).

Theorem 3.8. *Let $\mathscr{S}$ be the L^2-unit sphere defined in* (3.69)*. There exists a minimizer $u_* \in \mathscr{S} \subset \mathscr{V}$ such that*

$$E_* = \mathfrak{E}(u_*) = \min_{u \in \mathscr{S}} \mathfrak{E}(u).$$

Proof. In accordance with the strategy developed in [22, §2], we divide the proof into several steps.

1. Let $\{u_n\}_{n\geq 1}$ be a minimizing sequence, i.e., such that $u_n \in \mathscr{S}$ and $\lim_{n\to\infty} \mathfrak{E}(u_n) = E_*$. Then, there exists $C \in \mathbb{R}$ such that $\mathfrak{E}(u_n) \leq C$ for all $n \geq 1$. According to (3.68),
$$C \geq \mathfrak{E}(u_n) = \mathfrak{a}(u_n, u_n) \geq \frac{1}{4}\|u_n\|_{H^1}^2 - \Theta\|u_n\|_{L^2}^2.$$
From $\|u_n\|_{L^2} = 1$, we infer that $\|u_n\|_{H^1}^2 \leq 4(C + \Theta)$. Thus, the minimizing sequence $\{u_n\}_{n\geq 1}$ is bounded in $H^1(0, L)$.

2. We can therefore extract a subsequence, also denoted by $\{u_n\}_{n\geq 1}$, that converges weakly toward some element $u_* \in H^1(0, L)$. At this stage, we do not know whether or not u_* is L-periodic. The weak convergence
$$u_n \overset{H^1}{\rightharpoonup} u_*$$
and the convexity of the functionals $v \mapsto \int_0^L |v|^2$ and $v \mapsto \int_0^L |v'|^2$ implies that
$$\|u_*\|_{L^2}^2 \leq \liminf_{n\to+\infty} \|u_n\|_{L^2}^2 = 1, \tag{3.72a}$$
$$\|u_*'\|_{L^2}^2 \leq \liminf_{n\to+\infty} \|u_n'\|_{L^2}^2. \tag{3.72b}$$

3. By the Rellich-Kondrashov theorem [2], we have the compact embedding
$$H^1(0, L) \subset\subset C^0([0, L]).$$
This means that we can extract from $\{u_n\}_{n\geq 1}$ a subsequence, again denoted by $\{u_n\}_{n\geq 1}$, that converges strongly toward u_* in $C^0([0, L])$. This entails, in particular, that
$$u_*(0) = \lim_{n\to+\infty} u_n(0) = \lim_{n\to+\infty} u_n(L) = u_*(L).$$

In other words, u_* is L-periodic and $u_* \in \mathscr{V} = H^1_\#(0,L)$. The compact embedding also ensures that

$$u_*(X_I) = \lim_{n\to+\infty} u_n(X_I) \tag{3.73}$$

for all $1 \le I \le M$. Combining (3.72b) and (3.73), we end up with

$$\mathfrak{E}(u_*) \le \liminf_{n\to+\infty} \mathfrak{E}(u_n) = E_*. \tag{3.74}$$

4. By the Rellich-Kondrashov theorem [2], we also have the compact embedding

$$H^1(0,L) \subset\subset L^2(0,L)$$

This means that we can extract from $\{u_n\}_{n\ge1}$ a subsequence, again denoted by $\{u_n\}_{n\ge1}$, that converges strongly toward u_* in $L^2(0,L)$. This entails, in particular, that

$$\|u_*\|_{L^2} = \lim_{n\to+\infty} \|u_n\|_{L^2} = 1.$$

Thus, $u_* \in \mathscr{S}$ and $E_* \ge \mathfrak{E}(u_*)$. Combined with (3.74), this yields $E_* = \mathfrak{E}(u_*)$. □

REMARK 3.2. It is worth noting that, unlike the proof of Theorem 3.1 for an infinite domain, here the compact embedding $H^1(0,L) \subset\subset L^2(0,L)$ enables us to conclude without knowing the sign of E_*.

3.2.3 Properties of negative energy eigenstates

As explained in §3.1.3, we have to investigate the properties of all solutions of (3.58), called "eigenstates," before being able to go further. A better knowledge of these eigenstates, assuming that they exist, will help us establishing uniqueness of the minimizer u_*.

Kato's condition

Theorem 3.9. *The wave function u of every solution $(u,E) \in H^1_\#(0,L)\times\mathbb{R}$ of* (3.58)

1. *is infinitely differentiable on each interval*

$$(0,X_1),(X_1,X_2),\dots,(X_{M-1},X_M),(X_M,L);$$

2. *has a jump in derivative at each nucleus location X_I, $1 \le I \le M$, that satisfies the Kato condition*

$$\frac{u'(X_I^+) - u'(X_I^-)}{2} = -Z_I u(X_I); \tag{3.75}$$

3. *satisfies*

$$u'(0^+) = u'(L^-). \tag{3.76}$$

PROOF. The first two assertions are proven in a manner exactly identical to Theorem 3.2. For the third assertion, take a smooth periodic test function v such that

$$v(0) = v(L) \ne 0 \quad \text{and} \quad \operatorname{supp} v \subset [0,\epsilon) \cup (L-\epsilon, L]$$

with $\epsilon > 0$ small enough so that $[0, \epsilon) \cup (L - \epsilon, L]$ does not contain any nucleus X_I. Integration by parts yields

$$\int_0^\epsilon u'v' = -u'(0^+)v(0) - \int_0^\epsilon u''v,$$
$$\int_{L-\epsilon}^L u'v' = \quad u'(L^-)v(L) - \int_{L-\epsilon}^L u''v.$$

Summing these two equalities and arguing that $u'' = -2Eu$ holds in the classical sense on each open interval $(0, \epsilon)$ and $(L - \epsilon, L)$, and taking into account the fact that $v = 0$ over $(\epsilon, L - \epsilon)$, we end up with

$$\int_0^L u'v' = \int_0^\epsilon u'v' + \int_{L-\epsilon}^L u'v' = [u'(L^-) - u'(0^+)]v(0) + 2E\int_0^L uv.$$

After subtraction to the variational formulation (3.64a) multiplied by 2, the above equation leads to $[u'(L^-) - u'(0^+)]v(0) = 0$. A further division by $v(0) \neq 0$ gives rise to (3.76). □

Multi-Slater comb form

Contrary to the infinite model, eigenstates with non-negative energies $E \geq 0$ can no longer be ruled out on a periodic domain. This is because such solutions are now *bona fide* bound states. Since our primary interest lies in the fundamental state, which will be shown to have a negative energy, we shall content ourselves with the necessary conditions for eigenstates with $E < 0$. In preparation for various statements and formulae, it is convenient to introduce a few notions specific to a periodic domain.

Definition 3.1. Given a real number $y \in \mathbb{R}$, the *L-absolute value* of y is the non-negative number $|y|^\sim \in \mathbb{R}_+$ defined as

$$|y|^\sim = |y - nL|, \qquad \text{for } \frac{y}{L} \in \left[n - \frac{1}{2}, n + \frac{1}{2}\right],\ n \in \mathbb{Z}. \tag{3.77}$$

The value (3.77) is the distance from y to the closest integer multiple of L. It is always less than or equal to $L/2$. Equality occurs when y is a half-integer multiple of L, where $|\cdot|^\sim$ is continuous. Given two numbers $(x, y) \in [0, L]^2$, the L-distance $|y - x|^\sim$ is the shortest distance between x and y, seen as points on the closed manifold $[0, L]$ whose ends 0 and L have been identified. This distance is always less than $L/2$.

Definition 3.2. Given a real-valued function f defined over $\mathbb{R}$, the *L-periodization* or the *L-comb* of f is the L-periodic function $\widetilde{f}$ defined as

$$\widetilde{f}(x) = \sum_{n\in\mathbb{Z}} f(x + nL) \tag{3.78}$$

whenever the sum converges pointwise.

The infinite sum in the right-hand side (3.78) converges when f is compactly supported. Without assuming that f has compact support, Definition 3.2 still makes sense when $f \in L^2(\mathbb{R})$ and some additional condition, e.g., on the decay rate of f and $\widehat{f}$, is satisfied. The reader is referred to [33, Theorem 2.28, p. 48] for further details.

The L-comb of the Slater function $\mathrm{S}_{\zeta,X}$ defined in (3.22) can be shown to exist and to be equal to

$$\widetilde{\mathrm{S}}_{\zeta,X}(x) = \frac{\cosh(\zeta(|x-X|\tilde{}-L/2))}{\sinh(\zeta L/2)}. \tag{3.79}$$

To see this, set $y = x - X$ and start with $y \in (0, L/2)$. This allows one to write

$$\widetilde{\mathrm{S}}_{\zeta,X}(x) = \sum_{n\in\mathbb{Z}} \exp(-\zeta|y+nL|) = \sum_{n\geq 0} \exp(-\zeta(y+nL)) + \sum_{n\leq -1} \exp(\zeta(y+nL)).$$

Then, compute the two geometric series to get $\widetilde{\mathrm{S}}_{\zeta,X}(x) = \cosh(\zeta(y-L/2))/\sinh(\zeta L/2)$. Finally, extend the result by symmetry with respect to $y = 0$ and by periodicity.

Theorem 3.10. *Every negative energy solution $(u, E) \in H^1_{\#}(0, L) \times \mathbb{R}$ of* (3.58) *is necessarily of the form*

$$u = \sum_{J=1}^{M} \frac{Z_J}{\zeta} u(X_J)\, \widetilde{\mathrm{S}}_{\zeta,X_J}, \tag{3.80a}$$

$$E = -\frac{1}{2}\zeta^2, \tag{3.80b}$$

where

- $\zeta > 0$ *is a zero of the equation*

$$\det(\mathsf{C}^\zeta - \zeta \mathsf{I}) = 0, \tag{3.81a}$$

with I *the* $M \times M$ *identity matrix and* C^ζ *the* $M \times M$ matrix of compatibility *whose entries are*

$$\mathsf{C}^\zeta_{IJ} = Z_J \frac{\cosh(\zeta(|X_I - X_J|\tilde{}-L/2))}{\sinh(\zeta L/2)}; \tag{3.81b}$$

- *the vector* $\mathsf{u} \in \mathbb{R}^M$ *of components* $\mathsf{u}_J = u(X_J)$ *is a non-trivial solution of the* relations of compatibility

$$\mathsf{C}^\zeta \mathsf{u} = \zeta \mathsf{u}. \tag{3.82}$$

Furthermore, there are at most a finite number of distinct zeros ζ for (3.81a).

Proof. Assuming that the energy is negative, we look for $E = -\zeta^2/2$ with $\zeta > 0$. Instead of applying the Fourier transform, as was done in the proof of Theorem 3.3 for the infinite domain, we resort to the Fourier series for the periodic domain. We start by putting equation (3.58a) under the form

$$-\frac{1}{2}u'' - \sum_{J=1}^{M} Z_J u(X_J)\delta_{X_J} = Eu$$

and compute the Fourier coefficients of both sides. For $k \in \mathbb{Z}$, let

$$\widehat{u}_k = \frac{1}{L}\int_0^L u(x) \exp\left(-\frac{\mathrm{i}2\pi k x}{L}\right) \mathrm{d}x \tag{3.83}$$

be k-th Fourier coefficient of u. Recalling that

$$(\widehat{u''})_k = -\Big(\frac{2\pi k}{L}\Big)^2 \widehat{u}_k, \qquad (\widehat{\delta_X})_k = \frac{1}{L}\exp\Big(-\frac{\mathrm{i}2\pi kX}{L}\Big),$$

we have

$$\frac{1}{2}\Big(\frac{2\pi k}{L}\Big)^2 \widehat{u}_k = \frac{1}{L}\sum_{J=1}^{M} Z_J u(X_J)\exp\Big(-\frac{\mathrm{i}2\pi kX_J}{L}\Big) - \frac{1}{2}\zeta^2\widehat{u}_k,$$

from which we can extract

$$\begin{aligned}\widehat{u}_k &= \frac{2}{\zeta^2 + (2\pi k/L)^2}\cdot\frac{1}{L}\sum_{J=1}^{M} Z_J u(X_J)\exp\Big(-\frac{\mathrm{i}2\pi kX_J}{L}\Big)\\ &= \frac{1}{L}\sum_{J=1}^{M} Z_J u(X_J)\frac{2}{\zeta^2 + (2\pi k/L)^2}\exp\Big(-\frac{\mathrm{i}2\pi kX_J}{L}\Big)\\ &= \sum_{J=1}^{M}\frac{Z_J}{\zeta}u(X_J)\Big\{\frac{1}{L}\cdot\frac{2\zeta}{\zeta^2 + (2\pi k/L)^2}\exp\Big(-\frac{\mathrm{i}2\pi kX_J}{L}\Big)\Big\}. \end{aligned} \tag{3.84}$$

Applying the reconstruction formula and permuting the order of summation, we have

$$\begin{aligned}u(x) &= \sum_{k\in\mathbb{Z}}\widehat{u}_k\exp\Big(\frac{\mathrm{i}2\pi kx}{L}\Big)\\ &= \sum_{J=1}^{M}\frac{Z_J}{\zeta}u(X_J)\Big\{\frac{1}{L}\sum_{k\in\mathbb{Z}}\frac{2\zeta\exp(-\mathrm{i}X_J 2\pi k/L)}{\zeta^2 + (2\pi k/L)^2}\exp\Big(\frac{\mathrm{i}2\pi kx}{L}\Big)\Big\}. \end{aligned} \tag{3.85}$$

Thanks to (3.28), namely,

$$\widehat{\mathrm{S}_{\zeta,X_J}}(\xi) = \frac{2\zeta\exp(-\mathrm{i}X_J\xi)}{\zeta^2+\xi^2},$$

we can turn the expression in the brackets of (3.85) into

$$\frac{1}{L}\sum_{k\in\mathbb{Z}}\widehat{\mathrm{S}_{\zeta,X_J}}\Big(\frac{2\pi k}{L}\Big)\exp\Big(\frac{\mathrm{i}2\pi kx}{L}\Big) = \sum_{n\in\mathbb{Z}}\mathrm{S}_{\zeta,X_J}(x+nL) = \widetilde{\mathrm{S}}_{\zeta,X_J}(x) \tag{3.86}$$

using the Poisson summation formula [1]. Finally, the reconstruction (3.85) yields the desired superposition (3.80a), that is,

$$u(x) = \sum_{J=1}^{M}\frac{Z_J}{\zeta}u(X_J)\frac{\cosh(\zeta(|x-X_J|^{\sim} - L/2))}{\sinh(\zeta L/2)}. \tag{3.87}$$

The values $u(X_J)$ cannot be prescribed freely. They are subject to the relations of compatibility

$$u(X_I) = \sum_{J=1}^{M}\frac{Z_J}{\zeta}\frac{\cosh(\zeta(|X_I-X_J|^{\sim} - L/2))}{\sinh(\zeta L/2)}\,u(X_J),$$

that result from specifying $x = X_I$ in (3.87). Gathering all of these conditions for $I \in \{1,\dots,M\}$ and multiplying by ζ, we have the matrix-vector relation (3.82), where $\mathbf{C}^{\zeta}$ is defined in (3.81b). For a solution $\mathbf{u} \neq \mathbf{0}$ to exist, ζ must be an eigenvalue of the matrix $\mathbf{C}^{\zeta}$ (depending itself on ζ), whence the characterization (3.81a) for ζ.

Expanding the determinant (3.81a) using (3.81b), multiplying by $\sinh^M(\zeta L/2)$, using $2\cosh(\cdot) = \exp(\cdot) + \exp(-\cdot)$ and $2\sinh(\cdot) = \exp(\cdot) - \exp(-\cdot)$, regrouping terms multiplied by the same exponential, we see that it takes the form

$$f(\zeta) := \sinh^M(\zeta L/2)\det(\mathbf{C}^\zeta - \zeta\mathbf{I}) = \sum_{\ell=1}^{\mathcal{M}} \exp(q_\ell\zeta)P_\ell(\zeta),$$

in which $\mathcal{M}$ is some finite integer, P_ℓ is a non-zero polynomial and the exponents $q_\ell \in \mathbb{R}$ are distinct from each other. This is the exactly same form as in (3.30). Repeating the same argument as in the proof of Theorem 3.3, we can prove that f has at most a finite number of zeros. □

Should a negative energy solution of (3.58) exist, it is the superposition of M Slater combs $\widetilde{\mathrm{S}}_{\zeta,X_J}$ centered at each nucleus location and having the same orbital exponent ζ. The latter must be a zero of the "characteristic" equation (3.81a), whose number of solutions depend on the parameters (X_I, Z_I) of the problem. Even for a given "eigenvalue" $\zeta > 0$ of (3.81a), there might be several "eigenvectors" $\mathbf{u}$ that satisfy the compatibility system (3.82). The existence of a minimizer for $\mathfrak{E}$ does guarantee that there exists at least a solution ζ_* to (3.81a), the largest one (corresponding to the smallest E_*).

Sobolev regularity

As a consequence of the multi-Slater comb form (3.80a), we are going to determine the Sobolev regularity of any eigenstate u with a negative energy. Let us recall that, for a given $s \in \mathbb{R}$,

$$u \in H^s_\#(0,L) \Longleftrightarrow \|u\|^2_{H^s_\#(0,L)} := L\sum_{k\in\mathbb{Z}}(1+|2\pi k/L|^2)^s|\widehat{u}_n|^2 < \infty, \tag{3.88}$$

where $\widehat{u}_k$ is the k-th Fourier coefficient of u defined in (3.83). It is worth noting that for $s = 1$,

$$\|u\|^2_{H^1_\#(0,L)} = L\sum_{k\in\mathbb{Z}}(1+|2\pi k/L|^2)|\widehat{u}_k|^2 = \int_0^L |u|^2 + |u'|^2 = \|u\|^2_{H^1(0,L)}.$$

Corollary 3.5. *If $(u,E) \in H^1_\#(0,L)\times\mathbb{R}$ is a solution of (3.58) with $E < 0$, then*

$$u \in H^{3/2-\epsilon}_\#(0,L) \qquad \textit{for all } \epsilon > 0.$$

Proof. We first prove that any L-comb of Slater $\widetilde{\mathrm{S}}_{\zeta,X}$ belongs to $H^{3/2-\epsilon}_\#(0,L)$ for any $\epsilon > 0$. Indeed, as was seen in the Poisson summation formula (3.86),

$$(\widehat{\widetilde{\mathrm{S}}_{\zeta,X}})_k = \frac{1}{L}\widehat{\mathrm{S}_{\zeta,X}}\left(\frac{2\pi k}{L}\right).$$

But the (continuous) Fourier transform of $\mathrm{S}_{\zeta,X}$ is

$$\widehat{\mathrm{S}_{\zeta,X}}(\xi) = \frac{2\zeta\exp(-\mathrm{i}X\xi)}{\zeta^2+\xi^2}.$$

Taking the square of the modulus and setting $\xi = 2\pi k/L$, we have the equivalence

$$(1+|2\pi k/L|^2)^s|(\widehat{\widetilde{\mathrm{S}}_{\zeta,X}})_k|^2 \sim |2\pi k/L|^{2s}\cdot\frac{4\zeta^2}{L^2|2\pi k/L|^4} = \frac{4\zeta^2}{L^2}\left|\frac{L}{2\pi k}\right|^{4-2s}$$

when $k\to\infty$, assuming that $s\geq 0$. This assumption is justified since $\widetilde{\mathrm{S}}_{\zeta,X}\in L^2(0,L) = H^0_{\#}(0,L)$. Therefore,

$$\sum_{k\in\mathbb{Z}}(1+|2\pi k/L|^2)^s|(\widehat{\widetilde{\mathrm{S}}_{\zeta,X}})_k|^2<\infty \Longleftrightarrow \sum_{k\geq 1}\frac{1}{|k|^{4-2s}}<\infty.$$

The convergence of the latter sum is equivalent to $4-2s>1$, that is, $s<3/2$. As a finite linear combination (3.80a) of L-combs of Slater functions, u has at least the same Sobolev regularity $3/2-\epsilon$. □

This result will be useful in §4.3.2 of the next chapter for the a priori error estimate of the Galerkin approximation of the fundamental state by a basis of scaling functions.

3.2.4 Single- and double-delta potentials

From the general forms in Theorems 3.10, we can write down explicit formulae for the negative energy solutions in the special cases $M=1$ and $M=2$.

Definition 3.3. Given a real number $z>0$, the *L-alteration* of z is the unique positive zero $\widetilde{z}>0$ of the equation

$$\widetilde{z} = z\coth\left(\frac{\widetilde{z}L}{2}\right). \tag{3.89}$$

We always have $\widetilde{z}>z$. Since z and $\widetilde{z}$ are meant to be charges, $1/z$ and $1/\widetilde{z}$ are lengths and the inequality $1/\widetilde{z}<1/z$ expresses the fact that in a bounded periodic domain, the characteristic length is shorter than that of the infinite domain. When L is large, the hyperbolic cotangent is close to 1 and we have $\widetilde{z}\approx z$. It is not difficult to check that the function $z\mapsto\widetilde{z}$ is increasing on $\mathbb{R}_+^*$.

Single-delta

When $M=1$, the equations to be solved are

$$-\frac{1}{2}u'' - Z\delta_X u = Eu, \tag{3.90a}$$

$$\|u\|_{L^2(0,L)} = 1, \tag{3.90b}$$

for a given charge $Z>0$ and a given nucleus location $X\in(0,L)$. Periodic boundary conditions are implicitly included.

Theorem 3.11. *The only negative energy solution of problem* (3.90) *is, up to sign of the wave function, by*

$$u_*(x) = \frac{2^{1/2}}{[L+\widetilde{Z}^{-1}\sinh(L\widetilde{Z})]^{1/2}}\cosh(\widetilde{Z}(|x-X|\tilde{}-L/2)), \tag{3.91a}$$

$$E_* = -\frac{1}{2}\widetilde{Z}^2. \tag{3.91b}$$

It is thus also the minimizer of $\mathfrak{E}(u) = \frac{1}{2}\|u'\|_{L^2}^2 - Z|u(X)|^2$ *on* $\mathscr{S}$.

PROOF. The general form (3.80a) of Theorem 3.10 boils down in our case to

$$u(x) = \frac{Z}{\zeta} u(X) \frac{\cosh(\zeta(|x - X|^{\sim} - L/2))}{\sinh(\zeta L/2)},$$

while $u(X)$ is subject to the relation of compatibility (3.82), which reads

$$Z u(X) \coth(\zeta L/2) = \zeta u(X).$$

If $u(X) = 0$, then u vanishes identically. Therefore, $u(X) \neq 0$ and $\zeta = Z \coth(\zeta L/2)$ implies $\zeta = \widetilde{Z}$. There is only one solution with $E < 0$ and this solution is given by (3.91), after normalization of u. By Theorem 3.8, there exists a minimizer u_*. We also know that any minimizer must satisfy (3.90). Therefore, formulae (3.91) supply us with the only minimizer possible. □

When $L \to +\infty$, formulae (3.91) degenerate to their infinite domain counterparts (3.33). The knowledge of the solution for a single-delta potential is helpful for finding bounds on the ground state energy in the case of a multi-delta potential. In §3.2.5, we shall be using the first part of Theorem 3.11 under the following form: for all $\zeta > 0$ and all $X \in (0, L)$,

$$-\frac{1}{2}\widetilde{\zeta}^2 = \min_{\substack{v \in H^1_{\#}(0,L) \\ \|v\|_{L^2} = 1}} \frac{1}{2}\|v'\|^2_{L^2} - \zeta |v(X)|^2 = \frac{1}{2}\|(\widetilde{\mathrm{S}}_{\zeta,X})'\|^2_{L^2} - \zeta |\widetilde{\mathrm{S}}_{\zeta,X}(X)|^2. \tag{3.92}$$

Double-delta

When $M = 2$, the equations to be solved are

$$-\frac{1}{2}u'' - (Z_1 \delta_{X_1} + Z_2 \delta_{X_2})u = Eu, \tag{3.93a}$$

$$\|u\|_{L^2(0,L)} = 1. \tag{3.93b}$$

To determine all eigenstates of (3.93) with $E < 0$, we introduce the internuclear distance

$$R = |X_2 - X_1|^{\sim} \in (0, L/2]$$

and recall that $\Lambda_1 = Z_1^{-1}$ and $\Lambda_2 = Z_2^{-1}$ are characteristic lengths associated with the charges.

Theorem 3.12. *The negative energy solutions of problem* (3.93) *are given by*

$$u = \frac{Z_1}{\zeta} u(X_1)\, \widetilde{\mathrm{S}}_{\zeta,X_1} + \frac{Z_2}{\zeta} u(X_2)\, \widetilde{\mathrm{S}}_{\zeta,X_2}, \tag{3.94a}$$

$$E = -\frac{1}{2}\zeta^2 \tag{3.94b}$$

where ζ is a zero of the equation

$$(Z_1 - \zeta)(Z_2 - \zeta) = (Z_1 + Z_2)\zeta[\coth(\zeta L/2) - 1] + Z_1 Z_2 \frac{\sinh^2(\zeta(L/2 - R))}{\sinh^2(\zeta L/2)}, \tag{3.95a}$$

and where $(u(X_1), u(X_2))^T$ is a suitably normalized non-trivial vector satisfying

$$\begin{pmatrix} Z_1 \cosh(\zeta L/2) - \zeta \sinh(\zeta L/2) & Z_2 \cosh(\zeta(L/2 - R)) \\ Z_1 \cosh(\zeta(L/2 - R)) & Z_2 \cosh(\zeta L/2) - \zeta \sinh(\zeta L/2) \end{pmatrix} \begin{pmatrix} u(X_1) \\ u(X_2) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{3.95b}$$

These negative energy solutions consist of

1. *a fundamental state* (u_*, E_*)*, that corresponds to the unique zero* ζ_* *of* (3.102) *such that*

$$\max\{\widetilde{Z_1}, \widetilde{Z_2}\} < \zeta_* < \widetilde{Z_1 + Z_2}. \tag{3.96}$$

2. *a finite number of excited states* $(u_\sharp, E_\sharp)$*, that exist if and only if*

$$R\left(1 - \frac{R}{L}\right) > \frac{\Lambda_1 + \Lambda_2}{2} \tag{3.97}$$

and that correspond to the zeros $\zeta_\sharp$ *of* (3.102) *such that*

$$0 < \zeta_\sharp < \min\{Z_1, Z_2\}. \tag{3.98}$$

The idea of the proof is illustrated in Figure 3.4. In the same spirit as in Theorem 3.5 for the infinite model, we look for the intersection of the graphs representing the two sides of (3.95a): the parabola $\mathfrak{p}$ (in blue) for the left-hand side $(Z_1 - \zeta)(Z_2 - \zeta)$ and the curve $\mathfrak{g}$ (in red) for the right-hand side $g(\zeta)$, which has replaced the exponential $\exp(-2R\zeta)$. The trouble with g is that its third derivative does not have a constant sign, which prevents us from transposing the argument. This obstacle, however, can be overcome by means of an auxiliary function h whose graph $\mathfrak{h}$ (in green) lies below $\mathfrak{g}$ under some circumstances and whose third derivative does not vanish.

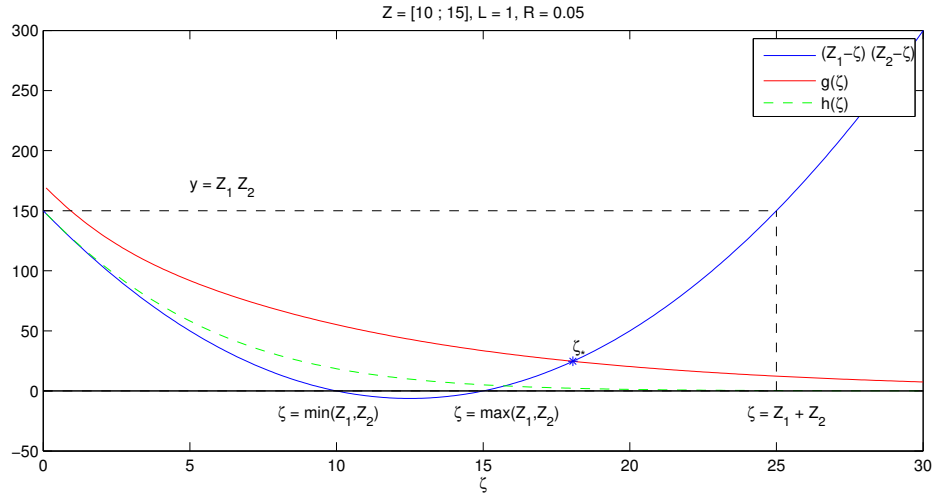

Figure 3.4: The parabola $\mathfrak{p} : \zeta \mapsto (Z_1 - \zeta)(Z_2 - \zeta)$, the curve $\mathfrak{g} : \zeta \mapsto g(\zeta)$ and the curve $\mathfrak{h} : \zeta \mapsto h(\zeta)$. If (3.100) holds, then $\mathfrak{h}$ lies below $\mathfrak{g}$.

Lemma 3.5. *Consider the functions*

$$g(\zeta) = (Z_1 + Z_2)\zeta[\coth(\zeta L/2) - 1] + Z_1 Z_2 \frac{\sinh^2(\zeta(L/2 - R))}{\sinh^2(\zeta L/2)}, \tag{3.99a}$$

$$h(\zeta) = (Z_1 + Z_2)\zeta[\coth(\zeta(\Lambda_1 + \Lambda_2)) - 1], \tag{3.99b}$$

defined on $\zeta > 0$. At $\zeta = 0^+$, it is possible to define by continuity

$$\begin{aligned} g(0) &= Z_1 Z_2 (1 - 2R/L)^2 + 2(Z_1 + Z_2)/L, \qquad & g'(0) &= -(Z_1 + Z_2), \\ h(0) &= Z_1 Z_2, & h'(0) &= -(Z_1 + Z_2). \end{aligned}$$

Moreover, if

$$R\Big(1 - \frac{R}{L}\Big) \leq \frac{\Lambda_1 + \Lambda_2}{2} \tag{3.100}$$

then for all $\zeta \geq 0$,

$$g(\zeta) \geq h(\zeta).$$

Proof. The values of $g(0)$, $g'(0)$ and $h(0)$, $h'(0)$ result from a direct calculation and from $\Lambda_1 + \Lambda_2 = (Z_1 + Z_2)/Z_1 Z_2$. Using the formula

$$\coth(\alpha) - \coth(\beta) = \frac{\sinh(\beta - \alpha)}{\sinh\alpha \sinh\beta},$$

we can express the difference between g and h as

$$g(\zeta) - h(\zeta) = \zeta(Z_1 + Z_2) \frac{\sinh(\zeta(\Lambda_1 + \Lambda_2 - L/2))}{\sinh(\zeta L/2)\sinh(\zeta(\Lambda_1 + \Lambda_2))} + Z_1 Z_2 \frac{\sinh^2(\zeta(L/2 - R))}{\sinh^2(\zeta L/2)}.$$

Two cases can arise. If $\Lambda_1 + \Lambda \geq L/2$, then obviously both terms in the right-hand side are non-negative and therefore $g(\zeta) - h(\zeta) \geq 0$. In the other case $\Lambda_1 + \Lambda_2 < L/2$, there exists a unique $r \in (0, 1)$ such that

$$\Lambda_1 + \Lambda_2 = \frac{L}{2}(1 - r^2).$$

Introduction of the dimensionless quantities

$$z = \frac{\zeta L}{2} \in \mathbb{R}_+, \quad s = 1 - \frac{2R}{L} \in (0, 1)$$

and some algebra yields

$$\Lambda_1 \Lambda_2 [g(\zeta) - h(\zeta)] = -(1 - r^2) z \frac{\sinh(r^2 z)}{\sinh(z)\sinh((1 - r^2)z)} + \frac{\sinh^2(sz)}{\sinh^2(z)}.$$

Plugging $R/L = (1 - s)/2$ into assumption (3.100), we find $1 - s^2 \leq 1 - r^2$, from which we deduce that $s \geq r$. It follows that

$$\Lambda_1 \Lambda_2 [g(\zeta) - h(\zeta)] \geq -(1 - r^2) z \frac{\sinh(r^2 z)}{\sinh(z)\sinh((1 - r^2)z)} + \frac{\sinh^2(rz)}{\sinh^2(z)}.$$

The positivity of the right-hand side is tantamount to

$$\sinh((1 - r^2)z)\sinh^2(rz) \geq (1 - r^2) z \sinh(z)\sinh(r^2 z)$$

or

$$\frac{\sinh((1 - r^2)z)}{(1 - r^2)z} \Big[\frac{\sinh(rz)}{rz}\Big]^2 \geq \frac{\sinh(z)}{z} \frac{\sinh(r^2 z)}{r^2 z}. \tag{3.101}$$

In view of the multiplicative nature of the inequality to be proved, it is judicious to resort to the infinite product [1]

$$\frac{\sinh(\pi y)}{\pi y} = \prod_{n=1}^{\infty} \left(1 + \frac{y^2}{n^2}\right).$$

Setting $z = \pi y$, the sought-for inequality (3.101) becomes

$$\prod_{n=1}^{\infty} \left(1 + \frac{(1-r^2)^2 y^2}{n^2}\right) \prod_{n=1}^{\infty} \left(1 + \frac{r^2 y^2}{n^2}\right)^2 \geq \prod_{n=1}^{\infty} \left(1 + \frac{y^2}{n^2}\right) \prod_{n=1}^{\infty} \left(1 + \frac{r^4 y^2}{n^2}\right).$$

All the products converge, thus we are allowed to multiply term by term to have

$$\prod_{n=1}^{\infty} \left(1 + \frac{(1+r^4)y^2}{n^2} + \frac{r^2(2-3r^2+2r^4)y^4}{n^4} + \frac{r^4(1-2r^2+r^4)y^6}{n^6}\right) \geq \prod_{n=1}^{\infty} \left(1 + \frac{(1+r^4)y^2}{n^2} + \frac{r^4 y^4}{n^4}\right).$$

It is enough to prove that each term on left is larger than the corresponding term on the right. This is true because their difference

$$\frac{2r^2(1-r^2)^2 y^4}{n^4} + \frac{r^4(1-r^2)^2 y^6}{n^6}$$

is obviously non-negative. □

Proof of Theorem 3.12. The general form (3.80a) of Theorem 3.10 reduces in our case to (3.94a). The relations of compatiblity (3.82) for $(u(X_1), u(X_2))$ become (3.95b). The characteristic equation $\det(\mathbf{C}^\zeta - \zeta \mathbf{I}) = 0$ reads

$$(Z_1 \coth(\zeta L/2) - \zeta)(Z_2 \coth(\zeta L/2) - \zeta) - Z_1 Z_2 \frac{\cosh^2(\zeta(L/2 - R))}{\sinh^2(\zeta L/2)} = 0. \tag{3.102}$$

Its equivalence with (3.95a) can be derived by expanding the product, inserting $-(Z_1 + Z_2)\zeta$, compensating and factorizing again. Let

$$\begin{aligned} f(\zeta) &= (Z_1 \coth(\zeta L/2) - \zeta)(Z_2 \coth(\zeta L/2) - \zeta) - Z_1 Z_2 \frac{\cosh^2(\zeta(L/2 - R))}{\sinh^2(\zeta L/2)} \\ &= (Z_1 - \zeta)(Z_2 - \zeta) - g(\zeta) \end{aligned}$$

the common left-hand side of (3.102) and (3.95a), where g is defined in Lemma 3.5.

— *Existence and uniqueness of* ζ_*. Setting $Z = \max\{Z_1, Z_2\}$, we notice on one hand that the quadratic polynomial $\zeta \mapsto (Z_1 - \zeta)(Z_2 - \zeta)$ is increasing on $(Z, +\infty)$. On the other hand, by brute force differentiation, we can show that

$$\zeta \mapsto \zeta(\coth(\zeta L/2) - 1) \quad \text{and} \quad \zeta \mapsto \frac{\sinh^2(\zeta(L/2 - R))}{\sinh^2(\zeta L/2)}$$

are decreasing functions of ζ on $\mathbb{R}_+$. As a result, f is an increasing function of ζ on $(Z, +\infty)$. Now, we observe that

$$f(\widetilde{Z_1}) = 0 - Z_1 Z_2 \frac{\widetilde{Z_1} \cosh^2((L/2 - R))}{\sinh^2(\widetilde{Z_1} L/2)} < 0$$

because $Z_1 \coth(\widetilde{Z_1}L/2) - \widetilde{Z_1} = 0$ by definition of $\widetilde{Z_1}$. Likewise, $f(\widetilde{Z_2}) < 0$ so that $f(\widetilde{Z}) < 0$. Setting $\mathcal{Z} = Z_1 + Z_2$ and expanding

$$f(\zeta) = \zeta[\zeta - \mathcal{Z}\coth(\zeta L/2)] + Z_1 Z_2 \Big(\coth^2(\zeta L/2) - \frac{\cosh^2(\zeta(L/2-R))}{\sinh^2(\zeta L/2)} \Big),$$

we have

$$f(\widetilde{\mathcal{Z}}) = 0 + Z_1 Z_2 \Big(1 - \frac{\sinh^2(\zeta(L/2-R))}{\sinh^2(\zeta L/2)} \Big) > 0.$$

By virtue of the intermediate value theorem, there exists $\zeta_* \in (\widetilde{Z}, \widetilde{\mathcal{Z}})$ such that $f(\zeta_*) = 0$. By monotonicity, this is the only zero of f on $(Z, +\infty)$ and gives therefore rise to the lowest energy. Because $\zeta_* > \widetilde{Z_1}$ and $\zeta_* > \widetilde{Z_2}$, the matrice $\mathbf{C}^{\zeta_*} - \zeta_* \mathbf{I}$ is not identically zero and thus there is just one "eigenvector" $\mathbf{u}_*$ associated with ζ_*, up to a normalization constant.

— *Non-existence of $\zeta_\sharp$ when $R(1 - R/L) \leq (\Lambda_1 + \Lambda_2)/2$.* From the auxiliary function h introduced in (3.99b) of Lemma 3.5, we define

$$d(\zeta) = (Z_1 - \zeta)(Z_2 - \zeta) - h(\zeta).$$

By construction, $d(0) = d'(0) = 0$. We are going to show that $d(\zeta) < 0$ for all $\zeta \in (0, Z)$. First, we note that by the same reasoning as above: (1) h is decreasing on $(0, +\infty)$ and d is decreasing on $(Z, +\infty)$; (2) there exists $\zeta_\flat \in (Z, +\infty)$ such that $d(\zeta_\flat) = 0$. Next, assume there is a $\zeta_\sharp \in (0, Z)$ such that $d(\zeta_\sharp) = 0$. In totality, d has already at least 3 zeros: 0, $\zeta_\sharp$ and $\zeta_\flat$. Applying Rolle's theorem, we can find two distinct zeros ζ_1 and ζ_2 for d' in $(0, \zeta_\flat)$. Since $d'(0) = 0$, this makes 3 distinct zeros for d': 0, ζ_1 and ζ_2. Successive applications of Rolle's theorem yield 2 distinct zeros for d'' and 1 zero for d'''. But

$$d'''(\zeta) = -h'''(\zeta) = \mathcal{Z}^3 \frac{4(\mathcal{Z}\zeta) - 3\sinh(2\mathcal{Z}\zeta) + 2(\mathcal{Z}\zeta)\cosh(2\mathcal{Z}\zeta)}{\sinh^4(\mathcal{Z}\zeta)}$$

and by studying the numerator, we can show that $d'''(\zeta) > 0$ for $\zeta > 0$. It follows by contradiction that d keeps a constant sign on $(0, Z)$. This sign is given by $d(\min\{Z_1, Z_2\}) = 0 - h(\min\{Z_1, Z_2\}) < 0$. If (3.100) holds, then Lemma 3.5 secures $g(\zeta) \geq h(\zeta)$. Hence, $f(\zeta) \leq d(\zeta) < 0$ for $\zeta \in (0, Z)$ and f has no other zero than ζ_*.

— *Existence of at most a finite number of $\zeta_\sharp$ when $R(1 - R/L) > (\Lambda_1 + \Lambda_2)/2$.* Otherwise, that is, when (3.97) holds,

$$f(0) = Z_1 Z_2 - g(0) = Z_1 Z_2 \Big[1 - \Big(1 - \frac{2R}{L} \Big)^2 \Big] - \frac{2(Z_1 + Z_2)}{L} > 0$$

and since $f(\min\{Z_1, Z_2\}) = 0 - g(\min\{Z_1, Z_2\}) < 0$, there is at least one zero $\zeta_\sharp \in (0, \min\{Z_1, Z_2\})$ for f. By Theorem 3.10, there are at most a finite number of such zeros. For each of these $\zeta_\sharp$, since $\zeta_\sharp < \min\{Z_1, Z_2\} < \min\{\widetilde{Z_1}, \widetilde{Z_2}\}$, the matrix $\mathbf{C}^{\zeta_\sharp} - \zeta_\sharp \mathbf{I}$ is not identically zero and there is only one eigenvector $\mathbf{u}_\sharp$ associated with $\zeta_\sharp$. □

When $L \to +\infty$, condition (3.97) for the appearance of an excited state degenerates to its infinite domain counterpart (3.39). For the infinite domain, we were able to prove (Theorem 3.5) that there is at most one excited state. For the periodic domain, despite

extensive numerical evidences, we do not have a rigorous proof that there is at most one negative energy excited state[2].

Conjecture 3.1. *When* (3.97) *occurs, there is exactly one negative energy excited state* $(u_\sharp, E_\sharp)$ *that corresponds to a unique zero* $\zeta_\sharp \in (0, \min\{Z_1, Z_2\})$ *of* (3.102).

For equal charges $Z_1 = Z_2 = Z$, we call the fundamental solution *gerade* because of the symmetry $u(X_1) = u(X_2)$. Unlike the infinite domain, there is no equivalent of the Lambert function to express the eigenstates in closed-form. All we can do is to extract the square root of (3.102) to have the simpler equation

$$\zeta \sinh(\zeta L/2) - Z\cosh(\zeta L/2) = \pm Z\cosh(\zeta(L/2-R)).$$

Like in the infinite domain, the wave function is highly "sensitive" to a slight perturbation of the charges. This phenomenon is described by the following statement, which is the counterpart of Proposition 3.3. Clearly, the amplification factor in the first-order expansion (3.103) degenerates to that of (3.51) when $L \to +\infty$.

Proposition 3.6. *In model* (3.93)*, consider the almost identical charges*

$$Z_1 = Z + \Delta Z, \qquad Z_2 = Z,$$

with ΔZ *a small variation, i.e.,* $|\Delta Z| \ll Z$*. Then, the values of the fundamental state* u *at the two cusps are in the ratio*

$$\frac{u(X_1)}{u(X_2)} = 1 + \left[\frac{\cosh(\zeta_* L/2)}{\cosh(\zeta_*(L/2-R))} - 1\right]\frac{\Delta Z}{2Z} + O(\Delta Z^2), \tag{3.103}$$

where ζ_* *is the* gerade *solution of the equal charges problem.*

Proof. Let $\zeta_* + \Delta\zeta_*$ be the gerade solution of the double-delta problem with almost identical charges. Then, $\zeta_* + \Delta\zeta_*$ is a root of

$$\begin{aligned}[(\zeta_*+\Delta\zeta_*)\sinh((\zeta_*+\Delta\zeta_*)L/2) - (Z+\Delta Z)\cosh((\zeta_*+\Delta\zeta_*)L/2)]\\ \cdot[(\zeta_*+\Delta\zeta_*)\sinh((\zeta_*+\Delta\zeta_*)L/2) - Z\cosh((\zeta_*+\Delta\zeta_*)L/2)]\\ - Z(Z+\Delta Z)\cosh^2((\zeta_*+\Delta\zeta_*)(L/2-R)) = 0.\end{aligned}$$

Carrying out the first-order Taylor expansion and dropping the zeroth-order terms (which cancel out each other), taking into account the property $\zeta_*\sinh(\zeta_* L/2) - Z\cosh(\zeta_* L/2) = Z\cosh(\zeta_*(L/2-R))$ for the gerade solution, we end up with

$$\Delta\zeta_* = -\frac{[\cosh(\zeta_* L/2) + \cosh(\zeta_*(L/2-R))]\Delta Z}{L(Z\sinh(\zeta_* L/2) - \zeta_*\cosh(\zeta_* L/2)) + (L-2R)Z\sinh(\zeta_*(L/2-R))} + O(\Delta Z^2) \tag{3.104}$$

after some (tedious) algebra. In view of the first line of (3.95b), the ratio of amplitudes $u(X_1)/u(X_2)$ is equal to

$$\frac{u(X_1)}{u(X_2)} = \frac{Z\cosh((\zeta_*+\Delta\zeta_*)(L/2-R))}{(\zeta_*+\Delta\zeta_*)\sinh((\zeta_*+\Delta\zeta_*)L/2) - (Z+\Delta Z)\cosh((\zeta_*+\Delta\zeta_*)L/2)}$$

[2]There are also countably many positive energy excited states for the periodic model, which we are not interested in.

Factorizing the numerator by $Z\cosh(\zeta_*(L/2-R))$, the denominator by $\zeta_*\sinh(\zeta_*L/2) - Z\cosh(\zeta_*L/2)$, and invoking again $\zeta_*\sinh(\zeta_*L/2) - Z\cosh(\zeta_*L/2) = Z\cosh(\zeta_*(L/2-R))$, we can transform the above ratio into

$$\frac{u(X_1)}{u(X_2)} = \frac{1+\tanh(\zeta_*(L/2-R))(L/2-R)\Delta\zeta_*}{1-\dfrac{\cosh(\zeta_*L/2)}{Z\cosh(\zeta_*(L/2-R))}\Delta Z + \dfrac{L[\zeta_*\cosh(\zeta_*L/2) - Z\sinh(\zeta_*L/2)]}{2Z\cosh(\zeta_*(L/2-R))}\Delta\zeta_*} + O(\Delta Z^2).$$

Pursuing the first-order expansion and inserting the value (3.104) for $\Delta\zeta_*$ into the new equation, simplifications occur and we finally obtain the ratio (3.103). □

As an illustration, we plot several wave functions in the case of a double-delta potential. In Figure 3.5, when the internuclear distance R is equal to $L/2$, the two deltas are symmetrically located on the "circle" $[0, L]$. The wave function is then symmetric with respect to each nucleus position X_1 and X_2, even if the two charges Z_1, Z_2 are different. If moreover $Z_1 = Z_2$, then the two cusps are identical.

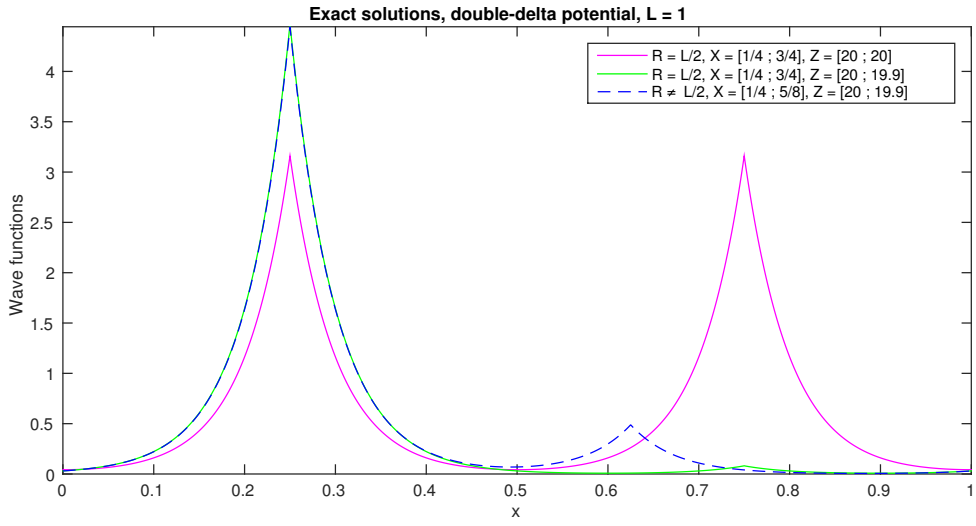

Figure 3.5: Wave function in the case of double-delta potential, periodic domain.

3.2.5 Uniqueness and other properties of the ground state

After having explored the properties of all eigenstates, defined to be solutions of (3.58), we return to the energy viewpoint (3.62).

Theorem 3.13. *The minimizer u_* of* (3.62) *is unique, up to a sign.*

Proof. The proof is very similar to that of Theorem 3.6. In summary:

1. First, we show that a minimizer u_* cannot vanish anywhere in $[0, L]$. Therefore, it must keep a constant sign. This is done by means of the Kato condition and the differential equation $v'' = -2Ev$ to be satisfied by $|u_*|$.

2. If u_* minimizes $\mathfrak{E}$ over $\mathscr{S}$, then $\rho_* = |u_*|^2$ remains strictly positive and minimizes the functional
$$\rho \mapsto \mathfrak{E}(\sqrt{\rho}) = \frac{1}{2}\int_{\mathbb{R}} |(\sqrt{\rho})'|^2 - \sum_{I=1}^{M} Z_I \rho(X_I)$$
over
$$\mathscr{C} = \left\{\rho > 0, \quad \sqrt{\rho} \in H^1_{\#}(0,L), \quad \int_{\mathbb{R}} \rho = 1\right\}.$$
By strict convexity of the functional $\rho \mapsto \mathfrak{E}(\sqrt{\rho})$ on the convex set $\mathscr{C}$, the minimizer ρ_* is unique. Thus, $u_* = \sqrt{\rho_*} > 0$ or $u_* = -\sqrt{\rho_*} < 0$. □

Let $E^{(1)}$ be another name for E_*, the "first eigenvalue." Analogously to what happens in an infinite domain, the "second eigenvalue"
$$E^{(2)} = \inf_{\substack{v \in (u_*)^\perp \subset \mathscr{V} \\ \|v\|_{L^2}=1}} \mathfrak{E}(v),$$
is larger than $E^{(1)}$ but cannot collapse to $E^{(1)}$ because of uniqueness.

Corollary 3.6. $E^{(2)} > E^{(1)}$.

Proof. The proof is very similar to that of Corollary 3.4. The basic idea is to apply the proof machinery for existence of a minimizer on the orthogonal complement $(u_*)^\perp$ in $\mathscr{V}$. Note that this time, we do not need to know the sign of E_* in advance. We also have to check that the minimizer $u^{(2)}$ is L-periodic, but this is most easy. □

The fact that $E^{(2)} > E^{(1)}$ ensures that the bilinear form $\mathfrak{a}(\cdot,\cdot) - E_*\mathfrak{b}(\cdot,\cdot)$ is L^2-coercive on $(u_*)^\perp$, and will be used for error estimates in §4.3.2 and §5.3.

