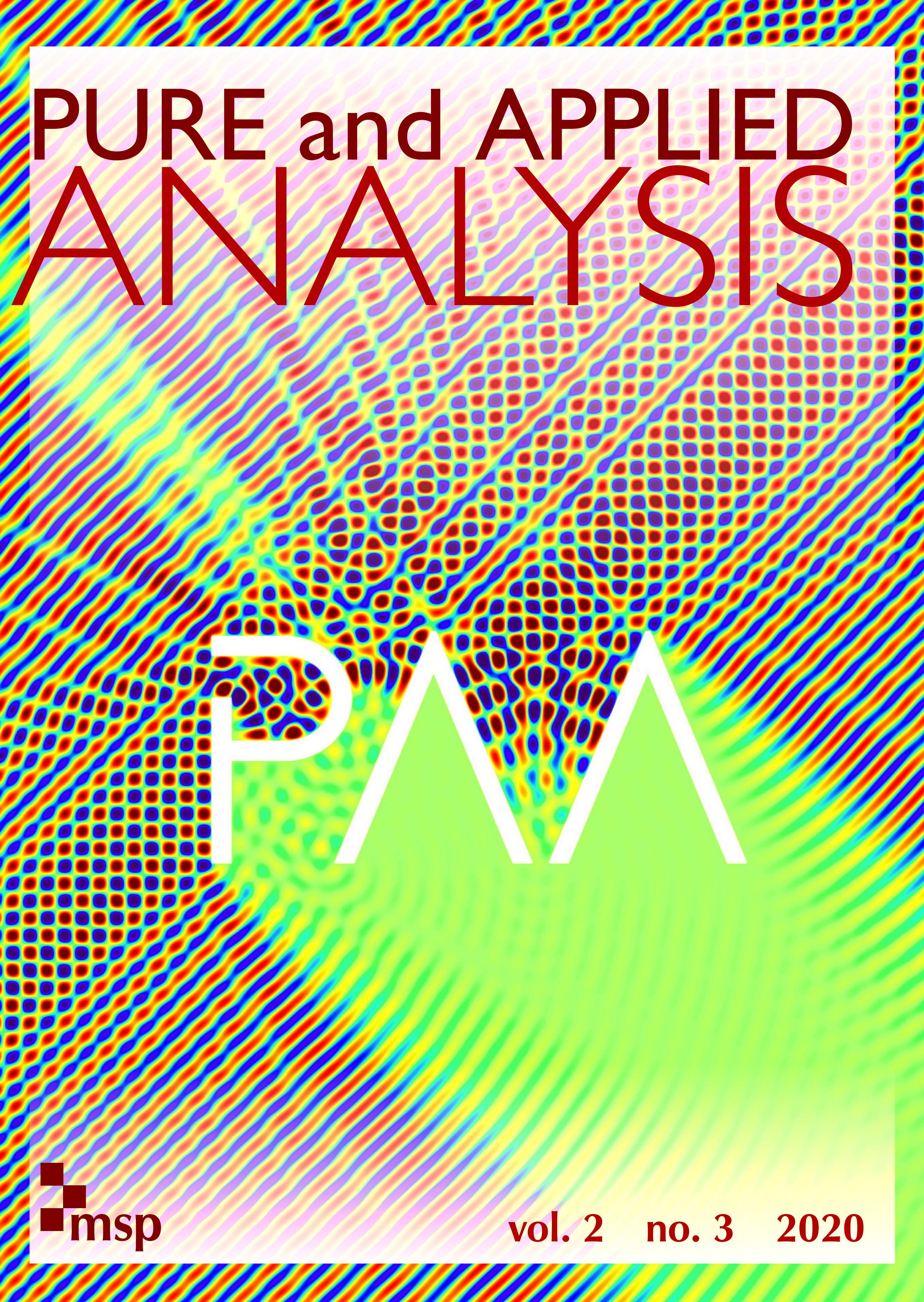
PURE and APPLIED ANALYSIS



PAA

msp

vol. 2 no. 3 2020

PURE and APPLIED
ANALYSIS

msp.org/paa

Cover image: The figure shows the outgoing scattered field produced by scattering a plane wave, coming from the northwest, off of the (stylized) letters P A A. The total field satisfies the homogeneous Dirichlet condition on the boundary of the letters. It is based on a numerical computation by Mike O'Neil of the Courant Institute.

See inside back cover or msp.org/paa for submission instructions.

The subscription price for 2020 is US $505/year for the electronic version, and $565/year (+$25, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Pure and Applied Analysis (ISSN 2578-5885 electronic, 2578-5893 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

PAA peer review and production are managed by EditFlow® from MSP.

PUBLISHED BY
mathematical sciences publishers
nonprofit scientific publishing

PURE and APPLIED
ANALYSIS

Vol. 2, No. 3, 2020
https://doi.org/10.2140/paa.2020.2.533

MEAN-FIELD MODEL FOR THE JUNCTION OF TWO QUASI-1-DIMENSIONAL QUANTUM COULOMB SYSTEMS

LING-LING CAO

Junctions appear naturally when one studies surface states or transport properties of quasi-1-dimensional materials such as carbon nanotubes, polymers and quantum wires. These materials can be seen as 1-dimensional systems embedded in the 3-dimensional space. We first establish a mean-field description of reduced Hartree–Fock-type for a 1-dimensional periodic system in the 3-dimensional space (a quasi-1-dimensional system), the unit cell of which is unbounded. With mild summability condition, we next show that a quasi-1-dimensional quantum system in its ground state can be described by a mean-field Hamiltonian. We also prove that the Fermi level of this system is always negative. A junction system is described by two different infinitely extended quasi-1-dimensional systems occupying separate half-spaces in three dimensions, where coulombic electron-electron interactions are taken into account and without any assumption on the commensurability of the periods. We prove the existence of the ground state for a junction system, the ground state is a spectral projector of a mean-field Hamiltonian, and the ground state density is unique.

1. Introduction

1A. ***Physical background and mathematical models.*** Atomic junctions of quasi-1-dimensional systems appear for instance when studying the surface states of 1-dimensional crystals [Aerts 1960; Shockley 1939], quantum thermal transport in nanostructures [Wang et al. 2008], and p-n junctions [Baugher et al. 2014; Pospischil et al. 2014], which are the foundation of the modern semiconductor electronic devices. Besides, electronic transport in carbon nanotubes [Laird et al. 2015] and in molecular wires [Nitzan and Ratner 2003], which recently attracted a lot of interest, is often modeled by the junction of two semi-infinite systems with different chemical potentials. In recent years, studies of various quantum Hall effects and topological insulators focused attention on 2-dimensional materials; see [Hasan and Kane 2010]. These 2-dimensional materials often possess periodicity in one dimension and can therefore be

MSC2010: 49S05, 58E99.
Keywords: mean field, junction, quasi-1-dimensional, quantum Coulomb system.

reduced to quasi-1-dimensional materials by momentum representation in the periodic direction [Hatsugai 1993]. Furthermore, when studying edge states properties (see [Hatsugai 1993; Avila et al. 2013]) of 2-dimensional materials, they can be seen as a junction with the vacuum.

Real world materials are often described by periodic [Catto et al. 2001; Cancès et al. 2008] or ergodically periodic [Cancès et al. 2013] systems in mathematical modeling. In this article we consider a junction of two different quasi-1-dimensional periodic systems without any assumption on the commensurability of the periods. Generally speaking, there are two regimes for the junction of two different periodic systems: when the chemical potentials of the underlying periodic systems are separated by some occupied bands (nonequilibrium regime, see Figure 3), and when the chemical potentials are in a common spectral gap (equilibrium regime, see Figure 4). The nonequilibrium regime models a persistent (nonperturbative) current in the junction system [Bruneau et al. 2015; 2016a; 2016b; Cornean et al. 2012], while the equilibrium regime can model either the ground state of the junction material or the presence of perturbative current in the linear response regime [Cornean et al. 2008]. In this article we consider the equilibrium regime, and only briefly comment on the nonequilibrium regime in Section 3B, as the study of this situation requires different techniques.

The most prominent feature of quasi-1-dimensional materials is the presence of strong electron-electron interactions due to low screening effect [Brus 2010; 2014] as electrons interact through the 3-dimensional space. For finite systems, one can use a N-body Schrödinger model to describe the electron-electron interactions. Nevertheless, this is impossible for infinite systems. Mean-field theory is a good candidate for infinite systems: it consists of replacing the N-body interactions by a 1-body interaction with an effective average field, leading to a quasiparticle description of the system. However, mean-field models are rarely available for quasi-1-dimensional periodic systems, as periodic systems are often considered either in the 3-dimensional space (see [Lieb and Simon 1977; Catto et al. 1998] for Thomas–Fermi-type models and [Catto et al. 2001] for Hartree–Fock-type models), or strictly in a 1-dimensional geometry (see [Blanc and Le Bris 2002] for Thomas–Fermi type models). To our knowledge, the work on Thomas–Fermi-type models [Blanc and Le Bris 2000] for polymers is the only literature available for a 1-dimensional periodic system with interactions through the 3-dimensional space. Furthermore, nonperiodic infinite systems are difficult to handle mathematically, as they do not possess any symmetry; hence the usual Bloch decomposition of periodic systems [Reed and Simon 1978; Catto et al. 2001] is not applicable, and the definition of the ground state energy needs to be examined [Blanc et al. 2003].

In this article, we establish a mean-field model to describe the junction of two different quasi-1-dimensional periodic systems (see Figure 2) in the 3-dimensional space with Coulomb interactions, under the framework of the reduced Hartree–Fock (rHF) description [Solovej 1991]. Note that the rHF model is strictly convex in the density, and can be seen as a good approximation of Kohn–Sham LDA model [Kohn and Sham 1965; Anantharaman and Cancès 2009; Lewin et al. 2020], which is widely used in condensed matter physics. This nonlinear model can be employed to describe the junction of two nanotubes, or a more realistic model of the junction of two quasi-1-dimensional crystals for electronic structure calculations. It can be further explored to study the linear response with respect to different Fermi levels between two semi-infinite chains: recall that the famous Landauer–Büttiker formalism [Landauer 1970; Büttiker

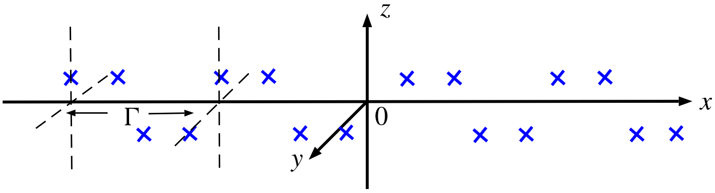


Figure 1. An example of nuclei configuration of a quasi-1-dimensional periodic system.

et al. 1985] for electronic (thermal) transport, which is based on the lead-device-lead description, can be seen as the junction of two different quasi-1-dimensional systems (leads) with different chemical (thermal) potentials, and the device as a perturbation of this junction. Note also that p-n junctions of carbon nanotubes without external battery [Léonard and Tersoff 1999; Lee et al. 2004] correspond to the equilibrium regime, and can thus be described by the model we consider. Furthermore, our model can also be easily adapted to describe 1-dimensional dislocation problems in the 3-dimensional space, while the linear 1-dimensional dislocation problems have been studied in [Korotyaev 2000; 2005] and some generalizations have been provided for higher-dimensional systems [Dohnal et al. 2011; Hempel and Kohlmann 2011; Hempel et al. 2015].

1B. ***Summary of main results.*** The organization of this article and the main results are as follows: in Section 2 we consider a quasi-1-dimensional periodic system, which is described by nuclei arranged periodically alongside the x-axis (see Figure 1) with electrons occupying the 3-dimensional space, as it is a building block for the junction system. We define a periodic rHF energy functional (2-15) by taking into account the real Coulomb interactions in the 3-dimensional space. In Theorem 2.6 we show that this rHF functional admits minimizers, and that the ground state electronic density is unique. Note that this is different from [Catto et al. 2001; Cancès et al. 2008] as the system is periodic only in the x-direction, the unit cell Γ being unbounded so that additional compactness proofs are needed when dealing with the ground state problem. With a mild summability condition (2-18) on the unique density of minimizers, we are able to obtain a mean-field Hamiltonian $H_{\text{per}} = -\frac{1}{2}\Delta + V_{\text{per}}$ to describe a quasi-1-dimensional periodic system, where V_{per} is the mean-field potential that tends to 0 in the $\boldsymbol{r} := (y, z)$-direction. In Theorem 2.7 we prove that the Fermi level ϵ_F of the quasi-1-dimensional system, which represents the highest energy occupied by electrons under this quasiparticle description, is always negative. We also prove that the unique minimizer is a spectral projector of the mean-field Hamiltonian $\gamma_{\text{per}} = \mathbb{1}_{(-\infty,\epsilon_F]}(H_{\text{per}})$.

In Section 3, under certain symmetry assumptions on the nuclear densities $\mu_{\text{per},L}$ and $\mu_{\text{per},R}$ of two different quasi-1-dimensional periodic systems, the junction system is described by considering the following nuclear configuration (see Figure 2):

$$\mu_J := \mathbb{1}_{x\leq 0} \cdot \mu_{\text{per},L} + \mathbb{1}_{x>0} \cdot \mu_{\text{per},R} + \nu,$$

where ν describes how the junction is initiated. We aim at establishing a quasiparticle description of this infinitely extended junction system with Coulomb interactions and show the existence of ground state. As we do not assume any commensurability of periods of the two quasi-1-dimensional systems, the junction

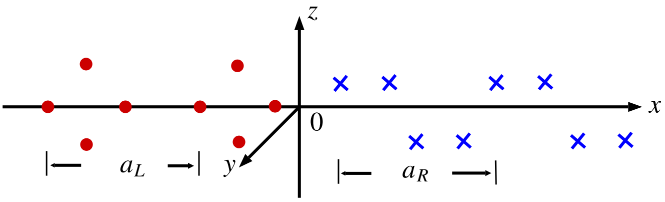


Figure 2. Nuclei configuration of the junction system with period a_L on $(-\infty, 0] \times \mathbb{R}^2$ and a_R on $(0, +\infty) \times \mathbb{R}^2$.

system does not possess any translation-invariant symmetry. The main idea is to establish a well-suited reference system based on the linear combination of periodic systems, and use perturbative techniques which have been widely used for mean-field type models [Hainzl et al. 2005a; 2007; 2009; Cancès et al. 2008; Frank et al. 2013] to justify the construction. More precisely, we define a reference Hamiltonian

$$H_\chi = \chi H_{\text{per},L} \chi + \sqrt{1-\chi^2} H_{\text{per},R} \sqrt{1-\chi^2}$$

with χ a smooth cut-off function approximating $\mathbb{1}_{x \le 0}$, where $H_{\text{per},L}$ and $H_{\text{per},R}$ are the mean-field Hamiltonians of the quasi-1-dimensional periodic systems. Denote by $\sigma(A) = \sigma_{\text{disc}}(A) \cup \sigma_{\text{ess}}(A)$ the spectrum of A, where $\sigma_{\text{disc}}(A)$ and $\sigma_{\text{ess}}(A)$ denote the discrete and essential spectra of an operator A respectively. Denote also by $\sigma_{\text{ac}}(A)$ the purely absolutely continuous spectrum of A. We first show in Proposition 3.1 that

$$\sigma_{\text{ess}}(H_\chi) = \sigma_{\text{ess}}(H_{\text{per},L}) \cup \sigma_{\text{ess}}(H_{\text{per},R}), \qquad \sigma_{\text{ess}}(H_\chi) \cap (-\infty, 0] \subseteq \sigma_{\text{ac}}(H_\chi).$$

This implies that the essential spectrum of the reference Hamiltonian is independent of the cut-off function χ, and the linear junction preserves the scattering channels of the underlying systems, since the purely absolutely continuous spectrum of the Hamiltonian has not been modified; hence the linear junction can be used to study the electronic conductance with the Landauer–Büttiker formalism (see for example [Bruneau et al. 2015; 2016a; 2016b] as well as the discussion following Proposition 3.1).

After introducing a reference state $\gamma_\chi := \mathbb{1}_{(-\infty,\epsilon_F)}(H_\chi)$, we show in Proposition 3.2 that the electronic density ρ_χ is close to the linear combination of the underlying periodic electronic densities, and that the difference with these reference densities decays exponentially fast. The quasiparticle description of the nonlinear junction state can be constructed by considering

$$\gamma_J = \gamma_\chi + Q_\chi,$$

where Q_χ is a trial density matrix which encodes the nonlinear effects of the junction system. Following the idea developed in [Cancès et al. 2008], we associate Q_χ with some minimization problem in Proposition 3.5, and denote by $\overline{Q}_\chi$ a minimizer. We prove in Theorem 3.6 that

$$\rho_{\gamma_J} = \rho_\chi + \rho_{\overline{Q}_\chi} \quad \text{is independent of } \chi.$$

This implies that the ground state of the junction system with Coulomb interactions exists and its density is independent of the choice of the reference state; see Corollary 3.6.1.

2. A reduced Hartree–Fock description of quasi-1-dimensional periodic systems

In this section we give a mathematical description of a quasi-1-dimensional periodic system in the framework of the reduced Hartree–Fock (rHF) approach. In Section 2A, we introduce some mathematical preliminaries. In Section 2B we construct a periodic rHF energy functional for a quasi-1-dimensional system.

Let us first introduce some notation. Unless otherwise specified, the functions on $\mathbb{R}^d$ considered in this article are complex-valued. Elements of $\mathbb{R}^3$ are denoted by $\boldsymbol{x} = (x, \boldsymbol{r})$, where $x \in \mathbb{R}$. For a given separable Hilbert space $\mathfrak{H}$, we denote by $\mathcal{L}(\mathfrak{H})$ the space of bounded linear operators acting on $\mathfrak{H}$, by $\mathcal{S}(\mathfrak{H})$ the space of bounded self-adjoint operators acting on $\mathfrak{H}$, and by $\mathfrak{S}_p(\mathfrak{H})$ the Schatten class of operators acting on $\mathfrak{H}$. For $1 \le p < \infty$, a compact operator A belongs to $\mathfrak{S}_p(\mathfrak{H})$ if and only if $\|A\|_{\mathfrak{S}_p} := (\mathrm{Tr}(|A|^p))^{1/p} < \infty$. Operators in $\mathfrak{S}_1(\mathfrak{H})$ and $\mathfrak{S}_2(\mathfrak{H})$ are respectively called trace-class and Hilbert–Schmidt. If $A \in \mathfrak{S}_1(L^2(\mathbb{R}^d))$, there exists a unique function $\rho_A \in L^1(\mathbb{R}^d)$ such that,

$$\text{for all } \phi \in L^\infty(\mathbb{R}^d), \quad \mathrm{Tr}(A\phi) = \int_{\mathbb{R}^d} \rho_A \phi.$$

The function ρ_A is called the density of the operator A. If the integral kernel $A(\boldsymbol{r}, \boldsymbol{r}')$ of A is continuous on $\mathbb{R}^d \times \mathbb{R}^d$, then $\rho_A(\boldsymbol{r}) = A(\boldsymbol{r}, \boldsymbol{r})$ for all $\boldsymbol{r} \in \mathbb{R}^d$. This relation still stands in some weaker sense for a generic trace-class operator.

An operator $A \in \mathcal{L}(L^2(\mathbb{R}^d))$ is called locally trace-class if the operator $\varrho A \varrho$ is trace-class for any $\varrho \in C_c^\infty(\mathbb{R}^d)$. The density of a locally trace-class operator $A \in \mathcal{L}(L^2(\mathbb{R}^d))$ is the unique function $\rho_A \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$ such that,

$$\text{for all } \phi \in C_c^\infty(\mathbb{R}^d), \quad \mathrm{Tr}(A\phi) = \int_{\mathbb{R}^d} \rho_A \phi.$$

Let $\mathscr{S}(\mathbb{R}^d)$ be the Schwartz space of rapidly decreasing functions on $\mathbb{R}^d$, and $\mathscr{S}'(\mathbb{R}^d)$ the space of tempered distributions on $\mathbb{R}^d$. We denote by $\hat{\phi}$ and $\check{\phi}$ the Fourier and inverse Fourier transforms on $\mathscr{S}'(\mathbb{R}^d)$, with the normalization,

$$\text{for all } \phi \in L^1(\mathbb{R}^d), \quad \hat{\phi}(\zeta) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \phi(x) \mathrm{e}^{-\mathrm{i}\zeta x}\, dx, \quad \check{\phi}(x) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \phi(\zeta) \mathrm{e}^{\mathrm{i}\zeta x}\, d\zeta.$$

The normalization ensures that the Fourier transform defines a unitary operator on $L^2(\mathbb{R}^d)$.

2A. ***Mathematical preliminaries.*** We first introduce a decomposition of the operator which is $\mathbb{Z}$-translation-invariant in the x-direction based on the partial Bloch transform. In order to describe the 1-dimensional periodic system in the 3-dimensional space, we next introduce a mixed Fourier transform. We also introduce a Green's function which is periodic only in the x-direction. Finally we introduce the kinetic energy space of density matrices and Coulomb interactions for quasi-1-dimensional systems.

Bloch transform in the x-direction. For $k \in \mathbb{Z}$, we denote by τ_k^x the translation operator in the x-direction acting on $L^2_{\mathrm{loc}}(\mathbb{R}^3)$:

$$\text{for all } u \in L^2_{\mathrm{loc}}(\mathbb{R}^3), \quad (\tau_k^x u)(\,\cdot\,, \boldsymbol{r}) = u(\,\cdot - k, \boldsymbol{r}) \quad \text{for a.a. } \boldsymbol{r} \in \mathbb{R}^2.$$

An operator A on $L^2(\mathbb{R}^3)$ is called $\mathbb{Z}$-translation-invariant in the x-direction if it commutes with τ_k^x for all $k \in \mathbb{Z}$. In order to decompose operators which are $\mathbb{Z}$-translation-invariant in the x-direction, let us without loss of generality choose a unit cell

$$\Gamma := \left[-\tfrac{1}{2}, \tfrac{1}{2}\right) \times \mathbb{R}^2,$$

and introduce the L^p spaces and H^1 spaces of functions which are 1-periodic in the x-direction: for $1 \le p \le +\infty$,

$$\begin{aligned} L^p_{\mathrm{per},x}(\Gamma) &:= \{u \in L^p_{\mathrm{loc}}(\mathbb{R}^3) \mid \|u\|_{L^p(\Gamma)} < +\infty,\ \tau_k^x u = u \text{ for all } k \in \mathbb{Z}\}, \\ H^1_{\mathrm{per},x}(\Gamma) &:= \{u \in L^2_{\mathrm{per},x}(\Gamma) \mid \nabla u \in (L^2_{\mathrm{per},x}(\Gamma))^3\}. \end{aligned}$$

Let us also introduce the constant fiber direct integral of Hilbert spaces [Reed and Simon 1978]

$$L^2(\Gamma^*; L^2_{\mathrm{per},x}(\Gamma)) := \int_{\Gamma^*}^{\oplus} L^2_{\mathrm{per},x}(\Gamma)\, \frac{d\xi}{2\pi},$$

with the base $\Gamma^* := [-\pi, \pi) \times \{0\}^2 \equiv [-\pi, \pi)$. The partial Bloch transform $\mathscr{B}$ is a unitary operator from $L^2(\mathbb{R}^3)$ to $L^2(\Gamma^*; L^2_{\mathrm{per},x}(\Gamma))$, defined on the dense subspace of $C_c^\infty(\mathbb{R}^3)$ of $L^2(\mathbb{R}^3)$:

$$\text{for all } (x, \boldsymbol{r}) \in \Gamma, \text{ for all } \xi \in \Gamma^*, \quad (\mathscr{B}\phi)_\xi(x, \boldsymbol{r}) := \sum_{k \in \mathbb{Z}} \mathrm{e}^{-\mathrm{i}(x+k)\xi} \phi(x+k, \boldsymbol{r}).$$

Its inverse is given, for $f_\bullet = (f_\xi)_{\xi \in \Gamma^*}$, by

$$\text{for all } k \in \mathbb{Z}, \ \text{ for a.a. } (x, \boldsymbol{r}) \in \Gamma, \quad (\mathscr{B}^{-1} f_\bullet)(x+k, \boldsymbol{r}) := \int_{\Gamma^*} \mathrm{e}^{\mathrm{i}(k+x)\xi} f_\xi(x, \boldsymbol{r})\, \frac{d\xi}{2\pi}.$$

The partial Bloch transform has the property that any operator A on $L^2(\mathbb{R}^3)$ which commutes with τ_k^x for $k \in \mathbb{Z}$ is decomposed by $\mathscr{B}$: for any $A \in \mathcal{L}(L^2(\mathbb{R}^3))$ such that $\tau_k^x A = A \tau_k^x$, there exists $A_\bullet \in L^\infty(\Gamma^*; \mathcal{L}(L^2_{\mathrm{per},x}(\Gamma)))$ such that, for all $u \in L^2(\mathbb{R}^3)$,

$$(\mathscr{B}(Au))_\xi = A_\xi (\mathscr{B}u)_\xi \quad \text{for a.a. } \xi \in \Gamma^*.$$

We hence use the following notation for the decomposition of an operator A which is $\mathbb{Z}$-translation-invariant in the x-direction:

$$A = \mathscr{B}^{-1}\left(\int_{\Gamma^*}^{\oplus} A_\xi \frac{d\xi}{2\pi}\right)\mathscr{B}.$$

In addition, $\|A\|_{\mathcal{L}(L^2(\mathbb{R}^3))} = \|\|A_\bullet\|_{\mathcal{L}(L^2_{\mathrm{per},x}(\Gamma))}\|_{L^\infty(\Gamma^*)}$. In particular, if A is positive and locally trace-class, then, for almost all $\xi \in \Gamma^*$, A_ξ is locally trace-class. The densities of these operators are related by the formula

$$\rho_A(\boldsymbol{x}) = \frac{1}{2\pi} \int_{\Gamma^*} \rho_{A_\xi}(\boldsymbol{x})\, d\xi. \tag{2-1}$$

If A is a (not necessarily bounded) self-adjoint operator such that $\tau_k^x (A+\mathrm{i})^{-1} = (A+\mathrm{i})^{-1}\tau_k^x$ for all $k \in \mathbb{Z}$, then A is decomposed by $\mathcal{U}$; see [Reed and Simon 1978, Theorems XIII.84 and XIII.85]. In particular, denoting by Δ the Laplace operator acting on $L^2(\mathbb{R}^3)$, the kinetic energy operator $-\frac{1}{2}\Delta$ on $L^2(\mathbb{R}^3)$ is decomposed by $\mathscr{B}$ as follows:

$$-\frac{1}{2}\Delta = \mathscr{B}^{-1}\left(\int_{\Gamma^*} -\frac{1}{2}\Delta_\xi \,\frac{d\xi}{2\pi}\right)\mathscr{B}, \quad -\Delta_\xi = (-\mathrm{i}\nabla_\xi)^2 = (\mathrm{i}\partial_x - \xi)^2 - \Delta_{\boldsymbol{r}}, \tag{2-2}$$

where $\Delta_{\boldsymbol{r}}$ is the Laplace operator acting on $L^2(\mathbb{R}^2)$.

Mixed Fourier transform. The mixed Fourier transform consists of a Fourier series transform in the x-direction and an integral Fourier transform in the $\boldsymbol{r}$-direction. Denote by $\mathscr{S}_{\mathrm{per},x}(\Gamma)$ the space of functions that are C^∞ on $\mathbb{R}^3$ and Γ-periodic, decaying faster than any power of $|\boldsymbol{r}|$ when $|\boldsymbol{r}|$ tends to infinity, as well as their derivatives. Denote by $\mathscr{S}'_{\mathrm{per},x}(\Gamma)$ the dual space of $\mathscr{S}_{\mathrm{per},x}(\Gamma)$. The mixed Fourier transform is the unitary transform $\mathscr{F} : L^2_{\mathrm{per},x}(\Gamma) \to \ell^2(\mathbb{Z}, L^2(\mathbb{R}^2))$ defined on the dense subspace $\mathscr{S}_{\mathrm{per},x}(\Gamma)$ of $L^2_{\mathrm{per},x}(\Gamma)$ by

$$\text{for all } \phi \in \mathscr{S}_{\mathrm{per},x}(\Gamma), \text{ for all } (n, \boldsymbol{k}) \in \mathbb{Z} \times \mathbb{R}^2, \quad \mathscr{F}\phi(n, \boldsymbol{k}) := \frac{1}{2\pi}\int_\Gamma \phi(x, \boldsymbol{r})\, \mathrm{e}^{-\mathrm{i}(2\pi n x + \boldsymbol{k}\cdot\boldsymbol{r})}\, dx\, d\boldsymbol{r}. \tag{2-3}$$

Its inverse is given by,

$$\text{for all } (\psi_n(\boldsymbol{k}))_{n\in\mathbb{Z},\boldsymbol{k}\in\mathbb{R}^2} \in \ell^2(\mathbb{Z}; L^2(\mathbb{R}^2)), \quad \mathscr{F}^{-1}\psi(x, \boldsymbol{r}) := \frac{1}{2\pi}\sum_{n\in\mathbb{Z}}\int_{\mathbb{R}^2} \psi_n(\boldsymbol{k})\, \mathrm{e}^{\mathrm{i}(2\pi n x + \boldsymbol{k}\cdot\boldsymbol{r})}\, d\boldsymbol{k}.$$

Note that $\mathscr{F}$ can be extended from $\mathscr{S}'_{\mathrm{per},x}(\Gamma)$ to $\mathscr{S}'(\mathbb{R}^3)$. One can easily see that $\mathscr{F}$ is an isometry from $L^2_{\mathrm{per},x}(\Gamma)$ to $\ell^2(\mathbb{Z}, L^2(\mathbb{R}^2))$ in the following sense:

$$\text{for all } f, g \in L^2_{\mathrm{per},x}(\Gamma), \quad \int_\Gamma \overline{f(x, \boldsymbol{r})}\, g(x, \boldsymbol{r})\, dx\, d\boldsymbol{r} = \sum_{n\in\mathbb{Z}}\int_{\mathbb{R}^2} \overline{\mathscr{F}f(n, \boldsymbol{k})}\, \mathscr{F}g(n, \boldsymbol{k})\, d\boldsymbol{k}. \tag{2-4}$$

Moreover, it is easy to verify that, for $f, g \in L^2_{\mathrm{per},x}(\Gamma)$,

$$\mathscr{F}(f \star_\Gamma g) = 2\pi(\mathscr{F}f)(\mathscr{F}g), \tag{2-5}$$

where $(f \star_\Gamma g)(\boldsymbol{x}) := \int_\Gamma f(\boldsymbol{x} - \boldsymbol{x}')g(\boldsymbol{x}')\, d\boldsymbol{x}'$. As an application of the mixed Fourier transform, let us introduce a Kato–Seiler–Simon-type inequality [Seiler and Simon 1975] for the operator $-\mathrm{i}\nabla_\xi = (-\mathrm{i}\partial_x + \xi, -\mathrm{i}\partial_{\boldsymbol{r}})$ for all $\xi \in \Gamma^*$, which will be repeatedly used in the proofs.

Lemma 2.1. *Fix $\xi \in \Gamma^*$. Let $2 \le p \le +\infty$ and $f, g \in L^p_{\mathrm{per},x}(\Gamma)$. Then*

$$\|f(-\mathrm{i}\nabla_\xi)g\|_{\mathfrak{S}_p(L^2_{\mathrm{per},x}(\Gamma))} \le (2\pi)^{-2/p}\|g\|_{L^p_{\mathrm{per},x}(\Gamma)}\left(\sum_{n\in\mathbb{Z}}\|f(2\pi n + \xi, \cdot)\|^p_{L^p(\mathbb{R}^2)}\right)^{1/p} \tag{2-6}$$

for any $2 \le p < \infty$ and

$$\|f(-\mathrm{i}\nabla_\xi)g\| \le \|g\|_{L^\infty_{\mathrm{per},x}(\Gamma)} \sup_{n\in\mathbb{Z}}\|f(2\pi n + \xi, \cdot)\|_{L^\infty(\mathbb{R}^2)},$$

when $p = +\infty$.

The proof is an easy adaptation of the proof of the classical Kato–Seiler–Simon inequality (4-1) by replacing the Fourier transform with the mixed Fourier transform $\mathscr{F}$. The detailed proof of this lemma can be read in [Cao 2019b, Lemma 2.1].

Periodic Green's function. We introduce a 3-dimensional Green's function which is 1-periodic in the x-direction in the same spirit as in [Blanc and Le Bris 2000; Lieb and Simon 1977].

Definition 2.2 (periodic Green's function). For $(x, \boldsymbol{r}) \in \mathbb{R}^3$, the periodic Green's function is defined as

$$G(x,\boldsymbol{r}) = -2\log(|\boldsymbol{r}|) + \widetilde{G}(x,\boldsymbol{r}), \quad \widetilde{G}(x,\boldsymbol{r}) := 4\sum_{n\ge 1} K_0(2\pi n|\boldsymbol{r}|)\cos(2\pi n x), \tag{2-7}$$

where $K_0(\alpha) := \int_0^{+\infty} \mathrm{e}^{-\alpha\cosh(t)}\,dt$ is the modified Bessel function of the second kind.

Note that $-2\log(|\boldsymbol{r}|)$ is the solution of the equation

$$-\Delta u = 4\pi\delta_{\boldsymbol{r}=0} \quad \text{in } \mathscr{S}'(\mathbb{R}^2),$$

where $\delta_a \in \mathscr{S}'(\mathbb{R}^d)$ is the Dirac distribution at $a \in \mathbb{R}^d$. It is also known as the 2-dimensional Coulomb kernel of the whole space Poisson equation. It appears here since we consider the Green's function which is periodic only in the x-direction and has a free boundary condition in the $\boldsymbol{r}$-direction. The following lemma summarizes the properties of the periodic Green's function defined in (2-7).

Lemma 2.3. (1) *The Green's function $G(x,\boldsymbol{r})$ defined in (2-7) satisfies the following Poisson equation*:

$$-\Delta G(x,\boldsymbol{r}) = 4\pi\sum_{n\in\mathbb{Z}} \delta_{(x,\boldsymbol{r})=(n,0)} \in \mathscr{S}'(\mathbb{R}^3).$$

Moreover $G \in \mathscr{S}'_{\mathrm{per},x}(\Gamma)$ and

$$\mathscr{F}(G)(n,\boldsymbol{k}) = \frac{2}{4\pi^2 n^2 + |\boldsymbol{k}|^2} \in \mathscr{S}'(\mathbb{R}^3). \tag{2-8}$$

(2) *The function $\widetilde{G}$ defined in (2-7) belongs to $L^p_{\mathrm{per},x}(\Gamma)$ for $1 \le p < 2$ and satisfies $\int_\Gamma \widetilde{G} \equiv 0$. Moreover, there exist positive constants d_1 and d_2 such that*

$$|\widetilde{G}(\,\cdot\,,\boldsymbol{r})| \le d_1\frac{\mathrm{e}^{-2\pi|\boldsymbol{r}|}}{\sqrt{|\boldsymbol{r}|}}$$

when $|\boldsymbol{r}| \to +\infty$, and $|\widetilde{G}(\,\cdot\,,\boldsymbol{r})| \le d_2/|\boldsymbol{r}|$ when $|\boldsymbol{r}| \to 0$, uniformly with respect to x. Finally, the function $\widetilde{G}(x,\boldsymbol{r})$ can also be written as

$$\widetilde{G}(x,\boldsymbol{r}) = \sum_{n\in\mathbb{Z}}\left(\frac{1}{\sqrt{(x-n)^2+|\boldsymbol{r}|^2}} - \int_{-1/2}^{1/2}\frac{1}{\sqrt{(x-y-n)^2+|\boldsymbol{r}|^2}}\,dy\right). \tag{2-9}$$

The proof of this lemma is an easy and direct computation and can be read in Section A1 of the Appendix.

One-body density matrices and kinetic energy space. In mean-field models, electronic states can be described by one-body density matrices; see, e.g., [Cancès et al. 2008; Frank et al. 2013]. Recall that for a finite system with N electrons, a density matrix is a trace-class self-adjoint operator $\gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) \cap \mathfrak{S}_1(L^2(\mathbb{R}^3))$ satisfying the Pauli principle $0 \le \gamma \le 1$ and the normalization condition $\mathrm{Tr}(\gamma) = \int_{\mathbb{R}^3}\rho_\gamma = N$. The kinetic energy of γ is given by $\mathrm{Tr}(-\frac{1}{2}\Delta\gamma) := \frac{1}{2}\mathrm{Tr}(|\nabla|\gamma|\nabla|)$; see [Catto et al. 2001; Cancès et al. 2008; Cancès and Stoltz 2012].

Consider a 1-dimensional periodic system in the 3-dimensional space, where atoms are arranged periodically in the x-direction with unit cell Γ and first Brillouin zone Γ^*. Since the rHF model is strictly convex in the density [Solovej 1991], we do not expect any spontaneous symmetry breaking. Therefore the electronic state of this quasi-1-dimensional system will be described by a one-body density matrix which commutes with the translations $\{\tau_k^x\}_{k\in\mathbb{Z}}$, and hence is decomposed by the partial Bloch transform $\mathscr{B}$. In view of the decomposition (2-2), we define the following admissible set of one-body density matrices, which guarantees that the number of electrons per unit cell and the kinetic energy per unit cell are finite:

$$\mathcal{P}_{\mathrm{per},x} := \left\{ \gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) \;\middle|\; 0 \le \gamma \le 1 \text{ for all } k \in \mathbb{Z}, \\ \tau_k^x \gamma = \gamma \tau_k^x,\ \int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}}\big(\sqrt{1-\Delta_\xi}\,\gamma_\xi \sqrt{1-\Delta_\xi}\big)\, d\xi < \infty \right\}, \tag{2-10}$$

where

$$\gamma = \mathscr{B}^{-1}\left(\int_{\Gamma^*} \gamma_\xi \, \frac{d\xi}{2\pi} \right) \mathscr{B}. \tag{2-11}$$

For any $\gamma \in \mathcal{P}_{\mathrm{per},x}$, it is easy to see that $\rho_\gamma \in L^1_{\mathrm{per},x}(\Gamma)$. Moreover, a Hoffmann-Ostenhof-type inequality [Hoffmann-Ostenhof and Hoffmann-Ostenhof 1977] can also be deduced from [Catto et al. 2001, equation (4.42)]:

$$\int_\Gamma |\nabla \sqrt{\rho_\gamma}|^2 \le \int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}}(-\Delta_\xi \gamma_\xi)\, \frac{d\xi}{2\pi}. \tag{2-12}$$

Therefore $\sqrt{\rho_\gamma}$ is in $H^1_{\mathrm{per},x}(\Gamma)$, and hence is in $L^6_{\mathrm{per},x}(\Gamma)$ by Sobolev embeddings, so that $\rho_\gamma \in L^p_{\mathrm{per},x}(\Gamma)$ for $1 \le p \le 3$ by an interpolation argument.

Coulomb interactions. Recall that the Coulomb interaction energy of charge densities f and g belonging to $L^{6/5}(\mathbb{R}^3)$ can be written in real and reciprocal space as

$$D(f,g) := \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{f(\boldsymbol{x})g(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d\boldsymbol{x}\, d\boldsymbol{x}' = 4\pi \int_{\mathbb{R}^3} \frac{\overline{\hat{f}(\boldsymbol{k})}\hat{g}(\boldsymbol{k})}{|\boldsymbol{k}|^2}.$$

In order to describe Coulomb interactions in the reciprocal space for a quasi-1-dimensional periodic system, we gather the results obtained in (2-4), (2-5) and (2-8), and define the Coulomb interaction energy per unit cell for charge densities f, g belonging to $\mathscr{S}_{\mathrm{per},x}(\Gamma)$ as

$$D_\Gamma(f,g) := 4\pi \sum_{n\in\mathbb{Z}} \int_{\mathbb{R}^2} \frac{\overline{\mathscr{F}(f)}(n,\boldsymbol{k})\mathscr{F}(g)(n,\boldsymbol{k})}{|\boldsymbol{k}|^2 + 4\pi^2 n^2}\, d\boldsymbol{k}. \tag{2-13}$$

It is easy to see that $D_\Gamma(\cdot\,,\cdot)$ is a positive definite bilinear form on $\mathscr{S}_{\mathrm{per},x}(\Gamma)$. Let us introduce the Coulomb space for the 1-dimensional periodic system in the 3-dimensional space as

$$\mathcal{C}_\Gamma := \{ f \in \mathscr{S}'_{\mathrm{per},x}(\Gamma) \mid \text{for all } n \in \mathbb{Z},\ \mathscr{F}(f)(n,\cdot) \in L^1_{\mathrm{loc}}(\mathbb{R}^2),\ D_\Gamma(f,f) < +\infty \}, \tag{2-14}$$

which is a Hilbert space endowed with the inner product $D_\Gamma(\cdot\,,\cdot)$.

Remark 2.4. Charge densities in $\mathcal{C}_\Gamma$ are neutral in some weak sense. Indeed, for $f \in \mathcal{C}_\Gamma \bigcap L^1_{\mathrm{per},x}(\Gamma)$, the condition

$$\int_{\mathbb{R}^2} \frac{|\mathscr{F}(f)(0,\boldsymbol{k})|^2}{|\boldsymbol{k}|^2}\, d\boldsymbol{k} < +\infty$$

implies that $\mathscr{F}(f)(0,\boldsymbol{0}) = \int_\Gamma f(x,\boldsymbol{r})\, dx\, d\boldsymbol{r} = 0$.

2B. ***Reduced Hartree–Fock description for a quasi-1-dimensional periodic system.*** Based on the kinetic energy space and Coulomb interactions defined in the previous section, we construct here an rHF energy functional for a quasi-1-dimensional periodic system which is 1-periodic only in the x-direction. We show that its ground state is given by the solution of some minimization problem. Denote by $Z \in \mathbb{N}^*$ the total nuclear charge in each unit cell. For the sake of technical reasons we model the nuclear density of a quasi-1-dimensional system by a smooth function (smeared nuclei) which is 1-periodic in the x-direction

$$\mu_{\text{per}}(x, \boldsymbol{r}) = \sum_{n \in \mathbb{Z}} Z\, m(x - n, \boldsymbol{r}),$$

where $m(x, \boldsymbol{r})$ is a nonnegative $C_c^\infty(\Gamma)$ function such that $\int_{\mathbb{R}^3} m = 1$. In particular $\int_\Gamma \mu_{\text{per}} = Z$.

For any trial density matrix γ which commutes with the translations τ_k^x in the x-direction, the periodic rHF energy functional for a quasi-1-dimensional system associated with the nuclear density μ_{per} is defined as

$$\mathcal{E}_{\text{per},x}(\gamma) := \frac{1}{2\pi} \int_{\Gamma^*} \operatorname{Tr}_{L^2_{\text{per},x}(\Gamma)} \Big(-\frac{1}{2}\Delta_\xi \gamma_\xi\Big) d\xi + \frac{1}{2} D_\Gamma(\rho_\gamma - \mu_{\text{per}}, \rho_\gamma - \mu_{\text{per}}). \tag{2-15}$$

Let us introduce the following set of admissible density matrices for this rHF energy functional, which guarantees that the kinetic energy and Coulomb interaction energy per unit cell are finite:

$$\mathcal{F}_\Gamma := \{\gamma \in \mathcal{P}_{\text{per},x} \mid \rho_\gamma - \mu_{\text{per}} \in \mathcal{C}_\Gamma\},$$

where $\mathcal{P}_{\text{per},x}$ is the kinetic energy space defined in (2-10) and $\mathcal{C}_\Gamma$ is the Coulomb space defined in (2-14).

Lemma 2.5. *The set $\mathcal{F}_\Gamma$ is not empty. Moreover, for any $\gamma \in \mathcal{F}_\Gamma$,*

$$\int_\Gamma \rho_\gamma = \int_\Gamma \mu_{\text{per}}. \tag{2-16}$$

The proof of Lemma 2.5 relies on an explicit construct of an element in $\mathcal{F}_\Gamma$ and can be read in Section 4A.

The periodic rHF ground state energy (per unit cell) of a quasi-1-dimensional system can then be written as the minimization problem

$$I_{\text{per}} = \inf\{\mathcal{E}_{\text{per},x}(\gamma) \mid \gamma \in \mathcal{F}_\Gamma\}. \tag{2-17}$$

The minimization problem similar to (2-17) under the Thomas–Fermi-type models has been studied in [Blanc and Le Bris 2000], where the authors proved the uniqueness of the minimizers, and justified the model by a thermodynamic limit argument. For a 3-dimensional periodic crystal, the minimization problem (2-17) has been examined in [Catto et al. 2001], where the authors showed the existence of minimizers and the uniqueness of the density of the minimizers. The characterization of the minimizers is given in [Cancès et al. 2008, Theorem 1]: the minimizer is unique and is a spectral projector satisfying a self-consistent equation. The following theorem provides similar results for a quasi-1-dimensional system: we show that the minimizer of (2-17) exists, and that the density of the minimizers is unique. Let us emphasize that the unit cell of a quasi-1-dimensional system is an unbounded domain Γ; hence we need to deal with the possible escaping of electrons in the $\boldsymbol{r}$-direction, a situation which need not be considered for bounded unit cells as in [Catto et al. 2001; Cancès et al. 2008].

Theorem 2.6 (existence of rHF ground state). *The minimization problem (2-17) admits a minimizer γ_{per} with density $\rho_{\gamma_{\text{per}}}$ belonging to $L^p_{\text{per},x}(\Gamma)$ for $1 \le p \le 3$. Additionally, all the minimizers share the same density.*

The proof of Theorem 2.6 relies on a classical variational argument and can be read in Section 4B.

In order to treat the junction of quasi-1-dimensional systems in Section 3, it is useful to define and study the mean-field potential $V_{\text{per}} = (\rho_{\gamma_{\text{per}}} - \mu_{\text{per}}) \star_\Gamma G$, where $\star_\Gamma$ denotes the convolution operator in the unit cell Γ and G is the x-periodic Green's function defined in (2-7). It is also critical to obtain some decay estimates of V_{per} in the $\boldsymbol{r}$-direction. However, since the Green's function G has a log-growth in the $\boldsymbol{r}$-direction, the L^p integrability of $\rho_{\gamma_{\text{per}}}$ obtained in Theorem 2.6 does not imply the decay of the mean-field potential V_{per} in the $\boldsymbol{r}$-direction. Moreover, the uniform bound given by the energy functional (2-15) does not provide any L^p bounds or a decay property of V_{per}. In view of this, we introduce the following assumption on $\rho_{\gamma_{\text{per}}}$. Note that this assumption, which called "summability condition", is common when treating the 2-dimensional Poisson equation [Lieb and Loss 2001, Theorem 6.21].

Assumption 1. *The unique ground state density $\rho_{\gamma_{\text{per}}}$ of the problem (2-17) satisfies*

$$\int_\Gamma |\boldsymbol{r}| \rho_{\gamma_{\text{per}}}(x, \boldsymbol{r})\, dx\, d\boldsymbol{r} < +\infty. \tag{2-18}$$

With this mild summability condition (2-18) on $\rho_{\gamma_{\text{per}}}$, we prove in Theorem 2.7 that $\rho_{\gamma_{\text{per}}}$ actually decays exponentially fast in the $|\boldsymbol{r}|$-direction (equation (2-22)), and the highest occupied energy (Fermi level) of electrons for a quasi-1-dimensional system in its ground state is always negative. This coincides with the physical reality: the additional summability condition on the density is sufficient to guarantee that the mean-field potential tends to 0 in the $\boldsymbol{r}$-direction. If the Fermi level is nonnegative, electrons can escape to infinity in the $\boldsymbol{r}$-direction, decreasing the energy of the system, and hence the system is not at ground state. Furthermore, we are able to characterize the unique minimizer as a spectral projector of the mean-field Hamiltonian. We comment on Assumption 1 in Remark 2.9, and we give an explicit example in which Assumption 1 is satisfied.

Example. If the mean-field potential V_{per} tends to 0 when $|\boldsymbol{r}|$ tends to $+\infty$, the density $\rho_{\gamma_{\text{per}}}$ decays exponentially fast in the $|\boldsymbol{r}|$-direction (see the proof of (2-22) in Section 4C); hence a posteriori Assumption 1 is satisfied.

Theorem 2.7 (properties of the rHF ground state with summability condition on the density). *Assume that Assumption 1 holds for the unique ground state density $\rho_{\gamma_{\text{per}}}$ of the minimization problem (2-17):*

(1) (the integrability of mean-field potential) *The mean-field potential*

$$V_{\text{per}} := (\rho_{\gamma_{\text{per}}} - \mu_{\text{per}}) \star_\Gamma G$$

belongs to $L^p_{\text{per},x}(\Gamma)$ for $2 < p \le +\infty$. Moreover, V_{per} is continuous and tends to zero in the $\boldsymbol{r}$-direction when $|\boldsymbol{r}| \to \infty$.

(2) (spectral properties of the mean-field Hamiltonian) *The mean-field Hamiltonian*

$$H_{\mathrm{per}} = \mathscr{B}^{-1}\left(\int_{\Gamma^*} H_{\mathrm{per},\xi}\,\frac{d\xi}{2\pi}\right)\mathscr{B} = -\frac{1}{2}\Delta + V_{\mathrm{per}}, \qquad H_{\mathrm{per},\xi} := -\frac{1}{2}\Delta_\xi + V_{\mathrm{per}}, \tag{2-19}$$

is a self-adjoint operator acting on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$ and form domain $H^1(\mathbb{R}^3)$. There exists $N_H \in \mathbb{N}^*$ which can be finite or infinite and a sequence $\{\lambda_n(\xi)\}_{\xi\in\Gamma^*,\,1\le n\le N_H}$ such that

$$\sigma_{\mathrm{ess}}(H_{\mathrm{per},\xi}) = [0,+\infty), \qquad \sigma_{\mathrm{disc}}(H_{\mathrm{per},\xi}) = \bigcup_{1\le n\le N_H} \lambda_n(\xi) \subset [-\|V_{\mathrm{per}}\|_{L^\infty}, 0).$$

Moreover, the following spectral decomposition holds:

$$\sigma(H_{\mathrm{per}}) = \sigma_{\mathrm{ess}}(H_{\mathrm{per}}) = \bigcup_{\xi\in\Gamma^*} \sigma(H_{\mathrm{per},\xi}), \qquad \bigcup_{\xi\in\Gamma^*} \sigma_{\mathrm{disc}}(H_{\mathrm{per},\xi}) \subseteq \sigma_{\mathrm{ac}}(H_{\mathrm{per}}). \tag{2-20}$$

In particular, $[0,+\infty) \subset \sigma_{\mathrm{ess}}(H_{\mathrm{per}})$.

(3) (the Fermi level is always negative) *The energy level counting function*

$$F(\kappa) : \kappa \mapsto \frac{1}{|\Gamma^*|}\int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}(\mathbb{1}_{(-\infty,\kappa]}(H_{\mathrm{per},\xi}))\,d\xi = \frac{1}{|\Gamma^*|}\sum_{n=1}^{N_H}\int_{\Gamma^*} \mathbb{1}(\lambda_n(\xi)\le\kappa)\,d\xi$$

is continuous and nondecreasing on $(-\infty,0]$. *The inequality*

$$N_H = F(0) \ge \int_\Gamma \mu_{\mathrm{per}}$$

holds, which means that there are always enough negative energy levels for the electrons. Moreover, there exists $\epsilon_F < 0$ called the Fermi level (chemical potential) such that $F(\epsilon_F) = \int_\Gamma \mu_{\mathrm{per}} = Z$, which represents the highest occupied energy level by electrons, and can be interpreted as the Lagrange multiplier associated with the charge neutrality condition (2-16).

(4) (the unique minimizer is a spectral projector) *The minimizer of the problem* (2-17) *is unique and satisfies the following self-consistent equation*:

$$\gamma_{\mathrm{per}} = \mathbb{1}_{(-\infty,\epsilon_F]}(H_{\mathrm{per}}) = \mathscr{B}^{-1}\left(\int_{\Gamma^*} \gamma_{\mathrm{per},\xi}\,\frac{d\xi}{2\pi}\right)\mathscr{B}, \qquad \gamma_{\mathrm{per},\xi} := \mathbb{1}_{(-\infty,\epsilon_F]}(H_{\mathrm{per},\xi}). \tag{2-21}$$

Furthermore, there exist positive constants C_{ϵ_F} and α_{ϵ_F} which depend on the Fermi level ϵ_F, such that

$$0 \le \rho_{\gamma_{\mathrm{per}}}(x,\boldsymbol{r}) \le C_{\epsilon_F}\mathrm{e}^{-\alpha_{\epsilon_F}|\boldsymbol{r}|}. \tag{2-22}$$

The proof of Theorem 2.7 can be read in Section 4C.

Remark 2.8. As the unit cell of the 1-dimensional system in the 3-dimensional space is an unbounded domain Γ, the decomposed mean-field Hamiltonian $H_{\mathrm{per},\xi}$ does not have a compact resolvent, which is a significant difference compared to the situation considered in [Catto et al. 2001; Cancès et al. 2008].

Remark 2.9. Let us comment on Assumption 1. The exponential decay of the density (2-22) implies the summability condition (2-18). However, we were not able to directly prove (2-18). This failure is mainly due to the lack of a priori summability bounds for the density matrices in $\mathcal{F}_\Gamma$. One might argue that we can add the condition (2-18) to the definition of $\mathcal{F}_\Gamma$. However, the set $\mathcal{F}_\Gamma$ with the condition (2-18) is not closed for the usual weak-$*$ topology when considering a minimizing sequence of (2-15). Another attempt is to use a Schauder fixed-point algorithm as in [Lions 1987; Cancès et al. 2009] to prove that (2-21) admits a solution. The most crucial step is to guarantee that there are enough negative bound states to meet the charge neutrality constraint (2-16). The number of bound states is controlled by the decay rate of potentials. With exponentially decaying densities we can show that [Blanc and Le Bris 2000, Lemma 2.5] there exists $C \in \mathbb{R}^+$ such that $|V_{\text{per}}(\cdot, \boldsymbol{r})| \le C|\boldsymbol{r}|^{-1}$. Nevertheless this condition is not sufficient to guarantee that there are enough bound states with negative energies, as the critical decay rate for numbers of bound states to be finite or infinite is $-|\boldsymbol{r}|^{-2}$ [Reed and Simon 1978, Theorem XIII.6]. In other words, we do not have a uniform bound over the Fermi level ϵ_F at each fixed-point iteration. On the other hand, the summability condition (2-18) is a sufficient but probably not a necessary condition for the negativity of the Fermi level and the characterization of the minimizers. The main difficult is to control the decay of the mean-field potential V_{per} in the $\boldsymbol{r}$-direction by just controlling the nuclear density μ_{per}, given that the Green's function defined in (2-7) has log-growth in the $\boldsymbol{r}$-direction. Furthermore, different decay scenarios of V_{per} in the $\boldsymbol{r}$-direction lead to different characterizations of the spectrum of the Hamiltonian H_{per}: if V_{per} is bounded from below, and positive with log-growth when $|\boldsymbol{r}| \to \infty$, one can show that the spectrum of $H_{\text{per},\xi}$ is purely discrete and the spectrum of H_{per} has a band structure. The Fermi level of the system could be positive in this case. We are not able to prove the above statements without Assumption 1. But we managed to show that the weak decay of Assumption 1 implies an exponential decay (2-22).

In order to describe the junction of quasi-1-dimensional systems, more specifically to guarantee that the Coulomb energy generated by the perturbative state is finite, the integrability of the mean-field potential provided in Theorem 2.7 is not sufficient in view of Lemma 3.3 below. In order to make use of this result, let us introduce a class of nuclear densities such that the x-averaged density is rotationally invariant in the $\boldsymbol{r}$-direction:

$$\mu_{\text{per,sym}}(x, \boldsymbol{r}) = \sum_{n\in\mathbb{Z}} Z\, m_{\text{s}}(x-n, \boldsymbol{r}),$$

where $m_{\text{s}}(x, \boldsymbol{r})$ is a nonnegative $C_c^\infty(\Gamma)$ function such that $\int_{\mathbb{R}^3} m_{\text{s}} = 1$. Moreover, there exists $m_{\text{sym}}(|\boldsymbol{r}|) \in C_c^\infty(\mathbb{R}^2)$ such that

$$\text{for all } \boldsymbol{r} \in \mathbb{R}^2, \qquad \int_{-1/2}^{1/2} m_{\text{s}}(x, \boldsymbol{r})\, dx \equiv m_{\text{sym}}(|\boldsymbol{r}|). \tag{2-23}$$

Lemma 2.10. *Suppose that Assumption 1 holds. Under the symmetry condition (2-23) on the nuclear density μ_{per}, all the results of Theorem 2.7 hold for the minimization problem (2-17). Additionally, the mean-field potential V_{per} belongs to $L^p_{\text{per},x}(\Gamma)$ for $1 < p \le +\infty$.*

The proof of this lemma can be read in Section A2. The nuclei of many actual materials can be modeled with a smear nuclear density satisfying the condition (2-23): for instance nanotubes and polymers with rotational symmetry in the $\boldsymbol{r}$-direction.

3. Mean-field stability for the junction of quasi-1-dimensional systems

In this section, we construct an rHF model for the junction of two different quasi-1-dimensional periodic systems. The junction system is described by periodic nuclei satisfying the symmetry condition (2-23) with different periodicities and possibly different charges per unit cell, occupying separately the left and right half-spaces (i.e., $(-\infty, 0] \times \mathbb{R}^2$ and $(0, +\infty) \times \mathbb{R}^2$); see Figure 2. We do not assume any commensurability of the different periodicities. The junction system is therefore a priori no longer periodic, making it impossible to define the periodic rHF energy. Inspired by perturbative approaches when treating infinitely extended systems [Hainzl et al. 2005a; 2005b; 2007; 2009; Cancès et al. 2008], the idea is to find an appropriate reference state which is close enough to the actual one. Section 3A gives a mathematical description of the junction system. Section 3B is devoted to a rigorous construction of a reference Hamiltonian H_χ and a reference one-particle density matrix defined as a spectral projector of H_χ. In Section 3C we construct a perturbative state, which encodes the nonlinear effects due to the electron-electron interaction in the rHF approximation, and associate the ground state energy of this perturbative state to some minimization problem in Section 3D.

3A. ***Mathematical description of the junction system.*** Consider two quasi-1-dimensional periodic systems with periods $a_L > 0$ and $a_R > 0$. The unit cells are respectively denoted by $\Gamma_L := [-a_L/2, a_L/2) \times \mathbb{R}^2$ and $\Gamma_R := [-a_R/2, a_R/2) \times \mathbb{R}^2$ with their duals $\Gamma_L^* := [-\pi/a_L, \pi/a_L)$ and $\Gamma_R^* := [-\pi/a_R, \pi/a_R)$. We consider nuclear densities fulfilling the symmetry condition (2-23) and suppose that Assumption 1 holds for the ground state densities of both quasi-1-dimensional periodic systems. More precisely, let $m_L(x, \boldsymbol{r})$ and $m_R(x, \boldsymbol{r})$ be nonnegative C_c^∞ functions with supports respectively in Γ_L and Γ_R such that $\int_{\mathbb{R}^3} m_L = 1$ and $\int_{\mathbb{R}^3} m_R = 1$. Assume that there exist $m_{\mathrm{sym},L}(|\boldsymbol{r}|), m_{\mathrm{sym},R}(|\boldsymbol{r}|) \in C_c^\infty(\mathbb{R}^2)$ such that,

$$\text{for all } \boldsymbol{r} \in \mathbb{R}^2, \quad \int_{-a_L/2}^{a_L/2} m_L(x, \boldsymbol{r})\, dx \equiv m_{\mathrm{sym},L}(|\boldsymbol{r}|), \quad \int_{-a_R/2}^{a_R/2} m_R(x, \boldsymbol{r})\, dx \equiv m_{\mathrm{sym},R}(|\boldsymbol{r}|).$$

Denoting by $Z_L, Z_R \in \mathbb{N}\backslash\{0\}$ the total charges of the nuclei per unit cells, the smeared periodic nuclear densities are respectively defined as

$$\mu_{\mathrm{per},L}(x, \boldsymbol{r}) := \sum_{n\in\mathbb{Z}} Z_L\, m_L(x - a_L n, \boldsymbol{r}), \quad \mu_{\mathrm{per},R}(x, \boldsymbol{r}) := \sum_{n\in\mathbb{Z}} Z_R\, m_R(x - a_R n, \boldsymbol{r}). \tag{3-1}$$

The periodic Green's functions with periods Γ_L and Γ_R are separately defined as

$$G_{a_L}(x, \boldsymbol{r}) = a_L^{-1} G\left(\frac{x}{a_L}, \boldsymbol{r}\right), \quad G_{a_R}(x, \boldsymbol{r}) = a_R^{-1} G\left(\frac{x}{a_R}, \boldsymbol{r}\right),$$

where $G(\cdot)$ is the periodic Green's function defined in (2-7). One can easily verify that

$$-\Delta G_{a_L}(x, \boldsymbol{r}) = 4\pi \sum_{n\in\mathbb{Z}} \delta_{(x,\boldsymbol{r})=(a_L n, \boldsymbol{0})} \in \mathscr{S}'(\mathbb{R}^3), \quad -\Delta G_{a_R}(x, \boldsymbol{r}) = 4\pi \sum_{n\in\mathbb{Z}} \delta_{(x,\boldsymbol{r})=(a_R n, \boldsymbol{0})} \in \mathscr{S}'(\mathbb{R}^3).$$

According to the results of Theorem 2.7, the following self-consistent equations uniquely define the ground states density matrices associated with the periodic nuclear densities $\mu_{\mathrm{per},L}$ and $\mu_{\mathrm{per},R}$:

$$\gamma_{\mathrm{per},L} := \mathbb{1}_{(-\infty,\epsilon_L]}(H_{\mathrm{per},L}), \quad H_{\mathrm{per},L} := -\frac{\Delta}{2} + V_{\mathrm{per},L}, \quad V_{\mathrm{per},L} := (\rho_{\mathrm{per},L} - \mu_{\mathrm{per},L}) \star_{\Gamma_L} G_{a_L},$$
$$\gamma_{\mathrm{per},R} := \mathbb{1}_{(-\infty,\epsilon_R]}(H_{\mathrm{per},R}), \quad H_{\mathrm{per},R} := -\frac{\Delta}{2} + V_{\mathrm{per},R}, \quad V_{\mathrm{per},R} := (\rho_{\mathrm{per},R} - \mu_{\mathrm{per},R}) \star_{\Gamma_R} G_{a_R},$$

where the negative constants ϵ_L and ϵ_R are the Fermi levels of the quasi-1-dimensional systems. The junction of the quasi-1-dimensional systems is described by considering the following nuclear density configuration (see Figure 2):

$$\mu_J(x,\boldsymbol{r}) := \mathbb{1}_{x\le 0}\cdot \mu_{\mathrm{per},L}(x,\boldsymbol{r}) + \mathbb{1}_{x>0}\cdot \mu_{\mathrm{per},R}(x,\boldsymbol{r}) + \upsilon(x,\boldsymbol{r}), \tag{3-2}$$

where $\upsilon(x,\boldsymbol{r}) \in L^{6/5}(\mathbb{R}^3)$ describes how the junction switches between the underlying nuclear densities. The assumption $\upsilon \in L^{6/5}(\mathbb{R}^3)$ ensures that $D(\upsilon,\upsilon) < +\infty$. Recall that

$$D(f,g) = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{f(x)g(y)}{|x-y|}\,dx\,dy = \frac{1}{4\pi}\int_{\mathbb{R}^3} \frac{\overline{\hat{f}(k)}\hat{g}(k)}{|k|^2}\,dk$$

describes the Coulomb interactions in the whole space. Once one sets the nuclear configuration (3-2), electrons are allowed to move in the 3-dimensional space. The *infinite* rHF energy functional for the junction system associated with a test density matrix γ_J formally reads

$$\mathcal{E}(\gamma_J) = \mathrm{Tr}\big(-\tfrac{1}{2}\Delta\gamma_J\big) + \tfrac{1}{2}D(\rho_{\gamma_J} - \mu_J, \rho_{\gamma_J} - \mu_J). \tag{3-3}$$

Let us also introduce the Coulomb space $\mathcal{C}$ and its dual $\mathcal{C}'$ (Beppo-Levi space [Cancès et al. 2008]):

$$\mathcal{C} := \{\rho \in \mathscr{S}'(\mathbb{R}^3) \mid \hat{\rho} \in L^1_{\mathrm{loc}}(\mathbb{R}^3),\ D(\rho,\rho) < \infty\}, \quad \mathcal{C}' := \{V \in L^6(\mathbb{R}^3) \mid \nabla V \in (L^2(\mathbb{R}^3))^3\}. \tag{3-4}$$

Note that the ground state energy of the junction system, if it exists, is infinite and there is no periodicity in this system; hence usual techniques which essentially consist of considering the energy per unit volume [Catto et al. 2001; Cancès et al. 2008] are not applicable. We next define a reference system such that the difference between the junction system and the reference can be considered as a perturbation. This perturbative approach has been used in [Hainzl et al. 2005a; 2007; 2009; Cancès et al. 2008] in various contexts. The next section is devoted to the rigorous mathematical construction of the reference state and its rHF energy functional.

3B. ***Reference state for the junction system.*** In this section, we construct a reference Hamiltonian obtained by a linear combination of the periodic mean-field potentials $V_{\mathrm{per},L}$ and $V_{\mathrm{per},R}$. We prove the validity of this approach by showing that the density generated by this reference state is close to the linear combination of the periodic densities $\rho_{\mathrm{per},L}$ and $\rho_{\mathrm{per},R}$.

Hamiltonian of the reference state. We introduce a class of smoothed cut-off functions. For $\boldsymbol{x} \in \mathbb{R}^3$, consider

$$\mathcal{X} := \Big\{\chi \in C^2(\mathbb{R}^3)\,\Big|\, 0 \le \chi \le 1,\ \chi(\boldsymbol{x}) = 1 \text{ if } \boldsymbol{x} \in \Big(-\infty, -\frac{a_L}{2}\Big]\times\mathbb{R}^2,\ \chi(\boldsymbol{x}) = 0 \text{ if } \boldsymbol{x} \in \Big[\frac{a_R}{2}, +\infty\Big)\times\mathbb{R}^2\Big\}. \tag{3-5}$$

Fix $\chi \in \mathcal{X}$, let us introduce a reference potential

$$V_\chi := \chi^2 V_{\mathrm{per},L} + (1-\chi^2) V_{\mathrm{per},R}.$$

We will show in Section 3D that the choice of $\chi \in \mathcal{X}$ is irrelevant. By Theorem 2.7 and Lemma 2.10 we know that V_χ belongs to $L^p_{\mathrm{loc}}(\mathbb{R}, L^p(\mathbb{R}^2))$ for $1 < p \leq \infty$, is continuous in all directions and tends to zero in the $\boldsymbol{r}$-direction. By the Kato–Rellich theorem (see for example [Helffer 2013, Theorem 9.10]), there exists a unique self-adjoint operator

$$H_\chi := -\tfrac{1}{2}\Delta + V_\chi \tag{3-6}$$

on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$ and form domain $H^1(\mathbb{R}^3)$. We next show that the essential spectrum of the reference Hamiltonian H_χ is the union of the essential spectra of $H_{\mathrm{per},L}$ and $H_{\mathrm{per},R}$, which implies that the reference system does not change essentially the unions of possible energy levels of quasiperiodic systems, and that there are no surface states which propagate along the junction surface in the $\boldsymbol{r}$-direction. Note that this is a priori not obvious as the cut-off function χ is $\boldsymbol{r}$-translation-invariant (hence not compact), therefore scattering states may occur at the junction surface and escape to infinity in the $\boldsymbol{r}$-direction. Standard techniques in scattering theory to prove this statement, such as Dirichlet decoupling [Deift and Simon 1976; Hempel et al. 2015], are not applicable in our situation since the junction surface is not compact.

Proposition 3.1 (spectral properties of the reference state H_χ). *For any $\chi \in \mathcal{X}$, the essential spectrum of H_χ defined in (3-6) satisfies*

$$\sigma_{\mathrm{ess}}(H_\chi) = \sigma_{\mathrm{ess}}(H_{\mathrm{per},L}) \cup \sigma_{\mathrm{ess}}(H_{\mathrm{per},R}).$$

In particular, $[0, +\infty) \subset \sigma_{\mathrm{ess}}(H_\chi)$ and $\sigma_{\mathrm{ess}}(H_\chi)$ does not depend on the shape of the cut-off function $\chi \in \mathcal{X}$ defined in (3-5).

The proof can be found in Section A3 of the Appendix. Note that Proposition 3.1 also implies that the reference system essentially preserves the scattering channels of the underlying quasi-1-dimensional systems, since the scattering involves the purely absolutely continuous spectrum of a Hamiltonian (see for example [Exner and Frank 2007; Bruneau et al. 2016a; 2016b]). However, this does not exclude the existence of embedded eigenvalues in the essential spectrum, which may cause additional scattering channels [Frank 2003; Frank and Shterenberg 2004]. We prove in Corollary 3.5.1 that the results in Proposition 3.1 still hold for the nonlinear junction.

Reference state as a spectral projector. Before constructing the reference state, let us discuss different regimes for junction system. From Theorem 2.7 we know that the chemical potentials (Fermi levels) ϵ_L and ϵ_R are negative. Introduce the energy interval $I_{\epsilon_F} := [\min(\epsilon_L, \epsilon_R), \max(\epsilon_L, \epsilon_R)]$. In view of Proposition 3.1, assume that the essential spectrum of H_χ below 0 is purely absolutely continuous, the nonequilibrium regime (Figure 3) corresponds to

$$\sigma_{\mathrm{ac}}(H_\chi) \cap I_{\epsilon_F} \neq \varnothing.$$

In this regime, steady state currents occur and the Landauer–Büttiker conductance can be calculated [Bruneau et al. 2015; 2016a; 2016b]. When $\mu_{\mathrm{per},L}$ and $\mu_{\mathrm{per},R}$ are identical, the junction system becomes

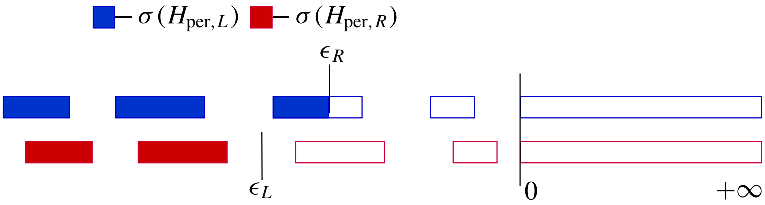


Figure 3. Spectrum of $H_{\mathrm{per},L}$, $H_{\mathrm{per},R}$ in the nonequilibrium regime.

periodic with different chemical potentials ϵ_L and ϵ_R on the left and right half-lines. In this case the Thouless conductance [Bruneau et al. 2015] can be defined and it is given by

$$C_T \frac{|\sigma_{\mathrm{ac}}(H_\chi) \bigcap I_{\epsilon_F}|}{|I_{\epsilon_F}|} > 0,$$

for some positive constant C_T. However it is not the aim of this article to discuss steady state currents for nonequilibrium systems. We instead consider the equilibrium regime (see Figure 4) with the following assumption.

Assumption 2. *The chemical potentials ϵ_L and ϵ_R are in a common spectral gap (Σ_a, Σ_b) (equilibrium regime, see Figure 4), where Σ_a is the maximum of the filled bands of $H_{\mathrm{per},L}$ and $H_{\mathrm{per},R}$, and Σ_b is the minimum of the unfilled bands of $H_{\mathrm{per},L}$ and $H_{\mathrm{per},R}$.*

Assumption 2 guarantees that the Fermi level of the junction system lies in a spectral gap of H_χ in view of Proposition 3.1, which is a common hypothesis [Cancès et al. 2008; Hainzl et al. 2005a; 2009] for 3-dimensional periodic insulating and semiconducting systems. We make this assumption for simplicity. Note that with approaches proposed in [Frank et al. 2011; 2013; Cancès et al. 2020] it is possible to extend the results to metallic systems provided that the junction system is in its ground state and no steady state current occurs.

Let us without loss of generality choose the Fermi level $\epsilon_F = \max(\epsilon_L, \epsilon_R) = \sup I_{\epsilon_F}$ and define the reference state γ_χ as the spectral projector associated with the states of H_χ below ϵ_F:

$$\gamma_\chi := \mathbb{1}_{(-\infty,\epsilon_F)}(H_\chi). \tag{3-7}$$

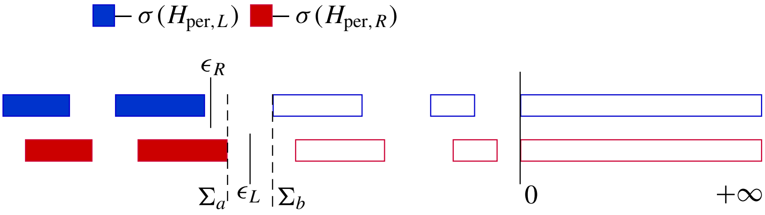


Figure 4. Spectrum of $H_{\mathrm{per},L}$, $H_{\mathrm{per},R}$ in the equilibrium regime.

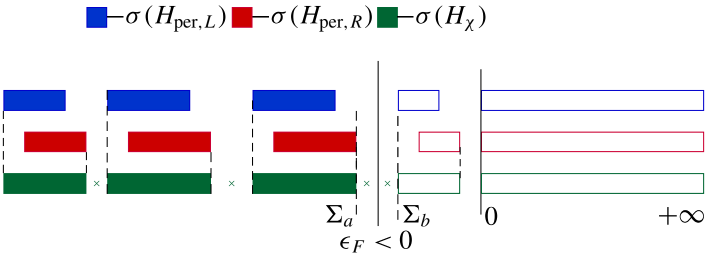


Figure 5. Spectrum of $H_{\mathrm{per},L}$, $H_{\mathrm{per},R}$ and H_χ below 0.

Note that H_χ can have discrete spectrum in the gap (Σ_a, Σ_b), with eigenvalues possibly accumulating at Σ_a and Σ_b, and ϵ_F can also be an eigenvalue of H_χ; see Figure 5. The definition of (3-7), however, excludes the possible bound states with energy ϵ_F.

The following proposition shows that the density ρ_χ of γ_χ is well-defined in $L^1_{\mathrm{loc}}(\mathbb{R}^3)$ and is close to the linear combination of the periodic densities $\rho_{\mathrm{per},L}$ and $\rho_{\mathrm{per},R}$, the difference decaying exponentially fast in the x-direction as $|x| \to \infty$.

Proposition 3.2 (exponential decay of density). *Under Assumption 2, the spectral projector γ_χ is locally trace class, so that its density ρ_χ is well-defined in $L^1_{\mathrm{loc}}(\mathbb{R}^3)$. Moreover,*

$$\chi^2\rho_{\mathrm{per},L} + (1-\chi^2)\rho_{\mathrm{per},R} - \rho_\chi \in L^p(\mathbb{R}^3) \quad \textit{for } 1 < p \le 2.$$

Furthermore, denote by w_a the characteristic function of the unit cube centered at $a \in \mathbb{R}^3$. There exist positive constants C and t such that for all

$$\alpha = (\alpha_x, 0, 0) \in \mathbb{R}^3, \quad \textit{with either } \operatorname{supp}(w_\alpha) \subset (-\infty, a_L/2] \times \mathbb{R}^2 \textit{ or } \operatorname{supp}(w_\alpha) \subset [a_R/2, +\infty) \times \mathbb{R}^2,$$

it holds

$$\int_{\mathbb{R}^3} \bigl| w_\alpha \bigl(\chi^2\rho_{\mathrm{per},L} + (1-\chi^2)\rho_{\mathrm{per},R} - \rho_\chi\bigr) w_\alpha \bigr| \le C e^{-t|\alpha|}.$$

The proof can be read in Section 4D.

Fictitious nuclear density of the reference state. The density ρ_χ associated with γ_χ is fixed once the Fermi level ϵ_F is chosen. We can therefore define a fictitious nuclear density μ_χ by imposing that the total electronic density $\rho_\chi - \mu_\chi$ generates the potential V_χ. The fictitious nuclear density μ_χ is given by

$$-\Delta V_\chi = 4\pi(\rho_\chi - \mu_\chi), \quad \mu_\chi := \rho_\chi - \bigl(\chi^2(\rho_{\mathrm{per},L} - \mu_{\mathrm{per},L}) + (1-\chi^2)(\rho_{\mathrm{per},R} - \mu_{\mathrm{per},R}) + \eta_\chi\bigr), \tag{3-8}$$

where η_χ has compact support in the x-direction:

$$\eta_\chi := -\frac{1}{4\pi}\bigl(\partial_x^2(\chi^2)(V_{\mathrm{per},L} - V_{\mathrm{per},R}) + 2\partial_x(\chi^2)\partial_x(V_{\mathrm{per},L} - V_{\mathrm{per},R})\bigr). \tag{3-9}$$

Let us emphasize that the Poisson's equation (3-8) is defined on the whole space $\mathbb{R}^3$.

The nuclear density of the junction is a fictitious nuclear density plus a perturbation. Once we have defined fictitious nuclear density, we can treat the difference between the real nuclear density of the junction system μ_J and the fictitious nuclear density μ_χ as a perturbative nuclear density. By doing so we can define a finite renormalized energy with respect to the perturbative nuclear density. Note that this idea is similar to the definition of the defect state in [Cancès et al. 2008] for defects in crystals, and the polarization of the vacuum in the Bogoliubov–Dirac–Fock model [Hainzl et al. 2005a; 2007; 2009]. Introduce

$$\nu_\chi := \mu_J - \mu_\chi = (\mathbb{1}_{x\le 0} - \chi^2)(\mu_{\mathrm{per},L} - \mu_{\mathrm{per},R}) + (\chi^2 \rho_{\mathrm{per},L} + (1-\chi^2)\rho_{\mathrm{per},R} - \rho_\chi) + \eta_\chi + \upsilon. \quad (3\text{-}10)$$

In order to guarantee that the perturbative state has a finite Coulomb energy, we need $D(\nu_\chi, \nu_\chi) < +\infty$. A sufficient condition is that ν_χ belongs to $L^{6/5}(\mathbb{R}^3)$. This motivates the following L^p-estimate on η_χ.

Lemma 3.3. *The function η_χ defined in (3-9) belongs to $L^p(\mathbb{R}^3)$ for $1 < p < 6$.*

The proof can be read in Section 4E. In view of Lemma 3.3 and Proposition 3.2, together with the fact that $(\mathbb{1}_{x\le 0} - \chi^2)(\mu_{\mathrm{per},L} - \mu_{\mathrm{per},R})$ has compact support and υ belongs to $L^{6/5}(\mathbb{R}^3)$, it is easy to see that ν_χ belongs to $L^{6/5}(\mathbb{R}^3)$, and hence to the Coulomb space $\mathcal{C}$ defined in (3-4). This means that the perturbative state generated by the nuclear density ν_χ has finite Coulomb energy.

Remark 3.4. The integrability of V_{per} provided by Lemma 2.10 is crucial to deduce Lemma 3.3.

3C. *Definition of the perturbative state.* In this section we define a perturbative state associated with the perturbative density ν_χ following the ideas developed in [Cancès et al. 2008]. We formally derive the rHF energy difference between the junction state γ_J and the reference state γ_χ by writing $\gamma_J = \gamma_\chi + Q_\chi$ with Q_χ a trial density state. In view of (3-3), we formally have

$$\begin{aligned}
&\mathcal{E}(\gamma_J) - \mathcal{E}(\gamma_\chi)\\
&= \mathrm{Tr}\big(-\tfrac12\Delta(\gamma_\chi + Q_\chi)\big) + \tfrac12 D(\rho_J - \mu_J, \rho_J - \mu_J) - \mathrm{Tr}\big(-\tfrac12\Delta\gamma_\chi\big) - \tfrac12 D(\rho_\chi - \mu_\chi, \rho_\chi - \mu_\chi)\\
&= \mathrm{Tr}(-\tfrac12\Delta Q_\chi) + D(\rho_\chi - \mu_\chi, \rho_{Q_\chi}) - D(\rho_{Q_\chi}, \nu_\chi) + \tfrac12 D(\rho_{Q_\chi}, \rho_{Q_\chi}) - D(\rho_\chi - \mu_\chi, \nu_\chi) + \tfrac12 D(\nu_\chi, \nu_\chi)\\
&= \mathrm{Tr}(H_\chi Q_\chi) - D(\rho_{Q_\chi}, \nu_\chi) + \tfrac12 D(\rho_{Q_\chi}, \rho_{Q_\chi}) - D(\rho_\chi - \mu_\chi, \nu_\chi) + \tfrac12 D(\nu_\chi, \nu_\chi). \qquad (3\text{-}11)
\end{aligned}$$

We next give a mathematical definition of the terms in the last equality of (3-11). We expect Q_χ to be a perturbation of the reference state γ_χ. More precisely, we expect Q_χ to be Hilbert–Schmidt. This is usually called the Shale–Stinespring condition [1965]; see [Lewin 2009; Solovej 2005] for a detailed discussion. Moreover, we also expect the kinetic energy of Q_χ to be finite. Let Π be an orthogonal projector on the Hilbert space $\mathfrak{H}$ such that both Π and $\Pi^\perp := 1 - \Pi$ have infinite ranks. A self-adjoint compact operator A on $\mathfrak{H}$ is said to be Π-trace class if $A \in \mathfrak{S}_2(\mathfrak{H})$ and both $\Pi A \Pi$ and $\Pi^\perp A \Pi^\perp$ are in $\mathfrak{S}_1(\mathfrak{H})$. For an operator A we define its Π-trace as

$$\mathrm{Tr}_\Pi(A) := \mathrm{Tr}(\Pi A \Pi) + \mathrm{Tr}(\Pi^\perp A \Pi^\perp),$$

and denote by $\mathfrak{S}_1^\Pi(\mathfrak{H})$ the associated set of Π-trace class operators. Since the reference state γ_χ defined in (3-7) is an orthogonal projector on $L^2(\mathbb{R}^3)$, we can define associated γ_χ-trace class operators. For any trial density matrix Q_χ, let us define by $Q_\chi^{++} := \gamma_\chi^\perp Q_\chi \gamma_\chi^\perp$ and $Q_\chi^{--} := \gamma_\chi Q_\chi \gamma_\chi$, and introduce a Banach space of operators with finite γ_χ-trace and finite kinetic energy as follows:

$$\mathcal{Q}_\chi := \big\{Q_\chi \in \mathfrak{S}_1^{\gamma_\chi}(L^2(\mathbb{R}^3)) \,\big|\, Q_\chi^* = Q_\chi,\ |\nabla| Q_\chi \in \mathfrak{S}_2(L^2(\mathbb{R}^3)),\\ |\nabla| Q_\chi^{++} |\nabla| \in \mathfrak{S}_1(L^2(\mathbb{R}^3)),\ |\nabla| Q_\chi^{--} |\nabla| \in \mathfrak{S}_1(L^2(\mathbb{R}^3))\big\},$$

equipped with its natural norm

$$\|Q_\chi\|_{\mathcal{Q}_\chi} := \|Q_\chi\|_{\mathfrak{S}_2} + \|Q_\chi^{++}\|_{\mathfrak{S}_1} + \|Q_\chi^{--}\|_{\mathfrak{S}_1} + \||\nabla| Q_\chi\|_{\mathfrak{S}_2} + \||\nabla| Q_\chi^{++} |\nabla|\|_{\mathfrak{S}_1} + \||\nabla| Q_\chi^{--} |\nabla|\|_{\mathfrak{S}_1}.$$

By construction $\operatorname{Tr}_{\gamma_\chi}(Q_\chi) = \operatorname{Tr}(Q_\chi^{++}) + \operatorname{Tr}(Q_\chi^{--})$. For Q to be an admissible perturbation of the reference state γ_χ, Pauli's principle requires that $0 \le \gamma_\chi + Q_\chi \le 1$. Let us introduce the following convex set of admissible perturbative states:

$$\mathcal{K}_\chi := \{Q_\chi \in \mathcal{Q}_\chi \mid -\gamma_\chi \le Q_\chi \le 1 - \gamma_\chi\}.$$

Note that $\mathcal{K}_\chi$ is not empty since it contains at least 0. Note also that $\mathcal{K}_\chi$ is the convex hull of states in $\mathcal{Q}_\chi$ of the special form $\gamma - \gamma_\chi$, where γ is an orthogonal projector [Cancès et al. 2008]. Furthermore, for any $Q_\chi \in \mathcal{K}_\chi$ a simple algebraic calculation shows that

$$Q_\chi^{++} \ge 0, \quad Q_\chi^{--} \le 0, \quad 0 \le Q_\chi^2 \le Q_\chi^{++} - Q_\chi^{--}.$$

As mentioned in the previous section, the Fermi level ϵ_F can be an eigenvalue of H_χ. Consider $N \in \mathbb{N}^*$ such that $\epsilon_F \in (\Sigma_{N,\chi}, \Sigma_{N+1,\chi}]$, where $\Sigma_{N,\chi} < \Sigma_{N+1,\chi}$ are two eigenvalues of H_χ in the gap (Σ_a, Σ_b), and let $\Sigma_{N,\chi} = \Sigma_a$ and $\Sigma_{N+1,\chi} = \Sigma_b$ whenever there is no such element. For any $\kappa \in (\Sigma_{N,\chi}, \epsilon_F)$, let us introduce the following rHF kinetic energy of a state $Q_\chi \in \mathcal{Q}_\chi$:

$$\operatorname{Tr}_{\gamma_\chi}(H_\chi Q_\chi) := \operatorname{Tr}\big(|H_\chi - \kappa|^{1/2}(Q_\chi^{++} - Q_\chi^{--})|H_\chi - \kappa|^{1/2}\big) + \kappa \operatorname{Tr}_{\gamma_\chi}(Q_\chi).$$

By [Cancès et al. 2008, Corollary 1], the above expression is independent of $\kappa \in (\Sigma_{N,\chi}, \epsilon_F)$. In view of the last line of (3-11) we introduce the following minimization problem

$$E_{\kappa,\chi} = \inf_{Q_\chi \in \mathcal{K}_\chi} \{\mathcal{E}_\chi(Q_\chi) - \kappa \operatorname{Tr}_{\gamma_\chi}(Q_\chi)\}, \tag{3-12}$$

where

$$\mathcal{E}_\chi(Q_\chi) := \operatorname{Tr}_{\gamma_\chi}(H_\chi Q_\chi) - D(\rho_{Q_\chi}, \nu_\chi) + \tfrac{1}{2} D(\rho_{Q_\chi}, \rho_{Q_\chi}). \tag{3-13}$$

3D. ***Properties of the junction system.*** The following result shows that the minimization problem (3-12) is well-posed and admits minimizers.

Proposition 3.5 (existence of the perturbative ground state). *Assume that Assumption 2 holds. Then there exist minimizers for the problem (3-12). There may be several minimizers, but they all share the same density. Moreover, any minimizer $\overline{Q}_\chi$ of (3-12) satisfies the following self-consistent equation:*

$$\begin{cases} \overline{Q}_\chi = \mathbb{1}_{(-\infty,\epsilon_F)}(H_{\overline{Q}_\chi}) - \gamma_\chi + \delta_\chi, \\ H_{\overline{Q}_\chi} = H_\chi + (\rho_{\overline{Q}_\chi} - \nu_\chi) \star |\cdot|^{-1}, \end{cases} \tag{3-14}$$

where δ_χ is a finite-rank self-adjoint operator satisfying $0 \le \delta_\chi \le 1$ and $\operatorname{Ran}(\delta_\chi) \subseteq \operatorname{Ker}(H_{\overline{Q}_\chi} - \epsilon_F)$.

The proof is a direct adaptation of several results obtained in [Cancès et al. 2008, Proposition 1, Lemma 2, Corollary 1, Corollary 2 and Theorem 2]. A short summary of main arguments can also be read in [Cao 2019b, Proposition 3.5]. Note that $(\rho_{\bar{Q}_\chi} - \nu_\chi) \star |\cdot|^{-1} \in L^6(\mathbb{R}^3)$ by [Cancès and Stoltz 2012, Lemma 16]; therefore $(1-\Delta)^{-1}(\rho_{\bar{Q}_\chi} - \nu_\chi) \star |\cdot|^{-1}$ belongs to $\mathfrak{S}_6$ by the Kato–Seiler–Simon inequality (4-1). Hence $(\rho_{\bar{Q}_\chi} - \nu_\chi) \star |\cdot|^{-1}$ is $-\Delta$-compact and thus H_χ-compact by the boundedness of V_χ, leaving the essential spectrum unchanged. Therefore in view of Proposition 3.1, the following corollary holds.

Corollary 3.5.1. *For any $\chi \in \mathcal{X}$ and $H_{\bar{Q}_\chi}$ a solution of (3-14), it holds*

$$\sigma_{\mathrm{ess}}(H_{\bar{Q}_\chi}) = \sigma_{\mathrm{ess}}(H_{\mathrm{per},L}) \cup \sigma_{\mathrm{ess}}(H_{\mathrm{per},R}), \quad \sigma_{\mathrm{ess}}(H_{\bar{Q}_\chi}) \cap (-\infty, 0] \subseteq \sigma_{\mathrm{ac}}(H_{\bar{Q}_\chi}).$$

In particular, $[0,+\infty) \subset \sigma_{\mathrm{ess}}(H_{\bar{Q}_\chi})$ and $\sigma_{\mathrm{ess}}(H_{\bar{Q}_\chi})$ does not depend on the shape of the cut-off function $\chi \in \mathcal{X}$ defined in (3-5).

The result of Proposition 3.5 can be interpreted as follows: given a cut-off function χ belonging to the class $\mathcal{X}$ defined in (3-5), we can construct a reference state γ_χ and a perturbative ground state $\bar{Q}_\chi$, the sum of which forms the ground state of the junction system. However it is artificial to introduce cut-off functions χ since there are infinitely many possible choices. In view of (3-2), the ground state of the junction system should not depend on the choice of cut-off functions. The following theorem shows that the electronic density of the junction system is indeed independent of the choice of the cut-off function χ.

Theorem 3.6 (independence of the reference state and uniqueness of ground state density). *The ground state density of the junction system with nuclear density defined in (3-2) under the rHF description is independent of the choice of the cut-off function $\chi \in \mathcal{X}$; i.e., the total electronic density $\rho_J = \rho_\chi + \rho_{Q_\chi}$ is independent of χ, where ρ_χ is the density associated with the spectral projector γ_χ defined in (3-7), and ρ_{Q_χ} is the unique density associated with the solution Q_χ of the minimization problem (3-14).*

The proof can be read in Section 4F. Theorem 3.6 and Proposition 3.5 together imply that:

Corollary 3.6.1. *The ground state of the junction system (3-2) is of the form*

$$\mathbb{1}_{(-\infty,\epsilon_F)}(H_\chi + (\rho_{\bar{Q}_\chi} - \nu_\chi)\star|\cdot|^{-1}) + \delta_\chi, \quad 0 \le \delta_\chi \le 1, \quad \mathrm{Ran}(\delta_\chi) \subseteq \mathrm{Ker}(H_\chi + (\rho_{\bar{Q}_\chi} - \nu_\chi)\star|\cdot|^{-1} - \epsilon_F),$$

and its density is independent of the choice of χ.

Note that an extension to junctions of 2-dimensional materials may be done by similar constructions as above; see [Cao 2019a] for more details.

4. Proofs of the results

In order to simplify the notation, in Sections 4A–4C when treating the quasi-1-dimensional periodic system we denote by $\mathfrak{S}_p$ the Schattern class $\mathfrak{S}_p(L^2_{\mathrm{per},x}(\Gamma))$ for $1 \le p \le +\infty$. Unless otherwise specified, starting from Section 4D we use $\mathfrak{S}_p$ instead of $\mathfrak{S}_p(L^2(\mathbb{R}^3))$ for the proofs of the junction system. First of all, let us recall the following Kato–Seiler–Simon (KSS) inequality:

Lemma 4.1 [Seiler and Simon 1975, Lemma 2.1]. *Let* $2 \leq p \leq \infty$. *For* g, f *belonging to* $L^p(\mathbb{R}^3)$, *the following inequality holds*:

$$\|f(-\mathrm{i}\nabla)g(x)\|_{\mathfrak{S}_p(L^2(\mathbb{R}^3))} \leq (2\pi)^{-3/p}\|g\|_{L^p(\mathbb{R}^3)}\|f\|_{L^p(\mathbb{R}^3)}. \tag{4-1}$$

4A. ***Proof of Lemma 2.5.*** We prove this lemma by an explicit construction of a density matrix belonging to $\mathcal{F}_\Gamma$. Consider a cut-off function $\psi \in C_c^\infty(\Gamma)$ such that $0 \leq \psi \leq 1$ and $\int_\Gamma \psi^2 = 1$, and define $\psi_{\text{per}} = \sum_{n\in\mathbb{N}} \psi(\cdot - n)$. Let $\omega \geq 0$ be a parameter to be made precise later. Define

$$\gamma_\omega = \mathbb{1}_{[0,\omega]}(-\Delta)\psi_{\text{per}}^2\mathbb{1}_{[0,\omega]}(-\Delta). \tag{4-2}$$

It is easy to see that $0 \leq \gamma_\omega \leq 1$, and that $\tau_k^x\gamma_\omega = \gamma_\omega\tau_k^x$ for all $k \in \mathbb{Z}$ by construction. It is also easy to see that the density of γ_ω is

$$\rho_{\gamma_\omega} = \frac{1}{(2\pi)^3}\int_{|k^2|\leq\omega} dk = (6\pi)^{-2}\omega^{3/2}\psi_{\text{per}}^2,$$

which is smooth and Γ-periodic. Moreover, it holds $\int_\Gamma \rho_{\gamma_\omega} = (6\pi)^{-2}\omega^{2/3}$. The kinetic energy per unit cell of γ_ω can be written as

$$\mathrm{Tr}_{L^2_{\text{per},x}(\Gamma)}\big(|\nabla|\mathbb{1}_{[0,\omega]}(-\Delta)\psi_{\text{per}}^2\mathbb{1}_{[0,\omega]}(-\Delta)|\nabla|\big) = \frac{1}{(2\pi)^3}\int_{|k|^2\leq\omega}|k|^2\,dk = \frac{1}{10\pi^2}\omega^{5/2}.$$

Hence for all finite ω, we have γ_ω belongs to $\mathcal{P}_{\text{per},x}$. Let us now show that there exists $\omega_* \geq 0$ such that $\rho_{\gamma_{\omega_*}} - \mu_{\text{per}} \in \mathcal{C}_\Gamma$. It is easy to see that there exists $\omega_* > 0$ such that

$$\int_\Gamma \rho_{\gamma_{\omega*}} = (6\pi)^{-2}\omega_*^{2/3} = \int_\Gamma \mu_{\text{per}}. \tag{4-3}$$

This condition is equivalent to $\mathscr{F}(\rho_{\gamma_{\omega_*}} - \mu_{\text{per}})(0, \mathbf{0}) = 0$. As $\boldsymbol{k} \mapsto \mathscr{F}(\rho_{\gamma_{\omega_*}} - \mu_{\text{per}})(0, \boldsymbol{k})$ is $C^1(\mathbb{R}^2)$ and bounded, the function $\boldsymbol{k} \mapsto |\boldsymbol{k}|^{-1}\mathscr{F}(\rho_{\gamma_{\omega_*}} - \mu_{\text{per}})(0, \boldsymbol{k})$ is in $L^2_{\text{loc}}(\mathbb{R}^2)$. In view of this, there exists a positive constant C such that

$$\begin{aligned}
&\sum_{n\in\mathbb{Z}}\int_{\mathbb{R}^2}\frac{|\mathscr{F}(\rho_{\gamma_{\omega_*}} - \mu_{\text{per}})(n, \boldsymbol{k})|^2}{|\boldsymbol{k}|^2 + 4\pi^2n^2}\,d\boldsymbol{k}\\
&\leq \int_{|\boldsymbol{k}|\leq 2\pi}\frac{|\mathscr{F}(\rho_{\gamma_{\omega_*}} - \mu_{\text{per}})(0, \boldsymbol{k})|^2}{|\boldsymbol{k}|^2}\,d\boldsymbol{k}\\
&\qquad + \frac{1}{4\pi^2}\bigg(\int_{|\boldsymbol{k}|>2\pi}|\mathscr{F}(\rho_{\gamma_{\omega_*}} - \mu_{\text{per}})(0, \boldsymbol{k})|^2\,d\boldsymbol{k} + \sum_{n\in\mathbb{Z}\setminus\{0\}}\int_{\mathbb{R}^2}|\mathscr{F}(\rho_{\gamma_{\omega_*}} - \mu_{\text{per}})(n, \boldsymbol{k})|^2\,d\boldsymbol{k}\bigg)\\
&\leq C + \frac{1}{4\pi^2}\int_\Gamma|\rho_{\gamma_{\omega_*}} - \mu_{\text{per}}|^2 < +\infty.
\end{aligned} \tag{4-4}$$

In view of the definition of the Coulomb energy (2-13), we can therefore conclude that

$$D_\Gamma(\rho_{\gamma_{\omega*}} - \mu_{\text{per}}, \rho_{\gamma_{\omega*}} - \mu_{\text{per}}) < +\infty.$$

This concludes the proof that $\gamma_{\omega_*} \in \mathcal{F}_\Gamma$. Hence $\mathcal{F}_\Gamma$ is not empty. As any density ρ_γ associated with $\gamma \in \mathcal{P}_{\text{per},x}$ is integrable, we can conclude that (2-16) holds in view of Remark 2.4.

4B. ***Proof of Theorem 2.6.*** We prove the existence of minimizers and the uniqueness of the density of minimizers for the problem (2-17) by considering a minimizing sequence of the energy functional, and show that there is no loss of compactness. This approach is rather classical for rHF-type models [Catto et al. 2001; Cancès et al. 2008; 2013; 2020]. But in our case we need to be careful as electrons might escape to infinity in the $\boldsymbol{r}$-direction. We show that this is impossible thanks to the Coulomb interactions.

First of all, it is convenient to introduce the following Banach space of operators which are $\mathbb{Z}$-translation-invariant in the x-direction:

$$\mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma) := \{\gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) \mid \tau_k^x \gamma = \gamma \tau_k^x \text{ for all } k \in \mathbb{Z},\ \underline{\mathrm{Tr}}_{L^2(\Gamma)}(|\gamma|) := \mathrm{Tr}_{L^2(\mathbb{R}^3)}(\mathbb{1}_\Gamma |\gamma| \mathbb{1}_\Gamma) < +\infty\},$$

equipped with the norm $\|\gamma\|_{\mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)} := \underline{\mathrm{Tr}}_{L^2(\Gamma)}(|\gamma|)$. In view of (2-1), it is clear that for any operator $\gamma \in \mathcal{F}_\Gamma$, it holds that

$$\mathrm{Tr}_{L^2(\mathbb{R}^3)}(\mathbb{1}_\Gamma \gamma \mathbb{1}_\Gamma) = \int_\Gamma \rho_\gamma = \frac{1}{2\pi}\int_{\Gamma^*}\left(\int_\Gamma \rho_{\gamma_\xi}\right) d\xi = \frac{1}{2\pi}\int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}(\gamma_\xi)\, d\xi.$$

Therefore any operator $0 \le \gamma \le 1$ belongs to $\mathcal{F}_\Gamma$ if and only if $\gamma \in \mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)$, $\sqrt{-\Delta}\gamma\sqrt{-\Delta} \in \mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)$ and $\rho_\gamma - \mu_{\mathrm{per}} \in \mathcal{C}_\Gamma$.

For any $\gamma \in \mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)$, define $\underline{\mathrm{Tr}}_{L^2(\Gamma)}\big(-\frac{1}{2}\Delta\gamma\big) := \frac{1}{2}\mathrm{Tr}_{L^2(\mathbb{R}^3)}(\mathbb{1}_\Gamma\sqrt{-\Delta}\gamma\sqrt{-\Delta}\mathbb{1}_\Gamma)$. Consider a minimizing sequence γ_n of the energy functional

$$\mathcal{E}_{\mathrm{per},x}(\gamma) = \underline{\mathrm{Tr}}_{L^2(\Gamma)}\big(-\tfrac{1}{2}\Delta\gamma\big) + \tfrac{1}{2}D_\Gamma(\rho_\gamma - \mu_{\mathrm{per}}, \rho_\gamma - \mu_{\mathrm{per}})$$

on $\mathcal{F}_\Gamma$. Therefore there exists $C > 0$ such that for all $n \ge 1$

$$0 \le \underline{\mathrm{Tr}}_{L^2(\Gamma)}\big(-\tfrac{1}{2}\Delta\gamma_n\big) \le C, \quad 0 \le D_\Gamma(\rho_{\gamma_n} - \mu_{\mathrm{per}}, \rho_{\gamma_n} - \mu_{\mathrm{per}}) \le C. \tag{4-5}$$

By Lemma 2.5, we also have

$$\underline{\mathrm{Tr}}_{L^2(\Gamma)}(\gamma_n) = \int_\Gamma \mu_{\mathrm{per}}. \tag{4-6}$$

In view of (4-5) and (4-6), we deduce that there exist (up to extraction) γ and T belonging to $\mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)$ such that $\gamma_n \overset{*}{\rightharpoonup} \gamma \in \mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)$ and $\sqrt{-\Delta}\gamma_n\sqrt{-\Delta} \overset{*}{\rightharpoonup} T \in \mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)$. Besides, the bounds (4-5) also imply that $\rho_{\gamma_n} - \mu_{\mathrm{per}} \rightharpoonup \tilde{\rho}_\gamma - \mu_{\mathrm{per}}$ weakly in $\mathcal{C}_\Gamma$, and hence in $\mathcal{D}'(\mathbb{R}^3)$ (see [Cao 2019a, Lemma 3.21, page 111] for details), the space of distributions.

The identification of $T \equiv \sqrt{-\Delta}\gamma\sqrt{-\Delta}$ is straightforward by testing these operators against operators with compact support and by the uniqueness of the weak limit in $\mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)$. Let us prove that $\rho_\gamma \equiv \tilde{\rho}_\gamma \in \mathcal{D}'(\mathbb{R}^3)$: since $\gamma_n \overset{*}{\rightharpoonup} \gamma \in \mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)$ and $\sqrt{-\Delta}\gamma_n\sqrt{-\Delta}$ is bounded in $\mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)$, the density ρ_{γ_n} is therefore bounded in $L^1_{\mathrm{per},x}(\Gamma)$ and $\nabla\sqrt{\rho_{\gamma_n}}$ is bounded in $L^2_{\mathrm{per},x}(\Gamma)$ by the Hoffmann-Ostenhof inequality (2-12). Therefore ρ_{γ_n} converges weakly in $L^r_{\mathrm{per},x}(\Gamma)$ for $1 < r \le 3$, and strongly locally in $L^q(\mathbb{R}^3)$ to ρ_γ for $1 \le q < 3$; hence $\rho_\gamma \equiv \tilde{\rho}_\gamma \in \mathcal{D}'(\mathbb{R}^3)$.

Thus the uniform bound (4-5) together with the weak convergence of $\sqrt{-\Delta}\gamma_n\sqrt{-\Delta} \overset{*}{\rightharpoonup} \sqrt{-\Delta}\gamma\sqrt{-\Delta} \in \mathfrak{S}_{1,\mathrm{per}}^{x}(\Gamma)$ implies

$$\underline{\mathrm{Tr}}_{L^2(\Gamma)}\big(-\tfrac{1}{2}\Delta\gamma\big) \le \liminf_{n\to\infty} \underline{\mathrm{Tr}}_{L^2(\Gamma)}\big(-\tfrac{1}{2}\Delta\gamma_n\big). \tag{4-7}$$

The inequality (4-7) also implies $\sqrt{\rho_\gamma} \in H^1_{\mathrm{per},x}(\Gamma)$ by the Hoffmann-Ostenhof inequality (2-12). Hence $\rho_\gamma \in L^p_{\mathrm{per},x}(\Gamma)$ for $1 \le p \le 3$.

Remark 4.2 (charge conservation in the limit). Since $\rho_\gamma - \mu_{\mathrm{per}}$ is an element in $\mathcal{C}_\Gamma$, this implies that the charge is neutral by Remark 2.4. That is, $\int_\Gamma \rho_\gamma = \int_\Gamma \mu_{\mathrm{per}}$.

As $D_\Gamma(\,\cdot\,,\cdot\,)$ defines an inner product on $\mathcal{C}_\Gamma$, by the weak convergence of $\rho_{\gamma_n} - \mu_{\mathrm{per}}$ to $\rho_\gamma - \mu_{\mathrm{per}}$ in $\mathcal{C}_\Gamma$ we obtain

$$D_\Gamma(\rho_\gamma - \mu_{\mathrm{per}}, \rho_\gamma - \mu_{\mathrm{per}}) \le \liminf_{n\to\infty} D_\Gamma(\rho_{\gamma_n} - \mu_{\mathrm{per}}, \rho_{\gamma_n} - \mu_{\mathrm{per}}). \tag{4-8}$$

In view of (4-7) and (4-8), we conclude that

$$\mathcal{E}_{\mathrm{per},x}(\gamma) \le \liminf_{n\to\infty} \mathcal{E}_{\mathrm{per},x}(\gamma_n),$$

which shows that the state γ is a minimizer of the problem (2-17).

Let us finally prove that all minimizers share the same density: Consider two minimizers $\bar\gamma_1$ and $\bar\gamma_2$. By the convexity of $\mathcal{F}_\Gamma$ it holds that $\frac12(\bar\gamma_1 + \bar\gamma_2) \in \mathcal{F}_\Gamma$. Moreover

$$\mathcal{E}_{\mathrm{per},x}\big(\tfrac12(\bar\gamma_1+\bar\gamma_2)\big) = \tfrac12\mathcal{E}_{\mathrm{per},x}(\bar\gamma_1) + \tfrac12\mathcal{E}_{\mathrm{per},x}(\bar\gamma_2) - \tfrac14 D_\Gamma(\rho_{\bar\gamma_1} - \rho_{\bar\gamma_2}, \rho_{\bar\gamma_1} - \rho_{\bar\gamma_2}),$$

which shows that $D_\Gamma(\rho_{\bar\gamma_1} - \rho_{\bar\gamma_2}, \rho_{\bar\gamma_1} - \rho_{\bar\gamma_2}) \equiv 0$; hence all the minimizers of the problem (2-17) share the same density.

4C. ***Proof of Theorem 2.7.*** We first define a mean-field Hamiltonian associated with the problem (2-17), and then show that the Fermi level is always negative. Moreover, the minimizer of (2-17) is uniquely given by the spectral projector of the mean-field Hamiltonian. In the end we show that the density of the minimizer decays exponentially fast in the $\boldsymbol{r}$-direction.

Properties of the mean-field potential and Hamiltonian. We begin with the definition of a mean-field potential and a mean-field Hamiltonian, and next study the spectrum of the mean-field Hamiltonian. Consider a minimizer γ_{per} of (2-17) with the unique density $\rho_{\gamma_{\mathrm{per}}} \in L^p_{\mathrm{per},x}(\Gamma)$, where $1 \le p \le 3$. Define the mean-field potential

$$V_{\mathrm{per}} := q_{\mathrm{per}} \star_\Gamma G, \quad q_{\mathrm{per}} := \rho_{\gamma_{\mathrm{per}}} - \mu_{\mathrm{per}},$$

which is the solution of Poisson's equation $-\Delta V_{\mathrm{per}} = 4\pi q_{\mathrm{per}}$. Let us prove that V_{per} belongs to $L^p_{\mathrm{per},x}(\Gamma)$ for $2 < p \le +\infty$. As μ_{per} is smooth and has compact support in the $\boldsymbol{r}$-direction and $\rho_{\gamma_{\mathrm{per}}}$ belongs to $L^p_{\mathrm{per},x}(\Gamma)$ for $1 \le p \le 3$; hence $\mathscr{F} q_{\mathrm{per}}(0,\cdot\,)$ belongs to $L^2(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2) \cap C^0(\mathbb{R}^2)$ by classical Fourier theory; see for example [Reed and Simon 1975]. Moreover, as $\int_\Gamma |\boldsymbol{r}|\rho_{\gamma_{\mathrm{per}}}(x,\boldsymbol{r}) < +\infty$ by (2-18),

$$\begin{aligned}\|\partial_{\boldsymbol{k}} \mathscr{F} q_{\mathrm{per}}(0,\boldsymbol{k})\|_{L^\infty(\mathbb{R}^2)} &= \left|\int_\Gamma \mathrm{e}^{-\mathrm{i}\boldsymbol{r}\cdot\boldsymbol{k}} \boldsymbol{r}\cdot q_{\mathrm{per}}(x,\boldsymbol{r})\,dx\,d\boldsymbol{r}\right| \\ &\le \left|\int_\Gamma |\boldsymbol{r}|\,|\rho_{\gamma_{\mathrm{per}}} + \mu_{\mathrm{per}}|(x,\boldsymbol{r})\,dx\,d\boldsymbol{r}\right| < +\infty.\end{aligned} \tag{4-9}$$

This implies that $\mathscr{F}q_{\mathrm{per}}(0,\boldsymbol{k})$ is $C^1(\mathbb{R}^2)$ and bounded. Note also that $\mathscr{F}q_{\mathrm{per}}(0,\mathbf{0})=0$ by the charge neutrality and that $\mathscr{F}q_{\mathrm{per}}(0,\cdot)$ belongs to $L^2(\mathbb{R}^2)\cap L^\infty(\mathbb{R}^2)\cap C^0(\mathbb{R}^2)$; hence, for $1\le\alpha<2$,

$$\begin{aligned}\int_{\mathbb{R}^2}|\mathscr{F}V_{\mathrm{per}}(0,\boldsymbol{k})|^\alpha\,d\boldsymbol{k}&=\int_{\mathbb{R}^2}\frac{|\mathscr{F}q_{\mathrm{per}}(0,\boldsymbol{k})|^\alpha}{|\boldsymbol{k}|^{2\alpha}}\,d\boldsymbol{k}\\&\le\int_{|\boldsymbol{k}|<1}\frac{|\mathscr{F}q_{\mathrm{per}}(0,\boldsymbol{k})|^\alpha}{|\boldsymbol{k}|^{2\alpha}}\,d\boldsymbol{k}+\biggl(\int_{|\boldsymbol{k}|\ge1}|\mathscr{F}q_{\mathrm{per}}(0,\boldsymbol{k})|^{2\alpha}\,d\boldsymbol{k}\biggr)^{1/2}\biggl(\int_{|\boldsymbol{k}|\ge1}\frac{1}{|\boldsymbol{k}|^{4\alpha}}\,d\boldsymbol{k}\biggr)^{1/2}\\&<+\infty.\end{aligned}\tag{4-10}$$

Note that the mixed Fourier transform $\mathscr{F}$ is an isometry from $L^2_{\mathrm{per},x}(\Gamma)$ to $\ell^2(\mathbb{Z},L^2(\mathbb{R}^2))$ by (2-4). On the other hand,

$$\text{for all }\phi\in\ell^1(\mathbb{Z},L^1(\mathbb{R}^2)),\quad\|\mathscr{F}^{-1}\phi\|_{L^\infty(\Gamma)}=\sup_{x\in\Gamma}\biggl|\frac{1}{2\pi}\sum_{n\in\mathbb{Z}}\int_{\mathbb{R}^2}\phi_n(\boldsymbol{k})\,\mathrm{e}^{\mathrm{i}(2\pi nx+\boldsymbol{k}\cdot\boldsymbol{r})}\,d\boldsymbol{k}\biggr|\le\frac{1}{2\pi}\|\phi\|_{\ell^1(\mathbb{Z},L^1(\mathbb{R}^2))}.$$

By the Riesz–Thorin interpolation theorem (see for example [Reed and Simon 1975; Lieb and Loss 2001, Theorem 5.7]), we can deduce a Hausdorff–Young inequality for $\mathscr{F}^{-1}$: for $1\le\alpha\le2$ there exists a constant C_α depending on α such that

$$\text{for all }\phi\in\ell^\alpha(\mathbb{Z},L^\alpha(\mathbb{R}^2)),\quad\|\mathscr{F}^{-1}\phi\|_{L^{\alpha'}_{\mathrm{per},x}(\Gamma)}\le C_\alpha\|\phi\|_{\ell^\alpha(\mathbb{Z},L^\alpha(\mathbb{R}^2))},\tag{4-11}$$

where $\alpha':=\alpha/(\alpha-1)$. Hence in view of (4-10) and (4-11), for $1\le\alpha<2$ and $2<\alpha'=\alpha/(\alpha-1)\le+\infty$ there exist positive constants $C_{\alpha,1}$, $C_{\alpha,2}$ and $C'_{\alpha,2}$ such that

$$\begin{aligned}\|V_{\mathrm{per}}\|^\alpha_{L^{\alpha'}_{\mathrm{per},x}(\Gamma)}&\le C_\alpha^\alpha\|\mathscr{F}V_{\mathrm{per}}\|^\alpha_{\ell^\alpha(\mathbb{Z},L^\alpha(\mathbb{R}^2))}=\frac{C_\alpha^\alpha}{2\pi}\sum_{n\in\mathbb{Z}}\int_{\mathbb{R}^2}|\mathscr{F}V_{\mathrm{per}}(n,\boldsymbol{k})|^\alpha\,d\boldsymbol{k}\\&=\frac{C_\alpha^\alpha}{2\pi}\int_{\mathbb{R}^2}|\mathscr{F}V_{\mathrm{per}}(0,\boldsymbol{k})|^\alpha\,d\boldsymbol{k}+\frac{C_\alpha^\alpha}{2\pi}\sum_{n\ne0}\int_{\mathbb{R}^2}\frac{|\mathscr{F}q_{\mathrm{per}}(n,\boldsymbol{k})|^\alpha}{(4\pi n^2+|\boldsymbol{k}|^2)^\alpha}\,d\boldsymbol{k}\\&\le C_{\alpha,1}+\frac{C_\alpha^\alpha}{2\pi}\biggl(\sum_{n\ne0}\int_{\mathbb{R}^2}|\mathscr{F}q_{\mathrm{per}}(n,\boldsymbol{k})|^{2\alpha}\,d\boldsymbol{k}\biggr)^{1/2}\biggl(\sum_{n\ne0}\int_{\mathbb{R}^2}\frac{1}{(4\pi n^2+|\boldsymbol{k}|^2)^{2\alpha}}\,d\boldsymbol{k}\biggr)^{1/2}\\&\le C_{\alpha,1}+C_{\alpha,2}\|\mathscr{F}q_{\mathrm{per}}\|^\alpha_{\ell^{2\alpha}(\mathbb{Z},L^{2\alpha}(\mathbb{R}^2))}\le C_{\alpha,1}+C'_{\alpha,2}\|q_{\mathrm{per}}\|^\alpha_{L^q_{\mathrm{per},x}(\Gamma)}<+\infty,\end{aligned}\tag{4-12}$$

where the last step we have used estimates similar to (4-11) for $\mathscr{F}$ and the fact that q_{per} belongs to $L^q_{\mathrm{per},x}(\Gamma)$ for $\frac{4}{3}<q:=2\alpha/(2\alpha-1)\le2$. Therefore V_{per} belongs to $L^p_{\mathrm{per},x}(\Gamma)$ for $2<p\le+\infty$. By the elliptic regularity we know that V_{per} belongs to the Sobolev space $W^{2,p}_{\mathrm{per},x}(\Gamma)$ for $2<p\le3$, where $W^{2,p}_{\mathrm{per},x}(\Gamma)$ is the space of functions, together with their gradients and hessians, belonging to $L^p_{\mathrm{per},x}(\Gamma)$. Approximating V_{per} by functions in $\mathcal{D}_{\mathrm{per},x}(\Gamma)$, we also deduce that V_{per} tends to 0 when $|\boldsymbol{r}|$ tends to infinity. The mean-field potential V_{per} defines a $-\Delta$-bounded operator on $L^2(\mathbb{R}^3)$ with relative bound zero; hence by the Kato–Rellich theorem (see for example [Helffer 2013, Theorem 9.10]) we know that $H_{\mathrm{per}}=-\frac{1}{2}\Delta+V_{\mathrm{per}}$ uniquely defines a self-adjoint operator on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$ and form

domain $H^1(\mathbb{R}^3)$. As H_{per} is $\mathbb{Z}$-translation-invariant in the x-direction,

$$H_{\text{per}} = \mathscr{B}^{-1}\left(\int_{\Gamma^*} H_{\text{per},\xi}\,\frac{d\xi}{2\pi}\right)\mathscr{B}, \qquad H_{\text{per},\xi} := -\tfrac{1}{2}\Delta_\xi + V_{\text{per}}.$$

Note that the decomposed Hamiltonian $H_{\text{per},\xi}$ does not have a compact resolvent as Γ is not a bounded domain. It is easy to see that $\sigma(-\Delta_\xi) = \sigma_{\text{ess}}(-\Delta_\xi) = [0, +\infty)$. On the other hand, by the inequality (2-6) we have

$$\|V_{\text{per}}(1-\Delta_\xi)^{-1}\|_{\mathfrak{S}_2} \le \frac{1}{2\pi}\|V_{\text{per}}\|_{L^2_{\text{per},x}(\Gamma)}\left(\sum_{n\in\mathbb{Z}}\int_{\mathbb{R}^2}\frac{1}{((2\pi n+\xi)^2+|\boldsymbol{k}|^2+1)^2}\,d\boldsymbol{k}\right)^{1/2} < +\infty.$$

In particular V_{per} is a compact perturbation of $-\Delta_\xi$, and therefore introduces at most countably many eigenvalues below 0 which are bounded from below by $-\|V_{\text{per}}\|_{L^\infty}$. Denote by $\{\lambda_n(\xi)\}_{1\le n\le N_H}$ these (negative) eigenvalues for $N_H \in \mathbb{N}^*$ (N_H can be finite or infinite). Then for all $\xi \in \Gamma^*$,

$$\sigma_{\text{ess}}(H_{\text{per},\xi}) = \sigma_{\text{ess}}(-\Delta_\xi) = [0, +\infty), \qquad \sigma_{\text{disc}}(H_{\text{per},\xi}) = \bigcup_{1\le n\le N_H}\lambda_n(\xi).$$

In view of the decomposition (2-19), a result of [Reed and Simon 1978, Theorem XIII.85] gives the spectral decomposition

$$\sigma_{\text{ess}}(H_{\text{per}}) \supseteq \bigcup_{\xi\in\Gamma^*}\sigma_{\text{ess}}(H_{\text{per},\xi}) = [0, +\infty), \qquad \sigma_{\text{disc}}(H_{\text{per}}) \subseteq \bigcup_{\xi\in\Gamma^*}\sigma_{\text{disc}}(H_{\text{per},\xi}) = \bigcup_{\xi\in\Gamma^*}\bigcup_{1\le n\le N_H}\lambda_n(\xi).$$

We also obtain from [Reed and Simon 1978, Theorem XIII.85(e)] that

$$\lambda \in \sigma_{\text{disc}}(H_{\text{per}}) \iff \{\xi \in \Gamma^* \mid \lambda \in \sigma_{\text{disc}}(H_{\text{per},\xi})\} \text{ has nontrivial Lebesgue measure.}$$

By the regular perturbation theory of the point spectra [Kato 1966] (see also [Reed and Simon 1978, Section XII.2]) and the approach of Thomas [1973, Lemma 1], we know that the eigenvalues $\lambda_n(\xi)$ below 0 are analytical functions of ξ and cannot be constant, so that $\{\xi \in \Gamma^* \mid \lambda \in \sigma_{\text{disc}}(H_{\text{per},\xi})\}$ has trivial Lebesgue measure, and the essential spectrum of H_{per} below 0 is purely absolutely continuous. As a conclusion,

$$\sigma(H_{\text{per}}) = \sigma_{\text{ess}}(H_{\text{per}}) = \bigcup_{\xi\in\Gamma^*}\sigma(H_{\text{per},\xi}).$$

The Fermi level is always negative. Let us prove that the inequality $N_H = F(0) \ge \int_\Gamma \mu_{\text{per}}$ always holds. The physical meaning of this statement is that the Fermi level of the quasi-1-dimensional system at ground state is always negative when the mean-field potential tends to 0 in the $\boldsymbol{r}$-direction. We prove this by contradiction: Assume that $F(0) < \int_\Gamma \mu_{\text{per}}$. Then we can always construct (infinitely many) states belonging to $\mathcal{F}_\Gamma$ with positive energies arbitrarily close to 0 and they decrease the ground state energy of the problem (2-17).

Let us first define a spectral projector representing all the states of H_{per} below 0: for any $\xi \in \Gamma^*$ and $H_{\text{per},\xi}$ defined in (2-19), define

$$\gamma_{\text{per}}^- := \mathbb{1}_{(-\infty,0]}(H_{\text{per}}), \qquad \gamma_{\text{per},\xi}^- := \mathbb{1}_{(-\infty,0]}(H_{\text{per},\xi}).$$

Therefore,

$$N_H = F(0) = \frac{1}{|\Gamma^*|}\int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}(\gamma^-_{\mathrm{per},\xi})\,d\xi = \int_\Gamma \rho_{\gamma^-_{\mathrm{per}}}. \tag{4-13}$$

The inequality $F(0) < \int_\Gamma \mu_{\mathrm{per}}$ implies

$$Z_{\mathrm{diff}} := \int_\Gamma \mu_{\mathrm{per}} - N_H = \int_\Gamma \mu_{\mathrm{per}} - \int_\Gamma \rho_{\gamma^-_{\mathrm{per}}} \quad \in \mathbb{N}^*. \tag{4-14}$$

The condition (4-14) implies in particular that $N_H < +\infty$, i.e., that there are at most finitely many states below 0. Let us construct states of H_{per} with positive energies. These states belonging to $L^2(\mathbb{R}^3)$ approximate the plane waves of H_{per} traveling in the $\boldsymbol{r}$-direction. For $R > 0$, recall that $\mathfrak{B}_R$ is the ball in $\mathbb{R}^3$ centered at 0. Consider a smooth function $t(x, \boldsymbol{r})$ supported in $\mathfrak{B}_1$, equal to 1 in $\mathfrak{B}_{1/2}$ and such that $\|t\|_{L^2(\mathbb{R}^3)} = 1$. For $n \in \mathbb{N}^*$, let us define

$$\psi_n(x, \boldsymbol{r}) := n^{-3/2} t\left(\frac{(x, \boldsymbol{r}) - (n^2, (n^2, n^2))}{n}\right).$$

It is easy to see that ψ_n belongs to $L^2(\mathbb{R}^3)$, converges weakly to 0 when n tends to infinity and $\|\psi_n\|_{L^2(\mathbb{R}^3)} = 1$. Moreover, as V_{per} tends to 0 in the $\boldsymbol{r}$-direction, for any $\epsilon > 0$ there exists an integer N_ϵ such that $|V_{\mathrm{per}}(\,\cdot\,, (n^2, n^2))| \le \epsilon$ when $n \ge N_\epsilon$. Denote by $\{\psi_{n,\xi}\}_{n\in\mathbb{N}^*, \xi\in\Gamma^*}$ the Bloch decomposition $\mathscr{B}$ in the x-direction (see Section 2A for the definition) of $\{\psi_n\}_{n\in\mathbb{N}^*}$ which belong to $L^2_{\mathrm{per},x}(\Gamma)$. For $n \ge N_\epsilon$, it holds

$$\begin{aligned} \|H_{\mathrm{per}}\psi_n\|^2_{L^2(\mathbb{R}^3)} &= \frac{1}{2\pi}\int_{\Gamma^*} \|H_{\mathrm{per},\xi}\psi_{n,\xi}\|^2_{L^2_{\mathrm{per},x}(\Gamma)}\,d\xi \\ &= \left\| -n^{-7/2}\Delta t\left(\frac{\cdot - (n^2, (n^2, n^2))}{n}\right) + V_{\mathrm{per}}\psi_n \right\|^2_{L^2(\mathbb{R}^3)} \le 2\left(\frac{1}{n^4} + \epsilon^2\right). \end{aligned} \tag{4-15}$$

Note that $\gamma^-_{\mathrm{per},\xi} H_{\mathrm{per},\xi} = \sum_{n=1}^{N_H} P_{\{\lambda_n(\xi)\}}(H_{\mathrm{per},\xi})$ is a compact operator, where $P_{\{\,\cdot\,\}}(H_{\mathrm{per},\xi})$ is the spectral projector of $H_{\mathrm{per},\xi}$. There exists an orthonormal basis $\{e_{n,\xi}\}_{n\ge 1}$ of $L^2_{\mathrm{per},x}(\Gamma)$ with elements in $H^1_{\mathrm{per},x}(\Gamma)$ such that $\gamma^-_{\mathrm{per},\xi} H_{\mathrm{per},\xi} e_{n,\xi} = \lambda_n(\xi) e_{n,\xi}$ for $1 \le n \le N_H$, and $\gamma^-_{\mathrm{per},\xi} H_{\mathrm{per},\xi} e_{n,\xi} \equiv 0$ for $n > N_H$. Let us construct test density matrices composed by all the states of H_{per} with negative energies and some states with positive energies. More precisely, for $N_0 \in \mathbb{N}^*$ to be made precise later, consider a test density matrix

$$\gamma_{N_0} = \mathscr{B}^{-1}\left(\int_{\Gamma^*} \gamma_{N_0,\xi}\,\frac{d\xi}{2\pi}\right)\mathscr{B},$$

where

$$\begin{aligned} \gamma_{N_0,\xi} &:= \gamma^-_{\mathrm{per},\xi} + \sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}} (1 - \gamma^-_{\mathrm{per},\xi})|\psi_{n,\xi}\rangle\langle\psi_{n,\xi}| \\ &= \sum_{n=1}^{+\infty} \gamma^-_{\mathrm{per},\xi}|e_{n,\xi}\rangle\langle e_{n,\xi}| + \sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}} (1 - \gamma^-_{\mathrm{per},\xi})|\psi_{n,\xi}\rangle\langle\psi_{n,\xi}|. \end{aligned}$$

Lemma 4.3. *For any $N_0 \in \mathbb{N}^*$, the state γ_{N_0} belongs to the admissible set $\mathcal{F}_\Gamma$.*

Proof. It is easy to see that $0 \le \gamma_{N_0} \le 1$. Note also that $\operatorname{Ran}(\gamma^-_{\mathrm{per},\xi}) = \operatorname{Span}\{e_{n,\xi}\}_{1\le n\le N_H}$ for all $\xi \in \Gamma^*$. The density of γ_{N_0} can be written as

$$\rho_{\gamma_{N_0}} = \frac{1}{2\pi}\int_{\Gamma^*}\sum_{n=1}^{N_H}|e_{n,\xi}|^2\,d\xi + \frac{1}{2\pi}\sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}}\int_{\Gamma^*}|\psi_{n,\xi}|^2\,d\xi.$$

The density $\rho_{\gamma_{N_0}}$ belongs to $L^p_{\mathrm{per}}(\Gamma)$ for $1 \le p \le 3$ as $\{e_{n,\xi}\}_{n\ge 1}$ and $\{\psi_{n,\xi}\}_{n\ge 1}$ belong to $H^1_{\mathrm{per},x}(\Gamma)$. Additionally, in view of (4-13),

$$\begin{aligned}\int_\Gamma \rho_{\gamma_{N_0}} &= \frac{1}{2\pi}\int_{\Gamma^*}\operatorname{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}(\gamma^-_{\mathrm{per},\xi})\,d\xi + \frac{1}{2\pi}\sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}}\int_\Gamma\int_{\Gamma^*}|\psi_{n,\xi}|^2\,d\xi\\ &= N_H + \sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}}\int_{\mathbb{R}^3}|\psi_n|^2 = N_H + Z_{\mathrm{diff}} = \int_\Gamma \mu_{\mathrm{per}}. \qquad (4\text{-}16)\end{aligned}$$

A simple calculation shows that $|\nabla|\gamma_{N_0,\xi}|\nabla|$ is trace-class on $L^2_{\mathrm{per},x}(\Gamma)$. Hence γ_{N_0} belongs to $\mathcal{P}_{\mathrm{per},x}$. Let us show that $\rho_{\gamma_{N_0}} - \mu_{\mathrm{per}}$ belongs to $\mathcal{C}_\Gamma$. Following calculations similar to the ones leading to (4-4), we only need to prove that $\boldsymbol{k} \mapsto |\boldsymbol{k}|^{-1}\mathscr{F}(\rho_{\gamma_{N_0}} - \mu_{\mathrm{per}})(0,\boldsymbol{k})$ is square-integrable near $\boldsymbol{k} = \boldsymbol{0}$ since $\rho_{\gamma_{N_0}} - \mu_{\mathrm{per}}$ belongs to $L^2_{\mathrm{per},x}(\Gamma)$. Note that

$$\mathscr{F}(\rho_{\gamma_{N_0}} - \mu_{\mathrm{per}})(0,\boldsymbol{0}) = \int_\Gamma \rho_{\gamma_{N_0}} - \mu_{\mathrm{per}} = 0$$

and

$$\begin{aligned}|\partial_{\boldsymbol{k}}\mathscr{F}(\rho_{\gamma_{N_0}} - \mu_{\mathrm{per}})(0,\boldsymbol{0})| &= \left|\int_\Gamma \boldsymbol{r}(\rho_{\gamma_{N_0}} - \mu_{\mathrm{per}})(x,\boldsymbol{r})\,dx\,d\boldsymbol{r}\right|\\ &\le \frac{1}{2\pi}\sum_{n=1}^{N_H}\int_{\Gamma^*}\int_\Gamma |\boldsymbol{r}|\,|e_{n,\xi}|^2(x,\boldsymbol{r})\,dx\,d\boldsymbol{r}\,d\xi\\ &\qquad + \sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}}\int_{\mathbb{R}^3}|\boldsymbol{r}|\,|\psi_n|^2(x,\boldsymbol{r})\,dx\,d\boldsymbol{r} + \int_\Gamma |\boldsymbol{r}|\mu_{\mathrm{per}}(x,\boldsymbol{r})\,dx\,d\boldsymbol{r} < +\infty,\end{aligned}$$

where we have used the fact that the eigenfunctions of $H_{\mathrm{per},\xi}$ associated with negative eigenvalues decay exponentially (see [Hislop and Sigal 1996, Theorem 3.4] and [Combes and Thomas 1973, Theorem 1]) so that $\int_\Gamma |\boldsymbol{r}|\,|e_{n,\xi}|^2(x,\boldsymbol{r})\,dx\,d\boldsymbol{r} < +\infty$ for $1 \le n \le N_H$ and $\xi \in \Gamma^*$, and the fact that $\{\psi_n\}_{N_0+1\le n\le N_0+Z_{\mathrm{diff}}}$ have compact support in the $\boldsymbol{r}$-direction by definition. Therefore $\mathscr{F}(\rho_{\gamma_{N_0}} - \mu_{\mathrm{per}})(0,\boldsymbol{k})$ is C^1 near $\boldsymbol{k} = \boldsymbol{0}$. The conclusion then follows by arguments similar to those leading to (4-4) in Section 4A. □

Lemma 4.3 implies that we can construct many admissible states in $\mathcal{F}_\Gamma$ by varying N_0. Let us show that we can always find N_0 such that γ_{N_0} decreases the ground state energy of (2-17) if $N_H < \int_\Gamma \mu_{\mathrm{per}}$.