Theorem 3.14. *The fundamental energy E_* is bounded by*
$$-\frac{1}{2}\widetilde{\mathcal{Z}}^2 \le E_* \le -\frac{1}{2}\widetilde{Z}^2, \tag{3.105}$$
where $\widetilde{\mathcal{Z}}$ is the L-alteration of the total charge $\mathcal{Z} = \sum_{J=1}^{M} Z_J$ and $\widetilde{Z}$ is the L-alteration of the greatest charge $Z = \max\{Z_1, \dots, Z_M\}$.

Proof. The proof is similar to that of Theorem 3.7 for the infinite domain. For a fixed $I \in \{1, \dots, M\}$, introduce
$$v_I(x) = \frac{2^{1/2}}{[L + (\widetilde{Z_I})^{-1}\sinh(L\widetilde{Z_I})]^{1/2}} \cosh(\widetilde{Z_I}(|x - X_I|^{\sim} - L/2))$$
From Theorem 3.11, we know that $\|v_I\|_{L^2} = 1$ and that from the reformulation (3.92), we have
$$\frac{1}{2}\|v_I'\|_{L^2}^2 - Z_I|v_I(X_I)|^2 = -\frac{1}{2}\widetilde{Z_I}^2.$$
Going back to the multi-delta potential problem at hand, we have
$$E_* = \min_{\substack{v \in H^1_{\#}(0,L) \\ \|v\|_{L^2}=1}} \mathfrak{E}(v) \le \mathfrak{E}(v_I).$$

But

$$\mathfrak{E}(v_I) = \frac{1}{2}\|v_I'\|_{L^2}^2 - Z_I|v_I(X_I)|^2 - \sum_{J\neq I} Z_J|v_I(X_J)|^2 = -\frac{1}{2}\widetilde{Z_I}^2 - \sum_{J\neq I} Z_J|v_I(X_J)|^2,$$

and thus $E_* \leq -\widetilde{Z_I}^2/2$. Since the function $z \mapsto \widetilde{z}$ is increasing on $\mathbb{R}_+^*$, the minimum of the last quantity over $I \in \{1,\dots,M\}$ yields the upper bound $E_* \leq -\widetilde{Z}^2/2$.

To derive the lower bound, let $\mathcal{Z} = \sum_{J=1}^M Z_J$ stand for the total charge and X some abscissa to be specified later. Again, by (3.34), we know that

$$-\frac{1}{2}\widetilde{\mathcal{Z}}^2 = \min_{\substack{v\in H^1_\#(0,L)\\ \|v\|_{L^2}=1}} \frac{1}{2}\|v'\|_{L^2}^2 - \mathcal{Z}|v(X)|^2 \leq \frac{1}{2}\|u_*'\|_{L^2}^2 - \mathcal{Z}|u_*(X)|^2.$$

But

$$E_* = \mathfrak{E}(u_*) = \frac{1}{2}\|u_*'\|_{L^2}^2 - \sum_{J=1}^M Z_J|u_*(X_J)|^2,$$

so that

$$-\frac{1}{2}\widetilde{\mathcal{Z}}^2 \leq E_* + \sum_{J=1}^M Z_J|u_*(X_J)|^2 - \mathcal{Z}|u_*(X)|^2 = E_* + \sum_{J=1}^M Z_J(|u_*(X_J)|^2 - |u_*(X)|^2).$$

By choosing $X = X_I$ such that $|u(X_I)| = \max_{1\leq J\leq M}|u(X_J)|$, we can make sure that every summand of the second term is non-positive. As a result, $-\widetilde{\mathcal{Z}}^2/2 \leq E_*$. □

When $L \to +\infty$ the sharpened bounds (3.105) degenerate to their infinite domain counterparts (3.56). The lower bound $E_* \geq \widetilde{\mathcal{Z}}^2/2$ will greatly help us devising a "good" norm in §5.3.4 for the practical computation of some *a posteriori* estimate.

Chapter 4

Numerical resolution of the multi-delta models

Contents

Nous appliquons la méthode de Ritz-Galerkin classique aux deux modèles 1-D (infini et périodique) introduits dans le chapitre §3. Ce faisant, nous mettons en avant le rôle prépondérant joué par le niveau d'énergie pour comparer des solutions approchées sans connaître la solution exacte.

Pour le modèle infini, nous proposons d'utiliser le critère d'énergie pour optimiser la base de gaussiennes pures associées à un atome isolé, ce qui débouche ainsi sur une nouvelle construction des "gaussiennes contractées". Cette démarche, qui semble peu performante en raison des difficultés liées à l'optimisation en plusieurs variables, sera reprise et améliorée au chapitre §6 en conjonction avec l'algorithme glouton et l'estimateur a posteriori *conçu au chapitre §5.*

Pour le modèle périodique, nous analysons l'ordre de convergence théorique dans une base de fonctions d'échelle, lequel est confirmé par les simulations numériques pour les potentiels simple-delta et double-delta. Nous présentons ensuite des calculs en base mixte, où les gaussiennes contractées élaborées précédemment sont directement injectées sans aucune adaptation à la base de fonctions d'échelle existante. Ce procédé, appelé "base mixte pré-optimisée", n'est certes pas optimal mais constitue un premier essai. Il sera aussi repris et amélioré dans les chapitres suivants.

4.1 Generalities on the Galerkin method

Let us start by recalling a few basic properties of the Ritz-Galerkin approximation, the method we shall be using for the discretization of the models introduced in §3. This method is, by far, the most popular one for quantum chemistry problems, as shown in the brief survey of §1.

Continuous and discrete variational formulations

The infinite model (3.1) and the periodic model (3.57)–(3.58) can be unified within the same framework. Let $\mathscr{V}$ be some functional space over which it there is a concept of H^1-norm. Over $\mathscr{V} \times \mathscr{V}$ are defined two bilinear forms $\mathfrak{a}(\cdot,\cdot)$ and $\mathfrak{b}(\cdot,\cdot)$. The variational eigenvalue problem is defined as the search for all pairs $(u, E) \in \mathscr{V} \times \mathbb{R}$ such that, for all $v \in \mathscr{V}$,

$$\mathfrak{a}(u, v) = E\, \mathfrak{b}(u, v), \tag{4.1a}$$

$$\mathfrak{b}(u, u) = 1, \tag{4.1b}$$

To be specific,

- for the infinite model, $\mathscr{V} = H^1(\mathbb{R})$ and

$$\mathfrak{a}(u, v) = \frac{1}{2} \int_{\mathbb{R}} u'v' - \sum_{I=1}^{M} Z_I\, u(X_I)v(X_I), \tag{4.2a}$$

$$\mathfrak{b}(u, v) = \int_{\mathbb{R}} uv =: \langle u, v\rangle_{L^2(\mathbb{R})}; \tag{4.2b}$$

- for the periodic model, $\mathscr{V} = H^1_{\#}(0, L)$ and

$$\mathfrak{a}(u, v) = \frac{1}{2} \int_0^L u'v' - \sum_{I=1}^{M} Z_I\, u(X_I)v(X_I), \tag{4.3a}$$

$$\mathfrak{b}(u, v) = \int_0^L uv =: \langle u, v\rangle_{L^2(0,L)}. \tag{4.3b}$$

Among the possible solutions (u, E) of problem (4.1), we are interested in the pair (u_*, E_*) with the smallest energy E possible. In §3, we prove that this ground state does exist (Theorem 3.1 and Theorem 3.8), is simple (Theorem 3.6 and Theorem 3.13), and can be characterized as the constrained minimization problem

$$E_* = \min_{\substack{u \in \mathscr{V} \\ \mathfrak{b}(u,u)=1}} \mathfrak{E}(u),$$

where $\mathfrak{E}(u) = \mathfrak{a}(u, u)$ represents the energy functional.

In the Galerkin approximation, instead of searching for the minimum over the whole space $\mathscr{V}$, we content ourselves with the minimum over a finite-dimensional subspace $\mathscr{V}_b \subset \mathscr{V}$, which gives rise to

$$E_b = \min_{\substack{u \in \mathscr{V}_b \\ \mathfrak{b}(u,u)=1}} \mathfrak{E}(u).$$

The optimality conditions for this problem lead to the discrete variational formulation: find $(u_b, E_b) \in \mathscr{V}_b \times \mathbb{R}$ such that, for all $v_b \in \mathscr{V}_b$,

$$\mathfrak{a}(u_b, v_b) = E_b\, \mathfrak{b}(u_b, v_b), \tag{4.4a}$$

$$\mathfrak{b}(u_b, u_b) = 1. \tag{4.4b}$$

Here, the subscript b refers to the basis of functions that span the subspace $\mathscr{V}_b$. As will be seen later, b could take various values such as g (in which case $\mathscr{V}_g$ is the space spanned by some Gaussians), or h (in which case $\mathscr{V}_h$ is the space of scaling functions defined on a regular mesh of size h), or h, g (in which case $\mathscr{V}_{h,g}$ is the space spanned by mixed basis consisting of the scaling functions and some periodized Gaussians). For the moment, $\mathscr{V}_b$ is best seen as some abstract subspace of $\mathscr{V}$.

The role of energy

Because the minimization of $\mathfrak{E}$ is performed over a smaller set, the minimal value is larger, i.e., $E_* \leq E_b$. To put it another way, the approximate energy level is always higher than the exact energy level. Pushing further the consequences of the minimization principle, let B be a basis that contains b. Then, $\mathscr{V}_b \subset \mathscr{V}_B$ and we have $E_* \leq E_B \leq E_b$. In other words, enlarging the basis has the effect of lowering the approximate energy and making it closer to the exact energy. Simultaneously, we improve the accuracy of the wave function as well. It is indeed expected that

$$E_b - E_* \simeq \|u_b - u_*\|^2_{\mathscr{V}}$$

for a sequence of subspaces $\mathscr{V}_b$ approaching $\mathscr{V}$ and satisfying some usual properties, the statement of which can be found in [7, 56]. By this equivalence, $E_B - E_* \leq E_b - E_*$ strongly suggests (although this is not a rigorous proof) that $\|u_B - u_*\|_{\mathscr{V}} \leq \|u_b - u_*\|_{\mathscr{V}}$ if $\mathscr{V}_b \subset \mathscr{V}_B$ are members of the same sequence approaching $\mathscr{V}$.

To shed more light on this issue, we recall that the rate of convergence of (u_b, E_b) towards (u_*, E_*) is classically related to the intrinsic quality of the $\mathscr{V}_b$'s in terms of the minimal distance projection

$$\xi_b := \min_{v \in \mathscr{V}_b} \|v - u_*\|_{\mathscr{V}},$$

which does not depend on the parameters involved in $\mathfrak{E}$. The following behavior is typical of linear eigenvalue problems: under suitable assumptions on the sequence $\mathscr{V}_b$, there exist constants C_0, C_1, C_2 such that

$$\|u_b - u_*\|_{\mathscr{V}} \leq C_0 \xi_b, \qquad C_1 \xi_b^2 \leq E_b - E_* \leq C_2 \xi_b^2. \tag{4.5}$$

Again, more details can be found in [7, 56]. In §4.3.2, we will investigate more carefully the proof and the significance of (4.5) for when $\mathscr{V}_b$ is $\mathscr{V}_h$, the space of scaling functions.

The conclusion we wish to draw from the above discussion is that we have a reliable criterion for comparing any two approximate solutions, as well as a good guide for enriching a given basis b by some new functions, which is none other than the energy level: the lower E_b is, the better. This basic tenet is especially useful when we do not know the value of the exact energy E_*.

4.2 Resolution of the infinite model on Gaussian bases

As a first exercice in Galerkin approximation for the 1-D models of §3, we consider the infinite model for the very simple situation of a single nucleus ($M = 1$) with charge $Z_1 = Z > 0$ located at $X_1 = 0$. The model (3.1) boils down to

$$-\frac{1}{2}u'' - Z\delta_0 u = Eu, \tag{4.6a}$$

$$\int_{\mathbb{R}} |u|^2 = 1. \tag{4.6b}$$

The variational formulation of (4.6) involves the bilinear forms

$$\mathfrak{a}(u,v) := \frac{1}{2}\int_{\mathbb{R}} u'v' - Zu(0)v(0), \tag{4.7a}$$

$$\mathfrak{b}(u,v) := \int_{\mathbb{R}} uv, \tag{4.7b}$$

defined on $\mathscr{V} \times \mathscr{V}$, as well as the energy functional

$$\mathfrak{E}(u) = \frac{1}{2}\int_{\mathbb{R}} |u'|^2 - Z|u(0)|^2, \tag{4.8}$$

defined on $\mathscr{V}$, where $\mathscr{V} = H^1(\mathbb{R})$. By virtue of Theorem 3.4, we know that the only solution of (4.6) is

$$u_*(x) = Z^{1/2}\,\mathrm{S}_{Z,0}(x) = Z^{1/2}\exp(-Z|x|)$$
$$E_* = -\frac{1}{2}Z^2$$

We consider the Galerkin approximation of this problem on a basis consisting of Gaussian functions centered at the nucleus position. The real motivation for this is to propose a new way of approximating the normalized Slater $Z^{1/2}\mathrm{S}_{Z,0}$ by a linear combination of Gaussians. This will be elaborated on in §4.2.2. For the moment, let us describe the details of the discrete eigenvalue problem.

4.2.1 Discrete eigenvalue problem

Let $Q \in \mathbb{N}^*$ be the number of Gaussians to be envisaged, and let

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_Q) \in (\mathbb{R}_+^*)^Q$$

be a Q-tuple representing the standard deviations of the Gaussians. For the moment, we do not know how to choose $\boldsymbol{\sigma}$, but this will be discussed in §4.2.2. The subspace on which the Galerkin method will be applied is

$$\mathscr{V}_b = \mathscr{V}_{\boldsymbol{\sigma}} := \mathrm{Span}\{g_{\sigma_q},\ \ 1 \le q \le Q\}, \tag{4.9}$$

where

$$g_\sigma(x) := \frac{1}{\sigma^{1/2}\pi^{1/4}}\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{4.10}$$

is the Gaussians centered at the origin, with standard deviation $\sigma > 0$, and normalized so that

$$\|g_\sigma\|_{L^2(\mathbb{R})} = 1.$$

The Galerkin approximate solution on $\mathscr{V}_{\boldsymbol{\sigma}}$ is designated by $(u_{\boldsymbol{\sigma}}, E_{\boldsymbol{\sigma}})$. Let us decompose the approximate wave function $u_{\boldsymbol{\sigma}}$ into a sum

$$u_{\boldsymbol{\sigma}} = \sum_{j=1}^{Q} \mathsf{u}_j^{\boldsymbol{\sigma}}\, g_{\sigma_j}$$

and encapsulate the set of coefficients $\mathsf{u}_j^{\boldsymbol{\sigma}}$, $1 \le j \le Q$, into the vector

$$\mathbf{u}^{\boldsymbol{\sigma}} = \begin{pmatrix} \mathsf{u}_1^{\boldsymbol{\sigma}} \\ \mathsf{u}_2^{\boldsymbol{\sigma}} \\ \dots \\ \mathsf{u}_Q^{\boldsymbol{\sigma}} \end{pmatrix} \in \mathbb{R}^Q.$$

Proposition 4.1. *The pair $(\mathbf{u}^{\boldsymbol{\sigma}}, E_{\boldsymbol{\sigma}})$ solves the smallest eigenvalue problem*

$$\mathbf{A}^{\boldsymbol{\sigma}}\mathbf{u}^{\boldsymbol{\sigma}} = E_{\boldsymbol{\sigma}}\mathbf{B}^{\boldsymbol{\sigma}}\mathbf{u}^{\boldsymbol{\sigma}}, \tag{4.11a}$$

$$(\mathbf{u}^{\boldsymbol{\sigma}})^T\mathbf{B}^{\boldsymbol{\sigma}}\mathbf{u}^{\boldsymbol{\sigma}} = 1, \tag{4.11b}$$

in which the $Q \times Q$ matrices $\mathbf{A}^{\boldsymbol{\sigma}}$ and $\mathbf{B}^{\boldsymbol{\sigma}}$ are given by

$$\mathsf{A}_{ij}^{\boldsymbol{\sigma}} = \sqrt{\frac{\sigma_i\sigma_j}{2(\sigma_i^2+\sigma_j^2)^3}} - \frac{Z}{\sqrt{\sigma_i\sigma_j\pi}}, \tag{4.12a}$$

$$\mathsf{B}_{ij}^{\boldsymbol{\sigma}} = \sqrt{\frac{2\sigma_i\sigma_j}{\sigma_i^2+\sigma_j^2}}\,, \tag{4.12b}$$

for $(i,j) \in \{1,\dots,Q\}^2$.

Proof. In the discrete variational formulation $\mathfrak{a}(u_{\boldsymbol{\sigma}}, v_{\boldsymbol{\sigma}}) = E_{\boldsymbol{\sigma}}\mathfrak{b}(u_{\boldsymbol{\sigma}}, v_{\boldsymbol{\sigma}})$ for all $v_{\boldsymbol{\sigma}} \in \mathscr{V}_{\boldsymbol{\sigma}}$, we specify $v_{\boldsymbol{\sigma}} = g_{\sigma_i}$ for a fixed $i \in \{1,2,\dots,Q\}$. This yields the i-th equation

$$\sum_{j=1}^{Q} \mathfrak{a}(g_{\sigma_j}, g_{\sigma_i})\mathsf{u}_j^{\boldsymbol{\sigma}} = E_{\boldsymbol{\sigma}} \sum_{j=0}^{Q} \mathfrak{b}(g_{\sigma_j}, g_{\sigma_i})\mathsf{u}_j^{\boldsymbol{\sigma}}.$$

Gathering all of these "row" equations, we obtain (4.11a) with

$$\mathsf{A}_{ij}^{\boldsymbol{\sigma}} = \mathfrak{a}(g_{\sigma_j}, g_{\sigma_i}), \qquad \mathsf{B}_{ij}^{\boldsymbol{\sigma}} = \mathfrak{b}(g_{\sigma_j}, g_{\sigma_i}).$$

The values (4.12) for the above entries result from (4.7) and from the identities

$$\int_{\mathbb{R}} g_{\sigma_i} g_{\sigma_j} = \sqrt{\frac{2\sigma_i\sigma_j}{\sigma_i^2+\sigma_j^2}}\,, \qquad \int_{\mathbb{R}} g'_{\sigma_i} g'_{\sigma_j} = \sqrt{\frac{2\sigma_i\sigma_j}{(\sigma_i^2+\sigma_j^2)^3}}\,, \tag{4.13}$$

which can be checked easily. □

4.2.2 Contracted Gaussians revisited

By construction, we have

$$E_* \leq E_{\boldsymbol{\sigma}} = \mathfrak{E}(u_{\boldsymbol{\sigma}}) = \min_{\substack{u \in \mathscr{V}_{\boldsymbol{\sigma}} \\ \mathfrak{b}(u,u)=1}} \mathfrak{E}(u)$$

for all $\boldsymbol{\sigma} \in (\mathbb{R}_+^*)^Q$. In the sense of energy, $u_{\boldsymbol{\sigma}}$ is the best approximation of u_* over $\mathscr{V}_{\boldsymbol{\sigma}}$. To find the best Q-tuple $\boldsymbol{\sigma}$ possible, we follow the energy minimizing principle recalled in §4.1 and perform an outer minimization problem

$$E_{\boldsymbol{\sigma}^*} = \min_{\boldsymbol{\sigma} \in (\mathbb{R}_+^*)^Q} E_{\boldsymbol{\sigma}} = \min_{\boldsymbol{\sigma} \in (\mathbb{R}_+^*)^Q} \min_{\substack{u \in \mathscr{V}_{\boldsymbol{\sigma}} \\ \mathfrak{b}(u,u)=1}} \mathfrak{E}(u). \tag{4.14}$$

Should the function $\boldsymbol{\sigma} \mapsto E_{\boldsymbol{\sigma}}$ reach a minimum at some $\boldsymbol{\sigma}^* \in (\mathbb{R}_+^*)^Q$, then $u_{\boldsymbol{\sigma}^*}$ would be the best approximation of u_* over all possible choices of $\boldsymbol{\sigma}$ for a given $Q \in \mathbb{N}^*$. Obviously, $u_{\boldsymbol{\sigma}^*}$ belongs to the following category of functions.

Definition 4.1. A centered *Q-G contracted Gaussian* is a linear combination of Q Gaussian primitives

$$\mathrm{CG}(\boldsymbol{\sigma}, \mathbf{v}, \cdot) := \sum_{q=1}^{Q} \mathsf{v}_q g_{\sigma_q}, \tag{4.15}$$

where the standard deviations $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_Q) \in (\mathbb{R}_+^*)^Q$ and the coefficients $\mathbf{v} = (\mathsf{v}_1, \mathsf{v}_2, ..., \mathsf{v}_Q) \in (\mathbb{R}^*)^Q$ are selected in such a way that $\mathrm{CG}(\boldsymbol{\sigma}, \mathbf{v}, \cdot)$ is the "best approximation" of a normalized Slater function $Z^{1/2}\,\mathrm{S}_{Z,0}$ in some sense to be precised.

Note that this Definition does not require $\|\mathrm{CG}(\boldsymbol{\sigma}, \mathbf{v}, \cdot)\|_{L^2}$ to be equal to 1. Our approximate solution

$$u_{\boldsymbol{\sigma}^*} = \mathrm{CG}(\boldsymbol{\sigma}^*, \mathbf{u}^{\boldsymbol{\sigma}^*}, \cdot)$$

does have unit L^2-norm. It is optimal in the sense of energy (4.14). Traditionally, contracted Gaussians are designed so as to be the "best" approximation of a Slater in one of these two senses:

1. *L^2-projection.* For a given number Q of primitives, the standard deviations $\boldsymbol{\sigma}^*$ are exact or approximate solutions of the least-squares problem

$$\min_{\boldsymbol{\sigma} \in (\mathbb{R}_+^*)^Q} \min_{v_{\boldsymbol{\sigma}} \in \mathscr{V}_{\boldsymbol{\sigma}}} \|v_{\boldsymbol{\sigma}} - u_*\|_{L^2}^2,$$

   as advocated by Hehre *et al.* [70] and Huzinaga [76]. Further assumptions on $\boldsymbol{\sigma}^*$ can be imposed in order to simplify the minimization problem. For instance, Longo [95] proposed to look for $\boldsymbol{\sigma}^*$ whose components make up a geometric series and whose geometric mean is a fixed value. The numerical order of convergence he measured is $\|u_{\boldsymbol{\sigma}^*}^Q - u_*\|_{L^2} \sim 1.1 \exp(-1.9\sqrt{Q})$.

2. *Numerical quadrature.* The integral transform [123]

$$\exp(-Z|x|) = \frac{Z}{2\sqrt{\pi}} \int_0^{+\infty} s^{-3/2} \exp(-Z^2/4s) \exp(-sx^2)\, \mathrm{d}s$$

expresses the Slater as a continuous sum of Gaussians. A numerical integration using a finite number Q of quadrature points would then give an approximate expansion of the Slater in a Gaussian basis. By carefully choosing the quadrature points, Kutzelnigg [86] was able to derive the *even-tempered basis set* formerly introduced by Feller and Ruedenberg [53] in an empirical way. The asymptotic behavior of the error is given by the energy difference $\mathfrak{E}(u^Q_{\boldsymbol{\sigma}^*}) - \mathfrak{E}(u_*) \sim \pi(3Q)^{3/2}\exp(-\pi\sqrt{3Q})$ in 3-D. Other quadrature rules are studied by Gomes and Custodio [66]. A systematic investigation of various families of expansion of a Slater into Gaussians can be found in the papers of Klahn and Bingel [81, 82].

In the literature, it seems that the energy minimization paradigm (4.14) has never been explored for constructing contracted Gaussians. The reason for this is probably that the evaluation of the objective function $\boldsymbol{\sigma} \mapsto E_{\boldsymbol{\sigma}}$ requires the computation of an eigenvalue and is therefore costly.

Like other minimization criteria, the energy minimization principle has a symmetry: it is invariant by rescaling, the meaning of which will be clarified in the upcoming Theorem. The proof of this seemingly obvious result crucially relies on a homogeneity property of the energy functional with respect to a scaling operator.

Theorem 4.1. *Let $\lambda > 0$. If $\boldsymbol{\sigma}^*$ solves* (4.14) *for Z, then $\lambda^{-1}\boldsymbol{\sigma}^*$ solves* (4.14) *for λZ. If $\mathrm{CG}(\boldsymbol{\sigma}^*, \mathbf{u}^{\boldsymbol{\sigma}^*}, \cdot)$ is the optimal contracted Gaussian for Z, then $\mathrm{CG}(\lambda^{-1}\boldsymbol{\sigma}^*, \mathbf{u}^{\boldsymbol{\sigma}^*}, \cdot)$ is the optimal contracted Gaussian for λZ.*

Proof. To emphasis the dependence on the charge Z, let us write the energy functional (4.8) as $\mathfrak{E}^Z$. Consider the scaling operator $R_\lambda : u \in H^1(\mathbb{R}) \mapsto R_\lambda u \in H^1(\mathbb{R})$ defined as

$$(R_\lambda u)(y) = \lambda^{1/2} u(\lambda y), \qquad y \in \mathbb{R}. \tag{4.16}$$

It is easily verified that $\|R_\lambda u\|_{L^2} = \|u\|_{L^2}$, $R_\lambda^{-1} = R_{1/\lambda}$, and

$$\begin{aligned}\mathfrak{E}^{\lambda Z}(R_\lambda u) &= \frac{1}{2}\int_R |(R_\lambda u)'|^2 - \lambda Z |(R_\lambda u)(0)|^2 \\ &= \frac{1}{2}\int_{\mathbb{R}} \lambda^3 |u'(\lambda y)|^2 dy - \lambda^2 Z |u(0)|^2 = \lambda^2 \mathfrak{E}^Z(u)\end{aligned}$$

after the change of variable $x = \lambda y$. Therefore, $\mathfrak{E}^{\lambda Z}(u) = \lambda^2 \mathfrak{E}^Z(R_{1/\lambda} u)$. By definition of the Gaussians, it is plain that if $u \in \mathscr{V}_{\boldsymbol{\sigma}}$, then $R_{1/\lambda} u \in \mathscr{V}_{\lambda\boldsymbol{\sigma}}$. As a result,

$$E^{\lambda Z}_\sigma = \min_{\substack{u \in \mathscr{V}_{\boldsymbol{\sigma}} \\ \mathfrak{b}(u,u)=1}} \mathfrak{E}^{\lambda Z}(u) = \min_{\substack{u \in \mathscr{V}_{\boldsymbol{\sigma}} \\ \mathfrak{b}(u,u)=1}} \lambda^2 \mathfrak{E}^Z(R_{1/\lambda} u) = \min_{\substack{v \in \mathscr{V}_{\lambda\boldsymbol{\sigma}} \\ \mathfrak{b}(v,v)=1}} \lambda^2 \mathfrak{E}^Z(v) = \lambda^2 E^Z_{\lambda\sigma}.$$

If $\boldsymbol{\sigma} \mapsto E^Z_{\boldsymbol{\sigma}}$ achieves its minimum at $\boldsymbol{\sigma}^*$, then $\boldsymbol{\sigma} \mapsto E^Z_{\lambda\boldsymbol{\sigma}}$ achieves its minimum at $\lambda^{-1}\boldsymbol{\sigma}^*$. Putting $\boldsymbol{\sigma} = \lambda^{-1}\boldsymbol{\sigma}^*$ in the second equality of the previous line and scrutinizing the minimal argument, we end up with $R_{1/\lambda} u^{\lambda Z}_{\lambda^{-1}\boldsymbol{\sigma}^*} = u^Z_{\boldsymbol{\sigma}^*}$. From this, we infer that

$$u^{\lambda Z}_{\lambda^{-1}\boldsymbol{\sigma}^*} = R_\lambda u^Z_{\boldsymbol{\sigma}^*} = R_\lambda \mathrm{CG}(\boldsymbol{\sigma}^*, \mathbf{u}^{\boldsymbol{\sigma}^*}, \cdot) = \mathrm{CG}(\lambda^{-1}\boldsymbol{\sigma}^*, \mathbf{u}^{\boldsymbol{\sigma}^*}, \cdot),$$

the last equality being due to $R_\lambda g_\sigma = g_{\sigma/\lambda}$. □

The characteristic length

$$\Lambda = Z^{-1}, \tag{4.17}$$

associated with the charge Z, measures the size of its "domain of influence". Application of Theorem 4.1 to $\lambda = \Lambda$ leads to the very important facts that

$$\begin{gathered}\boldsymbol{\sigma}^* \text{ is proportional to } \Lambda,\\ \mathbf{u}^{\boldsymbol{\sigma}^*} \text{ is independent of } \Lambda.\end{gathered}$$

Indeed, $\boldsymbol{\sigma}^*/\Lambda =: \boldsymbol{\tau}^*$ solves (4.14) for a charge equal to 1.

As was said in §1.3.2, contracted Gaussians are widely used by computational chemists, insofar as the CGTO (Contracted Gaussian Type Orbitals) bases are a good compromise between STO (Slater Type Orbitals) and GTO (Gaussian Type Orbitals). Let us describe how a CGTO basis looks like for the multi-delta model (3.1). Assume that for each $Q \in \mathbb{N}^*$, we have pre-computed the optimal contracted Gaussian

$$\mathrm{CG}(\boldsymbol{\tau}^*(Q), \mathbf{v}^*(Q), \cdot)$$

by the minimization principle (4.14) for the single-delta energy with charge $Z = 1$. Then, a CGTO basis for our multi-delta problem could be taken to be

$$\Big\{\mathrm{CG}(\Lambda_I \boldsymbol{\tau}^*(Q_I), \mathbf{v}^*(Q_I), \cdot - X_I),\ \ 1 \le I \le M\Big\},$$

where Q_I, the number of primitives at the I-th nucleus, remains to be tuned by the user.

4.2.3 Analytical and numerical results

Our task is now to determine the contracted Gaussians optimal in the sense of (4.14) for each $Q \in \mathbb{N}^*$. When $Q = 1$, the exact solution is given by an analytical formula.

Proposition 4.2. *For $Q = 1$, writing $\boldsymbol{\sigma} = \sigma$ and $\mathbf{u}^{\boldsymbol{\sigma}} = \mathsf{u}$,*

- *the solution $(u_{\boldsymbol{\sigma}}, E_{\boldsymbol{\sigma}}) \in \mathscr{V}_{\boldsymbol{\sigma}} \times \mathbb{R}$ of the inner minimization problem of (4.14) is*

$$u_{\boldsymbol{\sigma}} = g_\sigma, \tag{4.18a}$$

$$E_{\boldsymbol{\sigma}} = \frac{1}{4\sigma^2} - \frac{Z}{\sigma\sqrt{\pi}}; \tag{4.18b}$$

- *the solution $(\boldsymbol{\sigma}^*, \mathbf{u}^{\boldsymbol{\sigma}^*}) \in \mathbb{R} \times \mathbb{R}$ of the outer minimization problem of (4.14) is*

$$\sigma^* = \frac{\sqrt{\pi}}{2}\Lambda, \tag{4.19a}$$

$$\mathsf{u}^* = 1. \tag{4.19b}$$

Proof. Since $\mathscr{V}_{\boldsymbol{\sigma}} = \mathbb{R} g_\sigma$, we necessarily have $u_{\boldsymbol{\sigma}} = \alpha g_\sigma$ for some $\alpha \in \mathbb{R}$. To ensure $\|u_{\boldsymbol{\sigma}}\|_{L^2} = 1$, α must be ± 1. We consider $\alpha = 1$, the other choice being similar. The minimum energy is then

$$\begin{aligned} E_\sigma = \mathfrak{E}(u_{\boldsymbol{\sigma}}) &= \frac{1}{2}\int_{\mathbb{R}} |g_\sigma'|^2 - Z|g_\sigma(0)|^2 \\ &= \frac{1}{2}\int_{\mathbb{R}} \frac{x^2}{\sigma^5\sqrt{\pi}} \exp\Big(-\frac{x^2}{\sigma^2}\Big)\,\mathrm{d}x - Z\frac{1}{\sigma\sqrt{\pi}} = \frac{1}{4\sigma^2} - \frac{Z}{\sigma\sqrt{\pi}}, \end{aligned}$$

which proves (4.18). The derivative of E_σ with respect to σ reads

$$\frac{\mathrm{d}E_\sigma}{\mathrm{d}\sigma}(\sigma) = -\frac{1}{2\sigma^3} + \frac{Z}{\sigma^2\sqrt{\pi}}$$

and the only point at which it can be cancelled is $\sigma^* = \sqrt{\pi}/(2Z) = \Lambda\sqrt{\pi}/2$. At this point,

$$\frac{\mathrm{d}^2E_\sigma}{\mathrm{d}\sigma^2}(\sigma^*) = \frac{3}{2(\sigma^*)^4} - \frac{2Z}{(\sigma^*)^3\sqrt{\pi}} = \frac{8Z^4}{\pi^2} > 0,$$

which implies that $\sigma \mapsto E_\sigma$ reaches a local minimum at σ^*. A closer study of this function shows that this is also a global minimum over $\sigma \in \mathbb{R}_+^*$, which proves (4.19). □

The length $\sigma^* = (\sqrt{\pi}/2)\Lambda$ in (4.19a) is called ***reference standard deviation***. When $Q = 2$, only the inner minimization problem of (4.14) can be solved by a finite sequence of explicit formulae. The outer minimization problem of (4.14) has to be solved numerically.

Proposition 4.3. *For $Q = 2$, writing $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ and $\mathbf{u}^{\boldsymbol{\sigma}} = (\mathsf{u}_1, \mathsf{u}_2)^T$,*

- *the solution $(u_{\boldsymbol{\sigma}}, E_{\boldsymbol{\sigma}}) \in \mathscr{V}_{\boldsymbol{\sigma}} \times \mathbb{R}$ of the inner minimization problem of (4.14) is*

$$u_{\boldsymbol{\sigma}} = \mathsf{u}_1 g_{\sigma_1} + \mathsf{u}_2 g_{\sigma_2} \tag{4.20a}$$

$$E_{\boldsymbol{\sigma}} = \frac{1}{2}(\mathtt{C} - \sqrt{\mathtt{C}^2 - 4\mathtt{D}}) \tag{4.20b}$$

with

$$\mathbf{u}^{\boldsymbol{\sigma}} = \frac{1}{\{(\mathbf{v}^{\boldsymbol{\sigma}})^T \mathbf{B}^{\boldsymbol{\sigma}} \mathbf{v}^{\boldsymbol{\sigma}}\}^{1/2}} \mathbf{v}^{\boldsymbol{\sigma}}, \qquad \mathbf{v}^{\boldsymbol{\sigma}} = \left(E_{\boldsymbol{\sigma}}\mathsf{B}_{12}^{\boldsymbol{\sigma}} - \mathsf{A}_{12}^{\boldsymbol{\sigma}},\, \mathsf{A}_{11}^{\boldsymbol{\sigma}} - E_{\boldsymbol{\sigma}}\right)^T, \tag{4.21a}$$

$$\mathtt{C} = \frac{\mathsf{A}_{12}^{\boldsymbol{\sigma}} + \mathsf{A}_{22}^{\boldsymbol{\sigma}} - 2\mathsf{A}_{12}^{\boldsymbol{\sigma}}\mathsf{B}_{12}^{\boldsymbol{\sigma}}}{1 - (\mathsf{B}_{12}^{\boldsymbol{\sigma}})^2}, \qquad \mathtt{D} = \frac{\mathsf{A}_{11}^{\boldsymbol{\sigma}}\mathsf{A}_{22}^{\boldsymbol{\sigma}} - (\mathsf{A}_{12}^{\boldsymbol{\sigma}})^2}{1 - (\mathsf{B}_{12}^{\boldsymbol{\sigma}})^2}, \tag{4.21b}$$

where the entries of $\mathbf{A}^{\boldsymbol{\sigma}}$ and $\mathbf{B}^{\boldsymbol{\sigma}}$ are given in (4.12).

- *the solution $(\boldsymbol{\sigma}^*, \mathbf{u}^{\boldsymbol{\sigma}^*}) \in (\mathbb{R}_+^*)^2 \times \mathbb{R}^2$ of the outer minimization problem of (4.14) is*

$$(\sigma_1^*, \sigma_2^*) \approx (0.202009\Lambda,\, 1.013952\Lambda), \tag{4.22a}$$

$$(\mathsf{u}_1^*, \mathsf{u}_2^*) \approx (0.891491,\, 0.162129). \tag{4.22b}$$

Proof. As a solution of the eigenvalue problem (4.11), $E_{\boldsymbol{\sigma}}$ is a root of the second degree polynomial

$$\begin{aligned}\wp(E_{\boldsymbol{\sigma}}) &= \det(\mathbf{A}^{\boldsymbol{\sigma}} - E_{\boldsymbol{\sigma}}\mathbf{B}^{\boldsymbol{\sigma}}) \\ &= (\mathsf{A}_{11}^{\boldsymbol{\sigma}} - E_{\boldsymbol{\sigma}})(\mathsf{A}_{22}^{\boldsymbol{\sigma}} - E_{\boldsymbol{\sigma}}) - (\mathsf{A}_{12}^{\boldsymbol{\sigma}} - E_{\boldsymbol{\sigma}}\mathsf{B}_{12}^{\boldsymbol{\sigma}})^2 \\ &= (1 - (\mathsf{B}_{12}^{\boldsymbol{\sigma}})^2)E_{\boldsymbol{\sigma}}^2 - (\mathsf{A}_{11}^{\boldsymbol{\sigma}} + \mathsf{A}_{22}^{\boldsymbol{\sigma}} - 2\mathsf{A}_{12}^{\boldsymbol{\sigma}}\mathsf{B}_{12}^{\boldsymbol{\sigma}})E_{\boldsymbol{\sigma}} + (\mathsf{A}_{11}^{\boldsymbol{\sigma}}\mathsf{A}_{22}^{\boldsymbol{\sigma}} - (\mathsf{A}_{12}^{\boldsymbol{\sigma}})^2).\end{aligned}$$

Dividing $\wp$ by $1 - (\mathsf{B}_{12}^{\boldsymbol{\sigma}})^2$ and solving for the smaller zero, we obtain (4.20b), (4.21b). Searching for the corresponding eigenvector, we end up with (4.21a). As far as the outer minimization problem is concerned, it is solved numerically by Matlab. □

Q	$\boldsymbol{\tau}^* = \boldsymbol{\sigma}^*/\Lambda$	$\mathbf{v}^* = \mathbf{u}^{\boldsymbol{\sigma}^*}$	`err`
1	$\sqrt{\pi}/2$	1	0.3634
2	(0.202009, 1.013952)	(0.891491, 0.162129)	0.1240
3	(0.063854, 0.323840, 1.156697)	(0.030546, 0.253729, 0.788039)	0.0473

Table 4.1: Optimal parameters for contracted Gaussians in the sense of energy, with the corresponding relative energy error defined in (4.23).

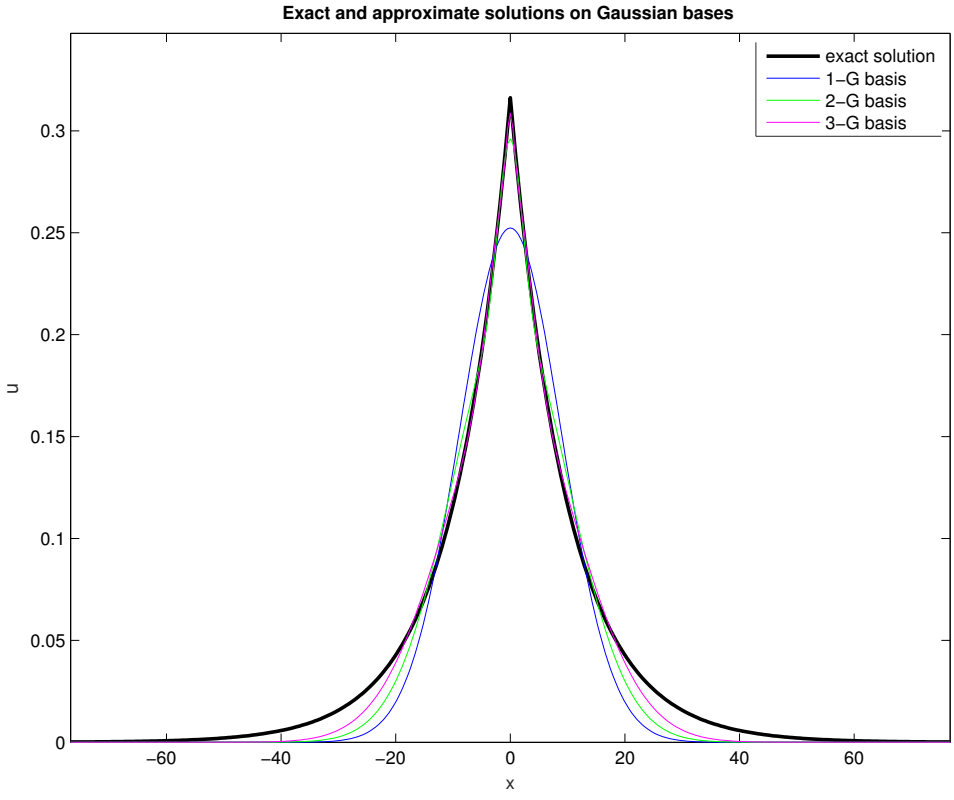

Figure 4.1: The Slater function and the first three contracted Gaussians optimized by the energy minimizing principle.

When $Q = 3$, the objective function $\boldsymbol{\sigma} \mapsto E_{\boldsymbol{\sigma}}$ can no longer be implemented explicitly. We must resort to an entirely numerical procedure. The overall computations are long, but the CPU time is still reasonable. The parameters for the optimal contracted Gaussian are reported in Table 4.1, where we recapitulate the standard deviations $\boldsymbol{\sigma}^*$ and the coefficients $\mathbf{u}^{\boldsymbol{\sigma}^*}$ for Q ranging from 1 to 3. The third column provides the relative errors

$$\texttt{err} = \frac{E_{\boldsymbol{\sigma}^*} - E_*}{|E_*|}. \tag{4.23}$$

For $Q = 1$, this relative error is quite large (36.35%). For $Q = 2$ and 3, it becomes more acceptable (12.40% and 4.73%). The contracted Gaussians $u_{\boldsymbol{\sigma}^*}$ are graphically displayed in Figure 4.1. For $Q = 2$ and 3, we see that the cusp at $X = 0$ is rather well approximated. In the rest of the domain, $u_{\boldsymbol{\sigma}^*}$ goes to 0 too fast to match u_*.

When $Q \geq 4$, the computations are extremely long and ordinary optimization does not seem feasible. This is why we leave the topic aside and will resume it in §6, with further simplifications.

4.3 Resolution of the periodic model on scaling function bases

We turn to the Galerkin approximation of the periodic model (3.57)–(3.58) in a basis associated with a mesh. Our objective is to become familiar with the specificities of wavelets before going to a mixed basis. For simplicity, we consider only one level in the multiresolution, consisting of scaling functions. In §4.3.1, we describe the discrete eigenvalue problem for an arbitrary number $M \geq 1$ of nuclei. The method is then analyzed in §4.3.2, where we establish a priori error estimates. Finally, numerical results are provided in §4.3.3 for $M = 1$, 2 and 3.

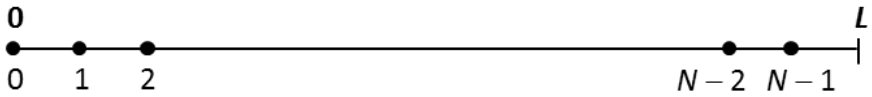

Figure 4.2: Nodes of the regular mesh of size $h = L/N$ for the periodic domain $[0, L]$.

4.3.1 Discrete eigenvalue problem

As depicted in Figure 4.2, the domain $[0, L]$ is divided into N equal intervals of length $h = L/N$, where $N = 2^{\jmath}$ is a power of 2, with a fixed $\jmath \in \mathbb{N}$. The nodes, numbered by $x_0, x_1, \ldots, x_{N-1}$, have abscissae

$$x_i = ih = i2^{-\jmath}L = \frac{iL}{N}, \qquad 0 \leq i \leq N - 1.$$

Because of periodicity, there is no degree of freedom at $x = L$, that is actually identified with $x = 0$.

Let ϕ be a Daubechies scaling function of order $\mathsf{M} \geq 3$, whose support is $[0, 2\mathsf{M}-1]$. The requirement $\mathsf{M} \geq 3$ is aimed at ensuring that $\phi \in H^1(\mathbb{R})$, so that the Hamiltonian matrix is well-defined (see Proposition 2.9 and Proposition 4.4). The periodization procedure described in Definition 2.4 supplies us with $N = 2^{\mathsf{J}}$ scaling functions $\widetilde{\phi}_{\mathsf{J},i}$, $0 \leq i \leq N-1$ on the interval $[0,1]$. To carry these over the interval $[0, L]$, let us consider

$$\widetilde{\chi}_i^h = R_{1/L}\widetilde{\phi}_{\mathsf{J},i}, \tag{4.24}$$

where R_λ stands for the scaling operator defined in (4.16). In other words,

$$\widetilde{\chi}_i^h(x) = \frac{1}{\sqrt{L}}\,\widetilde{\phi}_{\mathsf{J},i}\Big(\frac{x}{L}\Big)$$

for all $x \in [0, L]$. In view of (2.73), this can also be expressed as

$$\widetilde{\chi}_i^h(x) = \frac{1}{\sqrt{L}} 2^{\mathsf{J}/2} \sum_{k\in\mathbb{Z}} \phi\Big(2^{\mathsf{J}}\Big(\frac{x}{L}+k\Big) - i\Big) = \frac{1}{\sqrt{h}} \sum_{k\in\mathbb{Z}} \phi\Big(\frac{x - x_i + kL}{h}\Big).$$

Set

$$\mathscr{V}_h = \mathscr{V}_h := \mathrm{Span}\{\widetilde{\chi}_i^h, \quad 0 \leq i \leq N-1\}. \tag{4.25}$$

The Galerkin approximate solution on $\mathscr{V}_h$ is designated by (u_h, E_h). Let us decompose the approximate wave function u_h into a sum

$$u_h = \sum_{j=0}^{N-1} \mathsf{u}_j^h\, \widetilde{\chi}_j^h$$

and encapsulate the set of coefficients u_j^h, $0 \leq j \leq N-1$, into the vector

$$\mathbf{u}^h = \begin{pmatrix} \mathsf{u}_0^h \\ \mathsf{u}_1^h \\ \cdots \\ \mathsf{u}_{N-1}^h \end{pmatrix} \in \mathbb{R}^N.$$

Proposition 4.4. *If $N \geq 4\mathsf{M}-2$, then the pair $(\mathbf{u}^h, E_h) \in \mathbb{R}^N \times \mathbb{R}$ solves the smallest eigenvalue problem*

$$\mathbf{A}^h\mathbf{u}^h = E_h\mathbf{B}^h\mathbf{u}^h, \tag{4.26a}$$

$$(\mathbf{u}^h)^T\mathbf{B}^h\mathbf{u}^h = 1, \tag{4.26b}$$

in which the $N \times N$ matrices $\mathbf{A}^h$ and $\mathbf{B}^h$ are given by

$$\mathsf{A}_{ij}^h = \frac{1}{2h^2}\, a_{|i-j|^\sim} - \sum_{J=1}^{M} Z_J \widetilde{\chi}_i^h(X_J)\widetilde{\chi}_j^h(X_J), \tag{4.27a}$$

$$\mathsf{B}_{ij}^h = \delta_{ij}, \tag{4.27b}$$

for $(i,j) \in \{0,1,\ldots,N-1\}^2$, where δ_{ij} is the Kronecker symbol, $|i-j|^\sim$ the periodized distance of Definition 2.5 and a_k is the k-th connection coefficient of (2.59).

PROOF. In the discrete variational formulation $\mathfrak{a}(u_h, v_h) = E_h \mathfrak{b}(u_h, v_h)$ for all $v_h \in \mathscr{V}_h$, we specify $v_h = \widetilde{\chi}_i^h$ for a fixed $i \in \{0, 1, \ldots, N-1\}$. This yields the i-th equation

$$\sum_{j=0}^{N-1} \mathfrak{a}(\widetilde{\chi}_j^h, \widetilde{\chi}_i^h) \mathsf{u}_j^h = E_h \sum_{j=0}^{N-1} \mathfrak{b}(\widetilde{\chi}_j^h, \widetilde{\chi}_i^h) \mathsf{u}_j^h$$

Gathering all of these "row" equations, we obtain (4.26a) with

$$\mathsf{A}_{ij}^h = \mathfrak{a}(\widetilde{\chi}_j^h, \widetilde{\chi}_i^h), \qquad \mathsf{B}_{ij}^h = \mathfrak{b}(\widetilde{\chi}_j^h, \widetilde{\chi}_i^h).$$

The values (4.27) for the above entries result from

$$\int_0^L \widetilde{\chi}_i^h \, \widetilde{\chi}_j^h = \int_0^L (R_{1/L} \widetilde{\phi}_{\mathsf{J},i}) \, (R_{1/L} \widetilde{\phi}_{\mathsf{J},j}) = \int_0^1 \widetilde{\phi}_{\mathsf{J},i} \, \widetilde{\phi}_{\mathsf{J},j} = \delta_{i,j}$$

(see Theorem 2.6 for the orthonormality of the $\widetilde{\phi}_{\mathsf{J},i}$'s) and from

$$\int_0^L (\widetilde{\chi}_i^h)' \, (\widetilde{\chi}_j^h)' = \frac{1}{L^2} \int_0^L (R_{1/L} \widetilde{\phi}'_{\mathsf{J},i}) \, (R_{1/L} \widetilde{\phi}'_{\mathsf{J},j}) = \frac{1}{L^2} \int_0^1 \widetilde{\phi}'_{\mathsf{J},i} \, \widetilde{\phi}'_{\mathsf{J},j} = \frac{1}{L^2} \widetilde{a}_{i,j}^{\mathsf{J}} = \frac{2^{2\mathsf{J}}}{L^2} a_{|i-j|^\sim}$$

(see Proposition 2.13 for the last equality). □

It is common usage to call $\mathbf{A}^h$ and $\mathbf{B}^h$ respectively Hamiltonian matrix and mass matrix. From now on, we always assume $N \geq 4\mathsf{M} - 2 \geq 10$. In view of the specific form of $\mathfrak{a}(\cdot, \cdot)$, the Hamiltonian matrix can be further decomposed into

$$\mathbf{A}^h = \mathbf{T}^h + \mathbf{V}^h, \tag{4.28}$$

where

- $\mathbf{T}^h$ is the rigidity matrix associated with the ***kinetic*** part, which comes from $\int_0^L u'v'$ in the variational formulation;
- $\mathbf{V}^h$ the matrix of the delta ***potential*** part, which comes from $-\sum_{J=1}^M Z_J \, u(X_J) v(X_J)$.

It turns out that $\mathbf{T}^h$ is a symmetric circulant matrix, each line and column of which has $4\mathsf{M} - 3$ nonzero elements. When N grows, $\mathbf{T}^h$ becomes a sparse (band diagonal) matrix. Figure 4.3 shows $\mathbf{T}^h$ for $\mathsf{M} = 3$ and $N = 16$. On the other hand, as far as the potential matrix is concerned, it can be factorized as

$$\mathbf{V}^h = -\mathbf{\Phi}^h \mathbf{Z} (\mathbf{\Phi}^h)^T \tag{4.29}$$

where $\mathbf{\Phi}^h$ denotes the $N \times M$ matrix whose entries are $\Phi_{iJ}^h = \widetilde{\chi}_i^h(X_J)$ for $(i, J) \in \{0, \ldots, N-1\} \times \{1, \ldots, M\}$, and $\mathbf{Z} = \mathrm{Diag}(Z_1, Z_2, \ldots, Z_M)$.

EXAMPLE 4.1. Consider the single-delta potential case $M = 1$ with a nucleus of charge $Z_1 = Z$ located at the middle point $X_1 = L/2$. If the number of nodes N is even, that makes this middle point a node. The matrix $\mathbf{V}^h$ is then a $N \times N$ sparse matrix with only a $(2\mathsf{M} - 2) \times (2\mathsf{M} - 2)$ block $-(Z/h) \boldsymbol{\phi} \boldsymbol{\phi}^{\boldsymbol{T}}$ at the center, where the row vector $\boldsymbol{\phi}^T = (\phi(2\mathsf{M} - 2), \phi(2\mathsf{M} - 3), \ldots, \phi(2), \phi(1))$ contains values of the Daubechies scaling function ϕ at integer points in reverse order. Figure 4.4 shows $\mathbf{V}^h$ for $\mathsf{M} = 3$ and $N = 16$.

$$\frac{1}{2h^2}\begin{pmatrix}
a_0 & a_1 & a_2 & a_3 & a_4 & & & & & & & a_4 & a_3 & a_2 & a_1 \\
a_1 & a_0 & a_1 & a_2 & a_3 & a_4 & & & & & & & a_4 & a_3 & a_2 \\
a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & a_4 & & & & & & & a_4 & a_3 \\
a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & a_4 & & & & & & & a_4 \\
a_4 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & a_4 & & & & & & \\
 & a_4 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & a_4 & & & & & \\
 & & a_4 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & a_4 & & & & \\
 & & & a_4 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & a_4 & & & \\
 & & & & a_4 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & a_4 & & \\
 & & & & & a_4 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & a_4 & \\
 & & & & & & a_4 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & a_4 \\
a_4 & & & & & & & a_4 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 \\
a_3 & a_4 & & & & & & & a_4 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 \\
a_2 & a_3 & a_4 & & & & & & & a_4 & a_3 & a_2 & a_1 & a_0 & a_1 \\
a_1 & a_2 & a_3 & a_4 & & & & & & & a_4 & a_3 & a_2 & a_1 & a_0
\end{pmatrix}$$

Figure 4.3: The kinetic matrix $\mathbf{T}^h$ for $\mathsf{M} = 3$ (number of vanishing moments) and $N = 16$ (number of cells).

$$-\frac{Z}{h}\begin{pmatrix}
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & &
\end{pmatrix}$$

$$\begin{array}{cccc}
|\phi_4|^2 & \phi_4\phi_3 & \phi_4\phi_2 & \phi_4\phi_1 \\
\phi_3\phi_4 & |\phi_3|^2 & \phi_3\phi_2 & \phi_3\phi_1 \\
\phi_2\phi_4 & \phi_2\phi_3 & |\phi_2|^2 & \phi_2\phi_1 \\
\phi_1\phi_4 & \phi_1\phi_3 & \phi_1\phi_2 & |\phi_1|^2
\end{array}$$

Figure 4.4: The potential matrix $\mathbf{V}^h$ for $\mathsf{M} = 3$ (number of vanishing moments) and $N = 16$ (number of cells) assuming there is only one nucleus located at the middle of the domain, with the notation $\phi_\ell = \phi(\ell)$.

4.3.2 A priori error estimate

The distinctive feature of the subspace $\mathscr{V}_h$ generated by scaling functions is that, by Theorem 2.7, we have an *a priori* estimate of the best approximation error. This knowledge, in turn, enables us to derive *a priori* error estimates between (u_h, E_h) and (u_*, E_*). The central result of this section is Theorem 4.2, which roughly tells us that when $h \to 0$, $\|u_h - u_*\|_{H^1}$ behaves as $h^{1/2}$ and $E_h - E_*$ behaves as h for all orders $\mathsf{M} \geq 3$ of the scaling function ϕ. At first, this seems to contradict the intuition that the higher M is, the more polynomial exactness is ensured and the better the approximate solution should be. In reality, this "saturation" phenomenon stems from the low regularity of the exact eigenfunction u_* (Corollary 3.5).

Lemma 4.1. *Let $\epsilon > 0$ be an arbitrarily small real number. There exists $\Gamma_{\mathsf{M},\epsilon} > 0$ (dependent on ϵ and M but not on h) such that*

$$\min_{v_h \in \mathscr{V}_h} \|u_* - v_h\|_{H^1} \leq \Gamma_{\mathsf{M},\epsilon} \|u_*\|_{H^{3/2-\epsilon}} h^{1/2-\epsilon} \tag{4.30}$$

for all h close enough to 0.

Proof. Applying Theorem 2.7 with $s = 1$ to the 1-periodic function $u = R_L u_*$, where R_λ is the scaling operator introduced in (4.16), we obtain

$$\|R_L u_* - \widetilde{P}_\mathsf{J} R_L u_*\|_{H^1_\#(0,1)} \leq \widetilde{\Gamma}_{\mathsf{M},1,t} \|R_L u_*\|_{H^t_\#(0,1)} 2^{-\mathsf{J}(t-1)}$$

for $1 < t < \mathsf{M}$ if $R_L u_* \in H^t_\#(0,1)$. From Corollary 3.5, $u_* \in H^{3/2-\epsilon}_\#(0,L)$. This low regurality prevents us from taking t larger than $3/2 - \epsilon$ even when M is large. From definition (3.88) of $H^s_\#(0,L)$, it is easy to check that

$$\min\{1, \lambda^s\} \|v\|_{H^s_\#(0,L)} \leq \|R_\lambda v\|_{H^s_\#(0,L/\lambda)} \leq \max\{1, \lambda^s\} \|v\|_{H^s_\#(0,L)} \tag{4.31}$$

for all $s > 0$ and all L-periodic function $v \in H^s_\#(0,L)$. Hence, $R_L u_* \in H^{3/2-\epsilon}_\#(0,1)$ and

$$\|R_L(u_* - R_{1/L} \widetilde{P}_\mathsf{J} R_L u_*)\|_{H^1_\#(0,1)} \leq \widetilde{\Gamma}_{\mathsf{M},1,3/2-\epsilon} \|R_L u_*\|_{H^{3/2-\epsilon}_\#(0,1)} 2^{-\mathsf{J}(1/2-\epsilon)}.$$

Invoking (4.31) twice again and replacing $2^{-\mathsf{J}}$ by h/L, we end up with

$$\|u_* - R_{1/L} \widetilde{P}_\mathsf{J} R_L u_*\|_{H^1_\#(0,L)} \leq \widetilde{\Gamma}_{\mathsf{M},1,3/2-\epsilon} \frac{\max\{1, L^{3/2-\epsilon}\}}{\min\{1, L\} L^{1/2-\epsilon}} \|u_*\|_{H^{3/2-\epsilon}_\#(0,L)} h^{1/2-\epsilon}$$

From the observation that $R_{1/L} \widetilde{P}_\mathsf{J} R_L u_* \in \mathscr{V}_h$, we finally deduce (4.30). □

For the sake of clarity, we divide the upcoming exposition into two parts. In the first part, we take it for granted that some abstract assumptions, called "Standard Hypotheses," are satisfied. These allow us to establish Theorem 4.2. In the second part, we prove that the "Standard Hypotheses" are indeed fulfilled for our concrete problem. Our calculations are inspired from Chakir's thesis [26] but our presentation is a little different.

Standard Hypotheses

1. *There exists $K > 0$ (independent of h) such that for all $(v, w) \in \mathscr{V}^2$,*

$$|\mathfrak{a}(v, w) - E_* \mathfrak{b}(v, w)| \leq K \|v\|_{H^1} \|w\|_{H^1}. \tag{4.32}$$

2. *There exists $\beta > 0$ (independent of h) such that for any $v \in (u_*)^\perp$ (orthogonality in the L^2-sense) in $\mathscr{V}$,*
$$\beta \|v\|_{L^2}^2 \leq \mathfrak{a}(v,v) - E_* \mathfrak{b}(v,v). \tag{4.33}$$

3. *There exists $\gamma > 0$ (independent of h) such that for $e = u_h - u_*$,*
$$\gamma \|e\|_{H^1}^2 \leq \mathfrak{a}(e,e) - E_* \mathfrak{b}(e,e). \tag{4.34}$$

The second Standard Hypothesis (4.33) is not used in the first stage, where only (4.32) and (4.34) are required for proving Theorem 4.2. It appears as an intermediate step for proving the third Standard Hypothesis (4.34) in the second stage. However, we have deliberately conferred the status of Standard Hypothesis on (4.33) in order to highlight the L^2-coercivity property for $\mathfrak{a}(\cdot,\cdot) - E_* \mathfrak{b}(\cdot,\cdot)$ on a subspace of codimension 1.

Proposition 4.5. *If the Standard Hypotheses* (4.32)–(4.34) *are satisfied, then*
$$\gamma \|u_h - u_*\|_{H^1}^2 \leq E_h - E_* \leq K \|u_h - u_*\|_{H^1}^2. \tag{4.35}$$

Furthermore, there exists $C > 0$ (independent of h) such that, up to a negligible higher-order term in the upper bound, we have
$$\|u_h - u_*\|_{H^1} \leq C \min_{v_h \in \mathscr{V}_h} \|u_* - v_h\|_{H^1} \tag{4.36}$$

for all h close enough to 0.

Proof. Specifying $v = u_h$ in the discrete variational formulation and $v = u_*$ in the continuous variational formulation, we get the relations
$$\begin{aligned} \mathfrak{a}(u_h,u_h) &= E_h \mathfrak{b}(u_h,u_h) = E_h, \\ \mathfrak{a}(u_*,u_*) &= E_* \mathfrak{b}(u_*,u_*) = E_*. \end{aligned}$$

Their difference can be transformed as
$$\begin{aligned} E_h - E_* &= \mathfrak{a}(u_h,u_h) - \mathfrak{a}(u_*,u_*) \\ &= \mathfrak{a}(u_* + e, u_* + e) - \mathfrak{a}(u_*,u_*) \\ &= \mathfrak{a}(e,e) + 2\mathfrak{a}(u_*,e) \\ &= \mathfrak{a}(e,e) + 2E_* \mathfrak{b}(u_*,e) \\ &= \mathfrak{a}(e,e) - 2E_* \mathfrak{b}(e,e) + 2E_* \mathfrak{b}(u_h,e). \end{aligned}$$

Next, we prove that the last term of the right-hand side is half the middle term, i.e., $2\mathfrak{b}(u_h,e) = \mathfrak{b}(e,e)$. Indeed, since $\mathfrak{b}(u_h,u_h) = \mathfrak{b}(u_*,u_*) = 1$,we have
$$\begin{aligned} 2\mathfrak{b}(u_h,e) &= 2\mathfrak{b}(u_h,u_h) - 2\mathfrak{b}(u_h,u_*) \\ &= \mathfrak{b}(u_h,u_h) + \mathfrak{b}(u_*,u_*) - 2\mathfrak{b}(u_h,u_*) \\ &= \mathfrak{b}(u_h - u_*, u_b - u_*). \end{aligned}$$

Finally, we obtain
$$E_h - E_* = \mathfrak{a}(e,e) - E_* \mathfrak{b}(e,e). \tag{4.37}$$

Using the first Standard Hypothesis (4.32) and the third one (4.34), with e in the place of v, we have

$$\gamma\|e\|_{H^1}^2 \le E_h - E_* \le K\|e\|_{H^1}^2,$$

which completes the proof of (4.35).

To prove (4.36), we start from the third Standard Hypothesis (4.34) and attempt to insert an arbitrary $v_h \in \mathscr{V}_h$ in the right-hand side as

$$\begin{aligned}
\gamma\|e\|_{H^1}^2 &\le \mathfrak{a}(e,e) - E_*\mathfrak{b}(e,e)\\
&\le \mathfrak{a}(v_h-u_*,e) - E_*\mathfrak{b}(v_h-u_*,e) + \mathfrak{a}(u_h-v_h,e) - E_*\mathfrak{b}(u_h-v_h,e)\\
&\le \mathfrak{a}(v_h-u_*,e) - E_*\mathfrak{b}(v_h-u_*,e)\\
&\quad + E_h\mathfrak{b}(u_h-v_h,u_h) - E_*\mathfrak{b}(u_h-v_h,u_*) - E_*\mathfrak{b}(u_h-v_h,e)\\
&\le \mathfrak{a}(v_h-u_*,e) - E_*\mathfrak{b}(v_h-u_*,e) + (E_h-E_*)\mathfrak{b}(u_h-v_h,u_h).
\end{aligned}$$

The first difference can be bounded thanks to the continuity expressed by the first Standard Hypothesis (4.32), while the second difference can be bounded by means of (4.35). This yields

$$\begin{aligned}
\gamma\|e\|_{H^1}^2 &\le K\|v_h-u_*\|_{H^1}\|e\|_{H^1} + K\|e\|_{H^1}^2\|u_h-v_h\|_{L^2}\|u_h\|_{L^2}\\
&\le K\|v_h-u_*\|_{H^1}\|e\|_{H^1} + K\|e\|_{H^1}^2(\|e\|_{L^2}+\|u_*-v_h\|_{L^2})\|u_h\|_{L^2}\\
&\le K\|v_h-u_*\|_{H^1}\|e\|_{H^1}(1+\|e\|_{H^1}) + K\|e\|_{H^1}^3,
\end{aligned}$$

the last two lines being due to $\|\cdot\|_{L^2} \le \|\cdot\|_{H^1}$ and $\|u_h\|_{L^2} = 1$. Dividing both sides by $\gamma\|e\|_{H^1}$ and passing to the infimum in v_h, we obtain

$$\|e\|_{H^1} - \frac{K}{\gamma}\|e\|_{H^1}^2 \le \frac{K}{\gamma}(1+\|e\|_{H^1}) \min_{v_h\in\mathscr{V}_h}\|v_h-u_*\|_{H^1}.$$

Thus, the claim (4.36) would be true with $C = K/\gamma$ if we could "omit" $\|e\|_{H^1}$ in front of 1 in the right-hand side, as well as $\|e\|_{H^1}^2$ in front of $\|e\|_{H^1}$ in the left-hand side, for h close enough to 0. Let us prove that $\|e\|_{H^1} \to 0$ as $h \to 0$. Combining (4.34) and (4.37), we have

$$\gamma\|e\|_{H^1} \le \mathfrak{a}(e,e) - E_*\mathfrak{b}(e,e) = E_h - E_* = \mathfrak{E}(u_h) - \mathfrak{E}(u_*).$$

Since u_h minimizes $\mathfrak{E}(\cdot)$ over $\mathscr{V}_h$, for all $w_h \in \mathscr{V}_h$ we must have

$$\begin{aligned}
\gamma\|e\|_{H^1} &\le \mathfrak{E}(w_h) - \mathfrak{E}(u_*) = \mathfrak{a}(w_h,w_h) - \mathfrak{a}(u_*,u_*) = \mathfrak{a}(w_h-u_*,w_h+u_*)\\
&\le K\|w_h-u_*\|_{H^1}\|w_h+u_*\|_{H^1}.
\end{aligned}$$

Taking $w_h = \arg\min_{v_h\in\mathscr{V}_h}\|v_h-u_*\|_{H^1}$, we are guaranteed by virtue of Lemma 4.1 that when $h \to 0$, $\|w_h-u_*\|_{H^1} \to 0$ and $\|w_h\|_{H^1}$ remains bounded. This completes the proof of (4.36). □

With the constants $\Gamma_{m,\epsilon}$ (approximation), K (continuity) and C (Céa) introduced earlier, we are in a position to assert the main *a priori* error estimates.

Theorem 4.2. *Let $\epsilon > 0$ be an arbitrarily small real number. If the Standard Hypotheses* (4.32)–(4.34) *are satisfied, then*

$$\|u_h-u_*\|_{H^1} \le C\Gamma_{m,\epsilon}\|u_*\|_{H^{3/2-\epsilon}}\,h^{1/2-\epsilon} \tag{4.38a}$$

$$E_h - E_* \le KC^2\Gamma_{m,\epsilon}^2\|u_*\|_{H^{3/2-\epsilon}}^2\,h^{1-2\epsilon} \tag{4.38b}$$

for all h close enough to 0.

PROOF. The inequality (4.38a) derives from chaining the Céa-type inequality (4.36) and the best approximation property (4.30). As for inequality (4.38b), it is a straightforward consequence of (4.35). □

Theorem 4.2 shows that the order M of the scaling function does have some influence on the coefficient of the upper bounds, but does not alter their orders in h. In the limit $\epsilon \to 0$, these orders tend to 1/2 for $\|u_h - u_*\|_{H^1}$ and 1 for $E_h - E_*$. This will be corroborated by the numerical experiments in §4.3.3.

We now switch to the second part of this section and strive to verify that our problem does comply with the Standard Hypotheses (4.32)–(4.34).

Proposition 4.6. *The three Standard Hypotheses* (4.32)–(4.34) *are satisfied.*

The proof of Proposition 4.6 relies on specific properties of the problem at hand, contrary to the first part, whose proofs are totally abstract and generic. Most of these properties were derived in §3.2.2 and will be used again in §5.3 for the *a posteriori* estimate. We first need a technical result.

Lemma 4.2. *There exists $W > 0$ (independent of h) such that for all $v \in \mathscr{V}$, we have*

$$\mathfrak{a}(v,v) - E_*\mathfrak{b}(v,v) \geq \frac{1}{4}\|v\|_{H^1}^2 - W\|v\|_{L^2}^2. \tag{4.39}$$

PROOF. Let $v \in \mathscr{V}$. By equation (3.68) of Proposition 3.5,

$$\mathfrak{a}(v,v) \geq \frac{1}{4}\|v\|_{H^1}^2 - \Theta\|v\|_{L^2}^2,$$

where $\Theta > 0$ does not depend on h. This results in

$$\mathfrak{a}(v,v) - E_*\mathfrak{b}(v,v) \geq \frac{1}{4}\|v\|_{H^1}^2 - \Theta\|v\|_{L^2}^2 - E_*\|v\|_{L^2}^2 = \frac{1}{4}\|v\|_{H^1}^2 - (\Theta + |E_*|)\|v\|_{L^2}^2.$$

Thus, inequality (4.39) holds with the constant

$$W = \Theta + |E_*|, \tag{4.40}$$

which is independent of h. □

PROOF OF PROPOSITION 4.6. Let $(v,w) \in \mathscr{V}^2$. By the triangle inequality,

$$\begin{aligned} |\mathfrak{a}(v,w) - E_*\mathfrak{b}(v,w)| &\leq |\mathfrak{a}(v,w)| + |E_*||\mathfrak{b}(v,w) \\ &\leq \kappa\|v\|_{H^1}\|w\|_{H^1} + |E_*|\|v\|_{L^2}\|w\|_{L^2} \end{aligned}$$

by the H^1-continuity of $\mathfrak{a}(\cdot,\cdot)$, established in equation (3.67) of Proposition 3.5. Noting that $\|\cdot\|_{L^2} \leq \|\cdot\|_{H^1}$, we have

$$|\mathfrak{a}(v,w) - E_*\mathfrak{b}(v,w)| \leq (\kappa + |E_*|)\|v\|_{H^1}\|w\|_{H^1}.$$

Thus, the first Standard Hypothesis (4.32) holds, in which the constant

$$K = \kappa + |E_*| \tag{4.41}$$

does not depend on h.

To prove the second Standard Hypothesis (4.33), let

$$E^{(2)} = \inf_{\substack{v\in\mathscr{V}\\ v\perp u_*}} \frac{\mathfrak{a}(v,v)}{\mathfrak{b}(v,v)}$$

be the "second eigenvalue" of the continuous problem on $\mathscr{V}$. According to Corollary 3.6, $E^{(2)} > E^{(1)} = E_*$. For all $v \in (u_*)^\perp$, we then have

$$\mathfrak{a}(v,v) - E_*\mathfrak{b}(v,v) \geq (E^{(2)} - E^{(1)})\,\|v\|_{L^2}^2. \tag{4.42}$$

The constant $\beta = E^{(2)} - E^{(1)} > 0$ does not depend on h.

Finally, we prove the third Standard Hypothesis (4.34). By expanding the bilinear forms below, we get

$$\begin{aligned}
\mathfrak{a}(e,e) - E_*\mathfrak{b}(e,e) &= \mathfrak{a}(u_h-u_*,u_h-u_*) - E_*\mathfrak{b}(u_h-u_*,u_h-u_*)\\
&= \mathfrak{a}(u_h,u_h) - E_*\mathfrak{b}(u_h,u_h) + \mathfrak{a}(u_*,u_*) - E_*\mathfrak{b}(u_*,u_*)\\
&\quad - 2\big(\mathfrak{a}(u_*,u_h) - E_*\mathfrak{b}(u_*,u_h)\big)\\
&= \mathfrak{a}(u_h,u_h) - E_*\mathfrak{b}(u_h,u_h) + 0 - 2\cdot 0
\end{aligned} \tag{4.43}$$

thanks to the variational formulation on $\mathscr{V}$. For any $v \in \mathscr{V}$, decompose $v = v_1u_* + w$, with $w \in (u_*)^\perp$. Then $\|v\|_{L^2}^2 = v_1^2 + \|w\|_{L^2}^2$. We have

$$\begin{aligned}
\mathfrak{a}(v,v) - E_*\mathfrak{b}(v,v) &= \mathfrak{a}(v_1u_*+w,v_1u_*+w) - E_*\mathfrak{b}(v_1u_*+w,v_1u_*+w)\\
&= v_1^2(\mathfrak{a}(u_*,u_*) - E_*\mathfrak{b}(u_*,u_*)) + \mathfrak{a}(w,w) - E_*\mathfrak{b}(w,w)\\
&\quad + 2v_1(\mathfrak{a}(u_*,w) - E_*\mathfrak{b}(u_*,w))\\
&\geq 0 + \beta\|w\|_{L^2}^2 + 0
\end{aligned}$$

due to second Standard Hypothesis (4.33) for $w \in (u_*)^\perp$. So

$$\mathfrak{a}(v,v) - E_*\mathfrak{b}(v,v) \geq \beta\|w\|_{L^2}^2 = \beta(\|v\|_{L^2}^2 - v_1^2).$$

Take $v = u_h$ in this inequality, then $v_1 = \langle u_h, u_*\rangle_{L^2}$ and

$$\mathfrak{a}(u_h,u_h) - E_*\mathfrak{b}(u_h,u_h) \geq \beta(\|u_h\|_{L^2}^2 - |\langle u_h,u_*\rangle_{L^2}|^2).$$

According to equality (4.43),

$$\begin{aligned}
\mathfrak{a}(e,e) - E_*\mathfrak{b}(e,e) &= \mathfrak{a}(u_h,u_h) - E_*\mathfrak{b}(u_h,u_h)\\
&\geq \beta(\|u_h\|_{L^2}^2 - |\langle u_h,u_*\rangle_{L^2}|^2) \geq \beta(\|u_h\|_{L^2}^2 - |\langle u_h,u_*\rangle_{L^2}|)
\end{aligned}$$

because $|\langle u_h,u_*\rangle_{L^2}| \leq \|u_h\|_{L^2}\|u_*\|_{L^2} = 1$. Since both $\pm u_h$ and $\pm u_*$ satisfy their respective variational formulations, we can choose the signs of u_h and u_* in such a way that $\langle u_h,u_*\rangle_{L^2} \geq 0$. It is then possible to drop the absolute value to obtain

$$\begin{aligned}
\mathfrak{a}(e,e) - E_*\mathfrak{b}(e,e) &\geq \beta\{\|u_h\|_{L^2}^2 - \langle u_h,u_*\rangle_{L^2}\}\\
&\geq \frac{\beta}{2}\{\|u_h\|_{L^2}^2 + \|u_*\|_{L^2}^2 - 2\langle u_h,u_*\rangle_{L^2}\}\\
&\geq \frac{\beta}{2}\|u_h-u_*\|_{L^2}^2 = \frac{\beta}{2}\|e\|_{L^2}^2.
\end{aligned} \tag{4.44}$$

Let us combine this with

$$\mathfrak{a}(e,e) - E_*\mathfrak{b}(e,e) \geq \frac{1}{4}\|e\|_{H^1}^2 - W\|e\|_{L^2}^2, \tag{4.45}$$

which stems from (4.39) of Lemma 4.2 applied to $v = e$, in the following fashion: multiply (4.44) by W, multiply (4.45) by $\frac{1}{2}\beta$ and add them together. It follows that

$$\mathfrak{a}(e,e) - E_*\mathfrak{b}(e,e) \geq \frac{\beta}{4(2W+\beta)}\| e\|_{H^1}^2.$$

The constant $\gamma = \dfrac{\beta}{4(2W+\beta)} > 0$ does not depend on h. □

4.3.3 Numerical results

In practice, we should pay attention to the units of the quantities used in the equation. L, X_I and h have the same unit of length, as do $\Lambda_I = Z_I^{-1}$, since it is the visible "length" of the Slater function S_{Z_I}. Therefore we are going to work with the ratios X_I/L, $N = L/h$, Λ_I/h and L/Λ_I instead of L, X_I, Λ_I or h alone, because these quotients are adimensional and there are actual senses to them: for example, L/Λ_I is the relative size of the domain compared to the size of the Slater function S_{Z_I}, and Λ_I/h represents how fine the grid is in comparing with the cups. Moreover, numerical tests show that the solution only depends on these ratios and not really on L, X_I, Λ_I or h themselves.

We would consider large L/Λ_I (of value at least 10), since the model simulates an infinite domain. First, we plot the approximate wave function

$$u_h = \sum_{j=0}^{N-1} \mathsf{u}_j^h\, \widetilde{\chi}_j^h$$

over several bases of Daubechies scaling functions. Indeed, after obtaining the coefficients u_j^h's by Proposition 4.4, over a basis of N elements, the point values of the wave function are calculated from the point values of the $\widetilde{\chi}_j^h$'s, on a fine mesh of 1024 points, independently of N.

We also plot the relative error on the energy level E and look at its order of convergence, to see whether it's coherent with the result in Theorem 4.2. We are going to show an ample set of examples in the case of single-delta potentials, to observe how the solution and its approximations behave around the cusp or in the rest of the domain. Resolutions will also be given in the case of multi-delta potentials.

Single-delta potentials

Figure 4.5 plot the exact solution u_* and the approximate wave function u_h in the case of a potential of charge Z, over two Daubechies bases: `db4` and `db5`, with the parameters:

$$L = 1,\ L/\Lambda = 20,\ X/L = 0.5,\ N = 2^5.$$

We see that the scaling function basis of higher order (`db5`) does not necessarily produce a better resolution than the basis of lower order (`db4`) does. Later results on the energy error, over bases of many different orders, will dwell on this observation.

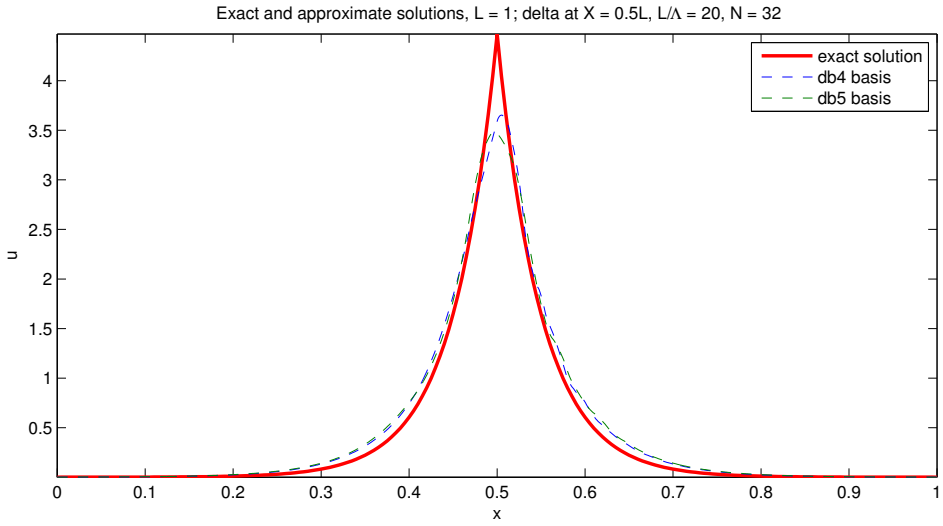

Figure 4.5: Wave functions on periodic domain, single-delta potential, over db4, db5 bases.

When choosing one among Daubechies scaling function families for our test, we pick db4 for some of its numerical properties, but other families work as well. Figure 4.6 plot the approximations for different Slater functions, corresponding to $L/\Lambda = 40, 20, 10$ respectively on the three panels downward. Each panel plots the solutions over bases of Daubechies scaling functions db4, with different meshes:

$$N \in \{2^5, 2^6, 2^7, 2^8, 2^9\}; \quad L = 1,\ L/\Lambda = 20,\ X/L = 0.5.$$

We see that the behavior of numerical solutions depends not on the number N of mesh points but on the relative grid step Λ/h instead; for example, for the same $N = 32$ the three blue curves are differently proportioned to the exact solutions (the red thick curves) in the three panels: the blue curve in the top one is rather bad compared to that in the bottom panel. On the other hand, the curves with the same Λ/h on each panel have the same proportion to the corresponding exact solutions. This result is intuitive: the narrower the Slater function becomes, the finer the mesh is required, for the cusp needs "seeing" enough of mesh points.

To have a more quantitative look at this observation, we consider the ***relative error*** on the energy E

$$\texttt{err}_h := \frac{E_h - E_*}{|E_*|}, \tag{4.46}$$

which is positive since $E_* < E_h$. In our cases of 1-D equations, the fundamental energy E_* can be calculated directly, as in chapter §3. Table 4.2 illustrates the fact that **the solution depends on Λ/h and not on L/Λ or N alone**. It lists the relative error on the energy over a db4 basis, when we keep Λ/h at a fixed value of 3.2 and changing N. The differences between the obtained $\texttt{err}_h$'s are very small and can be ignored.

Figure 4.5–4.6 also show that numerical solutions do not approach very well at the cusp when N is small - that will be an advantage of the mixed bases later on. Instead of refining the mesh and solving large linear systems, we shall add adequate Gaussian functions to the scaling function bases.

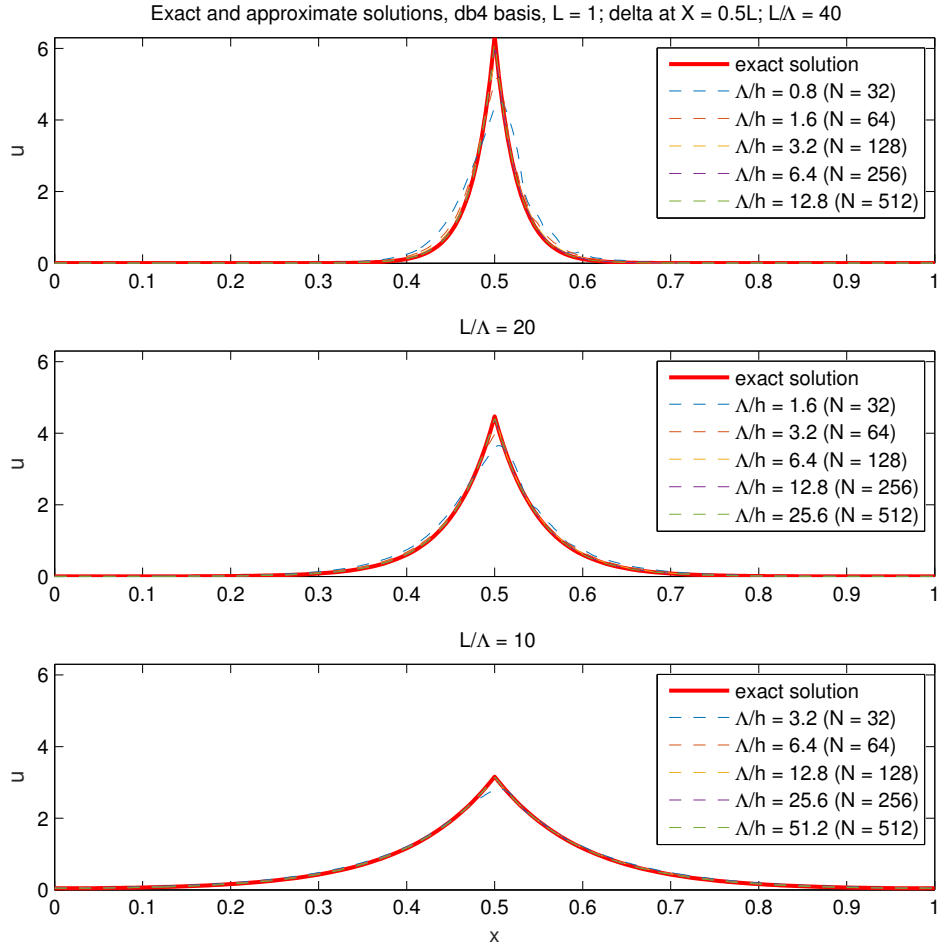

Figure 4.6: Wave functions on periodic domain, single-delta potential, over db4 basis, for $L/\Lambda = 40$ (top), 20 (middle) and 10 (bottom).

Outside of the cusp neighborhood, the effects of the boundary may be notable if the domain is not large enough. For example, if we take $L/\Lambda < 10$, numerical solutions are visibly bad at the boundary, as in Figure 4.7, with the parameters

$$N \in \{2^5, 2^6, 2^7\};\ L = 1,\ X/L = 0.5,\ L/\Lambda \in \{5, 10\}$$

That is why we would always consider $L/\Lambda \geq 10$, or we must refine the mesh.

The relative error $\texttt{err}_h$ is the main criterion to evaluate the approximate solutions. In the following figures, we plot $\texttt{err}_h$, with logarithmic scale, against the relative grid step Λ/h while refining the mesh. We also wish to see whether the fact that the delta potential located at a mesh point or not affects the results. With a single-delta potential always at $X = L/2$, we consider the case of N being even (X is a mesh point) and the case of N being odd (X is not at a mesh point) - if we work only with scaling functions and not wavelets, N can be any number in $\mathbb{N}^*$ and not necessarily a power of 2.

Figure 4.8 plots $\texttt{err}_h$ when N is even and

$$N = 18 \sim 400,\ L = 1,\ X/L = 0.5, L/\Lambda = 20,$$

over Daubechies scaling function bases of order $\mathsf{M} = 3, 4, 5$. To have a comparison with the usual finite element method, we also show the results over P_1 bases (the blue curve). Figure 4.9 plots the same quantities, except that N is now odd

$$N = 17 \sim 401,\ L = 1,\ X/L = 0.5, L/\Lambda = 20.$$

Figure 4.8 and 4.9 show that a basis of higher order does not necessarily give a better solution, as also attested by Figure 4.5. All the curves over scaling function bases are of slope (-1), so the errors on E are of order 1 for any M, which is coherent to the consequence of Theorem 4.2. This low order is due to the stiffness of the exact solution. P_1 bases has a big advantage in Figure 4.8, as the hat function is put right on the nucleus position to approximate the cusp.

Remark 4.1. In [115], using transparent boundary conditions and P_1 finite elements, we also observed an order of convergence equal to 1 for the energy difference $E_h - E_*$ and 1/2 for the solution error $\|u_h - u_*\|_{H^1}$. There is, however, an exceptional case for which $E_h - E_*$ is of order 2 and $\|u_h - u_*\|_{H^1}$ of order 1: this is when the nucleus coincides with a node of the mesh, as in Figure 4.8. Such a geometric configuration is indeed favorable to the approximation property of the exact solution.

Figure 4.10 and 4.11 plot the solutions on two types of scaling function bases: symmlet bases, which are symmetric (defined in §2.1.3 and denoted sym for order M) and Daubechies bases, which are of minimal phase (denoted dbM for order M). The left panels are for symmlet bases and the right panels are for Daubechies bases. In Figure 4.10, N is even and the nucleus coincides with a node:

$$N = 38 \sim 1038,\ L = 1,\ X/L = 0.5, L/\Lambda = 10,\ \mathsf{M} = 4 \sim 10$$

In Figure 4.11, N is odd and the nucleus does not coincide with a node:

$$N = 39 \sim 1039,\ L = 1,\ X/L = 0.5, L/\Lambda = 10,\ \mathsf{M} = 4 \sim 10.$$

All the curves are of slope (-1), so the errors on E remain of order 1.

X/L	Λ/h	N	Z	$\texttt{err}_h$
0.5	3.2	2^5	10	0.14929729
		2^6	20	0.14947930
		2^7	40	0.14947932

Table 4.2: Single delta, db4 basis, with different N, Z but fixed Λ/h.

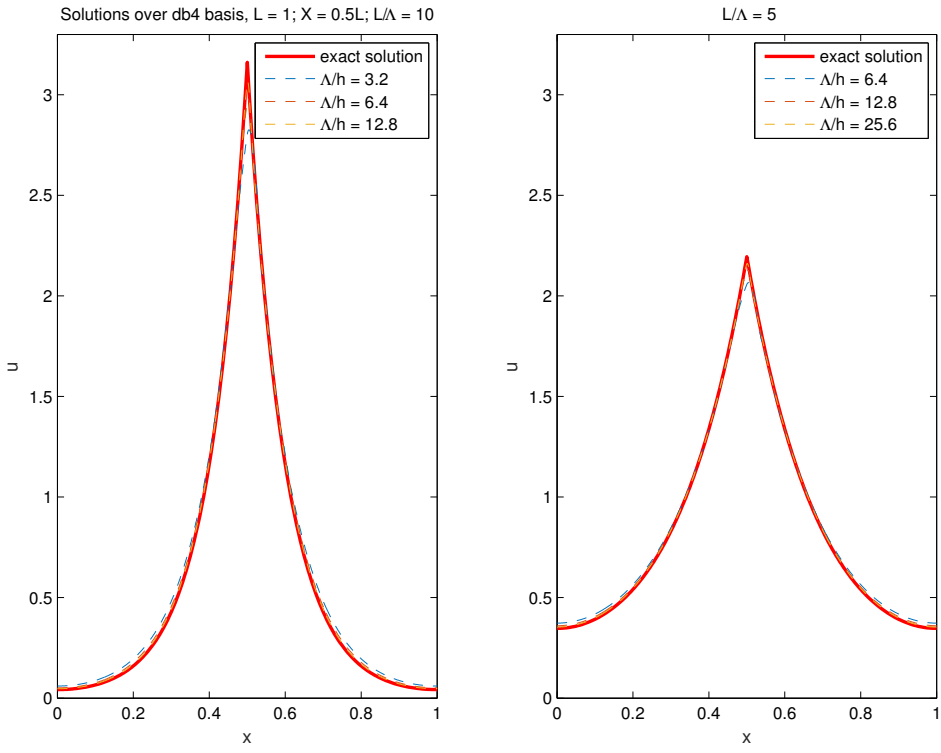

Figure 4.7: Wave functions on periodic domain, single-delta potential, over db4 basis, when L/Λ is small: $L/\Lambda = 10$ (left), $L/\Lambda = 5$ (right).

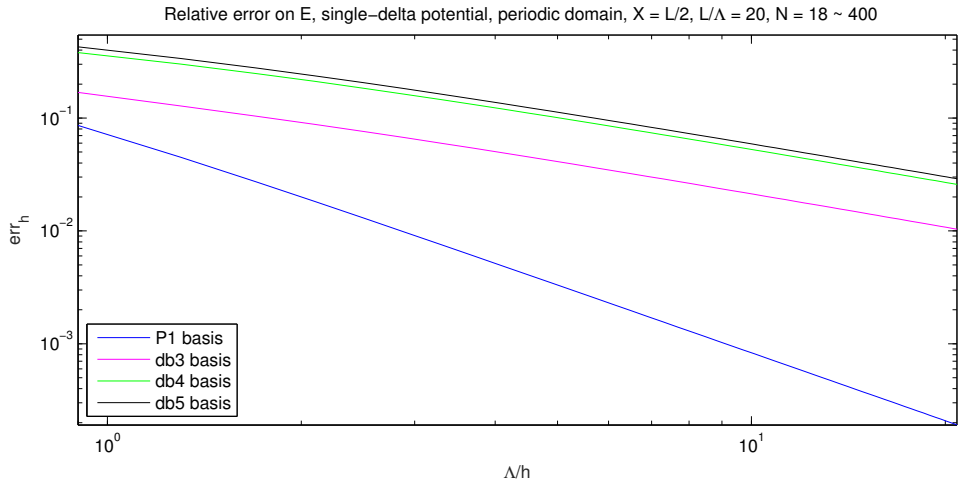

Figure 4.8: Relative error $\mathtt{err}_h$ when the single-delta potential is at a mesh node ($N = L/h$ is even), periodic domain.

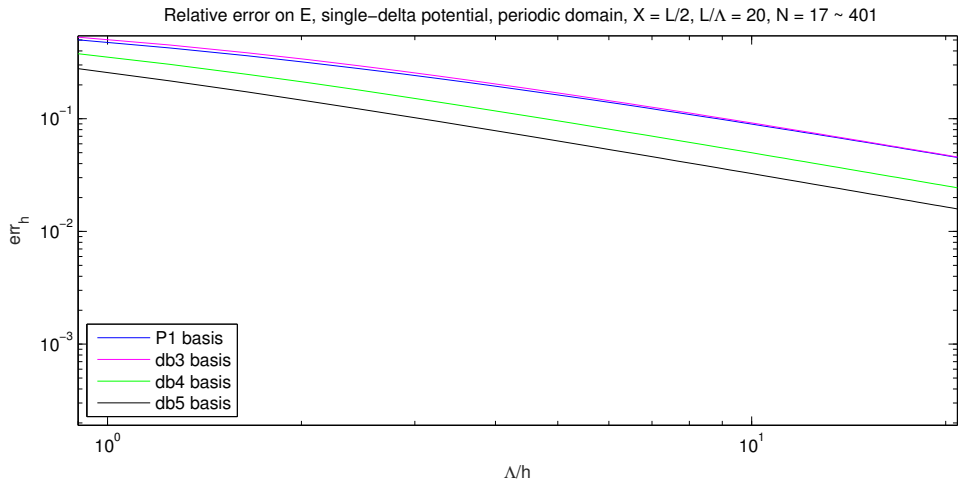

Figure 4.9: Relative error $\mathtt{err}_h$ when the single-delta potential is between two mesh nodes ($N = L/h$ is odd), periodic domain.

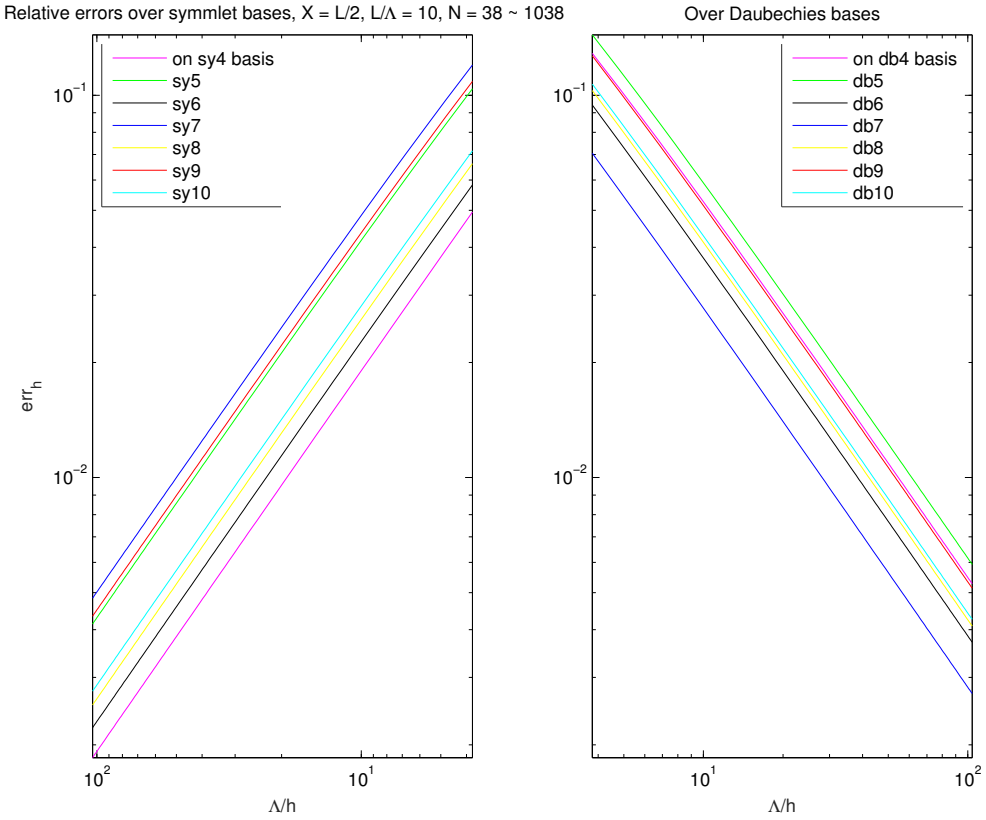

Figure 4.10: Relative error on E when the single-delta potential is on a node, $N = L/h$ is even.

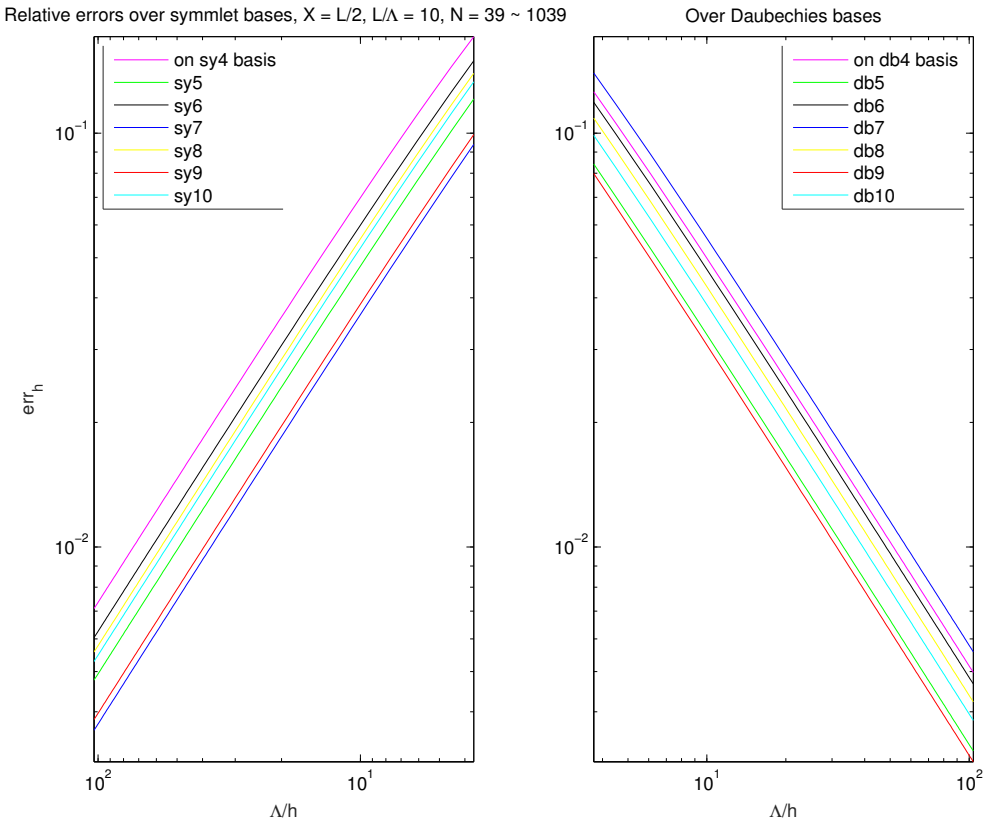

Figure 4.11: Relative error on E when the single-delta potential is at mid-step, $N = L/h$ is odd.

Multi-delta potentials

We continue to use the adimensional parameters X_I/L, $N = L/h$, Λ_I/h and L/Λ_I. When looking at the convergence of the solution, the ratio $\min_{1\le I\le M}(\Lambda_I)/h$ is considered to measure the fineness of the mesh with respect to the length of the Slater functions. For multi-delta potentials, there are new parameters joining in, which are the distances between nuclei.

For a **double-delta potential** at $[X_1; X_2]$, denote R the distance between the two nuclei:

$$R = X_2 - X_1.$$

This distance has an important effect on the solution, so the quotient $R/\max(\Lambda_1, \Lambda_2)$ will also be surveyed in our tests. Figure 4.12 plot the exact solution and approximate ones over `db4` scaling function bases, with fixed charges Z_1 and Z_2 but with different distances R. On each graph, we refine the mesh to see how the numerical solutions behave.

$$L = 1,\ Z_1 = 20,\ Z_2 = 19.9,\ X_1 = 1/4,\ X_2 \in \{3/8, 1/2, 5/8\},\ N \in \{64, 128, 256\}.$$

We see again here the *ionic bonding* phenomenon, where the electron is transferred from one ion to another by electrovalence. In the top panel of Figure 4.12, the two nuclei are far enough so the electron was "definitely" attracted to the one with a greater charge, even if the other charge is not much lesser. In that case, the approximate solutions tend to overestimate the smaller cusp and underestimate the bigger cup. In the bottom panel, the difference between the two cusps is smoothed out, because their distance R is reduced.

Figure 4.13 plots the solutions over `db4` scaling function bases, when the positions of the two nuclei are fixed but the charges Z_1 and Z_2 are reduced by 2 each time, equivalently the two cusps spread out twice more each time. On each graph, we also refine the mesh.

$$L = 1,\ X_1 = 1/4,\ X_2 = 1/2,\ N \in \{64, 128, 256\},\ [Z_1; Z_2] \in \{[40; 39.8], [20; 19.9], [10; 9.95]\}.$$

It once more shows the effect of $R/\max(\Lambda_1, \Lambda_2)$, or $\min(Z_1, Z_2)R$, on the solution. If R is fixed, the more the Z_I's are increased, the more the cusps are sensitive to the difference of charges. In the top panel, the approximate solutions also overestimate the smaller cusp and underestimate the bigger cup. The solutions are dependent upon the relative grid step $\min(\Lambda_1, \Lambda_2)/h$, as in the case of single-delta potentials.