Given a minimizer $\bar\gamma$ of (2-17), simple expansion of the energy functional around minimal shows that $\bar\gamma$ also minimizes the functional (see [Cancès et al. 2008])

$$\gamma \mapsto \int_{\Gamma^*}\operatorname{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}(H_{\mathrm{per},\xi}\gamma_\xi)\,d\xi$$

on $\mathcal{F}_\Gamma$. Therefore, given $N_0 \in \mathbb{N}^*$ we have

$$\begin{aligned}0 &\le \int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}(H_{\mathrm{per},\xi}(\gamma_{N_0,\xi} - \bar{\gamma}_\xi))\,d\xi \\ &= \int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}(\gamma^-_{\mathrm{per},\xi} H_{\mathrm{per},\xi}(\gamma_{N_0,\xi} - \bar{\gamma}_\xi))\,d\xi + \int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}((1-\gamma^-_{\mathrm{per},\xi}) H_{\mathrm{per},\xi}(\gamma_{N_0,\xi} - \bar{\gamma}_\xi))\,d\xi \\ &= M + \int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}((1-\gamma^-_{\mathrm{per},\xi}) H_{\mathrm{per},\xi}\gamma_{N_0,\xi})\,d\xi - \int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}((1-\gamma^-_{\mathrm{per},\xi}) H_{\mathrm{per},\xi}\bar{\gamma}_\xi)\,d\xi, \end{aligned} \tag{4-17}$$

where, since $0 \le \bar{\gamma}_\xi \le 1$ and $\{\lambda_n(\,\cdot\,)\}_{1\le n\le N_H} < 0$,

$$\begin{aligned} M &:= \int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}(\gamma^-_{\mathrm{per},\xi} H_{\mathrm{per},\xi}(\gamma_{N_0,\xi} - \bar{\gamma}_\xi))\,d\xi = \int_{\Gamma^*} \sum_{n=1}^{N_H} \lambda_n(\xi)\langle e_{n,\xi}|1-\gamma^-_{\mathrm{per},\xi}\bar{\gamma}_\xi|e_{n,\xi}\rangle\,d\xi \\ &= \int_{\Gamma^*} \sum_{n=1}^{N_H} \lambda_n(\xi)\langle e_{n,\xi}|1-\bar{\gamma}_\xi|e_{n,\xi}\rangle\,d\xi \le 0. \end{aligned} \tag{4-18}$$

In view of (4-15) and by a Cauchy–Schwarz inequality, we deduce that, for $N_0 \ge N_\epsilon$,

$$\begin{aligned} &\int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}((1-\gamma^-_{\mathrm{per},\xi}) H_{\mathrm{per},\xi}\gamma_{N_0,\xi})\,d\xi \\ &\qquad = \int_{\Gamma^*} \sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}} \sum_{m=1}^{+\infty} \langle e_{m,\xi}|H_{\mathrm{per},\xi}|\psi_{n,\xi}\rangle\langle\psi_{n,\xi}|e_{m,\xi}\rangle\,d\xi \\ &\qquad \le \int_{\Gamma^*} \sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}} \biggl(\sum_{m=1}^{+\infty} |\langle\psi_{n,\xi}|e_{m,\xi}\rangle|^2\biggr)^{1/2} \biggl(\sum_{m=1}^{+\infty} |\langle e_{m,\xi}|H_{\mathrm{per},\xi}|\psi_{n,\xi}\rangle|^2\biggr)^{1/2}\,d\xi \\ &\qquad = \int_{\Gamma^*} \sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}} \|\psi_{n,\xi}\|_{L^2_{\mathrm{per},x}(\Gamma)} \|H_{\mathrm{per},\xi}\psi_{n,\xi}\|_{L^2_{\mathrm{per},x}(\Gamma)}\,d\xi \\ &\qquad \le 2\pi \sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}} \biggl(\frac{1}{2\pi}\int_{\Gamma^*} \|\psi_{n,\xi}\|^2_{L^2_{\mathrm{per},x}(\Gamma)}\,d\xi\biggr)^{1/2} \biggl(\frac{1}{2\pi}\int_{\Gamma^*} \|H_{\mathrm{per},\xi}\psi_{n,\xi}\|^2_{L^2_{\mathrm{per},x}(\Gamma)}\,d\xi\biggr)^{1/2} \\ &\qquad \le 2\sqrt{2}\pi \sum_{n=N_0+1}^{N_0+Z_{\mathrm{diff}}} \biggl(\frac{1}{n^4}+\epsilon^2\biggr)^{1/2} \le 2\sqrt{2}\pi\, Z_{\mathrm{diff}} \biggl(\frac{1}{N_0^4}+\epsilon^2\biggr)^{1/2}. \end{aligned} \tag{4-19}$$

Moreover, by the definition of γ^-_{per},

$$\begin{aligned} &\int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}((1-\gamma^-_{\mathrm{per},\xi}) H_{\mathrm{per},\xi}\bar{\gamma}_\xi) \\ &\qquad = \int_{\Gamma^*} \mathrm{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}\bigl(|H_{\mathrm{per},\xi}|^{1/2}(1-\gamma^-_{\mathrm{per},\xi})\bar{\gamma}_\xi(1-\gamma^-_{\mathrm{per},\xi})|H_{\mathrm{per},\xi}|^{1/2}\bigr) \ge 0. \end{aligned} \tag{4-20}$$

We distinguish in the inequality (4-18) the cases $M \equiv 0$ or $M < 0$. When $M \equiv 0$, the inequality (4-18) implies that $\bar{\gamma}_\xi\gamma^-_{\mathrm{per},\xi} = \gamma^-_{\mathrm{per},\xi}$ for almost all $\xi \in \Gamma^*$. In view of the inequalities (4-19) and (4-20), the inequality (4-17) implies

$$\text{for all } N_0 \geq N_\epsilon, \quad 0 \leq \int_{\Gamma^*} \operatorname{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}((1-\gamma^-_{\mathrm{per},\xi}) H_{\mathrm{per},\xi}\bar{\gamma}_\xi)\, d\xi \leq 2\sqrt{2}\pi\, Z_{\mathrm{diff}}\left(\frac{1}{N_0^4}+\epsilon^2\right)^{1/2}. \tag{4-21}$$

By letting N_0 tend to infinity, it is easy to deduce that $(1-\gamma^-_{\mathrm{per},\xi})\bar{\gamma}_\xi = 0$ for almost all $\xi \in \Gamma^*$. Together with the fact that $\bar{\gamma}_\xi \gamma^-_{\mathrm{per},\xi} = \gamma^-_{\mathrm{per},\xi}$ we deduce that $\bar{\gamma}_\xi \equiv \gamma^-_{\mathrm{per},\xi}$ for almost all $\xi \in \Gamma^*$. In view of (4-13) and (4-14), by the charge neutrality we obtain that

$$Z_{\mathrm{diff}} = \int_\Gamma \mu_{\mathrm{per}} - N_H = \int_\Gamma \rho_{\bar{\gamma}} - N_H = \frac{1}{|\Gamma^*|}\int_{\Gamma^*} \operatorname{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}(\gamma^-_{\mathrm{per},\xi})\, d\xi - N_H \equiv 0.$$

Hence $\int_\Gamma \mu_{\mathrm{per}} = N_H = F(0)$. This also implies that the minimizer of the problem (2-17) is equal to γ^-_{per} when $N_H = F(0) = \int_\Gamma \mu_{\mathrm{per}}$ and $Z_{\mathrm{diff}} \equiv 0$. When $M < 0$ and $Z_{\mathrm{diff}} \neq 0$, we can always find $\epsilon > 0$ and $N_0 \geq N_\epsilon$ such that

$$2\sqrt{2}\pi\, Z_{\mathrm{diff}}\left(\frac{1}{N_0^4}+\epsilon^2\right)^{1/2} \leq -\frac{M}{2}.$$

In view of the inequalities (4-19) and (4-20), the inequality (4-17) implies

$$\text{for all } N_0 \geq N_\epsilon, \quad 0 \leq \int_{\Gamma^*} \operatorname{Tr}_{L^2_{\mathrm{per},x}(\Gamma)}((1-\gamma^-_{\mathrm{per},\xi}) H_{\mathrm{per},\xi}\bar{\gamma}_\xi)\, d\xi \leq \frac{M}{2} < 0,$$

which leads to a contradiction. We can finally conclude that $F(0) \geq \int_\Gamma \mu_{\mathrm{per}}$, so that the Fermi level of the quasi-1-dimensional system is always nonpositive. In the following paragraph we show that the Fermi level can be chosen to be strictly negative.

Form of the minimizer and decay of the density of minimizers. We have already shown that if $N_H = F(0) \equiv \int_\Gamma \mu_{\mathrm{per}}$ then the 1-dimensional system has a unique minimizer which is equal to γ^-_{per}. This also implies that for almost all $\xi \in \Gamma^*$, the operator $H_{\mathrm{per},\xi}$ has N_H strictly negative eigenvalues below 0; therefore we can always choose the Fermi level $\epsilon_F \in (\max_{\xi\in\Gamma^*} \lambda_{N_H}(\xi), 0)$. If $F(0) > \int_\Gamma \mu_{\mathrm{per}}$, it is clear that there exists $\epsilon_F < 0$ such that $F(\epsilon_F) = \int_\Gamma \mu_{\mathrm{per}}$ as $F(\kappa)$ is a nondecreasing function on $(-\infty, 0]$ with range in $[0, F(0)]$. The form of the minimizer and the uniqueness is a direct adaptation of [Cancès et al. 2008, Theorem 1] by using a spectral projector decomposition similar to (A.2) of that theorem; that is, the unique minimizer can be written as

$$\gamma_{\mathrm{per}} = \mathbb{1}_{(-\infty,\epsilon_F]}(H_{\mathrm{per}}) = \mathscr{B}^{-1}\left(\int_{\Gamma^*} \gamma_{\mathrm{per},\xi}\frac{d\xi}{2\pi}\right)\mathscr{B} = \mathscr{B}^{-1}\left(\int_{\Gamma^*}\sum_{n=1}^{N_H} \mathbb{1}(\lambda_n(\xi) \leq \epsilon_F)|e_{n,\xi}\rangle\langle e_{n,\xi}|\right)\mathscr{B}, \tag{4-22}$$

where $\gamma_{\mathrm{per},\xi} := \mathbb{1}_{(-\infty,\epsilon_F]}(H_{\mathrm{per},\xi})$. The Fermi level $\epsilon_F < 0$ can be considered as the Lagrange multiplier associated with the charge neutrality condition

$$F(\epsilon_F) = \int_\Gamma \rho_{\gamma_{\mathrm{per}}} = \int_\Gamma \mu_{\mathrm{per}}.$$

Once the unique minimizer is shown to be a spectral projector, we can use the exponential decay property of the eigenfunctions of $H_{\mathrm{per},\xi}$ in the $\boldsymbol{r}$-direction via the Combes–Thomas estimate [1973, Theorem 1]:

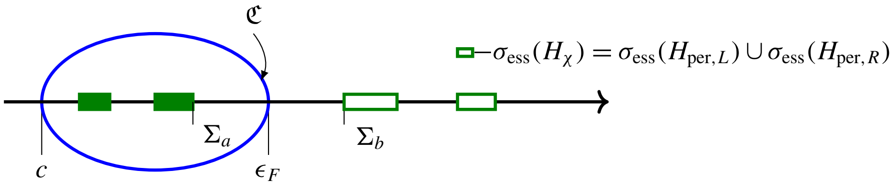

$\mathfrak{C}$
$\sigma_{\text{ess}}(H_\chi) = \sigma_{\text{ess}}(H_{\text{per},L}) \cup \sigma_{\text{ess}}(H_{\text{per},R})$
Σ_a
Σ_b
c
ϵ_F

Figure 6. The essential spectrum of H_χ, $H_{\text{per},L}$ and $H_{\text{per},R}$, and the contour $\mathfrak{C}$.

for almost all $\xi \in \Gamma^*$, there exist positive constants $C(\xi)$ and $\alpha(\xi)$ such that

$$\text{for all } (x, \boldsymbol{r}) \in \Gamma, \text{ for all } 1 \leq n \leq N_H, \quad |e_{n,\xi}(x, \boldsymbol{r})| \leq C(\xi)\mathrm{e}^{-\alpha(\xi)|\boldsymbol{r}|}.$$

On the other hand, the fact that $\int_\Gamma \mu_{\text{per}} = F(\epsilon_F) < +\infty$ implies that there exist only finitely many states of $H_{\text{per},\xi}$ below ϵ_F for all $\xi \in \Gamma^*$. Therefore there exist positive constants C_{ϵ_F} and α_{ϵ_F} such that

$$0 \leq \rho_{\gamma_{\text{per}}}(x, \boldsymbol{r}) \leq \frac{1}{2\pi} \int_{\Gamma^*} \sum_{n=1}^{N_H} \mathbb{1}(\lambda_n(\xi) \leq \epsilon_F) C^2(\xi) \mathrm{e}^{-2\alpha(\xi)|\boldsymbol{r}|}\, d\xi \leq C_{\epsilon_F} \mathrm{e}^{-\alpha_{\epsilon_F}|\boldsymbol{r}|}.$$

Note that this exponential decay property coincides with Assumption 1 that $\int_\Gamma |\boldsymbol{r}| \rho_{\gamma_{\text{per}}}(x, \boldsymbol{r})\, dx\, d\boldsymbol{r} < +\infty$.

4D. ***Proof of Proposition 3.2.*** In view of Proposition 3.1, consider a contour $\mathfrak{C}$ in the complex plane enclosing the spectrum of H_χ below the Fermi level ϵ_F without intersecting it, crossing the real axis at $c < \inf\{-\|V_{\text{per},L}\|_{L^\infty}, -\|V_{\text{per},R}\|_{L^\infty}\}$ (See Figure 6). This is possible even if ϵ_F is an eigenvalue: one can always slightly move the curve $\mathfrak{C}$ below ϵ_F in order bypass ϵ_F but still enclose all the spectrum of H_χ below ϵ_F. Let us introduce the following estimates, which are useful to characterize the decay property of densities. Since V_χ belongs to $L^\infty(\mathbb{R}^3)$, the following lemma is a direct adaption of [Cancès et al. 2008, Lemma 1]:

Lemma 4.4. *Under Assumption 2, there exist two positive constants c_1, c_2 such that,*

$$\textit{for all } \zeta \in \mathfrak{C}, \quad c_1(1 - \Delta) \leq |H_\chi - \zeta| \leq c_2(1 - \Delta)$$

as operators on $L^2(\mathbb{R}^3)$. In particular

$$\| |H_\chi - \zeta|^{1/2} (1 - \Delta)^{-1/2} \| \leq \sqrt{c_2}, \quad \| |H_\chi - \zeta|^{-1/2} (1 - \Delta)^{1/2} \| \leq \frac{1}{\sqrt{c_1}}.$$

Moreover, $(H_\chi - \zeta)(1 - \Delta)^{-1}$ and its inverse are bounded operators.

Let us turn to the proof of Proposition 3.2. First of all let us show that γ_χ is locally trace class. Consider $\varrho \in C_c^\infty(\mathbb{R}^3)$. Note that γ_χ is a spectral projector. In view of Lemma 4.4, by Cauchy's resolvent formula and the Kato–Seiler–Simon inequality (4-1), there exists a positive constant C_χ such that

$$\|\varrho \gamma_\chi \varrho\|_{\mathfrak{S}_1} = \|\varrho \gamma_\chi \gamma_\chi \varrho\|_{\mathfrak{S}_1} = \|\varrho \gamma_\chi\|_{\mathfrak{S}_2}^2 = \left\| \varrho \oint_{\mathfrak{C}} \frac{1}{2i\pi} \frac{1}{\zeta - H_\chi}\, d\zeta \right\|_{\mathfrak{S}_2}^2 \leq C_\chi \left\| \varrho \frac{1}{1 - \Delta} \right\|_{\mathfrak{S}_2}^2 \leq \frac{C_\chi}{4\pi} \|\varrho\|_{L^2(\mathbb{R}^3)}^2.$$

This implies that γ_χ is locally trace class so that its density ρ_χ is well-defined in $L^1_{\mathrm{loc}}(\mathbb{R}^3)$. Let us prove that $\chi^2\rho_{\mathrm{per},L} + (1-\chi^2)\rho_{\mathrm{per},R} - \rho_\chi$ belongs to $L^p(\mathbb{R}^3)$ for $1 < p \le 2$. It is difficult to directly compare the difference of $\chi^2\rho_{\mathrm{per},L} + (1-\chi^2)\rho_{\mathrm{per},R}$ and ρ_χ. We construct to this end a density operator γ_d whose density ρ_d is equal to $\chi^2\rho_{\mathrm{per},L} + (1-\chi^2)\rho_{\mathrm{per},R} - \rho_\chi$:

$$\gamma_d := \gamma_{d,1} + \gamma_{d,2}, \quad \gamma_{d,1} := \chi(\gamma_{\mathrm{per},L} - \gamma_\chi)\chi, \quad \gamma_{d,2} := \sqrt{1-\chi^2}(\gamma_{\mathrm{per},R} - \gamma_\chi)\sqrt{1-\chi^2}. \tag{4-23}$$

Note that if $\gamma_d \in \mathfrak{S}_1$, then $\mathrm{Tr}_{L^2(\mathbb{R}^3)}(\gamma_d) = \chi^2\rho_{\mathrm{per},L} + (1-\chi^2)\rho_{\mathrm{per},R} - \rho_\chi$.

The density ρ_d is in $L^p(\mathbb{R}^3)$ for $1 < p \le 2$. The proof that $\rho_d \in L^p(\mathbb{R}^3)$ relies on duality arguments: setting $q = p/(p-1) \in [2, +\infty)$, we prove that for any $W \in L^q(\mathbb{R}^3)$ there exists some $K_q > 0$ such that $|\mathrm{Tr}_{L^2(\mathbb{R}^3)}(\gamma_d W)| \le K_q \|W\|_{L^q}$. By Cauchy's formula we have

$$\begin{aligned} \gamma_{d,1} &= \frac{1}{2\pi \mathrm{i}} \oint_{\mathfrak{C}} \chi\left(\frac{1}{z - H_{\mathrm{per},L}} - \frac{1}{\zeta - H_\chi}\right)\chi\, d\zeta, \\ \gamma_{d,2} &= \frac{1}{2\pi \mathrm{i}} \oint_{\mathfrak{C}} \sqrt{1-\chi^2}\left(\frac{1}{\zeta - H_{\mathrm{per},R}} - \frac{1}{\zeta - H_\chi}\right)\sqrt{1-\chi^2}\, d\zeta. \end{aligned} \tag{4-24}$$

Let us prove that there exists $K_q^1 > 0$ such that

$$|\mathrm{Tr}_{L^2(\mathbb{R}^3)}(\gamma_{d,1} W)| \le K_q^1 \|W\|_{L^q}.$$

It is easily shown that a similar inequality holds for $\gamma_{d,2}$. Define $V_d := (1-\chi^2)(V_{\mathrm{per},L} - V_{\mathrm{per},R}) \in L^\infty(\mathbb{R}^3)$. Note that the function $V_d\chi$ has compact support in the x-direction and that $V_d\chi$ belongs to $L^r(\mathbb{R}^3)$ for $1 < r \le +\infty$ by Theorem 2.7. For any $\zeta \in \mathfrak{C}$, the integrand of $\gamma_{d,1}$ can be written as

$$D(\zeta) := \chi\left(\frac{1}{\zeta - H_{\mathrm{per},L}} - \frac{1}{\zeta - H_\chi}\right)\chi = \chi \frac{1}{\zeta - H_{\mathrm{per},L}} V_d \frac{1}{\zeta - H_\chi}\chi.$$

Note that since χ is translation-invariant in the $\boldsymbol{r}$-direction, it is not in any L^p space in $\mathbb{R}^3$, which prevents us from using the standard techniques such as calculating the commutator $[-\Delta, \chi]$ to give Schatten class estimates on $\gamma_{d,1}$. By writing $1 = \gamma_{\mathrm{per},L} + \gamma_{\mathrm{per},L}^\perp$ and $1 = \gamma_\chi + \gamma_\chi^\perp$, the following decomposition holds:

$$D(\zeta) = \chi \frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}} V_d \frac{1}{\zeta - H_\chi}\chi + \chi \frac{\gamma_{\mathrm{per},L}^\perp}{\zeta - H_{\mathrm{per},L}} V_d \frac{\gamma_\chi}{\zeta - H_\chi}\chi + \chi \frac{\gamma_{\mathrm{per},L}^\perp}{\zeta - H_{\mathrm{per},L}} V_d \frac{\gamma_\chi^\perp}{\zeta - H_\chi}\chi. \tag{4-25}$$

By the residue theorem,

$$\int_{\mathfrak{C}} \chi \frac{\gamma_{\mathrm{per},L}^\perp}{\zeta - H_{\mathrm{per},L}} V_d \frac{\gamma_\chi^\perp}{\zeta - H_\chi}\chi\, d\zeta \equiv 0. \tag{4-26}$$

To estimate other terms in (4-25) we rely on the following lemma:

Lemma 4.5. *For any $1 < p \le 2$, there exist positive constants $d_{p,1}$ and $d_{p,2}$, such that,*

$$\text{for all } \zeta \in \mathfrak{C}, \quad \left\|\chi \frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}} V_d\right\|_{\mathfrak{S}_p} \le d_{p,1}\|V_d\chi\|_{L^p(\mathbb{R}^3)}, \quad \left\|V_d \frac{\gamma_\chi}{\zeta - H_\chi}\chi\right\|_{\mathfrak{S}_p} \le d_{p,2}\|V_d\chi\|_{L^p(\mathbb{R}^3)}. \tag{4-27}$$

The proof of Lemma 4.5 can be read in Section A4 of the Appendix. Consider $W \in L^q(\mathbb{R}^3)$ for $q = p/(p-1) \in [2, +\infty)$. In view of (4-25), (4-26) and (4-27), by manipulations similar to the ones used in the proof of Lemma 4.5, and Hölder's inequality for Schatten class operators (see for example [Reed and Simon 1975, Proposition 5]), we obtain

$$\begin{aligned}\|\gamma_{d,1} W\|_{\mathfrak{S}_1} &= \frac{1}{2\pi}\left\| \oint_{\mathfrak{C}} D(\zeta)\, d\zeta\, W \right\|_{\mathfrak{S}_1} \\ &= \left\| \oint_{\mathfrak{C}} \chi \frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}} V_d \frac{1}{\zeta - H_\chi} \chi W + \chi \frac{\gamma^{\perp}_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}} V_d \frac{\gamma_\chi}{\zeta - H_\chi} \chi W \, d\zeta \right\|_{\mathfrak{S}_1} \\ &\le \left\| \oint_{\mathfrak{C}} \left(\chi \frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}} V_d \right) \frac{1}{\zeta - H_\chi} (1-\Delta) \left(\frac{1}{1-\Delta} \chi W \right) d\zeta \right\|_{\mathfrak{S}_1} \\ &\qquad\qquad + \left\| \oint_{\mathfrak{C}} \left(W \chi \frac{1}{1-\Delta} \right) (1-\Delta) \frac{\gamma^{\perp}_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}} \left(V_d \frac{\gamma_\chi}{\zeta - H_\chi} \chi \right) d\zeta \right\|_{\mathfrak{S}_1} \\ &\le C \|V_d \chi\|_{L^p(\mathbb{R}^3)} \left\| \frac{1}{1-\Delta} \chi W \right\|_{\mathfrak{S}_q} \le K^1_q \|W\|_{L^q(\mathbb{R}^3)}, \end{aligned} \tag{4-28}$$

where we have used the Kato–Seiler–Simon inequality (4-1) as well as the fact that $\|\chi\|_{L^\infty} = 1$. Similar estimates hold for $\gamma_{d,2}$. We therefore can conclude that $\rho_d = \chi^2 \rho_{\mathrm{per},L} + (1-\chi^2)\rho_{\mathrm{per},R} - \rho_\chi$ belongs to $L^p(\mathbb{R}^3)$ for $1 < p \le 2$.

Decay rate in the x-direction. Let us show that the density difference $\chi^2 \rho_{\mathrm{per},L} + (1-\chi^2)\rho_{\mathrm{per},R} - \rho_\chi$ decays exponentially fast in the x-direction. Note that there exists $N_L \in \mathbb{N}$ such that $N_L - 1 < a_L/2 \le N_L$. Setting $\mathbb{D}_{a_L} := [-a_L/2, +\infty) \times \mathbb{R}^2$, we prove the exponential decay when $\mathrm{supp}(w_\alpha) \subset \mathbb{R}^3 \backslash \mathbb{D}_{a_L}$. Let

$$\alpha := (\alpha_x, 0, 0) \in (\mathbb{R}, 0, 0), \qquad \beta = (\beta_x, \beta_y, \beta_z) \in \mathbb{Z}^3, \qquad \beta_x \ge -N_L.$$

We have

$$\mathbb{1}_{\mathbb{D}_{a_L}} \left(\sum_{\beta_x \ge -N_L}^{+\infty} \sum_{\beta_y, \beta_z \in \mathbb{Z}} w_\beta \right) = \mathbb{1}_{\mathbb{D}_{a_L}}, \qquad \mathbb{1}_{\mathbb{D}_{a_L}} V_d = V_d, \qquad \alpha_x < -\frac{a_L}{2} < -N_L + 1 \le \beta_x + 1.$$

The above relations imply, together with (4-27), the Combes–Thomas estimate (see for example [Klopp 1995; Combes and Thomas 1973; Germinet and Klein 2003]) and arguments similar to ones used in (4-28), that there exist positive constants C_1 and t_1 such that, for $1 < p \le 2$ and $q = p/(p-1) \ge 2$,

$$\begin{aligned}\|w_\alpha \gamma_d w_\alpha\|_{\mathfrak{S}_1} &= \|w_\alpha \gamma_{d,1} w_\alpha\|_{\mathfrak{S}_1} \\ &= \left\| \frac{1}{2\pi \mathrm{i}} \oint_{\mathfrak{C}} \left(w_\alpha \chi \frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}} V_d \frac{1}{\zeta - H_\chi} \chi\, w_\alpha + w_\alpha \chi \frac{\gamma^{\perp}_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}} V_d \frac{\gamma_\chi}{\zeta - H_\chi} \chi\, w_\alpha \right) d\zeta \right\|_{\mathfrak{S}_1} \\ &\le \left\| \frac{1}{2\pi \mathrm{i}} \oint_{\mathfrak{C}} \left(w_\alpha \chi \frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}} V_d \right) \left(\mathbb{1}_{\mathbb{D}_{a_L}} \frac{1}{\zeta - H_\chi} \chi\, w_\alpha \right) d\zeta \right\|_{\mathfrak{S}_1} \\ &\qquad\qquad + \left\| \frac{1}{2\pi \mathrm{i}} \oint_{\mathfrak{C}} (w_\alpha \chi \frac{\gamma^{\perp}_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}} \mathbb{1}_{\mathbb{D}_{a_L}}) \left(V_d \frac{\gamma_\chi}{\zeta - H_\chi} \chi\, w_\alpha \right) d\zeta \right\|_{\mathfrak{S}_1} \\ &\le K \sum_{\beta_x \ge -N_L}^{+\infty} \sum_{\beta_y, \beta_z \in \mathbb{Z}}^{+\infty} \mathrm{e}^{-t_1(\beta_x - \alpha_x)} \mathrm{e}^{-t_1|\beta_y|} \mathrm{e}^{-t_1|\beta_z|} \le C_1 \mathrm{e}^{-t_1|\alpha_x|}. \end{aligned}$$

The last step relies on the uniform distance of $\zeta \in \mathfrak{C}$ to $\sigma(H_\chi)$ and $\sigma(H_{\mathrm{per},L})$. Similar estimates hold when the support of w_α is in $[a_R/2, +\infty) \times \mathbb{R}^2$. There exist therefore positive constants C and t such that $\|w_\alpha \gamma_d w_\alpha\|_{\mathfrak{S}_1} = \int_{\mathbb{R}^3} |w_\alpha \rho_d w_\alpha| \le C\mathrm{e}^{-t|\alpha|}$, which concludes the proof.

4E. ***Proof of Lemma 3.3.*** From the last item of the Theorem 2.7 we know that $V_{\mathrm{per},L} \in L^p_{\mathrm{per},x}(\Gamma_L)$ (resp. $V_{\mathrm{per},R} \in L^p_{\mathrm{per},x}(\Gamma_R)$) for $1 < p \le +\infty$. Note also that $\partial_x^2(\chi^2)$, $\partial_x(\chi^2)$ are uniformly bounded and have support in $[-a_L/2, a_R/2] \times \mathbb{R}^2$. It therefore suffices to obtain the L^p-estimates on $\partial_x V_{\mathrm{per},L}$ and $\partial_x V_{\mathrm{per},R}$. We treat $\partial_x V_{\mathrm{per},L}$, the L^p-estimates of $\partial_x V_{\mathrm{per},R}$ following similar arguments. First of all in view of the form of the minimizer (4-22), by the Cauchy–Schwarz inequality

$$\begin{aligned}\partial_x \rho_{\mathrm{per},L} &= \partial_x \left(\frac{1}{2\pi} \int_{\Gamma_L^*} \sum_{n\ge1} \mathbb{1}(\lambda_n(\xi) \le \epsilon_L) |e_n(\xi,\cdot)|^2 \, d\xi \right) \\ &\le \frac{1}{\pi} \int_{\Gamma_L^*} \left(\sum_{n\ge1} \mathbb{1}(\lambda_n(\xi) \le \epsilon_L) |\partial_x |e_n|(\xi,\cdot)|^2 \right)^{1/2} \left(\sum_{n\ge1} \mathbb{1}(\lambda_n(\xi) \le \epsilon_L) |e_n(\xi,\cdot)|^2 \right)^{1/2} d\xi \\ &\le \frac{1}{\pi} \sqrt{K_{\xi,L}} \sqrt{\rho_{\mathrm{per},L}},\end{aligned}$$

where $K_{\xi,L}(\boldsymbol{x}) := \int_{\Gamma_L^*} \sum_{n\ge1} \mathbb{1}(\lambda_n(\xi) \le \epsilon_L) |\partial_x e_n(\xi,\boldsymbol{x})\, d\xi|^2$. We also have used the fact that $|\nabla |f|| \le |\nabla f|$ for any complex-valued function f. In view of the potential decomposition (A-5), the term $T(\boldsymbol{r})$ does not contribute to the x-directional derivative; hence

$$|\partial_x V_{\mathrm{per},L}| = |(\partial_x(\rho_{\mathrm{per},L} - \mu_{\mathrm{per},L})) \star \widetilde{G}_{a_L}| \le \left(\frac{1}{2\pi} \sqrt{K_{\xi,L}} \sqrt{\rho_{\mathrm{per},L}} + |\partial_x \mu_{\mathrm{per},L}| \right) \star |\widetilde{G}_{a_L}|.$$

On the other hand, finite kinetic energy condition (2-10) implies that $K_{\xi,L} \in L^1_{\mathrm{per},x}(\Gamma_L)$. Moreover, $\sqrt{\rho_{\mathrm{per},L}}$ belongs to $H^1_{\mathrm{per},x}(\Gamma_L)$ and hence to $L^s_{\mathrm{per},x}(\Gamma_L)$ for $2 \le s \le 6$. Therefore, by Hölder's inequality, for $p, m \ge 1$,

$$\int_{\Gamma_L} (K_{\xi,L} \rho_{\mathrm{per},L})^{p/2} \le \left(\int_{\Gamma_L} K_{\xi,L}^{pm/2} \right)^{1/m} \left(\int_{\Gamma_L} \rho_{\mathrm{per},L}^{pm/(2(m-1))} \right)^{(m-1)/m},$$

with the conditions $pm = 2$ and $2 \le pm/(m-1) \le 6$. This is the case for $\frac{4}{3} \le m \le 2$ and $1 \le p \le \frac{3}{2}$ so that $(K_{\xi,L}\rho_{\mathrm{per},L})^{1/2}$ belongs to $L^p_{\mathrm{per},x}(\Gamma_L)$ for $1 \le p \le \frac{3}{2}$. As $\partial_x \mu_{\mathrm{per},L}$ is in $L^p_{\mathrm{per},x}(\Gamma_L)$ for any $1 \le p \le +\infty$ and $\widetilde{G}_{a_L} \in L^q_{\mathrm{per},x}(\Gamma_L)$ for $1 \le q < 2$ by Lemma 2.3, we obtain by Young's convolution inequality that $\partial_x V_{\mathrm{per},L} \in L^s_{\mathrm{per},x}(\Gamma_L)$ for $1 \le s < 6$. This allows us to conclude the lemma.

4F. ***Proof of Theorem 3.6.*** We prove this theorem by taking two arbitrary cut-off functions χ_1, χ_2 belonging to $\mathcal{X}$ and proving that $\rho_{\chi_1} + \rho_{Q_{\chi_1}} = \rho_{\chi_2} + \rho_{Q_{\chi_2}}$. For $i = 1, 2$, consider the reference states associated with the Hamiltonian H_{χ_i}. Denote by γ_{χ_i} the spectral projector of H_{χ_i} below ϵ_F and by Q_{χ_i} the solutions of (3-14) associated with χ_i. Consider a test state

$$\widetilde{Q} := \gamma_{\chi_1} + Q_{\chi_1} - \gamma_{\chi_2}. \tag{4-29}$$

We show that $\widetilde{Q}$ is a minimizer of the problem (3-12) associated with the cut-off function χ_2, so that $\rho_{\widetilde{Q}} \equiv \rho_{Q_{\chi_2}}$ by the uniqueness of the density of the minimizer provided by Proposition 3.5. Note that Assumption 2 and Proposition 3.1 guarantee that there is a common spectral gap for H_{χ_i} and $\sigma_{\mathrm{ess}}(H_{\chi 1}) = \sigma_{\mathrm{ess}}(H_{\chi 2})$. We first show that the test state $\widetilde{Q}$ belongs to the convex set

$$\mathcal{K}_{\chi 2} := \{Q \in \mathcal{Q}_{\chi 2} \mid -\gamma_{\chi 2} \le Q \le 1 - \gamma_{\chi 2}\},$$

and hence is an admissible state for the minimization problem (3-12) associated with χ_2. We next show that $\widetilde{Q}$ is a minimizer.

The test state $\widetilde{Q}$ belongs to $\mathcal{K}_{\chi 2}$. We begin by proving that $\widetilde{Q}$ is in $\mathcal{Q}_{\chi 2}$. Let us prove that $\widetilde{Q}$ is γ_{χ_2}-trace class. The following lemma will be useful.

Lemma 4.6. *The difference of the spectral projectors $\gamma_{\chi_1} - \gamma_{\chi_2}$ belongs to $\mathfrak{S}_1^{\gamma_{\chi 2}}$. Moreover,*

$$|\nabla|(\gamma_{\chi 1} - \gamma_{\chi 2}) \in \mathfrak{S}_2, \quad (\gamma_{\chi 1} - \gamma_{\chi 2})|\nabla| \in \mathfrak{S}_2. \tag{4-30}$$

Proof. By Cauchy's resolvent formula and the Kato–Seiler–Simon inequality (4-1),

$$\begin{aligned}\|\gamma_{\chi 1} - \gamma_{\chi 2}\|_{\mathfrak{S}_2} &= \left\| \frac{1}{2i\pi} \oint_{\mathfrak{C}} (\zeta - H_{\chi 1})^{-1} (\chi_1^2 - \chi_2^2)(V_{\mathrm{per},L} - V_{\mathrm{per},R})(\zeta - H_{\chi 2})^{-1}\, d\zeta \right\|_{\mathfrak{S}_2} \\ &\le C \|(1-\Delta)^{-1}(\chi_1^2 - \chi_2^2)(V_{\mathrm{per},L} - V_{\mathrm{per},R})\|_{\mathfrak{S}_2} \\ &\le \frac{C}{2\sqrt{\pi}} \|(\chi_1^2 - \chi_2^2)(V_{\mathrm{per},L} - V_{\mathrm{per},R})\|_{L^2} < +\infty.\end{aligned} \tag{4-31}$$

The results of Lemma 4.4 imply that $|\nabla|(\zeta - H_{\chi_i})^{-1}$ is uniformly bounded with respect to $\zeta \in \mathfrak{C}$. By calculations similar to (4-31),

$$\| |\nabla|(\gamma_{\chi 1} - \gamma_{\chi 2})\|_{\mathfrak{S}_2} \le c_1 \|(\chi_1^2 - \chi_2^2)(V_{\mathrm{per},L} - V_{\mathrm{per},R})(1-\Delta)^{-1}\|_{\mathfrak{S}_2} < +\infty. \tag{4-32}$$

Hence $(\gamma_{\chi 1} - \gamma_{\chi 2})|\nabla|$ also belongs to $\mathfrak{S}_2$ since it is the adjoint of $|\nabla|(\gamma_{\chi_1} - \gamma_{\chi 2})$. On the other hand, as $\gamma_{\chi 1}$ is a bounded operator, in view of (4-31) and by writing $\gamma_{\chi_1} - \gamma_{\chi 2} = \gamma_{\chi 2}^{\perp} - \gamma_{\chi 1}^{\perp}$ and using the fact that $\gamma_{\chi_i} + \gamma_{\chi_i}^{\perp} = 1$,

$$\begin{aligned}\gamma_{\chi 2}^{\perp}(\gamma_{\chi 1} - \gamma_{\chi 2})\gamma_{\chi 2}^{\perp} &= \gamma_{\chi 2}^{\perp}\gamma_{\chi 1}\gamma_{\chi 2}^{\perp} = (\gamma_{\chi 1} - \gamma_{\chi 2})\gamma_{\chi 1}(\gamma_{\chi 1} - \gamma_{\chi 2}) \in \mathfrak{S}_1, \\ \gamma_{\chi 2}(\gamma_{\chi 1} - \gamma_{\chi 2})\gamma_{\chi 2} &= -\gamma_{\chi 2}\gamma_{\chi 1}^{\perp}\gamma_{\chi 2} = -(\gamma_{\chi 2} - \gamma_{\chi 1})\gamma_{\chi 1}^{\perp}(\gamma_{\chi 2} - \gamma_{\chi 1}) \in \mathfrak{S}_1.\end{aligned} \tag{4-33}$$

Together with (4-31) we conclude that $\gamma_{\chi_1} - \gamma_{\chi_2}$ belongs to $\mathfrak{S}_1^{\gamma_{\chi 2}}$. □

The following lemma is a consequence of [Hainzl et al. 2005a, Lemma 1] and the fact that $\gamma_{\chi 1} - \gamma_{\chi 2} \in \mathfrak{S}_2$.

Lemma 4.7. *Any self-adjoint operator A is in $\mathfrak{S}_1^{\gamma_{\chi 1}}$ if and only if A is in $\mathfrak{S}_1^{\gamma_{\chi 2}}$. Moreover* $\mathrm{Tr}_{\gamma_{\chi_1}}(A) = \mathrm{Tr}_{\gamma_{\chi_2}}(A)$.

The fact that $Q_{\chi 1} \in \mathfrak{S}_1^{\gamma_{\chi 1}}$ implies $|\nabla| Q_{\chi 1} \in \mathfrak{S}_2$, and $Q_{\chi 1} \in \mathfrak{S}_1^{\gamma_{\chi 2}}$ by Lemma 4.7. In view of this and Lemma 4.6 we know that $\widetilde{Q} = \gamma_{\chi 1} - \gamma_{\chi 2} + Q_{\chi 1}$ belongs to $\mathfrak{S}_1^{\gamma_{\chi 2}}$. The inequality (4-32) implies $|\nabla|\widetilde{Q} = |\nabla| Q_{\chi 1} + |\nabla|(\gamma_{\chi 1} - \gamma_{\chi 2}) \in \mathfrak{S}_2$. It remains to prove that $|\nabla|\gamma_{\chi 2}^{\perp}\widetilde{Q}\gamma_{\chi 2}^{\perp}|\nabla| \in \mathfrak{S}_1$ and $|\nabla|\gamma_{\chi 2}\widetilde{Q}\gamma_{\chi 2}|\nabla| \in \mathfrak{S}_1$.

In view of (4-29) we have

$$\begin{aligned}|\nabla|\gamma_{\chi_2}^{\perp}\widetilde{Q}\gamma_{\chi_2}^{\perp}|\nabla| &= |\nabla|\gamma_{\chi_2}^{\perp}Q_{\chi_1}\gamma_{\chi_2}^{\perp}|\nabla| + |\nabla|\gamma_{\chi_2}^{\perp}(\gamma_{\chi_1}-\gamma_{\chi_2})\gamma_{\chi_2}^{\perp}|\nabla|,\\ |\nabla|\gamma_{\chi_2}\widetilde{Q}\gamma_{\chi_2}|\nabla| &= |\nabla|\gamma_{\chi_2}Q_{\chi_1}\gamma_{\chi_2}|\nabla| + |\nabla|\gamma_{\chi_2}(\gamma_{\chi_1}-\gamma_{\chi_2})\gamma_{\chi_2}|\nabla|.\end{aligned} \tag{4-34}$$

We estimate (4-34) term by term. By Lemma A.8 we know $\||\nabla|\gamma_{\chi_i}\| \le \||\nabla|(1-\Delta)^{-1}\|\|(1-\Delta)\gamma_{\chi_i}\| < \infty$. Moreover, by writing $\gamma_{\chi_2}^{\perp} = \gamma_{\chi_1}^{\perp} + \gamma_{\chi_1} - \gamma_{\chi_2}$ we obtain

$$\begin{aligned}|\nabla|\gamma_{\chi_2}^{\perp}Q_{\chi_1}\gamma_{\chi_2}^{\perp}|\nabla| &= |\nabla|\gamma_{\chi_1}^{\perp}Q_{\chi_1}\gamma_{\chi_1}^{\perp}|\nabla| + |\nabla|\gamma_{\chi_1}^{\perp}Q_{\chi_1}(\gamma_{\chi_1}-\gamma_{\chi_2})|\nabla| + |\nabla|(\gamma_{\chi_1}-\gamma_{\chi_2})Q_{\chi_1}\gamma_{\chi_2}^{\perp}|\nabla|\\ &= |\nabla|\gamma_{\chi_1}^{\perp}Q_{\chi_1}\gamma_{\chi_1}^{\perp}|\nabla| + |\nabla|Q_{\chi_1}(\gamma_{\chi_1}-\gamma_{\chi_2})|\nabla| - |\nabla|\gamma_{\chi_1}Q_{\chi_1}(\gamma_{\chi_1}-\gamma_{\chi_2})|\nabla|\\ &\qquad + |\nabla|(\gamma_{\chi_1}-\gamma_{\chi_2})Q_{\chi_1}|\nabla| - |\nabla|(\gamma_{\chi_1}-\gamma_{\chi_2})Q_{\chi_1}\gamma_{\chi_2}|\nabla|.\end{aligned}$$

In view of (4-32), by the Cauchy–Schwarz inequality for Schatten operators,

$$\begin{aligned}\||\nabla|\gamma_{\chi_2}^{\perp}Q_{\chi_1}\gamma_{\chi_2}^{\perp}|\nabla|\|_{\mathfrak{S}_1} &\le \||\nabla|\gamma_{\chi_1}^{\perp}Q_{\chi_1}\gamma_{\chi_1}^{\perp}|\nabla|\|_{\mathfrak{S}_1} + \||\nabla|Q_{\chi_1}\|_{\mathfrak{S}_2}\|(\gamma_{\chi_1}-\gamma_{\chi_2})|\nabla|\|_{\mathfrak{S}_2}\\ &\quad + \||\nabla|\gamma_{\chi_1}\|\|Q_{\chi_1}\|_{\mathfrak{S}_2}\|(\gamma_{\chi_1}-\gamma_{\chi_2})|\nabla|\|_{\mathfrak{S}_2} + \||\nabla|(\gamma_{\chi_1}-\gamma_{\chi_2})\|_{\mathfrak{S}_2}\|Q_{\chi_1}|\nabla|\|_{\mathfrak{S}_2}\\ &\quad + \||\nabla|(\gamma_{\chi_1}-\gamma_{\chi_2})\|_{\mathfrak{S}_2}\|Q_{\chi_1}\|_{\mathfrak{S}_2}\|\gamma_{\chi_2}|\nabla|\| < \infty,\end{aligned}$$

and similarly by writing $\gamma_{\chi_2} = \gamma_{\chi_1} + \gamma_{\chi_2} - \gamma_{\chi_1}$ the following estimate holds:

$$\begin{aligned}\||\nabla|\gamma_{\chi_2}Q_{\chi_1}\gamma_{\chi_2}|\nabla|\|_{\mathfrak{S}_1} &\le \||\nabla|\gamma_{\chi_1}Q_{\chi_1}\gamma_{\chi_1}|\nabla|\|_{\mathfrak{S}_1} + \||\nabla|\gamma_{\chi_1}\|\|Q_{\chi_1}\|_{\mathfrak{S}_2}\|(\gamma_{\chi_1}-\gamma_{\chi_2})|\nabla|\|_{\mathfrak{S}_2}\\ &\quad + \|\gamma_{\chi_2}|\nabla|\|\|Q_{\chi_1}\|_{\mathfrak{S}_2}\||\nabla|(\gamma_{\chi_1}-\gamma_{\chi_2})\|_{\mathfrak{S}_2} < \infty,\end{aligned}$$

From (4-33) we know that

$$\begin{aligned}\||\nabla|\gamma_{\chi_2}^{\perp}(\gamma_{\chi_1}-\gamma_{\chi_2})\gamma_{\chi_2}^{\perp}|\nabla|\|_{\mathfrak{S}_1} &= \||\nabla|(\gamma_{\chi_1}-\gamma_{\chi_2})\gamma_{\chi_1}(\gamma_{\chi_1}-\gamma_{\chi_2})|\nabla|\|_{\mathfrak{S}_1} \le \||\nabla|(\gamma_{\chi_1}-\gamma_{\chi_2})\|_{\mathfrak{S}_2}^2 < \infty,\\ \||\nabla|\gamma_{\chi_2}(\gamma_{\chi_1}-\gamma_{\chi_2})\gamma_{\chi_2}|\nabla|\|_{\mathfrak{S}_1} &= \||\nabla|(\gamma_{\chi_2}-\gamma_{\chi_1})\gamma_{\chi_1}^{\perp}(\gamma_{\chi_2}-\gamma_{\chi_1})|\nabla|\|_{\mathfrak{S}_1} \le \||\nabla|(\gamma_{\chi_1}-\gamma_{\chi_2})\|_{\mathfrak{S}_2}^2 < \infty.\end{aligned}$$

This shows that $|\nabla|\gamma_{\chi_2}^{\perp}\widetilde{Q}\gamma_{\chi_2}^{\perp}|\nabla| \in \mathfrak{S}_1$ and $|\nabla|\gamma_{\chi_2}\widetilde{Q}\gamma_{\chi_2}|\nabla| \in \mathfrak{S}_1$. In view of (4-34), this allows us to conclude that $\widetilde{Q} \in \mathcal{Q}_{\chi_2}$. On the other hand, it is easy to see that $-\gamma_{\chi_2} \le \widetilde{Q} = \gamma_{\chi_1} + Q_{\chi_1} - \gamma_{\chi_2} \le 1 - \gamma_{\chi_2}$, which shows that $\widetilde{Q}$ belongs to the convex set $\mathcal{K}_{\chi_2}$.

The state $\widetilde{Q}$ is a minimizer. We now prove that $\widetilde{Q}$ is a minimizer of the problem (3-12) associated with the cut-off function χ_2. As $\widetilde{Q} \in \mathcal{K}_{\chi_2}$, the fact that Q_{χ_2} is a minimizer implies

$$\mathcal{E}_{\chi_2}(\widetilde{Q}) - \kappa\,\mathrm{Tr}_{\gamma_{\chi_2}}(\widetilde{Q}) \ge \mathcal{E}_{\chi_2}(Q_{\chi_2}) - \kappa\,\mathrm{Tr}_{\gamma_{\chi_2}}(Q_{\chi_2}). \tag{4-35}$$

Define $\Theta := \widetilde{Q} - Q_{\chi_2} = Q_{\chi_1} - Q_{\chi_2} + \gamma_{\chi_1} - \gamma_{\chi_2}$. The inequality (4-35) can therefore also be written as

$$\mathcal{E}_{\chi_2}(\Theta) - \kappa\,\mathrm{Tr}_{\gamma_{\chi_2}}(\Theta) + D(\rho_\Theta, \rho_{Q_{\chi_2}}) \ge 0. \tag{4-36}$$

It is easy to see that $-1 \le \Theta \le 1$ and Θ belongs to $\mathcal{Q}_{\chi_2}$ (but not necessarily to the convex set $\mathcal{K}_{\chi_2}$), which also implies that the density ρ_Θ of Θ is well-defined and belongs to the Coulomb space $\mathcal{C}$. Therefore (4-36) is well-defined. Introduce another state by exchanging the indices 1 and 2 in the definition of $\widetilde{Q}$:

$$\widetilde{\widetilde{Q}} := \gamma_{\chi_2} + Q_{\chi_2} - \gamma_{\chi_1}.$$

Proceeding as before, it can be shown that $\widetilde{\widetilde{Q}} \in \mathcal{K}_{\chi_1}$. By definition $Q_{\chi_1} = \Theta + \widetilde{\widetilde{Q}}$. Since Q_{χ_1} minimizes the problem (3-12) associated with χ_1 and $\widetilde{\widetilde{Q}} \in \mathcal{K}_{\chi_1}$,

$$\mathcal{E}_{\chi_1}(\widetilde{\widetilde{Q}}) - \kappa \operatorname{Tr}_{\gamma_{\chi_1}}(\widetilde{\widetilde{Q}}) \geq \mathcal{E}_{\chi_1}(\Theta + \widetilde{\widetilde{Q}}) - \kappa \operatorname{Tr}_{\gamma_{\chi_1}}(\Theta + \widetilde{\widetilde{Q}}).$$

The above equation can be simplified as

$$\mathcal{E}_{\chi_1}(\Theta) - \kappa \operatorname{Tr}_{\gamma_{\chi_1}}(\Theta) + D(\rho_\Theta, \rho_{\widetilde{\widetilde{Q}}}) \leq 0. \tag{4-37}$$

Let us show that the left-hand sides of (4-36) and (4-37) are equal. First of all as Θ belongs to $\mathcal{Q}_{\chi_2}$, we know that $\operatorname{Tr}_{\gamma_{\chi_2}}(\Theta) = \operatorname{Tr}_{\gamma_{\chi_1}}(\Theta)$ by Lemma 4.7. Note also that $\rho_{\widetilde{\widetilde{Q}}} = \rho_{\chi_2} - \rho_{\chi_1} + \rho_{Q_{\chi_2}}$. By Lemma 4.7

$$\begin{aligned}
&\mathcal{E}_{\chi_2}(\Theta) - \kappa \operatorname{Tr}_{\gamma_{\chi_2}}(\Theta) + D(\rho_\Theta, \rho_{Q_{\chi_2}}) - (\mathcal{E}_{\chi_1}(\Theta) - \kappa \operatorname{Tr}_{\gamma_{\chi_1}}(\Theta) + D(\rho_\Theta, \rho_{\widetilde{\widetilde{Q}}}))\\
&\quad= \operatorname{Tr}_{\gamma_{\chi_2}}((-\Delta + V_{\chi_2})\Theta) - D(\rho_\Theta, \nu_{\chi_2}) + \tfrac{1}{2} D(\rho_\Theta, \rho_\Theta) - \operatorname{Tr}_{\gamma_{\chi_1}}((-\Delta + V_{\chi_1})\Theta) + D(\rho_\Theta, \nu_{\chi_1})\\
&\qquad\qquad - \tfrac{1}{2} D(\rho_\Theta, \rho_\Theta) - \kappa(\operatorname{Tr}_{\gamma_{\chi_2}}(\Theta) - \operatorname{Tr}_{\gamma_{\chi_1}}(\Theta)) + D(\rho_\Theta, \rho_{Q_{\chi_2}}) - D(\rho_\Theta, \rho_{\widetilde{\widetilde{Q}}})\\
&\quad= \operatorname{Tr}_{\gamma_{\chi_2}}((-\Delta + V_{\chi_2})\Theta) - \operatorname{Tr}_{\gamma_{\chi_1}}((-\Delta + V_{\chi_1})\Theta) + D(\rho_\Theta, \rho_{\chi_1} + \nu_{\chi_1} - \rho_{\chi_2} - \nu_{\chi_2})\\
&\quad= \operatorname{Tr}_{\gamma_{\chi_2}}((V_{\chi_2} - V_{\chi_1})\Theta) + D(\rho_\Theta, \rho_{\chi_1} + \nu_{\chi_1} - \rho_{\chi_2} - \nu_{\chi_2}).
\end{aligned} \tag{4-38}$$

We show that (4-38) is equal to zero by first showing that $(V_{\chi_2} - V_{\chi_1})\Theta \in \mathfrak{S}_1 \subset \mathfrak{S}_1^{\gamma_{\chi_2}}$. Note that $V_{\chi_1} - V_{\chi_2} = (\chi_1^2 - \chi_2^2)(V_{\mathrm{per},L} - V_{\mathrm{per},R}) \in L^\infty(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$. By the definition of Θ, by the Kato–Seiler–Simon inequality and using calculations similar to (4-31),

$$\begin{aligned}
&\|(V_{\chi_1} - V_{\chi_2})\Theta\|_{\mathfrak{S}_1}\\
&\quad= \|(V_{\chi_1} - V_{\chi_2})(Q_{\chi_1} - Q_{\chi_2} + \gamma_{\chi_1} - \gamma_{\chi_2})\|_{\mathfrak{S}_1}\\
&\quad\leq \|(V_{\chi_1} - V_{\chi_2})(1-\Delta)^{-1}\|_{\mathfrak{S}_2}(\|(1-\Delta)(Q_{\chi_1} - Q_{\chi_2})\|_{\mathfrak{S}_2} + \|(1-\Delta)(\gamma_{\chi_1} - \gamma_{\chi_2})\|_{\mathfrak{S}_2})\\
&\quad\leq \frac{1}{2\sqrt{\pi}} \|V_{\chi_1} - V_{\chi_2}\|_{L^2}(\|(1-\Delta)(Q_{\chi_1} - Q_{\chi_2})\|_{\mathfrak{S}_2} + C\|(\chi_1^2 - \chi_2^2)(V_{\mathrm{per},L} - V_{\mathrm{per},R})(1-\Delta)^{-1}\|_{\mathfrak{S}_2}) < \infty,
\end{aligned}$$

where the fact that $(1-\Delta)Q_{\chi_i} \in \mathfrak{S}_2$ follows arguments similar to the ones used in the proof of [Cancès et al. 2008, Proposition 2]. Hence $(V_{\chi_1} - V_{\chi_2})\Theta$ belongs to $\mathfrak{S}_1$, and

$$\operatorname{Tr}_{\gamma_{\chi_2}}((V_{\chi_2} - V_{\chi_1})\Theta) = \operatorname{Tr}((V_{\chi_2} - V_{\chi_1})\Theta).$$

On the other hand, by the definition of V_{χ_i} in (3-8) and ν_i in (3-10) for $i = 1, 2$, we deduce that

$$\operatorname{Tr}((V_{\chi_2} - V_{\chi_1})\Theta) = D(\rho_\Theta, (\rho_{\chi_2} - \mu_{\chi_2}) - (\rho_{\chi_1} - \mu_{\chi_1})) = D(\rho_\Theta, \rho_{\chi_2} - \rho_{\chi_1} + \nu_{\chi_2} - \nu_{\chi_1}).$$

The above equation implies that the quantity (4-38) is equal to 0. Hence, in view of (4-36) and (4-37),

$$\mathcal{E}_{\chi_2}(\Theta) - \kappa \operatorname{Tr}_{\gamma_{\chi_2}}(\Theta) + D(\rho_\Theta, \rho_{Q_{\chi_2}}) = \mathcal{E}_{\chi_1}(\Theta) - \kappa \operatorname{Tr}_{\gamma_{\chi_1}}(\Theta) + D(\rho_\Theta, \rho_{\widetilde{\widetilde{Q}}}) \equiv 0.$$

We conclude with (4-35) that

$$\mathcal{E}_{\chi_2}(\widetilde{Q}) - \kappa \operatorname{Tr}_{\gamma_{\chi_2}}(\widetilde{Q}) \equiv \mathcal{E}_{\chi_2}(Q_{\chi_2}) - \kappa \operatorname{Tr}_{\gamma_{\chi_2}}(Q_{\chi_2}).$$

Therefore $\widetilde{Q}$ is a minimizer of the problem (3-12) associated with the cut-off function χ_2. From Proposition 3.5 we know that $\rho_{\widetilde{Q}} \equiv \rho_{Q_{\chi_2}}$, which is equivalent to $\rho_{Q_{\chi_2}} + \rho_{\chi_2} = \rho_{\chi_1} + \rho_{Q_{\chi_1}}$. By the arbitrariness of the choice of χ_1, χ_2 we deduce that $\rho_\chi + \rho_{Q_\chi}$ is independent of the cut-off function $\chi \in \mathcal{X}$.

Appendix

A1. ***Proof of Lemma 2.3.*** For $n \in \mathbb{Z}$, let us consider the 2-dimensional equation

$$-\Delta_{\boldsymbol{r}} G_n + 4\pi^2 n^2 G_n = 2\pi \delta_{\boldsymbol{r}=0} \quad \text{in } \mathscr{S}'(\mathbb{R}^2).$$

It is well known (see for example [Lieb and Loss 2001; Lahbabi 2014]) that the solution of the above equation is

$$G_n(|\boldsymbol{r}|) = \begin{cases} -\log(|\boldsymbol{r}|), & n \equiv 0, \\ K_0(2\pi|n||\boldsymbol{r}|), & |n| \geq 1, \end{cases}$$

where $K_0(\alpha) := \int_0^{+\infty} \mathrm{e}^{-\alpha\cosh(t)}\,dt$ is the modified Bessel function of the second kind. Therefore the Green's function $G(x, \boldsymbol{r})$ defined in (2-7) can be rewritten as

$$G(x, \boldsymbol{r}) = 2\sum_{n\in\mathbb{Z}} \mathrm{e}^{2\mathrm{i}\pi nx} G_n(\boldsymbol{r}) \in \mathscr{S}'_{\mathrm{per},x}(\Gamma). \tag{A-1}$$

This implies

$$-\Delta G(x, \boldsymbol{r}) = 4\pi \sum_{n\in\mathbb{Z}} \delta_{(x,\boldsymbol{r})=(n,0)}.$$

Taking the Fourier transform $\mathscr{F}$ on both sides of (A-1) we obtain (2-8). Let us now give some estimates on $\widetilde{G}$ defined in (2-7). Recall that there exist two positive constants C_0 and C_1 such that [Duffin 1971]

$$0 \leq K_0(\alpha) \leq \begin{cases} C_0|\log(\alpha)|, & \text{when } \alpha \leq 2\pi, \\ C_1\mathrm{e}^{-\alpha}(\pi/2\alpha)^{1/2}, & \text{when } \alpha > 2\pi. \end{cases}$$

For $|\boldsymbol{r}| > 1$, it holds that

$$|\widetilde{G}(x, \boldsymbol{r})| \leq 2C_1 \sum_{n=1}^{+\infty} \frac{\mathrm{e}^{-2\pi n|\boldsymbol{r}|}}{\sqrt{n|\boldsymbol{r}|}} \leq \frac{2C_1}{1-\mathrm{e}^{-2\pi}} \frac{\mathrm{e}^{-2\pi|\boldsymbol{r}|}}{\sqrt{|\boldsymbol{r}|}}. \tag{A-2}$$

For $|\boldsymbol{r}| \leq 1$ fixed, there exists $N \geq 1$ such that $N \leq 1/|\boldsymbol{r}| < N+1$. In particular, for $n > N+1$ we have $2\pi n|\boldsymbol{r}| > 2\pi$. There exists therefore a positive constant C such that

$$|\widetilde{G}(x, \boldsymbol{r})| \leq 4C_0\left|\sum_{n=1}^{N} \log(2\pi n|\boldsymbol{r}|)\right| + 2C_1 \sum_{n=N+1}^{\infty} \frac{\mathrm{e}^{-2\pi n|\boldsymbol{r}|}}{\sqrt{n|\boldsymbol{r}|}} \leq \frac{C}{|\boldsymbol{r}|}. \tag{A-3}$$

Together with (A-2) we deduce that $\widetilde{G}(x, \boldsymbol{r}) \in L^p_{\mathrm{per},x}(\Gamma)$ for $1 \leq p < 2$. Note that for all $\boldsymbol{r} \in \mathbb{R}^2\backslash\{0\}$, it holds $\int_{-1/2}^{1/2} \overline{G}(x, \boldsymbol{r})\,dx \equiv 0$. Consider, for $\boldsymbol{r} \neq 0$,

$$\overline{G}(x, \boldsymbol{r}) = \sum_{n\in\mathbb{Z}} \left(\frac{1}{\sqrt{(x-n)^2 + |\boldsymbol{r}|^2}} - \int_{-1/2}^{1/2} \frac{1}{\sqrt{(x-y-n)^2 + |\boldsymbol{r}|^2}}\,dy \right).$$

From [Blanc and Le Bris 2000, equation (1.8)],

$$-\Delta(\overline{G}(x,\boldsymbol{r})-2\log(|\boldsymbol{r}|))=4\pi\sum_{k\in\mathbb{Z}}\delta_{(x,\boldsymbol{r})=(k,0)}\in\mathscr{S}'(\mathbb{R}^3),$$

with $\overline{G}(x,\boldsymbol{r})=\mathcal{O}(1/|\boldsymbol{r}|)$ when $|\boldsymbol{r}|\to\infty$ by [Blanc and Le Bris 2000, Lemma 2.2]. Setting $u(x,\boldsymbol{r})=\widetilde{G}(x,\boldsymbol{r})-\overline{G}(x,\boldsymbol{r})$ we therefore obtain that $-\Delta u(x,\boldsymbol{r})\equiv 0$. As $u(x,\boldsymbol{r})$ belongs to $L^1_{\mathrm{loc}}(\mathbb{R}^3)$, by Weyl's lemma for the Laplace equation we obtain that $u(x,\boldsymbol{r})$ is $C^\infty(\mathbb{R}^3)$. On the other hand, by the decay properties of $\widetilde{G}$ and $\overline{G}$, we deduce that $|u(\,\cdot\,,\boldsymbol{r})|\to 0$ when $|\boldsymbol{r}|\to\infty$ uniformly in x, hence by the maximum modulus principle for harmonic functions we can conclude that $u\equiv 0$; hence $\widetilde{G}(x,\boldsymbol{r})=\overline{G}(x,\boldsymbol{r})$.

A2. ***Proof of Lemma 2.10.*** Assume that (2-23) holds, that is $\mu_{\mathrm{per}}(x,\boldsymbol{r})\equiv\mu_{\mathrm{per}}(x,|\boldsymbol{r}|)$ has radial symmetry in the $\boldsymbol{r}$-direction. It is clear that the results of Theorem 2.7 hold. We employ the same notation as in Theorem 2.7 in the sequel. By the uniqueness of density, $\rho_{\gamma_{\mathrm{per}}}$ enjoys the same radial symmetry in the $\boldsymbol{r}$-direction. Recall that $q_{\mathrm{per}}=\rho_{\gamma_{\mathrm{per}}}-\mu_{\mathrm{per}}$. Together with the facts that $\int_\Gamma|\boldsymbol{r}|\rho_{\gamma_{\mathrm{per}}}(x,\boldsymbol{r})\,dx\,d\boldsymbol{r}<+\infty$ and that μ_{per} has compact support in the $\boldsymbol{r}$-direction, the radial symmetry in the $\boldsymbol{r}$-direction implies

$$\int_\Gamma\boldsymbol{r}\cdot q_{\mathrm{per}}(x,\boldsymbol{r})\,dx\,d\boldsymbol{r}\equiv 0. \tag{A-4}$$

Note also that the exponential decay of density implies

$$\int_\Gamma|\boldsymbol{r}|^2|q_{\mathrm{per}}(x,\boldsymbol{r})|\,dx\,d\boldsymbol{r}<+\infty.$$

Following calculations similar to the those in (4-9), (4-10) and (4-12), it is easy to deduce that

$$\partial_{\boldsymbol{k}}\mathscr{F}q_{\mathrm{per}}(0,\boldsymbol{k})\equiv 0$$

and $\partial^2_{\boldsymbol{k}}\mathscr{F}q_{\mathrm{per}}(0,\boldsymbol{k})$ is continuous and bounded, so that $\mathscr{F}V_{\mathrm{per}}(0,\cdot\,)$ belongs to $L^2(\mathbb{R}^2)$, and V_{per} also belongs to $L^2_{\mathrm{per},x}(\Gamma)$. Let us prove that $V_{\mathrm{per}}\in L^p_{\mathrm{per},x}(\Gamma)$ for $1<p<2$ (for which we can conclude that V_{per} belongs to $L^p_{\mathrm{per},x}(\Gamma)$ for $1<p\le+\infty$). Let us rewrite V_{per} as

$$V_{\mathrm{per}}(x,\boldsymbol{r})=(q_{\mathrm{per}}\star_\Gamma G)(x,\boldsymbol{r})=(q_{\mathrm{per}}\star_\Gamma\widetilde{G})(x,\boldsymbol{r})+T(\boldsymbol{r}), \tag{A-5}$$

where

$$T(\boldsymbol{r})=-2\int_{\mathbb{R}^2}\bar{q}_{\mathrm{per}}(\boldsymbol{r}')\log(|\boldsymbol{r}-\boldsymbol{r}'|)\,d\boldsymbol{r}',\quad\bar{q}_{\mathrm{per}}(\boldsymbol{r}):=\int_{-1/2}^{1/2}q_{\mathrm{per}}(x,\boldsymbol{r})\,dx.$$

Recall that μ_{per} has compact support in the $\boldsymbol{r}$-direction; hence there exist positive constants C_q, α_q such that,

$$\text{for all }(x,\boldsymbol{r})\in\mathbb{R}^3,\quad|q_{\mathrm{per}}(\,\cdot\,,\boldsymbol{r})|\le C_q\mathrm{e}^{-\alpha_q|\boldsymbol{r}|},\quad|\bar{q}_{\mathrm{per}}(\boldsymbol{r})|\le C_q\mathrm{e}^{-\alpha_q|\boldsymbol{r}|}.$$

As $\widetilde{G}$ belongs to $L^p_{\mathrm{per},x}(\Gamma)$ for $1\le p<2$, by Young's convolution inequality we deduce that $q_{\mathrm{per}}\star_\Gamma\widetilde{G}$ belongs to $L^t_{\mathrm{per},x}(\Gamma)$ for $1\le t\le+\infty$.