Figure 4.14 plots the solutions over `db4` and `db5` bases, to show that higher orders do not necessarily give better resolutions. The two approximate wave functions nearly coincide with each other in the internuclear regions but differ at the cusps.

$$L = 1,\ Z_1 = 20,\ Z_2 = 19.9,\ X_1 = 3/8,\ X_2 = 5/8,\ N = 64.$$

The relative error $\texttt{err}_h$, defined in (4.46), will give a clearer look at the orders of approximations over different Daubechies bases. Figure 4.15 plots $\texttt{err}_h$ against $\min(\Lambda_1, \Lambda_2)/h$ when the mesh is being refined, over bases of `db4` $\sim$ `db10` scaling functions.

$$L = 1,\ Z_1 = 20,\ Z_2 = 19,\ X_1 = 1/2,\ X_2 = 3/4,\ N = 32 \sim 1024.$$

Coherent to the consequence of Theorem 4.2, all errors on E are of order 1 for any `M`.

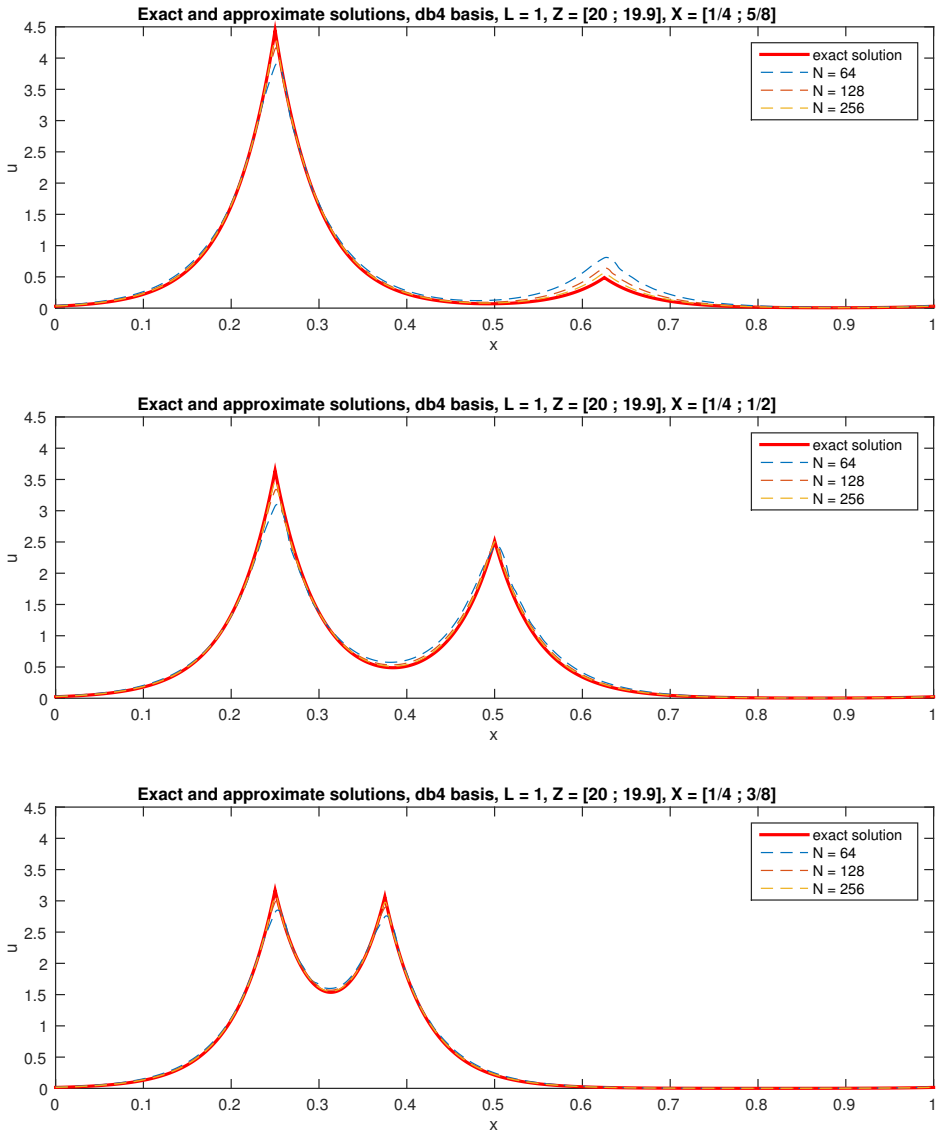

Figure 4.12: Exact and approximate wave functions on periodic domain, double-delta potential, over db4 basis, for $R = 3/8$ (top), $1/4$ (middle) and $1/8$ (bottom).

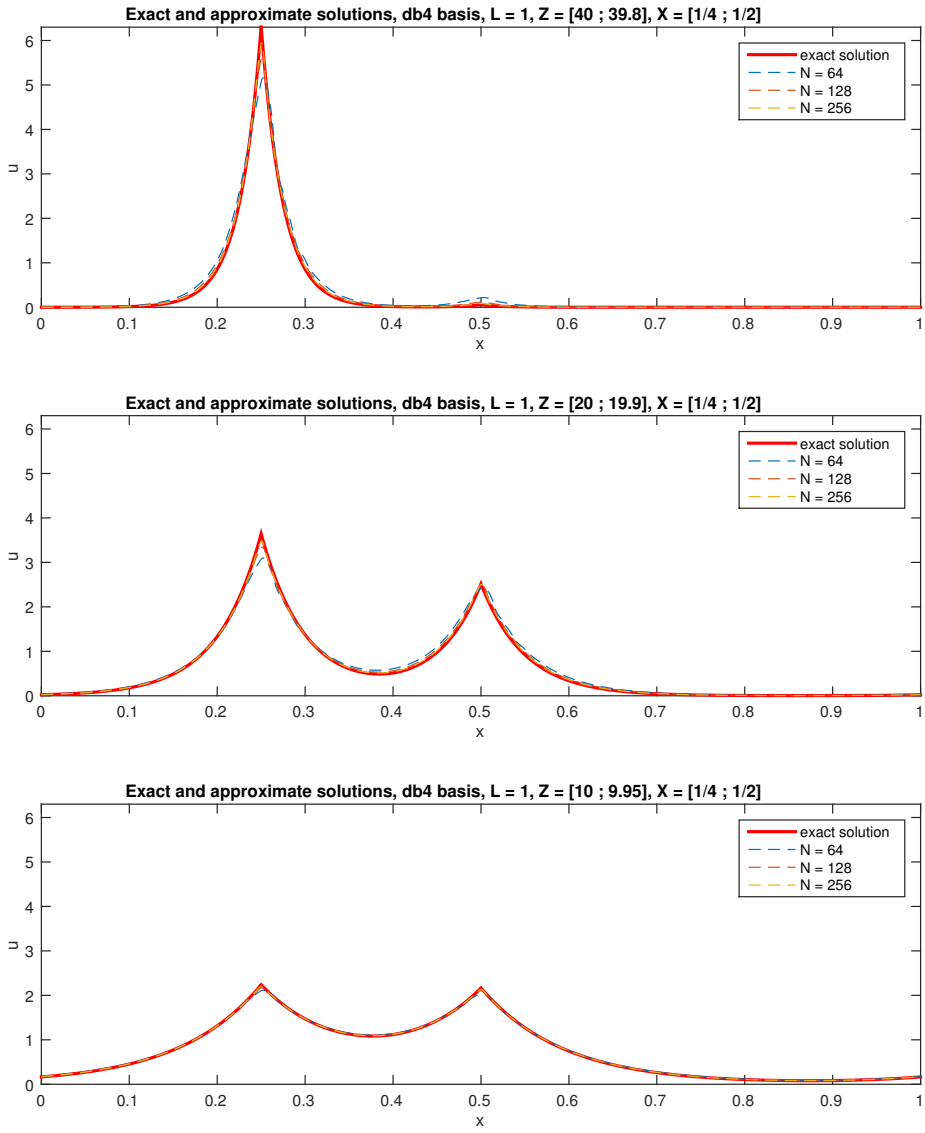

Figure 4.13: Exact and approximate wave functions on periodic domain, double-delta potential, over db4 basis, with fixed $[X_1; X_2]$ but varied $[Z_1; Z_2]$.

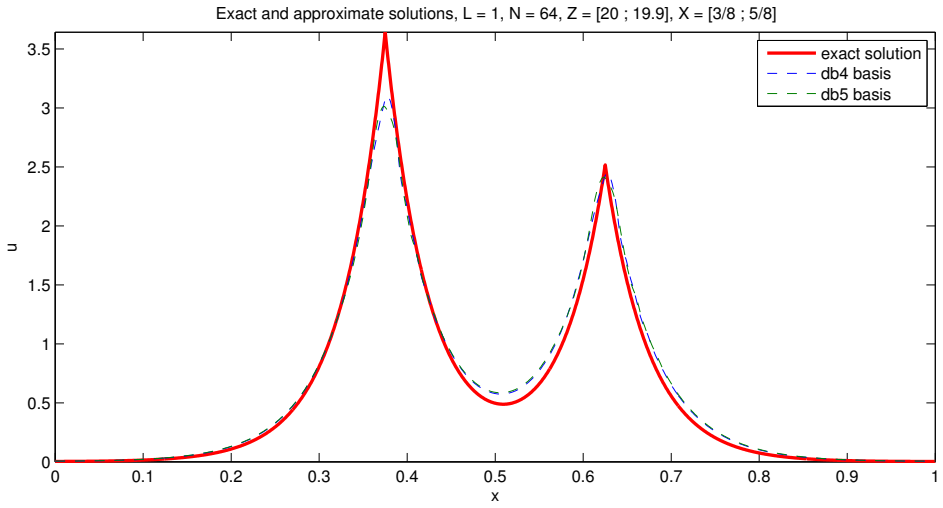

Figure 4.14: Exact and approximate wave functions on periodic domain, double-delta potential, over db4, db5 bases.

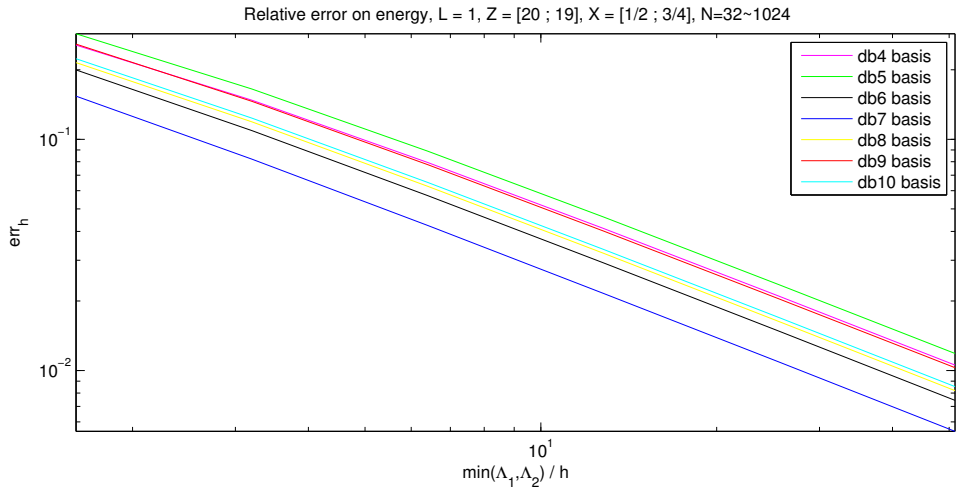

Figure 4.15: Relative error $\texttt{err}_h$ on periodic domain, double-delta potential, when $N = \dfrac{L}{h}$ is increasing.

Figure 4.16 illustrates the varying of the errors, over `db4` $\sim$ `db10` bases, when the internuclear distance R changes. The errors tend to rise when R increases, up until $R = L/2$. An explication might be that when the two nuclei are far from each other, the cusps become stark, so the scaling functions start to lose their grip on the cusps, as also seen in Figure 4.12. When $R > L/2$, the L-distance $|X_1 - X_2|^{\sim}$, as defined in (3.77), is actually less than $L/2$. The curves in Figure 4.16 are vertically symmetric across the line $R = L/2$, due to periodicity.

$$L = 1,\ Z_1 = 20,\ Z_2 = 19,\ X_1 = \frac{1}{32},\ R \in \left\{\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \frac{15}{16}\right\},\ N \in \{128, 256\}.$$

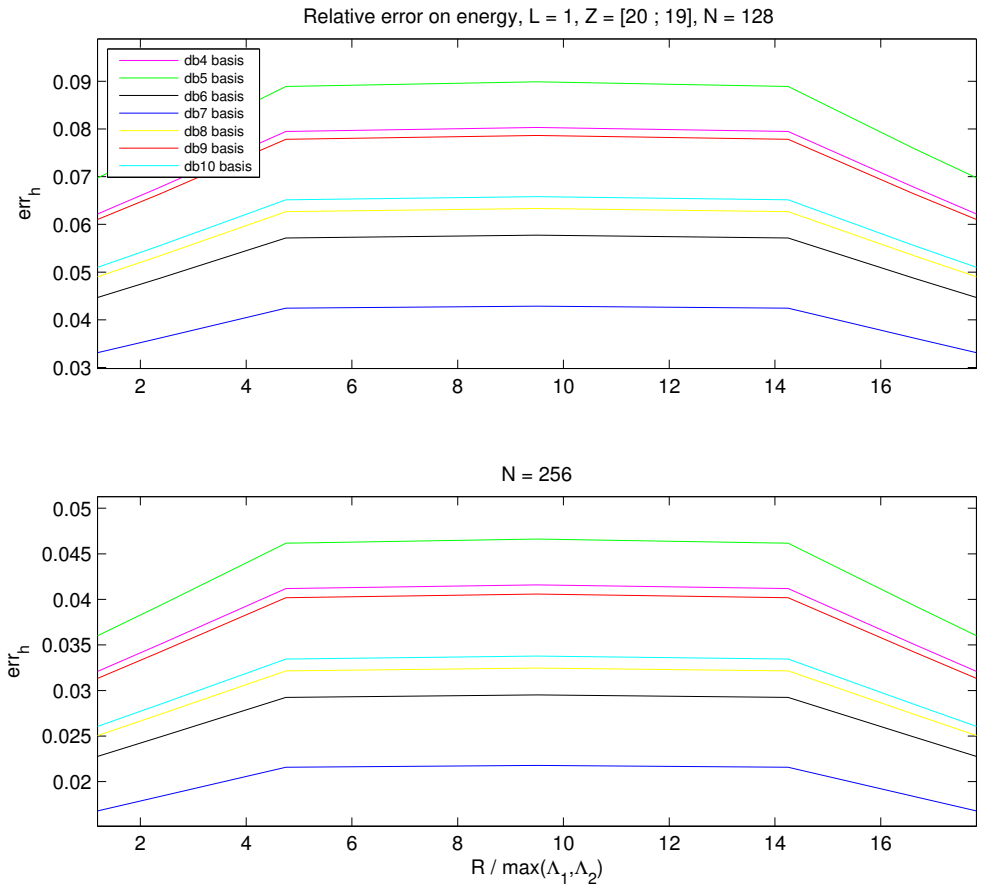

Figure 4.16: Relative error $\texttt{err}_h$ on periodic domain, double-delta potential, when the internuclear distance R is increasing, for $N = 128$ (top) and 256 (bottom).

In the case of **triple-delta potentials**, Figure 4.17 shows the solutions in different situations. When

$$X_3 - X_2 = X_2 - X_1 = L/3,$$

the three nuclei are equally located on the "circle" $[0, L]$. If, moreover, the three charges

are the same, then the three cusps are identical, for example as in the top panel:

$$N = 128,\ L = 1,\ Z_1 = Z_2 = Z_3 = 10,\ X_1 = 2/9,\ X_2 = 5/9,\ X_3 = 8/9.$$

The middle panel plots u on the same mesh, when the charges are still identical but the nuclei are differently located:

$$Z_1 = Z_2 = Z_3 = 10,\ X_1 = 2/9,\ X_2 = 3/9,\ X_3 = 5/9.$$

It shows that the two closer cusps tend to even out each other. The bottom panel keep the same nucleus positions but tried different charges:

$$Z_1 = 20,\ Z_2 = 15,\ Z_3 = 10,\ X_1 = 2/9,\ X_2 = 3/9,\ X_3 = 5/9,$$

then the smaller cusps has almost disappeared from the atomic orbitals. In all three cases, the db4 basis approaches the internuclear regions better than at the cusps.

4.4 Resolution of the periodic model on mixed bases

We are at last ready to attempt the Galerkin approximation of the periodic model (3.57)–(3.58) in a mixed basis. The first natural idea is to enrich a basis of Daubechies scaling functions by M periodized contracted Gaussians, each located at a nucleus position. The contracted Gaussians to be inserted are those previously designed in §4.2, without any modification apart from periodization. Of course, it is not expected that these *pre-optimized contracted Gaussians* remain optimal or quasi-optimal with the scaling functions for all mesh sizes. What we hope, however, is to show that they can already significantly improve the quality of the approximate solutions.

The discrete eigenvalue problem is described in §4.4.1 for an arbitrary number $M \geq 1$ of nuclei. The technical issues that arise from the computation of various elementary scalar products are discussed at length in §4.4.2. Finally, numerical results are provided in §4.4.3 for $M = 1$ and $M = 2$.

4.4.1 Discrete eigenvalue problem

In §4.2.2, we suggested the CGTO basis

$$\Big\{\mathrm{CG}_I := \mathrm{CG}(\boldsymbol{\sigma}^*(Q_I), \mathbf{v}^*(Q_I), \cdot - X_I), 1 \leq I \leq M\Big\}$$

for the infinite model (3.1). At each nucleus I, the number Q_I of primitives for the contracted Gaussian CG_I is to be set by the user. The reference standard deviations $\boldsymbol{\tau}^*(Q_I) = \boldsymbol{\sigma}^*(Q_I)/\Lambda_I \in (\mathbb{R}_+^*)^{Q_I}$ and the reference coefficients $\mathbf{v}^*(Q_I) \in \mathbb{R}^{Q_I}$ were optimized in the sense of (4.14) for a single-delta energy functional with $Z = 1$, and their values were given in Table 4.1. Consider the subspace

$$\mathscr{V}_g = \mathrm{Span}\{\widetilde{\mathrm{CG}}_I,\ \ 1 \leq I \leq M\}, \tag{4.47}$$

where

$$\widetilde{\mathrm{CG}}_I(\cdot) = \sum_{n\in\mathbb{Z}} \mathrm{CG}_I(\cdot + nL)$$

is the L-periodization of the I-th contracted Gaussian.

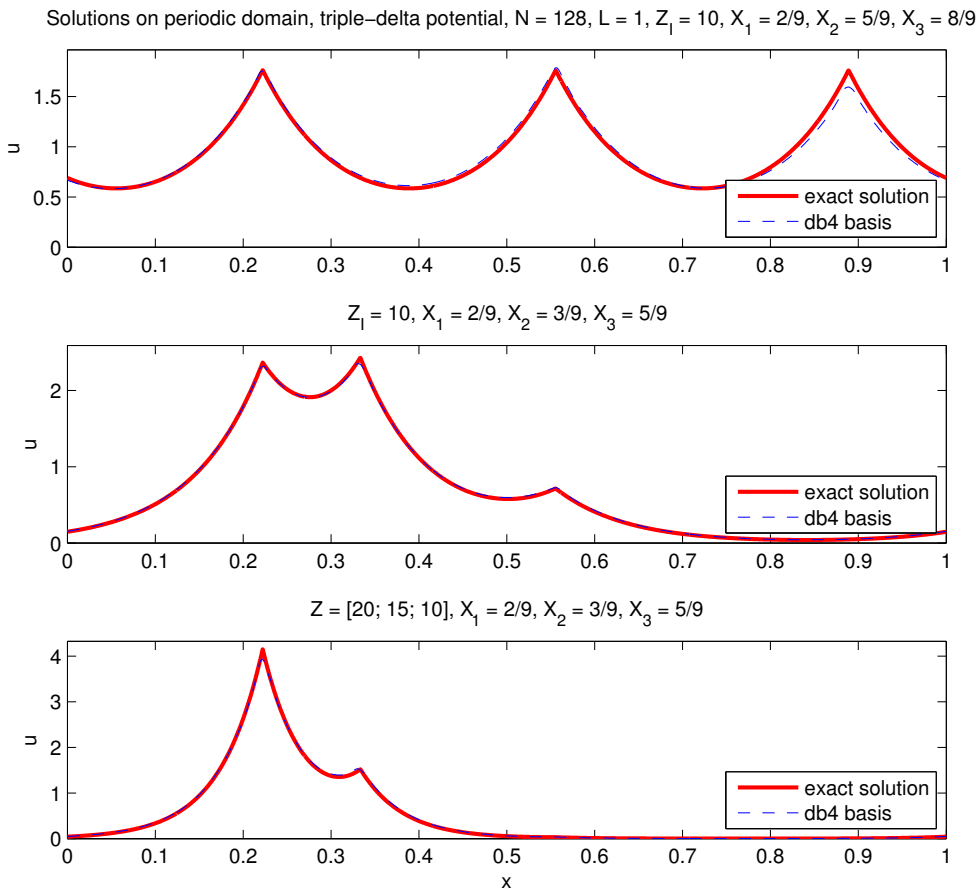

Figure 4.17: Exact and approximate wave functions on periodic domain, triple-delta potential, over db4 bases.

Since the periodization operator acts linearly, we have

$$\widetilde{\mathrm{CG}}_I(\cdot) = \sum_{q=1}^{Q_I} \mathsf{v}_q^*(Q_I)\, \widetilde{g}_{\sigma_q^*(Q_I)}(\cdot - X_I) = \sum_{q=1}^{Q_I} \mathsf{v}_q^*(Q_I)\, \widetilde{g}_{\sigma_q^*(Q_I),X_I},$$

the shorthand notation $\widetilde{g}_{\sigma,X} = \widetilde{g}_\sigma(\cdot - X)$ standing for the L-periodization of the shifted normalized Gaussian $g_{\sigma,X} = g_\sigma(\cdot - X)$. Set

$$\mathscr{V}_b = \mathscr{V}_{h,g} := \mathscr{V}_h \oplus \mathscr{V}_g, \tag{4.48}$$

where $\mathscr{V}_h$ was defined in (4.25) as the subspace spanned by the periodized scaling functions $\widetilde{\chi}_i^h$, $0 \le i \le N-1$. The Galerkin approximate solution on $\mathscr{V}_{h,g}$ is designated by $(u_{h,g}, E_{h,g})$. Let us decompose the approximate wave function $u_{h,g}$ into a sum

$$u_{h,g} = \sum_{j=0}^{N-1} \mathsf{u}_j^{h,g}\, \widetilde{\chi}_j^h + \sum_{J=1}^{M} \mathsf{u}_{N-1+J}^{h,g}\, \widetilde{\mathrm{CG}}_J$$

and encapsulate the set of coefficients $\mathsf{u}_j^{h,g}$, $0 \le j \le N+M-1$, into the block vector

$$\mathbf{u}^{h,g} = \begin{pmatrix} \mathbf{u}^h \\ \mathbf{u}^g \end{pmatrix} \in \mathbb{R}^{N+M}.$$

Proposition 4.7. *If $N \ge 4\mathsf{M}-2$, then the pair $(\mathbf{u}^{h,g}, E_{h,g}) \in \mathbb{R}^{N+M} \times \mathbb{R}$ solves the smallest eigenvalue problem*

$$\mathbf{A}^{h,g}\mathbf{u}^{h,g} = E_{h,g}\mathbf{B}^{h,g}\mathbf{u}^{h,g}, \tag{4.49a}$$

$$(\mathbf{u}^{h,g})^T\mathbf{B}^{h,g}\mathbf{u}^{h,g} = 1, \tag{4.49b}$$

in which the $(N+M)\times(N+M)$ matrices $\mathbf{A}^{h,g}$ and $\mathbf{B}^{h,g}$ are given by

$$\mathbf{A}^{h,g} = \begin{pmatrix} \mathbf{A}^h & \mathbf{A}^{hg} \\ \mathbf{A}^{gh} & \mathbf{A}^g \end{pmatrix}, \qquad \mathbf{B}^{h,g} = \begin{pmatrix} \mathbf{B}^h & \mathbf{B}^{hg} \\ \mathbf{B}^{gh} & \mathbf{B}^g \end{pmatrix}, \tag{4.50}$$

where $\mathbf{A}^h$ and $\mathbf{B}^h$ are respectively the $N \times N$ Hamiltonian and mass matrices defined in Proposition 4.4 for the pure scaling functions basis, the entries of $\mathbf{A}^g$ and $\mathbf{B}^g$ are

$$\begin{aligned} \mathsf{A}_{IJ}^g = {} & \frac{1}{2}\sum_{p=1}^{Q_I}\sum_{q=1}^{Q_J} \mathsf{v}_p^*(Q_I)\mathsf{v}_q^*(Q_J)\, \langle \widetilde{g}'_{\sigma_q^*(Q_J),X_J}, \widetilde{g}'_{\sigma_p^*(Q_I),X_I} \rangle_{L^2(0,L)} \\ & - \sum_{p=1}^{Q_I}\sum_{q=1}^{Q_J} \mathsf{v}_p^*(Q_I)\mathsf{v}_q^*(Q_I) \sum_{K=1}^{M} Z_K\, \widetilde{g}_{\sigma_q^*(Q_J),X_J}(X_K)\widetilde{g}_{\sigma_p^*(Q_I),X_I}(X_K) \end{aligned} \tag{4.51a}$$

$$\mathsf{B}_{IJ}^g = \sum_{p=1}^{Q_I}\sum_{q=1}^{Q_J} \mathsf{v}_p^*(Q_I)\mathsf{v}_q^*(Q_I)\, \langle \widetilde{g}_{\sigma_q^*(Q_J),X_J}, \widetilde{g}_{\sigma_p^*(Q_I),X_I} \rangle_{L^2(0,L)} \tag{4.51b}$$

for $(I,J)\in\{1,\dots,M\}^2$*, and the entries of* $\mathbf{A}^{hg}=(\mathbf{A}^{gh})^T$ *and* $\mathbf{B}^{hg}=(\mathbf{B}^{gh})^T$ *are*

$$\mathsf{A}_{iJ}^{hg}=\frac12\sum_{q=1}^{Q_J}\mathsf{v}_q^*(Q_J)\,\langle\widetilde{g}'_{\sigma_q^*(Q_J),X_J},(\widetilde{\chi}_i^h)'\rangle_{L^2(0,L)}$$
$$-\sum_{q=1}^{Q_J}\mathsf{v}_q^*(Q_I)\sum_{K=1}^{M}Z_K\,\widetilde{g}_{\sigma_q^*(Q_J),X_J}(X_K)\widetilde{\chi}_i^h(X_K)\tag{4.52a}$$

$$\mathsf{B}_{iJ}^{hg}=\sum_{q=1}^{Q_J}\mathsf{v}_q^*(Q_I)\,\langle\widetilde{g}_{\sigma_q^*(Q_J),X_J},\widetilde{\chi}_i^h\rangle_{L^2(0,L)}\tag{4.52b}$$

for $(i,J)\in\{0,\dots,N-1\}\times\{1,\dots,M\}$.

Proof. In the discrete variational formulation $\mathfrak{a}(u_{h,g},v_{h,g})=E_{h,g}\mathfrak{b}(u_{h,g},v_{h,g})$ for all $v_{h,g}\in\mathscr{V}_{h,g}$, we specify $v_{h,g}=\widetilde{\chi}_i^h$ for a fixed $i\in\{0,1,\le,N-1\}$. This yields the first N rows of (4.49a) with the matrices (4.50). Specifying $v_{h,g}=\widetilde{\mathrm{CG}}_I$ for a fixed $I\in\{1,\dots,M\}$ in the discrete variational formulation, we end up with the last M rows of (4.49a). □

In order to compute the four collections of scalar products over $[0,L]$ that arise in (4.51)–(4.52), namely,

$$\langle\widetilde{g}_{\sigma_q^*(Q_J),X_J},\widetilde{g}_{\sigma_p^*(Q_I),X_I}\rangle_{L^2(0,L)},\quad\langle\widetilde{g}_{\sigma_q^*(Q_J),X_J},\widetilde{\chi}_i^h\rangle_{L^2(0,L)},$$
$$\langle\widetilde{g}'_{\sigma_q^*(Q_J),X_J},\widetilde{g}'_{\sigma_p^*(Q_I),X_I}\rangle_{L^2(0,L)},\quad\langle\widetilde{g}'_{\sigma_q^*(Q_J),X_J},(\widetilde{\chi}_i^h)'\rangle_{L^2(0,L)},\tag{4.53}$$

the first step is to convert these into scalar products over $\mathbb{R}$.

Lemma 4.3. *For all* $(\sigma,\tau)\in(\mathbb{R}_+^*)^2$ *and* $(X,Y)\in\mathbb{R}^2$*, we have*

$$\int_0^L\widetilde{g}_{\sigma,X}\,\widetilde{g}_{\tau,Y}=\int_{\mathbb{R}}g_{\sigma,X}\,\widetilde{g}_{\tau,Y}\qquad\int_0^L\widetilde{g}'_{\sigma,X}\widetilde{g}'_{\tau,Y}=\int_{\mathbb{R}}g'_{\sigma,X}\,\widetilde{g}'_{\tau,Y},\tag{4.54a}$$

$$\int_0^L\widetilde{g}_{\sigma,X}\,\widetilde{\chi}_i^h=\int_{\mathbb{R}}g_{\sigma,X}\,\widetilde{\chi}_i^h\qquad\int_0^L\widetilde{g}'_{\sigma,X}\,(\widetilde{\chi}_i^h)'=-\int_{\mathbb{R}}g''_{\sigma,X}\,\widetilde{\chi}_i^h.\tag{4.54b}$$

Regarded as being defined over $\mathbb{R}$, the tilded functions in the right-hand sides do not belong to $L^2(\mathbb{R})$. Notwithstanding, the integrals do converge thanks to the rapid decay of $g_{\sigma,X}$ and $g''_{\sigma,X}$.

Proof. Let us prove the first identity, the remaining ones being similar. By switching the orders of summation and integration (justified by uniform convergence), we have

$$\int_0^L\widetilde{g}_{\sigma,X}\,\widetilde{g}_{\tau,Y}=\int_0^L\sum_{n\in\mathbb{Z}}g_{\sigma,X}(\cdot+nL)\,\widetilde{g}_{\tau,Y}$$
$$=\sum_{n\in\mathbb{Z}}\int_0^L g_{\sigma,X}(\cdot+nL)\,\widetilde{g}_{\tau,Y}=\sum_{n\in\mathbb{Z}}\int_{nL}^{(n+1)L}g_{\sigma,X}\,\widetilde{g}_{\tau,Y}(\cdot-nL),$$

after a change of variable for each integral. Since $\widetilde{g}_{\tau,Y}$ is L-periodic, $\widetilde{g}_{\tau,Y}(\cdot-nL)=\widetilde{g}_{\tau,Y}$ and therefore the last sum is none other than $\int_{\mathbb{R}}g_{\sigma,X}\,\widetilde{g}_{\tau,Y}$. The last identity of (4.54) results from an additional integration by parts. The boundary terms vanish at infinity. □

The second step towards evaluating the dot products (4.53) involves an approximation. Expressing

$$\int_{\mathbb{R}} g_{\sigma,X}\,\widetilde{g}_{\tau,Y} = \int_{\mathbb{R}} g_{\sigma,X} \sum_{n\in\mathbb{Z}} g_{\tau,Y}(\cdot+nL) = \sum_{n\in\mathbb{Z}} \int_{\mathbb{R}} g_{\sigma,X}\, g_{\tau,Y}(\cdot+nL),$$
$$\int_{\mathbb{R}} g_{\sigma,X}\,\widetilde{\chi}_i^h = \int_{\mathbb{R}} g_{\sigma,X} \sum_{n\in\mathbb{Z}} \chi_i^h(\cdot+nL) \;\; = \sum_{n\in\mathbb{Z}} \int_{\mathbb{R}} g_{\sigma,X}\, \chi_i^h(\cdot+nL),$$

and two similar quantities with $g''_{\sigma,X}$, we end up with

$$\langle \widetilde{g}_{\sigma,X}, \widetilde{g}_{\tau,Y}\rangle_{L^2(0,L)} = \sum_{n\in\mathbb{Z}} \langle g_{\sigma,X}, g_{\tau,Y}(\cdot+nL)\rangle_{L^2(\mathbb{R})}, \tag{4.55a}$$
$$\langle \widetilde{g}_{\sigma,X}, \widetilde{\chi}_i^h\rangle_{L^2(0,L)} = \sum_{n\in\mathbb{Z}} \langle g_{\sigma,X}, \chi_i^h(\cdot+nL)\rangle_{L^2(\mathbb{R})}, \tag{4.55b}$$
$$\langle \widetilde{g}'_{\sigma,X}, \widetilde{g}'_{\tau,Y}\rangle_{L^2(0,L)} = \sum_{n\in\mathbb{Z}} \langle g'_{\sigma,X}, g'_{\tau,Y}(\cdot+nL)\rangle_{L^2(\mathbb{R})}, \tag{4.55c}$$
$$-\langle \widetilde{g}'_{\sigma,X}, (\widetilde{\chi}_i^h)'\rangle_{L^2(0,L)} = \sum_{n\in\mathbb{Z}} \langle g''_{\sigma,X}, \chi_i^h(\cdot+nL)\rangle_{L^2(\mathbb{R})}. \tag{4.55d}$$

The idea is now to replace each infinite sum in the right-hand sides by a well-chosen finite sum. By "well-chosen" we mean that the number of terms selected in the finite sum can be easily determined so as to control the error within a prescribed threshold. Let us sketch out the general principle for(4.55a) and (4.55b). Consider the approximations

$$\langle \widetilde{g}_{\sigma,X}, \widetilde{g}_{\tau,Y}\rangle_{L^2(0,L)} \approx \sum_{n\in\mathcal{J}(\sigma,X,\tau,Y,L;\epsilon)} \langle g_{\sigma,X}, g_{\tau,Y}(\cdot+nL)\rangle_{L^2(\mathbb{R})}, \tag{4.56a}$$
$$\langle \widetilde{g}_{\sigma,X}, \widetilde{\chi}_i^h\rangle_{L^2(0,L)} \approx \sum_{n\in\mathcal{I}(\sigma,X,i,h,m,L;\epsilon)} \langle g_{\sigma,X}, \chi_i^h(\cdot+nL)\rangle_{L^2(\mathbb{R})}, \tag{4.56b}$$

where the subsets $\mathcal{J},\mathcal{I}\subset\mathbb{Z}$, consisting of consecutive integers, depend on various parameters and an error threshold ϵ. To fill in $\mathcal{J}$, we pick the integers n in increasing order of the distance $|X-Y+nL|$ between the peaks of the two Gaussians. To fill in $\mathcal{I}$, we pick the integers n in increasing order of the distance $|X-(i+\mathsf{M}-1/2)h+nL|$ between the peak of the Gaussian and the "center" of $\operatorname{supp}\chi_i^h(\cdot+nL)$. The number of integers to enter $\mathcal{J}$ must be sufficient so that the error between (4.56a) and (4.55a) be bounded in absolute value by ϵ. Likewise, the number of integers to enter $\mathcal{I}$ must be sufficient so that the error between (4.56b) and (4.55b) be bounded in absolute value by ϵ.

Thanks to the Cauchy-Schwarz inequality, it is possible to devise upper bounds for the absolute value of these errors and to "invert" the bounds to find the subsets $\mathcal{J}$ and $\mathcal{I}$ for a given $\epsilon>0$. Numerical experiments reveal that for $L\gg(2\mathsf{M}-1)h$ (the support of the scaling function is small relatively to the size of the domain), $L\gg\sigma$, $L\gg\tau$ (the Gaussians are narrow relatively to the size of the domain), and a reasonable ϵ, each of the subsets $\mathcal{J}$ and $\mathcal{I}$ often consists of a single element!

4.4.2 Wavelet-Gaussian scalar product

There remains a third step to accomplish before we can claim success in computing the products (4.53). After the first two steps, we are faced with the more elementary dot

products

$$\langle g_{\sigma,X}, g_{\tau,Y}\rangle_{L^2(\mathbb{R})},\ \ \langle g_{\sigma,X}, \chi_i^h\rangle_{L^2(\mathbb{R})},\ \ \langle g'_{\sigma,X}, g'_{\tau,Y}\rangle_{L^2(\mathbb{R})},\ \ \langle g''_{\sigma,X}, \chi_i^h\rangle_{L^2(\mathbb{R})} \tag{4.57}$$

(a shift of $g_{\tau,Y}$ or χ_i^h by $-nL$ does not change the type of the product). Those involving two Gaussians can be carried out analytically. Indeed,

$$\int_{\mathbb{R}} g_{\sigma,X}\, g_{\tau,Y} = \sqrt{\frac{2\sigma\tau}{\sigma^2+\tau^2}}\, \exp\Big\{-\frac{|X-Y|^2}{2(\sigma^2+\tau^2)}\Big\}, \tag{4.58a}$$

$$\int_{\mathbb{R}} g'_{\sigma,X}\, g'_{\tau,Y} = \sqrt{\frac{2\sigma\tau}{(\sigma^2+\tau^2)^3}}\Big[1-\frac{|X-Y|^2}{\sigma^2+\tau^2}\Big]\exp\Big\{-\frac{|X-Y|^2}{2(\sigma^2+\tau^2)}\Big\}. \tag{4.58b}$$

As for those involving a Gaussian and a scaling function, since Daubechies scaling functions do not have a closed-form expression, a numerical procedure is mandatory. Inspired from BigDFT, this procedure is based on a combination of quadrature rules and the two-scale relation.

Quadrature rules alone

Throughout the remainder of section §4.4.2, we shall omit the subscript $L^2(\mathbb{R})$ in the dot products $\langle\cdot,\cdot\rangle$. To compute the integrals

$$\langle g_{\sigma,X}, \chi_i^h\rangle = \int_{\mathbb{R}} g_{\sigma,X}(x)\chi_i^h(x)\,\mathrm{d}x, \qquad \langle g''_{\sigma,X}, \chi_i^h\rangle = \int_{\mathbb{R}} g''_{\sigma,X}(x)\chi_i^h(x)\,\mathrm{d}x,$$

we perform the change of variable $y = x/h - i$ so as to obtain

$$\langle g_{\sigma,X}, \chi_i^h\rangle = \langle g_{\sigma/h,X/h-i}, \phi\rangle, \qquad \langle g''_{\sigma,X}, \chi_i^h\rangle = \langle g''_{\sigma/h,X/h-i}, \phi\rangle.$$

Taking advantage of the fact that $\operatorname{supp}\phi = [0, 2\mathsf{M}-1]$, we have

$$\langle g_{\sigma/h,X/h-i}, \phi\rangle = \int_0^{2\mathsf{M}-1} g_{\sigma/h,X/h-i}(y)\phi(y)\,\mathrm{d}y, \tag{4.59a}$$

$$\langle g''_{\sigma/h,X/h-i}, \phi\rangle = \int_0^{2\mathsf{M}-1} g''_{\sigma/h,X/h-i}(y)\phi(y)\,\mathrm{d}y. \tag{4.59b}$$

This leads us to consider the more general problem of evaluating a product

$$\langle f, \phi\rangle = \int_{\mathbb{R}} f(x)\phi(x)\,\mathrm{d}x = \int_0^{2\mathsf{M}-1} f(x)\phi(x)\,\mathrm{d}x, \tag{4.60}$$

where $f \in L^2(\mathbb{R})$ is an infinitely differentiable function in place of $g_{\sigma/h,X/h-i}$ or $g''_{\sigma/h,X/h-i}$.

The first natural idea to approach the integral (4.60) is to resort to a quadrature rule of the general form

$$\langle\!\langle f, \phi\rangle\!\rangle := \sum_{\ell\in\mathbb{Z}} \omega_\ell f(\ell), \tag{4.61}$$

where the weights ω_ℓ are to be adjusted in order to guarantee more or less accuracy. The accuracy of the quadrature rule (4.61) is quantified by the notion of *degree of exactness*, which is defined to be the greatest integer $\mathsf{Q} \in \mathbb{N}$ such that formula (4.61) is exact for all polynomials of degree less than or equal to Q, i.e.,

$$\langle\!\langle x \mapsto x^q, \phi\rangle\!\rangle = \langle x \mapsto x^q, \phi\rangle \qquad \text{for all } \ 0 \le q \le \mathsf{Q}.$$

Let us review the quadrature rules most commonly used in conjunction with wavelets.

1. The *one-point mass*. The approximation

$$\langle\langle f, \phi \rangle\rangle = f(0) \tag{4.62}$$

is obtained by lumping ϕ to the Dirac mass δ_0. In general, the degree of exactness for (4.62) is

$$\mathsf{Q} = 0,$$

unless ϕ has some vanishing moments (except for the zeroth-order moment, of course), in other words unless ϕ is a *Coiflet*. The reader is referred to Daubechies' book [43, §8.2 and §7.4] for more details. We shall not use Coiflets in numerical simulations.

2. The *trapezoidal formula*. The approximation

$$\langle\langle f, \phi \rangle\rangle = \frac{1}{2}\phi(0)f(0) + \sum_{\ell=1}^{2\mathsf{M}-2} \phi(\ell)f(\ell) + \frac{1}{2}\phi(2\mathsf{M}-1)f(2\mathsf{M}-1) = \sum_{\ell=1}^{2\mathsf{M}-2} \phi(\ell)f(\ell) \tag{4.63}$$

comes from cutting the interval $[0, 2\mathsf{M}-1]$ into sub-intervals of length 1 and from applying the trapezoid rule on each sub-interval. The last equality is due to $\phi(0) = \phi(2\mathsf{M}-1) = 0$. In the general case of an arbitrary function ϕ, the degree of exactness for (4.63) is $\mathsf{Q} = 1$. However, if ϕ is a Daubechies scaling function of order M, then Sweldens and Piessens [127] proved that

$$\mathsf{Q} = \mathsf{M} - 1.$$

3. The *magic filter*. Neelov and Goedecker [110] recommend adjusting the weights ω_ℓ, $0 \leq \ell \leq 2\mathsf{M}-1$, in such a way that

$$\mathsf{Q} = 2\mathsf{M} - 1.$$

This is done by solving a $2\mathsf{M} \times 2\mathsf{M}$ Vandermonde linear system expressing exactness of (4.61) for the polynomials $\{1, x, \ldots, x^{2\mathsf{M}-1}\}$. The magic filter also serves many other purposes in BigDFT, notably for the computation of the nonlinear exchange-correlation term in the Kohn-Sham model [59].

The following classical result provides an exact representation, as well as an upper bound, for the quadrature error

$$\mathrm{e}(f) = \langle f, \phi \rangle - \langle\langle f, \phi \rangle\rangle.$$

Theorem 4.3. *If the numerical integration* (4.61) *has degree of exactness* $\mathsf{Q} \geq 0$ *and if* $f^{(\mathsf{Q}+1)} \in C^0(\mathbb{R}) \cap L^\infty(\mathbb{R})$*, then the quadrature error is equal to*

$$\mathrm{e}(f) = \frac{1}{\mathsf{Q}!} \int_0^{2\mathsf{M}-1} \mathfrak{K}(t) f^{(\mathsf{Q}+1)}(t)\, \mathrm{d}t,$$

where

$$\mathfrak{K}(t) := \mathrm{e}(x \mapsto (x-t)_+^{\mathsf{Q}}) \tag{4.64}$$

is the Peano kernel *and* $(x-t)_+ = \max\{x-t, 0\}$*. This error is bounded by*

$$|\mathrm{e}(f)| \leq \frac{1}{\mathsf{Q}!} \|\mathfrak{K}\|_{L^1} \|f^{(\mathsf{Q}+1)}\|_{L^\infty}. \tag{4.65}$$

PROOF. See standard textbooks, e.g., Crouzeix and Mignot [39, §2.2]. In a nutshell, it rests upon the Taylor formula with integral remainder

$$f(x) = f(0) + f'(0)x + \ldots + \frac{f^{(\mathtt{Q})}}{\mathtt{Q}!}x^{\mathtt{Q}} + \frac{1}{\mathtt{Q}!}\int_0^x (x-t)^{\mathtt{Q}} f^{(\mathtt{Q}+1)}(t)\,\mathrm{d}t,$$

to both sides of which the linear functional $\mathrm{e}(\cdot)$ is applied. □

Application of Theorem 4.3 to $f = g_{\sigma/h,X/h-i}$ and $f = g''_{\sigma/h,X/h-i}$ yields upper bounds of the quadrature errors for the two products of interest.

Corollary 4.1. *If the numerical integration* (4.61) *has degree of exactness* $\mathtt{Q} \geq 0$*, then*

$$|\mathrm{e}(g_{\sigma/h,X/h-i})| \leq \frac{1}{\mathtt{Q}!}\|\mathfrak{K}\|_{L^1}\|g_{1,0}^{(\mathtt{Q}+1)}\|_{L^\infty}\left(\frac{h}{\sigma}\right)^{\mathtt{Q}+3/2}, \tag{4.66a}$$

$$|\mathrm{e}(g''_{\sigma/h,X/h-i})| \leq \frac{1}{\mathtt{Q}!}\|\mathfrak{K}\|_{L^1}\|g_{1,0}^{(\mathtt{Q}+3)}\|_{L^\infty}\left(\frac{h}{\sigma}\right)^{\mathtt{Q}+7/2}. \tag{4.66b}$$

PROOF. Setting $f = g_{\sigma/h,X/h-i}$ and carefully taking out all of the factors h/σ, we end up with

$$\|f^{(\mathtt{Q}+1)}\|_{L^\infty} = \|g_{1,0}^{(\mathtt{Q}+1)}\|_{L^\infty}\left(\frac{h}{\sigma}\right)^{\mathtt{Q}+3/2},$$

the extra $1/2$ (added to $\mathtt{Q}+1$) being due to the normalization factor $(h/\sigma)^{1/2}$ in $g_{\sigma/h,X/h-i}$. The same argument holds for the second upper bound. □

Quadrature rules with two-scale relation

Even though we have considered only scaling functions on the coarsest level, it is possible to benefit from the multiresolution ladder (described in §2.1.1) to enhance the quality of the numerical integration. Instead of applying a quadrature rule abruptly on the coarsest level, the idea is to apply the quadrature rule on a finer level $\mathtt{L} \geq 1$, with the hope that the quadrature errors would be divided by $2^{\mathtt{L}(\mathtt{Q}+3/2)}$ in view of Corollary 4.1. Then, the results would be propagated downward via the two-scale relation (2.12).

Let $\phi_{\mathtt{L},k} = 2^{\mathtt{L}/2}\phi(2^{\mathtt{L}}\cdot - k)$, $k \in \mathbb{Z}$, be the scaling functions on level $\mathtt{L}$. The equality

$$\int_{\mathbb{R}} f(x)\phi_{\mathtt{L},k}(x)\,\mathrm{d}x = 2^{-\mathtt{L}/2}\int_0^{2\mathtt{M}-1} f(2^{-\mathtt{L}}(y+k))\phi(y)\,\mathrm{d}y$$

suggests us to define

$$\langle\!\langle f, \phi_{\mathtt{L},k}\rangle\!\rangle_0 := 2^{-\mathtt{L}/2}\langle\!\langle f(2^{-\mathtt{L}}(\cdot + k)), \phi\rangle\!\rangle \tag{4.67}$$

as an approximation of $\langle f, \phi_{\mathtt{L},k}\rangle$. This is what we mean by "applying the quadrature rule on level $\mathtt{L}$." On the other hand, the two-scale relation (2.12) implies

$$\begin{aligned}\phi_{\mathtt{L}-1,k} &= 2^{(\mathtt{L}-1)/2}\,\phi(2^{\mathtt{L}-1}\cdot - k)\\ &= 2^{(\mathtt{L}-1)/2}\,2^{1/2}\sum_{n\in\mathbb{Z}} h_n\phi(2(2^{\mathtt{L}-1}\cdot - k) - n) = \sum_{n\in\mathbb{Z}} h_n\phi_{\mathtt{L},n+2k}.\end{aligned}$$

This suggests us to define

$$\langle\!\langle f, \phi_{\mathtt{L}-1,k}\rangle\!\rangle_1 = \sum_{n\in\mathbb{Z}} h_n\langle\!\langle f, \phi_{\mathtt{L},n+2k}\rangle\!\rangle_0, \tag{4.68}$$

as an approximation of $\langle f, \phi_{\mathsf{L}-1,k}\rangle$. In (4.68), the subscript 1 in the left-hand side reflects the fact that the pieces of information (4.67) on level L have been propagated one level downward. Pursuing the descent process, we define

$$\langle\!\langle f, \phi_{\mathsf{L}-2,k}\rangle\!\rangle_2 = \sum_{n\in\mathbb{Z}} h_n \langle\!\langle f, \phi_{\mathsf{L},n+2k}\rangle\!\rangle_1 \tag{4.69}$$

and so on. At the end, after L stairs, we get

$$\langle\!\langle f, \phi_k\rangle\!\rangle_{\mathsf{L}} = \sum_{n\in\mathbb{Z}} h_n \langle\!\langle f, \phi_{1,n+2k}\rangle\!\rangle_{\mathsf{L}-1}. \tag{4.70}$$

For $k = 0$, the value $\langle\!\langle f, \phi\rangle\!\rangle_{\mathsf{L}}$ is what we refer to as "numerical quadrature with two-scale relation" of $\langle f, \phi\rangle$.

The following result [129] provides an exact representation, as well as an upper bound, for the L-quadrature error

$$\mathrm{e}_{\mathsf{L}}(f) = \langle f, \phi\rangle - \langle\!\langle f, \phi\rangle\!\rangle_{\mathsf{L}}.$$

Theorem 4.4. *If the numerical integration* (4.61) *has degree of exactness* $\mathsf{Q} \geq 0$ *and if* $f^{(\mathsf{Q}+1)} \in C^0(\mathbb{R}) \cap L^\infty(\mathbb{R})$*, then the* L*-quadrature error is equal to*

$$\mathrm{e}_{\mathsf{L}}(f) = \frac{1}{2^{\mathsf{L}(\mathsf{Q}+1)}\mathsf{Q}!} \int_0^{2\mathsf{M}-1} \mathfrak{K}_{\mathsf{L}}(t) f^{(\mathsf{Q}+1)}(t)\,\mathrm{d}t, \tag{4.71}$$

where $\mathfrak{K}_{\mathsf{L}}$ *is the* L*-th iterate of the recursion*

$$\mathfrak{K}_0(t) = \mathfrak{K}(t) \qquad [\textit{defined in } (4.64)],$$
$$\mathfrak{K}_{\mathsf{K}+1}(t) = \sqrt{2} \sum_{n=0}^{2\mathsf{M}-1} h_n \mathfrak{K}_{\mathsf{K}}(2t-n)$$

for $0 \leq \mathsf{K} \leq \mathsf{L}-1$*. This error is bounded by*

$$|\mathrm{e}_{\mathsf{L}}(f)| \leq \left(\frac{1}{2^{\mathsf{Q}+3/2}} \sum_{n=0}^{2\mathsf{M}-1} |h_n|\right)^{\mathsf{L}} \frac{1}{\mathsf{Q}!} \|\mathfrak{K}\|_{L^1} \|f^{(\mathsf{Q}+1)}\|_{L^\infty}. \tag{4.72}$$

Proof. Let us start with $\mathsf{L} = 1$. By (4.70),

$$\langle\!\langle f, \phi\rangle\!\rangle_1 = \sum_{n=0}^{2\mathsf{M}-1} h_n \langle\!\langle f, \phi_{1,n}\rangle\!\rangle = \frac{1}{\sqrt{2}} \sum_{n=0}^{2\mathsf{M}-1} h_n \langle\!\langle f((\cdot+n)/2), \phi\rangle\!\rangle,$$

the last equality resulting from an affine change of variable. For notational conciseness, we introduce the translation operator T_θ, defined as $T_\theta f = f(\cdot - \theta)$ for $\theta \in \mathbb{R}$, and the dilation operator S_ϑ, defined as $S_\vartheta f = f(\vartheta\cdot)$ for $\vartheta > 0$. Then,

$$\langle\!\langle f, \phi\rangle\!\rangle_1 = \frac{1}{\sqrt{2}} \sum_{n=0}^{2\mathsf{M}-1} h_n \langle\!\langle T_{-n} S_{1/2} f, \phi\rangle\!\rangle.$$

Subtracting this to

$$\langle f, \phi\rangle = \frac{1}{\sqrt{2}} \sum_{n=0}^{2\mathsf{M}-1} h_n \langle T_{-n} S_{1/2} f, \phi\rangle,$$

which is due to the refinement equation (2.12), we end up with

$$\mathrm{e}_1(f) = \frac{1}{\sqrt{2}} \sum_{n=0}^{2\mathsf{M}-1} h_n \mathrm{e}(T_{-n} S_{1/2} f), \tag{4.73}$$

where $\mathrm{e}(\cdot)$ denotes the error functional at level 0. By virtue of Theorem 4.3,

$$\mathrm{e}(T_{-n} S_{1/2} f) = \frac{1}{\mathsf{Q}!} \int_0^{2\mathsf{M}-1} \mathfrak{K}(t) (T_{-n} S_{1/2} f)^{(\mathsf{Q}+1)}(t) \, \mathrm{d}t = \frac{1}{\mathsf{Q}!} \int_{\mathbb{R}} \mathfrak{K}(t) (T_{-n} S_{1/2} f)^{(\mathsf{Q}+1)}(t) \, \mathrm{d}t,$$

the last equality being due to $\operatorname{supp} \mathfrak{K} \subset [0, 2\mathsf{M}-1]$. As

$$(T_{-n} S_{1/2} f)^{(\mathsf{Q}+1)}(t) = \frac{1}{2^{\mathsf{Q}+1}} f^{(\mathsf{Q}+1)}\left(\frac{t+n}{2}\right),$$

the previous equality becomes

$$\mathrm{e}(T_{-n} S_{1/2} f) = \frac{1}{2^{\mathsf{Q}+1} \mathsf{Q}!} \int_{\mathbb{R}} \mathfrak{K}(t) f^{(\mathsf{Q}+1)}\left(\frac{t+n}{2}\right) \mathrm{d}t = \frac{1}{2^{\mathsf{Q}} \mathsf{Q}!} \int_{\mathbb{R}} \mathfrak{K}(2s-n) f^{(\mathsf{Q}+1)}(s) \, \mathrm{d}s.$$

Plugging this into (4.73), we obtain

$$\mathrm{e}_1(f) = \frac{1}{2^{\mathsf{Q}+1} \mathsf{Q}!} \int_{\mathbb{R}} \mathfrak{K}_1(t) f^{(\mathsf{Q}+1)}(t) \, \mathrm{d}t,$$

with

$$\mathfrak{K}_1(t) = \sqrt{2} \sum_{n=0}^{2\mathsf{M}-1} h_n \mathfrak{K}(2t-n). \tag{4.74}$$

By induction on L, we prove the exact representation (4.71). Returning to $\mathsf{L} = 1$, we remark that

$$|\mathrm{e}_1(f)| \leq \frac{1}{2^{\mathsf{Q}+1} \mathsf{Q}!} \|\mathfrak{K}_1\|_{L^1} \|f^{(\mathsf{Q}+1)}\|_{L^\infty}. \tag{4.75}$$

By the triangle inequality,

$$\|\mathfrak{K}_1\|_{L^1} \leq \sqrt{2} \sum_{n=0}^{2\mathsf{M}-1} |h_n| \, \|\mathfrak{K}(2 \cdot - n)\|_{L^1} = \frac{1}{\sqrt{2}} \left(\sum_{n=0}^{2\mathsf{M}-1} |h_n| \right) \|\mathfrak{K}\|_{L^1}.$$

The chaining of this with (4.75) yields

$$|\mathrm{e}_1(f)| \leq \left(\frac{1}{2^{\mathsf{Q}+3/2}} \sum_{n=0}^{2\mathsf{M}-1} |h_n| \right) \frac{1}{\mathsf{Q}!} \|\mathfrak{K}\|_{L^1} \|f^{(\mathsf{Q}+1)}\|_{L^\infty}.$$

By induction on L, we derive the upper bounds (4.72). □

The quantity

$$\Xi_{\mathsf{Q},\mathsf{M}} = \frac{1}{2^{\mathsf{Q}+3/2}} \sum_{n=0}^{2\mathsf{M}-1} |h_n|, \tag{4.76}$$

whose L-th power appear in the bound (4.72), is the outcome of a competition between two opposing effects:

- a reduction of the quadrature error by the "ideal" factor $2^{\mathsf{Q}+3/2}$ for each higher level involved, thanks to a mesh size twice as small;
- an amplification of the quadrature error by the accumulation of the $2m$ quadrature errors in the linear combination that propagates information from one level to the lower one.

Thus, the L-quadrature method is interesting only if reduction wins, that is, $\Xi_{\mathsf{Q},\mathsf{M}} < 1$.

Proposition 4.8. *If the numerical integration* (4.61) *is the trapezoidal formula* ($\mathsf{Q} = \mathsf{M}-1$) *or the magic filter* ($\mathsf{Q} = 2\mathsf{M}-1$)*, then*

$$\Xi_{\mathsf{Q},\mathsf{M}} < 1.$$

Proof. We first notice that, by the Cauchy-Schwarz inequality,

$$\sum_{n=0}^{2\mathsf{M}-1} |h_n| < \left(\sum_{n=0}^{2\mathsf{M}-1} 1^2 \right)^{1/2} \left(\sum_{n=0}^{2\mathsf{M}-1} |h_n|^2 \right)^{1/2} = \sqrt{2\mathsf{M}},$$

strict inequality being due to the non-collinearity between $(h_0, \ldots, h_{2\mathsf{M}-1})$ and $(1, \ldots, 1)$. If $\mathsf{Q} \geq \mathsf{M}-1$ (trapezoidal formula or magic filter), then $\mathsf{Q}+3/2 \geq \mathsf{M}+1/2$ and

$$\Xi_{\mathsf{Q},\mathsf{M}} < \frac{\sqrt{2\mathsf{M}}}{2^{\mathsf{M}+1/2}} = \frac{\sqrt{\mathsf{M}}}{2^{\mathsf{M}}}.$$

But $\sqrt{\mathsf{M}} < 2^{\mathsf{M}}$ for all $\mathsf{M} \geq 1$. □

For the one-point mass quadrature rule, we do not have the theoretical guarantee that $\Xi_{\mathsf{Q},\mathsf{M}} < 1$ for Daubechies scaling functions, although preliminary experiments [49] testify to a convergence with respect to increasing L. For the numerical simulations, we decided to choose the trapezoidal rule as a good compromise between the theoretical guarantee of convergence with respect to L, the accuracy and the computational cost.

Corollary 4.2. *If the numerical integration* (4.61) *has degree of exactness* $\mathsf{Q} \geq 0$*, then*

$$|\mathrm{e}_{\mathsf{L}}(g_{\sigma/h,X/h-i})| \leq \left(\sum_{n=0}^{2\mathsf{M}-1} |h_n| \right)^{\mathsf{L}} \frac{1}{\mathsf{Q}!} \|\mathfrak{K}\|_{L^1} \|g_{1,0}^{(\mathsf{Q}+1)}\|_{L^\infty} \left(\frac{h}{2^{\mathsf{L}}\sigma} \right)^{\mathsf{Q}+3/2}, \tag{4.77a}$$

$$|\mathrm{e}_{\mathsf{L}}(g''_{\sigma/h,X/h-i})| \leq \left(4\sum_{n=0}^{2\mathsf{M}-1} |h_n| \right)^{\mathsf{L}} \frac{1}{\mathsf{Q}!} \|\mathfrak{K}\|_{L^1} \|g_{1,0}^{(\mathsf{Q}+3)}\|_{L^\infty} \left(\frac{h}{2^{\mathsf{L}}\sigma} \right)^{\mathsf{Q}+7/2}. \tag{4.77b}$$

Proof. Apply Theorem 4.4 to $f = g_{\sigma/h,X/h-i}$ and $f = g''_{\sigma/h,X/h-i}$. □

Example and practicalities

As an illustration of the efficiency of the L-numerical integration (combining quadrature rules and two-scale relation), we consider various approximations $\langle\!\langle g_{\sigma,X}, \phi \rangle\!\rangle_{\mathsf{L}}$ of $\langle g_{\sigma,X}, \phi \rangle$ for

$$X = 0.4, \quad \sigma = 0.05, \quad \mathsf{M} = 4, \quad \mathsf{L} \in \{0, 1, \ldots, 13\}.$$

This is a very narrow Gaussian, relatively to $h = 1$. The shapes of $g_{\sigma,X}$ and ϕ are depicted in the upper panel of Figure 4.18. In the lower panel of Figure 4.18, we plot the

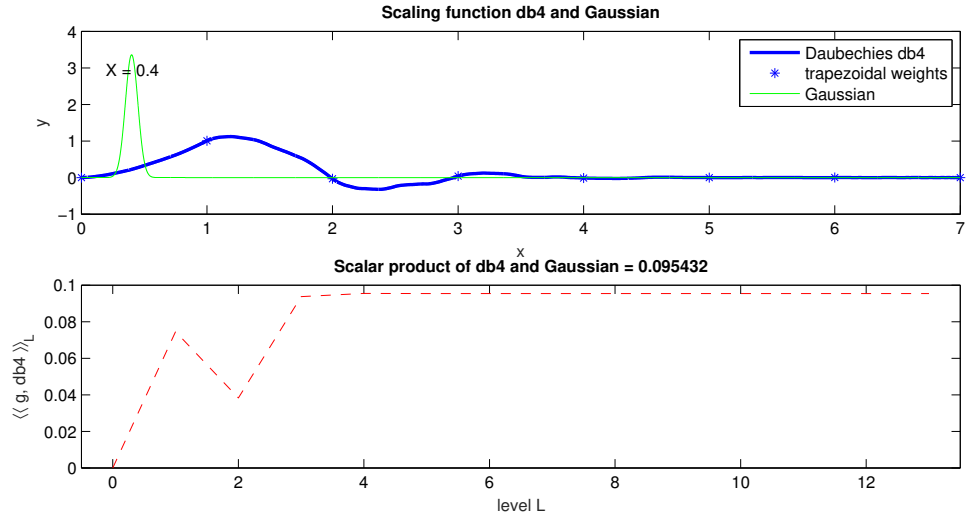

Figure 4.18: Approximate scalar product $\langle\langle g_{\sigma,X}, \phi\rangle\rangle_{\mathsf{L}}$ of a Gaussian and a Daubechies scaling function db4, versus L.

approximate scalar product $\langle\langle g_{\sigma,X}, \phi\rangle\rangle_{\mathsf{L}}$ as a function of the highest level L. The convergence to the exact product $\langle g_{\sigma,X}, \phi\rangle \approx 0.095432$ is quite fast.

An extensive study of the numerical computation of wavelet-Gaussian products can be found in Duchêne's report [49]. More aspects of theoretical error analysis, in particular the dynamical control of L using *a posteriori* error indicators, can be found in Tran's note [129]. In our simulations, for the sake of simplicity, we shall always be using $\mathsf{L} = 11$ since it provides enough accuracy. A main tool for our tests with wavelets and scaling functions is the package Wavelab850, in which the implementation of the low-pass downsampling and many others is available.

4.4.3 Numerical results

Given a scaling function basis on $[0, L]$, for each cusp of u_* we add one contracted Gaussian, which is centered at the cusp position, to the basis. These Gaussians might be the 1-G, 2-G or 3-G contracted Gaussian found in Table 4.1. We are going to see how these very few additional degrees of freedom enhance the resolution. We plot the approximate wave function

$$u_{h,g} = \sum_{j=0}^{N-1} \mathsf{u}_j^{h,g}\, \widetilde{\chi}_j^h + \sum_{J=1}^{M} \mathsf{u}_{N-1+J}^{h,g}\, \widetilde{\mathrm{CG}}_J,$$

then the curves of energy errors, over several mixed bases.

Single-delta potentials

Figure 4.19 plots the exact and Galerkin solutions over a db4 scaling function basis and a mixed basis with the 1-G Gaussian g_{σ^*}, on a mesh of 32 points. The added g_{σ^*} does improve the solution, especially at the cusp (see the cyan curve); but we shall have an even better approximation if we narrow it down by a dilation r: the Gaussian $g_{r\sigma^*}$ for

a random $r = 0.3$ works much better with the db4 basis on the given mesh. It proves that the contracted Gaussians can not be used straight away in mixed bases, and we need further optimizations.

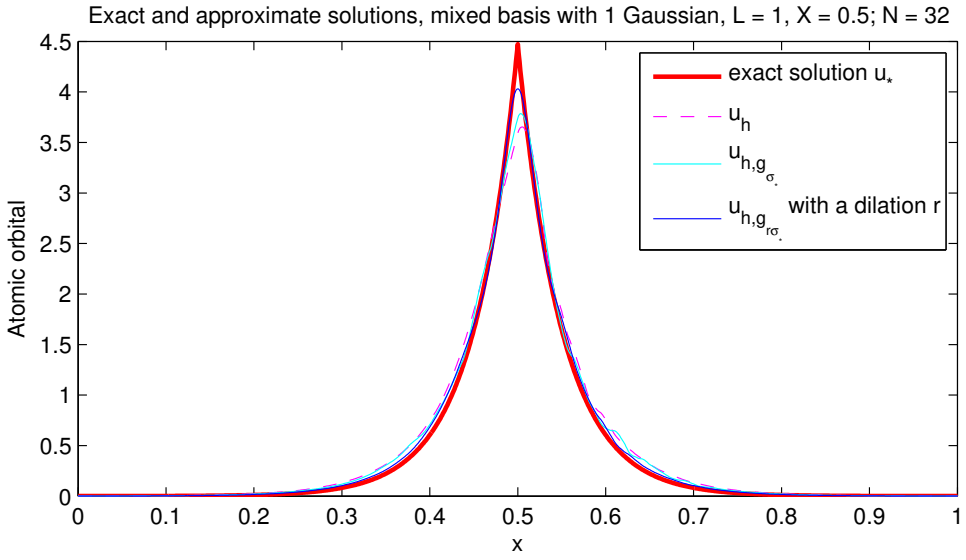

Figure 4.19: Wave functions on periodic domain, single-delta potential, over db4 basis and mixed basis with 1 Gaussian.

Figure 4.20 plots the relative errors on energy $\texttt{err}_h$ over db3, db4, db5 scaling function bases and mixed basis db4 with g_{σ^*}, when N varies:

$$L = 1,\ Z = 20,\ X = 1/2,\ N \in \{16, 32, \ldots, 256\}.$$

When N is large, the curves are of slope (-1); when N is small, the error is roughly divided by 2 for each additional primitive Gaussian. The error level of the Gaussian basis $\{g_{\sigma^*}\}$ is also put in comparison with other error curves. The test shows that the Gaussian works effectively when N is small, but if we refine the mesh several times, the "pre-optimal" Gaussian no longer adapts to the new mesh. This observation holds in the case of mixed bases with 2-G or 3-G contracted Gaussians, as illustrated in Figure 4.21:

$$L = 1,\ Z = 20,\ X = 1/2,\ N \in \{16, 32, \ldots, 2048\}.$$

Double-delta potentials

Given a scaling function basis on $[0, L]$, we enrich it by one contracted Gaussian at each cusp, preferably with the same number of primitives. The two additional elements might be of the 1-G, 2-G or 3-G type found in Table 4.1. So, overall, we only add two additional degrees of freedom.

Figure 4.22 plots the exact and approximate solutions over a db4 scaling function basis, a mixed basis with db4 scaling functions and one contracted Gaussian (1-G, 2-G or 3-G) placed at each cusp.

$$L = 1,\ Z_1 = 20,\ Z_2 = 19,\ X_1 = 3/8,\ X_2 = 1/2,\ N = 32.$$

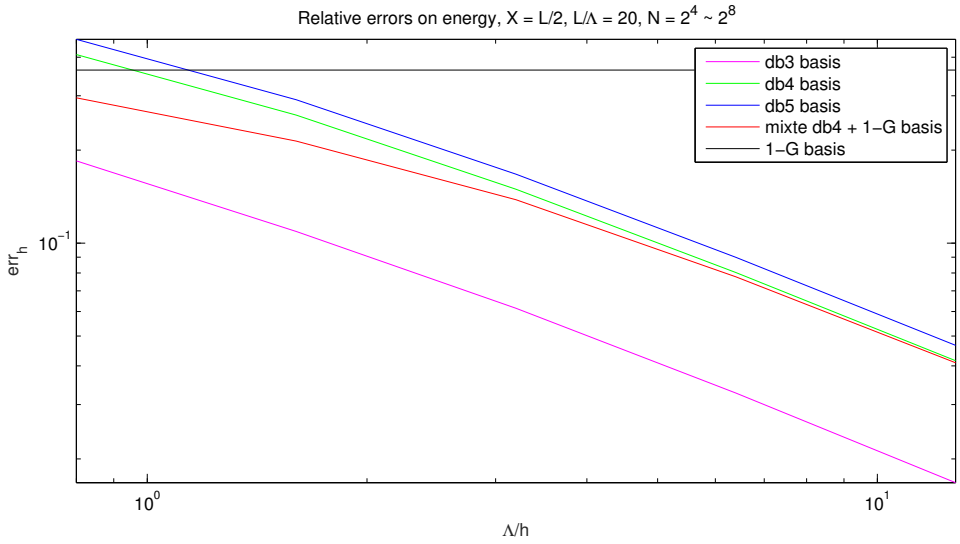

Figure 4.20: Relative error $\texttt{err}_h$ on periodic domain, single-delta potential, when $N = L/h$ is increasing.

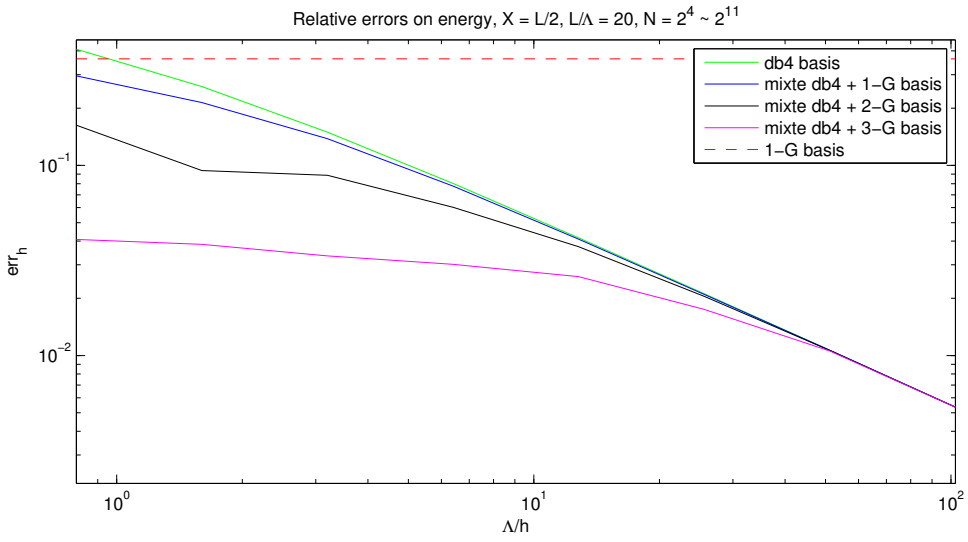

Figure 4.21: Relative error $\texttt{err}_h$ on periodic domain, single-delta potential, when $N = L/h$ is increasing.

Figure 4.23 plots the relative errors on energy $\texttt{err}_h$ over db4 scaling function bases and mixed basis db4 with one contracted Gaussian placed at each cusp (1-G, 2-G or 3-G, the numbers of primitive Gaussians at each cusp are equal), when N varies:

$$L = 1,\ Z_1 = 20,\ Z_2 = 19,\ X_1 = 3/8,\ X_2 = 1/2,\ N \in \{16, 32, \ldots, 2048\}.$$

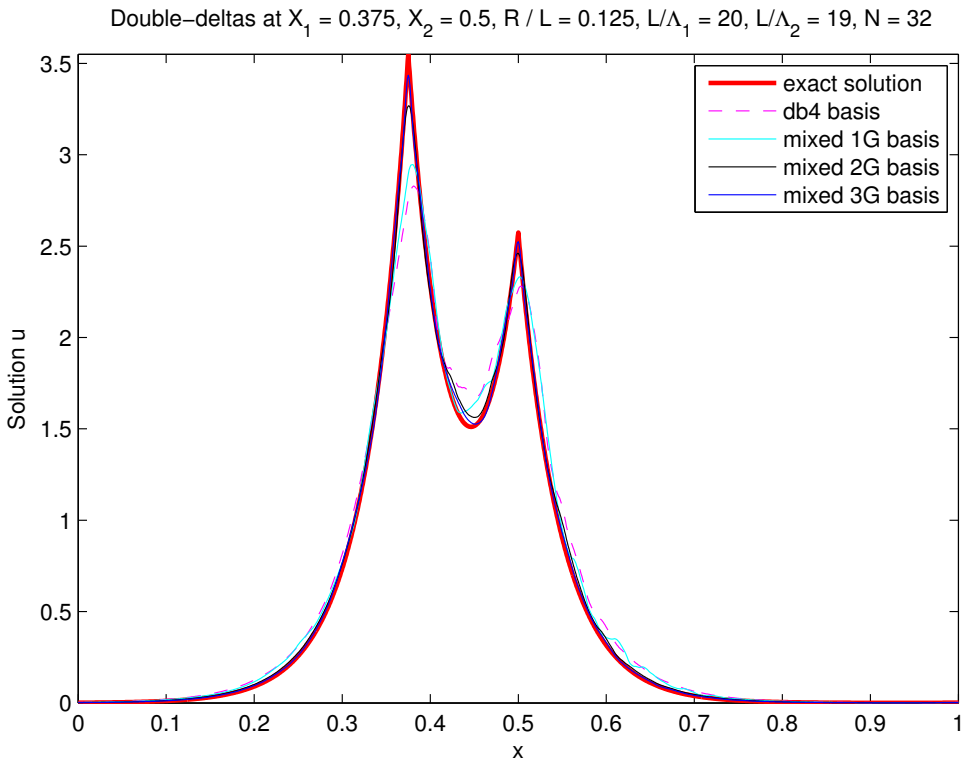

Figure 4.22: Wave functions on periodic domain, double-delta potential, over db4 basis and mixed basis with 1-G, 2-G, 3-G contracted Gaussians.

The error curves have similar forms and properties as in the case of single-delta potentials. In this test, the adding of a contracted Gaussian of Q primitives at each nucleus, $Q = 1, 2, 3$, to the scaling function basis divides the error on the energy by a factor up to 2^Q. This shows the advantage of wavelet-Gaussian mixed bases with respect to pure wavelet bases or pure Gaussian bases. In order to obtain the same error, one additional degree of freedom by the 1-G contracted Gaussian, in the case of single-delta potentials, is equivalent to doubling the dimension of the pure scaling function basis.

Note, however, that even though the pre-optimized contracted Gaussians are favored by quantum chemists for their low costs, they do suffer from two shortcomings:

- the number of primitives Q must be supplied as an input data; the numerical study above has shown that the error is roughly divided by 2 for each additional primitive, but this does not tell us when to stop, because we do not know the initial error.
- the pre-optimization does not take into account the presence of the scaling functions in $\mathscr{V}_h$. In fact, the contracted Gaussians are "optimal" without the scaling functions

and one may legitimately suspect that they do not "get along" well with the scaling functions.

As a consequence, we need another step of optimization to find the “best” Gaussians for the mixed bases. This will be done in chapter §5.

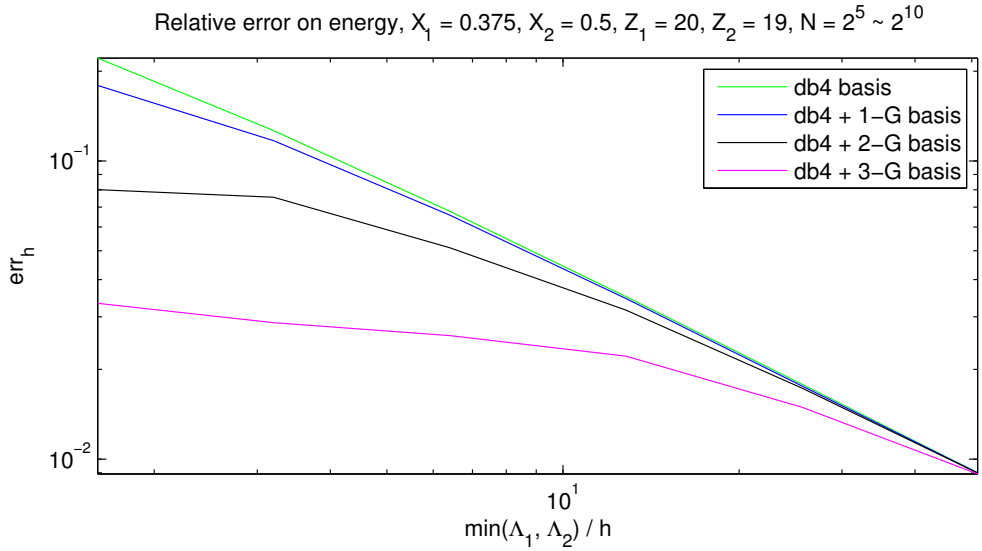

Figure 4.23: Relative error $\texttt{err}_h$ on periodic domain, double-delta potential, over db4 basis and mixed basis with 1-G, 2-G, 3-G contracted Gaussians, when $N = L/h$ is increasing.

Chapter 5

Strategy for an optimal choice of additional Gaussian functions

Contents

Nous exposons une stratégie pour enrichir une base existante par des gaussiennes supplémentaires (éventuellement périodisées). Conformément à l'esprit de l'approximation de Ritz-Galerkin que nous avons évoqué au chapitre §4, cette stratégie repose sur le critère de l'énergie. Il s'agit de régler les paramètres des fonctions à ajouter de sorte à minimiser le niveau d'énergie approché obtenu, ce qui revient aussi à maximiser la diminution d'énergie entre la solution en base initiale et la solution en base augmentée.

Pour ne pas perdre en efficacité, cette vision "idéale" doit être amendée sur deux points, au moyen de deux outils mathématiques. Le premier est qu'au lieu de la vraie diminution d'énergie (qui nécessite un calcul de valeur propre), on se contente d'un estimateur a posteriori *qu'on peut calculer sans avoir à connaître la solution en base agrandie. En adaptant les idées de Dusson et Maday [50], nous montrons que la norme duale du résidu peut servir d'estimateur* a posteriori *pour notre problème.*

Le deuxième outil à mettre en œuvre est destiné à approcher un problème d'optimisation à plusieurs variables par une suite incrémentale de problèmes d'optimisation à une variable. C'est l'algorithme glouton, dont l'utilisation conjointe avec un estimateur d'erreur a été proposée par Prud'homme et al. *[116] dans le cadre des méthodes de bases réduites. Ce que nous proposons ici s'apparente ainsi à une méthode "duale" des bases réduites.*

Notre stratégie finale se décline en deux algorithmes, qui se confondent pour les systèmes à un noyau mais qui diffèrent pour les systèmes à plusieurs noyaux. Dans le premier, c'est le glouton qui dicte l'ordre des noyaux où il faut intervenir. Dans le second, c'est nous qui imposons cet ordre en partant de la charge la plus élevée pour aller vers la plus petite.

5.1 Motivation

At the end of §4, we saw that a mixed basis consisting of scaling functions and a few (possibly contracted) Gaussians centered at the nuclei locations can bring about a significant improvement in accuracy. We now address the core issue of this thesis, which is how to optimally choose the additional Gaussians in a relatively inexpensive way.

We keep the notations from §4. Given the periodic scaling functions $\widetilde{\chi}_{h,i}$, $i = 0, \ldots, N-1$, spanning $\mathscr{V}_h$ in $\mathscr{V} = H^1_{\#}(0, L)$, on a mesh of size h for the discretization of the periodic equation

$$-\frac{1}{2}u'' + \Big(-\sum_{I=1}^{M} Z_I \delta_{X_I} \Big) u = Eu, \tag{5.1a}$$

$$\int_0^L |u|^2 = 1. \tag{5.1b}$$

For each $I \in \{1, \ldots, M\}$, we consider Q_I periodized Gaussians

$$\widetilde{g}_{\sigma_1^{(I)}, X_I}, \widetilde{g}_{\sigma_2^{(I)}, X_I}, \ldots, \widetilde{g}_{\sigma_{Q_I}^{(I)}, X_I}$$

centered at X_I and having standard deviations $\sigma_1^{(I)}, \sigma_2^{(I)}, \ldots, \sigma_{Q_I}^{(I)}$. The mixed basis is then

$$\{\widetilde{\chi}_{h,i}\}_{i=0}^{N-1} \bigcup \{\widetilde{g}_{\sigma_q^{(I)}, X_I}\}_{I \in \{1,\ldots,M\}}^{q \in \{1,\ldots,Q_I\}}, \tag{5.2}$$

and the corresponding subspace is

$$\mathscr{V}_{h,g} = \mathscr{V}_h \oplus \bigoplus_{I=1}^{M} \bigoplus_{q=1}^{Q_I} \Big(\mathbb{R}\, \widetilde{g}_{\sigma_q^{(I)}, X_I} \Big). \tag{5.3}$$

The question is to know how to best select the $(Q_I, \sigma_q^{(I)})$, $I \in \{1, \ldots, M\}$, $q \in \{1, \ldots, Q_I\}$.

Since the notations are somewhat heavy, let $\breve{g}_{\sigma_1}, \ldots, \breve{g}_{\sigma_Q}$ denote the additional periodized Gaussian functions, with

$$Q = \sum_{I=1}^{M} Q_I$$

being their total number. It it implicitly understood that each $\breve{g}_{\sigma_q}$ is centered at some nucleus position X_I. The mixed basis is now written as

$$\{\widetilde{\chi}_{h,i}\}_{i=0}^{N-1} \cup \{\breve{g}_{\sigma_q}\}_{q=1}^{Q}, \tag{5.4}$$

and the corresponding subspace is

$$\mathscr{V}_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_Q}} = \mathscr{V}_h \oplus \bigoplus_{q=1}^{Q} \big(\mathbb{R}\breve{g}_{\sigma_q} \big). \tag{5.5}$$

Let $(u_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_Q}}, E_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_Q}})$ be the Galerkin solution of the problem on $\mathscr{V}_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_Q}}$. Since

$$E_* \leq E_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_Q}} \leq E_h$$

the ideal choice of the $\breve{g}_{\sigma_q}$ amounts to minimize the energy level $E_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_Q}}$, or equivalently, to maximize the energy diminution $E_h - E_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_Q}}$. In other words, the optimal parameters result from the optimization problem

$$(\sigma_1^*, \ldots, \sigma_Q^*) = \arg \max_{(\sigma_1,\ldots,\sigma_Q)\in(\mathbb{R}_+^*)^Q} \left(E_h - E_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_Q}}\right) \tag{5.6}$$

As for the Q_I's (number of Gaussians at X_I), they may either be prescribed in advance according to the user's desire (as is customary in computational chemistry), or be themselves the variables of the optimization problem, subject to some threshold constraint on the energy decay. This point will be clarified in §5.2.

However, there are two difficulties associated with this maximization problem:

- The multivariable nature of the maximization problem makes it hard to solve. In particular, it is notorious that we may be trapped in a local (and not global) maximum. While global optimization methods exist, these are not economical unless we are dealing with one variable.

- To evaluate the value of the objective function, we have to compute $E_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_Q}}$, which requires us to solve an eigenvalue problem for every trial parameters tuple

  $$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_Q) \in (\mathbb{R}_+^*)^Q$$

  and which is of course very costly.

The above challenges are strikingly similar to those encountered in reduced basis techniques [71, 117]. The difference is that instead of reducing, we are enlarging the existing subspace $\mathscr{V}_h$. This observation suggests that we resort to the tools available in reduced basis techniques. More specifically, we will

1. Proceed incrementally by applying the greedy algorithm, which approximates the several-variable optimization problem by a sequence of one-variable optimization problems. Symbolically, the standard deviations of the additional Gaussians will be given by

   $$\begin{aligned}
   \sigma_1^* &= \arg \max_{\sigma_1\in\mathbb{R}_+^*} \left(E_h - E_{h,\breve{g}_{\sigma_1}}\right) \\
   \sigma_2^* &= \arg \max_{\sigma_2\in\mathbb{R}_+^*} \left(E_{h,\breve{g}_{\sigma_1^*}} - E_{h,\breve{g}_{\sigma_1^*},\breve{g}_{\sigma_2}}\right) \\
   &\ldots \\
   \sigma_q^* &= \arg \max_{\sigma_q\in\mathbb{R}_+^*} \left(E_{h,\breve{g}_{\sigma_1^*},\ldots,\breve{g}_{\sigma_{q-1}^*}} - E_{h,\breve{g}_{\sigma_1^*},\ldots,\breve{g}_{\sigma_{q-1}^*},\breve{g}_{\sigma_q}}\right), \quad q \geq 2
   \end{aligned}$$

2. Replace the energy decays $E_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_{q-1}}} - E_{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_{q-1}},\breve{g}_{\sigma_q}}$ by some *a posteriori* estimates

   $$\eta^2_{\{h,\breve{g}_{\sigma_1},\ldots,\breve{g}_{\sigma_{q-1}}\},\breve{g}_{\sigma_q}}$$

   to be designed based on the notion of residues, as advocated by Dusson and Maday [50] in another context.

5.2 Two greedy algorithms

The idea of using a greedy algorithm for reduced basis methods was first proposed by Prud'homme *et al.* [116]. Although it no longer yields the exact optimal solution, the greedy algorithm has proved to be an excellent trade-off between optimality and efficiency. Thanks to the gain in efficiency, it has succeeded in becoming a standard procedure in this area [71, 117]. By construction, it keeps adjoining one new "optimal" vector element to the existing basis at each iteration, until some stopping criterion is satisfied.

For convenience, let us denote the current basis by a formal b, and the corresponding subspace by $\mathscr{V}_b$. At the beginning, we start with a basis of scaling functions on a mesh of size h; we may write $b = \{h\}$. After $q-1$ iterations of the algorithm, the basis becomes

$$b = \{h, \breve{g}_{\sigma_1^*}, \ldots, \breve{g}_{\sigma_{q-1}^*}\} \ .$$

Assume that we have at our disposal an estimate $\eta^2_{b,\breve{g}_{\sigma_q}}$ of the energy decay $E_b - E_{b,\breve{g}_{\sigma_q}}$. The details of this estimate will be provided in the next section §5.3. For the time being, the only feature we need to know about $\eta^2_{b,\breve{g}_{\sigma_q}}$ is that it is an *a posteriori* one, that is, it can be inferred from (u_b, E_b) without having to compute $(u_{b,\breve{g}_{\sigma_q}}, E_{b,\breve{g}_{\sigma_q}})$.

According to the strategy sketched out in the previous section §5.1, let us write down below a first algorithm for choice of additional Gaussians. The algorithm stops when the relative decay of energy

$$\texttt{err} := \frac{E_b - E_{b,\breve{g}_{\sigma_q^*}}}{|E_b|}$$

becomes inferior to a certain tolerance $\epsilon_{tol} > 0$. Besides, we limit the searching for the standard deviations σ_q within a compact interval $\mathcal{I} \subset \mathbb{R}_+^*$ in order to ensure that the optimization problem associated with it is well-defined.

Algorithm 1 Full greedy algorithm

procedure GREEDY1(X_I, Z_I, ϵ_{tol})
  $b := \{h\}$ ▷ $\mathscr{V}_h$: existing basis of scaling function
  $q := 1$ ▷ initialization
  $Q_I = 0$ for $1 \le I \le M$
  repeat
    Compute or retrieve (u_b, E_b)
    $(I_q^*, \sigma_q^*) := \arg \max_{1 \le I_q \le M,\ \sigma_q \in \mathcal{I}} \eta_{b,\widetilde{g}_{\sigma_q, X_{I_q}}}$ ▷ optimization
    Compute $(u_{b,\widetilde{g}_{\sigma_q^*, X_{I_q^*}}}, E_{b,\widetilde{g}_{\sigma_q^*, X_{I_q^*}}})$
    $\texttt{err} := (E_b - E_{b,\widetilde{g}_{\sigma_q^*, X_{I_q^*}}})/|E_b|$ ▷ relative energy decay
    if ($\texttt{err} \ge \epsilon_{tol}$) **then**
      $b := b \cup \{\widetilde{g}_{\sigma_q^*, X_{I_q^*}}\}$ ▷ add new element to basis
      $q := q + 1$
      $Q_{I_q^*} = Q_{I_q^*} + 1$ ▷ number of Gaussians at activated cusp
    end if
  until ($\texttt{err} < \epsilon_{tol}$)
  $Q := q - 1$ ▷ total number of Gaussians
end procedure

A few remarks are in order. At each step q, Algorithm 1 singles out not only a standard deviation σ_q^* but also the cusp I_q^* at which it deems most relevant to act. This accounts for the name "Full greedy." Once it terminates, we can retrieve the values of Q and the Q_I's as outputs, and these numbers depend of course on ϵ_{tol}. This makes Algorithm 1 more interesting than other strategies where the Q_I's must be given by the user. The only parameter the user still has to specify is ϵ_{tol}, the minimal relative energy decay for the step to be acceptable. If ϵ_{tol} is too big, the procedure fails at the first step and no Gaussian is inserted.

The maximization problem involved at each step is in fact a mixed integer nonlinear programming (I_q is a discrete variable and σ_q a continuous variable), which splits into M maximization problems with respect to one continuous variable. In this respect, Algorithm 1 somehow remains expensive. We believe that it is possible to devise a simplified version of Algorithm 1, at each step of which we only have to carry out a single maximization problem with respect to one continuous variable. To do so, we impose the order in which the cusps must be "visited." We claim that this order is that of decreasing charges Z_I. Before justifying this insight, let us describe Algorithm 2 whose name "Partial greedy" expresses its being monitored by some *a priori* knowledge.

Algorithm 2 Partial greedy algorithm

procedure GREEDY2(X_I, Z_I, ϵ_{tol})
  $b := \{h\}$ ▷ $\mathscr{V}_h$: Existing basis of scaling function
  $q := 1$ ▷ Initialization
  $Q_I = 0$ for $1 \le I \le M$
  $\mathbb{M} :=$ sorted elements of $\{1, 2, \ldots, M\}$ in the descending order of $\{Z_1, Z_2, \ldots, Z_M\}$
  $i := 1$
  repeat
    $I := \mathbb{M}(i)$
    repeat
      Compute or retrieve (u_b, E_b)
      $\sigma_q^* := \arg\max_{\sigma_q \in \mathcal{I}} \eta_{b,\widetilde{g}_{\sigma_q, X_I}}$ ▷ one-variable optimization
      Compute $(u_{b,\widetilde{g}_{\sigma_q^*, X_I}}, E_{b,\widetilde{g}_{\sigma_q^*, X_I}})$
      $\mathtt{err} := (E_b - E_{b,\widetilde{g}_{\sigma_q^*, X_I}})/|E_b|$ ▷ relative energy decay
      if ($\mathtt{err} \ge \epsilon_{tol}$) **then**
        $b := b \cup \{\widetilde{g}_{\sigma_q^*, X_I}\}$ ▷ add new element to basis
        $q := q + 1$
        $Q_I := Q_I + 1$ ▷ number of Gaussians at current cusp
      end if
    until ($\mathtt{err} < \epsilon_{tol}$)
    $i := i + 1$
  until (($Q_I = 0$) or ($i > M$))
  $Q := q - 1$ ▷ total number of Gaussians
end procedure

As we have observed many times in the numerical experiments of §3 and §4, the amplitude $u(X_{I_*})$ of the wave function is always highest at the nucleus I_* having the largest charge Z_{I_*}. When this largest charge is sufficiently distinct from the other ones,

the values $u(X_I)$ of the wave function at the remaining nuclei $I \neq I_*$ are negligible—say, one order of magnitude less—compared to $u(X_{I_*})$; this is what we previously called the "ionization" phenomenon. Our intuition is then that the highest cusp gives the largest contribution to the error (in energy). Such a nucleus is the place that must be taken care of most "urgently."

In Algorithm 2, when it is no longer worth staying at the same nucleus (insofar as the energy no longer decreases sufficiently), we have another chance of making the energy decrease sufficiently by adding Gaussians at the "next worst" place. Of course, if our expectation is too high, i.e., if ϵ_{tol} is too big, we may fail at the first attempt on the new nucleus. In such a scenario, no Gaussian is inserted and the overall procedure ends itself.

REMARK 5.1. In practice, we shall be using Algorithm 2. The one-variable optimization at each step is performed by the function *fminbnd* of Matlab, which is based on golden section search and parabolic interpolation.

REMARK 5.2. In reduced basis methods, the larger space is fixed and the smaller space is to be optimized by minimizing the error estimate at each step of the greedy algorithm. In our strategy, the smaller space is fixed and the larger space is to be optimized by maximizing the energy decay estimate at each step of the greedy algorithm.

5.3 *A posteriori* estimate for the energy decay

We now elaborate on the missing ingredient of the above Algorithms, namely, the *a posteriori* estimate $\eta_{b,g}$ for the energy decay $E_b - E_{b,g}$. As in the previous section, $b = \{b_1, \ldots, b_{N_b}\}$ stands for the current basis (which may be a pure or already mixed one), while $g = \{\breve{g}_1, \ldots, \breve{g}_{N_g}\}$ represents here any set of new functions in $\mathscr{V}$ that could potentially be promoted into the basis (presumably Gaussians, but this does not change the abstract derivation of the estimate). Note that, while $N_g = 1$ is the only situation we actually need to consider for Algorithms 1 and 2, the more general situation $N_g \geq 1$ does not create any complication. The space spanned by b is designated by $\mathscr{V}_b$, and the space spanned by $b \cup g$ is designated by $\mathscr{V}_{b,g}$.

5.3.1 Residue and residual norm

Let (u_b, E_b) and $(u_{b,g}, E_{b,g})$ respectively be the Galerkin approximations of the periodic problem (5.1) over $\mathscr{V}_b$ and $\mathscr{V}_{b,g}$. The variational formulations

$$\mathfrak{a}(u_b, v) = E_b\, \mathfrak{b}(u_b, v) \quad \text{for all } v \in \mathscr{V}_b, \tag{5.8a}$$

$$\mathfrak{b}(u_b, u_b) = 1, \tag{5.8b}$$

and

$$\mathfrak{a}(u_{b,g}, v) = E_{b,g}\, \mathfrak{b}(u_{b,g}, v) \quad \text{for all } v \in \mathscr{V}_{b,g}, \tag{5.9a}$$

$$\mathfrak{b}(u_{b,g}, u_{b,g}) = 1, \tag{5.9b}$$

hold true. Following classical approaches in error estimation for finite element methods [3, 132], we define the estimate $\eta_{b,g}$ as the norm of some linear form known as residue.

Definition 5.1. The *residue* of (u_b, E_b) over $\mathscr{V}_{b,g}$, for the equation (5.1), is the linear form that maps any vector $v \in \mathscr{V}_{b,g}$ to the real number

$$\mathrm{Res}_{\mathscr{V}_{b,g}}[u_b, E_b](v) := \mathfrak{a}(u_b, v) - E_b\ \mathfrak{b}(u_b, v) \tag{5.10a}$$

$$= \frac{1}{2}\int_0^L u_b' v' - \sum_{J=1}^{M} Z_J\ u_b(X_J)v(X_J) - E_b \int_0^L u_b v\ . \tag{5.10b}$$

We will also write $\mathrm{Res}[u_b, E_b]$ instead of $\mathrm{Res}_{\mathscr{V}_{b,g}}[u_b, E_b]$ when there is no ambiguity. Remark 5.3 comes directly from the definition (5.10).

Remark 5.3. $\mathrm{Res}[u_b, E_b](v) = 0$ if $v \in \mathscr{V}_b$. Therefore, $\mathscr{V}_b \subset \mathrm{Ker}\ \mathrm{Res}[u_b, E_b]$.

We wish to define the estimate $\eta_{b,g}$ as the dual norm $\|\mathrm{Res}[u_b, E_b]\|_{\mathscr{V}'_{b,g}}$, but to be able to do so, we first have to ensure that the residue is continuous with respect to the norm induced by $\mathscr{V}$ on $\mathscr{V}_{b,g}$.

Proposition 5.1. *There exists a constant $C_b > 0$ (depending on b) such that for all $v \in \mathscr{V}_{b,g}$, we have*

$$|\mathrm{Res}[u_b, E_b](v)| \leq C_b \|v\|_{H^1}. \tag{5.11}$$

Proof. Since $\mathrm{Res}[u_b, E_b](v) = \mathfrak{a}(u_b, v) - E_b\mathfrak{b}(u_b, v)$,

$$\begin{aligned} |\mathrm{Res}[u_b, E_b](v)| &\leq |\mathfrak{a}(u_b, v)| + |E_b||\mathfrak{b}(u_b, v)| \\ &\leq \kappa \|u_b\|_{H^1}\|v\|_{H^1} + |E_b|\|u_b\|_{L^2}\|v\|_{L^2} \end{aligned}$$

by the H^1-continuity of $\mathfrak{a}(\cdot,\cdot)$, established in equation (3.67) of Proposition 3.5, and the Cauchy-Schwarz inequality. Noting that $\|\cdot\|_{L^2} \leq \|\cdot\|_{H^1}$, we end up with

$$|\mathrm{Res}[u_b, E_b](v)| \leq (\kappa + |E_b|)\|u_b\|_{H^1}\|v\|_{H^1}.$$

The constant $C_b = (\kappa + |E_b|)\|u_b\|_{H^1} > 0$ depends on b. □

We are now in a position to define the dual norm of the residue $\mathrm{Res}[u_b, E_b] \in \mathscr{V}'_{b,g}$. This quantity will serve as a measure for the discrepancy between the current-basis solution (u_b, E_b) and the virtually augmented-basis solution $(u_{b,g}, E_{b,g})$.

Definition 5.2. The *a posteriori estimate* between the solutions (u_b, E_b) and $(u_{b,g}, E_{b,g})$ is the dual norm

$$\eta_{b,g} := \|\mathrm{Res}_{\mathscr{V}_{b,g}}[u_b, E_b]\|_{\mathscr{V}'_{b,g}} = \sup_{v \in \mathscr{V}_{b,g}\setminus\{0\}} \frac{|\mathrm{Res}_{\mathscr{V}_{b,g}}[u_b, E_b](v)|}{\|v\|_{H^1}}\ . \tag{5.12}$$

5.3.2 Connection with the energy decay

Our next task, after defining $\eta_{b,g}$, is to show that this estimate faithfully reflects the behavior of the energy decay $E_b - E_{b,g}$ and of the error $\|u_b - u_{b,g}\|_{H^1}$. Usually, this is the difficult part since it is specific to the problem at hand. To achieve this purpose, we borrow and adapt some ideas from Cancès *et al.* [17] and Dusson and Maday [50]. It is worth mentioning that these authors consider a nonlinear Schrödinger equation but require the potential $V(\cdot)$ to be a classical function belonging to some L^p space, which make various quantities easier to bound. The main difficulty with our linear Schrödinger equation lies in

the singularity of the Dirac potentials δ_{X_I}. Fortunately, this can be overcome by invoking functional-analytic results that hold true exclusively for the one-dimensional space.

For the sake of clarity, we proceed by two stages. In the first stage, we take it for granted that some abstract assumptions, called "Standard Hypotheses," are satisfied. These enable us to derive the connection we are looking for (Proposition 5.2 and Theorem 5.1). In the second stage, we prove that the "Standard Hypotheses" are indeed fulfilled in the concrete case of our problem.

Standard Hypotheses

1. *There exists $K > 0$ (independent of b and g) such that for all $(v, w) \in \mathscr{V}_{b,g}^2$,*

$$|\mathfrak{a}(v, w) - E_{b,g}\, \mathfrak{b}(v, w)| \leq K\|v\|_{H^1}\|w\|_{H^1}. \tag{5.13}$$

2. *There exists $\beta_b > 0$ (dependent on b, not on g) such that for all $v \in (u_{b,g})^\perp$ (orthogonality in the L^2-sense) in $\mathscr{V}_{b,g}$,*

$$\beta_b\|v\|_{L^2}^2 \leq \mathfrak{a}(v, v) - E_{b,g}\, \mathfrak{b}(v, v). \tag{5.14}$$

3. *There exists $\gamma_b > 0$ (dependent on b, not on g) such that for $e = u_b - u_{b,g}$,*

$$\gamma_b\|e\|_{H^1}^2 \leq \mathfrak{a}(e, e) - E_{b,g}\, \mathfrak{b}(e, e). \tag{5.15}$$

The second Standard Hypothesis (5.14) is not used in the first part, where only (5.13) and (5.15) are required for proving Proposition 5.2 and Theorem 5.1. It appears as an intermediate step for proving the third Standard Hypothesis (5.15) in the second part. However, we have deliberately conferred the status of Standard Hypothesis on (5.14) in order to highlight the L^2-coercivity property for $\mathfrak{a}(\cdot,\cdot) - E_{b,g}\mathfrak{b}(\cdot,\cdot)$ on a subspace of codimension 1. In fact, $\mathfrak{a}(\cdot,\cdot) - E_{b,g}\mathfrak{b}(\cdot,\cdot)$ can be even shown to be H^1-coercive on $(u_{b,g})^\perp$.