It remains to prove that $T(\boldsymbol{r})$ belongs to $L^p(\mathbb{R}^2)$ for $1<p<2$. Let us use the partition $\mathbb{R}^2=\{|\boldsymbol{r}|\le 2R\}\cup\{|\boldsymbol{r}|>2R\}$ for the integration domain of $T(\boldsymbol{r})$. Note first that $\log(|\boldsymbol{r}|)$ is $L^t_{\mathrm{loc}}(\mathbb{R}^2)$ for

$1 \leq t < +\infty$. Therefore, by a Cauchy–Schwarz inequality, there exists a positive constant $C_{R,1}$ such that, for $p' = p/(p-1) \in (2, +\infty)$,

$$\begin{aligned}\Big(\int_{|\boldsymbol{r}|\leq 2R} |T(\boldsymbol{r})|^p \, d\boldsymbol{r}\Big)^{1/p} &= 2\Big(\int_{|\boldsymbol{r}|\leq 2R}\Big|\int_{\mathbb{R}^2} \bar{q}_{\mathrm{per}}(\boldsymbol{r}')\log(|\boldsymbol{r}-\boldsymbol{r}'|)\,d\boldsymbol{r}'\Big|^p\,d\boldsymbol{r}\Big)^{1/p}\\ &\leq 2\Big(\int_{|\boldsymbol{r}|\leq 2R}\Big|\int_{|\boldsymbol{r}'|\leq 3R} \bar{q}_{\mathrm{per}}(\boldsymbol{r}')\log(|\boldsymbol{r}-\boldsymbol{r}'|)\,d\boldsymbol{r}'\Big|^p\,d\boldsymbol{r}\Big)^{1/p}\\ &\qquad + 2C_q\Big(\int_{|\boldsymbol{r}|\leq 2R}\Big|\int_{|\boldsymbol{r}'|> 3R} \mathrm{e}^{-\alpha_q|\boldsymbol{r}'|}|\log(|\boldsymbol{r}-\boldsymbol{r}'|)|\,d\boldsymbol{r}'\Big|^p\,d\boldsymbol{r}\Big)^{1/p}\\ &\leq 2\Big(\int_{|\boldsymbol{r}'|\leq 3R} |\bar{q}_{\mathrm{per}}|^p\Big)^{1/p}\Big(\int_{|\boldsymbol{r}|\leq 2R}\Big(\int_{|\boldsymbol{r}'|\leq 3R} |\log(|\boldsymbol{r}-\boldsymbol{r}'|)|^{p'}\,d\boldsymbol{r}'\Big)^{p/p'}\,d\boldsymbol{r}\Big)^{1/p}\\ &\qquad + 2C_q\Big(\int_{|\boldsymbol{r}|\leq 2R}\Big|\int_{|\boldsymbol{r}'|> 3R} \mathrm{e}^{-\alpha_q|\boldsymbol{r}'|}\big|\log(|(R,R)-\boldsymbol{r}'|)\big|\,d\boldsymbol{r}'\Big|^p\,d\boldsymbol{r}\Big)^{1/p}\\ &\leq C_{R,1}. \end{aligned} \tag{A-6}$$

Let us look at the integration domain $\{|\boldsymbol{r}| > 2R\}$. Note that by the charge neutrality condition and the radial symmetry condition (A-4), it holds, for any $\boldsymbol{r} \neq 0$,

$$\int_{\mathbb{R}^2} \bar{q}_{\mathrm{per}}(\boldsymbol{r}')\log(|\boldsymbol{r}|)\,d\boldsymbol{r}' \equiv 0, \qquad \int_{\mathbb{R}^2} \bar{q}_{\mathrm{per}}(\boldsymbol{r}')\frac{\boldsymbol{r}'\boldsymbol{r}}{|\boldsymbol{r}|^2}\,d\boldsymbol{r}' \equiv 0.$$

Denote by

$$Q(\boldsymbol{r},\boldsymbol{r}') := \log(|\boldsymbol{r}-\boldsymbol{r}'|) - \log(|\boldsymbol{r}|) - \frac{\boldsymbol{r}'\boldsymbol{r}}{|\boldsymbol{r}|^2} = \frac{1}{2}\log\Big(1 - \frac{2\boldsymbol{r}\boldsymbol{r}'}{|\boldsymbol{r}|^2} + \frac{|\boldsymbol{r}'|^2}{|\boldsymbol{r}|^2}\Big) - \frac{\boldsymbol{r}'\boldsymbol{r}}{|\boldsymbol{r}|^2}.$$

Then

$$T(\boldsymbol{r}) = -2\int_{\mathbb{R}^2} \bar{q}_{\mathrm{per}}(\boldsymbol{r}')Q(\boldsymbol{r},\boldsymbol{r}')\,d\boldsymbol{r}'.$$

Note that when $|\boldsymbol{r}| > 2R$ and $|\boldsymbol{r}'|/|\boldsymbol{r}| \leq \epsilon_R$ for $\epsilon_R > 0$ fixed. A Taylor expansion shows that there exists a positive constant C such that $|Q(\boldsymbol{r},\boldsymbol{r}')| \leq C|\boldsymbol{r}'|^2/|\boldsymbol{r}|^2$. This motivates the following partition of $\mathbb{R}^2$ given $|\boldsymbol{r}| > 2R$:

$$\mathbb{R}^2 = \mathbb{B}_{\epsilon_R} \cup \mathbb{B}_{\epsilon_R}^{\complement}, \qquad \mathbb{B}_{\epsilon_R} := \Big\{\boldsymbol{r}' \in \mathbb{R}^2 \;\Big|\; \frac{|\boldsymbol{r}'|}{|\boldsymbol{r}|} \leq \epsilon_R\Big\}.$$

Hence

$$T(\boldsymbol{r}) = T_{\mathrm{int}}(\boldsymbol{r}) + T_{\mathrm{ext}}(\boldsymbol{r}), \quad T_{\mathrm{int}}(\boldsymbol{r}) := \int_{\mathbb{B}_{\epsilon_R}} \bar{q}_{\mathrm{per}}(\boldsymbol{r}')Q(\boldsymbol{r},\boldsymbol{r}')\,d\boldsymbol{r}', \quad T_{\mathrm{ext}}(\boldsymbol{r}) := \int_{\mathbb{B}_{\epsilon_R}^{\complement}} \bar{q}_{\mathrm{per}}(\boldsymbol{r}')Q(\boldsymbol{r},\boldsymbol{r}')\,d\boldsymbol{r}'.$$

Therefore, for $1 < p < 2$,

$$\begin{aligned}\int_{|\boldsymbol{r}|>2R} |T_{\mathrm{int}}(\boldsymbol{r})|^p\,d\boldsymbol{r} &\leq 2C^p\int_{|\boldsymbol{r}|>2R}\Big|\int_{\mathbb{B}_{\epsilon_R}} \bar{q}_{\mathrm{per}}(\boldsymbol{r}')\frac{|\boldsymbol{r}'|^2}{|\boldsymbol{r}|^2}\,d\boldsymbol{r}'\Big|^p\,d\boldsymbol{r}\\ &\leq 2C'\int_{|\boldsymbol{r}|>2R}\Big|\int_{|\boldsymbol{r}'|\leq\epsilon_R|\boldsymbol{r}|} \mathrm{e}^{-\alpha_q|\boldsymbol{r}'|}|\boldsymbol{r}'|^2\,d\boldsymbol{r}'\Big|^p|\boldsymbol{r}|^{-2p}\,d\boldsymbol{r} < +\infty. \end{aligned} \tag{A-7}$$

Similarly,

$$\int_{|\boldsymbol{r}|>2R} |T_{\mathrm{ext}}(\boldsymbol{r})|^p \, d\boldsymbol{r} \le C_1 \int_{|\boldsymbol{r}|>2R} \left| \int_{|\boldsymbol{r}'|>\epsilon_R|\boldsymbol{r}|} \mathrm{e}^{-\alpha_q|\boldsymbol{r}'|} |Q(\boldsymbol{r},\boldsymbol{r}')| \, d\boldsymbol{r}' \right|^p d\boldsymbol{r}$$
$$\le C_1 \int_{|\boldsymbol{r}|>2R} \left| \int_{|\boldsymbol{r}'|>\epsilon_R|\boldsymbol{r}|} \mathrm{e}^{-\alpha_q\epsilon_R|\boldsymbol{r}|/2} \mathrm{e}^{-\alpha_q|\boldsymbol{r}'|/2} |Q(\boldsymbol{r},\boldsymbol{r}')| \, d\boldsymbol{r}' \right|^p d\boldsymbol{r} < +\infty. \tag{A-8}$$

In view of (A-6), (A-7) and (A-8) we conclude that $T(\boldsymbol{r})$ belongs to $L^p(\mathbb{R}^2)$ for $1 < p < 2$. This leads to the conclusion that V_{per} belongs to $L^p_{\mathrm{per},x}(\Gamma)$ for $1 < p \le +\infty$.

A3. ***Proof of Proposition 3.1.*** Let us emphasize that the function χ being translation-invariant in the $\boldsymbol{r}$-direction makes it difficult to control the compactness in the $\boldsymbol{r}$-direction across the junction surface. Our geometry is very different from the cylindrical geometry considered in [Hempel et al. 2015] for instance which automatically provides compactness in the $\boldsymbol{r}$-direction.

The proof is organized as follows: we first prove that any $\lambda \in \sigma_{\mathrm{ess}}(H_{\mathrm{per},L}) \cup \sigma_{\mathrm{ess}}(H_{\mathrm{per},R})$ also belongs to $\sigma_{\mathrm{ess}}(H_\chi)$. We then prove that $\sigma_{\mathrm{ess}}(H_\chi)$ is included in $\sigma_{\mathrm{ess}}(H_{\mathrm{per},L}) \cup \sigma_{\mathrm{ess}}(H_{\mathrm{per},R})$ based on the IMS formula [Ismagilov 1961; Morgan 1979; Morgan and Simon 1980; Simon 1983].

The union of $\sigma_{\mathrm{ess}}(H_{\mathrm{per},L}) \cup \sigma_{\mathrm{ess}}(H_{\mathrm{per},R})$ *is included in* $\sigma_{\mathrm{ess}}(H_\chi)$. Without loss of generality we prove that λ_L belonging to $\sigma_{\mathrm{ess}}(H_{\mathrm{per},L})$ also belongs to $\sigma_{\mathrm{ess}}(H_\chi)$. Consider a Weyl sequence $\{w_n\}_{n\in\mathbb{N}^*}$ for $H_{\mathrm{per},L}$ associated with λ_L. Let us construct a Weyl sequence for H_χ from $\{w_n\}_{n\in\mathbb{N}^*}$. Fix $n \in \mathbb{N}^*$. There exists a sequence $\{v_{k,n}\}_{k\in\mathbb{N}^*}$ belonging to $C_c^\infty(\mathbb{R}^3)$ such that, for all $\epsilon > 0$, there exists a $K_n \in \mathbb{N}^*$ such that, for any $k \ge K_n$,

$$\|v_{k,n} - w_n\|_{H^2(\mathbb{R}^3)} \le \epsilon. \tag{A-9}$$

It is easy to see that $v_{K_n,n}$ tends weakly to 0 in $L^2(\mathbb{R}^3)$ as $n \to \infty$ since w_n converges weakly to 0. As $v_{K_n,n}$ has compact support, for any fixed $n \in \mathbb{N}^*$ and for $m \in \mathbb{N}^*$ large enough,

$$\mathrm{supp}(\tau^x_{a_L m} v_{K_n,n}) \cap \left(([-a_L/2, +\infty) \times \mathbb{R}^2) \bigcup \mathfrak{B}_n \right) = \varnothing, \tag{A-10}$$

where $\mathfrak{B}_n$ denotes the ball of radius n centered at 0 in $\mathbb{R}^3$. Note that the above equality also ensures that $\tau^x_{a_L m} v_{K_n,n}$ tends weakly to 0 in $L^2(\mathbb{R}^3)$ when $m \to +\infty$ for n fixed. In view of (A-9) and (A-10), we introduce $\tilde{w}_n := \tau^x_{a_L m_n} v_{K_n,n}$ for $n \in \mathbb{N}^*$ so that (A-10) is satisfied. This implies that $\tilde{w}_n$ tends weakly to 0 in $L^2(\mathbb{R}^3)$ when $n \to +\infty$. Moreover, in view of (A-9) and by the definition of the Weyl sequence

$$\|(H_\chi - \lambda_L)\tilde{w}_n\|_{L^2} = \|(H_\chi - \lambda_L)\tau^x_{a_L m_n} v_{K_n,n}\|_{L^2} = \|(H_{\mathrm{per},L} - \lambda_L)\tau^x_{a_L m_n} v_{K_n,n}\|_{L^2}$$
$$\le \|\tau^x_{a_L m_n}(H_{\mathrm{per},L} - \lambda_L)(v_{K_n,n} - w_n)\|_{L^2} + \|\tau^x_{a_L m_n}(H_{\mathrm{per},L} - \lambda_L)w_n\|_{L^2}$$
$$\le (1 + \|V_{\mathrm{per},L}\|_{L^\infty} + |\lambda_L|)\|v_{K_n,n} - w_n\|_{H^2} + \|(H_{\mathrm{per},L} - \lambda_L)w_n\|_{L^2} \xrightarrow[n\to\infty]{} 0.$$

Therefore the sequence $\tilde{w}_n/\|\tilde{w}_n\|_{L^2}$ is a Weyl sequence of H_χ associated with λ_L. This leads to the conclusion that

$$\sigma_{\mathrm{ess}}(H_{\mathrm{per},L}) \cup \sigma_{\mathrm{ess}}(H_{\mathrm{per},R}) \subseteq \sigma_{\mathrm{ess}}(H_\chi). \tag{A-11}$$

The essential spectrum $\sigma_{\mathrm{ess}}(H_\chi)$ *in included in* $\sigma_{\mathrm{ess}}(H_{\mathrm{per},L}) \cup \sigma_{\mathrm{ess}}(H_{\mathrm{per},R})$. We prove that $\lambda \in \sigma_{\mathrm{ess}}(H_\chi)$ also belongs to $\sigma_{\mathrm{ess}}(H_{\mathrm{per},L}) \cup \sigma_{\mathrm{ess}}(H_{\mathrm{per},R})$. We prove this statement by the IMS localization formula (see for example [Simon 1983, Lemma 3.1]), which states that for any smooth partition of unity $\{J_a\}_{a=0}^k$ (k does not need to be finite) such that $\sum_{a=0}^k J_a^2 = 1$, the following algebraic decomposition holds (on the proper form domain) for an operator H:

$$H = \sum_{a=0}^{k} \bigl(J_a H J_a + \tfrac{1}{2}[J_a, [J_a, H]]\bigr). \tag{A-12}$$

The proof of the above statement is based on the following algebraic relation:

$$\text{for all } 0 \le a \le k, \qquad J_a^2 H + H J_a^2 - 2 J_a H J_a \equiv [J_a, [J_a, H]].$$

Consider a partition of unity $\sum_{i=1}^3 f_i^2 = 1$ such that

$$\begin{aligned} f_1 &\equiv 1 \quad \text{on } (-\infty, -a_L) \times \mathbb{R}^2, & f_2 &\equiv 1 \quad \text{on } (a_R, +\infty) \times \mathbb{R}^2, \\ f_1 &\equiv 0 \quad \text{on } (-a_L/2, +\infty) \times \mathbb{R}^2, & f_2 &\equiv 0 \quad \text{on } (-\infty, a_R/2) \times \mathbb{R}^2. \end{aligned}$$

This implies that f_3 has support in $[-a_L, a_R] \times \mathbb{R}^2$.

Consider a Weyl sequence $(\phi_n)_{n \in \mathbb{N}^*} \in H^2(\mathbb{R}^3)$ of H_χ associated with $\lambda \in \sigma_{\mathrm{ess}}(H_\chi)$. For $R \in \mathbb{R}^+$ large enough, by applying (A-12) to the operator $(H_\chi - \lambda)^2$ it holds that

$$\begin{aligned} (H_\chi - \lambda)^2 &= f_1(x/R, \cdot)(H_{\mathrm{per},L} - \lambda)^2 f_1(x/R, \cdot) + f_2(x/R, \cdot)(H_{\mathrm{per},R} - \lambda)^2 f_2(x/R, \cdot) \\ &\quad + f_3(x/R, \cdot)(H_\chi - \lambda)^2 f_3(x/R, \cdot) + \frac{1}{2}\sum_{i=1}^{3} [f_i(x/R, \cdot), [f_i(x/R, \cdot), (H_\chi - \lambda)^2]]. \end{aligned}$$

In view of the above formula, for any ϕ_n,

$$\begin{aligned} \|(H_\chi - \lambda)\phi_n\|_{L^2}^2 &= \|(H_{\mathrm{per},L} - \lambda) f_1(x/R, \cdot)\phi_n\|_{L^2}^2 + \|(H_{\mathrm{per},R} - \lambda) f_2(x/R, \cdot)\phi_n\|_{L^2}^2 \\ &\quad + \bigl\|\bigl(-\tfrac{1}{2}\Delta - \lambda\bigr) f_3(x/R, \cdot)\phi_n\bigr\|_{L^2}^2 \\ &\quad + \|V_\chi f_3(x/R, \cdot)\phi_n\|_{L^2}^2 + 2\Re\langle V_\chi f_3(x/R, \cdot)\phi_n, \bigl(-\tfrac{1}{2}\Delta - \lambda\bigr) f_3(x/R, \cdot)\phi_n\rangle_{L^2} \\ &\quad + \frac{1}{2}\sum_{i=1}^{3} \langle \phi_n, [f_i(x/R, \cdot), [f_i(x/R, \cdot), (H_\chi - \lambda)^2]]\phi_n\rangle_{L^2}. \end{aligned} \tag{A-13}$$

Let us show that we can find a sequence of $R_n \to +\infty$ such that the last two terms of (A-13) tends to 0. First of all, remark that $\bigl[f_i(x/R, \cdot), \bigl[f_i(x/R, \cdot), -\tfrac{1}{2}\Delta\bigr]\bigr] = -|\nabla f_i(x/R, \cdot)|^2$. Hence, for $i = 1, 2, 3$,

$$\begin{aligned} &[f_i(x/R, \cdot), [f_i(x/R, \cdot), (H_\chi - \lambda)^2]] \\ &\quad = 2[H_\chi - \lambda, f_i(x/R, \cdot)]^2 + (H_\chi - \lambda)[f_i(x/R, \cdot), [f_i(x/R, \cdot), H_\chi - \lambda]] \\ &\qquad\qquad + [f_i(x/R, \cdot), [f_i(x/R, \cdot), H_\chi - \lambda]](H_\chi - \lambda) \\ &\quad = \tfrac{1}{2}\bigl(\Delta f_i(x/R, \cdot) + 2(\nabla f_i(x/R, \cdot)) \cdot \nabla\bigr)^2 - (H_\chi - \lambda)|\nabla f_i(x/R, \cdot)|^2 - |\nabla f_i(x/R, \cdot)|^2 (H_\chi - \lambda). \end{aligned}$$

Note also that there exists $C \in \mathbb{R}^+$ such that $|\nabla f_i(x/R)| \le C/R$, and $|\Delta f_i(x/R)| \le C/R^2$. Therefore there exists a positive constant C_r such that, for all $n \in \mathbb{N}^*$,

$$\left|\frac{1}{2}\sum_{i=1}^{3}\langle \phi_n, [f_i(x/R,\cdot), [f_i(x/R,\cdot), (H_\chi - \lambda)^2]]\phi_n\rangle\right| \le C_r \frac{\|\phi_n\|^2_{H^2(\mathbb{R}^3)}}{R}. \tag{A-14}$$

On the other hand, note that $f_3(x/R,\cdot)V_\chi$ tends to 0 in all directions. Hence, for every fixed $R \in \mathbb{R}^+$,

$$\begin{aligned}&\left|\|V_\chi f_3(x/R,\cdot)\phi_n\|^2_{L^2} + 2\Re\left\langle V_\chi f_3(x/R,\cdot)\phi_n, \left(-\tfrac{1}{2}\Delta - \lambda\right) f_3(x/R,\cdot)\phi_n\right\rangle_{L^2}\right| \\ &\qquad \le \|V_\chi f_3(x/R,\cdot)\phi_n\|^2_{L^2} + 2(1+|\lambda|)\|V_\chi f_3(x/R,\cdot)\phi_n\|_{L^2}\|\phi_n\|_{H^2(\mathbb{R}^3)} \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

In view of the above formula as well as (A-14), there exists a sequence $R_n \to +\infty$ such that the last two terms of (A-13) tends to 0. In particular, this implies

$$\|(H_{\mathrm{per},L}-\lambda)f_1(x/R_n,\cdot)\phi_n\|^2_{L^2}+\|(H_{\mathrm{per},R}-\lambda)f_2(x/R_n,\cdot)\phi_n\|^2_{L^2}+\left\|\left(-\tfrac{1}{2}\Delta-\lambda\right)f_3(x/R_n,\cdot)\phi_n\right\|^2_{L^2} \xrightarrow[n\to\infty]{} 0.$$

On the other hand,

$$\text{for all } n \in \mathbb{N}^*, \quad \|f_1(x/R_n,\cdot)\phi_n\|^2_{L^2} + \|f_2(x/R_n,\cdot)\phi_n\|^2_{L^2} + \|f_3(x/R_n,\cdot)\phi_n\|^2_{L^2} \equiv 1.$$

Therefore one of them cannot vanish in the limit; hence λ is in the spectrum of $H_{\mathrm{per},L}$, $H_{\mathrm{per},R}$ or $-\frac{1}{2}\Delta$. This allows us to conclude that

$$\sigma_{\mathrm{ess}}(H_\chi) \subseteq \sigma_{\mathrm{ess}}(H_{\mathrm{per},L}) \cup \sigma_{\mathrm{ess}}(H_{\mathrm{per},R}). \tag{A-15}$$

By gathering (A-11) and (A-15) we conclude that $\sigma_{\mathrm{ess}}(H_\chi) \equiv \sigma_{\mathrm{ess}}(H_{\mathrm{per},L}) \cup \sigma_{\mathrm{ess}}(H_{\mathrm{per},R})$. In particular, $\sigma_{\mathrm{ess}}(H_\chi)$ is independent of the function $\chi \in \mathcal{X}$.

A4. ***Proof of Lemma 4.5.*** First of all the following lemma will be useful.

Lemma A.8. *Consider a self-adjoint operator $H = -\Delta + V$ defined on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^2)$ and $V \in L^\infty(\mathbb{R}^3)$. For $E \in \mathbb{R}\backslash\sigma(H)$ denote by $\gamma = \mathbb{1}_{(-\infty,E]}(H)$. Then for any $a, b \in \mathbb{R}$, the operator $(1-\Delta)^a\gamma(1-\Delta)^b$ is bounded. Moreover, if $\gamma \in \mathfrak{S}_k$ for some $k \ge 1$, then $(1-\Delta)^a\gamma(1-\Delta)^b \in \mathfrak{S}_k$.*

Proof. Much as in Lemma 4.4 it can be shown that for any $\zeta \in \mathbb{R}\backslash\sigma(H)$ the operator $(\zeta - H)^{-a}(1-\Delta)^a$ and its inverse are bounded. Fix $\delta > 0$ and define $\lambda_0 := -\|V\|_{L^\infty} - \delta$. Then $\lambda_0 \notin \sigma(H)$. By writing $\gamma = \gamma^2$, there exists a positive constant C such that

$$\|(1-\Delta)^a\gamma(1-\Delta)^b\|_{\mathfrak{S}_k} \le C\|(\lambda_0 - H)^a\gamma(\lambda_0 - H)^b\|_{\mathfrak{S}_k} = C\|\gamma^2(\lambda_0 - H)^{a+b}\|_{\mathfrak{S}_k} < +\infty,$$

as $\gamma \in \mathfrak{S}_k$ and $\gamma(\lambda_0 - H)^{a+b}$ is a bounded operator. The proof of the boundedness in operator norm follows the same lines. □

Let us prove the statement for $\chi\gamma_{\mathrm{per},L}(\zeta - H_{\mathrm{per},L})^{-1}V_d$, the proof of the bound of $V_d\gamma_\chi(\zeta - H_\chi)^{-1}\chi$ follows similar arguments. Fix $R > 0$. Recall that $\mathfrak{B}_R$ is the ball in $\mathbb{R}^3$ centered at 0 with radius R. Denote by φ_R the characteristic function of $\mathfrak{B}_R$. For any $R > 0$, by the Kato–Seiler–Simon inequality (4-1) and the boundedness of $(1-\Delta)(\zeta - H_{\mathrm{per},L})^{-1}$ it is easy to see that $\varphi_R\chi(\zeta - H_{\mathrm{per},L})^{-1}$ and $(\zeta - H_{\mathrm{per},L})^{-1}V_d\varphi_R$ belong to $\mathfrak{S}_2$. The operator $\gamma_{\mathrm{per},L}(\zeta - H_{\mathrm{per},L})^m$ is bounded for any $m \in \mathbb{R}$ in view of Lemma 4.5. Therefore

$$\varphi_R\left(\chi\frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}}V_d\right)\varphi_R = \left(\varphi_R\chi\frac{1}{\zeta - H_{\mathrm{per},L}}\right)\gamma_{\mathrm{per},L}(\zeta - H_{\mathrm{per},L})\left(\frac{1}{\zeta - H_{\mathrm{per},L}}V_d\varphi_R\right) \in \mathfrak{S}_1.$$

Let us first prove that for any $1 \le p \le 2$, there exists a positive constant $d_{p,1}$ only depending on p such that, for any $R > 0$,

$$\left\|\varphi_R\chi\frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}}V_d\varphi_R\right\|_{\mathfrak{S}_p} \le d_{p,1}\|V_d\varphi_R^2\chi\|_{L^p(\mathbb{R}^3)}. \tag{A-16}$$

We first prove (A-16) for $p = 1$ and $p = 2$, and conclude by an interpolation argument for $1 \le p \le 2$. Consider $p = 1$. By the cyclicity of the trace and the Kato–Seiler–Simon inequality (4-1),

$$\begin{aligned}\left\|\varphi_R\chi\frac{\gamma_{\mathrm{per},L}}{\zeta-H_{\mathrm{per},L}}V_d\varphi_R\right\|_{\mathfrak{S}_1} &= \left\|\varphi_R\chi\frac{1}{\zeta-H_{\mathrm{per},L}}\gamma_{\mathrm{per},L}(\zeta-H_{\mathrm{per},L})\frac{1}{\zeta-H_{\mathrm{per},L}}V_d\varphi_R\right\|_{\mathfrak{S}_1}\\ &= \left\|\gamma_{\mathrm{per},L}(\zeta-H_{\mathrm{per},L})\frac{1}{\zeta-H_{\mathrm{per},L}}V_d\varphi_R^2\chi\frac{1}{\zeta-H_{\mathrm{per},L}}\right\|_{\mathfrak{S}_1}\\ &\le c\left\|\frac{1}{|1-\Delta|}|V_d\varphi_R^2\chi|\frac{1}{|1-\Delta|}\right\|_{\mathfrak{S}_1} = c\left\|\frac{1}{|1-\Delta|}|V_d\varphi_R^2\chi|^{1/2}\right\|_{\mathfrak{S}_2}^2 \le d_{1,1}\|V_d\varphi_R^2\chi\|_{L^1}.\end{aligned}$$

Let us next prove (A-16) for $p = 2$. Use again the cyclicity of the trace and the Kato–Seiler–Simon inequality (4-1),

$$\begin{aligned}\left\|\varphi_R\chi\frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}}V_d\varphi_R\right\|_{\mathfrak{S}_2}^2 &= \left\|V_d\varphi_R\frac{\gamma_{\mathrm{per},L}}{\bar{\zeta} - H_{\mathrm{per},L}}\varphi_R^2\chi^2\frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}}V_d\varphi_R\right\|_{\mathfrak{S}_1}\\ &= \left\|\frac{\gamma_{\mathrm{per},L}}{\bar{\zeta} - H_{\mathrm{per},L}}\varphi_R^2\chi^2\frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}}V_d^2\varphi_R^2\right\|_{\mathfrak{S}_1}\\ &\le c'\left\|\varphi_R^2\chi^2\frac{1}{\zeta - H_{\mathrm{per},L}}\gamma_{\mathrm{per},L}(\zeta - H_{\mathrm{per},L})\frac{1}{\zeta - H_{\mathrm{per},L}}V_d^2\varphi_R^2\right\|_{\mathfrak{S}_1}\\ &= c'\left\|\gamma_{\mathrm{per},L}(\zeta - H_{\mathrm{per},L})\frac{1}{\zeta - H_{\mathrm{per},L}}V_d^2\varphi_R^4\chi^2\frac{1}{\zeta - H_{\mathrm{per},L}}\right\|_{\mathfrak{S}_1}\\ &\le c''\left\|\frac{1}{\zeta - H_{\mathrm{per},L}}V_d^2\varphi_R^4\chi^2\frac{1}{\zeta - H_{\mathrm{per},L}}\right\|_{\mathfrak{S}_1}\\ &= c''\left\||V_d\varphi_R^2\chi|\frac{1}{\zeta - H_{\mathrm{per},L}}\right\|_{\mathfrak{S}_2}^2 \le d_{2,1}^2\|V_d\varphi_R^2\chi\|_{L^2}^2.\end{aligned}$$

By the interpolation arguments we can conclude (A-16) for $1 \le p \le 2$. Note that for $1 < p \le 2$ the following uniform bound holds:

$$\left\|\varphi_R\chi\frac{\gamma_{\mathrm{per},L}}{\zeta - H_{\mathrm{per},L}}V_d\varphi_R\right\|_{\mathfrak{S}_p} \le d_{p,1}\|V_d\varphi_R^2\chi\|_{L^p(\mathbb{R}^3)} \le d_{p,1}\|V_d\chi\|_{L^p(\mathbb{R}^3)}.$$

By passing the limit $R \to +\infty$ we can conclude the proof.

Acknowledgements

I would like to express my deep gratitude to Éric Cancès and Gabriel Stoltz for many useful discussions and advice for the article, as well as their critical readings of the manuscript. I would also like to thank the anonymous referees for their careful reading and appropriate suggestions of the manuscript, especially by pointing out simpler proofs of Theorem 2.6 and Proposition 3.1.

References

[Aerts 1960] E. Aerts, "Surface states of one-dimensional crystals, III", *Physica* **26**:12 (1960), 1063–1072. MR Zbl

[Anantharaman and Cancès 2009] A. Anantharaman and É. Cancès, "Existence of minimizers for Kohn–Sham models in quantum chemistry", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **26**:6 (2009), 2425–2455. MR Zbl

[Avila et al. 2013] J. C. Avila, H. Schulz-Baldes, and C. Villegas-Blas, "Topological invariants of edge states for periodic two-dimensional models", *Math. Phys. Anal. Geom.* **16**:2 (2013), 137–170. MR Zbl

[Baugher et al. 2014] B. W. H. Baugher, H. O. H. Churchill, Y. Yang, and P. Jarillo-Herrero, "Optoelectronic devices based on electrically tunable p-n diodes in a monolayer dichalcogenide", *Nature Nanotech.* **9** (2014), 262–267.

[Blanc and Le Bris 2000] X. Blanc and C. Le Bris, "Thomas–Fermi type theories for polymers and thin films", *Adv. Differential Equations* **5**:7-9 (2000), 977–1032. MR Zbl

[Blanc and Le Bris 2002] X. Blanc and C. Le Bris, "Periodicity of the infinite-volume ground state of a one-dimensional quantum model", *Nonlinear Anal.* **48**:6 (2002), 791–803. MR Zbl

[Blanc et al. 2003] X. Blanc, C. Le Bris, and P.-L. Lions, "A definition of the ground state energy for systems composed of infinitely many particles", *Comm. Partial Differential Equations* **28**:1-2 (2003), 439–475. MR Zbl

[Bruneau et al. 2015] L. Bruneau, V. Jakšić, Y. Last, and C.-A. Pillet, "Landauer–Büttiker and Thouless conductance", *Comm. Math. Phys.* **338**:1 (2015), 347–366. MR Zbl

[Bruneau et al. 2016a] L. Bruneau, V. Jakšić, Y. Last, and C.-A. Pillet, "Conductance and absolutely continuous spectrum of 1D samples", *Comm. Math. Phys.* **344**:3 (2016), 959–981. MR Zbl

[Bruneau et al. 2016b] L. Bruneau, V. Jakšić, Y. Last, and C.-A. Pillet, "Crystalline conductance and absolutely continuous spectrum of 1D samples", *Lett. Math. Phys.* **106**:6 (2016), 787–797. MR Zbl

[Brus 2010] L. Brus, "Commentary: carbon nanotubes, CdSe nanocrystals, and electron-electron interaction", *Nano Lett.* **10**:2 (2010), 363–365.

[Brus 2014] L. Brus, "Size, dimensionality, and strong electron correlation in nanoscience", *Accounts Chem. Res.* **47**:10 (2014), 2951–2959.

[Büttiker et al. 1985] M. Büttiker, Y. Imry, R. Landauer, and S. Pinhas, "Generalized many-channel conductance formula with application to small rings", *Phys. Rev. B* **31**:10 (1985), 6207–6215.

[Cancès and Stoltz 2012] É. Cancès and G. Stoltz, "A mathematical formulation of the random phase approximation for crystals", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **29**:6 (2012), 887–925. MR Zbl

[Cancès et al. 2008] É. Cancès, A. Deleurence, and M. Lewin, "A new approach to the modeling of local defects in crystals: the reduced Hartree–Fock case", *Comm. Math. Phys.* **281**:1 (2008), 129–177. MR Zbl

[Cancès et al. 2009] É. Cancès, G. Stoltz, G. E. Scuseria, V. N. Staroverov, and E. R. Davidson, "Local exchange potentials for electronic structure calculations", *Math. in Action* **2**:1 (2009), 1–42. MR Zbl

[Cancès et al. 2013] É. Cancès, S. Lahbabi, and M. Lewin, "Mean-field models for disordered crystals", *J. Math. Pures Appl.* (9) **100**:2 (2013), 241–274. MR Zbl

[Cancès et al. 2020] É. Cancès, L.-L. Cao, and G. Stoltz, "A reduced Hartree–Fock model of slice-like defects in the Fermi sea", *Nonlinearity* **33**:1 (2020), 156–195. MR Zbl

[Cao 2019a] L. Cao, "Analyse mathématique du transport thermo-électronique dans les solides désordonnés", preprint, 2019.

[Cao 2019b] L.-L. Cao, "Mean-field stability for the junction of quasi 1D systems with Coulomb interactions", preprint, 2019. arXiv

[Catto et al. 1998] I. Catto, C. Le Bris, and P.-L. Lions, *The mathematical theory of thermodynamic limits: Thomas–Fermi type models*, Oxford Univ. Press, 1998. MR Zbl

[Catto et al. 2001] I. Catto, C. Le Bris, and P.-L. Lions, "On the thermodynamic limit for Hartree–Fock type models", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **18**:6 (2001), 687–760. MR Zbl

[Combes and Thomas 1973] J. M. Combes and L. Thomas, "Asymptotic behaviour of eigenfunctions for multiparticle Schrödinger operators", *Comm. Math. Phys.* **34**:4 (1973), 251–270. MR Zbl

[Cornean et al. 2008] H. D. Cornean, P. Duclos, G. Nenciu, and R. Purice, "Adiabatically switched-on electrical bias and the Landauer–Büttiker formula", *J. Math. Phys.* **49**:10 (2008), art. id. 102106. MR Zbl

[Cornean et al. 2012] H. D. Cornean, P. Duclos, and R. Purice, "Adiabatic non-equilibrium steady states in the partition free approach", *Ann. Henri Poincaré* **13**:4 (2012), 827–856. MR Zbl

[Deift and Simon 1976] P. Deift and B. Simon, "On the decoupling of finite singularities from the question of asymptotic completeness in two body quantum systems", *J. Funct. Anal.* **23**:3 (1976), 218–238. MR Zbl

[Dohnal et al. 2011] T. Dohnal, M. Plum, and W. Reichel, "Surface gap soliton ground states for the nonlinear Schrödinger equation", *Comm. Math. Phys.* **308**:2 (2011), 511–542. MR Zbl

[Duffin 1971] R. J. Duffin, "Yukawan potential theory", *J. Math. Anal. Appl.* **35** (1971), 105–130. MR Zbl

[Exner and Frank 2007] P. Exner and R. L. Frank, "Absolute continuity of the spectrum for periodically modulated leaky wires in $\mathbb{R}^3$", *Ann. Henri Poincaré* **8**:2 (2007), 241–263. MR Zbl

[Frank 2003] R. L. Frank, "On the scattering theory of the Laplacian with a periodic boundary condition, I: Existence of wave operators", *Doc. Math.* **8** (2003), 547–565. MR Zbl

[Frank and Shterenberg 2004] R. L. Frank and R. G. Shterenberg, "On the scattering theory of the Laplacian with a periodic boundary condition, II: Additional channels of scattering", *Doc. Math.* **9** (2004), 57–77. MR Zbl

[Frank et al. 2011] R. L. Frank, M. Lewin, E. H. Lieb, and R. Seiringer, "Energy cost to make a hole in the Fermi sea", *Phys. Rev. Lett.* **106**:15 (2011), art. id. 150402.

[Frank et al. 2013] R. L. Frank, M. Lewin, E. H. Lieb, and R. Seiringer, "A positive density analogue of the Lieb–Thirring inequality", *Duke Math. J.* **162**:3 (2013), 435–495. MR Zbl

[Germinet and Klein 2003] F. Germinet and A. Klein, "Operator kernel estimates for functions of generalized Schrödinger operators", *Proc. Amer. Math. Soc.* **131**:3 (2003), 911–920. MR Zbl

[Hainzl et al. 2005a] C. Hainzl, M. Lewin, and É. Séré, "Existence of a stable polarized vacuum in the Bogoliubov–Dirac–Fock approximation", *Comm. Math. Phys.* **257**:3 (2005), 515–562. MR Zbl

[Hainzl et al. 2005b] C. Hainzl, M. Lewin, and É. Séré, "Self-consistent solution for the polarized vacuum in a no-photon QED model", *J. Phys. A* **38**:20 (2005), 4483–4499. MR Zbl

[Hainzl et al. 2007] C. Hainzl, M. Lewin, and J. P. Solovej, "The mean-field approximation in quantum electrodynamics: the no-photon case", *Comm. Pure Appl. Math.* **60**:4 (2007), 546–596. MR Zbl

[Hainzl et al. 2009] C. Hainzl, M. Lewin, and É. Séré, "Existence of atoms and molecules in the mean-field approximation of no-photon quantum electrodynamics", *Arch. Ration. Mech. Anal.* **192**:3 (2009), 453–499. MR Zbl

[Hasan and Kane 2010] M. Z. Hasan and C. L. Kane, "Colloquium: topological insulators", *Rev. Modern Phys.* **82**:4 (2010), 3045–3067.

[Hatsugai 1993] Y. Hatsugai, "Edge states in the integer quantum Hall effect and the Riemann surface of the Bloch function", *Phys. Rev. B* **48**:16 (1993), art. id. 11851.

[Helffer 2013] B. Helffer, *Spectral theory and its applications*, Cambridge Stud. Adv. Math. **139**, Cambridge Univ. Press, 2013. MR Zbl

[Hempel and Kohlmann 2011] R. Hempel and M. Kohlmann, "A variational approach to dislocation problems for periodic Schrödinger operators", *J. Math. Anal. Appl.* **381**:1 (2011), 166–178. MR Zbl

[Hempel et al. 2015] R. Hempel, M. Kohlmann, M. Stautz, and J. Voigt, "Bound states for nano-tubes with a dislocation", *J. Math. Anal. Appl.* **431**:1 (2015), 202–227. MR Zbl

[Hislop and Sigal 1996] P. D. Hislop and I. M. Sigal, *Introduction to spectral theory: with applications to Schrödinger operators*, Appl. Math. Sci. **113**, Springer, 1996. MR Zbl

[Hoffmann-Ostenhof and Hoffmann-Ostenhof 1977] M. Hoffmann-Ostenhof and T. Hoffmann-Ostenhof, "'Schrödinger inequalities' and asymptotic behavior of the electron density of atoms and molecules", *Phys. Rev. A* (3) **16**:5 (1977), 1782–1785. MR

[Ismagilov 1961] R. S. Ismagilov, "Conditions for the semiboundedness and discreteness of the spectrum in the case of one-dimensional differential operators", *Dokl. Akad. Nauk SSSR* **140** (1961), 33–36. In Russian; translated in *Soviet Math. Dokl.* **2** (1961), 1137–1140. MR Zbl

[Kato 1966] T. Kato, *Perturbation theory for linear operators*, Grundlehren der Math. Wissenschaften **132**, Springer, 1966. MR Zbl

[Klopp 1995] F. Klopp, "An asymptotic expansion for the density of states of a random Schrödinger operator with Bernoulli disorder", *Random Oper. Stochastic Equations* **3**:4 (1995), 315–331. MR Zbl

[Kohn and Sham 1965] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects", *Phys. Rev.* (2) **140**:4A (1965), 1133–1138. MR

[Korotyaev 2000] E. Korotyaev, "Lattice dislocations in a 1-dimensional model", *Comm. Math. Phys.* **213**:2 (2000), 471–489. MR Zbl

[Korotyaev 2005] E. Korotyaev, "Schrödinger operator with a junction of two 1-dimensional periodic potentials", *Asymptot. Anal.* **45**:1-2 (2005), 73–97. MR Zbl

[Lahbabi 2014] S. Lahbabi, "The reduced Hartree–Fock model for short-range quantum crystals with nonlocal defects", *Ann. Henri Poincaré* **15**:7 (2014), 1403–1452. MR Zbl

[Laird et al. 2015] E. A. Laird, F. Kuemmeth, G. A. Steele, K. Grove-Rasmussen, J. Nygård, K. Flensberg, and L. P. Kouwenhoven, "Quantum transport in carbon nanotubes", *Rev. Modern Phys.* **87**:3 (2015), 703–764. MR

[Landauer 1970] R. Landauer, "Electrical resistance of disordered one-dimensional lattices", *Philos. Mag.* (8) **21**:172 (1970), 863–867.

[Lee et al. 2004] J. U. Lee, P. P. Gipp, and C. M. Heller, "Carbon nanotube p-n junction diodes", *Appl. Phys. Lett.* **85** (2004), 145–147.

[Léonard and Tersoff 1999] F. Léonard and J. Tersoff, "Novel length scales in nanotube devices", *Phys. Rev. Lett.* **83**:24 (1999), 5174–5177. Correction in **89**:17 (2002), art. id. 179902.

[Lewin 2009] M. Lewin, *Large quantum systems: a mathematical and numerical perspective*, habilitation à diriger des recherches, University of Cergy-Pontoise, 2009, available at https://tel.archives-ouvertes.fr/tel-00394205/.

[Lewin et al. 2020] M. Lewin, E. H. Lieb, and R. Seiringer, "The local density approximation in density functional theory", *Pure Appl. Anal.* **2**:1 (2020), 35–73. MR Zbl

[Lieb and Loss 2001] E. H. Lieb and M. Loss, *Analysis*, 2nd ed., Grad. Studies in Math. **14**, Amer. Math. Soc., Providence, RI, 2001. MR Zbl

[Lieb and Simon 1977] E. H. Lieb and B. Simon, "The Thomas–Fermi theory of atoms, molecules and solids", *Adv. Math.* **23**:1 (1977), 22–116. MR Zbl

[Lions 1987] P.-L. Lions, "Solutions of Hartree–Fock equations for Coulomb systems", *Comm. Math. Phys.* **109**:1 (1987), 33–97. MR Zbl

[Morgan 1979] J. D. Morgan, III, "Schrödinger operators whose potentials have separated singularities", *J. Operator Theory* **1**:1 (1979), 109–115. MR Zbl

[Morgan and Simon 1980] J. D. Morgan, III and B. Simon, "Behavior of molecular potential energy curves for large nuclear separations", *Int. J. Quantum Chem.* **17**:6 (1980), 1143–1166.

[Nitzan and Ratner 2003] A. Nitzan and M. A. Ratner, "Electron transport in molecular wire junctions", *Science* **300**:5624 (2003), 1384–1389.

[Pospischil et al. 2014] A. Pospischil, M. M. Furchi, and T. Mueller, "Solar-energy conversion and light emission in an atomic monolayer p-n diode", *Nature Nanotech.* **9** (2014), 257–261.

[Reed and Simon 1975] M. Reed and B. Simon, *Methods of modern mathematical physics, II: Fourier analysis, self-adjointness*, Academic Press, New York, 1975. MR Zbl

[Reed and Simon 1978] M. Reed and B. Simon, *Methods of modern mathematical physics, IV: Analysis of operators*, Academic Press, New York, 1978. MR Zbl

[Seiler and Simon 1975] E. Seiler and B. Simon, "Bounds in the Yukawa2 quantum field theory: upper bound on the pressure, Hamiltonian bound and linear lower bound", *Comm. Math. Phys.* **45**:2 (1975), 99–114. MR

[Shale and Stinespring 1965] D. Shale and W. F. Stinespring, "Spinor representations of infinite orthogonal groups", *J. Math. Mech.* **14** (1965), 315–322. MR Zbl

[Shockley 1939] W. Shockley, "On the surface states associated with a periodic potential", *Phys. Rev.* **56**:4 (1939), 317–323. Zbl

[Simon 1983] B. Simon, "Semiclassical analysis of low lying eigenvalues, I: Nondegenerate minima: asymptotic expansions", *Ann. Inst. H. Poincaré Sect. A* (*N.S.*) **38**:3 (1983), 295–308. MR Zbl

[Solovej 1991] J. P. Solovej, "Proof of the ionization conjecture in a reduced Hartree–Fock model", *Invent. Math.* **104**:2 (1991), 291–311. MR Zbl

[Solovej 2005] J. P. Solovej, "Many-body quantum mechanics", lecture notes, Københavns Univ., 2005, available at http://www.math.ku.dk/~solovej/mbnotes.

[Thomas 1973] L. E. Thomas, "Time dependent approach to scattering from impurities in a crystal", *Comm. Math. Phys.* **33**:4 (1973), 335–343. MR

[Wang et al. 2008] J.-S. Wang, J. Wang, and J. T. Lü, "Quantum thermal transport in nanostructures", *Eur. Phys. J. B* **62** (2008), 381–404.

Received 24 Apr 2019. Revised 26 Mar 2020. Accepted 10 May 2020.

LING-LING CAO: caolingling0922@gmail.com
Université Paris-Est Marne-la-Vallée, CERMICS (ENPC), Marne-la-Vallée, France

PURE and APPLIED
ANALYSIS

Vol. 2, No. 3, 2020
https://doi.org/10.2140/paa.2020.2.581

OPTIMAL TRANSPORT AND BARYCENTERS FOR DENDRITIC MEASURES

YOUNG-HEON KIM, BRENDAN PASS AND DAVID J. SCHNEIDER

We introduce and study a variant of the Wasserstein distance on the space of probability measures, specially designed to deal with measures whose support has a dendritic, or tree-like structure with a particular direction of orientation. Our motivation is the comparison of and interpolation between plants' root systems. We characterize barycenters with respect to this metric, and establish that the interpolations of root-like measures, using this new metric, are also root-like, in a certain sense; this property fails for conventional Wasserstein barycenters. We also establish geodesic convexity with respect to this metric for a variety of functionals, some of which we expect to have biological importance.

1. Introduction

We introduce a new metric on the space of probability measures, the layerwise-Wasserstein distance. The motivation for this work is the need for a sound mathematical framework for describing the structure and diversity of dendritic structures in anisotropic environments. In particular, we are interested in the macroscopic structure of plant root systems developing under the influence of gravity and the stratification of chemical constituents, texture and microbial activity characteristic of soils. This biophysical context can be readily translated into mathematical terms. Plant tissues are composed of cells that physically partition $\mathbb{R}^3$ into two connected components – the "inside" and "outside". The resulting structure roughly corresponds to a CW complex (see, e.g., [Hatcher 2002]) describing the topology of the plant. Ignoring complex features present at microscopic scales, the external surface can be viewed as a smooth, connected two-dimensional manifold with genus zero embedded in $\mathbb{R}^3$. Computational representations of these external surfaces can be reconstructed using standard methods of optical, X-ray and neutron tomography.

This idealization misses two essential points: the above- and below-ground portions of plants display intricate structural forms that are remarkably resistant to quantitative analysis, and the form and function of these complicated structures are intimately related to the anisotropic environment in which they develop. The first condition implies the need to handle arbitrarily complicated distributions of mass in space subject to very modest restrictions on the behavior of the surface, while the second suggests the need to handle preferred directions in space.

Natural challenges include quantifying the difference between two or more roots, summarizing or describing the typical structure of a family of root systems (for instance, the roots of several genetically identical plants, grown in nearly identical environments, which often exhibit considerable variation in their structure) in a succinct way, quantifying the variation within that family and comparing the

MSC2010: primary 49K99; secondary 92C80.
Keywords: optimal transport, dendritic measures, layerwise-Wasserstein distance, plant root architecture.

structure exhibited by one family to another. A typical approach to these problems is to compute a family of *phenotypes* for each system (including, for example, total root length, rooting depth, and various topological invariants, such as the Horton–Strahler index) and compare and average among them; see, for instance, [Clark et al. 2011; 2013; Famoso et al. 2010; 2011; Piñeros et al. 2016]. Though this has met with some success in distinguishing between particular choices of root systems, it is not generally clear which phenotypes are most useful for this purpose, and the choice in different applications is often done in an ad hoc way. For virtually any collection of phenotypes, it is not hard to come up with drastically different root shapes sharing the same phenotypes.

Our approach focuses on roots as mass distributions in $\mathbb{R}^3$ where the vertical and horizontal directions have distinct roles (roots which are related by a rotation about the vertical axis are considered identical). Natural mathematical goals include constructing a metric between root shapes reflecting both their downward-pointing dendritic topology as well as the distances and sizes in the underlying space,[1] and producing a representative of a family of root systems which captures the average or typical structure of the family. After normalization for overall mass, root systems can be modeled as probability distributions; the Wasserstein distance from optimal transport [Villani 2003; 2009; Santambrogio 2015] is then one candidate for such a metric, and Wasserstein barycenters (Fréchet means with respect to this metric; see [Agueh and Carlier 2011]) are a corresponding candidate for a representative of a family. While this metric has proved fruitful in related problems involving comparing and averaging among shapes (image processing, for instance), we demonstrate in this paper that it is not ideally adapted to the downward dendritic structure prominent among root systems, in large part because optimal matchings don't generally exhibit monotonicity in the distinguished, vertical direction. While it is possible to incorporate vertical stratification in the usual definition of Wasserstein distance by penalizing transport in the vertical direction, the practical application of this formalism is limited by computational requirements. We propose a simple alternative based on a related metric, the layerwise-Wasserstein metric, derived from a variant of optimal transport in which monotonicity in the distinguished vertical direction is guaranteed; see Definition 2.2. The metric barycenter arising from this new metric is a natural candidate for a representative of a family of root systems. Furthermore, we suspect this distance may play a role in other applied problems featuring both tree-like and geometric structures (blood vessels in biology, river systems in topography, etc.). Our primary present goal is to develop the mathematical properties of the layerwise-Wasserstein distance and its interpolants, while the biological and methodological applications will be developed in subsequent work. However, we keep the motivating applications in mind as we go, and focus on properties of root systems that have potential biological relevance.

It is common in biology to model root systems by their *skeletons*, in which three-dimensional limbs are replaced by approximating one-dimensional curves [Bucksch 2014]; these skeletons retain the dendritic, or tree-like, structure of the root, but strip away its thickness (which is less crucial in some applications). As a corresponding mathematical object we introduce *skeletal measures*, which are essentially mass distributions supported on these skeletal structures; see Section 3. This gives a useful framework for

[1] Ideally, the metric should detect geometric differences, between, for instance, a short limb and a long one, as well as topological differences, between say, a forked limb and a straight one.

studying the topological properties of roots and their interpolations, while avoiding difficulties that arise when dealing with their (more realistic) three-dimensional structure.

When building interpolants to use as representatives of families of roots, a desirable property is that the dendritic structure is preserved: given several root systems, does their metric barycenter look like a root? We are able to give a fairly satisfactory affirmative answer to this question for skeletal root systems, using our layerwise-Wasserstein distance as the metric; see Theorem 3.6. On the other hand, we exhibit examples illustrating that when the conventional Wasserstein distance is used, interpolants of root systems may not resemble root systems at all; more precisely, we show that the Wasserstein barycenter of several skeletal roots can have high-dimensional support, so that the dendritic structure is broken; see Section 3B. We also establish comparisons between the total root length (essentially the one-dimensional Hausdorff measure of the support) of several skeletal root measures and their layerwise-Wasserstein barycenter; see Proposition 3.14. This type of result is impossible in general with the Wasserstein barycenter, as the support may be more than one-dimensional.

Aside from being natural for certain applications, the layerwise-Wasserstein distance also has computational advantages over its classical Wasserstein counterpart in certain situations, as the sorting in the distinguished direction is monotone, and so optimization problems arise only in spaces of codimension 1. In $\mathbb{R}^2$, for instance, the layerwise-Wasserstein distance essentially corresponds to the Knothe–Rosenblatt rearrangement [Knothe 1957; Rosenblatt 1952], which can be computed much more easily than the two-dimensional Wasserstein distance; previous connections between the Knothe–Rosenblatt rearrangement and optimal transport have been established in [Carlier et al. 2010] (where the Knothe–Rosenblatt rearrangement was shown to arise as a limit of optimal transport maps with anisotropic costs) and in [Backhoff et al. 2017] (where an interpolation of probability measures using the Knothe–Rosenblatt rearrangement was introduced).[2] More generally, the layerwise-Wasserstein distance is a special instance of the Monge–Knothe maps recently introduced in [Muzellec and Cuturi 2019]; in that work, properties of the corresponding metric, including interpolation between measures and convexity were not studied.

We also note that our layerwise-Wasserstein distance is similar in spirit to the Radon–Wasserstein distance found in [Bonneel et al. 2015], as both approaches involve disintegrating the measures and transporting their fibers. The difference lies in how the measures are disintegrated; we disintegrate with respect to a distinguished, vertical variable on the underlying space (which is natural in the applications we have in mind), whereas the disintegration in [Bonneel et al. 2015] is done with respect to Radon transformed variables.

In addition, it is worth commenting briefly on the relationship between this work and another recent series of papers relating optimal transport to plant root shapes [Bressan and Sun 2018; Bressan et al. 2020]. In those works, the objective is to identify and characterize root (and tree) shapes which optimize certain functionals, modeling absorption of nutrients and sunlight and the cost (via ramified optimal transport) of returning those nutrients to the base of plant, whereas our goal is to differentiate and interpolate between various root systems.

[2] Interpolating between two-dimensional measures is in fact not merely a mathematical simplification or toy model, but has actual agricultural applications, since experiments are sometimes done growing plants between two panes of glass, placed very close together, resulting in essentially two-dimensional root shapes.

The manuscript is organized as follows. In Section 2 we introduce the layerwise Wasserstein distance and barycenters, and establish some basic properties. Section 3 focuses on skeletal measures, while Section 4 is devoted to layerwise displacement interpolation and convexity.

2. Layerwise-Wasserstein distance

Let $M^+(X)$, denote the space of finite, nonnegative Borel measures with positive total mass (i.e., $\mu(X) > 0$ for each $\mu \in M^+(X)$) on a metric space X equipped with the weak-$*$ topology. Let $P(X)$ denote the space of finite, nonnegative Borel probability measures on a metric space X equipped with the weak-$*$ topology. Consider $M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ and let $M^+_{\text{ac}}(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ be its subset consisting of absolutely continuous measures (with respect to Lebesgue); let $P_{\text{ac}}(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ be the corresponding subset of probability measures. For $\mu \in M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$, let μ^V be its vertical marginal, defined by

$$\int_{\mathbb{R}_{\geq 0}} f(z)\mu^V(dz) = \int_{\mathbb{R}^d \times \mathbb{R}_{\geq 0}} f(z)\mu(dx, dy) \quad \text{for all } f \in C(\mathbb{R}_{\geq 0}).$$

Note that $|\mu^V| = |\mu|$, where $|\mu|$ denotes the total mass of μ. The following vertical rescaling of the measures in $M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ is a key step in our construction of the Wasserstein-type distance that uses the distinguished coordinate $\mathbb{R}_{\geq 0}$. Note also that measures may not necessarily have the same mass, so we also normalize them to be probability measures.

Definition 2.1 (vertical rescaling). Given $\mu \in M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$, we define its *vertically rescaled version*, namely,

$$\tilde{\mu} \in P(\mathbb{R}^d \times [0, 1]),$$

as follows: Let $F_\mu : \mathbb{R}_{\geq 0} \to [0, 1]$ be the cumulative function given by

$$F_\mu(y) = \frac{1}{|\mu|}\mu^V([0, y]).$$

Note that $(F_\mu)_\# \mu^V = |\mu|\mathcal{L}^1$, and F_μ is continuous for absolutely continuous μ^V.

Then, define

$$\tilde{\mu} = \frac{1}{|\mu|}(\text{id} \times F_\mu)_\# \mu,$$

where $\text{id} : \mathbb{R}^d \to \mathbb{R}^d$ is the identity map. Notice that the map $\mu \mapsto \tilde{\mu}$ from $M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ to $P(\mathbb{R}^d \times [0, 1])$ is continuous with respect to the weak-$*$ topology. In particular, this map pushes forward a given $\Omega \in P(M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0}))$ to its vertically rescaled version

$$\widetilde{\Omega} \in P(P(\mathbb{R}^d \times [0, 1])).$$

Note that the mapping F_μ depends on μ only through its vertical marginal, μ^V; we will sometimes abuse notation and write F_{μ^V} instead of F_μ.

This normalization allows us to define a Wasserstein-type distance that uses the disintegration along the vertical line. In the following, W_2^2 denotes the quadratic Wasserstein distance.

Definition 2.2 (layerwise-Wasserstein distance). Given $\mu, \nu \in M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$, define

$$d_{\mathrm{LW}}^2(\mu, \nu) = W_2^2\left(\frac{1}{|\mu^V|}\mu^V, \frac{1}{|\nu^V|}\nu^V\right) + \int_0^1 W_2^2(\tilde{\mu}_l, \tilde{\nu}_l)\, dl, \tag{2-1}$$

where $\tilde{\mu}$ and $\tilde{\nu}$ have disintegrations $\tilde{\mu}(dx, dl) = \tilde{\mu}_l(dx)\, dl$, $\tilde{\nu}(dx, dl) = \tilde{\nu}_l(dx)\, dl$ with respect to the Lebesgue measure dl on $[0, 1]$.

Remark 2.3. We note that strictly speaking d_{LW}^2 does not give a distance on $M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$, unless restricted to $P(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$, as different measures may have the same vertical rescaling $((1/|\mu^V|)\mu^V, \tilde{\mu})$; instead, it gives a metric on the set of equivalence classes, under the equivalence relation $\mu \sim \nu$ if $\mu/|\mu| = \nu/|\nu|$. To get a distance on $M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ one may add $(|\mu| - |\nu|)^2$ and consider the metric

$$W_2^2\left(\frac{1}{|\mu^V|}\mu^V, \frac{1}{|\nu^V|}\nu^V\right) + \int_0^1 W_2^2(\tilde{\mu}_l, \tilde{\nu}_l)\, dl + (|\mu| - |\nu|)^2.$$

In the following, however, we stick to (2-1) for simplicity (in fact, in subsequent sections, we restrict our attention entirely to $P(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$).

We now consider the metric barycenter corresponding to the layerwise-Wasserstein distance (2-1),[3] which we define below, and call it the layerwise-Wasserstein barycenter.

Definition 2.4 (layerwise Wasserstein barycenter). For $\Omega \in P(M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0}))$, a *layerwise Wasserstein barycenter* $\mathrm{Bar}^{\mathrm{LW}}(\Omega) \in P(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ is defined as an element of

$$\operatorname*{argmin}_{\mu \in P(\mathbb{R}^d \times \mathbb{R}_{\geq 0})} \int_{M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})} d_{\mathrm{LW}}^2(\mu, \nu)\, d\Omega(\nu).$$

To characterize layerwise-Wasserstein barycenters, we need a little more terminology. Define $\widetilde{\Omega}_l := (\nu \mapsto \tilde{\nu}_l)_\# \Omega$. A Wasserstein barycenter of $\widetilde{\Omega}_l$ is then a minimizer over $P(\mathbb{R}^d)$ of

$$\eta \mapsto \int_{P(\mathbb{R}^d)} W_2^2(\eta, \alpha)\, d\widetilde{\Omega}_l(\alpha) = \int_{M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})} W_2^2(\eta, \tilde{\nu}_l)\, d\Omega(\nu). \tag{2-2}$$

Similarly, defining $\Omega^V := (\nu \mapsto \nu^V/|\nu^V|)_\# \Omega$, a Wasserstein barycenter of Ω^V is a minimizer over $P(\mathbb{R}_{\geq 0})$ of

$$\eta \mapsto \int_{P(\mathbb{R}_{\geq 0})} W_2^2(\eta, \alpha)\, d\Omega^V(\alpha) = \int_{M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0})} W_2^2\left(\eta, \frac{\nu^V}{|\nu^V|}\right) d\Omega(\nu).$$

We then have the following:

Proposition 2.5. *A measure $\mu \in P(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ is a layerwise-Wasserstein barycenter of $\Omega \in P(M^+(\mathbb{R}^d \times \mathbb{R}_{\geq 0}))$ if and only if its vertical marginal μ^V is a Wasserstein barycenter of Ω^V and for almost every layer l, $\tilde{\mu}_l$ is a Wasserstein barycenter of $\widetilde{\Omega}_l$.*

[3] Strictly speaking, given the remark above, the metric barycenter is an equivalence class of measures; we choose as a representative the unique probability measure in a given class.

Proof. By definition, a layerwise Wasserstein barycenter μ must minimize

$$\begin{aligned}\int_{M^+(\mathbb{R}^d\times\mathbb{R}_{\geq 0})}&\Big[W_2^2\Big(\mu^V,\frac{1}{|\nu^V|}\nu^V\Big)+\int_0^1 W_2^2(\tilde{\mu}_l,\tilde{\nu}_l)\,dl\Big]\,d\Omega(\nu)\\ &=\int_{M^+(\mathbb{R}^d\times\mathbb{R}_{\geq 0})}W_2^2\Big(\mu^V,\frac{1}{|\nu^V|}\nu^V\Big)\,d\Omega(\nu)+\int_{M^+(\mathbb{R}^d\times\mathbb{R}_{\geq 0})}\int_0^1 W_2^2(\tilde{\mu}_l,\tilde{\nu}_l)\,dl\,d\Omega(\nu)\\ &=\int_{P(\mathbb{R}_{\geq 0})}W_2^2\Big(\mu^V,\frac{1}{|\nu^V|}\nu^V\Big)\,d\Omega^V(\nu^V)+\int_0^1\int_{P(\mathbb{R}^d)}W_2^2(\tilde{\mu}_l,\tilde{\nu}_l)\,d\widetilde{\Omega}_l(\tilde{\nu}_l)\,dl.\end{aligned}$$

By changing μ^V and $\tilde{\mu}_l$ independently, we see that μ minimizes the last line if and only if its vertical marginal μ^V minimizes the first term and, for almost every l, $\tilde{\mu}_l$ minimizes $\int_{P(\mathbb{R}^d)}W_2^2(\tilde{\mu}_l,\tilde{\nu}_l)\,d\widetilde{\Omega}_l(\tilde{\nu}_l)$; that is, μ^V is a Wasserstein barycenter of Ω^V and $\tilde{\mu}_l$ is a Wasserstein barycenter of $\widetilde{\Omega}_l$. □

The proposition gives a straightforward way to construct layerwise-Wasserstein barycenters; first construct the layers $\tilde{\mu}_l=\mathrm{Bar}^W(\widetilde{\Omega}_l)$ as Wasserstein barycenters of the $\widetilde{\Omega}_l$. Then letting $\mu^V=\mathrm{Bar}^W(\Omega^V)$ be the Wasserstein barycenter of Ω^V the layerwise Wasserstein barycenter $\mu=\mathrm{Bar}^{\mathrm{LW}}(\Omega)$ is defined by

$$d\mu(x,y)=d\tilde{\mu}_{F_{\mu^V}(y)}(x)\,d\mu^V(y).$$

Note that any $\mathrm{Bar}^{\mathrm{LW}}(\Omega)$ is written this way, and is uniquely determined if $\tilde{\mu}_l$ is uniquely determined for a.e. l. In particular, we have:

Corollary 2.6. *For $\Omega\in P(M^+_{\mathrm{ac}}(\mathbb{R}^d\times\mathbb{R}_{\geq 0}))$, there is a unique $\mu=\mathrm{Bar}^{\mathrm{LW}}(\Omega)$.*

Proof. As $\Omega\in P(M^+_{\mathrm{ac}}(\mathbb{R}^d\times\mathbb{R}_{\geq 0}))$, it also holds that $\widetilde{\Omega}_l\in P(P_{\mathrm{ac}}(\mathbb{R}^d))$ for a.e. l. Then uniqueness of $\tilde{\mu}_l$ follows from [Kim and Pass 2017]. □

The rescaled version $\widetilde{\mathrm{Bar}}^{\mathrm{LW}}(\Omega)\in P(\mathbb{R}^d\times[0,1])$ of the layerwise-Wasserstein barycenter $\mathrm{Bar}^{\mathrm{LW}}(\Omega)$ has the disintegration $d\widetilde{\mathrm{Bar}}^{\mathrm{LW}}(\Omega)(x,l)=d\widetilde{\mathrm{Bar}}_l^{\mathrm{LW}}(\Omega)\,dl$, where each $\mathrm{Bar}_l^{\mathrm{LW}}(\Omega)$ is a Wasserstein barycenter of the $\widetilde{\Omega}_l$. The rescaling mapping $F_{\mathrm{Bar}^{\mathrm{LW}}(\Omega)}$ satisfies

$$F_{\mathrm{Bar}^{\mathrm{LW}}(\Omega)}(y)=\Big[\int F_\nu^{-1}\,d\Omega(\nu)\Big]^{-1}(y).\tag{2-3}$$

We note here associativity, in the two-dimensional case, i.e., on $\mathbb{R}\times\mathbb{R}_{\geq 0}$, of the layerwise-Wasserstein barycenter of probability measures $\mu_1,\dots,\mu_N$ with weights $\lambda_1,\dots,\lambda_N$, where $\sum_{i=1}^N\lambda_i=1$ and each $\lambda_i\geq 0$.

Proposition 2.7 (associativity of two-dimensional layerwise-Wasserstein barycenters). *Assume that $d=1$ and $\lambda_1+\lambda_2+\lambda_3=1$. Then*

$$\mathrm{Bar}^{\mathrm{LW}}(\lambda_1\delta_{\mu_1}+\lambda_2\delta_{\mu_2}+\lambda_3\delta_{\mu_3})=\mathrm{Bar}^{\mathrm{LW}}((\lambda_1+\lambda_2)\delta_{\mathrm{Bar}^{\mathrm{LW}}(\delta_{\mu_1}\lambda_1/(\lambda_1+\lambda_2)+\delta_{\mu_2}\lambda_2/(\lambda_1+\lambda_2))}+\lambda_3\delta_{\mu_3}).$$

This proposition is potentially useful in certain computations. For example, when one adds a new sample μ_{N+1} root system to a family of N root systems with a (previously computed) barycenter $\bar{\mu}$, one can find the barycenter of the augmented family by computing the appropriately weighted barycenter of μ_{N+1} and $\bar{\mu}$, rather than the more difficult computation of the barycenter of the new family of $N+1$ systems.

Proof. The result follows immediately from the corresponding result in one dimension for Wasserstein barycenters. □

Remark 2.8. In our motivating application, we only distinguish between root systems up to rotation about the vertical axis; that is, we wish to identify two systems whenever we can transform one system to the other via a rotation fixing y. For actual root systems then, the following distance is relevant:

Definition 2.9 (Horizontally symmetrized layerwise-Wasserstein distance). We define the horizontally symmetrized layerwise-Wasserstein distance $d^2_{\mathrm{LW,symm}}(\mu,\nu)$ between μ and ν by

$$d^2_{\mathrm{LW,symm}}(\mu,\nu)=\min_{R\in\mathrm{SO}(d)} d^2_{\mathrm{LW}}(R_\#\mu,\nu),$$

where $\mathrm{SO}(d)$ denotes the special orthogonal group on the horizontal directions $\mathbb{R}^d$.

Note that $d_{\mathrm{LW,symm}}$ is a metric on the set of equivalence classes of probability measures under horizontal rotational equivalence (that is, $\nu\sim\mu$ if $\nu=R_\#\mu$ for some rotation $R\in\mathrm{SO}(d)$). A horizontally symmetrized Wasserstein barycenter $\mathrm{Bar}^{\mathrm{LW}}_{\mathrm{symm}}(\Omega)$ of a measure $\Omega\in P(M^+(\mathbb{R}^d\times\mathbb{R}_{\geq 0}))$ is then a metric barycenter with respect to this distance; that is, it is a minimizer of

$$\nu\mapsto\int_{M^+(\mathbb{R}^d\times\mathbb{R}_{\geq 0})} d^2_{\mathrm{LW,symm}}(\nu,\mu)\,d\Omega(\mu).$$

Equivalently, $\mathrm{Bar}^{\mathrm{LW}}_{\mathrm{symm}}(\Omega)$ minimizes

$$\nu\mapsto\min_{\substack{R_\mu\in\mathrm{SO}(d)\\ \text{for all }\mu\in M^+(\mathbb{R}^d\times\mathbb{R}_{\geq 0})}}\int_{M^+(\mathbb{R}^d\times\mathbb{R}_{\geq 0})} d^2_{\mathrm{LW}}(\nu,(R_\mu)_\#\mu)\,d\Omega(\mu).$$

Analogously, one could also consider rotationally symmetrized versions of the classical Wasserstein distance,

$$W^2_{2,\mathrm{symm}}(\mu,\nu):=\min_{R\in\mathrm{SO}(d)} W^2_2(\mu,R_\#\nu),$$

and corresponding barycenters, which are minimizers of

$$\nu\mapsto\min_{\substack{R_\mu\in\mathrm{SO}(d)\\ \text{for all }\mu\in M^+(\mathbb{R}^d\times\mathbb{R}_{\geq 0})}}\int_{P(\mathbb{R}^d\times\mathbb{R}_{\geq 0})} W^2_2(\nu,(R_\mu)_\#\mu)\,d\Omega(\mu). \tag{2-4}$$

Symmetrized Wasserstein barycenters are more natural for the root interpolation problem than classical Wasserstein barycenters. One of our goals in this paper is to demonstrate that symmetrized layerwise-Wasserstein barycenters are better suited for this problem than classical (symmetrized or unsymmetrized) Wasserstein barycenters; to this end, we provide examples in Section 3B of measures $\mu_1,\dots,\mu_m$ which are root-like in a certain sense (skeletons in the nomenclature of the next section), for which the symmetrized Wasserstein barycenter of $(1/m)\sum_{i=1}^m\delta_{\mu_i}$ does not resemble a root (that is, is not a skeleton). Their layerwise-Wasserstein barycenter, on the other hand, has a much more root-like structure (see Theorem 3.6 below).

3. Skeletal measures

Real plant root systems consist of limbs with thickness. However, biologists often approximate roots by their "skeletons", in which each limb is replaced by a one-dimensional curve, thus retaining the topological, or dendritic structure of the root, but losing its thickness. Below, we provide a formal mathematical definition of skeletons, and introduce skeletal measures, which are essentially distributions of mass supported on them.

Definition 3.1. Let $Y = [0, \bar{y}] \subset \mathbb{R}$, be an interval whose length $\bar{y}$ represents the vertical depth of the root. A *weak skeletal root* consists of the graphs of a finite union of curves,

$$\bigcup_{i=1}^{N} \operatorname{graph}(g_i),$$

where each $g_i : [\underline{y}_i, \bar{y}_i] \to \mathbb{R}^d$ is a Lipschitz function defined on a subinterval $[\underline{y}_i, \bar{y}_i] \subseteq Y$, satisfying the following properties:

(S1) (roots start from a common stem) $\underline{y}_1 = 0$ and $\underline{y}_i > 0$ for each $i = 2, \dots, N$.

(S2) (limbs emerge from older limbs) For each $i = 2, \dots, N$, there is some $j < i$ such that $\underline{y}_i \in (\underline{y}_j, \bar{y}_j)$ and $g_i(\underline{y}_i) = g_j(\underline{y}_i)$.

A *strong skeletal root* is a weak skeletal root which satisfies the additional condition:

(S3) (limbs never cross each other) For each $i \neq j$ and all $y \in (\underline{y}_i, \bar{y}_i] \cap (\underline{y}_j, \bar{y}_j]$, we have $g_i(y) \neq g_j(y)$.

We next define strong skeletal root measures.

Definition 3.2. A *strong skeletal root measure* is a probability measure whose support is an *entire* strong skeletal root, which is absolutely continuous with respect to the one-dimensional Hausdorff measure.

Strong skeletal root measures seem to be reasonable proxies for real roots. As we will see below, layerwise-Wasserstein barycenters of strong skeletal root measures preserve the one-dimensional structure of the support (this is an important distinction from conventional Wasserstein barycenters — see Example 3.12 below). Unfortunately, they are not always strong skeletal root measures for two reasons: the support may be disconnected, and the noncrossing property holds only in a weaker sense. This motivates the following definition:

Definition 3.3. A *weak skeletal root measure* is a probability measure supported on a weak skeletal root, which is absolutely continuous with respect to the one-dimensional Hausdorff measure, satisfying the following additional property:

(W3) For each $i \neq j$ and all $y \in (\underline{y}_i, \bar{y}_i] \cap (\underline{y}_j, \bar{y}_j]$, such that $g_i(y) = g_j(y)$, we have either

$$\lim_{z \to y^-} \mu_z(\{g_i(z)\}) = 0 \quad \text{or} \quad \lim_{z \to y^-} \mu_z(\{g_j(z)\}) = 0,$$

where $\mu_y = \tilde{\mu}_{F_\mu(y)}$ is the conditional probability of $d\mu(x, y) = d\mu_y(x)\, d\mu^V(y)$.

Note that by construction, for each l, the layer $\tilde{\mu}_l$ of a weak skeletal root measure μ, is a convex combination of Dirac masses.

Obviously strong skeletal root measures are weak skeletal root measures; weak skeletal root measures are essentially "roots with missing parts" and have a weaker version (W3) of the noncrossing condition. While the layerwise-Wasserstein barycenter of several strong skeletal root measures may not be a strong skeletal root measure, we are able to show below that it is a weak skeletal root measure.

Remark 3.4. Interpreting each graph(g_i) as a limb, condition (S3) expresses the natural expectation that limbs do not cross. We interpret (W3) as a weaker version of this: if $g_i(y) = g_j(y)$ and $\lim_{z\to y^-} \mu_z(\{g_i(z)\}) = 0$, we interpret g_i as consisting of two limbs: an upper limb g_i^1, defined by restricting g_i to $[\underline{y}_i, y]$, and a lower limb, g_i^2, obtained by restricting g_i to $[y, \bar{y}_i]$, emerging from the older limb y_j at the point y. This seems reasonable to us, since the hypothesis $\lim_{z\to y^-} \mu_z(\{g_i(z)\}) = 0$ means that there is no mass at y coming from the upper limb; the upper limb thus ends at the point y.

By interpreting a weak root as a tree in this sense, one can compute topological properties which are defined only for loop-free structures (including, for example, the Horton–Strahler index [Toroczkai 2002], often used by biologists to measure the topological complexity of root systems).

Remark 3.5. Skeletons can be computationally useful in practice. Algorithms are available to construct skeletons from real root system data, essentially by tracing the center of mass of the cross sections of each limb [Bucksch 2014]. Computing layerwise-Wasserstein barycenters of these skeletons is then much less computationally intensive than computing the barycenters of the original roots, since each layer is discretized by many fewer points, but may still provide valuable biological insight about the "average" topological structure of the family of root systems.

Assuming μ is a (weak or strong) skeletal root measure, supported on the skeletal root $\bigcup_{i=1}^N \operatorname{graph}(g_i)$, and the rescaling map F_μ is bi-Lipschitz, $\tilde{\mu}$ is also a (respectively weak or strong) skeletal root measure, supported on the skeletal root $\bigcup_{i=1}^N \operatorname{graph}(\tilde{g}_i)$, where $\tilde{g}_i := g_i \circ F_\mu^{-1}$. Note that the domain $[\underline{l}_i, \bar{l}_i] := [F_\mu(\underline{y}_i), F_\mu(\bar{y}_i)]$ of each rescaled limb $\tilde{g}_i$ is contained in $[0, 1]$. We call $\bigcup_{i=1}^N \operatorname{graph}(\tilde{g}_i)$ a *rescaled* skeletal root.

3A. *Layerwise-Wasserstein barycenters of skeletal root measures.*

We now prove that layerwise-Wasserstein barycenters of weak skeletal root measures are themselves weak skeletal root measures.

Theorem 3.6. *Let $\mu_1, \dots, \mu_m \in P(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ be compactly supported weak skeletal root measures such that $l \mapsto (\tilde{\mu}_i)_l$ is weak-$*$ left continuous and F_{μ_i} is bi-Lipschitz for each i, and $\lambda_1, \dots, \lambda_m > 0$ with $\sum_{i=1}^m \lambda_i = 1$. Then any layerwise-Wasserstein barycenter $\operatorname{Bar}^{\mathrm{LW}}\big(\sum_\alpha \lambda_\alpha \mu_\alpha\big)$ of $\mu_1, \dots, \mu_m$ with weights $\lambda_1, \dots, \lambda_m$ is also a weak skeletal root measure.*

Remark 3.7. We expect this result to play an important role in biological applications. As mentioned above, given a family of root systems, we will propose in future work interpreting the layerwise-Wasserstein barycenter as the best representative of that family. It is therefore desirable to compute certain biologically relevant traits of the barycenter, especially those traits that rely on its dendritic structure, for instance the total root length and the Horton–Strahler (HS) index [Toroczkai 2002]. The HS index in particular relies on the noncrossing property, and can be defined for weak skeletal root measures, thanks to (W3), but not for more general unions of graphs such as weak skeletal roots.

The key tool in the proof of this theorem is the barycentric ghost, which we define now.

Definition 3.8. For $\alpha = 1, 2, \dots, m$, let $\widetilde{S}_\alpha := \{\tilde{g}^\alpha_{i_\alpha} : i_\alpha = 1, 2, \dots, N_\alpha\}$ be a rescaled skeletal root, and let $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m)$ with $\lambda_1, \dots, \lambda_m > 0$ be a collection of weights with $\sum_{\alpha=1}^m \lambda_\alpha = 1$.

For fixed indices $i_1, \dots, i_m$, whenever the intersection $\bigcap_{\alpha=1}^m [\underline{l}^\alpha_{i_\alpha}, \bar{l}^\alpha_{i_\alpha}]$ of domains $[\underline{l}^\alpha_{i_\alpha}, \bar{l}^\alpha_{i_\alpha}]$ of the family $\{\tilde{g}^\alpha_{i\alpha}\}$ is nonempty, we define the curve

$$\widetilde{G}^\lambda_{i_1,i_2,\dots,i_m} := \sum_{\alpha=1}^m \lambda_\alpha \tilde{g}^\alpha_{i_\alpha}.$$

The *ghost* of the family $\{\widetilde{S}_\alpha\}$ with weights λ is then the collection of curves $\widetilde{G}^\lambda_{i_1,i_2,\dots,i_m}$.

At each slice $l \in [0, 1]$, the set $\widetilde{G}^\lambda_{i_1,i_2,\dots,i_m}(l)$ represents the Euclidean barycenters of all possible combinations of $\tilde{g}^\alpha_{i_\alpha}(l)$ in the supports of the discrete sliced layers. The Wasserstein barycenter of the layers is supported on these points; therefore, for skeletal root measures $\mu_1, \dots, \mu_m$, supported respectively on $\bigcup_{i_\alpha=1}^{N_\alpha} \operatorname{graph}(g^\alpha_{i_\alpha})$, we have

$$\text{if } (x, y) \in \operatorname{supp}\bigl(\mathrm{Bar}^{\mathrm{LW}}\bigl(\textstyle\sum_{\alpha=1}^m \lambda_\alpha \delta_{\mu_\alpha}\bigr)\bigr), \\ \text{then } x = \widetilde{G}^\lambda_{i_1,i_2,\dots,i_m}\Bigl(\bigl(\textstyle\sum_{\alpha=1}^m \lambda_\alpha F^{-1}_{\mu_\alpha}\bigr)^{-1}(y)\Bigr) \text{ for some choice of } i_1, \dots, i_m. \tag{3-1}$$

Given probability root measures, the ghost of their rescaled supports $\bigcup_{i_\alpha=1}^N \operatorname{graph}(\tilde{g}^\alpha_{i_\alpha})$ can be unrescaled via the mapping $y \mapsto \bigl(\sum_{\alpha=1}^m \lambda_\alpha F^{-1}_{\mu_\alpha}\bigr)^{-1}(y)$; the unrescaled ghost is then the union of the graphs

$$G^\lambda_{i_1,i_2,\dots,i_m}(y) := \widetilde{G}^\lambda_{i_1,i_2,\dots,i_m}\left(\left(\sum_{\alpha=1}^m \lambda_\alpha F^{-1}_{\mu_\alpha}\right)^{-1}(y)\right).$$

It is then easy to see that the layerwise-Wasserstein barycenter has support contained in the (unrescaled) ghost, though it typically won't fill it out. We think of the ghost sitting in the background; it is the largest possible potential support of the barycenter. We think of the actual support of the barycenter as sitting in the foreground on top of it.

Now, the ghost clearly satisfies (S1) (starting as stem) and (S2) (limbs emerge from older limbs) in the definition of skeletal roots. It does not generally satisfy (S3) (noncrossing). In order to verify that the layerwise-Wasserstein barycenter is a weak skeletal root measure, we must therefore show that it satisfies the weak noncrossing property (W3).

The following lemma essentially verifies (W3) for the rescaled barycenter; since it is clear that the bi-Lipschitz rescaling $\sum_{\alpha=1}^m \lambda_\alpha F^{-1}_{\mu_\alpha}$, which pushes $\widetilde{\mathrm{Bar}}^{\mathrm{LW}}\bigl(\sum_\alpha \lambda_\alpha \mu_\alpha\bigr)$ forward to $\mathrm{Bar}^{\mathrm{LW}}\bigl(\sum_\alpha \lambda_\alpha \mu_\alpha\bigr)$, preserves this property, the lemma implies Theorem 3.6.

Lemma 3.9. *Under the same assumptions as in Theorem 3.6, let* $\mu = \mathrm{Bar}^{\mathrm{LW}}\bigl(\sum_\alpha \lambda_\alpha \mu_\alpha\bigr)$ *be a layerwise-Wasserstein barycenter of the* $\{\mu_\alpha\}$*'s. Set* $l = \bigl(\sum_{\alpha=1}^m \lambda_\alpha F^{-1}_{\mu_\alpha}\bigr)^{-1}(y)$*, and suppose* $x = \widetilde{G}^\lambda_{j_1,j_2,\dots,j_m}(l) = \widetilde{G}^\lambda_{i_1,i_2,\dots,i_m}(l)$*, where* $j_\alpha \neq i_\alpha$ *for at least one* α *and* l *is not the minimal point in the domain of* $\widetilde{G}^\lambda_{i_1,i_2,\dots,i_m}$ *or* $\widetilde{G}^\lambda_{j_1,j_2,\dots,j_m}$*. Moreover, suppose that* $\lim_{z\to l^-} \tilde{\mu}_z(\widetilde{G}^\lambda_{i_1,i_2,\dots,i_m}(z)) > 0$*. Then,*

$$\lim_{z\to l^-} \tilde{\mu}_z(\{\widetilde{G}^\lambda_{j_1,j_2,\dots,j_m}(z)\} \setminus \{G^\lambda_{i_1,i_2,\dots,i_m}(z)\}) = 0.$$

The proof of this lemma leverages a connection between the Wasserstein barycenter of $\sum_\alpha \lambda_\alpha \delta_{(\tilde{\mu}_\alpha)_l}$ and the multimarginal extension of optimal transport, which is to minimize

$$\int_{(\mathbb{R}^d)^m} \sum_{\alpha,\beta} \lambda_\alpha \lambda_\beta |x_\alpha - x_\beta|^2 \, d\gamma(x_1, x_2, \dots, x_m) \tag{3-2}$$

among all probability measures γ on $(\mathbb{R}^d)^m$ whose marginals are the $(\tilde{\mu}_\alpha)_l$. It is well known that the mapping

$$\Delta_\lambda : (x_1, x_2, \dots, x_m) \to \sum_\alpha \lambda_\alpha x_\alpha$$

pushes each solution $\tilde{\gamma}_l$ forward to a Wasserstein barycenter $\tilde{\mu}_l$ [Agueh and Carlier 2011], and this mapping is invertible with a Lipschitz inverse on the support of $\tilde{\mu}_l$; see, e.g., [Kim and Pass 2017].

Proof of Lemma 3.9. For each layer l, we will denote by $\tilde{\gamma}_l \in P(\mathbb{R}^d \times \cdots \times \mathbb{R}^d)$ a solution to the multimarginal optimal transport problem (3-2). Assume that the conclusion of the lemma fails. Then there exists $\epsilon > 0$ and a sequence $l_k < l$ converging to l such that $\widetilde{G}^\lambda_{j_1,j_2,\dots,j_m}(l_k) \neq \widetilde{G}^\lambda_{i_1,i_2,\dots,i_m}(l_k)$ and $\tilde{\mu}_{l_k}(\widetilde{G}^\lambda_{j_1,j_2,\dots,j_m}(l_k)) > \epsilon$ for large enough k.

We first prove the lemma under the simplifying assumption that $\widetilde{G}^\lambda_{j_1,j_2,\dots,j_m}(l_k) \neq \widetilde{G}^\lambda_{j'_1,j'_2,\dots,j'_m}(l_k)$ for all $(j'_1, j'_2, \dots, j'_m) \neq (j_1, j_2, \dots, j_m)$.

This immediately implies

$$\tilde{\gamma}_{l_k}(\tilde{g}^1_{j_1}(l_k), \tilde{g}^2_{j_2}(l_k), \dots, \tilde{g}^m_{j_m}(l_k)) > \epsilon \tag{3-3}$$

and in particular

$$(\tilde{\mu}_\alpha)_{l_k}(g^\alpha_{j_\alpha}(l_k)) > \epsilon \tag{3-4}$$

for $\alpha = 1, 2 \dots, m$. After passing to a convergent subsequence, the $\tilde{\gamma}_{l_k}$ converge (in the weak-$*$ sense) to a measure $\tilde{\gamma}_l$ which is optimal in the multimarginal problem for the $(\tilde{\mu}_{i_\alpha})_l$, and

$$\tilde{\gamma}_l(\tilde{g}^1_{j_1}(l), \tilde{g}^2_{j_2}(l), \dots, \tilde{g}^m_{j_m}(l)) \neq 0.$$

Exactly the same argument implies the existence of a second minimizer $\tilde{\gamma}'_l$ to the multimarginal problem such that $\tilde{\gamma}'_l(\tilde{g}^l_{i_1}(l), \tilde{g}^2_{i_2}(l), \dots, \tilde{g}^m_{i_m}(l)) \neq 0$.

Although it is possible that $\tilde{\gamma}'_l \neq \tilde{\gamma}_l$, their linear average $\frac{1}{2}\tilde{\gamma}'_l + \frac{1}{2}\tilde{\gamma}_l$ is also optimal for the multimarginal problem and has both

$$(\tilde{g}^1_{i_1}(l), \tilde{g}^2_{i_2}(l), \dots, \tilde{g}^m_{i_m}(l)) \quad \text{and} \quad (\tilde{g}^1_{j_1}(l), \tilde{g}^2_{j_2}(l), \dots, \tilde{g}^m_{j_m}(l))$$

in its support, with the corresponding Wasserstein barycenter $\hat{\mu}_l = \frac{1}{2}\tilde{\mu}_l + \frac{1}{2}\tilde{\mu}'_l$. Notice that

$$\Delta_\lambda((\tilde{g}^1_{i_1}(l), \tilde{g}^2_{i_2}(l), \dots, \tilde{g}^m_{i_m}(l))) = x = \Delta_\lambda((\tilde{g}^1_{j_1}(l), \tilde{g}^2_{j_2}(l), \dots, \tilde{g}^m_{j_m}(l))).$$

Because Δ_λ has a Lipschitz inverse, we have

$$|(\tilde{g}^1_{i_1}(l), \tilde{g}^2_{i_2}(l), \dots, \tilde{g}^m_{i_m}(l)) - (\tilde{g}^1_{j_1}(l), \tilde{g}^2_{j_2}(l), \dots, \tilde{g}^m_{j_m}(l))| \leq C|x - x| = 0.$$

Now, letting α be such that $j_\alpha \neq i_\alpha$, the above implies that $\tilde{g}^\alpha_{i_\alpha}(l) = \tilde{g}^\alpha_{j_\alpha}(l)$. Since μ_α is a weak root measure, this means that, without loss of generality,

$$\lim_{z \to l^-} (\tilde{\mu}_\alpha)_z(\{\tilde{g}^\alpha_{j_\alpha}(z)\}) = 0.$$

This contradicts (3-4) and completes the proof under the additional assumption.

Now, if the assumption fails, instead of (3-3), we can conclude only that

$$\tilde{\gamma}_{l_k}(\tilde{g}^1_{j'_1}(l_k), \tilde{g}^2_{j_2}(l'_k), \dots, \tilde{g}^m_{j_m}(l'_k)) > \epsilon$$

for some $(j'_1, \dots, j'_m)$ with $G^\lambda_{j'_1, j'_2, \dots, j'_m}(l_k) = \widetilde{G}^\lambda_{j_1, j_2, \dots, j_m}(l_k)$, and by passing to a subsequence if necessary, we can take it to be the same $(j'_1, \dots, j'_m)$ for each k. As above, this implies

$$\tilde{\gamma}_l(\tilde{g}^1_{j'_1}(l), \tilde{g}^2_{j'_2}(l), \dots, \tilde{g}^m_{j'_m}(l)) \neq 0,$$

and since $\widetilde{G}^\lambda_{j'_1, j'_2, \dots, j'_m}(l_k) = \widetilde{G}^\lambda_{j_1, j_2, \dots, j_m}(l_k)$, passing to the limit implies

$$\widetilde{G}^\lambda_{j'_1, j'_2, \dots, j'_m}(l) = \widetilde{G}^\lambda_{j_1, j_2, \dots, j_m}(l) = \widetilde{G}^\lambda_{i_1, i_2, \dots, i_m}(l).$$

The rest of the proof follows exactly as in the special case above. □

Remark 3.10. If the solution $\tilde{\gamma}_l$ to the multimarginal problem (3-2) in the proof above is *unique*, and the μ_α are all strong root measures, then more is true: if $\lim_{z \to l^-} \tilde{\mu}_z(\{\widetilde{G}^\lambda_{j_1, j_2, \dots, j_m}(z)\}) \neq 0$ we actually have $\tilde{\mu}_l(G^\lambda_{i_1, i_2, \dots, i_m}(z)) = 0$ for $z < l$ sufficiently close to l.

To see this, note that as above, $\tilde{\gamma}_l(\tilde{g}^1_{j_1}(l), \tilde{g}^2_{j_2}(l), \dots, \tilde{g}^m_{j_m}(l)) \neq 0$. Since the root measures are strong, we must have $\tilde{g}^\alpha_{i_\alpha}(l) \neq \tilde{g}^\alpha_{j_\alpha}(l)$ for the α such that $i_\alpha \neq j_\alpha$. For $z < l$ with z close to l, *any* solution $\tilde{\gamma}_z$ to the multimarginal plan (3-2) must be weak-$*$ close to $\tilde{\gamma}_l$ (by uniqueness) and so must satisfy $\tilde{\gamma}_z(\tilde{g}^1_{j_1}(z), \tilde{g}^2_{j_2}(z), \dots, \tilde{g}^m_{j_m}(z)) \neq 0$. If such a solution satisfied $\tilde{\gamma}_z(\tilde{g}^1_{i_1}(z), \tilde{g}^2_{i_2}(z), \dots, \tilde{g}^m_{i_m}(z)) \neq 0$ as well, we would then have, by the Lipschitz property of Δ_λ^{-1},

$$|(g^1_{i_1}(z), g^2_{i_2}(z), \dots, g^m_{i_m}(z)) - (g^1_{j_1}(z), g^2_{j_2}(z), \dots, g^m_{j_m}(z))| \leq C|G^\lambda_{i_1, i_2, \dots, i_m}(z) - G^\lambda_{j_1, j_2, \dots, j_m}(z)|.$$

However, this is impossible since the right-hand side tends to 0 as z tends to l, but the left-hand side does not (as $\tilde{g}^\alpha_{i_\alpha}(l) \neq \tilde{g}^\alpha_{j_\alpha}(l)$ for at least one α, as described above). We conclude that we must have $\tilde{\gamma}_z(\tilde{g}^1_{i_1}(z), \tilde{g}^2_{i_2}(z), \dots, \tilde{g}^m_{i_m}(z)) = 0$ for any solution to the multimarginal problem and all $z < l$ sufficiently close to l; therefore, $\tilde{\mu}_z(G^\lambda_{i_1, i_2, \dots, i_m}(z)) = 0$ for any Wasserstein barycenter $\tilde{\mu}_z$ of the $\tilde{\mu}_1, \dots, \tilde{\mu}_m$.

This applies, for instance, when $d = 1$, in which case Wasserstein barycenters are always unique.

3B. ***Comparison with the Wasserstein barycenter.*** If we instead use the standard notion of the Wasserstein barycenter to interpolate between several root measures, the barycenter may not be a weak root measure, as the following examples show.

Example 3.11. Several constructions of [Santambrogio and Wang 2016] show that displacement interpolation does not generally preserve convexity of sets. In one of these, the two marginals measures are concentrated on line segments embedded in $\mathbb{R}^2$, while their displacement interpolant (or Wasserstein barycenter) is supported on a curve $y = f(x)$ with a strict local minimum, where y is the vertical direction;

see $\mu_{1/2}$ in Section 2 in [Santambrogio and Wang 2016]. In our context, the two line segments constitute simple strong skeletal root measures, whereas the displacement interpolant is not even a weak skeletal root measure (as the two limbs meeting at the minimum point x_0 of f violate (W3)). Note that this is precisely because the angle between the two limbs is greater than $\frac{\pi}{2}$, and so the optimal map is not monotone in the vertical direction.

Our second example is even less well-behaved; here we take three strong root measures for which the Wasserstein barycenter has three-dimensional support.

Example 3.12. Consider uniform measure on the mutually orthogonal segments

$$\begin{aligned} T &:= \{(t,t,t) : t \in [0,1]\},\\ R &:= \Big\{\Big(r, \big(\tfrac{-1+\sqrt{3}}{2}\big)r, \big(\tfrac{-1-\sqrt{3}}{2}\big)r\Big) : r \in [0,1]\Big\},\\ S &:= \Big\{\Big(s, \big(\tfrac{-1-\sqrt{3}}{2}\big)s, \big(\tfrac{-1+\sqrt{3}}{2}\big)s\Big) : s \in [0,1]\Big\} \end{aligned}$$

in $\mathbb{R}^3$.

Since the segments are orthogonal, the interaction terms $x_\alpha \cdot x_\beta = 0$ in the Gangbo–Swiech cost $(\mathbb{R}^3)^3$ (with, say, $\lambda_\alpha = 1/m,\ m = 3$)

$$\sum_{\alpha,\beta} \tfrac{1}{9}|x_\alpha - x_\beta|^2 = -\sum_{\alpha,\beta=1}^{3} \tfrac{4}{9} x_\alpha \cdot x_\beta + \sum_{\alpha=1}^{3} \tfrac{4}{9}|x_\alpha|^2$$

vanish.

Therefore, any measure with Lebesgue marginals supported on the product space $T \times R \times S$ is optimal in the multimarginal problem (3-2). The pushforward of any such measure γ by the mapping

$$\Big((t,t,t), \Big(r, \big(\tfrac{-1+\sqrt{3}}{2}\big)r, \big(\tfrac{-1-\sqrt{3}}{2}\big)r\Big), \Big(s, \big(\tfrac{-1-\sqrt{3}}{2}\big)s, \big(\tfrac{-1+\sqrt{3}}{2}\big)s\Big)\Big) \\ \mapsto \frac{(t,t,t) + (r,r,-2r) + \Big(r, \big(\tfrac{-1+\sqrt{3}}{2}\big)r, \big(\tfrac{-1-\sqrt{3}}{2}\big)r\Big) + \Big(s, \big(\tfrac{-1-\sqrt{3}}{2}\big)s, \big(\tfrac{-1+\sqrt{3}}{2}\big)s\Big)}{3}$$

is a Wasserstein barycenter. If γ is, for example, a product measure, this push forward is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^3$.

This is certainly not a skeletal measure, and cannot be interpreted as a root in any reasonable way.

As with the layerwise-Wasserstein distance, one might suggest that horizontal symmetrization of the classical Wasserstein distance is more appropriate for comparing root shapes. That is, one should consider minimizers of (2-4). In the preceding example, although the Wasserstein barycenter has three-dimensional support, the biologically more relevant horizontally symmetrized version is concentrated on a line segment (since after appropriate rotations, the three sample measures are the same). Below, we augment the sample measures to produce a horizontally symmetrized Wasserstein barycenter with three-dimensional support.

Example 3.13. Let μ_1, μ_2 and μ_3 be uniform measures on the respective domains S_i defined by

$$\begin{aligned} S_1 &:= \{(t,t,t) : t \in [0,1+\epsilon]\},\\ S_2 &:= \{(t,t,t) : t \in [0,1]\} \cup \Big\{\Big(1+t, 1+\big(\tfrac{-1+\sqrt{3}}{2}\big)t, 1+\big(\tfrac{-1-\sqrt{3}}{2}\big)t\Big) : t \in [0,\epsilon]\Big\},\\ S_3 &:= \{(t,t,t) : t \in [0,1]\} \cup \Big\{\Big(1+t, 1+\big(\tfrac{-1-\sqrt{3}}{2}\big)t, 1+\big(\tfrac{-1+\sqrt{3}}{2}\big)t\Big) : t \in [0,\epsilon]\Big\}. \end{aligned}$$

It is not hard to show that the identity rotation minimizes the Wasserstein distance between μ_i and $R_{\#}\mu_j$ among horizontal rotations R for sufficiently small ϵ.

Furthermore, the optimal plans between μ_i and μ_j couple the top limbs via the identify mappings and the bottom limbs via product measure (or any other coupling between the bottom limbs — the solution is nonunique). Therefore, the measure

$$\begin{aligned}\gamma = {} & (\mathrm{id}\times\mathrm{id}\times\mathrm{id})_{\#}(\mu_1|_{\{(t,t,t):t\in[0,1]\}}) \\ & +(\mu_1|_{\{(t,t,t):t\in[1,1+\epsilon]\}})(\mu_2|_{\{(1+t,1+(\frac{-1+\sqrt{3}}{2})t,1+(\frac{-1-\sqrt{3}}{2})t):t\in[0,\epsilon]\}})(\mu_3|_{\{(1+t,1+(\frac{-1-\sqrt{3}}{2})t,1+(\frac{-1+\sqrt{3}}{2})t):t\in[0,\epsilon]\}})\end{aligned}$$

is optimal in the multimarginal problem (3-2), and this plan has minimal cost among all multimarginal problems with marginals $(\mu_1, R_{2\#}\mu_2, R_{3\#}\mu_3)$ for horizontal rotations R_2 and R_3. Consequently the symmetrized Wasserstein barycenter from (2-4) is then the pushforward of this measure under the mapping $(x_1, x_2, x_3) \mapsto \frac{1}{3}(x_1 + x_2 + x_3)$; this consists of the uniform measure on $\{(t, t, t) : t \in [0, 1]\}$ and a measure constructed as in the previous example, with three-dimensional support, arising from coupling the three orthogonal lower limbs.

3C. *Total root length.* An important phenotype used by biologists to compare root systems is the total root length, which is well-defined for skeletal root systems.

Given a strong skeletal root measure μ supported on $\bigcup_{i=1}^{N} \mathrm{graph}(g_i)$ on $[0, \bar{y}]$, the total root length of μ is simply the one-dimensional Hausdorff measure of its support. Letting χ_i be the indicator function of the domain $[\underline{y}_i, \bar{y}_i] \subseteq [0, \bar{y}]$ of g_i, we note that the root length is

$$R(\mu) = \sum_{i=1}^{N} \int_0^{\bar{y}} \sqrt{1 + |(g_i)'(y)|^2}\,\chi_i(y)\,dy. \tag{3-5}$$

Here we establish a result comparing the total root lengths of several skeletal root systems and their layerwise-Wasserstein barycenter. Given strong skeletal root measures

$$\mu_\alpha \text{ supported on } \bigcup_{i=1}^{N_\alpha} \mathrm{graph}(g^\alpha_{i_\alpha}) \text{ on } [0, \bar{y}^\alpha] \text{ for } \alpha = 1, 2, \dots, m,$$

we compare their total root lengths to that of (a selected) layerwise-Wasserstein barycenter, with weights $\lambda_1, \dots, \lambda_m$. As above, we will also assume two sided bound

$$0 < L \le f^V_\alpha(y) \le U < \infty$$

on each μ_α, where f^V_α is the density of the vertical marginal μ^V_α. We have that $F'_{\mu_\alpha}(y) = f^V_\alpha(y)$, which implies that each rescaling change of variables is bi-Lipschitz.

Let $\bar{y} = \sum_{\alpha=1}^{m} \lambda_\alpha \bar{y}^\alpha$, so that any layerwise-Wasserstein barycenter of the μ_α is supported on $\mathbb{R}^d \times [0, \bar{y}]$. Assume that each $g^\alpha_{i_\alpha}$ is in $C^1([\underline{y}^\alpha_{i_\alpha}, \bar{y}^\alpha_{i_\alpha}])$ and let C be an upper bound on each $|(g^\alpha_{i_\alpha})'|$. We define the total root length of a layerwise-Wasserstein barycenter $\mathrm{Bar}^{\mathrm{LW}}\big(\sum_{\alpha=1}^{m} \lambda_\alpha \delta_{\mu_\alpha}\big)$ as the one-dimensional Hausdorff measure of its support, namely, the set

$$\left\{(x, y) : x \in \mathrm{spt}\left(\widetilde{\mathrm{Bar}}^{\mathrm{LW}}_l\left(\sum_{\alpha=1}^{m} \lambda_\alpha \delta_{\mu_\alpha}\right)\right), l = \left(\sum_{\alpha=1}^{m} \lambda_\alpha F^{-1}_{\mu_\alpha}\right)^{-1}(y)\right\},$$

where, as before $\widetilde{\mathrm{Bar}}_l^{\mathrm{LW}}\big(\sum_{\alpha=1}^m \lambda_\alpha \delta_{\mu_\alpha}\big)$ is the horizontal slice of the layerwise-Wasserstein barycenter at level l (that is, the Wasserstein barycenter of the $(\tilde{\mu}_\alpha)_l$).

Letting $G^\lambda_{i_1,i_2,\dots,i_m}$ be one of the graphs in the ghost, we let $\chi^\lambda_{i_1,i_2,\dots,i_m}$ be the indicator function of its active set; that is,

$$\chi^\lambda_{i_1,i_2,\dots,i_m}(y)=\begin{cases}1 & \text{if } G^\lambda_{i_1,i_2,\dots,i_m}(y) \text{ is well-defined and in the support of } \widetilde{\mathrm{Bar}}^{\mathrm{LW}}_{(\sum_{\alpha=1}^m\lambda_\alpha F^{-1}_{\mu_\alpha})^{-1}(y)}\big(\sum_{\alpha=1}^m\lambda_\alpha\delta_{\mu_\alpha}\big),\\ 0 & \text{otherwise.}\end{cases}$$

The root length is then

$$\sum_{i_1,\dots,i_m}\int_0^{\bar{y}} \sqrt{1+|(G^\lambda_{i_1,i_2,\dots,i_m})'(y)|^2}\,\chi^\lambda_{i_1,i_2,\dots,i_m}(y)\,dy.$$

Proposition 3.14. *Letting μ_α be skeletal roots for $\alpha = 1, 2, \dots, m$, there is a layerwise-Wasserstein barycenter* $\mathrm{Bar}^{\mathrm{LW}}\big(\sum_{\alpha=1}^m \lambda_\alpha\delta_{\mu_\alpha}\big)$ *of* $\sum_{\alpha=1}^m \lambda_\alpha\delta_{\mu_\alpha}$ *for which*

$$C_0R(\mu_\beta) \le R\left(\mathrm{Bar}^{\mathrm{LW}}\left(\sum_{\alpha=1}^m \lambda_\alpha\delta_{\mu_\alpha}\right)\right) \le C_1\left[C_2\sum_{\alpha=1}^m R(\mu_\alpha) - (m-1)\right]$$

for any $\beta = 1, 2, \dots, m$. The constants C_0, C_1 and C_2 depend only on $C = \sup_{\alpha,i_\alpha}\|(g^\alpha_{i_\alpha})'\|_{L^\infty}$, L and U.

The proof essentially consists of two steps: first, we establish a similar result for the rescaled versions. We then use bounds on the rescaling change of variables to translate the rescaled inequalities back to the original coordinates. We isolate the first step as a separate lemma.

Lemma 3.15. *Using the notation in the proposition above, there exists a layerwise Wasserstein barycenter such that*

$$\widetilde{C}_0R(\tilde{\mu}_\beta) \le R\left(\widetilde{\mathrm{Bar}}^{\mathrm{LW}}\left(\sum_{\alpha=1}^m \lambda_\alpha\delta_{\mu_\alpha}\right)\right) \le \widetilde{C}_1\left[\sum_{\alpha=1}^m R(\tilde{\mu}_\alpha) - (m-1)\right]$$

for any $\beta = 1, 2, \dots, m$.

Proof. Note that $\tilde{\mu}_\alpha$ is supported on the skeletal set $\bigcup_{i=1}^{N_\alpha}(\tilde{g}^\alpha_{i_\alpha})$, where $\tilde{g}^\alpha_{i_\alpha} := g^\alpha_{i_\alpha}\circ F^{-1}_{\mu_\alpha}$. The $\tilde{g}^\alpha_{i_\alpha}$ then have derivatives bounded by $\widetilde{C} = C/L$.

The ghost of the rescaled system consists of the limbs

$$\widetilde{G}^\lambda_{i_1,i_2,\dots,i_m} = \sum_{\alpha=1}^m \lambda_\alpha\tilde{g}^\alpha_{i_1},$$

which inherit the same derivative bounds as the $\tilde{g}^\alpha_{i_\alpha}$,

$$|(\widetilde{G}^\lambda_{i_1,i_2,\dots,i_m})'| \le \widetilde{C},$$

and at each $l \in [0, 1]$ it is shown in [Anderes et al. 2016] that there is a Wasserstein barycenter $\mathrm{Bar}^W\big(\sum_\alpha \lambda_\alpha\delta_{(\tilde{\mu}_\alpha)_l}\big)$ of the discrete measures $(\tilde{\mu}_\alpha)_l$ such that the number $S(l)$ of points in its support is at most $\sum_{\alpha=1}^m S_\alpha(y) - m + 1$, where $S_\alpha(l)$ is the number of points in the support of $(\tilde{\mu}_\alpha)_l$. We use this

Wasserstein barycenter in our construction of $\widetilde{\mathrm{Bar}}^{\mathrm{LW}}(\sum_{\alpha=1}^{m}\lambda_\alpha\delta_{\mu_\alpha})$. Therefore,

$$\begin{aligned}
R\left(\widetilde{\mathrm{Bar}}^{\mathrm{LW}}\left(\sum_{\alpha=1}^{m}\lambda_\alpha\delta_{\mu_\alpha}\right)\right) &= \sum_{i_1,\dots,i_m}\int_0^1 \sqrt{1+|(\widetilde{G}^{\lambda}_{i_1,i_2,\dots,i_m})'(l)|^2}\,\chi^{\lambda}_{i_1,i_2,\dots,i_m}(l)\,dl\\
&\le \int_0^1 \sqrt{1+\widetilde{C}^2}\left[\sum_{\alpha=1}^{m} S_\alpha(l)-m+1\right]dl\\
&\le \int_0^1 \sum_{\alpha=1}^{m}\sum_{i_\alpha=1}^{N_\alpha}\sqrt{1+|(\tilde{g}^{\alpha}_{i_\alpha})'(l)|^2}\sqrt{1+\widetilde{C}^2}\,\chi^{\alpha}_{i_\alpha}(l)\,dl-\sqrt{1+\widetilde{C}^2}(m-1)\\
&= \sqrt{1+\widetilde{C}^2}\sum_{\alpha=1}^{m}R(\tilde{\mu}_\alpha)-\sqrt{1+\widetilde{C}^2}(m-1).
\end{aligned}$$

Similarly, the $S(l)$ is bounded below by the support of each marginal, $S(l)\ge S_\beta(l)$, and so, for each β

$$\begin{aligned}
R\left(\widetilde{\mathrm{Bar}}^{\mathrm{LW}}\left(\sum_{\alpha=1}^{m}\lambda_\alpha\delta_{\mu_\alpha}\right)\right) &= \sum_{i_1,\dots,i_m}\int_0^1 \sqrt{1+|(\widetilde{G}^{\lambda}_{i_1,i_2,\dots,i_m})'(l)|^2}\,\chi^{\lambda}_{i_1,i_2,\dots,i_m}(l)\,dl\\
&\ge \int_0^1 S_\beta(l)\,dl\\
&\ge \int_0^1 \sum_{i=1}^{N_\beta}\frac{\sqrt{1+|(\tilde{g}^{\beta}_{i})'(l)|^2}}{\sqrt{1+\widetilde{C}^2}}\,\chi^{\beta}_{i}(l)\,dl\\
&= \frac{1}{\sqrt{1+\widetilde{C}^2}}R(\tilde{\mu}_\beta).
\end{aligned}$$

□

The proof of the proposition combines the lemma with straightforward estimates on the change of variables F_{μ_α}.

Proof of Proposition 3.14. The root length of each limb satisfies:

$$\begin{aligned}
\int_{\underline{y}_i}^{\bar{y}_i}\sqrt{1+[(g^{\alpha}_{i_\alpha})'(y)]^2}\,dy &= \int_{\underline{l}_i}^{\bar{l}_i}\sqrt{1+[(g^{\alpha}_{i_\alpha})'(F^{-1}_{\mu_\alpha}(l))]^2}\,[F^{-1}_{\mu_\alpha}]'(l)\,dl\\
&= \int_{\underline{l}_i}^{\bar{l}_i}\sqrt{[F^{-1}_{\mu_\alpha}]'(l)^2+[(g^{\alpha}_{i_\alpha})'(F^{-1}_{\mu_\alpha}(l))]^2[F^{-1}_{\mu_\alpha}]'(l)^2}\,dl\\
&\le K\int_{\underline{l}_i}^{\bar{l}_i}\sqrt{1+[(g^{\alpha}_{i_\alpha})'(F^{-1}_{\mu_\alpha}(l))]^2[F^{-1}_{\mu_\alpha}]'(l)^2}\,dl,
\end{aligned}$$

where $K=\max(1/L,1)$. The last term corresponds to the root length of the corresponding limb of $\tilde{\mu}_\alpha$. Adding over all limbs we get

$$R(\mu_\alpha)\le K\,R(\tilde{\mu}_\alpha),$$

while a symmetric argument yields

$$R(\mu_\alpha)\ge kR(\tilde{\mu}_\alpha),$$

with $k=\min(1/U,1)$.

Similarly, since the vertical rescaling for the barycenter

$$F_{\mathrm{Bar}^{\mathrm{LW}}(\sum_\alpha \lambda_\alpha \delta_{\mu_\alpha})} = \left(\sum_{\alpha=1}^{m} \lambda_\alpha F_{\mu_\alpha}^{-1}\right)^{-1}$$

inherits first derivative bounds from the μ_i, we also get

$$kR\left(\widetilde{\mathrm{Bar}}^{\mathrm{LW}}\left(\sum_{\alpha=1}^{m} \lambda_\alpha \delta_{\mu_\alpha}\right)\right) \le R\left(\mathrm{Bar}^{\mathrm{LW}}\left(\sum_{\alpha=1}^{m} \lambda_\alpha \delta_{\mu_\alpha}\right)\right) \le KR\left(\widetilde{\mathrm{Bar}}^{\mathrm{LW}}\left(\sum_{\alpha=1}^{m} \lambda_\alpha \delta_{\mu_\alpha}\right)\right).$$

Combined with the Lemma 3.15, these estimates yield the desired result. □

Remark 3.16. The result also holds, with essentially the same proof, for weak skeletal root measures, provided we take χ_i^α in (3-5) to be the indicator function of the subset of the domain $[\underline{y}_i^\alpha, \bar{y}_i^\alpha]$ where $(\tilde{\mu}_\alpha)_l(g_{i_\alpha}^\alpha(y)) > 0$ for $l = F_\mu(y)$.

Remark 3.17. It is unfortunately not possible to establish an upper bound on the root length of the layerwise Wasserstein barycenter which is *independent* of the number of samples m.

To see this, consider the skeletal root measures μ_α, each concentrated on two curves $g_{\alpha_1}, g_{\alpha_2} : [0,1] \to \mathbb{R}$, with $g_{\alpha_1}(y) = 0$ and $g_{\alpha_2}(y) = y$. We let the one-dimensional density of each μ_i be constant on each of the two limbs, with densities $1/\alpha$ on g_{α_1} and $1 - 1/\alpha$ on g_{α_2} (normalized to have total mass 1). The vertical marginals of each μ_α are then uniform, so $F_{\mu_\alpha}(y) = y$ and each $(\tilde{\mu}_\alpha)_l = (1/\alpha)\delta_0 + (1 - 1/\alpha)\delta_l$.

It is then not hard to see that the Wasserstein barycenter of $\sum_{\alpha=1}^{m}(1/m)(\tilde{\mu}_\alpha)_l$ is then concentrated on the $m+1$ points $0, l/m, 2l/m, \dots, l$, and so the support of the layerwise-Wasserstein barycenter of $\sum_{\alpha=1}^{m}(1/m)\mu_\alpha$ consists of the m curves $g_1, \dots, g_m : [0,1] \to \mathbb{R}$, with $g_\alpha(y) = \alpha y/m$; the total root length clearly grows with m.

Interpolating between a large number of marginals, or samples, m, can therefore result in weak skeletal measures with very large total root length,

4. Layerwise Wasserstein convexity

We will call a function $\mathcal{F} : P(\mathbb{R}^d \times \mathbb{R}_{\ge 0}) \to \mathbb{R}$ *layerwise-Wasserstein convex* if for any $\Omega \in P(P(\mathbb{R}^d \times \mathbb{R}_{\ge 0}))$,

$$\mathcal{F}(\mathrm{Bar}^{\mathrm{LW}}(\Omega)) \le \int \mathcal{F}(\mu)\, d\Omega(\mu).$$

This notion of convexity may potentially play an important role in applications. Given a family of root systems, corresponding to a family of genetically identical plants, grown under identical environmental conditions, we will in forthcoming work propose interpreting the layerwise-Wasserstein barycenter of the systems as the single root system which best represents the family. It is natural to compare phenotypes (for instance, center of mass, variance, entropy, total root length, etc.) of that barycenter with the phenotypes of the actual observed roots in the original family. If these phenotypes (interpreted as functionals on the space of measures) are layerwise-Wasserstein convex, the phenotype of the barycenter is always less than the average of the phenotypes of the samples.

The theory of layerwise convexity, which we begin to develop below, has a strong connection to the theory of displacement convexity, or convexity along geodesics with respect to the Wasserstein metric, introduced in [McCann 1997], and its extension to convexity over Wasserstein barycenters, introduced in [Agueh and Carlier 2011].

We begin with the Shannon entropy, perhaps the best known displacement convex functional. As we show below, it is also layerwise-Wasserstein convex. For roots, it can be regarded as a measure of the concentration of mass and therefore has potential biological interest. Given $\mu \in P(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$, with

$$\mu(x, y) = f(x, y)\, dx\, dy, \quad \text{where } x \in \mathbb{R}^d,\ y \in \mathbb{R}_{\geq 0},$$

the vertical marginal μ^V has density

$$f^V(y) = \int_{\mathbb{R}^d} f(x, y)\, dx.$$

Note that for fixed y, the probability measure

$$d\mu_y(x) = \frac{f(x, y)}{f^V(y)}\, dx$$

coincides with $\tilde{\mu}_l$ for $l = F_\mu(y)$. Recall that the Shannon entropy of μ is defined as

$$S(\mu) = \int_{\mathbb{R}^d} \int_{\mathbb{R}_{\geq 0}} f(x, y) \log f(x, y)\, dx\, dy.$$

This formula allows one to rewrite $S(\mu)$ using the layerwise decomposition.

Proposition 4.1. *The Shannon entropy $S(\mu)$ satisfies*

$$S(\mu) = \int_{\mathbb{R}_{\geq 0}} S(\mu_y)\, d\mu^V(y) + S(\mu^V) = \int_0^1 S(\tilde{\mu}_l)\, dl + S(\mu^V).$$

Proof. The proof is a calculation. In the following we use the standard convention that $0 \log(0) = 0$. We have

$$\begin{aligned}
S(\mu) &= \int_{\mathbb{R}_{\geq 0}} \int_{\mathbb{R}^d} f(x, y) \log\left[\frac{f(x, y) f^V(y)}{f^V(y)}\right] dx\, dy \\
&= \int_{\mathbb{R}_{\geq 0}} \int_{\mathbb{R}^d} f(x, y) \left[\log\left[\frac{f(x, y)}{f^V(y)}\right] + \log f^V(y)\right] dx\, dy \\
&= \int_{\mathbb{R}_{\geq 0}} \int_{\mathbb{R}^d} f(x, y) \log\left[\frac{f(x, y)}{f^V(y)}\right] dx\, dy + \int_{\mathbb{R}} \left(\int_{\mathbb{R}^d} f(x, y)\, dx\right) \log f^V(y)\, dy \\
&= \int_{\mathbb{R}_{\geq 0}} \int_{\mathbb{R}^d} f^V(y) \frac{f(x, y)}{f^V(y)} \log\left[\frac{f(x, y)}{f^V(y)}\right] dx\, dy + \int_{\mathbb{R}} f^V(y) \log f^V(y)\, dy \\
&= \int_{\mathbb{R}_{\geq 0}} \int_{\mathbb{R}^d} \left(\frac{f(x, y)}{f^V(y)} \log\left[\frac{f(x, y)}{f^V(y)}\right] dx\right) f^V(y)\, dy + S(\mu^V) \\
&= \int_{\mathbb{R}_{\geq 0}} S(\mu_y)\, d\mu^V(y) + S(\mu^V).
\end{aligned}$$

The final equality follows by noting that the cumulative distribution function F_μ satisfies $F'_\mu(y) = f^V(y)$ and by changing variables from y to $l = F_\mu(y)$. □

Corollary 4.2. *The Shannon entropy is layerwise-Wasserstein convex.*

Proof. Recall that from Proposition 2.5 the layerwise-Wasserstein interpolation of $\Omega \in P(P(\mathbb{R}^d \times \mathbb{R}_{\geq 0}))$ amounts to constructing the probability measure η whose vertical marginal η^V is the Wasserstein barycenter of Ω^V, and whose conditional probabilities $\tilde{\eta}_l$ are the Wasserstein barycenters of the $\widetilde{\Omega}_l$. Since the entropy $S(\mu)$ depends additively on μ^V and the $\tilde{\mu}_l$, Wasserstein convexity of the entropy (see [Kim and Pass 2017] for convexity with respect to general barycenters) yields the result. □

Many phenotypes of interest concern only the depth of the root, and not its horizontal distribution of mass (since, for instance, nutrient concentration in soil is largely determined by depth). Therefore the following simple observation is relevant.

Proposition 4.3. *Any Wasserstein convex function of the vertical marginal is layerwise Wasserstein convex.*

Proof. This follows immediately from the structure of the layerwise-Wasserstein distance, given in Proposition 2.5. □

Example 4.4. Let us list a few examples of functionals with possible biological applications, which are layerwise-Wasserstein convex by Proposition 4.3:

- The vertical mean

$$\mu \mapsto \bar{y} := \int_{\mathbb{R}^d \times \mathbb{R}_{\geq 0}} y \, d\mu(x, y) = \int_{\mathbb{R}_{\geq 0}} y \, d\mu^V(y);$$

one can verify easily that this is in fact affine along displacement (and hence layerwise-Wasserstein) interpolations.

- The vertical variance

$$\int_{\mathbb{R}^d \times \mathbb{R}_{\geq 0}} |y - \bar{y}|^2 \, d\mu(x, y) = \int_{\mathbb{R}_{\geq 0}} |y - \bar{y}|^2 \, d\mu^V(y),$$

a measure the spread of the mass in the vertical direction [Kim and Pass 2020].

- The vertical internal energy

$$\int_{\mathbb{R}_{\geq 0}} (f^V(y))^r \, dy \quad \text{for } r \geq 1.$$

- Vertical quantiles $F_\mu^{-1}(l)$ for each fixed $l \in (0, 1)$. For instance, the vertical median $\left(l = \frac{1}{2}\right)$ is the depth above which half the mass of the root lies. The 100th quantile (the maximal depth of the root) is often called the rooting depth, while the 87th quantile $\left(l = \frac{87}{100}\right)$ is a conventional phenotype often used as a measure of the root depth. The displacement convexity (and layerwise-Wasserstein convexity) of these follows immediately from the monotone structure of one-dimensional optimal couplings with respect to the distance-squared cost; in fact, it is layerwise-Wasserstein affine.

Although, unlike the examples above, it is not a functional of the vertical marginal, the structure of the layerwise-Wasserstein distance easily implies that the class of functionals in the following example are layerwise-Wasserstein convex as well.

Example 4.5. Any functional of the form

$$\bar{\mathcal{F}}(\mu) = \int_0^1 \bar{\mathcal{F}}_l(\tilde{\mu}_l)\, dl,$$

where each $\bar{\mathcal{F}}_l$ is Wasserstein convex, is layerwise-Wasserstein convex.

Acknowledgements

Kim and Pass are partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Discovery Grant program. Schneider is supported by the Canada Excellence in Research Chair in Food Systems and Security and the Global Institute for Food Security.

The authors gratefully acknowledge the support of the Pacific Institute of Mathematical Sciences (PIMS); in particular, they would like to thank PIMS director James Colliander for initiating this project by bringing the three of them together. Part of this work was done while the authors participated in a research-in-teams program at the Banff International Research Station (BIRS), and we thank BIRS for the excellent working environment provided there.

References

[Agueh and Carlier 2011] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space", *SIAM J. Math. Anal.* **43**:2 (2011), 904–924. MR Zbl

[Anderes et al. 2016] E. Anderes, S. Borgwardt, and J. Miller, "Discrete Wasserstein barycenters: optimal transport for discrete data", *Math. Methods Oper. Res.* **84**:2 (2016), 389–409. MR Zbl

[Backhoff et al. 2017] J. Backhoff, M. Beiglböck, Y. Lin, and A. Zalashko, "Causal transport in discrete time and applications", *SIAM J. Optim.* **27**:4 (2017), 2528–2562. MR Zbl

[Bonneel et al. 2015] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, "Sliced and Radon Wasserstein barycenters of measures", *J. Math. Imaging Vision* **51**:1 (2015), 22–45. MR Zbl

[Bressan and Sun 2018] A. Bressan and Q. Sun, "On the optimal shape of tree roots and branches", *Math. Models Methods Appl. Sci.* **28**:14 (2018), 2763–2801. MR Zbl

[Bressan et al. 2020] A. Bressan, M. Palladino, and Q. Sun, "Variational problems for tree roots and branches", *Calc. Var. Partial Differential Equations* **59**:1 (2020), art. id. 7. MR Zbl

[Bucksch 2014] A. Bucksch, "A practical introduction to skeletons for the plant sciences", *Appl. Plant Sci.* **2**:8 (2014), art. id. 1400005.

[Carlier et al. 2010] G. Carlier, A. Galichon, and F. Santambrogio, "From Knothe's transport to Brenier's map and a continuation method for optimal transport", *SIAM J. Math. Anal.* **41**:6 (2010), 2554–2576. MR Zbl

[Clark et al. 2011] R. T. Clark, R. B. MacCurdy, J. K. Jung, J. E. Shaff, S. R. McCouch, D. J. Aneshansley, and L. V. Kochian, "Three-dimensional root phenotyping with a novel imaging and software platform", *Plant Physiol.* **156**:2 (2011), 455–465.

[Clark et al. 2013] R. T. Clark, A. N. Famoso, K. Zhao, J. E. Shaff, E. J. Craft, C. D. Bustamente, S. R. McCouch, D. J. Aneshansley, and L. V. Kochian, "High-throughput two-dimensional root system phenotyping platform facilitates genetic analysis of root growth and development", *Plant Cell Environ.* **36**:2 (2013), 454–466.

[Famoso et al. 2010] A. N. Famoso, R. T. Clark, J. E. Shaff, E. Craft, S. R. McCouch, and L. V. Kochian, "Development of a novel aluminum tolerance phenotyping platform used for comparisons of cereal aluminum tolerance and investigations into rice aluminum tolerance mechanisms", *Plant Physiol.* **153**:4 (2010), 1678–1691.

[Famoso et al. 2011] A. N. Famoso, K. Zhao, R. T. Clark, C.-W. Tung, M. H. Wright, C. Bustamente, L. V. Kochian, and S. R. McCouch, "Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping", *PLoS Genetics* **7**:8 (2011), art. id. e1002221.

[Hatcher 2002] A. Hatcher, *Algebraic topology*, Cambridge Univ. Press, 2002. MR Zbl

[Kim and Pass 2017] Y.-H. Kim and B. Pass, "Wasserstein barycenters over Riemannian manifolds", *Adv. Math.* **307** (2017), 640–683. MR Zbl

[Kim and Pass 2020] Y.-H. Kim and B. Pass, "Nonpositive curvature, the variance functional, and the Wasserstein barycenter", *Proc. Amer. Math. Soc.* **148**:4 (2020), 1745–1756. MR Zbl

[Knothe 1957] H. Knothe, "Contributions to the theory of convex bodies", *Michigan Math. J.* **4** (1957), 39–52. MR Zbl

[McCann 1997] R. J. McCann, "A convexity principle for interacting gases", *Adv. Math.* **128**:1 (1997), 153–179. MR Zbl

[Muzellec and Cuturi 2019] B. Muzellec and M. Cuturi, "Subspace detours: building transport plans that are optimal on subspace projections", preprint, 2019. arXiv

[Piñeros et al. 2016] M. A. Piñeros, B. G. Larson, J. E. Shaff, D. J. Schneider, A. X. Falcão, L. Yuan, R. T. Clark, E. J. Craft, T. W. Davis, P.-L. Pradier, N. M. Shaw, I. Assaranurak, S. R. McCouch, C. Sturrock, M. Bennett, and L. V. Kochian, "Evolving technologies for growing, imaging and analyzing 3D root system architecture of crop plants", *J. Integr. Plant Biol.* **58**:3 (2016), 230–241.

[Rosenblatt 1952] M. Rosenblatt, "Remarks on a multivariate transformation", *Ann. Math. Stat.* **23** (1952), 470–472. MR Zbl

[Santambrogio 2015] F. Santambrogio, *Optimal transport for applied mathematicians*, Progr. Nonlin. Diff. Eq. Appl. **87**, Birkhäuser, Cham, 2015. MR Zbl

[Santambrogio and Wang 2016] F. Santambrogio and X.-J. Wang, "Convexity of the support of the displacement interpolation: counterexamples", *Appl. Math. Lett.* **58** (2016), 152–158. MR Zbl

[Toroczkai 2002] Z. Toroczkai, "Topological classification of binary trees using the Horton–Strahler index", *Phys. Rev. E* (3) **65**:1 (2002), art. id. 016130. MR

[Villani 2003] C. Villani, *Topics in optimal transportation*, Grad. Studies in Math. **58**, Amer. Math. Soc., Providence, RI, 2003. MR Zbl

[Villani 2009] C. Villani, *Optimal transport*: *old and new*, Grundlehren der Math. Wissenschaften **338**, Springer, 2009. MR Zbl

Received 10 Oct 2019. Revised 21 Feb 2020. Accepted 28 Mar 2020.

YOUNG-HEON KIM: yhkim@math.ubc.ca
Department of Mathematics, University of British Columbia, Vancouver, BC, Canada

BRENDAN PASS: pass@ualberta.ca
Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada

DAVID J. SCHNEIDER: dave.schneider@gifs.ca
Global Institute for Food Security, University of Saskatchewan, Saskatoon, SK, Canada

PURE and APPLIED
ANALYSIS

Vol. 2, No. 3, 2020
https://doi.org/10.2140/paa.2020.2.603

L^p ESTIMATES FOR BAOUENDI–GRUSHIN OPERATORS

GIORGIO METAFUNE, LUIGI NEGRO AND CHIARA SPINA

We prove L^p estimates for the Baouendi–Grushin operator $\Delta_x + |x|^\alpha \Delta_y$ in $L^p(\mathbb{R}^{N+M})$, $1 < p < \infty$, where $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$. When $p = 2$ more general weights belonging to the reverse Hölder class $B_2(\mathbb{R}^N)$ are allowed.

1. Introduction

We prove L^p estimates for the Baouendi–Grushin operator $L = \Delta_x + |x|^\alpha \Delta_y$ in $L^p(\mathbb{R}^{N+M})$, $1 < p < \infty$, where $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$; more specifically, we prove the L^p boundedness of the operators $D_{x_i x_j} L^{-1}$, $|x|^\alpha D_{y_i y_j} L^{-1}$ and $|x|^{\frac{\alpha}{2}} D_{x y_i} L^{-1}$. We use these results to characterize the domain of the operator L, denoted by $D_p(L)$, where the solution of the equation $\lambda u - Lu = f$ exists and is unique for any $f \in L^p(\mathbb{R}^{N+M})$ and $\lambda > 0$. In an equivalent way, we describe the domain under which L generates an (analytic and symmetric) semigroup in $L^p(\mathbb{R}^{N+M})$.

If α is an even integer, the operator is hypoelliptic but is not a sublaplacian in the sense of [Folland 1975] and our estimates seem to be known only for $\alpha = 2$ and even $N \geq 2$; see [Koch et al. 2015]. When α is an unrestricted positive real number, many results are known on local regularity of the equation $Lu = f$; see for example [Franchi and Serapioni 1987; Franchi et al. 1994; Franchi and Lanconelli 1984; Garofalo and Vassilev 2007] for the unique continuation property. We refer to [Robinson and Sikora 2008] for heat kernel estimates even in a more general context. However, we are not aware of global regularity results for the second derivatives of u, with the exception of [Wang 2003], where global Hölder regularity is proved for every $\alpha > 0$ and of [Kim 1999], where L^p estimates are proved, when $\alpha = 1$ and $N = 1$, in the half-plane $x > 0$ for the inhomogeneous problem $Lu = f$, $u(0, y) = g(y)$. When $g = 0$ the estimates in [Kim 1999] reduce to ours: even though our results are valid in the whole space, they can be rephrased when $N = 1$ in the half-space $x > 0$ for Dirichlet or Neumann boundary conditions by considering odd and even functions with respect to x, respectively.

We prove L^p estimates through an interpolation theorem in the absence of kernels in homogeneous spaces due to Z. Shen [2005, Theorem 3.1], see also [Auscher and Martell 2007, Theorem 3.14], and weighted mean value inequalities for subsolutions of the elliptic equation $Lu = 0$, with respect to the balls associated with the subelliptic distance defined by the operator, proved in [Franchi and Serapioni 1987; Chanillo and Wheeden 1986]. Some of these results can probably be generalized to the case when $|x|^\alpha$ is replaced by a weight function $\phi(x)$ belonging to the reverse Hölder class $B_p(\mathbb{R}^N)$. This is the

MSC2010: 35H20, 35J70, 47F05.
Keywords: Baouendi–Grushin operators, degenerate elliptic equations, subelliptic equations, L^p estimates.

case when $p = 2$, where the result is not obtained via integration by parts but using maximal inequalities due to P. Auscher and B. Ben Ali [2007] for Schrödinger operators with B_2 potentials. However, local estimates for subsolutions seem to be known only in special cases and they are crucial in our approach when $p \neq 2$. Another restriction comes from the estimates of the mixed derivatives, that is, for the operator $|x|^{\frac{\alpha}{2}} D_{xy_i} L^{-1}$, where our proof relies on scaling. In order to unify our approach and to improve the readability, we consider only the $|x|^\alpha$ case in our L^p estimates. Perturbation arguments from this model case allow us to treat different powers near 0 and ∞ or power-like behavior, but we prefer not to deal with these variants here.

The paper is organized as follows. In Section 2 we define the operator in $L^2(\mathbb{R}^{N+M})$ through a form and prove L^2 estimates via partial Fourier transform and maximal regularity results on Schrödinger operators. In Section 3 we briefly recall the subelliptic distance associated with L and the main geometrical objects needed in L^p estimates. The latter are proved in Section 4, where a separate subsection deals with mixed derivatives.