Proposition 5.2. *If the Standard Hypotheses* (5.13)–(5.15) *are satisfied, then*

$$\gamma_b\|u_b - u_{b,g}\|_{H^1}^2 \leq E_b - E_{b,g} \leq K\|u_b - u_{b,g}\|_{H^1}^2. \tag{5.16}$$

Proof. Specifying $v = u_b$ in (5.8) and $v = u_{b,g}$ in (5.9), we get the relations

$$\begin{aligned}\mathfrak{a}(u_b, u_b) &= E_b\mathfrak{b}(u_b, u_b) = E_b,\\ \mathfrak{a}(u_{b,g}, u_{b,g}) &= E_{b,g}\mathfrak{b}(u_{b,g}, u_{b,g}) = E_{b,g}.\end{aligned}$$

Their difference can be transformed as

$$\begin{aligned}E_b - E_{b,g} &= \mathfrak{a}(u_b, u_b) - \mathfrak{a}(u_{b,g}, u_{b,g})\\ &= \mathfrak{a}(u_{b,g} + e, u_{b,g} + e) - \mathfrak{a}(u_{b,g}, u_{b,g})\\ &= \mathfrak{a}(e, e) + 2\mathfrak{a}(u_{b,g}, e)\\ &= \mathfrak{a}(e, e) + 2E_{b,g}\mathfrak{b}(u_{b,g}, e)\\ &= \mathfrak{a}(e, e) - 2E_{b,g}\mathfrak{b}(e, e) + 2E_{b,g}\mathfrak{b}(u_b, e).\end{aligned}$$

Next, we prove that the last term of the right-hand side is half the middle term, i.e., $2\mathfrak{b}(u_b, e) = \mathfrak{b}(e, e)$. Indeed, since $\mathfrak{b}(u_b, u_b) = \mathfrak{b}(u_{b,g}, u_{b,g}) = 1$, we have

$$\begin{aligned} 2\mathfrak{b}(u_b, e) &= 2\mathfrak{b}(u_b, u_b) - 2\mathfrak{b}(u_b, u_{b,g}) \\ &= \mathfrak{b}(u_b, u_b) + \mathfrak{b}(u_{b,g}, u_{b,g}) - 2\mathfrak{b}(u_b, u_{b,g}) \\ &= \mathfrak{b}(u_b - u_{b,g}, u_b - u_{b,g}). \end{aligned}$$

Finally, we obtain

$$E_b - E_{b,g} = \mathfrak{a}(e, e) - E_{b,g}\mathfrak{b}(e, e).$$

Using the first Standard Hypothesis (5.13) with $v = w = e$ and the third one (5.15), we end up with

$$\gamma_b \|e\|_{H^1}^2 \leq E_b - E_{b,g} \leq K\|e\|_{H^1}^2,$$

which is the desired result. □

From Proposition 5.2, we can deduce the equivalence between the energy decrease and the squared estimate.

Theorem 5.1. *If the Standard Hypotheses* (5.13)–(5.15) *are satisfied, then up to a negligible higher-order term in the lower and upper bounds, we have*

$$\gamma_b \|u_{b,g} - u_b\|_{H^1} \leq \eta_{b,g} \leq K\|u_{b,g} - u_b\|_{H^1} \tag{5.17}$$

whenever (u_b, E_b) *and* $(u_{b,g}, E_{b,g})$ *are "close enough" to each other.*

Proof. Suppose that the Standard Hypotheses hold true. Let $v \in \mathcal{V}_{b,g}\backslash\{0\}$. By subtracting the two equalities

$$\begin{aligned} \mathrm{Res}[u_b, E_b](v) &= \mathfrak{a}(u_b, v) - E_b\mathfrak{b}(u_b, v), \\ 0 &= \mathfrak{a}(u_{b,g}, v) - E_{b,g}\mathfrak{b}(u_{b,g}, v), \end{aligned}$$

we obtain

$$\begin{aligned} \mathrm{Res}[u_b, E_b](v) &= \mathfrak{a}(u_b - u_{b,g}, v) - E_{b,g}\ \mathfrak{b}(u_b - u_{b,g}, v) + (E_{b,g} - E_b)\mathfrak{b}(u_b, v) \\ &= \mathfrak{a}(e, v) - E_{b,g}\mathfrak{b}(e, v) + (E_{b,g} - E_b)\mathfrak{b}(u_b, v). \end{aligned} \tag{5.18}$$

The triangle inequality, the first Standard Hypothesis (5.13) and the Cauchy-Schwarz inequality lead to

$$\begin{aligned} |\mathrm{Res}[u_b, E_b](v)| &\leq K\|e\|_{H^1}\|v\|_{H^1} + (E_{b,g} - E_b)\|u_b\|_{L^2}\|v\|_{L^2} \\ &\leq K\|e\|_{H^1}\|v\|_{H^1} + (E_b - E_{b,g})\|v\|_{H^1}, \end{aligned}$$

the second line being due to $\|u_b\|_{L^2} = 1$ and $\|v\|_{L^2} \leq \|v\|_{H^1}$. Dividing the inequality by $\|v\|_{H^1}$ and passing to the supremum in v, we end up with

$$\eta_{b,g} \leq K\|e\|_{H^1} + (E_b - E_{b,g}).$$

On the grounds of the assumed closeness between (u_b, E_b) and $(u_{b,g}, E_{b,g})$, $\|e\|_{H^1}$ and $E_b - E_{b,g}$ are small. Since $E_b - E_{b,g}$ is equivalent to $\|e\|_{H^1}^2$ (Proposition 5.2), this higher-order term can be "omitted" from the sum with $\|e\|_{H^1}$. That leads to $\eta_{b,g} \leq K\|e\|_{H^1}$.

We are now going to prove the other inequality. Specifying $v = e$ in (5.18) and invoking the third Standard Hypothesis (5.15), we arrive at

$$\begin{aligned}\mathrm{Res}[u_b, E_b](e) &= (\mathfrak{a}(e,e) - E_{b,g}\ \mathfrak{b}(e,e)) + (E_{b,g} - E_b)\mathfrak{b}(u_b, e)\\ &\geq \gamma_b \|e\|_{H^1}^2 - (E_b - E_{b,g})\ |\mathfrak{b}(u_b, e)|\\ &\geq \gamma_b \|e\|_{H^1}^2 - (E_b - E_{b,g})\ \|e\|_{H^1}.\end{aligned}$$

Upon dividing by $\|e\|_{H^1}$ and passing to the supremum,

$$\eta_{b,g} \geq \gamma_b \|e\|_{H^1} - (E_b - E_{b,g}).$$

As before, the second-order term $E_b - E_{b,g}$ can be neglected when the solutions are close to each other, so that $\eta_{b,g} \geq \gamma_b \|e\|_{H^1}$. □

Corollary 5.1. *If the Standard Hypotheses* (5.13)–(5.15) *are satisfied, then*

$$\frac{\gamma_b}{K^2}\,\eta_{b,g}^2 \leq E_b - E_{b,g} \leq \frac{K}{\gamma_b^2}\,\eta_{b,g}^2 \tag{5.19}$$

whenever (u_b, E_b) *and* $(u_{b,g}, E_{b,g})$ *are "close enough" to each other.*

Proof. Taking the square of (5.17), we end up with

$$\gamma_b^2 \|u_{b,g} - u_b\|_{H^1}^2 \leq \eta_{b,g}^2 \leq K^2 \|u_{b,g} - u_b\|_{H^1}^2.$$

Combining this with (5.16)

$$\gamma_b \|u_b - u_{b,g}\|_{H^1}^2 \leq E_b - E_{b,g} \leq K \|u_b - u_{b,g}\|_{H^1}^2$$

yields (5.19). □

It results from Corollary 5.1 that $\eta_{b,g}^2$ can serve as an estimate for $E_b - E_{b,g}$ when g ranges over some trial space. The independency of the constants γ_b/K^2 and K/γ_b^2 with respect to g testifies to the consistency of this estimate. It provides legitimacy to the maximization of $\eta_{b,g}^2$ with respect to all possible candidates g.

We will now enter the second stage of exposition and strive to verify that our problem does comply with the Standard Hypotheses (5.13)–(5.15) under some mild assumption on the current-basis solution.

Proposition 5.3. *Let*

$$E^{(2)} = \inf_{\substack{v \in \mathscr{V} \\ v \perp u_*}} \frac{\mathfrak{a}(v,v)}{\mathfrak{b}(v,v)} \tag{5.20}$$

be the "second eigenvalue" of the continuous problem on $\mathscr{V}$*. If*

$$E_b =: E_b^{(1)} < \min\{E^{(2)}, 0\}, \tag{5.21}$$

then the three Standard Hypotheses (5.13)–(5.15) *are satisfied.*

Before proving this Proposition, let us comment on assumption (5.21). As a consequence of Theorem 3.14, $E_* =: E^{(1)} < 0$. According to Corollary 3.6, $E^{(1)} < E^{(2)}$. Hence, $E^{(1)} < \min\{E^{(2)}, 0\}$. If $E_b =: E_b^{(1)}$ is the Galerkin approximation for $E_* = E^{(1)}$ on $\mathscr{V}_b$, then $E_b \geq E_*$ but it may happen that the approximation is really "bad," to such an extent that $E_b \geq \min\{E^{(2)}, 0\}$. In this light, condition (5.21) amounts to saying that the approximation in the current basis b must not be "too far" from the exact solution. If this closeness is guaranteed for the first step of the greedy algorithm ($b = h$), it will be automatically preserved as the basis is enlarged.

Lemma 5.1. *Under assumption* (5.21)*, there exists* $W > 0$ *(independent of* b *and* g*) such that for all* $v \in \mathscr{V}_{b,g}$*, we have*

$$\mathfrak{a}(v,v) - E_{b,g}\mathfrak{b}(v,v) \geq \frac{1}{4}\|v\|_{H^1}^2 - W\|v\|_{L^2}^2. \tag{5.22}$$

Proof. Let $v \in \mathscr{V}_{b,g}$. By equation (3.68) of Proposition 3.5,

$$\mathfrak{a}(v,v) \geq \frac{1}{4}\|v\|_{H^1}^2 - \Theta\|v\|_{L^2}^2,$$

where $\Theta > 0$ does not depend on b or g. This results in

$$\begin{aligned}\mathfrak{a}(v,v) - E_{b,g}\mathfrak{b}(v,v) &\geq \frac{1}{4}\|v\|_{H^1}^2 - \Theta\|v\|_{L^2}^2 - E_{b,g}\|v\|_{L^2}^2 \\ &\geq \frac{1}{4}\|v\|_{H^1}^2 - (\Theta + |E_{b,g}|)\|v\|_{L^2}^2.\end{aligned}$$

In view of assumption (5.21), $E_* \leq E_{b,g} \leq E_b < 0$, so that $|E_{b,g}| \leq |E_*|$. From this, we deduce that inequality (5.22) holds with the constant

$$W = \Theta + |E_*|, \tag{5.23}$$

which is independent of b and g. □

Proof of Proposition 5.3. Let $(v,w) \in \mathscr{V}_{b,g}^2$. By the triangle inequality,

$$\begin{aligned}|\mathfrak{a}(v,w) - E_{b,g}\mathfrak{b}(v,w)| &\leq |\mathfrak{a}(v,w)| + |E_{b,g}||\mathfrak{b}(v,w) \\ &\leq \kappa\|v\|_{H^1}\|w\|_{H^1} + |E_{b,g}|\|v\|_{L^2}\|w\|_{L^2}\end{aligned}$$

by the H^1-continuity of $\mathfrak{a}(\cdot,\cdot)$, established in equation (3.67) of Proposition 3.5. Noting that $\|\cdot\|_{L^2} \leq \|\cdot\|_{H^1}$, we have

$$|\mathfrak{a}(v,w) - E_{b,g}\mathfrak{b}(v,w)| \leq (\kappa + |E_{b,g}|)\|v\|_{H^1}\|w\|_{H^1}.$$

In view of assumption (5.21), $E_* \leq E_{b,g} \leq E_b < 0$, so that $|E_{b,g}| \leq |E_*|$. From this, we deduce that (5.13) holds with the constant

$$K = \kappa + |E_*|, \tag{5.24}$$

which does not depend on b or g.

To prove the second Standard Hypothesis (5.14), let

$$E_{b,g}^{(2)} = \inf_{\substack{v \in \mathscr{V}_{b,g} \\ v \perp u_{b,g}}} \frac{\mathfrak{a}(v,v)}{\mathfrak{b}(v,v)}.$$

be the "second eigenvalue" of the discrete problem on $\mathscr{V}_{b,g}$. From linear algebra, we know that this second eigenvalue can also be characterized by the Courant-Fischer principle as

$$E_{b,g}^{(2)} = \inf_{\substack{\mathscr{W}\subset\mathscr{V}_{b,g}\\ \dim\mathscr{W}=2}} \max_{w\in\mathscr{W}} \frac{\mathfrak{a}(w,w)}{\mathfrak{b}(w,w)}.$$

Compared to that of the continuous second eigenvalue[1]

$$E^{(2)} = \inf_{\substack{\mathscr{W}\subset\mathscr{V}\\ \dim\mathscr{W}=2}} \max_{w\in\mathscr{W}} \frac{\mathfrak{a}(w,w)}{\mathfrak{b}(w,w)}, \tag{5.25}$$

it is plain that $E^{(2)} \le E_{b,g}^{(2)}$. As a result, for all $v \in (u_{b,g})^\perp$,

$$\mathfrak{a}(v,v) - E_{b,g}\mathfrak{b}(v,v) \ge (E_{b,g}^{(2)} - E_{b,g}^{(1)})\,\mathfrak{b}(v,v) \ge (E^{(2)} - E_b^{(1)})\|v\|_{L^2}^2. \tag{5.26}$$

The constant $\beta_b = \lambda^{(2)} - \lambda_b^{(1)}$ is positive by assumption (5.21) and depends on b only.

Finally, we prove the third Standard Hypothesis (5.15). By expanding the bilinear forms below, we get

$$\begin{aligned}
\mathfrak{a}(e,e) - E_{b,g}\mathfrak{b}(e,e) &= \mathfrak{a}(u_b - u_{b,g}, u_b - u_{b,g}) - E_{b,g}\mathfrak{b}(u_b - u_{b,g}, u_b - u_{b,g})\\
&= \mathfrak{a}(u_b,u_b) - E_{b,g}\mathfrak{b}(u_b,u_b) + \mathfrak{a}(u_{b,g},u_{b,g}) - E_{b,g}\mathfrak{b}(u_{b,g},u_{b,g})\\
&\quad - 2(\mathfrak{a}(u_{b,g},u_b) - E_{b,g}\mathfrak{b}(u_{b,g},u_b))\\
&= \mathfrak{a}(u_b,u_b) - E_{b,g}\mathfrak{b}(u_b,u_b) + 0 - 2\cdot 0
\end{aligned} \tag{5.27}$$

thanks to the variational formulation on $\mathscr{V}_{b,g}$. For any $v \in \mathscr{V}_{b,g}$, decompose $v = v_1 u_{b,g} + w$, with $w \in (u_{b,g})^\perp$. Then, $\|v\|_{L^2}^2 = v_1^2 + \|w\|_{L^2}^2$ and we obtain

$$\begin{aligned}
\mathfrak{a}(v,v) - E_{b,g}\mathfrak{b}(v,v) &= \mathfrak{a}(v_1u_{b,g}+w, v_1u_{b,g}+w) - E_{b,g}\mathfrak{b}(v_1u_{b,g}+w, v_1u_{b,g}+w)\\
&= v_1^2(\mathfrak{a}(u_{b,g},u_{b,g}) - E_{b,g}\mathfrak{b}(u_{b,g},u_{b,g})) + \mathfrak{a}(w,w) - E_{b,g}\mathfrak{b}(w,w)\\
&\quad + 2v_1(\mathfrak{a}(u_{b,g},w) - E_{b,g}\mathfrak{b}(u_{b,g},w))\\
&\ge 0 + \beta_b\|w\|_{L^2}^2 + 0
\end{aligned}$$

due to second Standard Hypothesis (5.14) for $w \in (u_{b,g})^\perp$. So

$$\mathfrak{a}(v,v) - E_{b,g}\mathfrak{b}(v,v) \ge \beta_b\|w\|_{L^2}^2 = \beta_b(\|v\|_{L^2}^2 - v_1^2).$$

Take $v = u_b$ in this inequality, then $v_1 = \langle u_b, u_{b,g}\rangle_{L^2}$ and

$$\mathfrak{a}(u_b,u_b) - E_{b,g}\mathfrak{b}(u_b,u_b) \ge \beta_b(\|u_b\|_{L^2}^2 - |\langle u_b, u_{b,g}\rangle_{L^2}|^2).$$

According to equality (5.27),

$$\begin{aligned}
\mathfrak{a}(e,e) - E_{b,g}\mathfrak{b}(e,e) &= \mathfrak{a}(u_b,u_b) - E_{b,g}\mathfrak{b}(u_b,u_b)\\
&\ge \beta_b(\|u_b\|_{L^2}^2 - |\langle u_b, u_{b,g}\rangle_{L^2}|^2)\\
&\ge \beta_b(\|u_b\|_{L^2}^2 - |\langle u_b, u_{b,g}\rangle_{L^2}|)
\end{aligned}$$

[1]This may not seem obvious in an infinite dimensional space, but can be proven as follows: if $\dim\mathscr{W} = 2$, then $\mathscr{W} \cap (u_*)^\perp \neq \{0\}$. Take a vector $w \neq 0$ in this intersection and compare the right-hand sides of (5.25) and (5.20).

because $|\langle u_b, u_{b,g}\rangle_{L^2}| \leq \|u_b\|_{L^2}\|u_{b,g}\|_{L^2} = 1$. Since both $\pm u_b$ and $\pm u_{b,g}$ satisfy their respective variational formulations, we can choose the signs of u_b and $u_{b,g}$ in such a way that $\langle u_b, u_{b,g}\rangle_{L^2} \geq 0$. It is then possible to drop the absolute value to obtain

$$\begin{aligned}\mathfrak{a}(e,e) - E_{b,g}\mathfrak{b}(e,e) &\geq \beta_b\{\|u_b\|_{L^2}^2 - \langle u_b, u_{b,g}\rangle_{L^2}\}\\ &\geq \frac{\beta_b}{2}\{\|u_b\|_{L^2}^2 + \|u_{b,g}\|_{L^2}^2 - 2\langle u_b, u_{b,g}\rangle_{L^2}\}\\ &\geq \frac{\beta_b}{2}\|u_b - u_{b,g}\|_{L^2}^2 = \frac{\beta_b}{2}\|e\|_{L^2}^2. \end{aligned} \tag{5.28}$$

Let us combine this with

$$\mathfrak{a}(e,e) - E_{b,g}\mathfrak{b}(e,e) \geq \frac{1}{4}\|e\|_{H^1}^2 - W\|e\|_{L^2}^2 \tag{5.29}$$

—which stems from (5.22) of Lemma 5.1 applied to $v = e$— in the following fashion: multiply (5.28) by W, multiply (5.29) by $\frac{1}{2}\beta_b$ and add them together. It follows that

$$\mathfrak{a}(e,e) - E_{b,g}\mathfrak{b}(e,e) \geq \frac{\beta_b}{4(2W+\beta_b)}\|\,e\|_{H^1}^2.$$

The constant $\gamma_b = \dfrac{\beta_b}{4(2W+\beta_b)} > 0$ depends on b only. □

5.3.3 Practical computation of the estimate

The finite dimensionality of the subspaces $\mathscr{V}_b$ and $\mathscr{V}_{b,g}$ makes it possible for us to explicitly calculate the estimate $\eta_{b,g}$. Below, we give the details of this calculation and explain how it can be efficiently implemented. As in §5.3, we first consider the situation $N_g \geq 1$ for the sake of generality before delving into more specific details for $N_g = 1$, the only situation of interest for Algorithms 1 and 2.

Let us start by revisiting a basic result in constrained optimization.

Lemma 5.2. *Let* $\mathrm{N} \in \mathbb{N}^*$ *be a positive integer. Given a symmetric positive definite* $\mathrm{N}\times\mathrm{N}$*-matrix* $\mathbf{M}$ *and a vector* $\mathbf{r} \in \mathbb{R}^{\mathrm{N}}$*, we have*

$$\max_{\substack{\mathbf{v}\in\mathbb{R}^{\mathrm{N}}\\ \mathbf{v}^T\mathbf{M}\mathbf{v}=1}} \mathbf{r}^T\mathbf{v} = (\mathbf{r}^T\mathbf{M}^{-1}\mathbf{r})^{1/2}, \tag{5.30}$$

and this maximum value is achieved for

$$\mathbf{v}^* = \frac{1}{(\mathbf{r}^T\mathbf{M}^{-1}\mathbf{r})^{1/2}}\mathbf{M}^{-1}\mathbf{r}. \tag{5.31}$$

Proof. The symmetry and positive definiteness of $\mathbf{M}$ allow us to equip $\mathbb{R}^{\mathrm{N}}$ with the dot product

$$\langle \mathbf{p}, \mathbf{q}\rangle_{\mathsf{M}} = \mathbf{p}^T\mathbf{M}\mathbf{q}.$$

Application of the Cauchy-Schwarz inequality

$$\langle \mathbf{p}, \mathbf{q}\rangle_{\mathsf{M}} \leq \langle \mathbf{p}, \mathbf{p}\rangle_{\mathsf{M}}^{1/2}\langle \mathbf{q}, \mathbf{q}\rangle_{\mathsf{M}}^{1/2}$$

to $\mathbf{p} = \mathbf{M}^{-1}\mathbf{r}$ and $\mathbf{q} = \mathbf{v}$ yields

$$\mathbf{r}^T\mathbf{M}^{-T}\mathbf{M}\mathbf{v} \leq (\mathbf{r}^T\mathbf{M}^{-T}\mathbf{M}\mathbf{M}^{-1}\mathbf{r})^{1/2}(\mathbf{v}^T\mathbf{M}\mathbf{v})^{1/2}.$$

Upon simplification and taking into account the constraint on $\mathbf{v}$, we get

$$\mathbf{r}^T\mathbf{v} \leq (\mathbf{r}^T\mathbf{M}^{-1}\mathbf{r})^{1/2}.$$

The upper bound is the right-hand side of (5.30). It is reached when the vectors $\mathbf{p}$ and $\mathbf{q}$ are collinear, i.e., $\mathbf{v} = \lambda\mathbf{M}^{-1}\mathbf{r}$ for some $\lambda \in \mathbb{R}$. A further normalization of $\mathbf{v}$ gives the correct value of λ, whence (5.31). □

In the extended basis $b \cup g = \{b_1, \ldots, b_{N_b}\} \cup \{\breve{g}_1, \ldots, \breve{g}_{N_g}\}$, every function $v \in \mathscr{V}_{b,g}$ is decomposed as

$$v = \sum_{j=1}^{N_b} \mathsf{v}_j^b\, b_j + \sum_{q=1}^{N_g} \mathsf{v}_q^g\, \breve{g}_q. \tag{5.32}$$

The set of coefficients v_j^b, $1 \leq j \leq N_b$ and v_q^g, $1 \leq q \leq N_g$, is represented by the vector

$$\mathbf{v}^{b,g} = \begin{pmatrix} \mathbf{v}^b \\ \mathbf{v}^g \end{pmatrix} \in \mathbb{R}^{N_b+N_g}, \tag{5.33}$$

with $\mathbf{v}^b \in \mathbb{R}^{N_b}$ and $\mathbf{v}^g \in \mathbb{R}^{N_g}$. To alleviate notations, we shall be writing $\mathbf{v}$ instead of $\mathbf{v}^{b,g}$. The square of the H^1-norm of vector v can be computed from its coordinates as

$$\|v\|_{H^1}^2 = \mathbf{v}^T\mathbf{M}\mathbf{v} = (\mathbf{v}^b, \mathbf{v}^g) \begin{pmatrix} \mathbf{M}^b & \mathbf{M}^{bg} \\ \mathbf{M}^{gb} & \mathbf{M}^g \end{pmatrix} \begin{pmatrix} \mathbf{v}^b \\ \mathbf{v}^g \end{pmatrix}, \tag{5.34}$$

where $\mathbf{M}$ is the $(N_b + N_g) \times (N_b + N_g)$ Gram matrix expressing the H^1-norm in the mixed basis $b \cup g$. We have also used the symbol $\mathbf{M}$ instead of $\mathbf{M}^{b,g}$ in order to alleviate notations. The entries of $\mathbf{M}$ can be explicited as

$$\mathsf{M}_{ij}^b = \langle b_j, b_i \rangle_{H^1} \quad \text{for } (i,j) \in \{1, \ldots, N_b\}^2, \tag{5.35a}$$

$$\mathsf{M}_{pq}^g = \langle \breve{g}_q, \breve{g}_p \rangle_{H^1} \quad \text{for } (p,q) \in \{1, \ldots, N_g\}^2, \tag{5.35b}$$

$$\mathsf{M}_{iq}^{bg} = \langle \breve{g}_q, b_i \rangle_{H^1} \quad \text{for } (i,q) \in \{1, \ldots, N_b\} \times \{1, \ldots, N_g\}, \tag{5.35c}$$

$$\mathsf{M}_{pj}^{gb} = \langle b_j, \breve{g}_p \rangle_{H^1} \quad \text{for } (p,j) \in \{1, \ldots, N_g\} \times \{1, \ldots, N_b\}, \tag{5.35d}$$

where $\langle \cdot, \cdot \rangle_{H^1}$ denotes the H^1-dot product. All of the above entries are computable from the knowledge of the basis elements.

Theorem 5.2. *The square of the estimate $\eta_{b,g}$ introduced in Definition 5.10 is equal to*

$$\eta_{b,g}^2 = (\mathbf{r}^g)^T (\mathbf{M}^{-1})^g\, \mathbf{r}^g, \tag{5.36}$$

where

- *the vector $\mathbf{r}^g \in \mathbb{R}^{N_g}$ is given by*

$$\mathsf{r}_q^g = \mathrm{Res}_{\mathscr{V}_{b,g}}[u_b, E_b](\breve{g}_q) = \mathfrak{a}(u_b, \breve{g}_q) - E_b\mathfrak{b}(u_b, \breve{g}_q) \tag{5.37}$$

  for $1 \leq q \leq N_g$;

- *the matrix* $(\mathbf{M}^{-1})^g$ *is the lower-right* $N_g \times N_g$ *block of the full inverse* $\mathbf{M}^{-1}$*, that is,*

$$\mathbf{M}^{-1} = \begin{pmatrix} (\mathbf{M}^{-1})^b & (\mathbf{M}^{-1})^{bg} \\ (\mathbf{M}^{-1})^{gb} & (\mathbf{M}^{-1})^g \end{pmatrix}.$$

PROOF. By linearity of the residue form, we have

$$\operatorname{Res}[u_b, E_b](v) = \sum_{j=1}^{N_b} \mathsf{v}_j^b \operatorname{Res}[u_b, E_b](b_j) + \sum_{q=1}^{N_g} \mathsf{v}_q^g \operatorname{Res}[u_b, E_b](\breve{g}_q) \tag{5.38}$$

for any $\mathbf{v} \in \mathscr{V}_{b,g}$ having decomposition (5.32). But $b_j \in \mathscr{V}_b$ and $\mathscr{V}_b \subset \operatorname{Ker} \operatorname{Res}[u_b, E_b]$ by Remark 5.3, so $\operatorname{Res}[u_b, E_b](b_j) = 0$ and the equality (5.38) is reduced to

$$\operatorname{Res}[u_b, E_b](v) = \mathbf{r}^T \mathbf{v},$$

where $\mathbf{v}$ is the vector of components (5.33) and

$$\mathbf{r} = \begin{pmatrix} \mathbf{0} \\ \mathbf{r}^g \end{pmatrix}$$

using definition (5.37). On the other hand, by Definition 5.10 of the estimate,

$$\begin{aligned} \eta_{b,g} &= \sup_{\substack{v \in \mathscr{V}_{b,g} \\ \|v\|_{H^1}=1}} |\operatorname{Res}[u_b, E_b](v)| \\ &= \sup_{\substack{v \in \mathscr{V}_{b,g} \\ \|v\|_{H^1}^2=1}} \operatorname{Res}[u_b, E_b](v) \\ &= \sup_{\substack{\mathbf{v} \in \mathbb{R}^{N_b+N_g} \\ \mathbf{v}^T \mathbf{M} \mathbf{v}=1}} \mathbf{r}^T \mathbf{v}. \end{aligned}$$

As a Gram matrix, $\mathbf{M}$ is symmetric and positive definite. Applying Lemma 5.2 with $\mathrm{N} = N_b + N_g$, we infer that the maximum is attained and the square of its value is given by

$$\begin{aligned} \eta_{b,g}^2 &= \mathbf{r}^T \mathbf{M}^{-1} \mathbf{r} \\ &= (\mathbf{0}, (\mathbf{r}^g)^T) \begin{pmatrix} (\mathbf{M}^{-1})^b & (\mathbf{M}^{-1})^{bg} \\ (\mathbf{M}^{-1})^{gb} & (\mathbf{M}^{-1})^g \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{r}^g \end{pmatrix}, \end{aligned}$$

which implies (5.36). □

The residue vector $\mathbf{r}^g$ is of course computable from the data via (5.37). As for the computation of $(\mathbf{M}^{-1})^g$, although it seems to require the Gram matrix $\mathbf{M}$ to be inverted, only the last N_g components of the last N_g columns of $\mathbf{M}^{-1}$ are actually involved. Therefore, instead of taking the full inverse—which is tantanmount to solving $N_b + N_g$ linear systems—it is sufficient to solve N_g linear systems with matrix $\mathbf{M}$ and suitably varying right-hand sides. This observation allows for a significant saving when $N_g \ll N_b$. In particular, when $N_g = 1$ as in Algorithms 1 and 2, $\mathbf{r}^g$ becomes a scalar $\mathsf{r}^g := \mathsf{r}_1^g$ and only the last entry of $\mathbf{M}^{-1}$ is to be computed.

Corollary 5.2. *If $N_g = 1$, then*

$$\eta_{b,g}^2 = (\mathsf{r}^g)^2 (\mathsf{M}^{-1})_{N_b+1,N_b+1}, \tag{5.39}$$

where $(\mathsf{M}^{-1})_{N_b+1,N_b+1}$, the last entry of $\mathbf{M}^{-1}$, can be otained as the last component w^g of the solution to the linear system

$$\mathbf{M} \begin{pmatrix} \mathbf{w}^b \\ \mathsf{w}^g \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix}. \tag{5.40}$$

Proof. Left to the reader. □

5.3.4 Choice of an appropriate norm

In §5.3.3, the matrix $\mathbf{M}$ was introduced in (5.34) as the Gram matrix associated to the H^1-norm "of the mathematician"

$$\|u\|_{H^1}^2 := \|u\|_{L^2}^2 + \|u'\|_{L^2}^2. \tag{5.41}$$

As a matter of fact, two serious objections can be raised against this H^1-norm:

1. From the viewpoint of physics, the two summands of (5.41) do not have the same unit; this makes their sum meaningless and even "dangerous" to compute.
2. The norm (5.41) does not "see" the parameters of the original problem; this accounts for the following observation in preliminary numerical tests: despite the theoretical equivalence proved in Corollary 5.1, the maximum argument of $\eta_{b,g}^2$ may in practice lie rather far from that $E_b - E_{b,g}$.

One way to better match the maximum arguments of $\eta_{b,g}^2$ and $E_b - E_{b,g}$ is to make the constants K and γ_p of (5.19) as close to 1 as possible. In this spirit, we recommend a different norm on the subspace $\mathscr{V}_{b,g}$ which is not only equivalent to the H^1-norm but also consistent with the parameters of the problem. By this, we mean that in the formulae (5.35) for the entries of $\mathbf{M}$, it is judicious to replace $\langle \cdot, \cdot \rangle_{H^1}$ by a new dot product $(\!(\cdot, \cdot)\!)_\varepsilon$, defined over $\mathscr{V} \times \mathscr{V}$ as

$$\begin{aligned} (\!(v, w)\!)_\varepsilon :&= \mathfrak{a}(v, w) - (E_* - \varepsilon)\, \mathfrak{b}(v, w) \\ &= \frac{1}{2} \int_0^L v'w' - \sum_{I=1}^M Z_I v(X_I) w(X_I) - (E_* - \varepsilon) \int_0^L vw, \end{aligned} \tag{5.42}$$

for some $\varepsilon \geq 0$. The exact energy level E_* is generally unknown, but what must be known to define (5.42) is $E_* - \varepsilon$, a lower bound of the exact energy (this entails that the value of ε is generally unknown as well). As pointed out in Theorem 3.14, one such lower bound is available to us, namely,

$$E_* \geq -\frac{1}{2} \widetilde{\mathcal{Z}}^2 =: E_* - \varepsilon.$$

Naturally, if we happen to know the exact energy E_* (for instance, when $M = 1$ and $M = 2$), we can think of setting $\varepsilon = 0$ in $(\!(\cdot, \cdot)\!)\varepsilon$. The new norm

$$|\!|\!| v |\!|\!|_\varepsilon = (\!(v, v)\!)_\varepsilon^{1/2} \tag{5.43}$$

enjoys the following equivalence property.

Proposition 5.4. *If $\varepsilon > 0$, then $|||\cdot|||_\varepsilon$ is equivalent to $\|\cdot\|_{H^1}$ on the whole space $\mathscr{V}$. There exist two constants $\gamma_\varepsilon > 0$ and $K_\varepsilon > 0$, both depending on ε, such that for all $v \in \mathscr{V}$, we have*

$$\gamma_\varepsilon \|v\|_{H^1}^2 \leq |||v|||_\varepsilon^2 \leq K_\varepsilon \|v\|_{H^1}^2. \tag{5.44}$$

If $\varepsilon = 0$ and $E_* < E_{b,g}$, then $|||\cdot|||_0$ is equivalent to $\|\cdot\|_{H^1}$ on the subspace $\mathscr{V}_{b,g}$. There exist two constants $\gamma_{b,g} > 0$ and $K > 0$, with $\gamma_{b,g}$ depending on $\mathscr{V}_{b,g}$, such that for all $v \in \mathscr{V}_{b,g}$, we have

$$\gamma_{b,g} \|v\|_{H^1}^2 \leq |||v|||_0^2 \leq K \|v\|_{H^1}^2. \tag{5.45}$$

Proof. First, let us deal with $\varepsilon > 0$. For $v \in \mathscr{V}$, expressing

$$(\!(v,v)\!)_\varepsilon = \mathfrak{a}(v,v) - E_* \mathfrak{b}(v,v) + \varepsilon \mathfrak{b}(v,v) \tag{5.46}$$

and invoking the first Standard Hypothesis (4.32) (Proposition 4.6), we obtain

$$\mathfrak{a}(v,v) - E_* \mathfrak{b}(v,v) \leq K \|v\|_{H^1}^2 \tag{5.47}$$

with the same K as in (4.41) or (5.24). Furthermore, since $\mathfrak{b}(v,v) = \|v\|_{L^2}^2 \leq \|v\|_{H^1}^2$, it is plain that

$$(\!(v,v)\!)_\varepsilon \leq (K + \varepsilon) \|v\|_{H^1}^2.$$

This is the upper-bound of (5.44) with $K_\varepsilon = K + \varepsilon$.

By virtue of Lemma 4.2,

$$\mathfrak{a}(v,v) - E_* \mathfrak{b}(v,v) \geq \frac{1}{4} \|v\|_{H^1}^2 - W \|v\|_{L^2}^2$$

with the same W as in (4.40) or (5.23). Therefore, by (5.46),

$$(\!(v,v)\!)_\varepsilon \geq \frac{1}{4} \|v\|_{H^1}^2 - W \|v\|_{L^2}^2.$$

Combining this equation with

$$(\!(v,v)\!)_\varepsilon \geq \varepsilon \|v\|_{L^2}, \tag{5.48}$$

which is due to E_* being the smallest eigenvalue, we deduce that

$$(\!(v,v)\!)_\varepsilon \geq \frac{\varepsilon}{4(W + \varepsilon)} \|v\|_{H^1}^2.$$

This is the lower-bound of (5.44) with $\gamma_\varepsilon = \varepsilon / 4(W + \varepsilon)$.

We now tackle the case $\varepsilon = 0$. Because (5.47) remains valid, it is obvious that

$$(\!(v,v)\!)_\varepsilon \leq K \|v\|_{H^1}^2.$$

Writing

$$(\!(v,v)\!)_\varepsilon = \mathfrak{a}(v,v) - E_{b,g} \mathfrak{b}(v,v) + (E_{b,g} - E_*) \mathfrak{b}(v,v) \tag{5.49}$$

and invoking Lemma 5.1, we have on one hand

$$(\!(v,v)\!)_\varepsilon \geq \frac{1}{4} \|v\|_{H^1}^2 - W \|v\|_{L^2}^2. \tag{5.50}$$

On the other hand, for $v \in \mathscr{V}_{b,g}$, we have $\mathfrak{a}(v,v) - E_{b,g}\mathfrak{b}(v,v) \geq 0$ due to $E_{b,g}$ being the smallest eigenvalue of the discrete problem over $\mathscr{V}_{b,g}$. As a result,

$$(\!(v,v)\!)_\varepsilon \geq (E_{b,g} - E_*)\|v\|_{L^2}^2. \tag{5.51}$$

Combining (5.50) and (5.51), we arrive at

$$(\!(v,v)\!)_\varepsilon \geq \frac{E_{b,g} - E_*}{4(W + E_{b,g} - E_*)}\|v\|_{H^1}^2.$$

Thanks to the assumption $E_{b,g} > E_*$, the constant

$$\gamma_{b,g} = \frac{E_{b,g} - E_*}{4(W + E_{b,g} - E_*)}$$

is strictly positive. □

By Proposition 5.4, it is in principle not advisable to use $\varepsilon = 0$, insofar as one of the equivalence constants would depend on g, that is, on the additional functions to be adjusted. In practice, however, numerical experiments demonstrate that this is still a good choice for the new norm.

Chapter 6

Numerical results for the mixed basis

Contents

Nous mettons la stratégie proposée au chapitre §5 à l'épreuve des tests numériques. Comme au chapitre §4, nous commençons par le cas des potentiels simple-delta en domaine infini dans une base de gaussiennes, ce qui permet de reprendre sous une nouvelle perspective la construction des gaussiennes contractées, laissée en suspens à la fin de §4.2.2.

Nous abordons ensuite les potentiels double-delta en domaine périodique dans une base mixte ondelettes-gaussiennes. Une analyse étape par étape du comportement des deux algorithmes du chapitre §5 est entamée, à l'issue de laquelle nous mettrons en avant un troisième algorithme, plus empirique mais plus économique. Ce dernier ressemble aussi davantage aux orbitales atomiques puisqu'il s'appuie la plupart du temps sur le transfert atome/molécule des gaussiennes contractées construites en présence d'une base d'ondelettes existante par les deux premiers algorithmes. La limite de validité de ce troisième algorithme apparaît lorsque la distance entre les deux noyaux est faible au regard de leurs longueurs caractéristiques, auquel cas la transférabilité est remise en cause. Dans ce cas, on revient à l'estimateur a posteriori *pour déterminer les gaussiennes à ajouter.*

Nous esquissons à la fin du chapitre ce qu'il reste à faire pour les problèmes avec un potentiel de trois deltas ou plus.

6.1 Estimates in the case of single-delta potential

We begin by stating a paradigm that will qualify the criterion of maximizing $\eta^2_{b,g}$.

Paradigm 6.1. *With a fixed basis b, the additional element g that maximizes the squared* a posteriori *estimate $\eta^2_{b,g}$ is also such that the corresponding energy decay $E_b - E_{b,g}$ is very close to its maximum value.*

This paradigm will be illustrated in the following sections. We are going to show analytic calculations and/or approximate results for two types of bases, where $\mathscr{V}_b$ is either a pure Gaussian basis or a mixed basis. On the other hand, we will demonstrate that it is faster to compute $\eta^2_{b,g}$ than $E_b - E_{b,g}$; therefore, we are able to look for additional elements g using the $\eta^2_{b,g}$ criterion.

6.1.1 Gaussian bases

This section serves two purposes:

- to show, numerically, the good agreement between the two optimization criteria — the *a posteriori* estimate and the energy decay— therefore justify the statement of Paradigm 6.1.
- to demonstrate the theoretical computation of the estimate in a simple case.

As explained in §4.2, for this type of bases, in order to have explicit formulae for the estimate, we consider the equation with a single-delta potential at $X = 0$ on the infinite domain:

$$-\frac{1}{2}u'' - Z\delta_0 u = E\,u, \tag{6.1a}$$

$$\int_{\mathbb{R}} |u|^2 dx = 1, \tag{6.1b}$$

for a given $Z > 0$. Again, the space of solutions is

$$\mathscr{V} = H^1(\mathbb{R}).$$

Recall that the mass form and rigid form of the equation (6.1) are:

$$\mathfrak{b}(u,v) := \int_{\mathbb{R}} uv = (u,v), \tag{6.2a}$$

$$\mathfrak{a}(u,v) := \frac{1}{2}\int_{\mathbb{R}} u'v' - Zu(0)v(0). \tag{6.2b}$$

For a set of standard deviations $\boldsymbol{\sigma} = \{\sigma_q, 1 \le q \le Q\}$, let us consider the subspace of the Galerkin approximation

$$\mathscr{V}_b = \mathscr{V}_{\boldsymbol{\sigma}} = \mathrm{Span}\{g_{\sigma_q}, 1 \le q \le Q\},$$

where the g_{σ_q}'s are L^2-normalized Gaussians centered at 0, defined in (4.10)

$$g_\sigma(x) := \frac{1}{\sigma^{1/2}\pi^{1/4}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{6.3}$$

In a similar spirit with §5.3.4, we define a new H^1-norm on $\mathscr{V}_g$ from the product in (5.42) with $\varepsilon = 0$.

Definition 6.1. We define a H^1 inner product on $\mathscr{V}_g$ as

$$(\!(v,w)\!) := \mathfrak{a}(v,w) - E_*\mathfrak{b}(v,w) \tag{6.4a}$$

$$= \frac{1}{2}\int_{\mathbb{R}} v'w' - Zv(0)w(0) - E_*\int_{\mathbb{R}} vw, \tag{6.4b}$$

for all $v, w \in \mathscr{V}_g$, where E_* is the exact energy level of (6.1), calculated in Theorem 3.4

$$\boxed{E_* = -\frac{Z^2}{2}.} \tag{6.5}$$

The following Lemma 6.1 rewrites the identities (4.13) into a form that suits this chapter better.

Lemma 6.1. *For all $\sigma_p, \sigma_q \in \mathbb{R}_+$, with $r_{pq} := \sigma_q/\sigma_p$, we have*

$$\int_{-\infty}^{\infty} g_{\sigma_p} g_{\sigma_q} dx = \sqrt{\frac{2\sigma_q/\sigma_p}{1+(\sigma_q/\sigma_p)^2}} = \sqrt{\frac{2r_{pq}}{1+r_{pq}^2}},$$

$$\int_{-\infty}^{\infty} g'_{\sigma_p} g'_{\sigma_q} dx = \frac{1}{\sigma_p^2+\sigma_q^2}\sqrt{\frac{2\sigma_q/\sigma_p}{1+(\sigma_q/\sigma_p)^2}} = \frac{1}{\sigma_p^2}\frac{1}{1+r_{pq}^2}\sqrt{\frac{2r_{pq}}{1+r_{pq}^2}},$$

$$\int_{-\infty}^{\infty} |g'_{\sigma_p}|^2 dx = \frac{1}{2\sigma_p^2}.$$

With Lemma 6.1 in mind and

$$g_{\sigma_p}(0) g_{\sigma_q}(0) = \frac{1}{\sqrt{\sigma_p\sigma_q\pi}} = \frac{1}{\sigma_p\sqrt{\pi r_{pq}}},$$

we obtain the mass form and rigid form of the Gaussians

$$\mathfrak{a}(g_{\sigma_p}, g_{\sigma_q}) = \frac{1}{2\sigma_p^2}\frac{1}{1+r_{pq}^2}\sqrt{\frac{2r_{pq}}{1+r_{pq}^2}} - \frac{1}{\sigma_p\sqrt{\pi r_{pq}}}, \tag{6.6a}$$

$$\mathfrak{a}(g_{\sigma_p}, g_{\sigma_p}) = \frac{1}{4\sigma_p^2} - \frac{1}{\sigma_p\sqrt{\pi}}, \tag{6.6b}$$

$$\mathfrak{b}(g_{\sigma_p}, g_{\sigma_q}) = \sqrt{\frac{2r_{pq}}{1+r_{pq}^2}}. \tag{6.6c}$$

For fixed indices $p, q \in \mathbb{N}^*$, the ratio $r_{pq} = \sigma_q/\sigma_p$ is called the ***dilation*** of g_{σ_q} compared to g_{σ_p}. When without ambiguity, we omit the indices p, q in r_{pq} and simply write r. Denote

$$\Lambda := \frac{1}{Z}.$$

We are going to use the listed equalities to compute the estimate when $Q \leq 3$.

Basis of 1 Gaussian

For $Q = 1$, we consider a basis of one centered Gaussian g_{σ_1} where the standard deviation σ_1 is an unknown parameter. From (4.18) and (4.19), we know that the reference standard deviation

$$\boxed{\sigma_* = \frac{\sqrt{\pi}}{2}\Lambda} \tag{6.7}$$

corresponds to the minimum energy level, which is

$$\boxed{E_{1*} = -\frac{Z^2}{\pi},}$$

and which has to be compared to the exact value (6.5). This would be the first optimal Gaussian by the energy criterion, while, by definition, there is not yet an *a posteriori* estimate. Choose the first element $g_{\sigma_{1*}}$ of our basis to be this optimal g_{σ_*}.

Passing from 1 Gaussian to 2 Gaussians ($Q = 2$)

Since in all our calculations the standard deviations are homogenous with Λ, we may suppose that $\Lambda = 1$ and sustitute all σ/Λ by σ.

The unique solution, up to a sign, of the equation (6.1) on the subspace $\mathbb{R}g_{\sigma_{1*}}$ is also $g_{\sigma_{1*}}$. Using Algorithm 2, we look for the most suitable g_{σ_2} to enrich the basis $\{g_{\sigma_{1*}}\}$. Denote

$$\mathscr{V}_2 := \mathrm{Span}\{g_{\sigma_{1*}}, g_{\sigma_2}\},$$

with σ_2 to be determined later, and E_2 the approximate energy levels on $\mathscr{V}_2$. The ideal g_{σ_2} would be the Gaussian which maximizes the energy decay $E_{1*} - E_2$; but, by Paradigm 6.1, we may also look for the g_{σ_2} that maximizes the *a posteriori* estimate

$$\eta_{\sigma_{1*},\sigma_2} := \|\mathrm{Res}_{\mathscr{V}_2}[g_{\sigma_{1*}}, E_{1*}]\|_{\mathscr{V}_2'}.$$

Computation of the a posteriori *estimate.* We calculate $\eta_{\sigma_{1*},\sigma_2}$ by the formula (5.39):

$$\eta^2_{\sigma_{1*},\sigma_2} = \left(\mathrm{Res}_{\mathscr{V}_2}[g_{\sigma_{1*}}, E_{1*}](g_{\sigma_2})\right)^2 \left(\mathsf{M}_2^{-1}\right)_{2,2}, \tag{6.8}$$

where $\mathbf{M}_2 \in \mathbb{R}^{2\times 2}$ is the Gram matrix of the basis $\{g_{\sigma_{1*}}, g_{\sigma_2}\}$ by the new H^1-norm, and $\left(\mathsf{M}_2^{-1}\right)_{2,2}$ is the last entry of its inverse matrix. We proceed to calculate $\mathbf{M}_2$ with the new inner product $(\!(\cdot,\cdot)\!)$ on $\mathscr{V}_2$, defined in (6.4).

$$\mathbf{M}_2 := \begin{pmatrix} (\!(g_{\sigma_{1*}}, g_{\sigma_{1*}})\!) & (\!(g_{\sigma_{1*}}, g_{\sigma_2})\!) \\ \\ (\!(g_{\sigma_2}, g_{\sigma_{1*}})\!) & (\!(g_{\sigma_2}, g_{\sigma_2})\!) \end{pmatrix}.$$

Formula (5.39) gives us

$$\eta^2_{\sigma_{1*},\sigma_2} = |\mathrm{Res}[g_{\sigma_{1*}}, E_{1*}](g_{\sigma_2})|^2 \frac{(\!(g_{\sigma_{1*}}, g_{\sigma_{1*}})\!)}{\det \mathbf{M}_2}. \tag{6.9}$$

Notice that the exact energy level in this case is

$$E_* = -1/2,$$

so the inner product defined in (6.4) becomes

$$(\!(v,w)\!) = \mathfrak{a}(v,w) + \frac{1}{2}\mathfrak{b}(v,w). \tag{6.10}$$

By definition of the residue of $(g_{\sigma_{1*}}, E_{1*})$ over $\mathscr{V}_2$, we have

$$\mathrm{Res}_{\mathscr{V}_2}[g_{\sigma_{1*}}, E_{1*}](g_{\sigma_2}) = \mathfrak{a}(g_{\sigma_{1*}}, g_{\sigma_2}) - E_{1*}\mathfrak{b}(g_{\sigma_{1*}}, g_{\sigma_2}). \tag{6.11}$$

We look for the optimal dilation $r = \sigma_2/\sigma_{1*}$ which maximizes the estimate $\eta_{\sigma_{1*},\sigma_2}$ or the energy decay $E_{1*} - E_2$, while σ_2 is varying. Insert $\sigma_p = \sigma_{1*} = \sqrt{\pi}/2$ into the equalities (6.6) and $E_{1*} = -1/\pi$ into (6.11), we have

$$\begin{aligned}\mathrm{Res}[g_{\sigma_{1*}}, E_{1*}](g_{\sigma_2}) &= \frac{2}{\pi}\frac{1}{1+r^2}\sqrt{\frac{2r}{1+r^2}} - \frac{2}{\pi\sqrt{r}} + \frac{1}{\pi}\sqrt{\frac{2r}{1+r^2}} \\ &= \frac{1}{\pi}\left(\frac{3+r^2}{1+r^2}\sqrt{\frac{2r}{1+r^2}} - \frac{2}{\sqrt{r}}\right).\end{aligned} \tag{6.12}$$

Using (6.6) again with $\sigma_p = \sigma_{1*} = \sqrt{\pi}/2$, we obtain

$$(\!(g_{\sigma_{1*}}, g_{\sigma_{1*}})\!) = \frac{\pi-2}{2\pi}; \tag{6.13a}$$

$$(\!(g_{\sigma_2}, g_{\sigma_2})\!) = \frac{1}{\pi r^2} - \frac{2}{\pi r} + \frac{1}{2}; \tag{6.13b}$$

$$(\!(g_{\sigma_{1*}}, g_{\sigma_2})\!) = \frac{2}{\pi}\left(\frac{\pi r^2+\pi+4}{4(1+r^2)}\sqrt{\frac{2r}{1+r^2}} - \frac{1}{\sqrt{r}}\right); \tag{6.13c}$$

and

$$\begin{aligned}\det \mathbf{M}_2 &= (\!(g_{\sigma_{1*}}, g_{\sigma_{1*}})\!)(\!(g_{\sigma_2}, g_{\sigma_2})\!) - |(\!(g_{\sigma_{1*}}, g_{\sigma_2})\!)|^2 \\ &= \frac{\pi-2}{2\pi^2}\left(\frac{1}{r^2} - \frac{2}{r} + \frac{\pi}{2}\right) - \frac{4}{\pi^2}\left(\frac{\pi r^2+\pi+4}{4(1+r^2)}\sqrt{\frac{2r}{1+r^2}} - \frac{1}{\sqrt{r}}\right)^2.\end{aligned} \tag{6.14}$$

Insert (6.12), (6.13a) and (6.14) into (6.9), we obtain an analytic form for the squared estimate $\eta^2_{\sigma_{1*},\sigma_2}$.