Notation. We use L^p for $L^p(\mathbb{R}^{N+M})$ and C_c^∞ for $C_c^\infty(\mathbb{R}^{N+M})$. L_c^∞ stands for the space of all bounded measurable functions on $\mathbb{R}^{N+M}$ having compact support. $\mathcal{S}$ is the Schwartz space and $\mathcal{S}'$ the space of tempered distributions. We also write $B(r) := \{x \in \mathbb{R}^N : |x| < r\}$, $B(x_0, r) = x_0 + B(r)$.

2. L^2 estimates

Let $\phi : \mathbb{R}^N \to [0, +\infty[$ be a nonnegative continuous function and set

$$\mathbb{R}^N \supseteq \Omega_N = \{x \in \mathbb{R}^N : \phi(x) > 0\}, \qquad \Omega = \Omega_N \times \mathbb{R}^M.$$

Let L be the operator defined on smooth functions by

$$L = \Delta_x + \phi(x)\Delta_y,$$

where $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$. Setting

$$a = \left(\begin{array}{c|c} I_N & 0 \\ \hline 0 & \phi I_M \end{array}\right) = (a_{ij})$$

or

$$a_{ij}(x, y) = \begin{cases} 1 & \text{if } i = j \leq N, \\ \phi(x) & \text{if } N + 1 \leq i = j \leq N + M \end{cases} \tag{1}$$

and 0 elsewhere, we can write

$$L = \operatorname{div}(a\,\nabla)$$

and, therefore, L is formally self-adjoint with respect to the Lebesgue measure.

Remark 2.1. L is nondegenerate in the x-direction but degenerates in the y-direction outside Ω. Accordingly, $\nabla_x u$ will denote the distributional gradient (with respect to x) of u in the whole space $\mathbb{R}^{N+M}$ and $\nabla_y u$ only its distributional gradient (with respect to y) in Ω.

We give a formal definition of L through a symmetric form:

Definition 2.2. Consider the sesquilinear form $\mathfrak{a}$ in L^2 defined by

$$\mathfrak{a}(u, v) := \int_{\mathbb{R}^{N+M}} [\langle \nabla_x u, \nabla_x \bar{v}\rangle + \phi(x)\langle \nabla_y u, \nabla_y \bar{v}\rangle]\, dx\, dy,$$
$$D(\mathfrak{a}) := \{u \in L^2 : u \in H^1_{\text{loc}}(\Omega),\ \nabla_x u,\ \phi^{\frac{1}{2}} \nabla_y u \in L^2\}.$$

According to the remark above, we require that the weak gradient $\nabla_y u$ exists only in Ω.

We summarize in the following lemma the main properties of a. Note that, due to the assumptions on ϕ, $\mathfrak{a}$ is locally uniformly elliptic on Ω.

Lemma 2.3. *The form $\mathfrak{a}$ is densely defined, nonnegative, symmetric and closed in L^2 and the following properties hold:*

(i) *If Q is an orthogonal matrix in $\mathbb{R}^M$, $y_0 \in \mathbb{R}^M$ and $I_{Q+y_0}u(x, y) = u(x, Qy + y_0)$, then for every $u, v \in D(\mathfrak{a})$, one has $I_{Q+y_0}u$, $I_{Q+y_0}v \in D(\mathfrak{a})$ and*

$$\mathfrak{a}(I_{Q+y_0}u, I_{Q+y_0}v) = \mathfrak{a}(u, v).$$

(ii) *If ϕ is homogeneous of degree $\alpha \geq 0$, i.e., $\phi(sx) = s^\alpha \phi(x)$ for $x \in \mathbb{R}^N$, $s > 0$, then defining the dilation*

$$I_s u(x, y) = u(sx, s^{\frac{2+\alpha}{2}} y),$$

for every $u, v \in D(\mathfrak{a})$ one has $I_s u$, $I_s v \in D(\mathfrak{a})$ and

$$\mathfrak{a}(I_s u, I_s v) = s^{2-N-\frac{2+\alpha}{2}M} \mathfrak{a}(u, v).$$

Proof. Clearly, due to the positivity of ϕ, $\mathfrak{a}$ is a nonnegative symmetric form in L^2. The closedness of the form follows easily since $\mathfrak{a}$ is locally uniformly elliptic in Ω. The proofs of (i) and (ii) follow by a straightforward computation. □

Let $-L$ be the operator associated with $\mathfrak{a}$, that is,

$$D(L) := \{u \in D(\mathfrak{a}) : \text{there exists } v \in L^2 \text{ such that } \mathfrak{a}(u, w) = \textstyle\int_{\mathbb{R}^{N+M}} v\bar{w}\, dx\, dy \text{ for all } w \in D(\mathfrak{a})\},$$
$$-Lu := v. \tag{2}$$

The basic properties of L are listed below.

Proposition 2.4. *The operator $-L$ defined in (2) is nonnegative and self-adjoint. Moreover:*

(i) *$C_c^\infty \hookrightarrow D(L) \hookrightarrow \{u \in L^2 \cap W^{2,2}_{\text{loc}}(\Omega)\ ;\ Lu \in L^2\}$ and for every $u \in C_c^\infty$*

$$Lu = \Delta_x u + \phi(x)\Delta_y u.$$

(ii) *L generates a contractive analytic semigroup $\{e^{zL} : z \in \mathbb{C}_+\}$ in L^2.*

(iii) *The semigroup $\{e^{tL} : t > 0\}$ is submarkovian; i.e., it is positive and L^∞-contractive.*

(iv) *If Q is an orthogonal matrix in $\mathbb{R}^M$ and $y_0 \in \mathbb{R}^M$, then*

$$L = I^{-1}_{Q+y_0} L I_{Q+y_0}, \quad I_{Q+y_0}u(x, y) = u(x, Qy + y_0).$$

(v) *If ϕ is homogeneous of degree α, then*

$$s^2L = I_s^{-1}LI_s, \quad I_su(x, y) = u(sx, s^{\frac{2+\alpha}{2}}y), \quad s > 0.$$

Proof. Part (i) is clear by construction and from interior elliptic regularity (see, however, the proof of Theorem 2.9 for justifying the integration by parts). The generation property of L follows by standard results, see [Ouhabaz 2005, Chapter 1, Section 4]; the positivity of e^{tL} as well its L^∞-contractivity is a consequence of the Beurling–Deny criteria satisfied by the form $\mathfrak{a}$, see [Ouhabaz 2005, Corollary 2.18], and note that $\mathfrak{a}$ is real; that is, $\mathfrak{a}(u, v) \in \mathbb{R}$ whenever u, v are real functions. Concerning (iv) and (v), let $u \in D(L)$, $v \in D(\mathfrak{a})$ and $s > 0$. Then

$$\begin{aligned}\mathfrak{a}(I_su, v) &= s^{2-N-\frac{2+\alpha}{2}M}\mathfrak{a}(u, I_{s^{-1}}v)\\ &= -s^{2-N-\frac{2+\alpha}{2}M}\int_{\mathbb{R}^{N+M}}(Lu)I_{s^{-1}}\bar{v}\,dx\,dy = -s^2\int_{\mathbb{R}^{N+M}}(I_sLu)\bar{v}\,dx\,dy;\end{aligned}$$

hence $I_su \in D(L)$ and $LI_su = s^2I_sLu$. The proof for I_{Q+y_0} is similar. □

The following proposition shows that C_c^∞ is dense in $D(L)$ with respect to the graph norm.

Proposition 2.5. *C_c^∞ is a core for the operator $(L, D(L))$ and the form $\mathfrak{a}$.*

Proof. Since $I - L$ is invertible we have to show that $(I - L)(C_c^\infty)$ is dense in L^2 or, equivalently, that $(I - L)(C_c^\infty)^\perp = \{0\}$. To this aim let $v \in L^2$ such that

$$\int_{\mathbb{R}^{N+M}}(I - L)u\,\bar{v}\,dx\,dy = 0 \quad \text{for all } u \in C_c^\infty.$$

Taking the partial Fourier transform with respect to the y-variable and applying the Fubini and Plancherel theorems we get

$$\int_{\mathbb{R}^{N+M}}[\hat{u}(x,\xi) - \Delta_x\hat{u}(x,\xi) + \phi(x)|\xi|^2\hat{u}(x,\xi)]\,\bar{\hat{v}}(x,\xi)\,dx\,d\xi = 0 \quad \text{for all } u \in C_c^\infty.$$

Choosing $u = A(x)B(y) \in C_c^\infty$ we have $\hat{u}(x,\xi) = A(x)\widehat{B}(\xi)$ and

$$\int_{\mathbb{R}^{N+M}}[A(x) - \Delta_xA(x) + \phi(x)|\xi|^2A(x)]\,\widehat{B}(\xi)\bar{\hat{v}}(x,\xi)\,dx\,d\xi = 0. \tag{3}$$

Fix $\xi_0 \in \mathbb{R}^M$, $r > 0$ and let

$$w(\xi) = \frac{1}{|B(\xi_0, r)|}\chi_{B(\xi_0,r)} \in L^2(\mathbb{R}^M).$$

Let $(B_n)_n \in C_c^\infty(\mathbb{R}^M)$ be a sequence of test functions such that $B_n \to \check{w}$ in $L^2(\mathbb{R}^M)$; then $\widehat{B}_n \to w$ in $L^2(\mathbb{R}^M)$ and taking the limit as $n \to \infty$ in (3) with $\widehat{B}$ replaced by $\widehat{B}_n$ we obtain

$$\frac{1}{|B(\xi_0, r)|}\int_{B(\xi_0,r)}d\xi\int_{\mathbb{R}^N}[A(x) - \Delta_xA(x) + \phi(x)|\xi|^2A(x)]\,\bar{\hat{v}}(x,\xi)\,dx = 0.$$

Letting $r \to 0$ and using the Lebesgue differentiation theorem, we have for a.e. $\xi_0 \in \mathbb{R}^M$

$$\int_{\mathbb{R}^N}[A(x) - \Delta_xA(x) + \phi(x)|\xi_0|^2A(x)]\,\bar{\hat{v}}(x,\xi_0)\,dx = 0,$$

which, since u was arbitrary, is valid for every $A \in C_c^\infty(\mathbb{R}^N)$. The operator $\Delta_x - \phi(\,\cdot\,)|\xi|^2$ is a Schrödinger operator in $L^2(\mathbb{R}^N)$ with nonpositive potential $-\phi|\xi|^2 \in L^2_{\text{loc}}(\mathbb{R}^N)$ and $C_c^\infty(\mathbb{R}^N)$ is dense in the domain $D(\Delta_x - \phi(\,\cdot\,)|\xi|^2)$ with respect to the graph norm; see [Kato 1972]. The last equation then implies $\hat{v}(\,\cdot\,, \xi_0) = 0$ for a.e. $\xi_0 \in \mathbb{R}^M$, which proves the required claim.

Since $D(L)$ is dense in $D(L^{1/2}) = D(\mathfrak{a})$, see [Kato 1966, Theorem VI.2.23], the second statement follows from the first. □

In order to prove the main result of this section we recall the definition of B_p-weights. Let $1 < p \le \infty$. Then $\omega \in B_p(\mathbb{R}^N)$, where $B_p(\mathbb{R}^N)$ is the class of the reverse Hölder weights of order p, if $\omega \in L^p_{\text{loc}}$, $\omega > 0$ a.e. and there exists a positive constant C such that the inequality

$$\left(\frac{1}{|B|}\int_B \omega^p\right)^{\frac{1}{p}} \le \frac{C}{|B|}\int_B \omega \tag{4}$$

holds for every ball B. If $p = \infty$, the left-hand side of the inequality above has to be replaced by the essential supremum of ω on B. The smallest positive constant C such that (4) holds is the B_p constant of ω. We recall that powers $|x|^\alpha$ belong to $B_\infty(\mathbb{R}^N)$ whenever $\alpha \ge 0$. This is easily seen first considering balls of radius 1 (and large centers) and then scaling.

Remark 2.6. In the proof of the following result we need the maximal L^2 inequalities for Schrödinger operators $\Delta - V$, $0 \le V \in B_2(\mathbb{R}^N)$, shown in [Auscher and Ben Ali 2007, Theorem 1.1, Corollary 1.3]. They say that the operator $V(\Delta - V)^{-1}$ is bounded in $L^2(\mathbb{R}^N)$; moreover, the norm of $V(\Delta - V)^{-1}$ is bounded by a constant which depends only on N and the B_2 constant of V. This last fact is not explicitly stated in [Auscher and Ben Ali 2007], even though it follows from the proofs, but can be found in [Carbonaro et al. 2008, Theorem 3.6] in the more general setting of parabolic Schrödinger operators $D_t - \Delta + V$. In fact the norm of $V(\Delta - V)^{-1}$ depends on C in (4) and the constant in the Harnack inequality for the Laplacian in $\mathbb{R}^N$. See also [Shen 1995, Theorem 0.3] where, however, $N \ge 3$ and $V \in B_q$ for some $q \ge \frac{N}{2}$.

Theorem 2.7. *Assume that $\phi : \mathbb{R}^N \to [0, +\infty[$ belongs to $B_2(\mathbb{R}^N)$. Then for every $1 \le i, j \le N$, $1 \le h, k \le M$, one has*

$$\|D_{x_i x_j} u\|_2 + \|\phi D_{y_h y_k} u\|_2 \le C\|Lu\|_2, \quad u \in D(L).$$

Moreover

$$\|\nabla_x u\|_2 + \|\phi^{\frac{1}{2}} \nabla_y u\|_2 \le C(\|Lu\|_2 + \|u\|_2), \quad u \in D(L).$$

Proof. By Proposition 2.5 we may assume that $u \in C_c^\infty$. Consider the partial Fourier transform with respect to the y-variable. Let $v(x, \xi) = \hat{u}(x, \xi)$. Then, setting $Lu = f$, we have

$$\Delta_x v(x, \xi) - \phi(x)|\xi|^2 v(x, \xi) = \hat{f}(x, \xi) \in L^2.$$

Observe now that, for every fixed $\xi \in \mathbb{R}^M$, $\Delta_x - \phi(\,\cdot\,)|\xi|^2$ is a Schrödinger operator in $\mathbb{R}^N$ with potential $\phi|\xi|^2$. Moreover, since $\phi \in B_2(\mathbb{R}^N)$, it immediately follows that $\phi|\xi|^2$ satisfies the reverse Hölder condition with the same constant as ϕ. By Remark 2.6 above, we have

$$|\xi|^4 \int_{\mathbb{R}^N} \phi(x)^2 |v(x,\xi)|^2\, dx \le C \int_{\mathbb{R}^N} |\hat{f}(x,\xi)|^2\, dx,$$

with a constant C not depending on ξ. Integrating the last inequality over $\mathbb{R}^M$, we get

$$\int_{\mathbb{R}^{N+M}} |\xi|^4 \phi(x)^2 |v(x,\xi)|^2\, dx\, d\xi \le C \int_{\mathbb{R}^{N+M}} |\hat{f}(x,\xi)|^2\, dx\, d\xi.$$

Since $|\cdot|^2 v(x,\cdot) = \widehat{\Delta_y u}(x,\cdot)$ we get, using Fubini's theorem and the Plancherel theorem in $\mathbb{R}^M$,

$$\int_{\mathbb{R}^{N+M}} \phi(x)^2 |\Delta_y u(x,y)|^2\, dx\, dy \le C \int_{\mathbb{R}^{N+M}} |f(x,y)|^2\, dx\, dy,$$

which reads as $\|\phi \Delta_y u\|_2 \le C\|Lu\|_2$; by difference we also get $\|\Delta_x u\|_2 \le C\|Lu\|_2$.

The Calderón–Zygmund theorem applied separately to each variable implies

$$\|D_{x_i x_j} u\|^2_{L^2(\mathbb{R}^N)} \le C(N)\|\Delta_x u\|^2_{L^2(\mathbb{R}^N)}, \quad 1 \le i, j \le N,$$
$$\|D_{y_h y_k} u\|^2_{L^2(\mathbb{R}^M)} \le C(M)\|\Delta_y u\|^2_{L^2(\mathbb{R}^M)}, \quad 1 \le h, k \le M.$$

Integrating the previous inequalities (with the last one multiplied by $\phi(x)^2$) over $\mathbb{R}^M$ and $\mathbb{R}^N$, respectively, we get the first claim.

Concerning the gradient estimates, it is enough to observe that, by interpolation, for every $\epsilon > 0$,

$$\|\nabla_x\|_{L^2(\mathbb{R}^N)} \le \epsilon \sum_{i,j=1}^{N} \|D_{x_i x_j} u\|_{L^2(\mathbb{R}^N)} + \frac{C}{\epsilon}\|u\|_{L^2(\mathbb{R}^N)}.$$

The estimates for the first-order derivatives with respect to x immediately follow after integration over $\mathbb{R}^M$ and by using the first part of the theorem. For the gradient with respect to y, we start, analogously, from

$$\|\nabla_y\|_{L^2(\mathbb{R}^M)} \le \epsilon \sum_{h,k=1}^{M} \|D_{y_h y_k} u\|_{L^2(\mathbb{R}^M)} + \frac{C}{\epsilon}\|u\|_{L^2(\mathbb{R}^M)}.$$

Choosing $\epsilon = \phi(x)^{1/2}$, the claim follows after the integration over $\mathbb{R}^N$ and by using the first part of the theorem. □

Remark 2.8. If $\phi \in B_N(\mathbb{R}^N)$ and $N \ge 3$, it can be proved along the same lines that

$$\|\phi^{\frac{1}{2}} D_{x_i y_h} u\|_2 \le C\|Lu\|_2, \quad u \in D(L).$$

In fact, the partial Fourier transform of $D_{x_i y_h} u$ is $-i\xi_h D_{x_i}\hat{u}(x,\xi)$ and $\phi|\xi|^2$ satisfies the $B_N(\mathbb{R}^N)$ reverse Hölder condition with the same constant as ϕ. By [Shen 1995, Theorem 0.8]

$$|\xi|^2 \int_{\mathbb{R}^N} \phi(x) |\nabla_x v(x,\xi)|^2\, dx \le C \int_{\mathbb{R}^N} |\hat{f}(x,\xi)|^2\, dx, \tag{5}$$

with a constant C not depending on ξ. Integrating over $\mathbb{R}^M$ and using Plancherel's theorem, we get

$$\int_{\mathbb{R}^{N+M}} \phi(y) |D_{x_i y_h} u(x,y)|^2\, dx\, dy \le C \int_{\mathbb{R}^{N+M}} |f(x,y)|^2\, dx\, dy.$$

The above result probably holds also for $N = 1, 2$ since the maximal inequality (5) is discussed in [Auscher and Ben Ali 2007] (see the comments after Corollary 1.5); the authors say that their methods give the result for all N but a detailed proof is not given. Since we shall not use this remark in what follows we omit further investigation.

In the following theorem we characterize the domain of the operator L.

Theorem 2.9. *If $\phi \in B_2(\mathbb{R}^N)$ then the domain of the operator L defined in (2) satisfies*

$$D(L) = \{u \in L^2 : \nabla_x u,\ D_{x_i x_j} u,\ \phi^{\frac{1}{2}} \nabla_y u,\ \phi D_{y_h y_k} u \in L^2\}. \tag{6}$$

Proof. Let $\widetilde{D}(L)$ be the set defined in the right-hand side of equality (6). Theorem 2.7 then implies $D(L) \subseteq \widetilde{D}(L)$. To prove the equality it is then enough to prove that the operator $(L, \widetilde{D}(L))$ is dissipative since in this case $I - L : \widetilde{D}(L) \to L^2$ is an injective extension of the resolvent operator $I - L : D(L) \to L^2$ and so both operators must coincide. Let $u \in \widetilde{D}(L)$; then, by the definition, for every compact set $\omega \Subset \Omega$, we have $u, D^2 u \in L^2(\omega)$; hence $u \in H^2_{\text{loc}}(\Omega)$. Moreover a section argument, see for example [Ziemer 1989, Theorem 2.1.4], shows that for a.e. $x \in \Omega_N$, we have $u(x, \cdot) \in H^2(\mathbb{R}^M)$ and

$$\int_{\mathbb{R}^M} u\, \Delta_y u\, dy = -\int_{\mathbb{R}^M} |\nabla_y u|^2\, dy \quad \text{for a.e. } x \in \Omega_N.$$

Then multiplying by ϕ, integrating in x and using Fubini's theorem we get

$$\int_{\mathbb{R}^{N+M}} \phi(x)\, u\, \Delta_y u\, dx\, dy = -\int_{\mathbb{R}^{N+M}} \phi(x) |\nabla_y u|^2\, dx\, dy.$$

Analogous reasoning applied to the y-sections shows that

$$\int_{\mathbb{R}^{N+M}} u\, \Delta_x u\, dx\, dy = -\int_{\mathbb{R}^{N+M}} |\nabla_x u|^2\, dx\, dy.$$

The last two inequalities imply

$$\int_{\mathbb{R}^{N+M}} u\, Lu\, dx\, dy = -\int_{\mathbb{R}^{N+M}} (|\nabla_y u|^2 + \phi(x) |\nabla_y u|^2)\, dx\, dy \le 0,$$

which, since $u \in \widetilde{D}(L)$ was arbitrary, implies the dissipativity of $(L, \widetilde{D}(L))$. □

The next proposition provides regularity properties of the solution of the resolvent equation with respect to the y-variables.

Proposition 2.10. *Let $u \in D(L)$ be such that $u - Lu = f \in C_c^\infty$. Then for every multiindex α one has $D_y^\alpha u \in D(L)$ and*

$$D_y^\alpha u - L D_y^\alpha u = D_y^\alpha f.$$

Proof. Let $u \in D(L)$ be such that $u - Lu = f \in C_c^\infty$. Then

$$\int_{\mathbb{R}^{N+M}} (uv + \langle \nabla_x u, \nabla_x v \rangle + \phi(x) \langle \nabla_y u, \nabla_y v \rangle)\, dx\, dy = \int_{\mathbb{R}^{N+M}} f v\, dx\, dy \quad \text{for every } v \in D(\mathfrak{a}). \tag{7}$$

For $h \in \mathbb{R}^M$ let $D_h g(z) := (g(x, y+h) - g(x, y))$ and let us take, in the last equation, $v = D_{-h} D_h u \in D(\mathfrak{a})$. Then, since $D_{-h} = D_h^*$, one has

$$\begin{aligned}\int_{\mathbb{R}^{N+M}} (|D_h u|^2 + |D_h \nabla_x u|^2 \phi(x) + |D_h \nabla_y u|^2)\, dx\, dy &= \int_{\mathbb{R}^{N+M}} D_h f\, D_h u\, dx\, dy \\ &\le \|D_h f\|_2 \|D_h u\|_2 \le \tfrac{1}{2}(\|D_h f\|_2^2 + \|D_h u\|_2^2). \quad (8)\end{aligned}$$

In particular for every $\omega \Subset \Omega$ there exists some positive constant $C = C(\omega)$ such that

$$\|D_h \nabla u\|_{L^2(\omega)} \le C|h| \|\nabla f\|_{L^2(\omega)}$$

for sufficiently small h; this proves that ∇u is weakly differentiable in ω in the y-variable and that $D_{y_i} u \in H^1_{\mathrm{loc}}(\Omega)$. Moreover, if $e_1, \dots, e_M$ is the standard basis of $\mathbb{R}^M$, $t \ne 0$, and $h = te_i$, then dividing both sides of (8) by t and taking the limit for $t \to 0$ we obtain

$$\frac{1}{2} \int_{\mathbb{R}^{N+M}} (|D_{y_i} u|^2 + |D_{y_i} \nabla_x u|^2 \phi(x) + |D_{y_i} \nabla_y u|^2)\, dx\, dy \le \int_{\mathbb{R}^{N+M}} |D_{y_i} f|^2\, dx\, dy,$$

which proves that $D_{y_i} u \in D(\mathfrak{a})$. Let us fix now $v \in C_c^\infty$; using (7) with v replaced by $D_{-te_i} v$ we get

$$\begin{aligned}\int_{\mathbb{R}^{N+M}} D_{te_i} f\, v\, dx\, dy &= \int_{\mathbb{R}^{N+M}} f\, D_{-te_i} v\, dx\, dy \\ &= \int_{\mathbb{R}^{N+M}} \big(u D_{-te_i} v + \langle \nabla_x u, \nabla_x D_{-te_i} v\rangle + \phi(x) \langle \nabla_y u, \nabla_y D_{-te_i} v\rangle\big)\, dx\, dy \\ &= \int_{\mathbb{R}^{N+M}} \big(D_{te_i} u v + \langle D_{te_i} \nabla_x u, \nabla_x v\rangle + \phi(x) \langle D_{te_i} \nabla_y u, \nabla_y v\rangle\big)\, dx\, dy.\end{aligned}$$

Dividing both sides of the last equation by t and taking the limit for $t \to 0$ we obtain

$$\int_{\mathbb{R}^{N+M}} D_{y_i} f\, v\, dx\, dy = \int_{\mathbb{R}^{N+M}} \big(D_{y_i} u v + \langle D_{y_i} \nabla_x u, \nabla_x v\rangle + \phi(x) \langle D_{y_i} \nabla_y u, \nabla_y v\rangle\big)\, dx\, dy.$$

Since by Proposition 2.5 C_c^∞ is a core for $\mathfrak{a}$ and since v is arbitrary in the last equation, we have that $D_{y_i} u \in D(L)$ and $D_{y_i} u - L(D_{y_i} u) = D_{y_i} f$, which is the required claim for $|\alpha| = 1$. An inductive argument easily proves the claim for any multiindex α. Moreover, since $D_y^\alpha u = (I - L)^{-1} D_y^\alpha f$, for some $C = C(\alpha) > 0$ we have

$$\|D_y^\alpha u\|_2 \le C \|D_y^\alpha f\|_2. \qquad \square$$

We end this section by proving a version of Kato's inequality adapted to L, which will be used for proving L^p-estimates.

Proposition 2.11. *Let $u \in D(L)$ and let us define*

$$\operatorname{sign}(u) = \begin{cases} 0 & \text{if } u(x) = 0, \\ u(x)/|u(x)| & \text{if } u(x) \ne 0. \end{cases}$$

Then $|u|$ satisfies the distributional inequality

$$-\mathfrak{a}(|u|, \varphi) \ge \int_{\mathbb{R}^{N+M}} \operatorname{sign}(u)\, Lu\, \phi\, dx\, dy \quad \text{for any } 0 \le \varphi \in C_c^\infty.$$

Proof. We suppose first that $u \in C_c^\infty$. If

$$u_\epsilon(x) = \sqrt{|u|^2 + \epsilon^2}$$

then $u_\epsilon \geq |u|$ and

$$u_\epsilon(a\,\nabla u_\epsilon) = u(a\,\nabla u) \tag{9}$$

(here a is the matrix defined in (1)). Thus (9) implies

$$\begin{aligned} |\nabla_x u_\epsilon| &\leq |u|\,|u_\epsilon|^{-1}|\nabla_x u| \leq |\nabla_x u|, \\ \phi(x)|\nabla_y u_\epsilon| &\leq |u|\,|u_\epsilon|^{-1}\phi(x)|\nabla_y u| \leq \phi(x)|\nabla_y u|. \end{aligned} \tag{10}$$

Taking the divergence of (9) we obtain

$$u_\epsilon L u_\epsilon + |\nabla_x u_\epsilon|^2 + \phi(x)|\nabla_y u_\epsilon|^2 = uLu + |\nabla_x u|^2 + \phi(x)|\nabla_y u|^2,$$

so by (10)

$$Lu_\epsilon \geq \frac{u}{u_\epsilon} Lu. \tag{11}$$

Integrating by parts the right-hand side of (11), it follows that

$$-\mathfrak{a}(u_\epsilon, \varphi) \geq \int_{\mathbb{R}^{N+M}} \frac{u}{u_\epsilon}\, Lu\, \varphi\, dx\, dy \quad \text{for any } 0 \leq \varphi \in C_c^\infty.$$

Letting $\epsilon \to 0$ we get

$$-\mathfrak{a}(|u|, \varphi) \geq \int_{\mathbb{R}^{N+M}} \operatorname{sign}(u)\, Lu\, \varphi\, dx\, dy \quad \text{for any } 0 \leq \varphi \in C_c^\infty.$$

Let now $u \in D(L)$ and let $u_n \in C_c^\infty$ be such that $u_n \to u$ in $D(L)$. Up to a subsequence, if necessary, we can also suppose that $u_n \to u$ almost everywhere. Since also $u_n \to u$ in $D(\mathfrak{a})$ by the last inequalities

$$-\mathfrak{a}(|u_n|, \varphi) \geq \int_{\mathbb{R}^{N+M}} \operatorname{sign}(u_n)\, Lu_n\, \varphi\, dx\, dy \quad \text{for any } 0 \leq \varphi \in C_c^\infty;$$

the claim follows letting $n \to \infty$. □

3. The distance d associated with L

Let $\alpha > 0$ and let

$$L = \Delta_x + |x|^\alpha \Delta_y$$

be the self-adjoint operator defined in Section 2 with $\phi(x) = |x|^\alpha$. In this section we introduce a natural metric d on $\mathbb{R}^{N+M}$ associated with L and which makes the triple $(\mathbb{R}^{N+M}, d, \mathcal{L})$, consisting of $\mathbb{R}^{N+M}$ equipped with the distance d and the Lebesgue measure $\mathcal{L}$, a homogeneous space in the sense of [Coifman and Weiss 1971; 1977].

Definition 3.1. Let $\gamma : [0, T] \to \mathbb{R}^{N+M}$ be an absolutely continuous curve. We say that γ is a subunit curve if for a.e. $t \in [0, T]$ one has

$$\langle \dot\gamma(t), \xi \rangle^2 \leq |\xi_x|^2 + |x|^\alpha |\xi_y|^2 \quad \text{for every } \xi = (\xi_x, \xi_y) \in \mathbb{R}^{N+M}.$$

For every $z_1, z_2 \in \mathbb{R}^{N+M}$ we define

$$\begin{aligned} d(z_1, z_2) &= \inf\{T \in \mathbb{R}^+ : \text{ there exists a subunit curve } \gamma : [0, T] \to \mathbb{R}^{N+M},\ \gamma(0) = z_1,\ \gamma(T) = z_2\} \\ &= \sup\{\psi(z_2) - \psi(z_1) : \psi \in W^{1,\infty}(\mathbb{R}^{N+M}),\ |\nabla_x \psi|^2 + |x|^\alpha |\nabla_y \psi|^2 \le 1\}. \end{aligned} \tag{12}$$

We remark that d is a well-defined distance and that any pair of points $z_1, z_2 \in \mathbb{R}^{N+M}$ can be joined by a subunit curve; see [Franchi and Serapioni 1987, Section 2, Example 3.6] and [Franchi and Lanconelli 1984, Definition 2.4]. A proof of the equality in (12) can be found in [Jerison and Sánchez-Calle 1987, Proposition 3.1].

For $z_0 \in \mathbb{R}^{N+M}$, $r > 0$, we write

$$S(z_0, r) := \{z \in \mathbb{R}^{N+M} : d(z_0, z) < r\}$$

to denote the balls of $\mathbb{R}^{N+M}$ with respect to the metric d. In the next proposition we clarify the structure of the metric and define an equivalent system of balls which are explicit and easier to work with. For $z_0 = (x_0, y_0) \in \mathbb{R}^{N+M}$, $r > 0$, let us define the cylindrical set

$$Q(z_0, r) := B(x_0, r) \times B(y_0, r(x_0)), \quad r(x_0) := r(r + |x_0|)^{\frac{\alpha}{2}}. \tag{13}$$

Proposition 3.2. *There exist two positive constants $C_1, C_2 > 0$ such that the distance function d satisfies for every $z_1 = (x_1, y_1)$, $z_1 = (x_2, y_2) \in \mathbb{R}^{N+M}$*

$$C_1 F(z_1, z_2) \le d(z_1, z_2) \le C_2 F(z_1, z_2),$$

where

$$F(z_1, z_2) = |x_1 - x_2| + \left(\frac{|y_1 - y_2|}{(|x_1| + |x_2|)^{\frac{\alpha}{2}}} \wedge |y_1 - y_2|^{\frac{2}{2+\alpha}} \right).$$

In particular

$$|S(z_0, r)| \simeq \begin{cases} r^{N+M(1+\frac{\alpha}{2})} & \text{if } r \ge |x_0|, \\ r^{N+M} |x_0|^{M\frac{\alpha}{2}} & \text{if } r \le |x_0|, \end{cases}$$

and the metric balls satisfy the doubling property

$$|S(z_0, sr)| \le C s^{N+M(1+\frac{\alpha}{2})} |S(z_0, r)| \quad \text{for every } z_0 \in \mathbb{R}^{N+M},\ s \ge 1.$$

Furthermore there exists a constant $c > 1$ such that for every $z_0 = (x_0, y_0) \in \mathbb{R}^{N+M}$, $r > 0$,

$$Q(z_0, c^{-1} r) \subseteq S(z_0, r) \subseteq Q(z_0, cr).$$

In particular $|S(z_0, r)| \simeq r^{N+M}(r + |x_0|)^{M\alpha/2}$ and $(\mathbb{R}^{N+M}, d, \mathcal{L})$ is a metric space of homogeneous type.

Proof. The first part of the statement is proved in [Robinson and Sikora 2008, Proposition 5.1, Corollary 5.2] $\big($take in that paper $\delta_1 = \delta_1' = 0,\ \delta_2 = \delta_2' = \frac{\alpha}{2},\ D = D' = N + M\big(1 + \frac{\alpha}{2}\big)\big)$. A proof of the second part can be found in [Franchi and Serapioni 1987, Proposition 2.7, Example 3.6] and [Franchi et al. 1994, Proposition 1]. □

4. L^p estimates

Let $1 < p < \infty$. In this section we assume that $\phi(x) = |x|^\alpha$, with $\alpha > 0$, and consider therefore the operator

$$L = \Delta_x + |x|^\alpha \Delta_y$$

in L^p with $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$. Property (iii) of Proposition 2.4 shows that the symmetric semigroup $(e^{tL})_{t\geq 0}$ generated by L in L^2 is submarkovian. Then by a standard result, see for example [Ouhabaz 2005, Chapter 3], it induces a consistent family of strongly continuous semigroups on L^p for any $1 < p < \infty$, still denoted by $(e^{tL})_{t\geq 0}$. Moreover $(e^{tL})_{t\geq 0}$ extends to a contractive holomorphic semigroup on a sector; see [Ouhabaz 2005, Theorem 3.13].

Definition 4.1. For any $p \in (1, \infty)$ we define the sectorial operator $(L, D_p(L))$ as the generator of the extrapolated semigroup $(e^{tL})_{t\geq 0}$ in L^p. We also write $D_2(L) = D(L)$.

$D_p(L) \cap D(L)$ is dense in L^p; in fact, if $f \in C_c^\infty$, then $e^{tL} f \in D_p(L) \cap D(L)$ and converges to f in L^p. Then $D_p(L) \cap D(L)$, being a dense invariant set, is by construction a core for $(L, D_p(L))$.

Theorem 2.7 holds in the specific situation since $|x|^\alpha \in B_\infty(\mathbb{R}^N)$ and we prove that those estimates extend to $1 < p < \infty$.

We recall that $|x|^\beta$ belongs to $A_t(\mathbb{R}^N)$, the class of Muckenhoupt weights of order $t \geq 1$, whenever $0 \leq \beta < N(t-1)$. This means that

$$\left(\frac{1}{|B|}\int_B |x|^\beta\,dx\right)\left(\frac{1}{|B|}\int_B |x|^{\beta(1-t')}\,dx\right)^{t-1} \leq C$$

for any ball (or cube) B of $\mathbb{R}^N$; see for example [Duoandikoetxea 2001, Chapther 7.3]. However, we need Muckenhoupt weights in $(\mathbb{R}^{N+M}, d, \mathcal{L})$ with respect to the metric defined in Section 3. Since $|x|^\beta$ is independent of y and since the balls S in this space are equivalent to the cylinders $Q(z, r)$ defined in (13), which are products of balls in R^N and $\mathbb{R}^M$ respectively, one easily verifies that

$$\left(\frac{1}{|S|}\int_S |x|^\beta\,dx\,dy\right)\left(\frac{1}{|S|}\int_S |x|^{\beta(1-t')}\,dx\,dy\right)^{t-1} \leq C$$

for every ball S (or cylinder) in $(\mathbb{R}^{N+M}, d, \mathcal{L})$.

A theory on these classes of weights in homogeneous spaces is presented for example in [Strömberg and Torchinsky 1989, Chapter I], to which we refer for the proofs of the results needed in what follows. In particular, we recall that Muckenhoupt weights induce doubling measures. The following well-known consequence of the definition is crucial in our approach.

Lemma 4.2. *If $\phi(x, y) = |x|^\beta$, $t \geq 1$, and $\beta < N(t-1)$, there exists $c > 0$ such that the inequality*

$$\left(\frac{1}{|Q|}\int_Q g\right)^t \leq \frac{c}{\phi(Q)}\int_Q g^t \phi \tag{14}$$

holds for all nonnegative functions g and all cylinders Q in $(\mathbb{R}^{N+M}, d, \mathcal{L})$. Here

$$\phi(Q) = \int_Q \phi.$$

Proof. By Hölder's inequality one has

$$\left(\frac{1}{|Q|}\int_Q g\right)^t = \left(\frac{1}{|Q|}\int_Q g\phi^{\frac{1}{t}}\phi^{-\frac{1}{t}}\right)^t \le \left(\frac{1}{|Q|}\int_Q g^t\phi\right)\left(\frac{1}{|Q|}\int_Q \phi^{1-t'}\right)^{t-1}$$

and the claim follows from the A_t property of $|x|^\beta$ in $(\mathbb{R}^{N+M}, d, \mathcal{L})$. □

The A_t property of $\phi = |x|^\alpha$, combined with mean value inequalities for Baouendi–Grushin operators, allows us to characterize the domain of the operator. We prove the following result.

Theorem 4.3. *For every $1 \le i, j \le N$, $1 \le h, k \le M$, the operators $|x|^\alpha D_{y_h y_k}(I-L)^{-1}$, $D_{x_i x_j}(I-L)^{-1}$, originally defined in L^2, extend to bounded operators in L^p.*

The main tool is the following result due to Shen [2005, Theorem 3.1], which can be considered as a version of the Calderón–Zygmund theorem in the absence of kernels. The original proof, where Euclidean balls are used, can be modified to work also for our space $(\mathbb{R}^{N+M}, d, \mathcal{L})$. Indeed an improved version of Shen's result in more general homogeneous spaces, which covers the cases of our interest, can be found in [Auscher and Martell 2007, Theorem 3.14 and Section V].

Theorem 4.4. *Let $1 \le p_0 < q_0 \le \infty$. Suppose that T is a sublinear bounded operator on L^{p_0}. Suppose moreover that there exist $\alpha_2 > \alpha_1 > 1$, $C > 0$, such that*

$$\left(\frac{1}{|Q|}\int_Q |Tf|^{q_0}\right)^{\frac{1}{q_0}} \le C\left(\frac{1}{|\alpha_1 Q|}\int_{\alpha_1 Q} |Tf|^{p_0}\right)^{\frac{1}{p_0}}$$

for all cylinders Q and for all $f \in C_c^\infty$, with support in $\mathbb{R}^{N+M} \setminus \alpha_2 Q$. Then, for $p_0 \le p < q_0$, there exists a positive constant C_p such that for all $f \in C_c^\infty$

$$\|Tf\|_p \le C_p \|f\|_p.$$

We briefly describe our strategy of proof of Theorem 4.3. We first prove the a priori estimates for $p \ge 2$ by applying the above theorem to the operator $T = |x|^\alpha D_{y_h y_k}(I-L)^{-1}$, with $p_0 = 2$, arbitrary $q_0 > 2$ and $\alpha_1 = 3$, $\alpha_2 = 4$. Therefore we have to prove that, if Q is a cylinder and $f \in C_c^\infty$ has support in $\mathbb{R}^{N+M} \setminus 4Q$, then $u = (I-L)^{-1} f$ satisfies

$$\left(\frac{1}{|Q|}\int_Q ||x|^\alpha D_{y_h y_k} u|^{q_0}\right)^{\frac{1}{q_0}} \le C\left(\frac{1}{|3Q|}\int_{3Q} ||x|^\alpha D_{y_h y_k} u|^2\right)^{\frac{1}{2}}$$

for some positive C independent of f. Observe that u satisfies in $4Q$ the equation

$$u - Lu = u - \Delta_x u - |x|^\alpha \Delta_y u = 0.$$

Moreover, by Proposition 2.10, the operator L commutes with the second-order derivatives with respect to y and $v = D_{y_h y_k} u$ satisfies the same equation in $4Q$.

To get the a priori estimates in the case $1 < p \le 2$, we apply Shen's theorem to the adjoint operator T^*.

As a first step we recall a mean value inequality for subsolutions of L, that is, for functions v satisfying the inequality $Lv \ge 0$ in Q, in a weak sense. This means that $\mathfrak{a}(u, \varphi) \le 0$ for any $0 \le \varphi \in C_c^\infty(Q)$.

Lemma 4.5 (see [Franchi and Serapioni 1987, Theorem 5.7]). *There exists a positive constant C such that, if v is a local subsolution of L in $4Q$, then*

$$\sup_Q |v| \le C\left(\frac{1}{|3Q|}\int_{3Q} v^2\right)^{\frac{1}{2}}.$$

The previous mean value inequality remains true also for $0 < r < \infty$. It follows from the self-improvement of the right-hand side in weak reverse Hölder estimates; see for example [Chanillo and Wheeden 1986, Theorem 4.1]. We give a simple self-contained proof which is a simplification of [Bernicot et al. 2016, Theorem B.1] for the sup-norm. Note that the case $r > 2$ follows from Hölder's inequality.

Lemma 4.6. *For every $0 < r < \infty$, there exists a positive constant C_r such that, if v is local subsolution of L in $4Q$, then*

$$\sup_Q |v| \le C_r\left(\frac{1}{|3Q|}\int_{3Q} |v|^r\right)^{\frac{1}{r}}.$$

Proof. Let $r < 2$ and $\mathcal{Q}$ be the collection of all cylinders Q' contained in Q. For $\epsilon \in (0, 1)$ let

$$C_r(\epsilon) := \sup_{Q'\in\mathcal{Q}} \frac{\sup_{Q'} |v|}{\left(\frac{1}{|3Q'|}\int_{3Q'} |v|^r\right)^{\frac{1}{r}} + \epsilon} \le \epsilon^{-1} \sup_Q |v|.$$

Let us fix $Q' \in \mathcal{Q}$ and let Q'' be a cylinder centered at some point of Q' and such that $9Q'' \subseteq 3Q'$. We assume that the radii $r(Q'')$, $r(Q')$ of Q'' and Q' satisfy $c^{-1}r(Q') \le r(Q'') \le cr(Q')$ for some fixed constant $c \in (0, 1)$. Applying Lemma 4.5 in Q'' we get

$$\begin{aligned}\sup_{Q''} |v| &\le C_2\left(\frac{1}{|3Q''|}\int_{3Q''} v^2\right)^{\frac{1}{2}} \le C_2\big(\sup_{3Q''} |v|\big)^{1-\frac{r}{2}}\left(\frac{1}{|3Q''|}\int_{3Q''} |v|^r\right)^{\frac{1}{2}}\\ &\le C_2C_r(\epsilon)^{1-\frac{r}{2}}\left[\left(\frac{1}{|9Q''|}\int_{9Q''} |v|^r\right)^{\frac{1}{r}} + \epsilon\right]^{1-\frac{r}{2}}\left(\frac{1}{|3Q''|}\int_{3Q''} |v|^r\right)^{\frac{1}{2}}\\ &\le C'C_2C_r(\epsilon)^{1-\frac{r}{2}}\left[\left(\frac{1}{|9Q''|}\int_{9Q''} |v|^r\right)^{\frac{1}{r}} + \epsilon\right] \le C''C_2C_r(\epsilon)^{1-\frac{r}{2}}\left[\left(\frac{1}{|3Q'|}\int_{3Q'} |v|^r\right)^{\frac{1}{r}} + \epsilon\right].\end{aligned}$$

Since Q'' is arbitrary

$$\sup_{Q'} |v| \le C''C_2C_r(\epsilon)^{1-\frac{r}{2}}\left[\left(\frac{1}{|3Q'|}\int_{3Q'} |v|^r\right)^{\frac{1}{r}} + \epsilon\right].$$

Taking the supremum over $Q' \in \mathcal{Q}$ we get $C_r(\epsilon) \le (C''C_2)^{2/r}$ and the thesis follows letting $\epsilon \to 0$. □

Now we prove that Lemma 4.6 holds if we replace the Lebesgue measure with $|x|^\beta\, dx,\ \beta > 0$.

Lemma 4.7. *Fix $0 < s < \infty$ and v as in Lemma 4.5. Then*

$$\sup_Q |v| \le \left(\frac{C}{\phi(3Q)}\int_{3Q}\phi|v|^s\right)^{\frac{1}{s}},$$

where C depends only on s, p and the B_p constant of $\phi(x) = |x|^\beta$ and

$$\phi(3Q) = \int_{3Q}\phi.$$

Proof. Let $0 < s < \infty$ and Q be a cylinder of $\mathbb{R}^{N+M}$. We fix t as in Lemma 4.2. By using Lemma 4.6 with $r = \frac{s}{t}$ and (14) we obtain

$$\sup_Q |v| \le C\left(\frac{1}{|3Q|}\int_{3Q}|v|^{\frac{s}{t}}\right)^{\frac{t}{s}} \le C\left(\frac{1}{\phi(3Q)}\int_{3Q}\phi|v|^s\right)^{\frac{1}{s}}. \qquad \square$$

By combining the estimate in Lemma 4.7 and the B_p property we deduce the following.

Corollary 4.8. *Let $0 < s < \infty$, $1 < p < \infty$ and v as in Lemma 4.5. Then*

$$\left(\frac{1}{|Q|}\int_Q(|x|^\beta|v|^s)^p\right)^{\frac{1}{p}} \le \frac{C}{|3Q|}\int_{3Q}|x|^\beta|v|^s,$$

where C depends only on s, p and the B_p constant of $|x|^\beta$.

Proof. Using the B_p property of $\phi = |x|^\beta$ and Lemma 4.7 we obtain

$$\left(\frac{1}{|Q|}\int_Q(\phi|v|^s)^p\right)^{\frac{1}{p}} \le \left(\frac{1}{|Q|}\int_Q\phi^p\right)^{\frac{1}{p}}\sup_Q|v|^s \le C\left(\frac{1}{|Q|}\int_Q\phi\right)\sup_Q|v|^s \le \frac{C}{|3Q|}\int_{3Q}\phi|v|^s. \qquad \square$$

We can now prove our main result.

Proof of Theorem 4.3. We first consider the operators $|x|^\alpha D_{y_hy_k}(I-L)^{-1}$.

Let us preliminarily treat the case $p \ge 2$. Let us fix $q_0 > 2$ and let Q be a cylinder in $\mathbb{R}^{N+M}$ and $f \in C_c^\infty$ a smooth function with support in $\mathbb{R}^{N+M}\setminus 4Q$. We set

$$T = |x|^\alpha D_{y_hy_k}(I-L)^{-1}, \quad u = (I-L)^{-1}f, \quad v = D_{y_hy_k}(I-L)^{-1}f.$$

By Theorem 2.7, T is bounded on L^2. Since $f = 0$ in $4Q$, by Proposition 2.10 $v - Lv = 0$ in $4Q$. Combining the last equality with Kato's inequality of Proposition 2.11, we get

$$-\mathfrak{a}(|v|,\varphi) \ge \int \operatorname{sign} v\, Lv\,\varphi = \int |v|\,\varphi \ge 0 \quad \text{for all } 0 \le \varphi \in C_c^\infty(4Q).$$

It follows that $|v|$ is a local subsolution of L. Note that v is a local solution of $v - Lv = 0$ but we cannot assert that it is a local subsolution of L, that is, $Lv \ge 0$, since its sign is not given. By Corollary 4.8 with $s = 2$ and $\beta = 2\alpha$ we have

$$\left(\frac{1}{|Q|}\int_Q(|x|^{2\alpha}|v|^2)^q\right)^{\frac{1}{q}} \le \frac{C}{|3Q|}\int_{3Q}|x|^{2\alpha}|v|^2, \quad 1 < q < \infty,$$

or, equivalently,

$$\left(\frac{1}{|Q|}\int_Q (|x|^\alpha|v|)^{2q}\right)^{\frac{1}{2q}} \le \left(\frac{C}{|3Q|}\int_{3Q} (|x|^\alpha|v|)^2\right)^{\frac{1}{2}}, \quad 1<q<\infty.$$

It follows that for $2 \le q_0 < \infty$

$$\begin{aligned}\left(\frac{1}{|Q|}\int_Q |Tf|^{q_0}\right)^{\frac{1}{q_0}} &= \left(\frac{1}{|Q|}\int_Q (|x|^\alpha|D_{y_hy_k}u|)^{q_0}\right)^{\frac{1}{q_0}} \\ &\le \left(\frac{C}{|3Q|}\int_{3Q} (|x|^\alpha|D_{y_hy_k}u|)^2\right)^{\frac{1}{2}} = \left(\frac{C}{|3Q|}\int_{3Q} |Tf|^2\right)^{\frac{1}{2}}.\end{aligned}$$

By Theorem 4.4, T extends to a bounded operator in L^p for every $2 \le p < q_0$. Since we can choose q_0 arbitrarily, the case $2 \le p < \infty$ follows.

To treat the case $p < 2$ we consider the adjoint operator

$$T^* = D_{y_hy_k}(I-L)^{-1}|x|^\alpha,$$

which is bounded in L^2. By duality, the boundedness of T in L^p for every $1 < p \le 2$ is equivalent to that of T^* in L^p for every $p \ge 2$. As before we fix $q_0 > 2$ and we prove that T^* satisfies Shen's assumption for every $2 \le q_0 < \infty$. Let Q be a cylinder in $\mathbb{R}^{N+M}$ and $f \in C_c^\infty$ with support in $\mathbb{R}^{N+M} \setminus 4Q$; set

$$u = (I-L)^{-1}(|x|^\alpha f), \quad v = D_{y_hy_k}u.$$

Then v satisfies $v - Lv = 0$ in $4Q$. By arguing as above, $|v|$ is a local subsolution of L; hence Lemma 4.5 yields a positive constant C such that

$$\sup_Q |v| \le C\left(\frac{1}{|3Q|}\int_{3Q} v^2\right)^{\frac{1}{2}}.$$

It follows that

$$\begin{aligned}\left(\frac{1}{|Q|}\int_Q |T^*f|^{q_0}\right)^{\frac{1}{q_0}} &= \left(\frac{1}{|Q|}\int_Q |D_{y_hy_k}u|^{q_0}\right)^{\frac{1}{q_0}} \\ &\le \sup_Q |v| \le C\left(\frac{1}{|3Q|}\int_{3Q} v^2\right)^{\frac{1}{2}} = \left(\frac{C}{|3Q|}\int_{3Q} |T^*f|^2\right)^{\frac{1}{2}}\end{aligned}$$

and the proof is complete by Theorem 4.4 applied to T^*.

By difference the operator $\Delta_x(I-L)^{-1}$ is bounded on L^p and, integrating with respect to y the classical Calderón–Zygmund estimates in the x-variables we deduce the L^p boundedness of $D_{x_ix_j}(I-L)^{-1}$. □

Remark 4.9. Theorem 4.3 holds for every $0 \le \phi \in B_2(\mathbb{R}^N)$ when $1 < p \le 2$. In fact, the properties of the powers $|x|^\alpha$ stated in Lemma 4.7 and Corollary 4.8 have been used in the above proof only when $p > 2$.

We can now give a partial description of the domain $D_p(L)$, which will be strengthened in Theorem 4.21 after proving L^p-estimates for mixed derivatives.

Proposition 4.10. *Let $p \in (1, \infty)$. Then one has*

$$D_p(L) = \{u \in L^p : \nabla_x u,\ D_{xx}u,\ |x|^{\frac{\alpha}{2}}\nabla_y u,\ |x|^\alpha D_{yy}u \in L^p\}. \tag{15}$$

Moreover

$$\|D_{x_i x_j}u\|_p + \||x|^\alpha D_{y_h y_k}u\|_p \le C\|Lu\|_p, \quad u \in D_p(L). \tag{16}$$

Proof. Let $1 < p < \infty$ and let $\widetilde{D}_p(L)$ be the set defined in the right-hand side of equality (15).

Let us preliminarily prove that $D_p(L) \subseteq \widetilde{D}_p(L)$.

Theorem 4.3 and the consistency of the resolvent operators in L^2 and in L^p imply that

$$\|D_{x_i x_j}u\|_p + \||x|^\alpha D_{y_h y_k}u\|_p \le C\|(I-L)u\|_p \tag{17}$$

for any $u \in (I-L)^{-1}(C_c^\infty)$ which is dense in $D_p(L)$ with respect to the graph norm. This implies that (17) extends to $D_p(L)$, proving that u has pure second-order distributional derivatives which satisfy $D_{x_i x_j}$, $|x|^\alpha D_{y_h y_k} \in L^p$ and that

$$\|D_{x_i x_j}u\|_p + \||x|^\alpha D_{y_h y_k}u\|_p \le C\|(I-L)u\|_p \le C(\|u\|_p + \|Lu\|_p), \quad u \in D_p(L). \tag{18}$$

As in Theorem 2.7, an interpolation argument shows that $\nabla_x u$, $|x|^{\frac{\alpha}{2}}\nabla_y u \in L^p(\mathbb{R}^{N+M})$; i.e., $u \in \widetilde{D}_p(L)$.

To get homogeneous estimates, we use Proposition 2.4(v), and apply (18) to $u(x, y) = v(sx, s^{(2+\alpha)/2}y)$, $s > 0$, thus obtaining

$$\|D_{x_i x_j}u\|_p + \||x|^\alpha D_{y_h y_k}u\|_p \le C(\|Lv\|_p + s^{-2}\|v\|_p).$$

Letting $s \to \infty$ we obtain (16).

To prove that $\widetilde{D}_p(L) = D(L)$, we proceed as in the proof of Theorem 2.9 and show that the operator $(L, \widetilde{D}_p(L))$ is dissipative. Let $u \in \widetilde{D}_p(L)$; then the same sectional argument of (6) shows that for a.e. $x \in \Omega_N$, we have $u(x, \cdot) \in W^{2,p}(\mathbb{R}^M)$ and from [Metafune and Spina 2008]

$$\int_{\mathbb{R}^M} u|u|^{p-2}\Delta_y u\, dy = -(p-1)\int_{\mathbb{R}^M} |\nabla_y u|^2|u|^{p-2}\, dy \quad \text{for a.e. } x \in \Omega_N.$$

Then multiplying by $|x|^\alpha$, integrating in x and using Fubini's theorem we get

$$\int_{\mathbb{R}^{N+M}} |x|^\alpha u|u|^{p-2}\Delta_y u\, dx\, dy = -(p-1)\int_{\mathbb{R}^{N+M}} |x|^\alpha|\nabla_y u|^2|u|^{p-2}\, dx\, dy.$$

Analogously

$$\int_{\mathbb{R}^{N+M}} u|u|^{p-2}\Delta_x u\, dx\, dy = -(p-1)\int_{\mathbb{R}^{N+M}} |\nabla_x u|^2|u|^{p-2}\, dx\, dy.$$

The last two inequalities imply

$$\int_{\mathbb{R}^{N+M}} u|u|^{p-2}\, Lu\, dx\, dy = -(p-1)\int_{\mathbb{R}^{N+M}} (|\nabla_x u|^2 + |x|^\alpha|\nabla_y u|^2)|u|^{p-2}\, dx\, dy \le 0,$$

which, since $u \in \widetilde{D}_p(L)$ is arbitrary, implies the dissipativity of $(L, \widetilde{D}_p(L))$. □

Remark 4.11. The previous theorem implies, in particular, the equivalence between the graph norm $\|u\|_p + \|Lu\|_p$ and the norm

$$\|u\|_p + \|\nabla_x u\|_p + \||x|^{\frac{\alpha}{2}}\nabla_y u\|_p + \|D_{xx}u\|_p + \||x|^{\alpha}D_{yy}u\|_p,$$

since this last clearly dominates $\|Lu\|_p$.

The following proposition shows that C_c^∞ also is a core for $(L, D_p(L))$.

Proposition 4.12. *For any $p \in (1,\infty)$, C_c^∞ is a core for the operator $(L, D_p(L))$.*

Proof. Let $u \in D_p(L)$; we preliminarily approximate u with functions in $D_p(L)$ having compact support in $\mathbb{R}^{N+M}$. Let $\eta \in C_c^\infty(\mathbb{R}^N)$ be a smooth function such that $\chi_{B_1} \le \eta \le \chi_{B_2}$ and, for every $n \in \mathbb{N}$, $x \in \mathbb{R}^N$, define $\eta_n(x) = \eta\big(\frac{x}{n}\big)$. Set $u_n = \eta_n u$. Now u_n has, by construction, compact support in x and, using the characterization in (15), one can easily recognize that $u_n \in D_p(L)$. Lebesgue's theorem immediately implies that u_n, $|x|^{\alpha/2}\nabla_y u_n$, $|x|^\alpha D_{yy}u_n$ tend to u, $|x|^{\alpha/2}\nabla_y u$, $|x|^\alpha D_{yy}u$ in L^p, respectively. Concerning the x-gradient, we have

$$\begin{aligned}\|\nabla_x(\eta_n u) - \nabla_x(u)\|_p^p &\le \int_{\mathbb{R}^{N+M}} |\eta_n - 1|^p |\nabla_x u|^p \, dx\, dy + \int_{\mathbb{R}^{N+M}} |\nabla_x \eta_n|^p |u|^p \, dx\, dy \\ &\le \int_{\mathbb{R}^{N+M}} |\eta_n - 1|^p |\nabla u|^p \, dx\, dy + Cn^{-p} \int_{\{n \le |x| \le 2n\}} |u|^p \, dx\, dy,\end{aligned}$$

which tends to 0 by dominated convergence. Similarly $D_{xx}u_n$ tends to $D_{xx}u$ in L^p. By Proposition 4.10 and Remark 4.11, this proves that u_n tends to u in $D_p(L)$. Using a similar argument with η replaced by an analogous cut-off function $\eta \in C_c^\infty(\mathbb{R}^M)$, we can approximate u with functions in $D_p(L)$ having compact support also in the y-variable.

Let us suppose first that $p < 2$ and let $u \in D_p(L)$ (the case $p = 2$ is already proved in Proposition 2.5). From the first part of the proof, we can suppose u has compact support. Let $\eta \in C_c^\infty$ such that $\eta = 1$ on the support of u and, using Proposition 2.5, let $(u_n)_{n\in\mathbb{N}}$ be a sequence of C_c^∞ functions such that $u_n \to u$ in $D_2(L)$ as $n \to \infty$. This implies

$$\|L(\eta u_n - u)\|_p + \|\eta u_n - u\|_p \le C[\|L(\eta u_n - u)\|_2 + \|\eta u_n - u\|_2] \to 0 \quad \text{as } n \to \infty,$$

where C depends on the measure of the support of u. This proves the claim for $p < 2$.

The proof of the case $p > 2$ can be carried out by slightly adapting the arguments used in the proof of Proposition 2.5. We equivalently show that $(I - L)(C_c^\infty)$ is dense in L^p and to this aim let $v \in L^{p'}(\mathbb{R}^{N+M})$ such that

$$\int_{\mathbb{R}^{N+M}} (I - L)u\, \bar{v}\, dx\, dy = 0 \quad \text{for all } u \in C_c^\infty.$$

Since $1 < p' < 2$, the partial Fourier transform of $v(x,\cdot) \in L^{p'}(\mathbb{R}^M)$, with respect to the y-variable, exists as a function in $L^p(\mathbb{R}^M)$ for a.e. $x \in \mathbb{R}^N$. Therefore, taking in the last equality the Fourier transform with respect to the y-variable and applying the Fubini and Plancherel theorems, we get

$$\int_{\mathbb{R}^{N+M}} [\hat{u}(x,\xi) - \Delta_x \hat{u}(x,\xi) + |x|^\alpha |\xi|^2 \hat{u}(x,\xi)]\, \bar{\hat{v}}(x,\xi)\, dx\, d\xi = 0 \quad \text{for all } u \in C_c^\infty.$$

By [Semenov 1977, Theorem 1.1], since the potential is nonnegative and in L^p_{loc}, we know $C_c^\infty(\mathbb{R}^N)$ is a core for $\Delta_x - \phi(\,\cdot\,)|\xi|^2$ in L^p; proceeding then as in the proof of Proposition 2.5 we conclude that $\hat{v}(\,\cdot\,,\xi)=0$ for a.e. $\xi\in\mathbb{R}^M$ and the proof follows. □

Mixed derivatives. By using classical covering results and Rellich inequalities we obtain here L^p estimates for the mixed second-order derivatives. To simplify the notation we write x for any of the variables x_i, $i=1,\dots,N$, and y for any of the variables y_h, $h=1,\dots,M$.

Theorem 4.13. *For every $u\in D_p(L)$*

$$\||x|^{\frac{\alpha}{2}}D_{xy}u\|_p\le C\|Lu\|_p.$$

We need a Rellich-type inequality.

Lemma 4.14. *Let $p\neq N,\frac{N}{2}$. There exist a positive constant C such that for $u\in C_c^\infty$ satisfying $u(0,y)=0$, $\nabla_x u(0,y)=0$ for all $y\in\mathbb{R}^M$, we have*

$$\left\|\frac{u}{|x|^2}\right\|_p\le C\|Lu\|_p.$$

Proof. Let $u\in C_c^\infty$ such that $u(0,y)=0$, $\nabla_x u(0,y)=0$ so that $\|u/|x|^2\|_p<\infty$. Then by [Metafune et al. 2019, Theorem 4.2; 2015, Theorem 3.3]

$$\int_{\mathbb{R}^N}\left|\frac{u}{|x|^2}\right|^p dx\le C\int_{\mathbb{R}^N}|\Delta_x u|^p\,dx.$$

Integrating the previous inequality over $\mathbb{R}^M$ and using Theorem 4.3,

$$\left\|\frac{u}{|x|^2}\right\|_p\le C\|\Delta_x u\|_p\le C(\|Lu\|_p+\||x|^\alpha\Delta_y u\|_p)\le C\|Lu\|_p.$$ □

Remark 4.15. The above Rellich inequality uses Theorem 4.3 to replace Δ_x with the operator L. However, even its version in dimension 1 (that is, for D_{xx} rather than L) is not obvious and probably cannot be obtained by integration by parts; see, e.g., [Metafune et al. 2019], where it is shown that Rellich inequalities can be proved for the Laplacian in $L^p(\mathbb{R}^N)$ when $p<\frac{N}{2}$, a condition which is never verified in dimension 1.

We first prove mixed derivatives estimates assuming that both u and $\nabla_x u$ vanish for $x=0$. We need the following covering result.

Proposition 4.16 [Cupini and Fornaro 2004, Proposition 6.1]. *For every $0\le k<\frac{1}{2}$ there exists a natural number $\zeta=\zeta(N,k)$ with the following property. Given $\mathcal{F}=\{x+B(\rho(x))\}_{x\in\mathbb{R}^N}$, where $\rho:\mathbb{R}^N\to\mathbb{R}_+$ is a Lipschitz continuous function with Lipschitz constant k, there exist a countable subcovering $\{x_n+B(\rho(x_n))\}_{n\in\mathbb{N}}$ of $\mathbb{R}^N$ such that at most ζ among the double balls $\{x_n+B(2\rho(x_n))\}_{n\in\mathbb{N}}$ overlap.*

The following lemma is an immediate consequence of the classical Calderón–Zygmund inequalities for the Laplacian.

Lemma 4.17. *Let $1 < p < \infty$. Then for every $a \in \mathbb{R}$, $u \in C_c^\infty$, one has*

$$\|D_{xx}u\|_p + \|a^2 D_{yy}u\|_p + \|aD_{xy}u\|_p \le C\|\Delta_x u + a^2\Delta_y u\|_p.$$

Proof. It is sufficient to apply the Calderón–Zygmund inequalities to $v(x, y) = u(x, ay)$. □

Lemma 4.18. *Let $p \neq \frac{N}{2}$, N. One has*

$$\|D_{xx}u\|_p + \||x|^\alpha D_{yy}u\|_p + \||x|^{\frac{\alpha}{2}} D_{xy}u\|_p \le C\|Lu\|_p$$

for every $u \in C_c^\infty$ such that $u(0, y) = 0$, $\nabla_x u(0, y) = 0$ for all $y \in \mathbb{R}^M$.