Computation of the energy decay. The quantity $E_{1*} - E_2$ can also be obtained by a closed-form expression. After Proposition 4.3 in chapter §4, the approximate energy level E_2 is calculated as

$$E_2 = \frac{1}{2}\left(\mathtt{C} - \sqrt{\mathtt{C}^2 - 4\mathtt{D}}\right) \tag{6.15}$$

with

$$\mathtt{C} = \frac{\mathsf{A}^{\sigma}_{12} + \mathsf{A}^{\sigma}_{22} - 2\mathsf{A}^{\sigma}_{12}\mathsf{B}^{\sigma}_{12}}{1 - \left(\mathsf{B}^{\sigma}_{12}\right)^2}, \qquad \mathtt{D} = \frac{\mathsf{A}^{\sigma}_{11}\mathsf{A}^{\sigma}_{22} - \left(\mathsf{A}^{\sigma}_{12}\right)^2}{1 - \left(\mathsf{B}^{\sigma}_{12}\right)^2}, \tag{6.16}$$

where the entries of the matrices $\mathbf{A}^{\sigma}$ and $\mathbf{B}^{\sigma}$ are given in (4.12), with σ_{1*} and σ_2 in place of σ_p and σ_q.

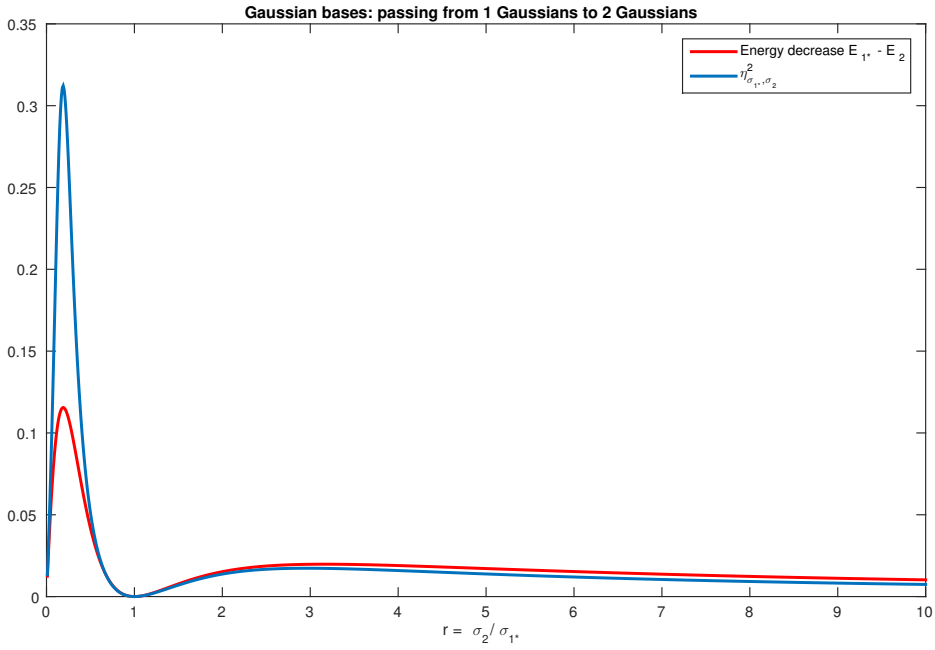

Figure 6.1: Energy decay $E_{1*} - E_2$ and squared estimate $\eta^2_{\sigma_{1*},\sigma_2}$ in terms of $r = \frac{\sigma_2}{\sigma_{1*}}$, Gaussian bases of dimension 2.

Comparison of the two methods. We now compare the two ways of computing the optimal dilation factor r. It is up to us to choose an interval in $\mathbb{R}_+^*$ for r, in this chapter we will test with r in $(0, 10]$.

Using the above analytic forms, Figure 6.1 plot the curves of $E_{1*} - E_2$ and $\eta^2_{\sigma_{1*},\sigma_2}$ when r varies in $(0, 10]$, then find their maximal points. The two curves show a good agreement between themselves, especially in term of the peaks. There might be several local peaks for each curve but we pay attention only to the global maxima on the interval $(0, 10]$, of which the arguments are:

$$r^{\mathfrak{E}}_{1*} = 0.188624,$$
$$r^{\eta}_{1*} = 0.189930,$$

where we have denoted

$$r^{\mathfrak{E}}_{1*} := \arg \max_{\sigma_2 = r\sigma_{1*},\ 0<r\leq 10} (E_{1*} - E_2),$$
$$r^{\eta}_{1*} := \arg \max_{\sigma_2 = r\sigma_{1*},\ 0<r\leq 10} \eta^2_{\sigma_{1*},\sigma_2}.$$

We see that the difference between these two numbers is small: it is of order 10^{-3}. When we use the usual "mathematical" H^1-norm or some equivalent norms instead of the "physical" H^1-norm $(\!(\cdot,\cdot)\!)$, the tests do not give such promising results. It is because the definition formula of $(\!(\cdot,\cdot)\!)$ is coherent with that of the residue, as we have intentionally chosen it.

Choose the second Gaussian in our basis to be the optimal $g_{\sigma_{2*}}$ by the estimate criterion, with

$$\sigma_{2*} = r_{1*}^{\eta}\sigma_{1*}.$$

Passing from 2 Gaussians to 3 Gaussians ($Q = 3$)

Continuing by Algorithm 2, we look for an optimal Gaussian g_{σ_3}. Denote:

- (u_{2*}, E_{2*}) the solution of the equation on the new-found subspace $\mathscr{V}_2 = \mathrm{Span}\{g_{\sigma_{1*}}, g_{\sigma_{2*}}\}$; in particular, thanks to Proposition 4.3, the approximate atomic orbital u_{2*} can be calculated by a closed form: if we denote $(\mathsf{u}_1, \mathsf{u}_2)$ the coefficients of u_{2*} in the basis $\{g_{\sigma_{1*}}, g_{\sigma_{2*}}\}$, then

$$(\mathsf{u}_1, \mathsf{u}_2)^T = \frac{1}{\{(\mathbf{v}^{\boldsymbol{\sigma}})^T \mathbf{B}^{\boldsymbol{\sigma}} \mathbf{v}^{\boldsymbol{\sigma}}\}^{1/2}} \mathbf{v}^{\boldsymbol{\sigma}}, \qquad \mathbf{v}^{\boldsymbol{\sigma}} = \left(E_{\boldsymbol{\sigma}} \mathsf{B}_{12}^{\boldsymbol{\sigma}} - \mathsf{A}_{12}^{\boldsymbol{\sigma}},\, \mathsf{A}_{11}^{\boldsymbol{\sigma}} - E_{2*}\right)^T, \tag{6.17}$$

  where the entries of the matrices $\mathbf{A}^{\boldsymbol{\sigma}}$ and $\mathbf{B}^{\boldsymbol{\sigma}}$ are given in (4.12), with σ_{1*} and σ_{2*} in place of σ_p and σ_q.

- $\mathscr{V}_3 := \mathrm{Span}\{g_{\sigma_{1*}}, g_{\sigma_{2*}}, g_{\sigma_3}\}$ with σ_3 to be determined later,

- E_3 the approximate energy levels on $\mathscr{V}_3$.

The appropriate σ_3 is the one that maximizes $\eta^2_{\{\sigma_{1*},\sigma_{2*}\},\sigma_3}$, where

$$\eta^2_{\{\sigma_{1*},\sigma_{2*}\},\sigma_3} := \|\mathrm{Res}_{\mathscr{V}_3}[u_{2*}, E_{2*}]\|_{\mathscr{V}_3'}.$$

We calculate $\eta^2_{\{\sigma_{1*},\sigma_{2*}\},\sigma_3}$ by the formula (5.39):

$$\eta^2_{\sigma_{1*},\sigma_2} = \left(\mathrm{Res}_{\mathscr{V}_3}[u_{2*}, E_{2*}](g_{\sigma_3})\right)^2 \left(\mathsf{M}_3^{-1}\right)_{3,3}, \tag{6.18}$$

where $\mathbf{M}_3 \in \mathbb{R}^{3\times 3}$ is the Gram matrix of the basis $\{g_{\sigma_{1*}}, g_{\sigma_{2*}}, g_{\sigma_3}\}$ by the new H^1-norm $(\!(\cdot,\cdot)\!)$. To calculate $\mathbf{M}_3$, we notice that, if $r_p = \sigma_p/\sigma_{1*}$ for any $p \in \mathbb{N}^*$, then $(\!(g_{\sigma_p}, g_{\sigma_q})\!)$ is a function of (r_p, r_q):

$$(\!(g_{\sigma_p}, g_{\sigma_q})\!) = f(r_p, r_q), \tag{6.19}$$

where

$$f(x,y) := \frac{2}{\pi}\left(\sqrt{\frac{2xy}{(x^2+y^2)^3}} - \frac{1}{\sqrt{xy}}\right) + \frac{1}{2}\sqrt{\frac{2xy}{x^2+y^2}}\,. \tag{6.20}$$

Insert the formulas (6.19) and (6.20) into $\mathbf{M}_3$, we can easily compute its determinant. For $\mathrm{Res}_{\mathscr{V}_3}[u_{2*}, E_{2*}](g_{\sigma_3})$, it is a linear combination of

$$\mathrm{Res}_{\mathscr{V}_3}[g_{\sigma_{i*}}, E_{2*}](g_{\sigma_3}) = \mathfrak{a}(g_{\sigma_{i*}}, g_{\sigma_3}) - E_{2*}\,\mathfrak{b}(g_{\sigma_{i*}}, g_{\sigma_3}), \quad i = 1, 2. \tag{6.21}$$

Again, insert (6.19) and (6.20) into (6.21), we can compute the residue and consequently, the squared estimate $\eta^2_{\sigma_{1*},\sigma_2}$.

Figure 6.2 plots the two curves of $E_{2*}-E_3$ and $\eta^2_{\{\sigma_{1*},\sigma_{2*}\},\sigma_3}$ when the dilation $r = \sigma_3/\sigma_{1*}$ varies in $(0, 10]$, then zooms in near the global maximal points, of which the arguments are:

$$r^{\mathfrak{E}}_{2*} = 0.033629,$$
$$r^{\eta}_{2*} = 0.033624.$$

We see that the two curves behave alike, and the two maximal arguments are again very close. The difference between them is even smaller this time: it is of order 10^{-6}. These examples lead us to another numerical assumption which is more concrete than Paradigm 6.1.

Paradigm 6.2. *If the basis is "rich" enough, the maximal arguments of the two criteria —the energy decay and the* a posteriori *estimate— are nearly identical, in the sense that their difference can be ignored.*

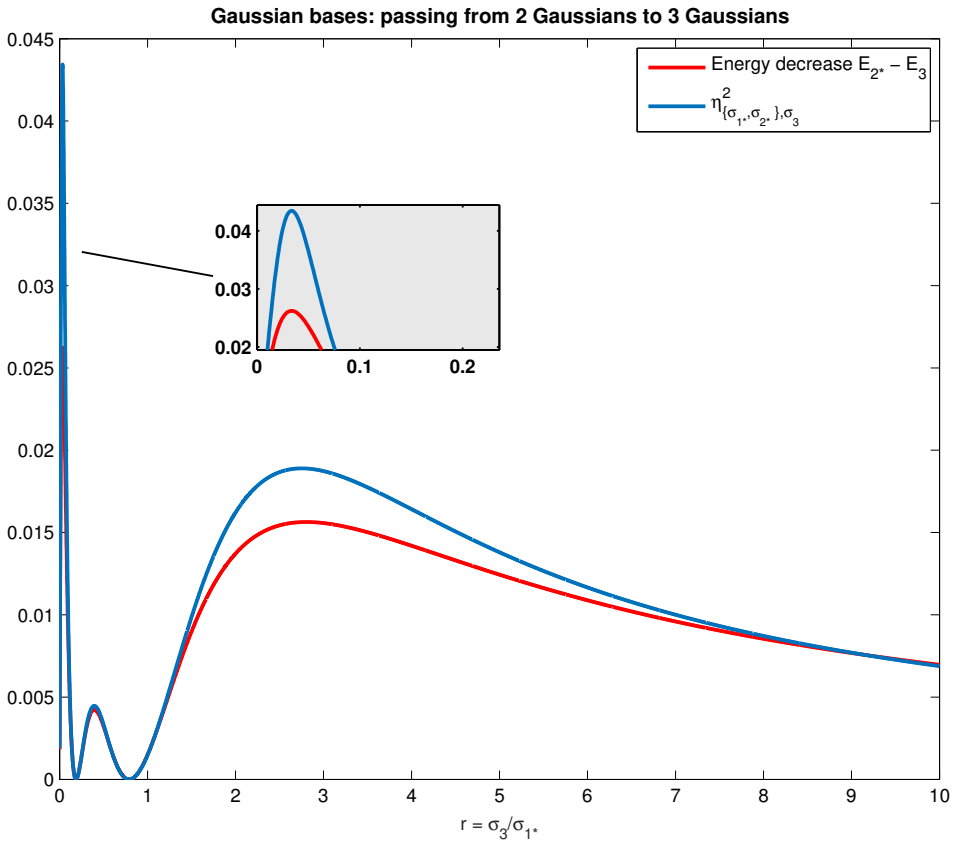

Figure 6.2: Energy decrease $E_{2*}-E_3$ and squared estimate $\eta^2_{\{\sigma_{1*},\sigma_{2*}\},\sigma_3}$ in terms of $r = \frac{\sigma_3}{\sigma_{1*}}$, Gaussian bases of dimension 3.

Bases of more than 3 Gaussians

In Table 6.1, which corresponds to the equation (6.1) on $\mathbb{R}$, for $1 \leq Q \leq 6$, we list the optimal dilation r for the Q-th Gaussian found by the greedy algorithm, assuming that the first $Q-1$ Gaussians were already found, **using the estimate criterion**. Denote

$$r_*^{\mathfrak{E}} := \arg \max_{\sigma_Q = r\sigma_{1*},\ 0<r\leq 10} \left(E_{(Q-1)*} - E_Q\right),$$
$$r_*^{\eta} := \arg \max_{\sigma_Q = r\sigma_{1*},\ 0<r\leq 10} \eta^2_{\{\sigma_{1*},\ldots,\sigma_{(Q-1)*}\},\sigma_Q},$$

$\mathrm{err}(r)$ the relative difference between the two dilations at each step:

$$\mathrm{err}(r) := \frac{\left|r_*^{\mathfrak{E}} - r_*^{\eta}\right|}{r_*^{\eta}},$$

$\mathtt{err}\left(E^{\mathfrak{E}}\right)$ and $\mathtt{err}\left(E^{\eta}\right)$ respectively the relative energy decay obtained by using $r_*^{\mathfrak{E}}$ or r_*^{η}, while the relative energy decay is understood as

$$\mathtt{err}(E) := \frac{E_{(Q-1)} - E_Q}{\left|E_{(Q-1)}\right|},$$

$\mathrm{err}(\mathtt{err})$ the relative difference between the two energy decays, to quantify their proximity:

$$\mathrm{err}(\mathtt{err}) := \frac{\left|\mathtt{err}\left(E^{\mathfrak{E}}\right) - \mathtt{err}\left(E^{\eta}\right)\right|}{\mathtt{err}\left(E^{\eta}\right)},$$

$t^{\mathfrak{E}}$ and t^{η} respectively the computation time (in seconds) in Matlab to find $r_*^{\mathfrak{E}}$ or r_*^{η}. We choose $\epsilon_{tol} = 0.01$ and stop the algorithm when

$$\mathtt{err}\left(E^{\eta}\right) < \epsilon_{tol},$$

which is a criterion that leads to maximum 6 Gaussians in this case. The standard deviation of the new-found Gaussian would be

$$\sigma = r_*^{\eta}\sigma_{1*} \text{ where } \sigma_{1*} = \frac{\sqrt{\pi}}{2}\Lambda.$$

Recalling the other parameters for the table

$$Z = 1,\ X = 0.$$

Table 6.1 shows the good agreement between the estimate and energy decrease: not only are the optimal dilations $r_*^{\mathfrak{E}}$ and r_*^{η} close to each other, but also the energies obtained by using them, $\mathtt{err}\left(E^{\mathfrak{E}}\right)$ and $\mathtt{err}\left(E^{\eta}\right)$, are not much different. This fact is confirmed by the relative difference $\mathrm{err}(\mathtt{err})$. Moreover, the time to compute r_*^{η} is shorter than that to compute $r_*^{\mathfrak{E}}$, since the latter evokes eigenvalue problems. Therefore, we assume that we can use r_*^{η} instead of $r_*^{\mathfrak{E}}$ for all of its advantages. We then test this assumption on mixed bases.

Q	$r_*^{\mathfrak{E}}$	r_*^{η}	err(r)	$\mathtt{err}(E^{\mathfrak{E}})$	$\mathtt{err}(E^{\eta})$	err$(\mathtt{err})$	$t^{\mathfrak{E}}$ (s)	t^{η} (s)
1	1	1	0					
2	0.189930	0.188624	6.9E-03	0.362954	0.362942	3.3E-05	0.005	0.002
3	0.033629	0.033624	1.3E-04	0.060456	0.060456	2.8E-08	0.006	0.003
4	2.879014	2.852703	9.2E-03	0.028234	0.028231	8.2E-05	0.008	0.005
5	0.454340	0.453914	9.4E-04	0.031925	0.031925	1.6E-06	0.012	0.007
6	0.005779	0.005779	6.6E-07	0.009398	0.009398	1.0E-07	0.018	0.011

Table 6.1: Optimal dilations for Gaussian bases by the greedy algorithm, for the energy criterion ($r_*^{\mathfrak{E}}$) and the estimate criterion (r_*^{η}), and the corresponding computational time.

6.1.2 Scaling function-Gaussian mixed bases

For mixed bases, we can no longer give analytic calculations, yet the wavelet-Gaussian scalar product can be evaluated numerically with arbitrary precision; we may therefore provide reliable numerical results of our procedure. We switch back to the periodic model

$$-\frac{1}{2}u'' - Z\delta_X u = E\,u, \tag{6.22a}$$

$$\int_0^L |u|^2 = 1, \tag{6.22b}$$

for a fix point $X \in [0, L]$ and a charge $Z > 0$, with weak periodic boundary conditions. Given a periodic scaling function basis on $[0, L]$

$$b = \{\widetilde{\chi}_i^h, i = 0, 1, \ldots, N-1\}$$

in

$$\mathscr{V} = H^1_{\#}([0, L]),$$

the corresponding subspace is

$$\mathscr{V}_b = \mathscr{V}_h := \mathrm{Span}\{\widetilde{\chi}_i^h, i = 0, 1, \ldots, N-1\}.$$

The respective approximate solution is denoted (u_h, E_h).

In the case of a single-delta potential, Algorithm 1 and Algorithm 2 in chapter §5 are the same. We will just use the greedy algorithm to gradually add one Gaussian at a time to the basis, until it reaches some threshold. In this simple case, we are able to take a close look at each new-found Gaussian.

Optimal choice of the first Gaussian

First, we add one periodic Gaussian $\widetilde{g}_{\sigma_1,X}$, with σ_1 unknown, to the basis b and denote

$$\mathscr{V}_{h,g_{\sigma_1}} := \mathrm{Span}\{\widetilde{g}_{\sigma_1,X}, \widetilde{\chi}_i^h, i = 0, 1, \ldots, N-1\}$$

the augmented subspace, $(u_{h,g_{\sigma_1}}, E_{h,g_{\sigma_1}})$ the approximate solution on $\mathscr{V}_{h,g_{\sigma_1}}$, and $\eta_{h,g_{\sigma_1}}$ the *a posteriori* estimate between the solutions (u_h, E_h) and $(u_{h,g_{\sigma_1}}, E_{h,g_{\sigma_1}})$:

$$\eta_{h,g_{\sigma_1}} := \|\mathrm{Res}_{\mathscr{V}_{h,g_{\sigma_1}}}[u_h, E_h]\|_{\mathscr{V}'_{h,g_{\sigma_1}}}. \tag{6.23}$$

Again, we calculate $\eta_{h,g_{\sigma_1}}$ numerically by the formula (5.39), with the new H^1-norm defined in (6.4). We look for the optimal σ_1 that maximizes the estimate $\eta_{h,g_{\sigma_1}}$ or, we might hope after Theorem 5.1 and Paradigm 6.1, that the same standard deviation would "almost" maximize the energy decrease $E_h - E_{h,g_{\sigma_1}}$.

For the following numerical tests we use Daubechies scaling functions of order $\mathsf{M} = 4$. Considering the reference standard deviation

$$\boxed{\sigma_* = \frac{\sqrt{\pi}}{2}\Lambda} \tag{6.24}$$

with $\Lambda = 1/Z$. Instead of searching for σ_1, we will look for the optimal dilation $r_1 = \sigma_1/\sigma_*$. We measure both $\eta^2_{h,g_{\sigma_1}}$ and $E_h - E_{h,g_{\sigma_1}}$ while r_1 varies in $(0, 10]$ to see the good agreement between the two quantities. When

$$N = 2^7,\ L = 1,\ X = L/2,\ \mathsf{M} = 4, \tag{6.25}$$

the maximal arguments of the two quantities are:

$$r^{\mathfrak{E}}_{1*} = 0.048105, \tag{6.26}$$
$$r^{\eta}_{1*} = 0.048104, \tag{6.27}$$

where we have denoted

$$r^{\mathfrak{E}}_{1*} := \arg \max_{\sigma_1 = r\sigma_*,\ 0<r\leq 10} (E_h - E_{h,g_{\sigma_1}}),$$
$$r^{\eta}_{1*} := \arg \max_{\sigma_1 = r\sigma_*,\ 0<r\leq 10} \eta^2_{h,g_{\sigma_1}}.$$

These two values are very close, so Paradigm 6.1 and 6.2 hold true. With the parameters given in (6.25), Figure 6.3 zooms in the two curves around the points $r^{\mathfrak{E}}_{1*}$ and r^{η}_{1*}, when r_1 varies in the interval $(0, 2]$; it shows that the shape of the two curves are quite similar. More results for different tests will be listed in data tables 6.2 and 6.3 at the end of the section.

The proximity of the two maximal arguments allows us to use r^{η}_{1*} instead of $r^{\mathfrak{E}}_{1*}$. To proceed, we keep the optimal Gaussian $\widetilde{g}_{\sigma_{1*},X}$ for the basis, with

$$\sigma_{1*} = r^{\eta}_{1*}\sigma_*$$

We will plot the solutions over these bases to see their behavior. With the parameters as above, Figure 6.4 shows the exact solution u_*, the approximate solution u_h over the scaling function basis, the solution $u_{h,g_{\sigma_*}}$ over the mixed basis with the reference Gaussian $\widetilde{g}_{\sigma_*,X}$, and the solution $u_{h,g_{\sigma_{1*}}}$ over the mixed basis with the optimal Gaussian $\widetilde{g}_{\sigma_{1*},X}$. The zoom-in at the cusp shows that the reference Gaussian $\widetilde{g}_{\sigma_*,X}$ is not the best choice when combining with scaling functions; dilating this Gaussian by r^{η}_{1*} has much improved the accuracy.

Optimal choice of the second Gaussian

We continue to enrich the basis $\{b, \widetilde{g}_{\sigma_{1*},X}\}$ by a second Gaussian of standard deviation σ_2, always centered at X. We search for the optimal ratio $r_2 = \sigma_2/\sigma_*$ which maximizes the energy diminution $E_{h,g_{\sigma_1}} - E_{h,g_{\sigma_1},g_{\sigma_2}}$ or $\eta^2_{h,g_{\sigma_1},g_{\sigma_2}}$.

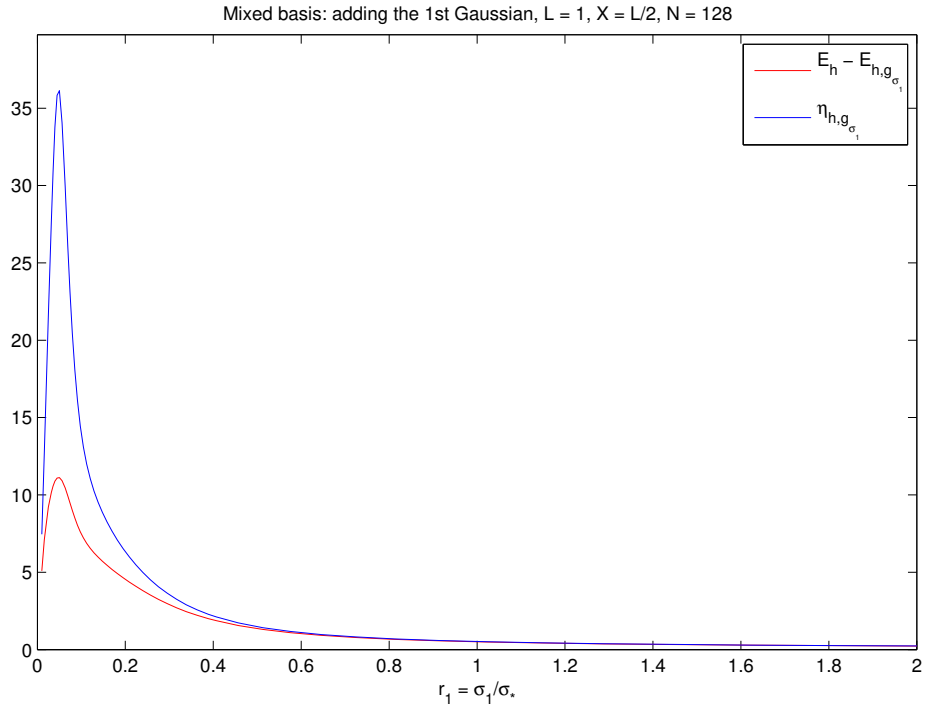

Figure 6.3: Energy decrease $E_h - E_{h,g_{\sigma_1}}$ and squared estimate $\eta^2_{h,g_{\sigma_1}}$ in terms of $r_1 = \frac{\sigma_1}{\sigma_*}$, mixed basis with 1 Gaussian $\widetilde{g}_{\sigma_1,X}$, single-delta potential.

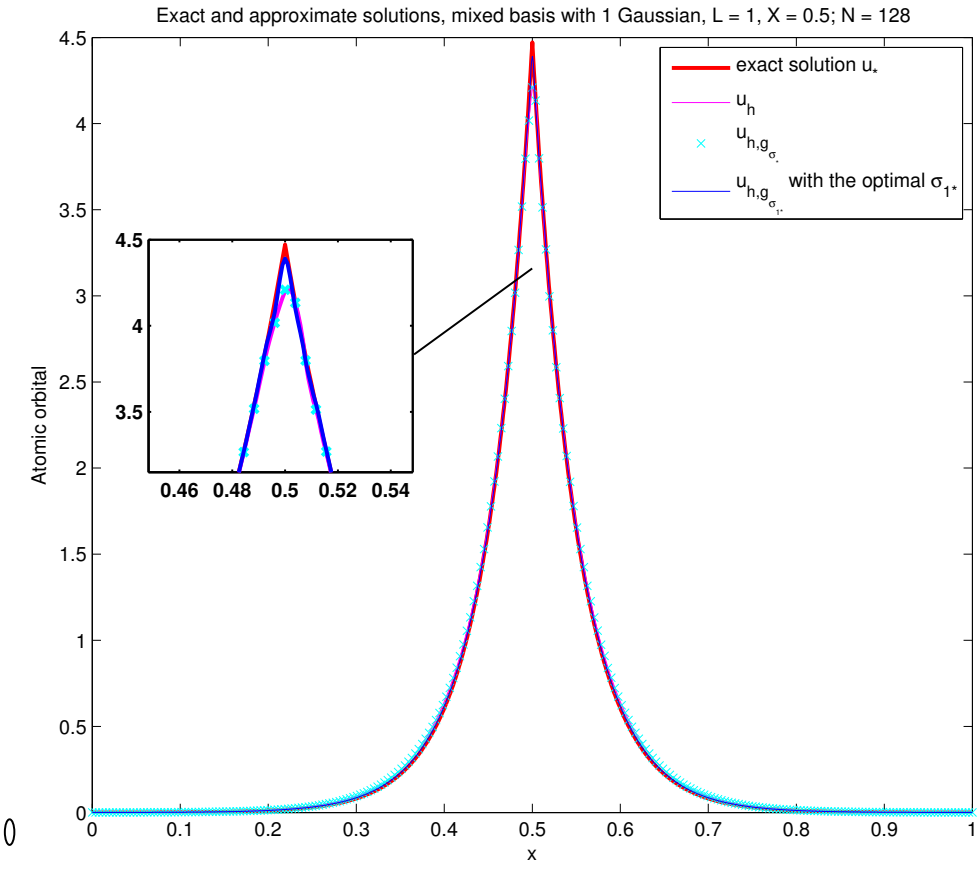

Figure 6.4: Exact and approximate wave functions, single-delta potential.

By definition, $\eta_{h,g_{\sigma_1},g_{\sigma_2}}$ is the error estimate of the solution $(u_{h,g_{\sigma_1},g_{\sigma_2}}, E_{h,g_{\sigma_1},g_{\sigma_2}})$ compared to $(u_{h,g_{\sigma_1}}, E_{h,g_{\sigma_1}})$. Once again, the two maximal arguments are very close:

$$r_{2*}^{\mathfrak{E}} = 0.00866519 \tag{6.28}$$

$$r_{2*}^{\eta} = 0.00866518 \tag{6.29}$$

Their proximity (difference of 10^{-8}) is better than when there was only one Gaussian in the basis. Take this optimal value r_{2*}^{η} and the respective Gaussian $\widetilde{g}_{\sigma_{2*}}$

$$\sigma_{2*} := r_{2*}^{\eta}\sigma_*,$$

and continue with the greedy algorithm, we obtain the following data tables.

Tables of numerical results

The notations are similar to those of Table 6.1. For the following tables, which correspond to the periodic equation on $[0, L]$ with a single-delta potential at X, over a mixed basis of scaling functions `db4` and Q Gaussians (centered at X), on a mesh of N points; we list the optimal dilation r for the Q-th Gaussian found by the greedy algorithm, assuming that the first $Q-1$ Gaussians were already found, **using the estimate criterion**.

As mentioned in chapter §4, because of physical units, we consider L/Λ and X/L instead of L, Λ or X alone. We check our paradigms in Table 6.2, where we fix

$$L/\Lambda = 20,\ X/L = 1/2,\ \epsilon_{tol} = 0.01\,.$$

The greedy algorithm stops when

$$\texttt{err}(E^{\eta}) < \epsilon_{tol},$$

which is a criterion that leads to maximum 4 Gaussians for $N = 2^5$, and 3 Gaussians for $N = 2^7$.

L/Λ	X/L	N	Q	$r_*^{\mathfrak{E}}$	r_*^{η}	err(r)	$\texttt{err}(E^{\mathfrak{E}})$	$\texttt{err}(E^{\eta})$	err(`err`)
20	1/2	2^5	1	0.184912	0.184543	2.0E-03	0.23213519	0.23213451	2.9E-06
			2	0.033204	0.033202	5.6E-05	0.05243842	0.05243842	1.8E-09
			3	0.572892	0.572903	1.9E-05	0.02090214	0.02090214	7.7E-10
			4	0.005756	0.005756	1.7E-06	0.00931463	0.00931463	1.4E-11
		2^7	1	0.048106	0.048104	2.9E-05	0.06053981	0.06053981	5.8E-10
			2	0.008665	0.008665	1.0E-06	0.01376319	0.01376319	1.3E-10
			3	0.154949	0.154949	4.1E-07	0.00586680	0.00586680	1.3E-11

Table 6.2: Single-delta, mixed bases with Q Gaussians functions.

Table 6.2 shows that, once the basis is augmented enough, or once the mesh is refined, $r_*^{\mathfrak{E}}$ and r_*^{η} are very close; more importantly, the energy decrease that they yield, $\texttt{err}(E^{\mathfrak{E}})$ and $\texttt{err}(E^{\eta})$, are almost identical, so we can use one in place of the other.

Table 6.3 lists the optimal dilations for the first and second added Gaussians, corresponding to two different charges Z, while other parameters are fixed. This data will be useful in the next section §6.2 for multi-delta potentials. We also see from the table that the numerical results do not move too much if the value of the charge Z is changed by a small quantity.

X/L	N	Q	L/Λ	$r_*^{\mathfrak{E}}$	r_*^{η}	err(r)	$\mathtt{err}(E^{\mathfrak{E}})$	$\mathtt{err}(E^{\eta})$	err$(\mathtt{err})$
1/2	2^7	1	20	0.048106	0.048104	2.9E-05	0.06053981	0.06053981	5.8E-10
			19	0.045710	0.045709	2.4E-05	0.05753699	0.05753699	4.5E-10
		2	20	0.008665	0.008665	1.0E-06	0.01376319	0.01376319	1.3E-10
			19	0.008234	0.008234	1.3E-06	0.01308025	0.01308025	3.0E-12

Table 6.3: Single delta, mixed bases with 1 or 2 Gaussians, with different charges $Z = 1/\Lambda$.

6.2 Optimal strategy in the case of double-delta potentials

When we put the case of double-delta potentials next to the case of single-delta potentials, it represents the reality of general molecules versus isolated atoms. If the bases found in the previous section are of good enough quality, we can consider their "transferability", i.e. their possibility of being used in other environments. It is worth to mention that chemists have been applying the bases constructed for isolated atoms to the case of general molecules; when nuclei in the molecule are far from each other, which is most of the time, the bases for atoms of the same charges can be used. This approach allows us to make full advantage of the results found in §6.1 and save computation time.

To justify this strategy, first we use Algorithm 2 in chapter §5 to search for new Gaussians by a partial greedy procedure, then we will build a new algorithm, compare it with Algorithm 2 and show that the new one is even more efficient.

6.2.1 By the partial greedy algorithm

For a double-delta potential at $X = [X_1; X_2]$, first, we are going to add Gaussians centered at the bigger cusp to the scaling function basis; then, once the error "at" the bigger cusp is reduced enough, we will move to the smaller cusp.

Without loss of generality we may assume that the bigger cusp is at X_1, or, equivalently, the charges satisfy

$$Z_1 \geq Z_2.$$

Figure 6.5 plots the two curves of the energy decay and the squared estimate when we add a first Gaussian g_{σ_1} at X_1. With $r_1 = \frac{\sigma_1}{\sigma_*}$ varying, the parameters are

$$Q = 1,\ N = 128,\ L = 1,\ X_1 = 1/2,\ X_2 = 3/4,\ Z = [20; 19],\ r_1 \in (0, 2].$$

We see that the curves are quite similar to Figure 6.3, in the case of a single-delta potential of a same charge.

For the following tables, which correspond to the periodic equation (3.58) over a mixed basis of scaling functions `db4` and Q_1 Gaussians at X_1 plus Q_2 Gaussians at X_2, on a mesh of N points, we list the $(Q_1 + Q_2)$-th optimal dilation r on each line, assuming that the first $Q_1 + Q_2 - 1$ dilations were already found on previous lines. Other notations are similar with the case of single-delta potentials. Table 6.4 lists the obtained data when we gradually add one Gaussian centered at X_1 or X_2 to the scaling function basis, until there are 2 Gaussians at each cusps.

There are several observations from Table 6.4 that might evolve our strategy, especially with the case of single-delta potentials in mind:

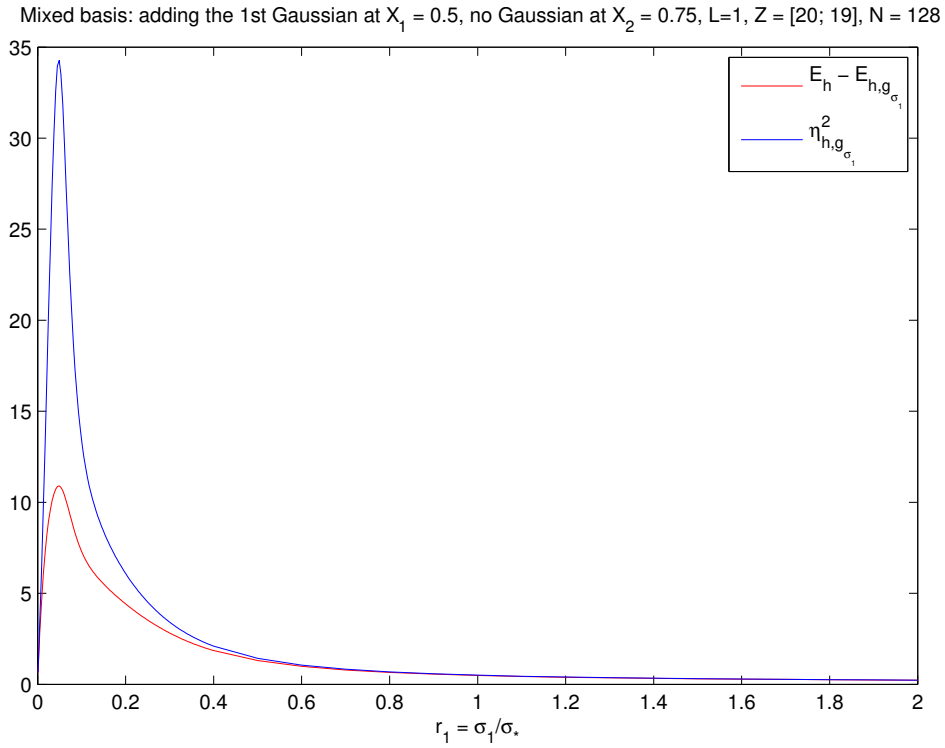

Figure 6.5: Energy decrease $E_h - E_{h,g_{\sigma_1}}$ and squared estimate $\eta^2_{h,g_{\sigma_1}}$ in terms of $r_1 = \frac{\sigma_1}{\sigma_*}$, mixed basis with 1 Gaussian $\widetilde{g}_{\sigma_1,X_1}$ at the bigger cusp, double-delta potential.

L/Λ	X/L	N	Q	$r_*^{\mathfrak{E}}$	r_*^{η}	err(r)	$\mathtt{err}(E^{\mathfrak{E}})$	$\mathtt{err}(E^{\eta})$	err$(\mathtt{err})$
20; 19	1/2; 3/4	2^7	1G+0G	0.048106	0.048104	2.9E-05	0.059091	0.059091	6.5E-10
			1G+1G	0.045697	0.045696	4.0E-05	0.000682	0.000682	1.8E-09
			2G+1G	0.008665	0.008665	3.4E-06	0.013425	0.013425	2.3E-07
			2G+2G	0.008231	0.008231	5.3E-06	0.000199	0.000199	1.5E-10

Table 6.4: Double-delta, mixed bases with $[Q_1; Q_2]$ Gaussian functions at $[X_1; X_2]$, added one by one from the previous line.

- Paradigm 6.1 and Paradigm 6.2 still hold true, i.e., the proximity of the two dilations by the energy criterion and the η criterion takes place.
- The partial greedy Algorithm 2 proves to be more effective than the full greedy Algorithm 1. In particular, adding an optimal Gaussian at the bigger cusp reduces the energy error much better than adding an optimal Gaussian at the smaller cusp.
- The most intriguing observation comes from comparing Table 6.4 to Table 6.3 in the single-delta case. In Table 6.4, at $x = X_1$ or $x = X_2$, the values of the optimal dilation are not very far from the optimal values in Table 6.3 with the corresponding charges.

Even when we test with different meshes (different N), the third observation holds true: the two sets of optimal r (for a double-delta potential and a single-delta one with the same charge) remain quite close. That fact is illustrated in Figure 6.6, where we plot the optimal r^η obtained for the first Gaussian in the single-delta case (the magenta line), for the second Gaussian in the single-delta case (the teal line), and for the first Gaussian at X_1 in the double-delta case (the blue stars). The parameters are:

$$L = 1,\ N = 16 \sim 128,\ Z = 20,\ X = L/2 \text{ in the single-delta case,}$$
$$Z = [20; 19],\ X = [L/2; 3L/4] \text{ in the double-delta case.}$$

We detect an "almost" linear dependence of the optimal r on the number of points N when other parameters are fixed; it helps predicting r for future cases. We also see that the blue stars are all on the magenta line, or **the bases for general molecules are very close to the bases for isolated atoms**.

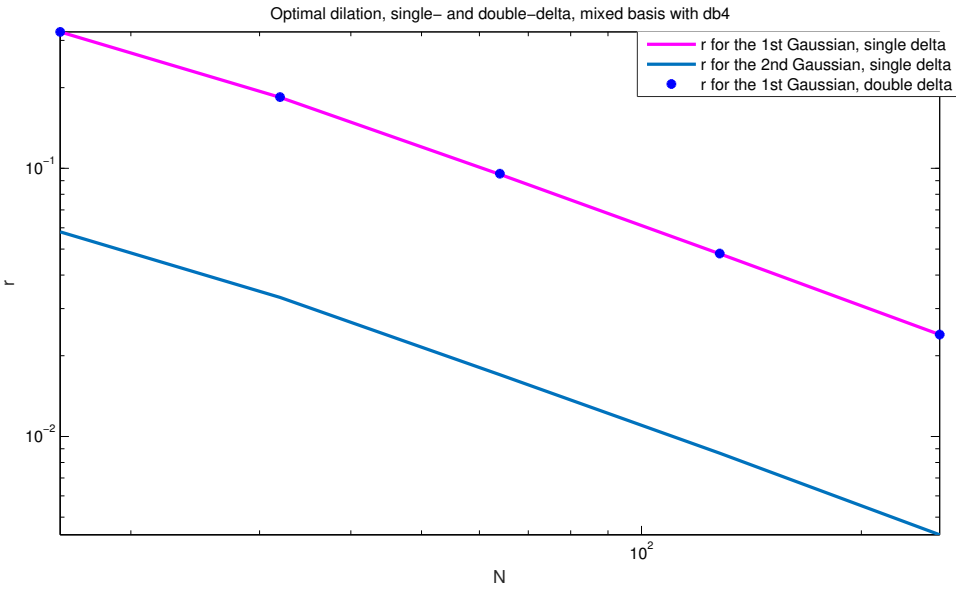

Figure 6.6: Dependence of the optimal r_*^η on the number of points N when the charge Z is fixed.

With this observation, we might be able to predict the bases for the double-delta case from the single-delta case, or even better, reuse the bases found in the single-delta case. We come up with a new algorithm which attemps to use equally Q' Gaussians at X_1 and Q' Gaussians at X_2 (with $Q = 2Q'$), of which the standard deviations are taken from the single-delta case, with proper re-scaling to the respective charges.

6.2.2 Algorithm of Independent Optimization

The new algorithm for the case of double-delta potentials is written as Algorithm 3. Then, we will show the explicit procedure on a concrete example.

Algorithm 3 Independent Optimization for double nuclei

1: **procedure** INDEPENDENT-DOUBLE($X_1, X_2, Z_1, Z_2, \epsilon_{tol}$)
2: $b := \{h\}$ ▷ $\mathscr{V}_h$: Existing basis of scaling function
3: $Q := 0$ ▷ Initialization
4: $Q' = 0$
5: **repeat**
6: Compute or retrieve (u_b, E_b)
7: Retrieve $\sigma^*_{Z_I}$ from Algorithm 2 ▷ Off-line, for the case $M = 1, Z = Z_I$
8: $\tilde{b} := b \cup \widetilde{g}_{\sigma^*_{Z_1},X_1} \cup \widetilde{g}_{\sigma^*_{Z_2},X_2}$ ▷ add new elements to basis
9: Compute $(u_{\tilde{b}}, E_{\tilde{b}})$
10: $\texttt{err} := (E_b - E_{\tilde{b}})/|E_b|$ ▷ relative energy decay
11: $b := \tilde{b}$
12: $Q' := Q' + 1$ ▷ number of Gaussians at each cusp
13: $Q := Q + 2$ ▷ total number of Gaussians
14: **until** ($\texttt{err} < \epsilon_{tol}$)
15: **end procedure**

Considering

$$Q' = 1,\ L = 1,\ N = 2^7,\ [X_1; X_2] = [\frac{1}{2}; \frac{3}{4}],\ [Z_1; Z_2] = [20; 19],$$

then, for a mixed 1G+1G basis by Algorithm 2, the added Gaussians should be $\{\widetilde{g}_{\sigma_1,X_1}, \widetilde{g}_{\sigma_2,X_2}\}$ where

$$\sigma_1 = 0.048104\,\frac{\sqrt{\pi}}{2}\Lambda_1,\ \sigma_2 = 0.045696\,\frac{\sqrt{\pi}}{2}\Lambda_2,\ \text{with } \Lambda_I = \frac{1}{Z_I},\ I = 1, 2.$$

as in Table 6.4, with respect to the corresponding charges Z_I. In the new algorithm, the added Gaussians would be $\{\widetilde{g}_{\sigma_1,X_1}, \widetilde{g}_{\sigma_2,X_2}\}$ where

$$\sigma_1 = 0.048104\,\frac{\sqrt{\pi}}{2}\Lambda_1,\ \sigma_2 = 0.045709\,\frac{\sqrt{\pi}}{2}\Lambda_2,$$

both numbers coming from Table 6.3 (the single-delta case) with respect to the corresponding charges Z_I. Comparisons between the two numerical results will be given in the next §6.2.3.

In general, at each step, the new algorithm adds 2 pre-calculated Gaussians to the basis, one Gaussian for each cusp. The energy decay is evaluated; the procedure stops

when the decay goes below some threshold ϵ_{tol}. This strategy has its limit of validity: it is only applicable if the two nuclei are "separated" from each other, or equivalently, if the internuclear distance R is big enough. This point will be made clear in §6.2.3.

6.2.3 Comparison between the old and new algorithms

In Table 6.5, we compare the results obtained by Algorithm 2 with the results obtained by Algorithm 3 (Independent Optimization), when the distance R between the two cusps is considered to be far enough, with respect to the charges. The energy decrease is not a suitable criterion to compare the two procedures, since Algorithm 3 does not *incrementally* find the sequence of Gaussians as Algorithm 2 does. Therefore, relative errors on energy

$$\texttt{Err}(E) := \frac{E_{h,g_{\sigma_1,X_\cdot},\ldots,g_{\sigma_Q,X_\cdot}} - E_*}{|E_*|} \tag{6.30}$$

and computation times are listed as to judge the advantages of each algorithm. For each transferred basis in Algorithm 3, a line is added in the table for the total time of off-line optimization and processing Algorithm 3. In practice, the optimization has been done beforehand for isolated atoms and will not be repeated. There is also a column err(`Err`) which is the relative difference between the errors of the two methods.

N	ZL	X/L	Q	Algorithm	`Err`(E)	err(`Err`)	t (s)
2^7	[20 ;19]	$[\frac{1}{2};\frac{3}{4}]$	1G+1G	Algorithm 2	0.024406		1009
				Algorithm 3 + off-line optim.	0.024406	1.8E-09	585
				Algorithm 3			10
			2G+2G	Algorithm 2	0.011112		1768
				Algorithm 3 + off-line optim.	0.011112	1.3E-05	835
				Algorithm 3			19

Table 6.5: Mixed bases with $[Q';Q']$ Gaussian functions at $[X_1;X_2]$.

In Table 6.5, Algorithm 3, which reuses the bases optimized for isolated atoms, produces an error of almost the same value as Algorithm 2 does, while enormously economizing computation time. But if the distance R is small in comparison with the charges (for example if $R/\max(\Lambda_1,\Lambda_2) < 1/2$), Algorithm 3 may not be so optimal. In Figure 6.7, we look at different R's and how they affect the gap between the two algorithms, with the same parameters as in the example in §6.2.2 except R:

$$Q' = 1,\ L = 1,\ N = 2^7,\ [Z_1;Z_2] = [20;19],\ R \in \left\{\frac{1}{64};\frac{1}{32};\frac{1}{16};\frac{1}{8};\frac{1}{4}\right\}.$$

The blue line represents err(`Err`), which is the relative difference between the `Err`(E), defined by (6.30) and obtained by Algorithm 2, and the `Err`(E) obtained by Algorithm 3. The dashed black line represents the former `Err`(E), since it is the energy error over the best basis available. We see that, the larger the distance R is, the better Algorithm 3 gets in term of closing the gap with Algorithm 2. Therefore, depending on our threshold ϵ_{tol}, we might choose to work with this new algorithm when $R/\max(\Lambda_1,\Lambda_2)$ is large enough.

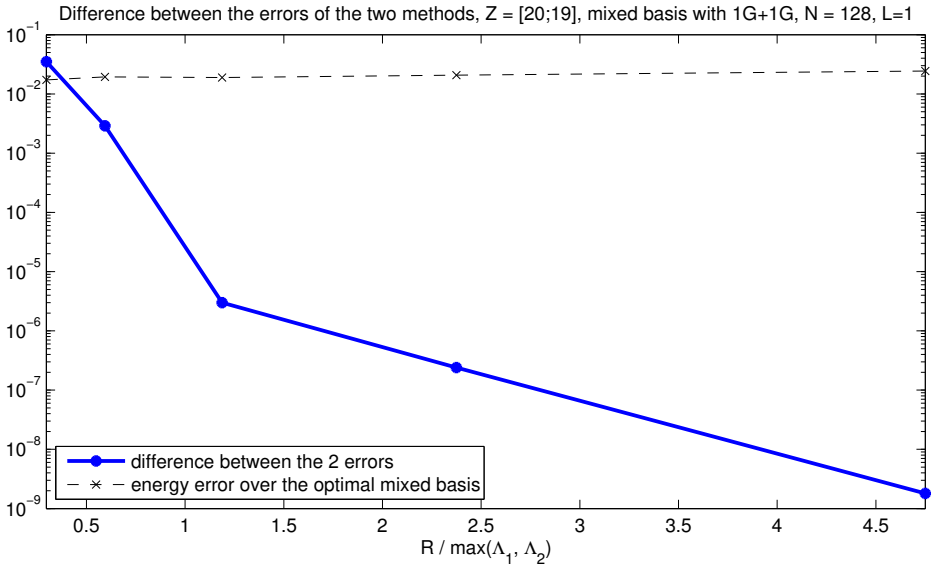

Figure 6.7: Difference between the energy errors of Algorithms 2 and 3.