Proof. We fix $x_0 \in \mathbb{R}^N$ and choose $\vartheta \in C_c^\infty(\mathbb{R}^N)$ such that $0 \le \vartheta \le 1$, $\vartheta(x) = 1$ for $x \in B(0, 1)$ and $\vartheta(x) = 0$ for $x \in \mathbb{R}^N \setminus B(0, 2)$. Moreover, we set $\vartheta_\rho(x) = \vartheta((x - x_0)/\rho)$, where $\rho = \frac{1}{4}|x_0|$. We apply Lemma 4.17 to the function $\vartheta_\rho u$ and obtain

$$\|D_{xx}(\vartheta_\rho u)u\|_p + \||x_0|^\alpha D_{yy}(\vartheta_\rho u)\|_p + \||x_0|^{\frac{\alpha}{2}} D_{xy}(\vartheta_\rho u)\|_p \le C\|\Delta_x(\vartheta_\rho u) + |x_0|^\alpha \Delta_y(\vartheta_\rho u)\|_p.$$

By the classical interpolation inequalities for the gradient we get, for every $\eta > 0$,

$$\begin{aligned}
&\|D_{xx}u\|_{L^p(B(x_0,\rho))} + \||x_0|^\alpha D_{yy}u\|_{L^p(B(x_0,\rho))} + \||x_0|^{\frac{\alpha}{2}} D_{xy}u\|_{L^p(B(x_0,\rho))}\\
&\le C\bigg(\|\Delta_x u + |x_0|^\alpha \Delta_y u\|_{L^p(B(x_0,2\rho))} + \frac{1}{\rho}\|\nabla_x u\|_{L^p(B(x_0,2\rho))} + \frac{1}{\rho^2}\|u\|_{L^p(B(x_0,2\rho))} + \frac{|x_0|^{\frac{\alpha}{2}}}{\rho}\|\nabla_y u\|_{L^p(B(x_0,2\rho))}\bigg)\\
&\le C\bigg(\|\Delta_x u + |x_0|^\alpha \Delta_y u\|_{L^p(B(x_0,2\rho))} + \eta\|\Delta_x u\|_{L^p(B(x_0,2\rho))} + \eta\||x_0|^\alpha \Delta_y u\|_{L^p(B(x_0,2\rho))} + \frac{1}{\eta\rho^2}\|u\|_{L^p(B(x_0,2\rho))}\bigg).
\end{aligned}$$

Since

$$\rho = \tfrac{1}{4}|x_0|, \quad \tfrac{1}{2}|x_0| \le |x| \le \tfrac{3}{2}|x_0|, \quad x \in B(x_0, 2\rho),$$

we get

$$\begin{aligned}
&\|D_{xx}u\|_{L^p(B(x_0,\rho))} + \||x|^\alpha D_{yy}u\|_{L^p(B(x_0,\rho))} + \||x|^{\frac{\alpha}{2}} D_{xy}u\|_{L^p(B(x_0,\rho))}\\
&\quad\le C\bigg(\|\Delta_x u + |x|^\alpha \Delta_y u\|_{L^p(B(x_0,2\rho))} + \eta\|\Delta_x u\|_{L^p(B(x_0,2\rho))}\\
&\qquad\qquad + \eta\||x|^\alpha \Delta_y u\|_{L^p(B(x_0,2\rho))} + \frac{1}{\eta}\bigg\|\frac{u}{|x|^2}\bigg\|_{L^p(B(x_0,2\rho))}\bigg). \qquad (19)
\end{aligned}$$

Let $\{B(x_n, \rho(x_n))\}$ be a countable covering of $\mathbb{R}^N$ as in Proposition 4.16 such that at most ζ among the double balls $\{B(x_n, 2\rho(x_n))\}$ overlap. Writing (19) with x_n instead of x_0 and summing over n it follows that

$$\begin{aligned}
&\|D_{xx}u\|_p + \||x|^\alpha D_{yy}u\|_p + \||x|^{\frac{\alpha}{2}} D_{xy}u\|_p\\
&\qquad\le C\bigg(\|\Delta_x u + |x|^\alpha \Delta_y u\|_p + \eta\|\Delta_x u\|_p + \eta\||x|^\alpha \Delta_y u\|_p + \frac{1}{\eta}\bigg\|\frac{u}{|x|^2}\bigg\|_p\bigg).
\end{aligned}$$

By choosing η small enough we get

$$\|D_{xx}u\|_p + \||x|^\alpha D_{yy}u\|_p + \||x|^{\frac{\alpha}{2}} D_{xy}u\|_p \le C\bigg(\|\Delta_x u + |x|^\alpha \Delta_y u\|_p + \bigg\|\frac{u}{|x|^2}\bigg\|_p\bigg).$$

and the claim follows from Lemma 4.14. □

Next we prove mixed derivatives estimates assuming that either u or $\nabla_x u$ vanishes for $x=0$.

Lemma 4.19. *If $p \neq \frac{2N}{2-\alpha}, \frac{N}{2}, N$, then*

$$\||x|^{\frac{\alpha}{2}} D_{xy} u\|_p \le C\|Lu\|_p$$

for every $u \in C_c^\infty$ such that $u(0,y)=0$ or $u_x(0,y)=0$ for all $y \in \mathbb{R}^M$.

Proof. Let $u \in C_c^\infty$ such that $u(0,y)=0$ and let $v(x,y)=\frac{1}{\lambda}u(\lambda x, y)$. Then $v(0,y)=0$ and $\nabla_x v(0,y) = \nabla_x u(0,y)$. This implies that $w=u-v$ satisfies $w(0,x)=w_x(0,x)=0$. Moreover

$$\||x|^{\frac{\alpha}{2}} D_{xy} v\|_p = \lambda^{-\frac{\alpha}{2}-\frac{N}{p}}\||x|^{\frac{\alpha}{2}} D_{xy} u\|_p$$

and, applying Theorem 4.3,

$$\begin{aligned}\|Lv\|_p &\le \|v_{xx}\|_p + \||x|^\alpha \Delta_y v\|_p\\ &= \lambda^{1-\frac{N}{p}}\|u_{xx}\|_p + \lambda^{-\alpha-1-\frac{N}{p}}\||x|^\alpha \Delta_y u\|_p \le C(\lambda)\|Lu\|_p.\end{aligned}$$

Applying Lemma 4.18 to w we then have

$$\begin{aligned}\||x|^{\frac{\alpha}{2}} D_{xy} u\|_p &\le \||x|^{\frac{\alpha}{2}} D_{xy} w\|_p + \||x|^{\frac{\alpha}{2}} D_{xy} v\|_p \le C(\|Lw\|_p + \||x|^{\frac{\alpha}{2}} D_{xy} v\|_p)\\ &\le C(\|Lu\|_p + \|Lv\|_p + \||x|^{\frac{\alpha}{2}} D_{xy} v\|_p) \le C(\lambda)\|Lu\|_p + C\||x|^{\frac{\alpha}{2}} D_{xy} v\|_p\\ &= C(\lambda)\|Lu\|_p + C\lambda^{-\frac{\alpha}{2}-\frac{N}{p}}\||x|^{\frac{\alpha}{2}} D_{xy} u\|_p.\end{aligned}$$

The claim then follows by choosing λ large enough such that $C\lambda^{-\alpha/2-N/p} \le \frac{1}{2}$.

Assume now $u_x(0,y)=0$ and let $v(x,y)=u(\lambda x,y)$. Then $u(0,y)=v(0,y)$ and $v_x(0,y) = \lambda u_x(0,y)=0$. Moreover

$$\||x|^{\frac{\alpha}{2}} D_{xy} v\|_p = \lambda^{1-\frac{\alpha}{2}-\frac{N}{p}}\||x|^{\frac{\alpha}{2}} D^2_{xy} u\|_p.$$

It follows that $w=u-v$ satisfies $w(0,x)=w_x(0,x)=0$. Hence an analogous argument yields

$$\||x|^{\frac{\alpha}{2}} D_{xy} u\|_p \le C(\lambda)\|Lu\|_p + C\lambda^{1-\frac{\alpha}{2}-\frac{N}{p}}\||x|^{\frac{\alpha}{2}} D_{xy} u\|_p.$$

Choosing λ large enough or small enough according to $1-\frac{\alpha}{2}-\frac{N}{p}>0$ or $1-\frac{\alpha}{2}-\frac{N}{p}<0$ we get the claim for $1-\frac{\alpha}{2}-\frac{N}{p}\neq 0$ or, equivalently, $p \neq \frac{2N}{2-\alpha}$. □

Proof of Theorem 4.13. Let us suppose, preliminarily, $p \neq \frac{2N}{2-\alpha}$, $p\neq \frac{N}{2}$, $p \neq N$ and let $u \in C_c^\infty$. We introduce the operators

$$Pu(x,y) = \frac{u(x,y)+u(-x,y)}{2}, \qquad Qu(x,y) = \frac{u(x,y)-u(-x,y)}{2}$$

Observe that

$$u = Pu + Qu, \qquad Qu(0,y)=0, \qquad \nabla_x(Pu)(0,y)=0, \qquad y \in \mathbb{R}^M,$$

P and Q commute with the second-order derivatives and $\|P(Lu)\|_p + \|Q(Lu)\|_p$ is equivalent to $\|Lu\|_p$. Moreover

$$L(Pu) = P(Lu), \qquad L(Qu) = Q(Lu).$$

We can therefore apply the results in Lemma 4.19 to Pu and Qu. For the mixed second-order derivatives we get

$$\||x|^{\frac{\alpha}{2}}D_{xy}u\|_p \le \||x|^{\frac{\alpha}{2}}P(D_{xy}u)\|_p + \||x|^{\frac{\alpha}{2}}Q(D_{xy}u)\|_p = \||x|^{\frac{\alpha}{2}}D_{xy}(Pu)\|_p + \||x|^{\frac{\alpha}{2}}Q_{xy}(Qu)\|_p$$
$$\le C(\|L(Pu)\|_p + \|L(Qu)\|_p) = C(\|P(Lu)\|_p + \|Q(Lu)\|_p) \le C\|Lu\|_p.$$

By density the proof extends to $u \in D_p(L)$.

Suppose now $p = \frac{2N}{2-\alpha}$. Observe that, by the previous part of the proof, the operator $|x|^{\alpha/2}D_{xy}(I-L)^{-1}$ is bounded in L^p for some $p_1 < \frac{2N}{2-\alpha} < p_2$, with $p_1,\ p_2 \neq \frac{N}{2}, N$. The Riesz–Thorin interpolation theorem then yields the boundedness of $|x|^{\alpha/2}D_{xy}(I-L)^{-1}$ also for $p = \frac{2N}{2-\alpha}$; the same scaling argument used in the proof of Proposition 4.10 then proves the required claim. We can argue similarly for $p = N$, $p = \frac{N}{2}$. □

As a corollary we improve gradient estimates near $x = 0$ showing that $|x|^{\alpha/2-1}\nabla_y u \in L^p$, $u \in D_p(L)$, when $\left(\frac{\alpha}{2}-1\right)p + N > 0$. This last condition is necessary for the above integrability, since otherwise the weight $|x|^{\alpha/2-1}$ is not locally p-summable near $x = 0$. We also recall that $|x|^{\alpha/2}\nabla_y u \in L^p$ by Proposition 4.10.

Corollary 4.20. *Let $\left(\frac{\alpha}{2}-1\right)p + N > 0$. Then for every $u \in D_p(L)$*

$$\||x|^{\frac{\alpha}{2}-1}\nabla_y u\|_p \le C\|Lu\|_p.$$

Proof. By density we may assume that $u \in C_c^\infty$. By the Hardy inequality, see for example [Metafune et al. 2015, Proposition 8.1],

$$\int_{\mathbb{R}^N} |x|^{(\frac{\alpha}{2}-1)p}|\nabla_y u|^p\,dx \le \left(\frac{p}{\left(\frac{\alpha}{2}-1\right)p+N}\right)^p \int_{\mathbb{R}^N} |x|^{\frac{\alpha}{2}p}|D_{xy}u|^p\,dx.$$

Integrating over $\mathbb{R}^M$ and using Theorem 4.13, the claim follows. □

We can now strengthen Proposition 4.10.

Theorem 4.21. *Let $p \in (1,\infty)$. Then one has*

$$D_p(L) = \{u \in L^p : \nabla_x u,\ D_{xx}u \in L^p,\ |x|^{\frac{\alpha}{2}}\nabla_y u,\ |x|^{\frac{\alpha}{2}}D_{xy}u,\ |x|^{\alpha}D_{yy}u \in L^p\}.$$

In particular the graph norm $\|u\|_p + \|Lu\|_p$ is equivalent to

$$\|u\|_p + \|\nabla_x u\|_p + \||x|^{\frac{\alpha}{2}}\nabla_y u\|_p + \|D_{xx}u\|_p + \||x|^{\alpha}D_{yy}u\|_p + \||x|^{\frac{\alpha}{2}}D_{xy}u\|_p.$$

Moreover, if $\left(\frac{\alpha}{2}-1\right)p + N > 0$, then also $|x|^{\frac{\alpha}{2}-1}\nabla_y u \in L^p$.

Proof. The proof follows from Proposition 4.10, Theorem 4.13 and Corollary 4.20. □

References

[Auscher and Ben Ali 2007] P. Auscher and B. Ben Ali, "Maximal inequalities and Riesz transform estimates on L^p spaces for Schrödinger operators with nonnegative potentials", *Ann. Inst. Fourier* (*Grenoble*) **57**:6 (2007), 1975–2013. MR Zbl

[Auscher and Martell 2007] P. Auscher and J. M. Martell, "Weighted norm inequalities, off-diagonal estimates and elliptic operators, I: General operator theory and weights", *Adv. Math.* **212**:1 (2007), 225–276. MR Zbl

[Bernicot et al. 2016] F. Bernicot, T. Coulhon, and D. Frey, "Gaussian heat kernel bounds through elliptic Moser iteration", *J. Math. Pures Appl.* (9) **106**:6 (2016), 995–1037. MR Zbl

[Carbonaro et al. 2008] A. Carbonaro, G. Metafune, and C. Spina, "Parabolic Schrödinger operators", *J. Math. Anal. Appl.* **343**:2 (2008), 965–974. MR Zbl

[Chanillo and Wheeden 1986] S. Chanillo and R. L. Wheeden, "Harnack's inequality and mean-value inequalities for solutions of degenerate elliptic equations", *Comm. Partial Differential Equations* **11**:10 (1986), 1111–1134. MR Zbl

[Coifman and Weiss 1971] R. R. Coifman and G. Weiss, *Analyse harmonique non-commutative sur certains espaces homogènes*: *étude de certaines intégrales singulières*, Lecture Notes in Mathematics **242**, Springer, 1971. MR Zbl

[Coifman and Weiss 1977] R. R. Coifman and G. Weiss, "Extensions of Hardy spaces and their use in analysis", *Bull. Amer. Math. Soc.* **83**:4 (1977), 569–645. MR Zbl

[Cupini and Fornaro 2004] G. Cupini and S. Fornaro, "Maximal regularity in $L^p(\mathbb{R}^N)$ for a class of elliptic operators with unbounded coefficients", *Differential Integral Equations* **17**:3-4 (2004), 259–296. MR Zbl

[Duoandikoetxea 2001] J. Duoandikoetxea, *Fourier analysis*, Graduate Studies in Mathematics **29**, American Mathematical Society, Providence, RI, 2001. MR Zbl

[Folland 1975] G. B. Folland, "Subelliptic estimates and function spaces on nilpotent Lie groups", *Ark. Mat.* **13**:2 (1975), 161–207. MR Zbl

[Franchi and Lanconelli 1984] B. Franchi and E. Lanconelli, "An embedding theorem for Sobolev spaces related to nonsmooth vector fields and Harnack inequality", *Comm. Partial Differential Equations* **9**:13 (1984), 1237–1264. MR Zbl

[Franchi and Serapioni 1987] B. Franchi and R. Serapioni, "Pointwise estimates for a class of strongly degenerate elliptic operators: a geometrical approach", *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **14**:4 (1987), 527–568. MR Zbl

[Franchi et al. 1994] B. Franchi, C. E. Gutiérrez, and R. L. Wheeden, "Two-weight Sobolev–Poincaré inequalities and Harnack inequality for a class of degenerate elliptic operators", *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei* (9) *Mat. Appl.* **5**:2 (1994), 167–175. MR Zbl

[Garofalo and Vassilev 2007] N. Garofalo and D. Vassilev, "Strong unique continuation properties of generalized Baouendi–Grushin operators", *Comm. Partial Differential Equations* **32**:4-6 (2007), 643–663. MR Zbl

[Jerison and Sánchez-Calle 1987] D. Jerison and A. Sánchez-Calle, "Subelliptic, second order differential operators", pp. 46–77 in *Complex analysis, III* (College Park, MD, 1985–86), edited by C. A. Berenstein, Lecture Notes in Math. **1277**, Springer, 1987. MR Zbl

[Kato 1966] T. Kato, *Perturbation theory for linear operators*, Die Grundlehren der mathematischen Wissenschaften **132**, Springer, 1966. MR Zbl

[Kato 1972] T. Kato, "Schrödinger operators with singular potentials", *Israel J. Math.* **13** (1972), 135–148. MR Zbl

[Kim 1999] J. U. Kim, "An L^p a priori estimate for the Tricomi equation in the upper half space", *Trans. Amer. Math. Soc.* **351**:11 (1999), 4611–4628. MR Zbl

[Koch et al. 2015] H. Koch, A. Petrosyan, and W. Shi, "Higher regularity of the free boundary in the elliptic Signorini problem", *Nonlinear Anal.* **126** (2015), 3–44. MR Zbl

[Metafune and Spina 2008] G. Metafune and C. Spina, "An integration by parts formula in Sobolev spaces", *Mediterr. J. Math.* **5**:3 (2008), 357–369. MR Zbl

[Metafune et al. 2015] G. Metafune, M. Sobajima, and C. Spina, "Weighted Calderón–Zygmund and Rellich inequalities in L^p", *Math. Ann.* **361**:1-2 (2015), 313–366. MR Zbl

[Metafune et al. 2019] G. Metafune, L. Negro, M. Sobajima, and C. Spina, "Rellich inequalities in bounded domains", *Math. Ann.* (online publication December 2019).

[Ouhabaz 2005] E. M. Ouhabaz, *Analysis of heat equations on domains*, London Mathematical Society Monographs Series **31**, Princeton University Press, 2005. MR Zbl

[Robinson and Sikora 2008] D. W. Robinson and A. Sikora, "Analysis of degenerate elliptic operators of Grušin type", *Math. Z.* **260**:3 (2008), 475–508. MR Zbl

[Semenov 1977] Y. A. Semenov, "Schrödinger operators with L^p_{loc}-potentials", *Comm. Math. Phys.* **53**:3 (1977), 277–284. MR Zbl

[Shen 1995] Z. W. Shen, "L^p estimates for Schrödinger operators with certain potentials", *Ann. Inst. Fourier* (*Grenoble*) **45**:2 (1995), 513–546. MR Zbl

[Shen 2005] Z. Shen, "Bounds of Riesz transforms on L^p spaces for second order elliptic operators", *Ann. Inst. Fourier* (*Grenoble*) **55**:1 (2005), 173–197. MR Zbl

[Strömberg and Torchinsky 1989] J.-O. Strömberg and A. Torchinsky, *Weighted Hardy spaces*, Lecture Notes in Mathematics **1381**, Springer, 1989. MR Zbl

[Wang 2003] L. Wang, "Hölder estimates for subelliptic operators", *J. Funct. Anal.* **199**:1 (2003), 228–242. MR Zbl

[Ziemer 1989] W. P. Ziemer, *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*, Graduate Texts in Mathematics **120**, Springer, 1989. MR Zbl

Received 27 Nov 2019. Revised 24 Jan 2020. Accepted 15 Mar 2020.

GIORGIO METAFUNE: giorgio.metafune@unisalento.it
Dipartimento di Matematica "Ennio De Giorgi", Università del Salento, Lecce, Italy

LUIGI NEGRO: luigi.negro@unisalento.it
Dipartimento di Matematica "Ennio De Giorgi", Università del Salento, Lecce, Italy

CHIARA SPINA: chiara.spina@unisalento.it
Dipartimento di Matematica "Ennio De Giorgi", Università del Salento, Lecce, Italy

PURE and APPLIED
ANALYSIS

Vol. 2, No. 3, 2020
https://doi.org/10.2140/paa.2020.2.627

msp

STABILIZATION OF WAVE EQUATIONS ON THE TORUS WITH ROUGH DAMPINGS

NICOLAS BURQ AND PATRICK GÉRARD

For the damped wave equation on a compact manifold with *continuous* dampings, the geometric control condition is necessary and sufficient for uniform stabilization. On the two-dimensional torus, in the special case where $a(x)=\sum_{j=1}^N a_j 1_{x\in R_j}$ (R_j are polygons), we give a very simple necessary and sufficient geometric condition for uniform stabilization. We also propose a natural generalization of the geometric control condition which makes sense for L^∞ dampings. We show that this condition is always necessary for uniform stabilization (for any compact (smooth) manifold and any L^∞ damping), and we prove that it is sufficient in our particular case on $\mathbb{T}^2$ (and for our particular dampings).

Pour l'équation des ondes amortie sur une variété compacte, dans le cas d'un amortissement *continu*, la condition de contrôle géométrique est nécessaire et suffisante pour la stabilisation uniforme. Sur le tore $\mathbb{T}^2$ et dans le cas où $a(x)=\sum_{j=1}^N a_j 1_{x\in R_j}$ (R_j sont des polygones), nous exhibons une condition géométrique nécessaire et suffisante très simple. Nous proposons aussi une généralisation naturelle de la condition de contrôle géométrique, pour un amortissement seulement L^∞. Cette généralisation est toujours nécessaire pour la stabilisation uniforme (sur toute variété compacte régulière), et nous démontrons qu'elle est suffisante dans notre cas particulier du tore $\mathbb{T}^2$ (et pour nos fonctions d'amortissement particulières).

1. Notation and main results

Let (M,g) be a (smooth) compact Riemannian manifold endowed with the metric g, Δ_g the Laplace operator on functions on M and for $a\in L^\infty(M)$, let us consider the damped wave (or Klein–Gordon) equation

$$(\partial_t^2-\Delta+a(x)\,\partial_t+m)u=0,\quad (u|_{t=0},\partial_t u|_{t=0})=(u_0,u_1)\in(H^1\times L^2)(M),\tag{1-1}$$

where $0\le m\in L^\infty(M)$. If $a\ge 0$ a.e. it is well known that the energy

$$E_m(u)(t)=\int_M(|\nabla_g u|_g^2+|\partial_t u|^2+m|u|^2)\,d\mathrm{vol}_g\tag{1-2}$$

is decaying and satisfies

$$E_m(u)(t)=E_m(u)(0)-\int_0^t\int_M 2a(x)|\partial_t u|^2\,d\mathrm{vol}_g.$$

We shall say that the *uniform stabilisation* holds for the damping a if one of the following equivalent properties holds (see Appendix B for the equivalence):

MSC2010: 34L20, 35L05, 35Q93, 35S05, 58J40, 93D15.
Keywords: wave equations, control, stabilization, second microlocalization, geometric control.

(1) There exists a rate $f(t)$ such that $\lim_{t\to+\infty} f(t)=0$ and for any $(u_0,u_1)\in(H^1\times L^2)(M)$

$$E_m(u)(t)\le f(t)E_m(u)(0).$$

(2) There exists $C,c>0$ such that for any $(u_0,u_1)\in(H^1\times L^2)(M)$

$$E_m(u)(t)\le Ce^{-ct}E_m(u)(0).$$

(3) There exists $T>0$ and $c>0$ such that for any $(u_0,u_1)\in(H^1\times L^2)(M)$, if u is the solution to the damped wave equation (1-1), then

$$E_m(u)(0)\le C\int_0^T\int_M 2a(x)|\partial_t u|^2\,d\mathrm{vol}_g.$$

(4) There exists $T>0$ and $c>0$ such that for any $(u_0,u_1)\in(H^1\times L^2)(M)$, if u is the solution to the undamped wave equation

$$(\partial_t^2-\Delta+m)u=0,\quad (u|_{t=0},\partial_t u|_{t=0})=(u_0,u_1)\in(H^1\times L^2)(M),\tag{1-3}$$

then

$$E_m(u)(0)\le C\int_0^T\int_M 2a(x)|\partial_t u|^2\,d\mathrm{vol}_g.$$

The following result is classical; see [Rauch and Taylor 1974; 1975; Babich and Popov 1981; Babich and Ulin 1981; Ralston 1982; Bardos, Lebeau and Rauch 1992; Burq and Gérard 1997; Lebeau 1996; Koch and Tataru 1995; Sjöstrand 2000; Hitrik 2003.

Theorem 1 [Bardos, Lebeau and Rauch 1992; Burq and Gérard 1997]. *Let $m\ge 0$. Assume that the damping a is continuous. For $\rho_0=(x_0,\xi_0)\in S^*M$ denote by $\gamma_{\rho_0}(s)$ the geodesic starting from x_0 in (co-)direction ξ_0. Then the damping a stabilizes uniformly the wave equation if and only if the following geometric condition is satisfied*:

$$\exists T,c>0\quad\textit{such that}\quad \inf_{\rho_0\in S^*M}\int_0^T a(\gamma_{\rho_0}(s))\,ds\ge c.\tag{GCC}$$

When the damping a is not continuous but merely L^∞, an adaptation of the same techniques gives:

Theorem 2. *Assume that $a\in L^\infty(M)$. Then the strong geometric condition*

$$\begin{aligned}\exists T,c>0\quad&\textit{such that}\quad\forall\rho_0\in S^*M,\quad\exists s\in(0,T),\quad\exists\delta>0\\ &\textit{such that}\quad a\ge c\ \textit{ a.e. on } B(\gamma_{\rho_0}(s),\delta)\end{aligned}\tag{SGCC}$$

is ***sufficient*** *for uniform stabilisation, and the weak geometric condition*

$$\exists T>0\quad\textit{such that}\quad\forall\rho_0\in S^*M,\quad\exists s\in(0,T)\quad\textit{such that}\quad\gamma_{\rho_0}(s)\in\operatorname{supp}(a),\tag{WGCC}$$

where supp(a) *is the support* (*in the distributional sense*) *of a, is* ***necessary*** *for uniform stabilisation.*

Though the question appears to be very natural, until the present work, the only known case in between was essentially an example of [Lebeau 1992, pp. 15–16] (from an idea of J. Rauch) where $M = \mathbb{S}^d$ and a is the characteristic function of the half-sphere (notice however some refinements of (WGCC) in [Humbert, Privat and Trélat 2019; Trélat, Humbert and Privat 2017]). In the case of the half-sphere, uniform stabilisation holds (see [Zhu 2018] for a more detailed proof and a generalization of this result).

Theorem 3 [Lebeau 1992]. *On the d-dimensional sphere,*

$$\mathbb{S}^d = \{x = (x_0, \dots, x_d) \subset \mathbb{R}^{d+1} : \|x\| = 1\},$$

uniform stabilisation holds for the characteristic function of the half-sphere

$$\mathbb{S}^d_+ = \{x = (x_0, \dots, x_d) \subset \mathbb{R}^{d+1} : \|x\| = 1,\ x_0 > 0\}.$$

Remark 1.1. Notice that in this case, all the geodesics enter the interior of the support of a, and hence fulfill the (SGCC) requirements, except the family of geodesics included in the boundary of the support of a, the $(d-1)$-dimensional sphere,

$$\partial\mathbb{S}^d_+ = \{x = (x_0, \dots, x_d) \subset \mathbb{R}^{d+1} : \|x\| = 1,\ x_0 = 0\}.$$

When the manifold is a two-dimensional torus (rational or irrational) and the damping a is a linear combination of characteristic functions of polygons, i.e., there exists N, R_j, $j = 1, \dots, N$, (disjoint and not necessarily vertical) polygons and $0 < a_j$, $j = 1, \dots, N$, such that

$$a(x) = \sum_{j=1}^{N} a_j 1_{x \in R_j}, \tag{1-4}$$

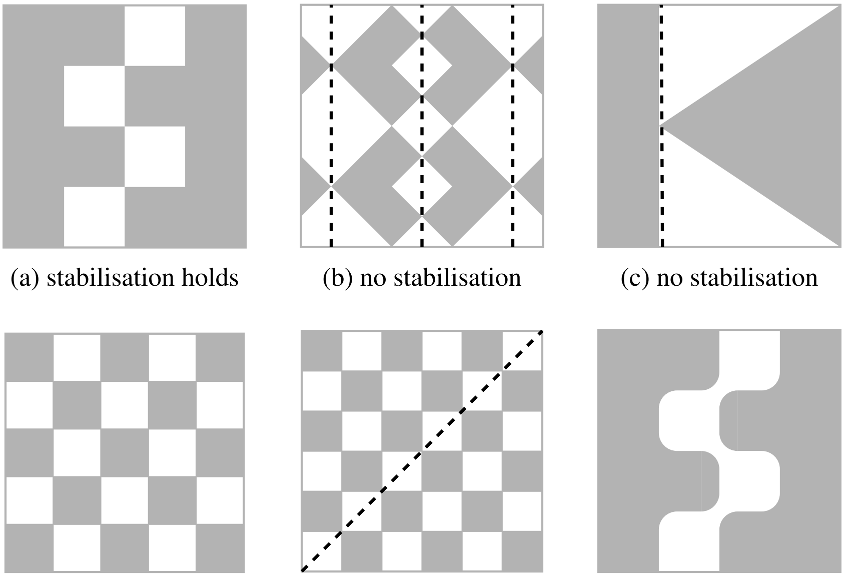
we can state another natural simple geometric condition. Let us endow the torus with an orientation (i.e., we see the torus as a surface in $\mathbb{R}^3$ and define at each point a normal vector $n(x)$). For any initial point x_0 and any norm-1 tangent vector X_0, let γ be the geodesic starting from x_0 in direction X_0 and parametrized by arc length. Let $\nu(\gamma(s))$ be the unique vector normal to γ in the torus and such that $(n(\gamma(s)), \dot\gamma(s), \nu(\gamma(s)))$ is a direct orthonormal frame. By convention, we shall say that ν points to the left of the geodesic (and $-\nu$ to the right).

Assumption 1.2. Assume that the manifold is a two-dimensional torus $\mathbb{T}^2 = \mathbb{R}^2/A\mathbb{Z} \times B\mathbb{Z}$, $A, B > 0$. Assume that there exists $T > 0$ such that all geodesics (straight lines) of length T either encounter the interior of one of the polygons or follow for some time one of the sides of a polygon R_{j_1} *on the left* and for some time one of the sides of a polygon R_{j_2} (possibly the same) *on the right*.

Our main result is the following:

Theorem 4. *The damping a stabilize uniformly the wave equation if and only if Assumption 1.2 is satisfied.*

Corollary 1.3. *Stabilisation holds for the examples* (a) *and* (d) *of Figure 1, but not for* (b), (c) *and* (e).


(a) stabilisation holds (b) no stabilisation (c) no stabilisation

(d) stabilisation holds (e) no stabilisation (f) stabilisation holds

Figure 1. Checkerboards: the damping a is equal to 1 in the gray region, and 0 elsewhere. For all these examples (WGCC) is satisfied but not (SGCC). The dashed lines are geodesics which violate Assumption 1.2.

Remark 1.4. In Assumption 1.2, as soon as the damping is nontrivial (i.e., we have at least one polygon), all nonclosed geodesics will enter the interior of this polygon (because any nonclosed geodesic is dense in the torus). As a consequence, the second part of the assumption has to be checked only for closed geodesics. Actually, closed geodesics corresponding to directions $(\xi, \eta) = (p, q)/\sqrt{p^2+q^2},\ p \wedge q = 1$, will also enter the polygon as soon as p^2+q^2 is large enough. As a consequence, the second part of Assumption 1.2 has to be checked only for a *finite number* of closed geodesics.

Remark 1.5. As pointed out by a referee, our proof actually gives a sufficient condition for stabilisation in a more general setting where the R_j need not be polygons, but are open subsets, and we assume that all but a finite number of closed geodesics are damped (in the sense that they enter the interior of one of the R_j's) and the remaining closed geodesics satisfy the left/right property on intervals of positive measure (which implies that the boundaries of the open sets R_{j_1} and R_{j_2} have some flat parts). As a consequence, stabilisation holds for Figure 1(f).

Remark 1.6. Stabilisation implies that exact controllability holds for some finite $T > 0$. However our proof relies on a contradiction argument and resolvent estimates. It gives no geometric interpretation for this controllability time. This contradiction argument allows us on tori to avoid a particularly delicate regime at the edge of the uncertainty principle (see Section 3A). Giving a geometric interpretation of the time necessary for control would require dealing with this regime; see [Burq ≥ 2020].

The plan of the paper is the following: In Section 2 we give a proof of Theorem 2 using classical methods. In Section 3, we focus on the model case of the checkerboard in Figure 1(a). We first reduce the question of uniform stabilisation to the proof of an observation estimate for high-frequency solutions of Helmholtz equations. We proceed by contradiction and construct good quasimodes, for the study of which we perform a microlocalisation which shows that the only obstruction is the vertical geodesic in the middle of the board. Then we prove a nonconcentration estimate which shows that solutions of Helmholtz equations (quasimodes) cannot concentrate too fast on this trajectory. This is essentially the only point in the proof which is specific to the torus and it relies on the special geometric structure of the torus which was previously used in the context of Schrödinger equations [Burq and Zworski 2004; 2005; 2012; Macià 2010; Anantharaman and Macià 2014; Bourgain, Burq and Zworski 2013] and also for wave equations [Burq and Hitrik 2007; Anantharaman and Léautaud 2014]. Finally, by means of a second microlocalisation with respect to this vertical geodesic, we obtain a contradiction. In Section 4, we show how the general case can be reduced to this model case. Finally, in the last section we introduce a generalized version of (GCC) that makes sense for $a \in L^\infty$ and which is equivalent to Assumption 1.2 in our particular case. We prove that this generalized geometric control condition is always necessary (on any Riemannian manifold and for any damping $0 \le a \in L^\infty$) and we conjecture that it is always sufficient. For the convenience of the reader, we gathered in Appendix A quite a few classical results about the link between resolvent estimates and stabilisation.

The *second microlocalisation* procedure has a well-established history starting with [Laurent 1979; 1985; Kashiwara and Kawai 1980; Sjöstrand 1982; Lebeau 1985] in the analytic context (see also [Bony and Lerner 1989] in the C^∞ framework and [Sjöstrand and Zworski 1999] in the semiclassical setting) and in the framework of defect measures by [Fermanian-Kammerer 2000; Miller 1996; 1997; 2000; Nier 1996; Fermanian-Kammerer and Gérard 2002; 2003; 2004]. Notice that most of these previous works in the framework of measures dealt with lagrangian or involutive submanifolds, and it is worth comparing our contribution with these previous works, in particular [Nier 1996; Anantharaman and Macià 2014]. Here we are interested in the wave equation, while the authors in [Nier 1996; Anantharaman and Macià 2014] were interested in the Schrödinger equation, and (compared to [Anantharaman and Macià 2014]) we are dealing with worse quasimodes ($o(h)$ instead of $o(h^2)$). Another difference is that we perform a second microlocalisation along a symplectic submanifold (namely $\{(x=0, y, \xi=0, \eta) \in T^*\mathbb{T}^2\}$), while they consider an isotropic submanifold $\{x = 0\}$ in [Nier 1996] or $\{(x', x'', \xi'=0, \xi'') \in T^*\mathbb{T}^d\}$ in [Anantharaman and Macià 2014]. An exception is [Fermanian-Kammerer 2005], to which our construction is very close. On the other hand, a feature shared by the present work and [Nier 1996; Anantharaman and Macià 2014] is that in all cases the analysis requires working at the edges of the uncertainty principle and using refinements of some exotic Weyl–Hörmander classes ($S^{1,1}$ in [Nier 1996], $S^{0,0}$ in [Anantharaman and Macià 2014] and $S^{1/2,1/2}$ in the present work); see [Hörmander 1985; Léautaud and Lerner 2017] for related work. Another worthwhile comparison is with [Burq and Hitrik 2007; Anantharaman and Léautaud 2014] on the damped wave equation on the torus when the control domain is arbitrary (in this case (WGCC) is in general not satisfied). However, though both works use some kind of second microlocalisation and deal with the wave equation, in [Burq and Hitrik 2007; Anantharaman and Léautaud

2014] the approaches use Schrödinger equations methods (strong quasimodes) transposed to get wave equations results and consequently lead to much weaker results (polynomial decay vs. exponential decay) under much weaker assumptions (arbitrary open sets).

2. First microlocalisation, proof of Theorem 2

In this section we work on an arbitrary compact manifold M with an arbitrary damping function $a \in L^\infty(M)$ and outline the classical propagation arguments which show that (SGCC) is sufficient for stabilisation, while (WGCC) is necessary. Let us assume (SGCC) holds. According to Proposition A.5, we need to prove (A-5)

$$\begin{gathered}\exists h_0 > 0 \quad \text{such that} \quad \forall\, 0 < h < h_0, \quad \forall (u, f) \in H^2(M) \times L^2(M), \quad (h^2\Delta + 1)u = f, \\ \|u\|_{L^2(M)} \leq C\Big(\|a^{\frac{1}{2}}u\|_{L^2} + \frac{1}{h}\|f\|_{L^2}\Big).\end{gathered} \tag{A-5}$$

To prove this estimate we argue by contradiction and obtain sequences $(h_n) \to 0$ and (u_n, f_n) such that

$$(h_n^2\Delta + 1)u_n = f_n, \quad \|u_n\|_{L^2} = 1, \quad \|a^{\frac{1}{2}}u_n\|_{L^2} = o(1)_{n\to+\infty}, \quad \|f_n\|_{L^2} = o(h_n)_{n\to+\infty}.$$

Extracting a subsequence, we can assume that the sequence (u_n) has a semiclassical measure ν on $T^*\mathbb{T}^2$. For $q \in C_0^\infty(T^*M)$, we define $\mathrm{Op}_h(q)$ by the following procedure. Using a partition of unity, we can assume that q is supported in a local chart. Then, in this chart, we define

$$\mathrm{Op}_h(q)(u) = \frac{1}{(2\pi h)^d}\int e^{\frac{i}{h}(x-y)\cdot\xi} q(x,\xi)\zeta(y)u(y)\,dy\,d\xi, \tag{2-1}$$

where $\zeta = 1$ in a neighborhood of the support of q (note that modulo smoothing $O(h^\infty)$ errors, this quantisation does not depend on the choice of the cut-off ζ). Then a semiclassical measure for the sequence (u_n) satisfies

$$\lim_{n\to+\infty}(\mathrm{Op}_{h_n}(q)u_n, u_n)_{L^2(M)} = \langle \nu, q\rangle.$$

In our case, it is supported in the characteristic set

$$\{(X, \Xi) \in S^*M : \|\Xi\|^2 = 1\}.$$

Furthermore, this measure has total mass 1 and is invariant by the bicharacteristic flow:

$$\Xi \cdot \nabla_X \nu = 0.$$

We refer to [Gérard and Leichtnam 1993; Burq 1997; 2002, Section 3] for the definition of semiclassical calculus on manifolds and semiclassical measures and a proof of these results in a very similar context (see also [Zworski 2012]). Let

$$S = \{x \in M : \text{there exists } \delta > 0, c > 0 \text{ such that } a \geq c \text{ on } B(x,\delta)\}.$$

Since $\|a^{1/2}u_n\|_{L^2} = o(1)_{n\to+\infty}$, we get that the measure ν vanishes in a neighborhood of every point $\rho \in S_x^*M$ for all $x \in S$. The assumption (SGCC) ensures that every bicharacteristic contains at least one point in S. Hence ν is identically 0, which contradicts the fact that it has total mass 1!

For the sake of completeness, let us now prove that stabilisation implies (WGCC). We are going to use that for any geodesic there exists a sequence of solutions to the wave equation which concentrates on this geodesic. This can be obtained through geometric optics constructions (see Proposition 5.1) or systematic use of semiclassical measures (see [Gérard 1996] in the special case of constant coefficients). In fact, we shall in Section 5A refine this approach and use *quantitative* estimates for this concentration to show that actually stabilisation implies the stronger condition (GGCC).

Proposition 2.1. *Assume that (WGCC) does not hold. Let $T > 0$. Consider a geodesic γ of length T which does not encounter the support of the damping function a. Then there exists a sequence (u_n) of solutions to the wave equation (1-3) which satisfies*

$$\lim_{n\to+\infty} E_m(u_n) = 1, \qquad \lim_{n\to+\infty} \int_0^T \int_M a(x)|\partial_t u|^2(t,x)\,dx\,dt = 0. \tag{2-2}$$

First by compactness, there exists $\delta > 0$ such that

$$\operatorname{dist}(\gamma([-\delta, T+\delta]), \operatorname{supp}(a)) \geq \delta.$$

Then, according to Proposition 5.1, there exists a sequence of approximate solutions (v_n) to the wave equation (with $m = 0$) which concentrate on the geodesic γ and satisfy (5-2) and (5-3). From (5-3), we deduce that

$$\|v_n\|_{L^2(M)} = O(h_n), \tag{2-3}$$

and from the concentration on the geodesic γ (and the δ separation with the support of a)

$$\int_0^T \int_M a(x)|\partial_t v_n|^2(t,x)\,dx\,dt = o(1)_{n\to+\infty} \tag{2-4}$$

uniformly with respect to $t \in [-1, T+1]$. The solution u_n to the wave equation (with m) (1-3) with the same initial data satisfies

$$(\partial_t^2 - \Delta + m)(u_n - v_n) = -m v_n, \quad (u_n - v_n)|_{t=0} = 0, \quad \partial_t(u_n - v_n)|_{t=0} = 0.$$

As a consequence of Duhamel's formula and (2-3),

$$\sup_{t\in[0,T]} \|u_n - v_n\|^2_{H^1(M)} + \|\partial_t u_n - \partial_t v_n\|^2_{L^2(M)} = O(h_n^2).$$

This implies according to (2-4)

$$E_m(u_n) = 1 + O(h_n^2), \qquad \int_0^T \int_M a(x)|\partial_t v_n|^2(t,x)\,dx\,dt = O(h_n^2).$$

3. The model case of a checkerboard

In this section we prove Theorem 4 for the model in Figure 2 on the two-dimensional torus $\mathbb{T}^2 = \mathbb{R}^2/(2\mathbb{Z})^2$. We shall later microlocally reduce the general case to this model.

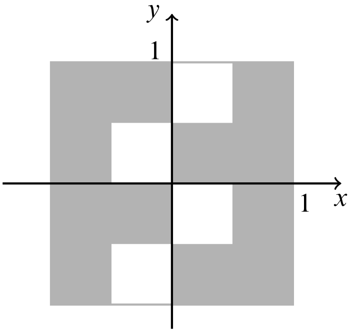


Figure 2. The checkerboard: a microlocal model where the damping a is equal to 1 in the gray region, and 0 elsewhere.

As previously, we prove (A-5) via a contradiction argument and construct the associated semiclassical measure ν. According to the results in Section 2, since the only two bicharacteristics which do not enter the interior of the set where $a = 1$ are

$$\{(x = 0, \xi = 0, \eta = \pm 1)\},$$

we know that ν is supported on the union of these two bicharacteristics.

3A. ***A priori nonconcentration estimate.*** In this section we show that (u_n) cannot concentrate on too small neighbourhoods around $\{x = 0\}$. This is the key (and only) point where we use the particular structure of the torus as a product manifold.

Let us recall that $\|(h_n^2\Delta + 1)u_n\|_{L^2} = o(h_n)$. Define

$$\epsilon(h_n) = \max\left(h_n^{\frac{1}{6}}, \left(\frac{\|(h_n^2\Delta + 1)u_n\|}{h_n}\right)^{\frac{1}{6}}\right), \tag{3-1}$$

so that

$$h_n^{-1}\epsilon^{-6}(h_n)\|(h_n^2\Delta + 1)u_n\|_{L^2} \le 1, \qquad \lim_{n\to+\infty} \epsilon(h_n) = 0. \tag{3-2}$$

The purpose of this section is to prove the following nonconcentration result which is actually related to Kakeya–Nikodym bounds; see [Sogge 2011; Blair and Sogge 2015; Miao, Sogge, Xi and Yang 2016].

Proposition 3.1. *Assume that $\|u_n\|_{L^2} = \mathcal{O}(1)$ and (3-2) holds. Then there exists $C > 0$ such that*

$$\forall n \in \mathbb{N}, \quad \|u_n\|_{L^2(\{|x|\le h_n^{1/2}\epsilon^{-2}(h_n)\})} \le C\epsilon^{\frac{1}{2}}(h_n).$$

The proposition follows from the following one-dimensional propagation estimate; see [Burq and Zuily 2015] for related estimates.

Proposition 3.2. *There exist $C > 0$ and h_0 such that for any $0 < h < h_0$, $1 \le \beta \le h^{-1/2}$, and any $(u, f) \in H^2 \times L^2$ a solution of*

$$(h^2(\partial_x^2 + \partial_y^2) + 1)u = f,$$

we have

$$\|u\|_{L^\infty(\{|x|\le\beta h^{1/2}\};L^2_y)}\le C\beta^{-\frac{1}{2}}h^{-\frac{1}{4}}(\|u\|_{L^2_{x,y}(\{\beta h^{1/2}\le|x|\le2\beta h^{1/2}\})}+h^{-1}\beta^2\|f\|_{L^2_{x,y}(\{|x|\le2\beta h^{1/2}\})}). \quad (3\text{-}3)$$

Let us first show that Proposition 3.1 follows from Proposition 3.2. Indeed, choosing $\beta=\epsilon^{-3}(h)$, Hölder's inequality gives

$$\begin{aligned}\|u\|_{L^2(\{|x|\le h^{1/2}\epsilon^{-2}(h)\})}&\\ &\le h^{\frac{1}{4}}\epsilon^{-1}(h)\|u\|_{L^\infty(\{|x|\le h^{1/2}\epsilon^{-3}(h\};L^2_y)}\\ &\le C\epsilon^{\frac{1}{2}}(h)(\|u\|_{L^2_{x,y}(\{h^{1/2}\epsilon^{-3}(h)\le|x|\le2h^{1/2}\epsilon^{-3}(h)\})}+h^{-1}\epsilon^{-6}(h)\|f\|_{L^2_{x,y}(\{|x|\le2\beta h^{1/2}\})})\\ &\le C\epsilon^{\frac{1}{2}}(h)(\|u\|_{L^2}+h^{-1}\epsilon^{-6}(h)\|f\|_{L^2})\le 2C\epsilon^{\frac{1}{2}}(h),\end{aligned} \quad (3\text{-}4)$$

where in the last inequality we used (3-2).

Now we can prove Proposition 3.2. Denote by v (resp. g) the partial Fourier transform of u (resp. f) with respect to y. For fixed x,

$$\|v(x,\cdot)\|_{L^2_\eta}=(2\pi)^{\frac{1}{2}}\|u(x,\cdot)\|_{L^2_y}.$$

We deduce that (3-3) is equivalent to

$$\begin{aligned}\|v\|_{L^\infty(\{|x|\le\beta h^{1/2}\};L^2_\eta)}&\\ &\le C\beta^{-\frac{1}{2}}h^{-\frac{1}{4}}(\|v\|_{L^2(\{\beta h^{1/2}\le|x|\le2\beta h^{1/2}\};L^2_\eta)}+h^{-1}\beta^2\|f\|_{L^2(\{|x|\le2\beta h^{1/2}\};L^2_\eta)}).\end{aligned} \quad (3\text{-}5)$$

Now, by Minkowski's inequality,

$$\|v\|_{L^\infty_x;L^2_\eta}\le\|v\|_{L^2_\eta;L^\infty_x}$$

and we deduce that (3-5) is implied by the following one-dimensional result.

Proposition 3.3. *There exist $C>0$ and h_0 such for any $0<h<h_0$, $\eta\in\mathbb{R}$, $1\le\beta\le h^{-1/2}$, and any (v,g) a solution of*

$$\left(h^2\frac{d^2}{dx^2}+1-h^2\eta^2\right)v=g,$$

we have

$$\|v\|_{L^\infty(\{|x|\le\beta h^{1/2}\})}\le C\beta^{-\frac{1}{2}}h^{-\frac{1}{4}}(\|v\|_{L^2(\{\beta h^{1/2}\le|x|\le2\beta h^{1/2}\})}+h^{-1}\beta^2\|g\|_{L^2(\{|x|\le2\beta h^{1/2}\})}). \quad (3\text{-}6)$$

We change variables $x=\beta h^{1/2}z$, and it is enough to prove, for solutions of

$$(h\beta^{-2}\,\partial_z^2+1-h^2\eta^2)v=g,$$

that

$$\|v\|_{L^\infty(\{|z|\le1\})}\le C(\|v\|_{L^2(\{1\le|z|\le2\})}+h^{-1}\beta^2\|g\|_{L^2(\{|z|\le2\})}). \quad (3\text{-}7)$$

Finally, this latter estimate follows (with $\tau=\beta^2h^{-1}(1-h^2\eta^2)$) from the result below, which is a generalization of [Burq and Zuily 2015, Proposition 3.2] (note that taking advantage of the dimension 1, we can replace the L^2 norm in the left of (3.3) of that work by an L^∞ norm).

Lemma 3.4. *There exists $C > 0$ such that, for any $\tau \in \mathbb{R}$ and any solution (v, k) on $(-2, 2)$ of*

$$(\partial_z^2 + \tau)v = k,$$

we have

$$\|v\|_{L^\infty(-1,1)} \le C\left(\|v\|_{L^2(\{1\le|z|\le 2\})} + \frac{1}{\sqrt{1+|\tau|}}\|k\|_{L^1(-2,2)}\right).$$

Let $\chi \in C_0^\infty(-2, 2)$ equal 1 on $(-1, 1)$. Then $u = \chi v$ satisfies

$$(\partial_z^2 + \tau)u = \chi k + 2\partial_z(\chi' v) - \chi'' v. \tag{3-8}$$

We distinguish two regimes.

Elliptic regime: $\tau \le -1$. Then, multiplying by u and integrating by parts gives

$$\begin{aligned}\|\partial_z u\|^2_{L^2(-2,2)} + |\tau|\,\|u\|^2_{L^2(-2,2)} &= -(\chi k + 2\partial_z(\chi' v) - \chi'' v, u)_{L^2}\\ &= -(\chi k - \chi'' v, u)_{L^2} + 2(\chi' v, \partial_z u)_{L^2},\end{aligned} \tag{3-9}$$

which implies

$$\begin{aligned}&\|\partial_z u\|^2_{L^2(-2,2)} + |\tau|\,\|u\|^2_{L^2(-2,2)}\\ &\qquad \le C\big(\|k\|_{L^1(-2,2)}\|u\|_{L^\infty} + \|v\|_{L^2(\{1\le|z|\le2\})}(\|u\|_{L^2(\{1\le|z|\le2\})} + \|\partial_z u\|_{L^2(-2,2)})\big),\end{aligned} \tag{3-10}$$

and the one-dimensional Gagliardo–Nirenberg inequality

$$\|u\|_{L^\infty} \le C\|\partial_z u\|^{\frac12}_{L^2}\|u\|^{\frac12}_{L^2}$$

allows us to conclude in this regime.

Hyperbolic regime: $\tau \ge -1$. Let $\sigma = \sqrt{\tau} \in \mathbb{R}^+ \cup i[0, 1]$. The solution of (3-8) is

$$u(x) = \int_{y=-2}^{x} e^{-i\sigma(x-y)}\int_{z=-2}^{y} e^{i\sigma(y-z)}g(z)\,dz\,dy = \int_{z=-2}^{x} g(z)\int_{y=z}^{x} e^{i\sigma(2y-x-z)}\,dy\,dz,$$

where $g = \chi k - \chi'' v + 2\partial_z(\chi' v) = g_1 + \partial_z g_2$. Since, for $x, z \in [-2, 2]$,

$$\left|\int_{y=z}^{x} e^{i\sigma(2y-x-z)}\,dy\right| \le \frac{C}{1+|\sigma|},$$

the contribution of g_1 is uniformly bounded by

$$\frac{C}{1+|\tau|}(\|\chi k\|_{L^1(-2,2)} + \|v\|_{L^1(\{1\le|z|\le2\})}).$$

Integrating by parts in the integral involving $\partial_z g_2$, we see that similarly, the contribution of $\partial_z g_2$ is bounded by

$$C\|\chi' v\|_{L^1(-2,2)}.$$

3B. ***Second microlocalisation.*** In this section we develop the tools required to understand the concentration properties of our sequence (u_n) on the symplectic submanifold $\{x=0, \xi=0\}$ of the phase space $T^*\mathbb{T}^2$. The construction is very close to the one in [Fermanian-Kammerer 2005].

3B1. *Symbols and operators.* We define S^m to be the class of smooth functions of the variables $(X, \Xi, z, \zeta) \in \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}$ which have compact supports with respect to the (X, Ξ)-variables and are polyhomogeneous of degree m with respect to the (z, ζ)-variables, with limits in the radial direction

$$\lim_{r\to+\infty} \frac{1}{r^m} a\left(X, \Xi, \frac{(rz, r\zeta)}{\|(z,\zeta)\|}\right) = \tilde{a}\left(X, \Xi, \frac{(z,\zeta)}{\|(z,\zeta)\|}\right).$$

When $m = 0$, via the change of variables

$$(z,\zeta) \mapsto (\tilde{z}, \tilde{\zeta}) = \frac{(z,\zeta)}{\sqrt{1+|z|^2+|\zeta|^2}},$$

such functions are identified with smooth compactly supported functions on $\mathbb{R}^4_{(X,\Xi)} \times \overline{B(0,1)}_{\tilde{z},\tilde{\zeta}}$, where $\overline{B(0,1)}$ denotes the closed unit ball in $\mathbb{R}^2$.

Let $\epsilon(h)$ satisfy

$$\lim_{h\to 0} \epsilon(h) = 0, \quad \epsilon(h) \geq h^{\frac{1}{2}}.$$

In order to perform the second microlocalisation around the submanifold given by the equations $x = 0$, $\xi = 0$, we define, for $a \in S^m$,

$$\operatorname{Op}_h(a) = a\left(x,\, y,\, hD_x,\, hD_y,\, \frac{\epsilon(h)}{h^{\frac{1}{2}}} x,\, \epsilon(h) h^{\frac{1}{2}} D_x\right),$$

where $X = (x, y)$ and $\Xi = (\xi, \eta)$. Notice that this quantification is the usual one [Hörmander 1985, Chapter 18.1], associated to the symbol

$$a\left(x,\, y,\, \xi,\, \eta,\, \frac{\epsilon(h)}{h^{\frac{1}{2}}} x,\, \epsilon(h) h^{\frac{1}{2}} \xi\right).$$

A simple calculation shows that since $\epsilon(h) \geq h^{1/2}$, the latter symbol belongs to the class

$$S((1+\epsilon^2(h)h^{-1}x^2 + \epsilon^2(h)h\xi^2)^{\frac{m}{2}}, g)$$

of the Weyl–Hörmander calculus [Hörmander 1985, Chapter 18.5] for the metric

$$g = \frac{\epsilon^2(h)}{h} \frac{dx^2}{1+\epsilon^2(h)h^{-1}x^2+\epsilon^2(h)h\xi^2} + \epsilon^2(h)h \frac{d\xi^2}{1+\epsilon^2(h)h^{-1}x^2+\epsilon^2(h)h\xi^2} + \frac{dy^2}{1+y^2+h^2\eta^2} + h^2 \frac{d\eta^2}{1+y^2+h^2\eta^2}. \tag{3-11}$$

As a consequence, we deduce that the operators such defined enjoy good properties and we have a good symbolic calculus; namely for all $a \in S^0$, the operator $\operatorname{Op}(a)$ is bounded on $L^2(\mathbb{R}^2)$ uniformly with respect to h, and

$$\forall a \in S^p,\, b \in S^q,\, ab \in S^{p+q}, \quad \operatorname{Op}(a)\operatorname{Op}(b) = \operatorname{Op}(ab) + \epsilon^2(h) r,$$

where $r \in \operatorname{Op}(S^{p+q-1})$, and

$$\forall a \in S^0,\ a \geq 0 \quad \Longrightarrow \quad \exists C > 0 \text{ such that } \operatorname{Re}(\operatorname{Op}(a)) \geq -C\epsilon^2(h) \text{ and } \|\operatorname{Im}(\operatorname{Op}(a))\| \leq C\epsilon^2(h).$$

3B2. *Definition of the second semiclassical measure.* In this section, we consider a sequence (u_n) of functions on the two-dimensional torus $\mathbb{T}^2$ such that

$$(h_n^2\Delta+1)u_n=\mathcal{O}(1)_{L^2}. \tag{3-12}$$

We identify u_n with a periodic function on $\mathbb{R}^2$. Now, using the symbolic calculus properties in Section 3B1, and in particular Gårding's inequality and the L^2 boundedness of operators, we can extract a subsequence (still denoted by (u_n)) such that there exists a positive measure $\tilde{\mu}$ on $T^*\mathbb{T}^2\times\overline{N}$ — $\overline{N}$ denotes the sphere compactification of $N=\mathbb{R}^2_{z,\zeta}$ — such that, for any symbol $a\in S^0$,

$$\lim_{n\to+\infty}(\mathrm{Op}_{h_n}(a)u_n,u_n)_{L^2}=\langle\tilde{\mu},\tilde{a}\rangle,$$

where the continuous function $\tilde{a}$ on $T^*\mathbb{R}^2\times\overline{N}$ is naturally defined in the interior by the value of the symbol a and on the sphere at infinity by

$$\tilde{a}(x,y,\xi,\eta,\tilde{z},\tilde{\zeta})=\lim_{r\to+\infty}a(x,y,\xi,\eta,r\tilde{z},r\tilde{\zeta})$$

(which exists because a is polyhomogeneous of degree 0). The measure $\tilde{\mu}$ is of course periodic, and hence defines naturally a measure μ on $T^*\mathbb{T}^2\times\overline{N}$, and using (3-12), it is easy to see that there is no loss of mass at infinity in the Ξ-variable:

$$\mu(T^*\mathbb{T}^2\times\overline{N})=\lim_{n\to+\infty}\|u_n\|^2_{L^2(\mathbb{T}^2)}. \tag{3-13}$$

3B3. *Properties of the second semiclassical measure.* We now turn to the sequence constructed in Section 2 and study refined properties of the second semiclassical measure constructed above, for the choice $\epsilon(h)$ given by (3-1). Notice that compared to (3-12) the sequence considered here satisfies the stronger

$$(h_n^2\Delta+1)u_n=o(h_n)_{L^2}.$$

Proposition 3.5. *The measure μ satisfies the following properties*:

(1) *Assume only that*

$$(h_n^2\Delta+1)u_n=O(1)_{L^2}.$$

Then the measure μ has total mass $1=\|u_n\|^2_{L^2}$ *(h_n-oscillation).*

(2) *Assume now that*

$$(h_n^2\Delta+1)u_n=o(h_n)_{L^2}$$

and $\|au_n\|_{L^2}=o(1)$. *Then, since the projection of the measure μ on the (x,y,ξ,η)-variables is the measure ν of Section 2 which is invariant by the bicharacteristic flow, we get that the measure μ is supported on the set*

$$\{(x,y,\xi,\eta,z,\zeta);x=0,\xi=0,\eta=\pm1\}.$$

(3) *Assume now that*

$$(h_n^2\Delta+1)u_n=O(h_n\epsilon(h_n))_{L^2}.$$

Then the measure μ is supported on the sphere at infinity in the (z,ζ)-variables.

(4) *Assume now that*

$$(h_n^2\Delta+1)u_n = O(h_n\epsilon(h_n))_{L^2}, \quad \|1_R u_n\|_{L^2} = o(1),$$

where R is a polygon. Then the measure μ vanishes 2-microlocally at each point of ∂R on the side where the polygon R lies. Namely in our geometry, the measure μ vanishes 2-microlocally on the right on $\{x=0,\ y\in(0,\frac{1}{2})\cup(-1,-\frac{1}{2})\}$ and 2-microlocally on the left on $\{x=0, y\in(-\frac{1}{2},0)\cup(\frac{1}{2},1)\}$; more precisely,

$$\begin{aligned}&\mu\big(\{(x,y,\xi,\eta,z,\zeta):x=0,\ y\in\big(0,\tfrac{1}{2}\big)\cup\big(-1,-\tfrac{1}{2}\big),\ z>0\}\big)=0,\\ &\mu\big(\{(x,y,\xi,\eta,z,\zeta):x=0,\ y\in\big(-\tfrac{1}{2},0\big)\cup\big(\tfrac{1}{2},1\big),\ z<0\}\big)=0.\end{aligned} \tag{3-14}$$

(5) *According to point (3) above, if we identify the sphere at infinity in the (z,ζ)-variables with $\mathbb{S}^1$ by means of the choice of variables $z=r\cos(\theta)$, $\zeta=r\sin(\theta)$, $r\to+\infty$, the measure μ can be seen as a measure in (x,y,ξ,η,θ)-variables, supported on $x=0,\ \xi=0,\ \eta=\pm1$. In this coordinate system,*

$$(\eta\,\partial_y-\sin^2(\theta)\,\partial_\theta)\mu=0. \tag{3-15}$$

Remark 3.6. In Proposition 3.5, the only point where we use crucially the particular geometry of the torus (Proposition 3.1) is point (3). For more general geometries, this point is not true. However, for the part on the sphere at infinity of the measure, we can still get an analog of (3-15) for more general geometries, involving the curvature of the surface along the geodesic; see [Burq ≥ 2020].

Proof. The proof of point (1) follows from (3-13). To prove point (2), we just remark that the choice of test functions $a(x,\Xi,z,\zeta)=a(X,\Xi)$ shows that the direct image $\pi_*(\mu)$ of μ by the map

$$\pi:(X,\Xi,z,\zeta)\mapsto(X,\Xi)$$

is actually the (first) semiclassical measure ν constructed in Section 2, and consequently, this property follows from Section 2. To prove point (3), we recall that from Proposition 3.1, we have that for any $\chi\in C_0^\infty$, bounded by 1 and supported in $(-A,A)$,

$$\begin{aligned}\|\chi(h_n^{-\frac{1}{2}}\epsilon(h_n)x)u_n\|_{L^2}^2 &\le \|u_n\|_{L^2(\{|x|\le Ah^{1/2}\epsilon^{-1}(h)\}}\\ &\le \|u_n\|_{L^2(\{|x|\le h^{1/2}\epsilon^{-2}(h)\}}\le C\epsilon^{\frac{1}{2}}(h_n) \quad\Longrightarrow\quad \langle\mu,\chi(z)\rangle=0.\end{aligned} \tag{3-16}$$

To prove point (4), recall from Figure 2 that the damping a is equal to 1 on $(0,\frac{1}{2})\times(0,\frac{1}{2})$ and that

$$\|au_n\|_{L^2}=\|a^{\frac{1}{2}}u_n\|_{L^2}=o(1)_{n\to+\infty}.$$

Point (4) will follow from:

Proposition 3.7. *Assume that*

$$\|au_n\|_{L^2}=o(1)_{n\to+\infty},$$

and that the damping is equal to 1 on $(0,\delta)\times(c,d)$ (resp. $(-\delta,0)\times(c,d)$). Then the measure μ vanishes two-microlocally on the right (resp. on the left) above $T^*\mathbb{T}^2|_{\{0\}\times(c,d)}$:

$$\begin{aligned}&\mu(\{(x,y,\xi,\eta,z,\zeta):x=0,\ y\in(c,d),\ \eta=\pm1,\ z>0\})=0,\\ (\textit{resp.}\quad &\mu(\{(x,y,\xi,\eta,z,\zeta):x=0,\ y\in(c,d),\ \eta=\pm1,\ z<0\})=0),\end{aligned} \tag{3-17}$$

Let $\psi \in C^\infty(\mathbb{R})$ be supported in $\{1 < r\}$ and equal 1 for $r \geq 2$. Let $\chi \in C_0^\infty(-1,1)$ equal 1 on $\left(-\frac{1}{2}, \frac{1}{2}\right)$, and $\tilde{\chi} \in C_0^\infty(c,d)$ equal 1 on $c+\delta_0, d-\delta_0$, $\delta > 0$. Consider the symbol

$$b(x,y,\xi,\eta,z,\zeta) = \chi\left(\frac{2x}{\delta}\right)\tilde{\chi}(y)\chi(\xi)\chi(\eta-1)\psi\left(\frac{z}{\delta_0|\zeta|}\right)\psi(z^2+\zeta^2).$$

On the other hand, since $\chi(2x/\delta)\tilde{\chi}(y)$ is supported on $\left(-\frac{\delta}{2}, \frac{\delta}{2}\right)_x \times (c,d)_y$ and since $\psi(z/\delta|\zeta|)$ is supported in $z > 0$, we infer that the range of $\operatorname{Op}_{h_n}(b)$ is supported in the domain $\left(0, \frac{\delta}{2}\right)_x \times (c,d)_y$ and consequently

$$\begin{aligned}(\operatorname{Op}_{h_n}(b)u_n, u_n) &= (1_{x\in(0,\frac{\delta}{2})}1_{y\in(c,d)}\operatorname{Op}_{h_n}(b)u_n, u_n) = (\operatorname{Op}_{h_n}(b)u_n, 1_{x\in(0,\frac{\delta}{2})}1_{y\in(c,d)}u_n)\\ &= (\operatorname{Op}_{h_n}(b)u_n, 1_{x\in(0,\frac{\delta}{2})}1_{y\in(c,d)}au_n) = o(1)_{n\to+\infty}.\end{aligned} \tag{3-18}$$

This implies

$$\mu(\{(x,y,\xi,\eta,z,\zeta) : x=0,\ y\in(c+\delta_0, d-\delta_0),\ \eta=1,\ z\geq 2\delta_0|\zeta|\}) = 0.$$

Taking $\delta_0 > 0$ arbitrarily small, we deduce that on the (z,ζ)-sphere at infinity which contains the support of μ, we have

$$\mu(\{(x,y,\xi,\eta,z,\zeta) : x=0,\ y\in(c,d),\ \eta=1,\ z>0\}) = 0.$$

The case $\eta = -1$ and the other properties in (3-14) follow similarly.

To prove the last property, we write for $q \in S^0$

$$\begin{aligned}&\frac{1}{2ih_n}[h_n^2\Delta+1, \operatorname{Op}_{h_n}(q)]\\ &\qquad = \operatorname{Op}_{h_n}((\xi\,\partial_x+\eta\,\partial_y+\zeta\,\partial_z)q) - i\frac{h_n}{2}\operatorname{Op}_{h_n}(\Delta_{x,y}q) - i\frac{h_n}{2}(\epsilon(h_n)h_n^{-\frac{1}{2}})^2\operatorname{Op}_{h_n}(\partial_z^2 q).\end{aligned} \tag{3-19}$$

Since unfolding the bracket shows that, as $n\to\infty$,

$$\frac{1}{2ih_n}([h_n^2\Delta+1, \operatorname{Op}_{h_n}(q)]u_n, u_n) \to 0,$$

we get

$$o(1)_{n\to\infty} = (\operatorname{Op}_{h_n}((\xi\,\partial_x+\eta\,\partial_y+\zeta\,\partial_z)q)u_n, u_n). \tag{3-20}$$

Let us compute the limit on the sphere at infinity of $(\xi\,\partial_x+\eta\,\partial_y+\zeta\,\partial_z)q$. We denote by $\tilde{q}$ the function q in the (r,θ)-coordinate system. In this system of coordinates, the operator $\zeta\,\partial_z$ is given by

$$-\sin^2(\theta)\,\partial_\theta + r\cos(\theta)\sin(\theta)\,\partial_r.$$

Now we use that, for a polyhomogeneous symbol q of *degree* 0, the main part of q at infinity does not depend on r. As a consequence, the symbol $r\,\partial_r q$ is polyhomogeneous of *degree* -1 (while homogeneity would dictate *degree* 0). Therefore we get, for any polyhomogeneous symbol q of degree 0,

$$\begin{aligned}\zeta\,\partial_z q|_{\mathbb{S}^1} &= \lim_{r\to+\infty}(-\sin^2(\theta)\,\partial_\theta + r\cos(\theta)\sin(\theta)\,\partial_r)\tilde{a}(x,y,\xi,\eta,r,\theta)\\ &= -\sin^2(\theta)\,\partial_\theta\lim_{r\to+\infty}\tilde{q}(x,y,\xi,\eta,r,\theta).\end{aligned} \tag{3-21}$$

Since the measure $\tilde{\mu}$ is supported in $\xi = 0$, (3-15) follows from (3-20). □

We can now conclude the contradiction argument, and end the proof of the resolvent estimate (A-5). Notice that the two fixed points for the flow of

$$\dot{\theta} = -\sin^2(\theta)$$

are given by $\theta = 0(\pi)$. We want to show that the measure $\tilde{\mu}$ vanishes identically to get a contradiction with point (1) in Proposition 3.5. For $(x{=}0, \xi{=}0, y_0, \eta_0{=}\pm 1, \theta_0)$ in the support of $\tilde{\mu}$, let us denote by $\phi_s(\theta_0)$ the solution of

$$\frac{d}{ds}\phi_s(\theta_0) = -\sin^2(\phi_s(\theta_0)), \quad \phi_0(\theta_0) = \theta_0,$$

so that $\phi_s(\theta_0) = \operatorname{Arccotan}(s + \operatorname{cotan}(\theta_0))$. From the invariance (3-15) of the measure $\tilde{\mu}$, we deduce that

$$\forall s \in \mathbb{R}, \quad (x = 0,\, y_s = y_0 + s\eta_0 \pmod{2\pi},\, \xi = 0,\, \eta_0,\, \theta_s = \phi_s(\theta_0)) \in \operatorname{supp}(\tilde{\mu}).$$

Consequently, if $\theta_0 \in [0, \pi)$, there exists $s > 0$ such that $y_s \in \left(0, \frac{1}{2}\right) \pmod{2\pi}$ if $\theta_s \in \left[0, \frac{\pi}{2}\right)$, while, if $\theta_0 \in [-\pi, 0)$, there exists $s > 0$ such that $y_s \in \left(-\frac{1}{2}, 0\right) \pmod{2\pi}$ if $\theta_s \in \left[-\pi, -\frac{\pi}{2}\right)$. This is impossible according to (3-14).