Figure 6.8 plots the exact solution and the approximate one over a mixed basis with one optimal Gaussian at each delta position, obtained by Algorithm 3, with the same parameters as in §6.2.2:

$$Q' = 1,\ L = 1,\ N = 2^7,\ [Z_1; Z_2] = [20; 19],\ [X_1; X_2] = [1/2; 3/4].$$

The mixed 1G+1G approximation is very close to the exact atomic orbital; it gives the more reason to use Algorithm 3.

6.3 Methodology for multi-delta potentials

We can adapt Algorithm 3 to the case of multi-delta potentials, keeping in mind that there is a trade-off between accuracy and computation time. Suppose that the optimal Gaussians for mixed bases in the case of single-delta potentials have already been found. Then, in a slightly different environment, we transfer these Gaussians to the multi-delta case by Algorithm 4.

At each step, the new algorithm adds M pre-calculated Gaussians to the basis, one Gaussian for each cusp. The energy decay is evaluated; the procedure stops when the decay goes below some threshold ϵ_{tol}. The pre-calculated Gaussians are retrieved from Algorithm 2. The optimization process

$$\sigma_q^* := \arg\max_{\sigma_q \in \mathcal{I}} \eta_{b, \widetilde{g}_{\sigma_q, X_I}}, \tag{6.31}$$

for a fixed interval $\mathcal{I}$ in $\mathbb{R}_+^*$ and an isolated atom of nuclear charge Z_I, which is the part that takes up the most CPU cycles, has been done beforehand. The only task left in this approach is the resolution over the new mixed basis and the re-evaluation of the error, which costs very little compared to the former.

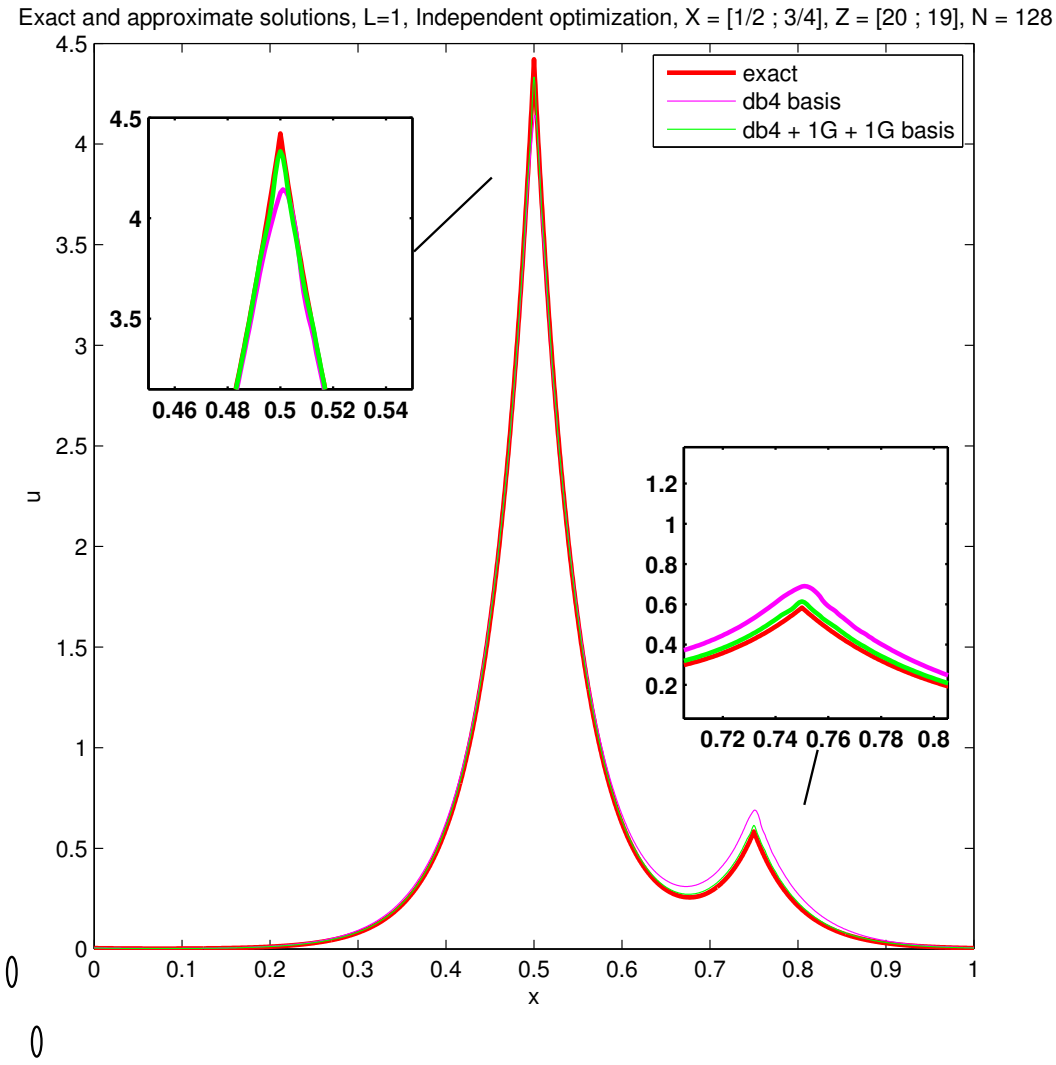

Figure 6.8: Exact and approximate wave functions, double delta.

Algorithm 4 Independent Optimization for multi nuclei

procedure INDEPENDENT(X_I, Z_I, ϵ_{tol})
  $b := \{h\}$ ▷ $\mathscr{V}_h$: Existing basis of scaling function
  $Q := 0$ ▷ Initialization
  $Q' = 0$
  repeat
    Compute or retrieve (u_b, E_b)
    Retrieve $\sigma^*_{Z_I}$ from Algorithm 2 ▷ Off-line, for the case $M = 1, Z = Z_I$
    $\tilde{b} := b \cup \bigcup_{I=1}^{M} \{\widetilde{g}_{\sigma^*_{Z_I}, X_I}\}$ ▷ add new elements to basis
    Compute $(u_{\tilde{b}}, E_{\tilde{b}})$
    `err` $:= (E_b - E_{\tilde{b}})/|E_b|$ ▷ relative energy decay
    $b := \tilde{b}$
    $Q' := Q' + 1$ ▷ number of Gaussians at each cusp
    $Q := Q + M$ ▷ total number of Gaussians
  until (`err` $< \epsilon_{tol}$)
end procedure

For example, in Table 6.5, instead of spending more than 1000s by Algorithm 2, we get a favorable approximation in 10s by Algorithm 3, of a difference in accuracy as small as 10^{-9}. Even if the optimal Gaussians for isolated atoms have not yet been calculated, i.e. the line (6.31) has to be counted into computation time, it is still more economic than Algorithm 2, as attested by Table 6.5 (585s instead of 1009s). Indeed, the step (6.31) evokes an optimization for single-delta potentials, which will be a lot faster than optimizations for multi-delta potentials where there are more degrees of freedom (Gaussians) in the basis, the matrices are more complex, which leads to much more CPU time.

However, if there are two nuclei in the molecule very close to each other, we shall use Algorithm 4 with cautions. As seen in chapter §3, when $R/\max(\Lambda_1, \Lambda_2)$ is small, the two cusps are not much different even if their charges are. In physical sense, when two nuclei get too close, they affect each other and the electron is pulled toward both of them. Transferring two different bases of isolated atoms to this case will not illustrate the situation. Moreover, numerical tests have been showing that the real optimized bases might be far more deviating from transferred bases in this case.

Figure 6.7 —the image of how Algorithm 4 catches up with the real optimization when $R/\max(\Lambda_1, \Lambda_2)$ increases— shows an idea how to proceed. It gives us a rough value of err(`Err`), the difference between the two methods, corresponding to the given R. If

$$\text{err}(\texttt{Err}) < \epsilon_{tol}$$

then we may consider that the two nuclei are far enough and use the Independent Optimization bases.

Other than the user-defined parameter ϵ_{tol}, a basis-dependent threshold could be available, given by the energy error `Err`. The dashed black line in Figure 6.7 gives an indication of the brink from which the difference between the two errors might become significant, as it would be higher than the errors themselves. We denote by R_* the value of R at which the two lines cross. When all the internuclear distances are greater than R_*, we may use

Algorithm 4; when a distance between two nuclei is too small, Algorithm 2 is the way to go.

Therefore, in the future, when we encounter a 1-D linear Schrödinger periodic equation with a multi-delta potential, with a pre-defined basis, we might calculate the *a posteriori* estimate and be assured that it is well-related to the energy decay. Then, using this estimate in Algorithm 2 or Algorithm 4 —depending on how close the nuclei are to each other, in reality they are often not— we can decide which one to choose between the two procedures, or, whether to trade-off the accuracy with computation time. With this approach we do not need many degrees of freedom and still have good approximations both at the cusp and the internuclear region.

Conclusion and perspectives

Summary of key results

Theoretical analysis of multi-delta 1-D models

In view of the objective initially stated in the Introduction, it might come as a surprise that so much room has been given for the mathematical analysis of the multi-delta 1-D models in chapter §3. Far from being a dry academic exercice, this "aperitif" has revealed the richness of the toy models considered and prepares in the best possible way the ground for the *a posteriori* estimates in chapter §5.

There is an almost perfect parallelism between the results for the infinite model and those for the periodic model. Let us single out the most prominent ones.

- The existence and uniqueness of the ground state (Theorems 3.1, 3.6 and Theorems 3.8, 3.13) provides us with the guarantee that we are working with good models, for which it is worth talking about a numerical approximation.
- The generic forms of the eigenfunctions (Theorem 3.3 and Theorem 3.10) shed light on the structure of the exact solutions and lead to a more refined knowledge of their Sobolev regularity (Corollary 3.2 and Corollary 3.5) as well as the special cases of one or two nuclei (Theorems 3.4, 3.5 and Theorems 3.11, 3.12).
- The lower and upper bounds on the fundamental energy (Theorem 3.7 and Theorem 3.14) are not only useful for defining an appropriate norm for the actual computation of the estimate (section §5.3.4) but also for the initialization of any numerical solver for the exact ground state.
- The strict separation between the first and second energy levels (Corollary 3.4 and Corollary 3.6) is a very favorable feature for the relevance of the *a posteriori* error estimates proposed in chapter §5.

In comparison with their counterparts for the infinite model, the formulae for the periodic model are lengthier and the proofs more intricate. The periodic model also contains several families of excited states with positive energies that we have not addressed, as this would take us too far from the scope of this dissertation.

Practical construction of the mixed basis

In response to the objective stated in the Introduction, we have brought a mathematically sound and computationally effective answer to the question of constructing the (contracted) Gaussians to be inserted into an existing scaling functions basis. The methology

we developed in chapter §5 to obtain this answer rests upon a combination of two tools commonly used in other areas: *a posteriori* estimates for the energy decay and the greedy algorithm. Numerical tests of chapter §6 attested to the feasibility of the wavelet-Gaussian mixed basis on the multi-delta 1-D models.

Of the two algorithms proposed in chapter §5, only Algorithm 2 was actually meant to be used in practice. Indeed, Algorithm 1 is a (too) direct application of the original greedy algorithm and appears unnatural to the chemists, in that it controls entirely the order of atoms to be visited. In Algorithm 2, some *a priori* knowledge was instilled in order to get back the control of the order of atoms to be visited and, simultaneously, to be more natural as well as less expensive. In this respect, perhaps the most astonishing fact is that Algorithm 3, discovered in the course of numerical experiments and later extended into Algorithm 4, performs much better in terms of efficiency, at least when the nuclei are far enough from each other. Algorithms 3 and 4, which essentially rely on the idea of *transferability* from atom to molecule, appear to be more empirical but also more in line with the habits of the chemists.

The dual residual norm (5.10) used to define the *a posteriori* estimate is a very classical notion that can be generalized quite easily to nonlinear models [50]. The difficulty for such a generalization, however, is to show that the proposed estimate is well related to the actual energy decay. In other words, the hard part is to establish a result similar to Corollary 5.1, in which the equivalence constants must depend only on b (the pre-existing basis) and not on g (the additional basis elements). We have been lucky enough to do so for the multi-delta 1-D models. Based on recent works such as [28, 30, 50], we are confident this should be also possible for nonlinear models.

Recommendations for future research

Nonlinear 1-D models

An immediate natural extension of the linear multi-delta 1-D models studied in this thesis is their nonlinear counterparts

$$-\frac{1}{2}u'' - \sum_{I=1}^{M} Z_I \delta_{X_I} u + \beta |u|^2 u = \Upsilon u, \tag{6.32a}$$

$$\|u\|_{L^2} = 1, \tag{6.32b}$$

obtained by adding the Gross-Pitaevskii term $\beta|u|^2u$ to the left-hand side for some $\beta \in \mathbb{R}$. The ground state solution (u_*, Υ_*) would then minimize the energy functional

$$\mathfrak{E}(u) = \frac{1}{2}\int |u|^2 - \sum_{I=1}^{M} Z_I |u(X_I)|^2 + \frac{\beta}{2}\int |u|^4. \tag{6.33}$$

The difficulties enumerated below are expected.

- If $\beta \neq 0$, then $\mathfrak{E}(u_*) \neq \Upsilon_*$. In other words, the energy level $E = \mathfrak{E}(u)$ and the eigenvalue Υ are no longer the same quantity. This makes the error analysis more delicate.

- If (u_b, Υ_b) is a Galerkin approximation of (u_*, Υ_*), then

$$E_b - E_* := \mathfrak{E}(u_b) - \mathfrak{E}(u_*) \not\simeq \|u_b - u_*\|_{H^1}^2$$

  for $\beta \neq 0$. Instead, according to existing works on similar nonlinear problems [17,50], we would have

$$E_b - E_* \simeq \|u_b - u_*\|_{H^1}.$$

  This deteriorates the convergence of the energy level, but the latter can still be used as the guiding principle for the construction of mixed bases.

- The numerical computation of (u_b, Υ_b) in any basis involves a SCF (Self-Consistent Field) loop, which is similar to what was explained in §1.2.3 and for which appropriate algorithms should be devised in the same spirit as those reviewed in [16] and [22, §6.2.5]. This difficulty was observed by Duchêne [48] using a pure P_1 basis in a preliminary study of the single-delta equation

$$-\frac{1}{2}u'' - Z\delta_0 u + \beta|u|^2 u = \Upsilon u, \tag{6.34a}$$

$$\|u\|_{L^2} = 1, \tag{6.34b}$$

  whose ground state solution (u_*, Υ_*) exists and is analytically known for $\beta \leq 2Z$. As β approaches $2Z$, it becomes slower and slower for any attempted SCF algorithm to converge, including those inspired from [20, 21, 24].

To achieve a higher level of model sophistication while remaining in 1-D, it can be envisaged a system with two unknown functions. The simplest instance of such systems reads

$$-\frac{1}{2}\varphi_1'' - \sum_{I=1}^{M} Z_I \delta_{X_I}\varphi_1 + \beta\big(|\varphi_1|^2 + |\varphi_2|^2\big)\varphi_1 = \Upsilon_1\varphi_1, \tag{6.35a}$$

$$-\frac{1}{2}\varphi_2'' - \sum_{I=1}^{M} Z_I \delta_{X_I}\varphi_2 + \beta\big(|\varphi_1|^2 + |\varphi_2|^2\big)\varphi_2 = \Upsilon_2\varphi_2, \tag{6.35b}$$

$$\langle\varphi_i, \varphi_j\rangle_{L^2} = \delta_{ij}, \tag{6.35c}$$

which results from simplifying the 3-D Hartree-Fock model for two electrons, represented by the molecular orbitals φ_1 and φ_2. System (6.35) enables us to get closer to a "real-life" chemical system. In particular, it provides us with the opportunity to go beyond the s-channel and extend the construction of mixed bases to the higher-channels (p, d, f...).

Nonlinear, semilocal 3-D DFT models

In our opinion, rather than spending too much time with the toy system (6.35), it is more rewarding to study mixed bases on traditional 3-D DFT models with semilocal functionals, as it would already represent a considerable achievement. For this kind of 3-D models, we have indeed identified two challenges.

1. When the molecular orbitals φ are expanded in a mixed basis, it is crucial to come up with a smart way to *accurately express the charge density* ρ_Φ, defined in (1.31), as well as the corresponding nonlinear potential

$$\rho_\Phi \star \frac{1}{|\mathbf{x}|} + v_{\mathrm{xc}}(\rho_\Phi).$$

   This issue was pointed out by Longo [95, §10.2.3] as the very first step in the journey towards merging mixed-basis prototypes into the BigDFT production code. In BigDFT, the density is known from its discrete values on a grid using the magic filter [59, 110], which creates a map between Daubechies scaling functions and the Deslauriers-Dubuc interpolating scaling functions. It turns out that the contribution of the Gaussians to the density cannot be collocated on the same grid, because the latter might be too sparse to capture a narrow Gaussian.

2. As mentioned earlier for 1-D nonlinear models, it is capital to have a *competitive SCF algorithm* for the nonlinear eigenvalue system (1.35). Currently, the SCF algorithm used in BigDFT is a direct minimization of the total energy, accelerated by either a preconditioned steepest-descent algorithm a preconditioned DIIS (Direct Inversion in the Iterative Subspace) method [22, §6.2.5]. For this SCF algorithm to work well, a *sine qua non* ingredient is a good preconditioner. Such a preconditioner was especially designed [60] for a pure wavelet basis using a low-pass filter. This preconditioner cannot be applied as such to a mixed basis, where high-frequency contents are naturally introduced through the (contracted) Gaussians. The design of a new preconditioner suitable to a mixed basis appears to be an ambitious task that would by itself justify a new PhD thesis!

Bibliography

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions: with Formulas, Graphs and Mathematical Tables*, Dover Publications, New York, 1970. 107, 113

[2] R. A. Adams, *Sobolev Spaces*, vol. 65 of Pure and Applied Mathematics, Academic Press, New York, 1975. 78, 80, 82, 101, 103, 104

[3] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, no. 37 in Pure and Applied Mathematics, Wiley, 2000. 172

[4] A. Anantharaman and E. Cancès, *Existence of minimizers for Kohn-Sham models in quantum chemistry*, Ann. Inst. H. Poincaré, 26 (2009), pp. 2425–2455, https://doi.org/10.1016/j.anihpc.2009.06.003. 29

[5] T. A. Arias, *Multiresolution analysis of electronic structure: semicardinal and wavelet bases*, Rev. Mod. Phys., 71 (1999), pp. 267–311, https://doi.org/10.1103/RevModPhys.71.267. 41

[6] M. Aubert, N. Bessis, and G. Bessis, *Prolate-spheroidal orbitals for homonuclear and heteronuclear diatomic molecules. II. Shielding effects for the two-electron problem*, Phys. Rev. A, 10 (1974), pp. 61–70, https://doi.org/10.1103/PhysRevA.10.61. 91

[7] I. Babuška and J. Osborn, *Eigenvalue problems*, in Finite Element Methods (Part 1), P. G. Ciarlet and J. L. Lions, eds., vol. 2 of Handbook of Numerical Analysis, Elsevier, Amsterdam, 1991, pp. 64–787, https://doi.org/10.1016/S1570-8659(05)80042-0. 121

[8] G. Bao, G. Hu, and D. Liu, *An h-adaptive finite element solver for the calculations of the electronic structures*, J. Comput. Phys., 231 (2012), pp. 4967–4979, https://doi.org/10.1016/j.jcp.2012.04.002. 15

[9] G. Bao, G. Hu, and D. Liu, *Numerical solution of the Kohn-Sham equation by finite element methods with an adaptive mesh redistribution technique*, J. Sci. Comput., 55 (2013), pp. 372–391, https://doi.org/10.1007/s10915-012-9636-1. 15

[10] G. Beylkin, *On the representation of operators in bases of compactly supported wavelets*, SIAM J. Numer. Anal., 29 (1992), pp. 1716–1740, https://doi.org/10.1137/0729097. 55, 62, 65

[11] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms I*, Comm. Pure Appl. Math., 44 (1991), pp. 141–183, https://doi.org/10.1002/cpa.3160440202. 15

[12] G. Beylkin and R. Cramer, *A multiresolution approach to regularization of singular operators and fast summation*, SIAM J. Sci. Comput., 24 (2002), pp. 81–117, https://doi.org/10.1137/S1064827500379227. 15

[13] D. R. Bowler and T. Miyazaki, *$O(N)$ methods in electronic structure calculations*, Rep. Prog. Phys., 75 (2012), p. 036503, https://doi.org/10.1088/0034-4885/75/3/036503. 13, 41

[14] S. F. Boys, *Electronic wave functions. I. A general method of calculation for the stationary states of any molecular system*, Proc. R. Soc. Lond. A, 200 (1950), pp. 542–554, https://doi.org/10.1098/rspa.1950.0036. 13, 34

[15] M. Brewster, G. I. Fann, and Z. Yang, *Wavelets for electronic structure calculations*, J. Math. Chem., 22 (1997), pp. 117–142, https://doi.org/10.1023/A:1019171830287. 41

[16] E. Cancès, *Self-Consistent Field (SCF) algorithms*, in Encyclopedia of Applied and Computational Mathematics, B. Engquist, ed., Springer, Berlin, Heidelberg, 2015, pp. 1310–1316, https://doi.org/10.1007/978-3-540-70529-1_256. 31, 211

[17] E. Cancès, R. Chakir, and Y. Maday, *Numerical analysis of nonlinear eigenvalue problems*, J. Sci. Comput., 45 (2010), pp. 90–117, https://doi.org/10.1007/s10915-010-9358-1. 17, 173, 211

[18] E. Cancès, R. Chakir, and Y. Maday, *Numerical analysis of the planewave discretization of orbital-free and Kohn-Sham models*, M2AN, 46 (2012), pp. 341–388, https://doi.org/10.1051/m2an/2011038. 17

[19] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday, *Computational quantum chemistry: A primer*, in Special Volume: Computational Chemistry, P. G. Ciarlet and C. Le Bris, eds., vol. X of Handbook of Numerical Analysis, North-Holland, Elsevier, Amsterdam, 2003, pp. 3–270, https://doi.org/10.1016/S1570-8659(03)10003-8. 13, 27, 29, 35

[20] E. Cancès and C. Le Bris, *Can we outperform the DIIS approach for electronic structure calculations?*, Int. J. Quant. Chem., 79 (2000), pp. 82–90, https://doi.org/10.1002/1097-461X(2000)79:2<82::AID-QUA3>3.0.CO;2-I. 31, 211

[21] E. Cancès and C. Le Bris, *On the convergence of SCF algorithms for the Hartree-Fock equations*, M2AN, 34 (2000), pp. 749–774, https://doi.org/10.1051/m2an:2000102. 31, 211

[22] E. Cancès, C. Le Bris, and Y. Maday, *Méthodes mathématiques en chimie quantique*, vol. 53 of Mathématiques et Applications, Springer, Berlin Heidelberg, February 2006. 13, 23, 31, 79, 82, 95, 103, 211, 212

[23] E. Cancès and M. Lewin, *Modèles à N corps*. Notes du cours M2 : Méthodes variationnelles en physique quantique, Février 2008, http://lewin.u-cergy.fr/M2/Ncorps.pdf. 23

[24] E. Cancès and K. Pernal, *Projected gradient algorithms for Hartree-Fock and density-matrix functional theory*, J. Chem. Phys., 128 (2008), p. 134108, https://doi.org/10.1063/1.2888550. 31, 211

[25] K. Capelle, *A bird's-eye view of density-functional theory*, Braz. J. Phys., 36 (2006), pp. 1318–1343, https://doi.org/10.1590/S0103-97332006000700035. 13

[26] R. Chakir, *Contribution à l'analyse numérique de quelques problèmes en chimie quantique et mécanique*, PhD thesis, Université Pierre et Marie Curie, 2009, https://tel.archives-ouvertes.fr/tel-00459149/. 133

[27] C. Chauvin, *Les ondelettes comme fonctions de base dans le calcul des structures électroniques*, PhD thesis, Institut National Polytechnique de Grenoble, Novembre 2005, http://tel.archives-ouvertes.fr/tel-00139349/. 14, 41

[28] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou, *Adaptive finite element approximations for Kohn-Sham models*, Multiscale Model. Simul., 12 (2014), pp. 1828–1869, https://doi.org/10.1137/130916096. 15, 17, 210

[29] H. Chen, X. Gong, L. He, Z. Yang, and A. Zhou, *Numerical analysis of finite dimensional approximations of Kohn–Sham models*, Adv. Comput. Math., 38 (2013), pp. 225–256, https://doi.org/10.1007/s10444-011-9235-y. 17

[30] H. Chen, L. He, and A. Zhou, *Finite element approximations of nonlinear eigenvalue problems in quantum physics*, Comput. Meth. Appl. Mech. Engrg, 200 (2011), pp. 1846–1865, https://doi.org/10.1016/j.cma.2011.02.008. 15, 17, 210

[31] C. Chizallet, S. Lazare, D. Bazer-Bachi, F. Bonnier, V. Lecocq, E. Soyer, A.-A. Quoineaud, and N. Bats, *Catalysis of transesterification by a nonfunctionalized metal-organic framework: Acido-basicity at the external surface of ZIF-8 probed by FTIR and ab initio calculations*, J. Am. Chem. Soc., 132 (2010), pp. 12365–12377, https://doi.org/10.1021/ja103365s. 13

[32] K. Cho, T. A. Arias, J. D. Joannopoulos, and P. K. Lam, *Wavelets in electronic structure calculations*, Phys. Rev. Lett., 71 (1993), pp. 1808–1811, https://doi.org/10.1103/PhysRevLett.71.1808. 41

[33] C. K. Chui, *An introduction to wavelets*, vol. 1 of Wavelet Analysis and its Applications, Academic Press, Boston, 1992. 105

[34] C. K. Chui and J.-Z. Wang, *A cardinal spline approach to wavelets*, Proc. Amer. Math. Soc., 113 (1991), pp. 785–793, https://doi.org/10.1090/S0002-9939-1991-1077784-X. 43

[35] A. Cohen, *Ondelettes, analyses multirésolutions et filtres miroirs en quadrature*, Ann. Inst. H. Poincaré, Analyse non linéaire, 7 (1990), pp. 439–459, http://www.numdam.org/item?id=AIHPC_1990__7_5_439_0. 47

[36] A. Cohen, *Numerical analysis of wavelet methods*, vol. 32 of Studies in Mathematics and Its Applications, North-Holland, Elsevier, Amsterdam, April 2003, http://www.sciencedirect.com/science/bookseries/01682024/32. 53

[37] A. Cohen and I. Daubechies, *A new technique to estimate the regularity of refinable functions*, Rev. Mat. Iberoamericana, 12 (1996), pp. 527–591. 56

[38] A. Cohen, I. Daubechies, and J.-C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560, https://doi.org/10.1002/cpa.3160450502. 43

[39] M. Crouzeix and A. L. Mignot, *Analyse numérique des équations différentielles*, Collection Mathématiques Appliquées pour la Maîtrise, Masson, Paris, 1984. 158

[40] W. Dahmen and C. A. Micchelli, *Using the refinement equation for evaluating integrals of wavelets*, SIAM J. Numer. Anal., 30 (1993), pp. 507–537, https://doi.org/10.1137/0730024. 62

[41] X. Dai, J. Xu, and A. Zhou, *Convergence and optimal complexity of adaptive finite element eigenvalue computations*, Numer. Math., 110 (2008), pp. 313–355, https://doi.org/10.1007/s00211-008-0169-3. 15

[42] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996, https://doi.org/10.1002/cpa.3160410705. 43, 47, 53

[43] I. Daubechies, *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, 1992, https://doi.org/10.1137/1.9781611970104. 41, 45, 46, 49, 54, 55, 61, 157

[44] I. Daubechies and J. C. Lagarias, *Two-scale difference equations. II. Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079, https://doi.org/10.1137/0523059. 56

[45] I. P. Daykov, T. A. Arias, and T. D. Engeness, *Robust ab initio calculation of condensed matter: Transparent convergence through semicardinal multiresolution analysis*, Phys. Rev. Lett., 90 (2003), p. 216402, https://doi.org/10.1103/PhysRevLett.90.216402. 41

[46] T. J. M. de Bruin, L. Magna, P. Raybaud, and H. Toulhoat, *Hemilabile ligand induced selectivity: a DFT study on ethylene trimerization catalyzed by titanium complexes*, Organometallics, 22 (2003), pp. 3404–3413, https://doi.org/10.1021/om030255w. 13

[47] M. de Llano, A. Salazar, and M. A. Solís, *Two-dimensional delta potential wells and condensed-matter physics*, Rev. Mex. Fis., 51 (2005), pp. 626–632, http://ref.scielo.org/dnbpw6. 77

[48] P. Duchêne, *Résolution numérique d'une équation de type Schrödinger 1-D non linéaire*. Note de travail IFPEN, avril 2013. 211

[49] P. Duchêne, *Calcul numérique du produit scalaire entre une gaussienne et la fonction d'échelle d'une ondelette*. Note de travail IFPEN, juin 2014. 161, 162

[50] G. Dusson and Y. Maday, *A posteriori analysis of a nonlinear Gross-Pitaevskii type eigenvalue problem*, IMA J. Numer. Anal., 37 (2017), pp. 94–137, https://doi.org/10.1093/imanum/drw001. 17, 19, 167, 169, 173, 210, 211

[51] T. D. Engeness and T. A. Arias, *Multiresolution analysis for efficient, high precision all-electron density-functional calculations*, Phys. Rev. B, 65 (2002), p. 165106, https://doi.org/10.1103/PhysRevB.65.165106. 41

[52] R. N. Euwema, *Rapid convergence of crystalline energy bands by use of a plane-wave-Gaussian mixed basis set*, Int. J. Quant. Chem., 5 (1971), pp. 471–487, https://doi.org/10.1002/qua.560050855. 16

[53] D. F. Feller and K. Ruedenberg, *Systematic approach to extended even-tempered orbital bases for atomic and molecular calculations*, Theor. Chim. Acta, 52 (1979), pp. 231–251, https://doi.org/10.1007/BF00547681. 125

[54] P. Fischer and M. Defranceschi, *Looking at atomic orbitals through Fourier and wavelet transforms*, Int. J. Quant. Chem., 45 (1993), pp. 619–636, https://doi.org/10.1002/qua.560450612. 41

[55] P. Fischer and M. Defranceschi, *Iterative process for solving Hartree–Fock equations by means of a wavelet transform*, Appl. Comput. Harm. Anal., 1 (1994), pp. 232–241, https://doi.org/10.1006/acha.1994.1010. 41

[56] G. J. Fix, *Eigenvalue approximation by the finite element method*, Adv. Math., 10 (1973), pp. 300–316, https://doi.org/10.1016/0001-8708(73)90113-8. 121

[57] A. A. Frost, *Delta-function model. I. Electronic energies of hydrogen-like atoms and diatomic molecules*, J. Chem. Phys., 1150 (1956), https://doi.org/10.1063/1.1743167. 18, 75, 76

[58] S. Geltman, *Bound states in delta function potentials*, J. Atom. Molec. Opt. Phys., (2011), pp. 1–4, https://doi.org/10.1155/2011/573179. 77

[59] L. Genovese and T. Deutsch, *Multipole-preserving quadratures for the discretization of functions in real-space electronic structure calculations*, Phys. Chem. Chem. Phys., 17 (2015), pp. 31582–31591, https://doi.org/10.1039/C5CP01236H. 157, 212

[60] L. Genovese, A. Neelov, S. Goedecker, T. Deutsch, S. A. Ghasemi, A. Willand, D. Caliste, O. Zilberberg, M. Rayson, A. Bergman, and R. Schneider, *Daubechies wavelets as a basis set for density functional pseudopotential calculations*, J. Chem. Phys., 129 (2008), p. 014109, https://doi.org/10.1063/1.2949547. 13, 41, 71, 72, 212

[61] L. Genovese, M. Ospici, T. Deutsch, J.-F. Méhaut, A. Neelov, and S. Goedecker, *Density functional theory calculation on many-cores hybrid central processing unit-graphic processing unit architectures*, J. Chem. Phys., 131 (2009), p. 034103, https://doi.org/10.1063/1.3166140. 14, 41

[62] S. Goedecker, *Wavelets and their application for the solution of partial differential equations in physics*, vol. 4 of Cahiers de physique, Presses Polytechniques et Universitaires Romandes, Lausanne, 1998. 63

[63] S. Goedecker, *Linear scaling electronic structure methods*, Rev. Mod. Phys., 71 (1999), pp. 1085–1123, https://doi.org/10.1103/RevModPhys.71.1085. 13, 41

[64] S. Goedecker and O. V. Ivanov, *Linear scaling solution of the Coulomb problem using wavelets*, Solid State Commun., 105 (1998), pp. 665–669, https://doi.org/10.1016/S0038-1098(97)10241-1. 41, 71

[65] S. Goedecker and O. V. Ivanov, *Frequency localization properties of the density matrix and its resulting hypersparsity in a wavelet representation*, Phys. Rev. B, 59 (1999), pp. 7270–7273, https://doi.org/10.1103/PhysRevB.59.7270. 41, 71

[66] A. S. P. Gomes and R. Custodio, *Exact Gaussian expansions of Slater-type atomic orbitals*, J. Comput. Chem., 23 (2002), pp. 1007–1012, https://doi.org/10.1002/jcc.10090. 125

[67] X. Gonze, B. Amadon, P.-M. Anglade, J.-M. Beuken, F. Bottin, P. Boulanger, F. Bruneval, D. Caliste, R. Caracas, M. Côté, T. Deutsch, L. Genovese, P. Ghosez, M. Giantomassi, S. Goedecker, D. R. Hamann, P. Hermet, F. Jollet, G. Jomard, S. Leroux, M. Mancini, S. Mazevet, M. J. T. Oliveira, G. Onida, Y. Pouillon, T. Rangel, G.-M. Rignanese, D. Sangalli, R. Shaltaf, M. Torrent, M. J. Verstraete, G. Zerah, and J. W. Zwanziger, *ABINIT: First-principles approach to material and nanosystem properties*, Comput. Phys. Commun., 180 (2009), pp. 2582–2615, https://doi.org/10.1016/j.cpc.2009.07.007. 13

[68] R. J. Harrison, G. I. Fann, T. Yanai, Z. Gan, and G. Beylkin, *Multiresolution quantum chemistry: Basic theory and initial applications*, J. Chem. Phys., 121 (2004), pp. 11587–11598, https://doi.org/10.1063/1.1791051. 15, 41

[69] W. J. Hehre, W. A. Lathan, R. Ditchfield, M. D. Newton, and J. A. Pople, *Gaussian 70, quantum chemistry program exchange, program no. 237*, 1970. 13

[70] W. J. Hehre, R. F. Stewart, and J. A. Pople, *Self-consistent molecular-orbital methods. I. Use of Gaussian expansions of Slater-type atomic orbitals*, J. Chem. Phys., 51 (1969), pp. 2657–2664, https://doi.org/10.1063/1.1672392. 13, 36, 124

[71] J. S. Hesthaven, G. Rozza, and B. Stamm, *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*, SpringerBriefs in Mathematics, Springer, Heidelberg, 2016, https://doi.org/10.1007/978-3-319-22470-1. 17, 169, 170

[72] M. Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, and T. Østergaard Sørensen, *Electron wavefunctions and densities for atoms*, Ann. Inst. H. Poincaré, 2 (2001), pp. 77–100, https://doi.org/10.1007/PL00001033. 25

[73] M. Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, and H. Stremnitzer, *Local properties of Coulombic wave functions*, Commun. Math. Phys., 163 (1994), pp. 185–215, https://doi.org/10.1007/BF02101740. 25

[74] P. Hohenberg and W. Kohn, *Inhomogeneous electron gas*, Phys. Rev., 136 (1964), pp. B864–B871, https://doi.org/10.1103/PhysRev.136.B864. 27

[75] E. Hunsicker, V. Nistor, and J. O. Sofo, *Analysis of periodic Schrödinger operators: Regularity and approximation of eigenfunctions*, J. Math. Phys., 49 (2008), p. 083501, https://doi.org/http://dx.doi.org/10.1063/1.2957940. 15

[76] S. Huzinaga, *Gaussian-type functions for polyatomic systems. I*, J. Chem. Phys., 42 (1965), pp. 1293–1302, https://doi.org/10.1063/1.1696113. 124

[77] L. Jacques, L. Duval, C. Chaux, and G. Peyré, *A panorama on multiscale geometric representations, intertwining spatial, directional and frequency selectivity*, Signal Process., 91 (2011), pp. 2699–2730, https://doi.org/10.1016/j.sigpro.2011.04.025. 71

[78] T. Kato, *Fundamental properties of Hamiltonian operators of Schrödinger type*, Trans. Amer. Math. Soc., 70 (1951), pp. 195–211, https://doi.org/10.1090/S0002-9947-1951-0041010-X. 23

[79] T. Kato, *On the eigenfunctions of many-particle systems in quantum mechanics*, Comm. Pure Appl. Math., 10 (1957), pp. 151–177, https://doi.org/10.1002/cpa.3160100201. 15, 25, 77

[80] T. Kato, *Perturbation Theory for Linear Operators*, vol. 132 of Grundlehren der mathematischen Wissenschaften, Springer, Berlin, 1980. 23

[81] B. Klahn and W. A. Bingel, *The convergence of the Rayleigh-Ritz method in quantum chemistry. I. The criteria of convergence*, Theor. Chim. Acta, 44 (1977), pp. 9–26, https://doi.org/10.1007/BF00548026. 125

[82] B. Klahn and W. A. Bingel, *The convergence of the Rayleigh-Ritz method in quantum chemistry. II. Investigation of the convergence for special systems of Slater, Gauss and two-electron functions*, Theor. Chim. Acta, 44 (1977), pp. 27–43, https://doi.org/10.1007/BF00548027. 125

[83] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev., 140 (1965), pp. A1133–A1138, https://doi.org/10.1103/PhysRev.140.A1133. 28

[84] G. Kresse and J. Furthmüller, *Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set*, Phys. Rev. B, 54 (1996), pp. 11169–11186, https://doi.org/10.1103/PhysRevB.54.11169. 13

[85] H.-C. Kreusler and H. Yserentant, *The mixed regularity of electronic wave functions in fractional order and weighted Sobolev spaces*, Numer. Math., 121 (2012), pp. 781–802, https://doi.org/10.1007/s00211-012-0447-y. 26

[86] W. Kutzelnigg, *Theory of the expansion of wave functions in a Gaussian basis*, Int. J. Quant. Chem., 51 (1994), pp. 447–463, https://doi.org/10.1002/qua.560510612. 125

[87] B. Langwallner, C. Ortner, and E. Süli, *Existence and convergence results for the Galerkin approximation of an electronic density functional*, Math. Models Methods. Appl. Sci., 20 (2010), pp. 2237–2265, https://doi.org/10.1142/S021820251000491X. 17

[88] C. Le Bris, *Quelques problèmes mathématiques en chimie quantique moléculaire*, PhD thesis, École Polytechnique, 1993, http://cat.inist.fr/?aModele=afficheN&cpsidt=158663. 29

[89] Z. H. Levine and J. W. Wilkins, *An energy-minimizing mesh for the Schrödinger equation*, J. Comput. Phys., 83 (1989), pp. 361–372, https://doi.org/10.1016/0021-9991(89)90124-1. 15

[90] H. Li and V. Nistor, *Analysis of a modified Schrödinger operator in 2D: Regularity, index, and FEM*, J. Comput. Appl. Math., 224 (2009), pp. 320–338, https://doi.org/10.1016/j.cam.2008.05.009. 15

[91] E. H. Lieb, *Thomas-Fermi and related theories of atoms and molecules*, Rev. Mod. Phys., 53 (1981), pp. 603–641, https://doi.org/10.1103/RevModPhys.53.603. 27

[92] E. H. Lieb, *Density functionals for Coulomb systems*, Int. J. Quant. Chem., 24 (1983), pp. 243–277, https://doi.org/10.1002/qua.560240302. 27

[93] E. H. Lieb, *Bound on the maximum negative ionization of atoms and molecules*, Phys. Rev. A, 29 (1984), pp. 3018–3028, https://doi.org/10.1103/PhysRevA.29.3018. 24

[94] G. Lippert, J. Hutter, and M. Parrinello, *A hybrid Gaussian and plane wave density functional scheme*, Mol. Phys., 92 (1997), pp. 477–488, https://doi.org/10.1080/002689797170220. 16

[95] F. Longo, *Gaussian and wavelet bases in electronic structure calculations*, PhD thesis, Politecnico di Torino, March 2011. 16, 71, 124, 212

[96] Y. Maday, *$h - P$ finite element approximation for full-potential electronic structure calculations*, Chin. Ann. Math., Ser. B, 35 (2014), pp. 1–24, https://doi.org/10.1007/s11401-013-0819-3. 15

[97] Y. Maday, *Numerical analysis of eigenproblems for electronic structure calculations*, in Encyclopedia of Applied and Computational Mathematics, B. Engquist, ed., Springer, Berlin, Heidelberg, 2015, pp. 1042–1047, https://doi.org/10.1007/978-3-540-70529-1_258. 17

[98] Y. Maday, *A priori and a posteriori error analysis in chemistry*, in Encyclopedia of Applied and Computational Mathematics, B. Engquist, ed., Springer, Berlin, Heidelberg, 2015, pp. 5–10, https://doi.org/10.1007/978-3-540-70529-1_255. 17

[99] Y. Maday and G. Turinici, *Error bars and quadratically convergent methods for the numerical simulation of the Hartree-Fock equations*, Numer. Math., 94 (2003), pp. 739–770, https://doi.org/10.1007/s002110100358. 17

[100] S. Mallat, *Multiresolution approximation and wavelet orthonormal bases of $L^2(\mathbb{R})$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87, https://doi.org/10.1090/S0002-9947-1989-1008470-5. 44

[101] S. Mallat, *A wavelet tour of signal processing: The sparse way*, Academic Press, 2008. 45, 46, 47, 49, 51, 52, 53, 55

[102] A. J. Markvoort, R. Pino, and P. A. J. Hilbers, *Interpolating wavelets in Kohn-Sham electronic structure calculations*, in Computational Science — ICCS 2001: International Conference San Francisco, CA, USA, May 28–30, 2001 Proceedings, Part I, V. N. Alexandrov, J. J. Dongarra, B. A. Juliano, R. S. Renner, and C. J. K. Tan, eds., vol. 2073 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2001, pp. 541–550, https://doi.org/10.1007/3-540-45545-0_63. 41

[103] M. Mehra and K. Goyal, *Algorithm 929: A suite on wavelet differentiation algorithms*, ACM Trans. Math. Soft., 39 (2013), p. 27, https://doi.org/10.1145/2491491.2491497. 60, 61

[104] Y. Meyer, *Ondelettes et opérateurs*, Hermann, 1990. 44

[105] S. Mohr, L. E. Ratcliff, P. Boulanger, L. Genovese, D. Caliste, T. Deutsch, and S. Goedecker, *Daubechies wavelets for linear scaling density functional theory*, J. Chem. Phys., 140 (2014), p. 204110, https://doi.org/10.1063/1.4871876. 72

[106] P. Monasse and V. Perrier, *Orthonormal wavelet bases adapted for partial differential equations with boundary conditions*, SIAM J. Math. Anal., 29 (1998), pp. 1040–1065, https://doi.org/10.1137/S0036141095295127. 99

[107] P. Motamarri, M. Nowak, K. Leiter, J. Knap, and V. Gavini, *Higher-order adaptive finite-element methods for Kohn-Sham density functional theory*, J. Comput. Phys., 253 (2013), pp. 308–343, https://doi.org/10.1016/j.jcp.2013.06.042. 15

[108] S. Muraki, *Volume data and wavelet transforms*, IEEE Comput. Graph. Appl., 13 (1993), pp. 50–56, https://doi.org/10.1109/38.219451. 71

[109] S. Nagy and J. Pipek, *A wavelet-based adaptive method for determining eigenstates of electronic systems*, Theor. Chem. Acc., 125 (2010), pp. 471–479, https://doi.org/10.1007/s00214-009-0653-6. 15

[110] A. I. Neelov and S. Goedecker, *An efficient numerical quadrature for the calculation of the potential energy of wavefunctions expressed in the Daubechies wavelet basis*, J. Comput. Phys., 217 (2006), pp. 312–339, https://doi.org/10.1016/j.jcp.2006.01.003. 157, 212

[111] O. M. Nielsen, *Wavelets in Scientific Computing*, PhD thesis, Technical University of Denmark, Lyngby, 1998, https://hal.archives-ouvertes.fr/tel-00803835/. 67

[112] A. M. N. Niklasson, C. J. Tymczak, and H. Röder, *Multiresolution density-matrix approach to electronic structure calculations*, Phys. Rev. B, 66 (2002), p. 155120, https://doi.org/10.1103/PhysRevB.66.155120. 41

[113] R. G. Parr and W. Yang, *Density-Functional Theory of atoms and molecules*, vol. 16 of International Series of Monographs on Chemistry, Oxford University Press, New York, 1989. 13

[114] V. Perrier and C. Basdevant, *Periodical wavelet analysis, a tool for inhomogeneous field investigations: Theory and algorithms*, Rech. Aérosp., (1989), pp. 53–67. 68

[115] D. H. Pham, *Bases mixtes ondelettes–gaussiennes pour le calcul des structures électroniques*, tech. report, IFP Energies nouvelles, Rueil-Malmaison, September 2013. 99, 141

[116] C. Prud'homme, D. V. Rovas, K. Veroy, L. Machiels, Y. Maday, A. T. Patera, and G. Turinici, *Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods*, J. Fluids Engrg., 124 (2002), pp. 70–80, https://doi.org/10.1115/1.1448332. 19, 167, 170

[117] A. Quarteroni, A. Manzoni, and F. Negri, *Reduced Basis Methods for Partial Differential Equations: An Introduction*, UNITEXT, La Matematica per il 3+2, Springer International Publishing, Cham, 2015. 17, 169, 170

[118] J. M. Restrepo and G. Leaf, *Inner product computations using periodized Daubechies wavelets*, Int. J. Numer. Meth. Engng., 40 (1998), pp. 3557–3578, https://doi.org/10.1002/(SICI)1097-0207(19971015)40:19<3557::AID-NME227>3.0.CO;2-A. 69

[119] J. M. Restrepo, G. K. Leaf, and G. Schlossnagle, *Periodized Daubechies wavelets*, tech. report, Argonne National Laboratory, Illinois, USA, February 1996, https://www.osti.gov/scitech/biblio/211651. 68, 69

[120] M. Schechter, *Operator methods in quantum mechanics*, North-Holland, Amsterdam, 1981. 23

[121] T. C. Scott, J. F. Babb, A. Dalgarno, and J. D. Morgan, *The calculation of exchange forces: General results and specific models*, J. Chem. Phys., 99 (1993), pp. 2841–2854, https://doi.org/10.1063/1.465193. 92

[122] T. C. Scott, R. Mann, and R. E. Martinez II, *General relativity and quantum mechanics: Towards a generalization of the Lambert W function*, AAECC, 17 (2006), pp. 41–47, https://doi.org/10.1007/s00200-006-0196-1. 92

[123] I. Shavitt and M. Karplus, *Gaussian-transform method for molecular integrals. I. Formulation for energy integrals*, J. Chem. Phys., 43 (1964), pp. 398–414, https://doi.org/10.1063/1.1696757. 124

[124] I. M. Sigal, *How many electrons can a nucleus bind?*, Ann. Phys., 157 (1984), pp. 307–320, https://doi.org/10.1016/0003-4916(84)90062-9. 24

[125] J. C. Slater, *Atomic shielding constants*, Phys. Rev., 36 (1930), pp. 57–64, https://doi.org/10.1103/PhysRev.36.57. 33

[126] J. C. Slater, *Wave functions in a periodic potential*, Phys. Rev., 51 (1937), pp. 846–851, https://doi.org/10.1103/PhysRev.51.846. 16, 40

[127] W. Sweldens and R. Piessens, *Quadrature formulae and asymptotic error expansions for wavelet approximations of smooth functions*, SIAM J. Numer. Anal., 31 (1994), pp. 1240–1264, https://doi.org/10.1137/0731065. 52, 157

[128] A. Szabo and N. S. Ostlund, *Modern quantum chemistry: An introduction to advanced electronic structure theory*, MacMillan, 1982. 37

[129] Q. H. Tran, *Une analyse de l'erreur de quadrature du produit scalaire gaussienne-fonction d'échelle*. Note de travail, Janvier 2014. 159, 162

[130] M. Unser, P. Thévenaz, and A. Aldroubi, *Shift-orthogonal wavelet bases using splines*, IEEE Signal Processing Letters, 3 (1996), pp. 85–88, https://doi.org/10.1109/97.481163. 43

[131] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, *Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach*, Comput. Phys. Commun., 167 (2005), pp. 103–128, https://doi.org/10.1016/j.cpc.2004.12.014. 16

[132] R. Verfürth, *A review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Advances in numerical mathematics, Wiley-Teubner, New York, 1996. 172

[133] G. G. Walter and L. Cai, *Periodic wavelets from scratch*, J. Comput. Anal. Appli., 1 (1999), pp. 25–41, https://doi.org/10.1023/A:1022614519335. 69

[134] T. Yanai, G. I. Fann, Z. Gan, R. J. Harrison, and G. Beylkin, *Multiresolution quantum chemistry in multiwavelet bases: Analytic derivatives for Hartree-Fock and density functional theory*, J. Chem. Phys., 121 (2004), pp. 2866–2876, https://doi.org/10.1063/1.1768161. 41

[135] T. Yanai, G. I. Fann, Z. Gan, R. J. Harrison, and G. Beylkin, *Multiresolution quantum chemistry in multiwavelet bases: Hartree-Fock exchange*, J. Chem. Phys., 121 (2004), p. 6680, https://doi.org/10.1063/1.1790931. 41

[136] H. Yserentant, *On the regularity of the electronic Schrödinger equation in Hilbert spaces of mixed derivatives*, Numer. Math., 98 (2004), pp. 731–759, https://doi.org/10.1007/s00211-003-0498-1. 26

[137] D. Zhang, L. Shen, A. Zhou, and X.-G. Gong, *Finite element method for solving Kohn-Sham equations based on self-adaptive tetrahedral mesh*, Phys. Lett. A, 372 (2008), pp. 5071–5076, https://doi.org/10.1016/j.physleta.2008.05.075. 15

[138] G. M. Zhislin, *Discussion of the spectrum of Schrödinger operators for systems of many particles*, Trudy Mosk. Mat. Obs., 9 (1960), pp. 81–120. 24

[139] A. Zhou, *An analysis of finite-dimensional approximations for the ground state solution of Bose-Einstein condensates*, Nonlinearity, 17 (2004), pp. 541–550, https://doi.org/10.1088/0951-7715/17/2/010. 17

[140] A. Zhou, *Finite dimensional approximations for the electronic ground state solution of a molecular system*, Math. Methods Appl. Sci., 30 (2007), pp. 429–447, https://doi.org/10.1002/mma.793. 17