4. Back to the general case

Let us work on the torus $\mathbb{T}^2 = \mathbb{R}^2/A\mathbb{Z}\times B\mathbb{Z}$ with $A > 0$, $B > 0$. Since the irrational directions $\Xi = (A\xi, B\eta)$, with $\xi/\eta \notin \mathbb{Q}$, correspond to dense geodesics, and since a is bounded from below on an open set, we deduce that the measure ν defined in Section 2 is supported — in the Ξ-variables — on the set of finitely many rational directions

$$\Xi = (A\xi, B\eta), \quad \text{with } \frac{\xi}{\eta} \in \mathbb{Q},$$

satisfying moreover the elliptic regularity condition, $|\Xi|^2 = 1$, which do not enter the interior of the rectangles. Hence, there exists an isolated direction Ξ_0, so that $(X_0, \Xi_0) \in \operatorname{supp}(\nu)$, which can be written as

$$\Xi_0 = \frac{1}{\sqrt{n^2A^2 + m^2B^2}}(nA, mB), \quad \Xi_0^{\perp} = \frac{1}{\sqrt{n^2A^2 + m^2B^2}}(-mB, nA), \tag{4-1}$$

where the integers n, m may be chosen to have gcd 1. The change of coordinates in $\mathbb{R}^2$,

$$F : (x, y) \longmapsto X = F(x, y) = x\Xi_0^{\perp} + y\Xi_0, \tag{4-2}$$

is orthogonal and hence $-\Delta_X = D_x^2 + D_y^2$.

We have the following simple lemma [Burq and Zworski 2012, Lemma 2.7], which can be deduced from an elementary calculation.

Lemma 4.1. *Suppose that Ξ_0 and F are given by (4-1) and (4-2). If $u = u(x, y)$ is periodic with respect to $A\mathbb{Z} \times B\mathbb{Z}$ then $F^*u := u \circ F$ satisfies*

$$F^*u(x + k\alpha, y + \ell\beta) = F^*u(x, y - k\gamma), \quad k, \ell \in \mathbb{Z},\ (x, y) \in \mathbb{R}^2, \tag{4-3}$$

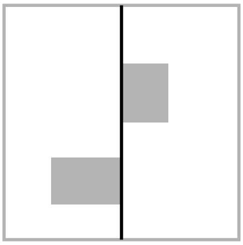


Figure 3. The microlocal model: on the left the rectangle R_1, on the right the rectangle R_2, in the middle the bicharacteristic in the support of μ.

where, for fixed $p, q \in \mathbb{Z}$ *such that* $qn - pm = 1$,

$$\alpha = \frac{AB}{\sqrt{n^2A^2 + m^2B^2}}, \quad \beta = \sqrt{n^2A^2 + m^2B^2}, \quad \gamma = -\frac{pnA^2 + qmB^2}{\sqrt{n^2A^2 + m^2B^2}}.$$

When $B/A = r/s \in \mathbb{Q}$, $r, s \in \mathbb{Z} \setminus \{0\}$,

$$F^*u(x + k\tilde{\alpha}, y + \ell\beta) = F^*u(x, y), \quad k, \ell \in \mathbb{Z},\ (x, y) \in \mathbb{R}^2, \tag{4-4}$$

for $\tilde{\alpha} = (n^2s^2 + m^2r^2)\alpha$.

In this new coordinate system, we know that there exists x_0 such that $(x_0, y_0, 0, 1)$ is in the support of the measure $F^*\mu$. By translation invariance, we can assume that $x_0 = 0$. Since

$$(\xi\, \partial_x + \eta\, \partial_y)F^*\mu = 0,$$

we infer that actually the whole line $(x_0{=}0, \mathbb{R} \pmod{2\pi}, 0, 1)$ belongs to the support of $F^*\mu$. If this bicharacteristic curve enters the interior of the support of a (i.e., encounters a point in a neighborhood of which a is bounded away from 0), then by propagation, no point of this bicharacteristic curve lies in the support of μ, which gives a contradiction. On the other hand since Assumption 1.2 is satisfied, we know that there exist two (at least) polygons R_1, R_2 so that the right side of R_1 is $\{0\} \times [\alpha, \beta]$, while the left side of R_2 is $\{0\} \times [\gamma, \delta]$. We may shrink these polygons to rectangles having the same property.

In other words, we are microlocally reduced to the study of the checkerboard in Figure 3. Notice that the change of variables we used in Lemma 4.1 does not keep periodicity with respect to the x-variables but transforms it into some pseudoperiodicity condition (see (4-3)). However, for the study of the checkerboard model in Section 3, we only used periodicity with respect to the y-variables (to prove Proposition 3.1) — which is preserved. The rest of the contradiction argument follows the same lines as in Section 3.

5. Generalized geometric condition

For a general Riemannian manifold and a general damping function $a \in L^\infty(M)$, a natural substitute to (GCC) is the following generalized geometric condition:

$$\exists T, c > 0 \quad \text{such that} \quad \liminf_{\epsilon \to 0} \inf_{\rho_0 \in S^*M} \frac{1}{\mathrm{Vol}(\Gamma_{\rho_0,\epsilon,T})} \int_{\Gamma_{\rho_0,\epsilon,T}} a(x)\, dx \geq c, \tag{GGCC}$$

where $\Gamma_{\rho_0,\epsilon,T}$ is the set of points at distance less than ϵ from the geodesic segment $\{\gamma_{\rho_0}(s) : s \in (0,T)\}$. At first glance, (GGCC) might seem to be a strong condition, difficult to fulfill. We shall prove below that it cannot be relaxed as, on any manifold and for any $a \in L^\infty(M)$, it is a necessary condition for uniform stabilization. On the other hand, we also prove below that in the case of two-dimensional tori it is equivalent to Assumption 1.2. We conjecture that on a general manifold and for general $a \in L^\infty$, uniform stabilization holds if and only if (GGCC) holds. The results in this article show that it is indeed the case on two-dimensional tori, if a satisfies (1-4). For general dampings it is easy to show that (GGCC) implies (WGCC), while the compactness of S^*M shows that it is implied by (SGCC) (δ in (SGCC) can, by compactness, be chosen the same for all $\rho_0 \in S^*M$).

5A. *The generalized geometric condition is necessary for stabilization.*

Theorem 5. *Uniform stabilization implies (GGCC).*

Proof. The proof of this result relies on geometric optics constructions (with complex phases) for the wave equation of [Ralston 1982, Section 2.1] that we recast in our wave equation context.

Proposition 5.1. *Let M be a compact manifold without boundary endowed with a smooth metric g and a smooth density κ. Let*

$$\Delta = \operatorname{div}_\kappa \nabla_g \tag{5-1}$$

be the Laplace operator. Let $\big(t_0, x_0, \tau_0{=}\tfrac{1}{2}, \xi_0\big) \in \operatorname{Char}(\partial_t^2 - \Delta)$, the characteristic manifold. Denote by $\big(t(s){=}t_0 + s, \tau_0{=}\tfrac{1}{2}, \gamma(s), \xi(s)\big)$ the bicharacteristic starting from $\big(t_0, x_0, \tfrac{1}{2}, \xi_0\big)$. Then for any $N > 0$ there exists a family of approximate solutions $v_{h,N}(t,x)$ defined for $0 < h < h_0$ to the wave equation

$$(\partial_t^2 - \Delta)v_{h,N} = \mathcal{O}(h^N)_{L^2(M)}, \quad E(v_{h,N}) = \int_M (|\nabla_g v_{h,N}|^2 + |\partial_t v_{h,N}|^2)\kappa\, dx = 1 + o(h) \tag{5-2}$$

with error terms locally uniformly controlled in time, and which are (locally in time) exponentially localized in $\mathbb{R}_t \times M$ near $(t(s), x(s))$:

$$\begin{gathered}\forall T > 0, \quad \exists C, \alpha > 0 \quad \textit{such that} \quad \forall t \in [0,T], \quad \forall x \in M,\\ (|v_{h,N}| + |h\nabla_x v_{h,N}| + |h\,\partial_t v_{h,N}|)(t(s), x) \leq C h^{1-\frac{d}{4}} e^{-\alpha \frac{\operatorname{dist}(x,x(s))^2}{h}}.\end{gathered} \tag{5-3}$$

Consequently, if we denote by $\Gamma_T = \gamma([0,T])$ the image of the geodesic in M,

$$\begin{gathered}\forall T > 0, \quad \exists C, \alpha > 0 \quad \textit{such that} \quad \forall x \in M,\\ \int_0^T (|\nabla_g v_{h,N}|^2 + |\partial_t v_{h,N}|^2)(t,x)\, dt \leq C h^{-\frac{d-1}{2}} e^{-\alpha \frac{\operatorname{dist}(x,\Gamma_T)^2}{h}}.\end{gathered} \tag{5-4}$$

Let us first show how we can deduce Theorem 5 from Proposition 5.1. We are going to test the observation estimates (1-3) on such sequences of solutions.

Let us we assume that (GGCC) does not hold. Fix $T > 0$. Then there exist $\eta_n = (x_n, \xi_n) \in S^*M$, $\epsilon_n \to 0$ such that

$$\lim_{n \to +\infty} \kappa_n = 0, \quad \kappa_n := \frac{1}{\epsilon_n^{d-1}} \int_{\Gamma_{\eta_n,\epsilon_n,T}} a(x)\, dx.$$

Let $t_n = 0$. Let $\rho_n = \bigl(t_n, \tau_n{=}\frac{1}{2}, x_n, \xi_n\bigr)$, fix $N = 1$ (we actually need a crude version of Ralston's construction) and let v_h^n be the approximate solution of the wave equation constructed in Proposition 5.1, with initial point ρ_n. We shall use that the family of solutions which depends on two parameters h and the initial point in the cotangent bundle is uniformly controlled with respect to this latter parameter, which will follow from the proof of Proposition 5.1 given below. Since, according to Proposition 5.1, we have

$$\|(v_h^n, \partial_t v_h^n)|_{t=0}\|_{\dot{H}^1\times L^2} = 1 + o(1)_{n\to+\infty},$$

and according to (5-2) and Duhamel's formula, v_h^n is, modulo a $\mathcal{O}(h_n^N)$ error in energy space, equal to the solution to the exact wave equation with the same initial data, to show that uniform stabilization does not hold, it is now enough to show that for a properly chosen sequence $h_n \to 0$

$$\lim_{n\to+\infty} \int_0^T \int_M a(x)|\partial_t v_{h_n,n}|^2\,dx\,dt = 0. \tag{5-5}$$

Extracting a subsequence, we can assume that the sequence of initial points ρ_n converges to $\rho = \bigl(t_0{=}0, \xi_0{=}\frac{1}{2}, x_0, \xi_0\bigr)$. The only point we shall use about our approximate solutions is the upper bound (5-4), which implies

$$\begin{aligned}\int_0^T\int_M a(x)|\partial_t v_{h_n}|^2\,dx\,dt\\ &= \int_0^T\int_M a(x)|\partial_t v_{h_n}|^2\,dx\,dt\\ &\le C\int_{\Gamma_{\rho_n,\epsilon_n,T}} a(x)h_n^{-\frac{d-1}{2}} e^{-\alpha\frac{\operatorname{dist}(x,\Gamma_{\rho_n,T})^2}{h_n}}\,dx + \int_{\Gamma^c_{\rho_n,\epsilon_n,T}} a(x)h_n^{-\frac{d-1}{2}} e^{-\alpha\frac{\operatorname{dist}(x,\Gamma_{\rho_n,T})^2}{h_n}}\,dx.\end{aligned} \tag{5-6}$$

The contribution of the first term is bounded by

$$Ch_n^{-\frac{d-1}{2}}\int_{\Gamma_{\rho_n,\epsilon_n,T}} a(x)\,dx \le \kappa_n\left(\frac{\epsilon_n^2}{h_n}\right)^{\frac{d-1}{2}}.$$

On the other hand, the second term is bounded by

$$\|a\|_{L^\infty}\int_{\Gamma^c_{\rho_n,\epsilon_n,T}} h_n^{-\frac{d-1}{2}} e^{-\alpha\frac{\operatorname{dist}(x,\Gamma_{\rho_n,T})^2}{h_n}}\,dx. \tag{5-7}$$

To estimate this integral we work in (a finite set of) coordinate systems. In such local coordinates, $\Gamma_{\rho_n,T}$ is a finite union of smooth arcs of geodesics (because the geodesic can self-intersect) and it is enough to estimate (5-7) where we replaced $\operatorname{dist}(x,\Gamma_{\rho_n,T})$ by the distance to any such arc. We can change again coordinates such that locally the considered arc of geodesic is

$$\{(y_1 = 0, y' \in \mathbb{R}^{d-1})\},$$

and the distance to the arc γ_n satisfies

$$\exists C > 0 \quad \text{such that} \quad \frac{1}{C}|y'| \le \operatorname{dist}(y,\gamma_n) \le C|y'|.$$

This leads to the estimate (if $\epsilon_n \geq C^2\sqrt{h_n}$)

$$\int_{\operatorname{dist}(x,\gamma)\geq\epsilon_n} h_n^{-\frac{d-1}{2}} e^{-\alpha\frac{\operatorname{dist}(x,\gamma_n)^2}{h_n}}\,dx = \int_{|x'|\geq\frac{\epsilon_n}{C}} h_n^{-\frac{d-1}{2}} e^{-\alpha\frac{|x'|^2}{Ch_n}}\,dx \\ = C'\int_{|y'|\geq\frac{\epsilon_n}{C^2\sqrt{h_n}}} e^{-\alpha|y'|^2}\,dy' \leq C'e^{-\alpha\frac{\epsilon_n^2}{C^4h_n}}. \tag{5-8}$$

We now choose

$$h_n = \kappa_n^{\frac{1}{d-1}}\epsilon_n^2 \to 0$$

such that

$$\frac{\epsilon_n^2}{h_n} = \kappa_n^{-\frac{1}{d-1}} \to +\infty, \quad \kappa_n\left(\frac{\epsilon_n^2}{h_n}\right)^{\frac{d-1}{2}} = \kappa_n^{\frac{1}{2}} \to 0. \tag{5-9}$$

This choice implies

$$\int_0^T\int_M a(x)|\partial_t w_{h_n}|^2\,dx\,dt = o(1)_{n\to+\infty},$$

which contradicts (1-3) because the energy of the initial data $(w_{h_n}, \partial_t w_{h_n})$ is constant and nonzero. This completes the proof of Theorem 5. □

Let us now come back to the proof of Proposition 5.1. This is basically done in [Ralston 1982, Section 2.1]. The idea is to define oscillating solutions (phase and symbol) by constructing the germs on the bicharacteristic curve. Let $\rho_0 = \big(t_0, \tau_0{=}\frac{1}{2}, x_0, \xi_0\big)$ be a point in the characteristic variety of the wave equation

$$\operatorname{Char} = \{(t,x,\tau,\xi) \in T^*M : |\tau|^2 = |\xi|^2 = 1\}.$$

Let

$$\Gamma = \big\{t(s), \gamma(s), \tau(s) = \tfrac{1}{2}, \xi(s))\big\}$$

be the bicharacteristic curve issued from ρ_0. For any $T < +\infty$, we can choose systems along the geodesic γ and get an immersion

$$i : (-\epsilon, T+\epsilon) \times B(0,\epsilon) \subset \mathbb{R}\times\mathbb{R}^{d-1} \to M,$$

along which the bicharacteristic takes the form

$$\gamma(s) = \big(t = s,\ x_1 = s,\ x' = 0,\ \tau = \tfrac{1}{2},\ \xi_1 = -\tfrac{1}{2},\ \xi' = 0\big),$$

which allows us to reduce the analysis to $\mathbb{R}^d$. In this coordinate system, (5-1) takes the form

$$\Delta = \frac{1}{\kappa(x)}\sum_{i,j}\frac{\partial}{\partial x_i}g^{i,j}(x)\kappa(x)\frac{\partial}{\partial x_j}.$$

We now write $y = (t,x)$ and seek approximate solutions of the wave equation with the form

$$v_h(t,x) = e^{\frac{i}{h}\psi(t,x)}\sigma(t,x,h), \tag{5-10}$$

where $\sigma(t{=}0)=\sigma_0\in C_0^\infty(\mathbb{R}^d)$ has sufficiently small compact support near 0. Applying the operator $\partial_t^2-\Delta_x$ we get

$$\begin{aligned}(\partial_t^2-\Delta_x)v_h = &-\frac{1}{h^2}\Big((\partial_t\psi)^2-\sum_{1\le k,j\le n} g^{k,j}(x)\,\partial_{x_k}\Psi\,\partial_{x_j}\Psi\Big)\sigma e^{\frac{i}{h}\psi}\\ &+e^{\frac{i}{h}\psi}\frac{i}{h}\Big(2\partial_t\psi\,\partial_t\sigma-2\sum_{1\le k,j\le n} g^{k,j}(x)\,\partial_k\psi\,\partial_j\sigma-\frac{1}{\kappa}\sum_{1\le k,j\le n}\partial_k(g^{k,j}\kappa)(x)\sigma\,\partial_j\psi\Big)\\ &+e^{\frac{i}{h}\psi}(\partial_t^2-\Delta\psi)\sigma. \qquad (5\text{-}11)\end{aligned}$$

Ralston [1982, Section 2.1] then showed that provided

$$\psi(t(s),x(s))=t(s)-x_1(s)+cste \iff \partial_{t,x}\psi(t(s),x(s))=(\tau(s),\xi(s)),$$

and choosing

$$\operatorname{Im}\Big(\frac{\partial^2\psi}{\partial x^2}\Big)\Big|_{\gamma(0)}\ge c\mathrm{Id},\quad c>0, \qquad (5\text{-}12)$$

it is possible to solve both the eikonal equation

$$p=\Big((\partial_t\psi)^2-\sum_{1\le k,j\le n} g^{k,j}(x)\partial_{x_k}\psi\,\partial_{x_j}\psi\Big)=0$$

and the transport equation

$$T=\Big(2\partial_t\psi\,\partial_t\sigma-2\sum_{1\le k,j\le n} g^{k,j}(x)\,\partial_k\psi\,\partial_j\sigma-\frac{1}{\kappa}\sum_{1\le k,j\le n}\partial_k(g^{k,j}\kappa)(x)\sigma\,\partial_j\psi\Big)e^{\frac{i}{h}\psi}+h(\partial_t^2-\Delta\psi)\sigma=0,$$

with

$$\operatorname{Im}\Big(\frac{\partial^2\psi}{\partial x^2}\Big)\Big|_{\gamma(s)}\ge c(s)\mathrm{Id},\quad c(s)>0, \qquad (5\text{-}13)$$

to arbitrarily large order on the bicharacteristic γ by choosing

$$\sigma=\sum_p h^p\sigma_p.$$

Here by solving to arbitrarily large order, we mean that we can cancel an arbitrarily large number of (t,x) derivatives on γ.

On the torus $\mathbb{T}^d$, these constructions can be performed explicitly and we get

$$\psi(t,x)=t-x_1+i((t-x_1)^2+g(t)|x'|^2)+O(|t-x_1|^3+|x'|^3), \qquad (5\text{-}14)$$

with g solving

$$2ig'(t)+4g^2(t)=0 \iff g(t)=\frac{g(0)}{1-2itg(0)},\ \ g(0):=1.$$

Notice in particular that

$$\operatorname{Re}(g(t))=\frac{1}{1+4t^2}>0, \qquad (5\text{-}15)$$

and we can choose a symbol

$$\sigma(t,x_1,x') = \frac{\sigma_0(t-x_1,x')}{(1-2it)^{\frac{d-1}{2}}} + O(h) + O(|t-x_1|+|x'|).$$

Finally, it remains to cut off the symbol such constructed near the geodesic (taking advantage of (5-13), we see that this truncation will add an exponentially small error), and to normalize by multiplying by

$$ch^{1-\frac{d}{4}}$$

to ensure the normalization of the energy in (5-2) and the error bound (5-3). We leave the details to the reader.

5B. *Assumption 1.2 and* (GGCC).

On two-dimensional tori and for dampings a satisfying (1-4) we have:

Proposition 5.2. *On a two-dimensional torus $\mathbb{T}^2$, if the damping a satisfies* (1-4), *then* (GGCC) *is equivalent to Assumption 1.2.*

Proof. Since Assumption 1.2 implies uniform stabilization (Theorem 4) which in turn implies (GGCC) (Theorem 5), it is enough to show that (GGCC) implies Assumption 1.2.

Let us assume (GGCC). If Assumption 1.2 were not satisfied, then there would, for any $T>0$, exist a geodesic curve γ of length T which either does not encounter $\mathcal{R}=\bigcup_{j=1}^{N}\bar{R}_j$, or does encounter $\mathcal{R}$ only at corners or only on the left (or only on the right). In the first case, by compactness, the geodesic curve remains at distance ϵ_0 of $\mathcal{R}$, and consequently for $0<\epsilon<\epsilon_0$ we have

$$\int_{\Gamma_{\rho_0,\epsilon,T}} a(x)\,dx = 0.$$

In the second case (see checkerboard in Figure 1(b)), by compactness, the geodesic curve encounters only a finite number of corners, and consequently ($d=2$)

$$\int_{\Gamma_{\rho_\sigma,\epsilon,T}} a(x)\,dx = \mathcal{O}(\epsilon^2),$$

with a constant $c>0$ depending on the angles of the corners, while

$$\operatorname{Vol}(\Gamma_{\rho_0,\epsilon,T}) \sim C\epsilon,$$

which implies that (GGCC) does not hold. In the last case (see the right checkerboard in Figure 1(c) and Figure 4), let us consider the family of geodesics $\gamma_\sigma=\{\gamma_{\rho_\sigma}(s):s\in(0,T)\}$, $\sigma\in[0,1)$, parallel on the right to $\gamma_0=\{\gamma_{\rho_0}(s):s\in(0,T)\}$ (i.e., if $\rho_0=(X_0,\Xi_0)$, then $\rho_\sigma=X_0+\sigma\Xi_0^\perp$, where $\Xi_0^\perp$ is the unit vector orthogonal to Ξ_0, pointing on the right of γ_0). Since on the right γ_0 encounters no side of any rectangle R_j, it may encounter only (finitely many) corner points. As a consequence, for any $\sigma>0$

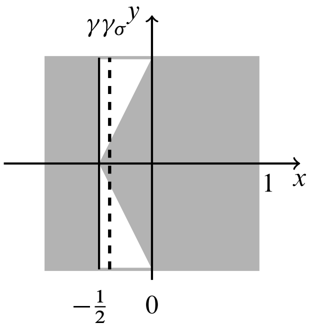


Figure 4. Proof of Proposition 5.2, last case.

sufficiently small, and $0 < \epsilon \ll \sigma$,

$$\int_{\Gamma_{\rho_\sigma,\epsilon,T}} a(x)\,dx \sim c\sigma\epsilon \quad (\epsilon \to 0),$$
$$\mathrm{Vol}(\Gamma_{\rho_\sigma,\epsilon,T}) \sim C\epsilon \quad (\epsilon \to 0).$$

We deduce that

$$\lim_{\epsilon\to 0} \frac{1}{\mathrm{Vol}}(\Gamma_{\rho_\sigma,\epsilon,T}) \int_{\Gamma_{\rho_\sigma,\epsilon,T}} a(x)\,dx = c\sigma,$$

and letting $\sigma \to 0$ shows that (GGCC) does not hold. □

Appendix A: Resolvent estimates and stabilization

In this appendix, we collect a few classical results on resolvent estimates.

A1. ***Resolvent estimates and stabilization.*** It is classical [Gearhart 1978] that stabilization or observability of a self-adjoint evolution system is equivalent to resolvent estimates; see also [Burq and Zworski 2004; Miller 2012; Anantharaman and Léautaud 2014]. For completeness we shall give below a proof (only the fact that resolvent estimates imply stabilization).

Proposition A.1. *Consider a strongly continuous semigroup e^{tA} on a Hilbert space H, with infinitesimal generator A defined on $D(A)$. The following two properties are equivalent*:

(1) *There exist $C, \delta > 0$ such that the resolvent of A, $(A-\lambda)^{-1}$ exists for* $\mathrm{Re}\,\lambda \geq -\delta$ *and satisfies*

$$\exists C > 0 \quad \textit{such that} \quad \forall \lambda \in \mathbb{C}^\delta = \{z \in \mathbb{C} : \mathrm{Re}\, z \geq -\delta\}, \quad \|(A-\lambda)^{-1}\|_{\mathcal{L}(H)} \leq C.$$

(2) *There exist $M, \delta > 0$ such that for any $t > 0$*

$$\|e^{tA}\|_{\mathcal{L}(H)} \leq Me^{-\delta t}.$$

Proof. Let us first prove that (2) implies (1). We start with the resolvent equality (always true for $\mathrm{Re}\,\lambda \gg 0$)

$$(A-\lambda)^{-1} f = -\int_0^{+\infty} e^{t(A-\lambda)} f\,dt,$$

and we deduce that if $\|e^{tA}\| \leq Ce^{-\beta t}$, then (1) is satisfied for any $\delta < \beta$. To prove that (1) implies (2), for $u_0 \in D(A)$, and $\chi \in C^\infty(\mathbb{R})$ equal to 0 for $t \leq -1$ and to 1 for $t \geq 0$, consider

$$u(t) = \chi(t)e^{t(A-\omega)}u_0.$$

For ω large enough, u belongs to $L^\infty(\mathbb{R}; H)$ because strongly continuous semigroups of operators satisfy

$$\exists C, c > 0 \quad \text{such that} \quad \forall t > 0, \quad \|e^{tA}\| \leq Ce^{ct},$$

and u satisfies

$$(\partial_t + \omega - A)u(t) = \chi'(t)e^{t(A-\omega)}u_0 =: v(t).$$

Taking Fourier transforms in the time variable, we get

$$(i\tau + \omega - A)\hat{u}(\tau) = \hat{v}(\tau). \tag{A-1}$$

Since $v(t)$ is supported in $t \in [-1, 0]$, the right-hand side in (A-1) is holomorphic and bounded in any domain

$$\mathbb{C}_\alpha = \{\tau \in \mathbb{C} : \operatorname{Im}\tau \geq \alpha,\ \alpha \in \mathbb{R}\}.$$

From the assumption on the resolvent, we deduce that $\hat{u}$ admits a holomorphic extension to $\{\tau : \operatorname{Im}\tau \leq \delta + \omega\}$ which satisfies

$$\|\hat{u}(\tau)\|_H \leq C\|\hat{v}(\tau)\|_H.$$

We deduce that

$$\begin{aligned} \|e^{(\omega+\delta)t}u\|_{L^2(\mathbb{R}_t;H))} &= \|\hat{u}(\tau + i(\omega+\delta))\|_{L^2(\mathbb{R}_\tau;H)} \leq C\|\hat{v}(\tau + i(\omega+\delta))\|_{L^2(\mathbb{R}_\tau;H)} \\ &\leq C\|e^{(\omega+\delta)t}v\|_{L^2(\mathbb{R}_t;H)} \leq C'\|u_0\|_H. \end{aligned} \tag{A-2}$$

This implies exponential decay of $e^{tA}u_0$ in the L^2_t norm, with the weight $e^{\delta t}$. Now consider $w(t) := \chi(t-T)e^{tA}u_0$, which satisfies

$$(\partial_t - A)w = \chi'(t-T)e^{tA}u_0, \quad w|_{t=T-1} = 0.$$

From Duhamel formula, we deduce

$$w(T) = \int_{T-1}^{T} e^{(T-s)A}\chi'(t-T)e^{sA}u_0\,ds,$$

and consequently (recall that the semigroup norm is locally bounded in time)

$$\begin{aligned} \|w(T)\|_H &\leq \int_{T-1}^{T} \|e^{(T-s)A}\chi'(t-T)e^{sA}u_0\|_H\,ds \\ &\leq C \sup_{\sigma\in[0,1]} \|e^{\sigma A}\|_{\mathcal{L}(H)} \int_{T-1}^{T} \|e^{sA}u_0\|_H\,ds \leq C'e^{-\delta T}\|e^{\delta s}e^{sA}u_0\|_{L^2(T-1,T);H} \\ &\leq C''e^{-\delta T}\|u_0\|_H. \end{aligned}$$

□

A2. ***Semigroups for damped wave equations.*** The solution to (1-1) is given very classically by

$$\begin{pmatrix} u \\ \partial_t u \end{pmatrix} = e^{tA}\begin{pmatrix} u_0 \\ u_1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & \mathrm{Id} \\ \Delta - m & -a \end{pmatrix}, \tag{A-3}$$

where A is defined on $\mathcal{H} = H^1(M) \times L^2(M)$ with domain $H^2(M) \times H^1(M)$. When $m > 0$, since

$$E(u) = \|u\|_{H^1}^2 + \|\partial_t u\|_{L^2}^2,$$

to study the decay of the energy, we can apply directly the characterization given by Proposition A.1. When $m = 0$, the semigroup e^{tA} is not a contraction semigroup on $H^1 \times L^2$ (because the energy (1-2) does not control the H^1 norm). The main difference from the case $m = 0$ and m nontrivial comes from:

Lemma A.2. *Assume that* $0 \le m \in L^\infty(M)$ *and* m *is not trivial* $\big(\int_M m(x)\,dx > 0\big)$. *Then the norms*

$$\|u\|_{H^1} = (\|\nabla_x u\|_{L^2}^2 + \|u\|_{L^2}^2)^{\frac{1}{2}}, \quad \|u\| = \sqrt{E(u)} = (\|\nabla_x u\|_{L^2}^2 + \|m^{\frac{1}{2}} u\|_{L^2}^2)^{\frac{1}{2}}$$

are equivalent.

Indeed, as for a classical proof of Poincaré's inequality, we proceed by contradiction to prove the only nontrivial inequality ($\|u\|_{H^1} \le C\|u\|$), and get a sequence $(u_n) \in H^1(M)$ such that

$$\|u_n\|_{H^1} = 1, \quad \|u_n\| \to_{n\to+\infty} 0.$$

By the weak compactness of the unit ball in H^1 we can extract a subsequence (still denoted by (u_n)) which converges weakly in H^1, and hence because M is compact strongly in L^2, to a limit $u \in H^1$. Since $\|u_n\| \to +\infty$ we get that the sequence actually converges strongly in H^1 and

$$\|u\| = 0 \to \nabla_x u = 0, \quad m^{\frac{1}{2}} u = 0.$$

We deduce that u is constant in M and since $\int_M mu = 0$, we finally get $u = 0$, which contradicts the fact that $\|u_n\|_{H^1} = 1$ (and the strong convergence of u_n to 0).

For $s = 1, 2$, we define $\dot{H}^s = H^s(M)/\mathbb{R}$ to be the quotient space of $H^s(M)$ by the constant functions, endowed with the norm

$$\|\dot{u}\|_{\dot{H}^1} = \|\nabla u\|_{L^2}, \quad \|\dot{u}\|_{\dot{H}^2} = \|\Delta u\|_{L^2}.$$

We define the operator

$$\dot{A} = \begin{pmatrix} 0 & \Pi \\ \dot{\Delta} & -a \end{pmatrix}$$

on $\dot{H}^1 \times L^2$ with domain $\dot{H}^2 \times H^1$, where Π is the canonical projection $H^1 \to \dot{H}^1$ and $\dot{\Delta}$ is defined by

$$\dot{\Delta}\dot{u} = \Delta u$$

(independent of the choice of $u \in \dot{u}$). The operator $\dot{A}$ is maximal dissipative and hence defines a semigroup of contractions on $\dot{\mathcal{H}} = \dot{H}^1 \times L^2$. Indeed for $U = \binom{\dot{u}}{v}$,

$$\mathrm{Re}(\dot{A}U, U)_{\dot{\mathcal{H}}} = \mathrm{Re}(\nabla u, \nabla v)_{L^2} + (\Delta u - av, n)_{L^2} = -(av, v)_{L^2},$$

and

$$(\dot{A}-\mathrm{Id})\begin{pmatrix}\dot{u}\\ v\end{pmatrix}=\begin{pmatrix}\dot{f}\\ g\end{pmatrix} \quad\Longleftrightarrow\quad \Pi v-\dot{u}=\dot{f},\quad \dot{\Delta}\dot{u}-(a+1)v=g$$
$$\Longleftrightarrow\quad \Pi v-\dot{u}=\dot{f},\quad \Delta v-(1+a)v=g+\Delta f\in H^{-1}(M), \tag{A-4}$$

and we solve this equation by variational theory. Notice that this shows that the resolvent $(\dot{A}-\mathrm{Id})^{-1}$ is well-defined and continuous from $\dot{H}^1\times L^2$ to $\dot{H}^2\times H^1$.

Lemma A.3. *The injection $\dot{H}^2\times H^1$ to $\dot{H}^1\times L^2$ is compact.*

This follows from identifying $\dot{H}^n$ with the kernel of the linear form $u\mapsto\int_M u$.

Corollary A.4. *The operator $(\dot{A}-\mathrm{Id})^{-1}$ is compact on $\dot{\mathcal{H}}$.*

On the other hand, it is very easy to show that for $(u_0,u_1)\in H^1\times L^2$

$$\begin{pmatrix}\Pi & 0\\ 0 & \mathrm{Id}\end{pmatrix}e^{tA}=e^{t\dot{A}}\begin{pmatrix}\Pi & 0\\ 0 & \mathrm{Id}\end{pmatrix},$$

and consequently, stabilization is equivalent to the exponential decay (in norm) of $e^{t\dot{A}}$ (and consequently, according to Proposition A.1 equivalent to resolvent estimates for $\dot{A}$).

A3. *Reduction to high-frequency observation estimates.* In this section, we show that for $m\geq 0$, stabilization is equivalent to semiclassical observation estimates; see [Miller 2012].

Proposition A.5. *Assume that $0\leq a\in L^\infty$ is nontrivial $\left(\int_M a>0\right)$. Then stabilization holds for (1-1) if and only if*

$$\exists h_0>0\quad \textit{such that}\quad \forall\, 0<h<h_0,\quad \forall\,(u,f)\in H^2(M)\times L^2(M),\quad (h^2\Delta+1)u=f,$$
$$\|u\|_{L^2(M)}\leq C\Big(\|a^{\frac{1}{2}}u\|_{L^2}+\frac{1}{h}\|f\|_{L^2}\Big). \tag{A-5}$$

We prove the proposition for $m=0$. The proof for $m\not\equiv 0$ is similar (slightly simpler as we do not have to work with the operator $\dot{A}$ but can stick with A). From Proposition A.1, stabilization is equivalent to the fact that the resolvent $(\dot{A}-\lambda)^{-1}$ is bounded on $\mathbb{C}^\delta$. Since $\dot{A}$ is maximal dissipative, its resolvent is defined (and bounded) on any domain $\mathbb{C}^{-\epsilon}$ $(\epsilon>0)$. We deduce that it is equivalent to prove that it is uniformly bounded on $i\mathbb{R}$ (and consequently by perturbation, on a δ neighborhood of $i\mathbb{R}$). Since

$$(\dot{A}-\lambda)=(1+(1-\lambda)(\dot{A}-1)^{-1})(\dot{A}-1),$$

and $(\dot{A}-1)^{-1}$ is compact on $\dot{\mathcal{H}}$ (see Corollary A.4), the operator $(1+(1-\lambda)(\dot{A}-1)^{-1})$ is Fredholm with index 0 and consequently, $\dot{A}-\lambda$ is invertible if and only if it is injective. As a consequence, stabilization is equivalent to the following a priori estimates

$$\exists C>0\quad \text{such that}\quad \forall\lambda\in\mathbb{R},\ U\in\dot{H}^2\times H^1,\ F\in\dot{H}^1\times L^2,$$
$$(\dot{A}-i\lambda)U=F\quad\Longrightarrow\quad \|U\|_{\dot{\mathcal{H}}}\leq C\|F\|_{\dot{\mathcal{H}}}. \tag{A-6}$$

A3.1. *High-frequency resolvent estimates imply stabilization.* We argue by contradiction. We assume (A-5) holds and assume that (A-6) does not hold. Then there exist sequences (λ_n), (U_n), (F_n) such that

$$(\dot{A} - i\lambda_n)U_n = F_n, \quad \|U_n\|_{\dot{\mathcal{H}}} > n\|F_n\|_{\dot{\mathcal{H}}}.$$

Since $U_n \neq 0$, we can assume $\|U_n\|_{\dot{\mathcal{H}}} = 1$. Extracting subsequences we can also assume that $\lambda_n \to \lambda \in \mathbb{R} \cup \{\pm\infty\}$ as $n \to \infty$. We write

$$U_n = \begin{pmatrix} \dot{u}_n \\ v_n \end{pmatrix}, \quad F_n = \begin{pmatrix} \dot{f}_n \\ g_n \end{pmatrix},$$

and distinguish according to three cases:

Zero frequency: $\lambda = 0$. In this case, we have

$$\dot{A}U_n = o(1)_{\dot{\mathcal{H}}} \quad \Longleftrightarrow \quad \Pi v_n = o(1)_{\dot{H}^1}, \quad \Delta\dot{u}_n - av_n = o(1)_{L^2}.$$

We deduce that there exists $c_n \in \mathbb{C}$ such that

$$v_n - c_n = o(1)_{H^1}, \quad \Delta u_n - ac_n = o(1)_{L^2}.$$

But

$$\int_M \Delta u_n = 0 \quad \Longrightarrow \quad c_n \int_M a = o(1) \quad \Longrightarrow \quad c_n = o(1).$$

As a consequence, we get $v_n = o(1)_{L^2}$ and $\Delta u_n = o(1)_{L^2} \Rightarrow \dot{u}_n = o(1)_{\dot{H}^1}$. This contradicts $\|U_n\|_{\dot{\mathcal{H}}} = 1$.

Low frequency: $\lambda \in \mathbb{R}^*$. In this case, we have

$$(\dot{A} - i\lambda)U_n = o(1)_{\dot{\mathcal{H}}} \quad \Longleftrightarrow \quad \Pi v_n - i\lambda\dot{u}_n = o(1)_{\dot{H}^1}, \quad \Delta\dot{u}_n - (i\lambda + a)v_n = o(1)_{L^2}.$$

We deduce

$$\Delta v_n - i\lambda(a + i\lambda)v_n = o(1)_{L^2} + \Delta(o(1)_{\dot{H}^1}) = o(1)_{H^{-1}}.$$

Since (v_n) is bounded in L^2, from this equation, we deduce that Δv_n is bounded in H^{-1} and consequently v_n is bounded in H^1. Extracting another subsequence, we can assume that v_n converges in L^2 to v which satisfies

$$\Delta v + \lambda^2 v - i\lambda av = 0.$$

Taking the imaginary part of the scalar product with v in L^2 gives (since $\lambda \neq 0$) $\int_M a|v|^2 = 0$, and consequently $av = 0$, which implies that v is an eigenfunction of the Laplace operator. But since the zero set of nontrivial eigenfunctions has Lebesgue measure 0 in M, $av = 0$ implies that $v = 0$ (and consequently $v_n = o(1)_{L^1}$). Now, we have

$$\Delta\dot{u}_n = (i\lambda + a)v_n + o(1)_{L^2} = o(1)_{L^2} \quad \Longrightarrow \quad \dot{u}_n = o(1)_{\dot{H}^1}.$$

This contradicts $\|U_n\|_{\dot{\mathcal{H}}} = 1$.

High frequency: $\lambda_n \to \pm\infty$. We study the case $\lambda_n \to +\infty$; the other case is obtained by considering $\overline{U}_n$. Let $h_n = \lambda_n^{-1}$. Then

$$\begin{aligned}
&(\dot{A} - i\lambda_n)U_n = o(1)_{\dot{\mathcal{H}}} \\
&\qquad\iff\quad -i\lambda_n \dot{u}_n + \Pi v_n = o(1)_{\dot{H}^1}, \quad \Delta \dot{u}_n - (i\lambda_n + a)v_n = o(1)_{L^2} \\
&\qquad\iff\quad \dot{u}_n = -ih_n \Pi v_n + o(h_n)_{\dot{H}^1}, \quad (h_n^2\Delta + 1 - ih_n a)v_n = o(h_n)_{L^2} + o(h_n^2)_{H^{-1}}. \qquad \text{(A-7)}
\end{aligned}$$

To conclude in this regime, we need:

Lemma A.6. *The observation inequality* (A-5) *implies the more general*

$$\begin{gathered}
\exists h_0 > 0 \quad \textit{such that} \quad \forall\, 0<h<h_0, \quad \forall (u, f_1, f_2) \in H^2(M)\times L^2(M)\times H^{-1}(M), \quad (h^2\Delta+1)u = f_1 + f_2, \\
\|h\nabla_x u\|_{L^2} + \|u\|_{L^2(M)} \le C\left(\|a^{\frac12}u\|_{L^2} + \frac{1}{h}\|f_1\|_{L^2} + \frac{1}{h^2}\|f_2\|_{H^{-1}}\right). \qquad \text{(A-8)}
\end{gathered}$$

Proof. Let $P_h^{\pm} = h^2\Delta + 1 \pm iha$ be defined on L^2 with domain H^2. Writing

$$P_h^{\pm} = (1 + (2 \pm iha)(h^2\Delta - 1)^{-1})(h^2\Delta - 1),$$

and since $(h^2\Delta - 1)^{-1}$ is compact on L^2, we deduce that $(1 + (2 \pm iha)(h^2\Delta - 1)^{-1})$ is Fredholm with index 0; hence $P_h^{\pm}$ is invertible if and only if it is injective. On the other hand we have

$$h\|a^{\frac12}u\|_{L^2}^2 = \pm\operatorname{Im}(P_h^{\pm}u, u)_{L^2} \le \|P_h^{\pm}u\|_{L^2}\|u\|_{L^2},$$

which combined with (A-5) implies $((h^2\Delta + 1)u = P_h^{\pm}u \mp ihau)$

$$\begin{aligned}
\|u\|_{L^2}^2 &\le C\left(\|a^{\frac12}u\|_{L^2}^2 + \frac{1}{h^2}(\|P_h^{\pm}u\|_{L^2}^2 + h^2\|au\|_{L^2}^2)\right) \\
&\le \frac{C'}{h}\|P_h^{\pm}u\|_{L^2}\|u\|_{L^2} + \frac{C'}{h^2}\|P_h^{\pm}u\|_{L^2}^2 \quad\Longrightarrow\quad \|u\|_{L^2} \le \frac{C''}{h}\|P_h^{\pm}u\|_{L^2}. \qquad \text{(A-9)}
\end{aligned}$$

Since

$$\big|\|u\|_{L^2}^2 - \|h\nabla_x u\|_{L^2}^2\big| = |\operatorname{Re}(P_h^{\pm}u, u)_{L^2}| \le \|P_h^{\pm}u\|_{L^2}\|u\|_{L^2}, \qquad \text{(A-10)}$$

we deduce that $P_h^{\pm}$ is injective and hence bijective from H^2 to L^2 with inverse bounded by C''/h from L^2 to L^2 and by C/h^2 from L^2 to H^1. We now proceed by duality to obtain (A-8). The adjoint of $P_h^{\pm}$ is $P_h^{\mp}$ and is consequently bounded from H^{-1} to L^2 by C/h^2. Using again the identity (A-10) we get

$$P_h^{\pm}u = f_1 + f_2 \quad\Longrightarrow\quad \|h\nabla_x u\|_{L^2} + \|u\|_{L^2(M)} \le \frac{C}{h}\|f_1\|_{L^2} + \frac{C}{h^2}\|f_2\|_{H^{-1}}.$$

Finally

$$(h^2\Delta + 1)u = f_1 + f_2 \quad\Longrightarrow\quad P_h^{+}u = iahu + f_1 + f_2,$$

and we get

$$\begin{aligned}
\|h\nabla_x u\|_{L^2} + \|u\|_{L^2(M)} &\le C\left(\frac{1}{h}\|ihau + f_1\|_{L^2} + \frac{1}{h^2}\|f_2\|_{H^{-1}}\right) \\
&\le C'\left(\|a^{\frac12}u\|_{L^2} + \frac{1}{h}\|f_1\|_{L^2} + \frac{1}{h^2}\|f_2\|_{H^{-1}}\right). \qquad \square
\end{aligned}$$

We now come back to our sequence satisfying (A-7). From (A-8), (A-7) implies

$$\|h_n \nabla_x v_n\|_{L^2} + \|v_n\|_{L^2} = o(1)_{n\to+\infty},$$

and in turn

$$\|\nabla_x u_n\|_{L^2} = o(1)_{n\to+\infty}.$$

This contradicts $\|U_n\|_{\dot{\mathcal{H}}} = 1$.

A3.2. *Stabilization implies resolvent estimates.* Consider now $U = \binom{\dot u}{v}$, $F = \binom{\dot f}{g}$ such that

$$(\dot A - i\lambda)U = F \quad\Longleftrightarrow\quad -i\lambda\dot u + \Pi v = \dot f \quad\text{and}\quad (\Delta v + \lambda^2 - i\lambda a)v = i\lambda g + \Delta f.$$

From (A-5) with $h = \lambda^{-1}$, we get

$$\|v\|_{L^2} + \|h\nabla_x v\|_{L^2} \le C\|g\|_{L^2} + C\|\Delta f\|_{H^{-1}} \le C(\|g\|_{L^2} + C\|\nabla_x f\|_{L^2}),$$

and also

$$\|\nabla_x u\|_{L^2} = h\|\nabla_x(v-f)\| \le C(\|g\|_{L^2} + C\|\nabla_x f\|_{L^2}).$$

Appendix B: Characterization of stabilization

Here we shall prove that the properties (1), (2), (3) and (4) of the Introduction are equivalent. The implication (2) ⇒ (1) is trivial. To show (1) ⇒ (3) we fix T such that $f(T) \le \frac{1}{2}$. Then since

$$E_m(u)(T) = E_m(u)(0) - \int_0^T\int_M a(x)|\partial_t u|^2(t,x)\,dx\,dt \le \frac{1}{2}E_m(u)(0),$$

we deduce

$$E_m(u)(0) \le 2\int_0^T\int_M a(x)|\partial_t u|^2(t,x)\,dx\,dt,$$

which is (3). Conversely, if (3) is satisfied, we get

$$E_m(u)(T) = E_m(u)(0) - \int_0^T\int_M a(x)|\partial_t u|^2(t,x)\,dx\,dt \le \Big(1 - \frac{1}{C}\Big)E_m(u)(0).$$

Let $\delta = \big(1 - \frac{1}{C}\big) < 1$. Applying the previous estimate between 0 and T, then T and $2T$, etc., we get

$$E_m(u)(nT) \le \delta^n E_m(u)(0),$$

and hence the exponential decay along the discrete sequence of times nT. Finally, writing $nT \le t < (n+1)T$, we get

$$E_m(u)(t) \le E_m(u)(nT) \le \delta^n E_m(u)(0) \le e^{\log(\delta)(\frac{t}{T}-1)}E_m(u)(0),$$

which is (2). It remains to prove that (3) and (4) are equivalent. We shall actually prove that if (3) holds for some $T > 0$, then (4) holds for the same $T > 0$. Let us fix $T > 0$ and assume that (4) does not hold;

i.e., there exist sequences $(u_0^n, u_1^n) \in H^1 \times L^2$ such that the corresponding solutions to the undamped wave equation (1-3) satisfy

$$E_m(u_n)(0) > n \int_0^T \int_M a(x) |\partial_t u^n|^2(t,x)\, dt\, dx.$$

This implies that u_n is not identically 0 and dividing u_n by $\sqrt{E_m(u_n)(0)}$), we can assume $E_m(u_n)(0) = 1$, and

$$\int_0^T \int_M a(x) |\partial_t u_n|^2(t,x)\, dt\, dx \le \frac{1}{n}. \tag{B-1}$$

Consider now (v_n) the sequence of solutions to the damped wave equation (1-1), with the same initial data (u_0^n, u_1^n), and $w_n = u_n - v_n$ a solution to

$$(\partial_t^2 - \Delta + a\,\partial_t + m) w_n = -a\,\partial_t u_n, \quad (w_n|_{t=0}, \partial_t w_n|_{t=0}) = (0,0) \in (H^1 \times L^2)(M).$$

From Duhamel's formula and (B-1) we deduce

$$\begin{aligned} \|(w_n, \partial_t w_n)\|_{L^\infty((0,T);H^1(M)\times L^2(M))} &\le \|a\,\partial_t u_n\|_{L^1(0,T);L^2(M)} \\ &\le \|a\|_{L^\infty}^{\frac{1}{2}} \|a^{\frac{1}{2}}\,\partial_t u_n\|_{L^1(0,T);L^2(M)} = o(1)_{n\to+\infty}. \end{aligned} \tag{B-2}$$

We deduce

$$E_m(v_n)(0) = 1 + o(1)_{n\to+\infty}$$

and

$$\int_0^T \int_M a(x) |\partial_t v_n|^2(t,x)\, dt\, dx = \|a^{\frac{1}{2}} v_n\|_{L^2((0,T)\times M)} = o(1)_{n\to+\infty},$$

which implies that (3) does not hold. As a consequence, we just proved (3) ⇒ (4). The proof of (4) ⇒ (3) is similar.

Acknowledgements

Agence Nationale de la Recherche partially supported Burq and Gérard through project ANAÉ ANR-13-BS01-0010-03 and Burq through project ISDEEC ANR-16-CE40-0013. We warmly thank a referee whose comments and suggestions lead to significant improvements of this article.

References

[Anantharaman and Léautaud 2014] N. Anantharaman and M. Léautaud, "Sharp polynomial decay rates for the damped wave equation on the torus", *Anal. PDE* **7**:1 (2014), 159–214. MR Zbl

[Anantharaman and Macià 2014] N. Anantharaman and F. Macià, "Semiclassical measures for the Schrödinger equation on the torus", *J. Eur. Math. Soc.* (*JEMS*) **16**:6 (2014), 1253–1288. MR Zbl

[Babich and Popov 1981] V. M. Babich and M. M. Popov, "Propagation of concentrated sound beams in a three-dimensional inhomogeneous medium", *Akust. Zh.* **27**:6 (1981), 828–825. In Russian. MR

[Babich and Ulin 1981] V. M. Babich and V. V. Ulin, "The complex space-time ray method and 'quasiphotons'", pp. 5–12 in Математические вопросы теории распространения волн **12**, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov **117**, 1981. In Russian; translated in *J. Soviet Math.* **24**:3 (1984), 269–273. MR Zbl

[Bardos, Lebeau and Rauch 1992] C. Bardos, G. Lebeau, and J. Rauch, "Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary", *SIAM J. Control Optim.* **30**:5 (1992), 1024–1065. MR Zbl

[Blair and Sogge 2015] M. D. Blair and C. D. Sogge, "Refined and microlocal Kakeya–Nikodym bounds for eigenfunctions in two dimensions", *Anal. PDE* **8**:3 (2015), 747–764. MR Zbl

[Bony and Lerner 1989] J.-M. Bony and N. Lerner, "Quantification asymptotique et microlocalisations d'ordre supérieur, I", *Ann. Sci. École Norm. Sup.* (4) **22**:3 (1989), 377–433. MR Zbl

[Bourgain, Burq and Zworski 2013] J. Bourgain, N. Burq, and M. Zworski, "Control for Schrödinger operators on 2-tori: rough potentials", *J. Eur. Math. Soc.* (*JEMS*) **15**:5 (2013), 1597–1628. MR Zbl

[Burq 1997] N. Burq, "Mesures semi-classiques et mesures de défaut", pp. 167–195 in *Séminaire Bourbaki* 1996/1997 (exposé 826), Astérisque **245**, Société Mathématique de France, Paris, 1997. MR Zbl

[Burq 2002] N. Burq, "Semi-classical estimates for the resolvent in nontrapping geometries", *Int. Math. Res. Not.* **2002**:5 (2002), 221–241. MR Zbl

[Burq $\geq$ 2020] N. Burq, "Wave control and second-microlocalization on geodesics", in preparation.

[Burq and Gérard 1997] N. Burq and P. Gérard, "Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes", *C. R. Acad. Sci. Paris Sér. I Math.* **325**:7 (1997), 749–752. MR Zbl

[Burq and Hitrik 2007] N. Burq and M. Hitrik, "Energy decay for damped wave equations on partially rectangular domains", *Math. Res. Lett.* **14**:1 (2007), 35–47. MR Zbl

[Burq and Zuily 2015] N. Burq and C. Zuily, "Laplace eigenfunctions and damped wave equation on product manifolds", *Appl. Math. Res. Express. AMRX* **2015**:2 (2015), 296–310. MR Zbl

[Burq and Zworski 2004] N. Burq and M. Zworski, "Geometric control in the presence of a black box", *J. Amer. Math. Soc.* **17**:2 (2004), 443–471. MR Zbl

[Burq and Zworski 2005] N. Burq and M. Zworski, "Bouncing ball modes and quantum chaos", *SIAM Rev.* **47**:1 (2005), 43–49. MR Zbl

[Burq and Zworski 2012] N. Burq and M. Zworski, "Control for Schrödinger operators on tori", *Math. Res. Lett.* **19**:2 (2012), 309–324. MR Zbl

[Fermanian-Kammerer 2000] C. Fermanian-Kammerer, "Mesures semi-classiques 2-microlocales", *C. R. Acad. Sci. Paris Sér. I Math.* **331**:7 (2000), 515–518. MR Zbl

[Fermanian-Kammerer 2005] C. Fermanian Kammerer, "Analyse à deux échelles d'une suite bornée de L^2 sur une sous-variété du cotangent", *C. R. Math. Acad. Sci. Paris* **340**:4 (2005), 269–274. MR Zbl

[Fermanian-Kammerer and Gérard 2003] C. Fermanian Kammerer and P. Gérard, "A Landau–Zener formula for non-degenerated involutive codimension 3 crossings", *Ann. Henri Poincaré* **4**:3 (2003), 513–552. MR Zbl

[Fermanian-Kammerer and Gérard 2004] C. Fermanian Kammerer and P. Gerard, "A Landau–Zener formula for two-scaled Wigner measures", pp. 167–177 in *Dispersive transport equations and multiscale models* (Minneapolis, MN, 2000), edited by N. Ben Abdallah et al., IMA Vol. Math. Appl. **136**, Springer, 2004. MR Zbl

[Fermanian-Kammerer and Gérard 2002] C. Fermanian-Kammerer and P. Gérard, "Mesures semi-classiques et croisement de modes", *Bull. Soc. Math. France* **130**:1 (2002), 123–168. MR Zbl

[Gearhart 1978] L. Gearhart, "Spectral theory for contraction semigroups on Hilbert space", *Trans. Amer. Math. Soc.* **236** (1978), 385–394. MR Zbl

[Gérard 1996] P. Gérard, "Oscillations and concentration effects in semilinear dispersive wave equations", *J. Funct. Anal.* **141**:1 (1996), 60–98. MR Zbl

[Gérard and Leichtnam 1993] P. Gérard and E. Leichtnam, "Ergodic properties of eigenfunctions for the Dirichlet problem", *Duke Math. J.* **71**:2 (1993), 559–607. MR Zbl

[Hitrik 2003] M. Hitrik, "Eigenfrequencies and expansions for damped wave equations", *Methods Appl. Anal.* **10**:4 (2003), 543–564. MR Zbl

[Hörmander 1985] L. Hörmander, *The analysis of linear partial differential operators, III: Pseudodifferential operators*, Grundlehren der Mathematischen Wissenschaften **274**, Springer, 1985. MR Zbl

[Humbert, Privat and Trélat 2019] E. Humbert, Y. Privat, and E. Trélat, "Observability properties of the homogeneous wave equation on a closed manifold", *Comm. Partial Differential Equations* **44**:9 (2019), 749–772. MR Zbl

[Kashiwara and Kawai 1980] M. Kashiwara and T. Kawai, "Second-microlocalization and asymptotic expansions", pp. 21–76 in *Complex analysis, microlocal calculus and relativistic quantum theory* (*Proc. Internat. Colloq., Centre Phys., Les Houches, 1979*), Lecture Notes in Phys. **126**, Springer, 1980. MR Zbl

[Koch and Tataru 1995] H. Koch and D. Tataru, "On the spectrum of hyperbolic semigroups", *Comm. Partial Differential Equations* **20**:5-6 (1995), 901–937. MR Zbl

[Laurent 1979] Y. Laurent, "Double microlocalisation et problème de Cauchy dans le domaine complexe", exposé 11 in *Journées: équations aux dérivées partielles* (Saint-Cast, 1979), École Polytech., Palaiseau, 1979. MR Zbl

[Laurent 1985] Y. Laurent, *Théorie de la deuxième microlocalisation dans le domaine complexe*, Progress in Mathematics **53**, Birkhäuser, Boston, 1985. MR Zbl

[Léautaud and Lerner 2017] M. Léautaud and N. Lerner, "Energy decay for a locally undamped wave equation", *Ann. Fac. Sci. Toulouse Math.* (6) **26**:1 (2017), 157–205. MR Zbl

[Lebeau 1985] G. Lebeau, "Deuxième microlocalisation sur les sous-variétés isotropes", *Ann. Inst. Fourier* (*Grenoble*) **35**:2 (1985), 145–216. MR Zbl

[Lebeau 1992] G. Lebeau, "Control for hyperbolic equations", pp. 24 in *Journées "équations aux dérivées partielles"* (Saint-Jean-de-Monts, 1992), École Polytech., Palaiseau, 1992. MR Zbl

[Lebeau 1996] G. Lebeau, "Équation des ondes amorties", pp. 73–109 in *Algebraic and geometric methods in mathematical physics* (Kaciveli, 1993), edited by A. Boutet de Monvel and V. Marchenko, Math. Phys. Stud. **19**, Kluwer Acad., Dordrecht, 1996. MR Zbl

[Macià 2010] F. Macià, "High-frequency propagation for the Schrödinger equation on the torus", *J. Funct. Anal.* **258**:3 (2010), 933–955. MR Zbl

[Miao, Sogge, Xi and Yang 2016] C. Miao, C. D. Sogge, Y. Xi, and J. Yang, "Bilinear Kakeya–Nikodym averages of eigenfunctions on compact Riemannian surfaces", *J. Funct. Anal.* **271**:10 (2016), 2752–2775. MR Zbl

[Miller 1996] L. Miller, *Propagation d'ondes semi-classiques à travers une interface et mesures 2-microlocales*, Ph.D. thesis, École Polytechnique, 1996.

[Miller 1997] L. Miller, "Réfraction d'ondes semi-classiques par des interfaces franches", *C. R. Acad. Sci. Paris Sér. I Math.* **325**:4 (1997), 371–376. MR Zbl

[Miller 2000] L. Miller, "Refraction of high-frequency waves density by sharp interfaces and semiclassical measures at the boundary", *J. Math. Pures Appl.* (9) **79**:3 (2000), 227–269. MR Zbl

[Miller 2012] L. Miller, "Resolvent conditions for the control of unitary groups and their approximations", *J. Spectr. Theory* **2**:1 (2012), 1–55. MR Zbl

[Nier 1996] F. Nier, "A semi-classical picture of quantum scattering", *Ann. Sci. École Norm. Sup.* (4) **29**:2 (1996), 149–183. MR Zbl

[Ralston 1982] J. Ralston, "Gaussian beams and the propagation of singularities", pp. 206–248 in *Studies in partial differential equations*, edited by W. Littman, MAA Stud. Math. **23**, Math. Assoc. America, Washington, DC, 1982. MR Zbl

[Rauch and Taylor 1974] J. Rauch and M. Taylor, "Exponential decay of solutions to hyperbolic equations in bounded domains", *Indiana Univ. Math. J.* **24** (1974), 79–86. MR Zbl

[Rauch and Taylor 1975] J. Rauch and M. Taylor, "Decay of solutions to nondissipative hyperbolic systems on compact manifolds", *Comm. Pure Appl. Math.* **28**:4 (1975), 501–523. MR Zbl

[Sjöstrand 1982] J. Sjöstrand, "Singularités analytiques microlocales", pp. 1–166 Astérisque **95**, Soc. Math. France, Paris, 1982. MR Zbl

[Sjöstrand 2000] J. Sjöstrand, "Asymptotic distribution of eigenfrequencies for damped wave equations", *Publ. Res. Inst. Math. Sci.* **36**:5 (2000), 573–611. MR Zbl

[Sjöstrand and Zworski 1999] J. Sjöstrand and M. Zworski, "Asymptotic distribution of resonances for convex obstacles", *Acta Math.* **183**:2 (1999), 191–253. MR Zbl

[Sogge 2011] C. D. Sogge, “Kakeya–Nikodym averages and L^p-norms of eigenfunctions”, *Tohoku Math. J.* (2) **63**:4 (2011), 519–538. MR Zbl

[Trélat, Humbert and Privat 2017] E. Trélat, E. Humbert, and Y. Privat, “A sufficient condition for observability of waves by measurable subsets”, preprint, 2017, available at https://hal.archives-ouvertes.fr/hal-01652890v1.

[Zhu 2018] H. Zhu, “Stabilization of damped waves on spheres and Zoll surfaces of revolution”, *ESAIM Control Optim. Calc. Var.* **24**:4 (2018), 1759–1788. MR Zbl

[Zworski 2012] M. Zworski, *Semiclassical analysis*, Graduate Studies in Mathematics **138**, American Mathematical Society, Providence, RI, 2012. MR Zbl

Received 9 Dec 2019. Revised 2 Feb 2020. Accepted 15 Mar 2020.

NICOLAS BURQ: nicolas.burq@math.u-psud.fr
Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, Orsay, France CNRS, Institut Universitaire de France,

PATRICK GÉRARD: patrick.gerard@math.u-psud.fr
Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, Orsay, France, CNRS, UMR 8628

PURE and APPLIED
ANALYSIS

Vol. 2, No. 3, 2020
https://doi.org/10.2140/paa.2020.2.659

ANALYSIS OF A SIMPLE EQUATION FOR THE GROUND STATE ENERGY OF THE BOSE GAS

ERIC A. CARLEN, IAN JAUSLIN AND ELLIOTT H. LIEB

In 1963 a partial differential equation with a convolution nonlinearity was introduced in connection with a quantum mechanical many-body problem, namely the gas of bosonic particles. This equation is mathematically interesting for several reasons. Although the equation was expected to be valid only for small values of the parameters, further investigation showed that predictions based on the equation agree well over the *entire range* of parameters with what is expected to be true for the solution of the true many-body problem. Additionally, the novel nonlinearity is easy to state but seems to have almost no literature up to now. Finally, the earlier work did not prove existence and uniqueness of a solution, which we provide here along with properties of the solution such as decay at infinity.

1. Introduction

This paper is devoted to the study of an integrodifferential equation introduced in [Lieb 1963] in connection with the study of the Bose gas, a many-body problem in quantum mechanics. The equation is

$$(-\Delta + 4e + \mathcal{V}(x))u(x) = \mathcal{V}(x) + 2e\rho(u * u)(x), \tag{1-1}$$

with $x \in \mathbb{R}^d$ and $*$ denoting convolution: $u * u(x) := \int u(x-y)u(y)\,dy$. Here, $\mathcal{V}$ is a given function (called the *potential*) in $L^1(\mathbb{R}^d) \cap L^p(\mathbb{R}^d)$, with $p > d/2$ for $d \geqslant 2$ and $p > 1$ for $d = 1$. We assume $\mathcal{V}$ to be nonnegative. (This corresponds to a repulsive interaction between the particles in the underlying quantum system.) The two parameters e and ρ are nonnegative numbers, and they are related by a constraint, namely

$$e = \frac{\rho}{2}\int (1-u(x))\mathcal{V}(x)\,dx. \tag{1-2}$$

MSC2010: primary 35Q40; secondary 82B10.
Keywords: Bose gas, partial differential equations.

We are interested in solutions of (1-1) that satisfy the constraint (1-2), or, in other words, solutions of the system (1-1)–(1-2). We are particularly interested in the case $d = 3$, though other dimensions are also of interest. As explained in [Lieb 1963], the parameter ρ corresponds to the particle density N/V of the underlying Bose gas in the large volume and large particle number limit, and $e = E/N$ stands for the energy per particle.

One would like to fix a value ρ for the density, and then one expects, on the basis of the arguments in [Lieb 1963], that there will be a unique value of $e = e(\rho)$ such that there is a solution of (1-1)–(1-2) with u taking values in [0, 1]. This value of e is then the energy per particle of the Bose gas in its ground state.

The problem of determining this ground state energy per particle, as a function of the density, has attracted the attention of a great many researchers since the pioneering work [Lenz 1929]. In that paper and subsequent work [Bogolubov 1947; Lee et al. 1957], an asymptotic expansion of $e(\rho)$ for $d = 3$ and small ρ was obtained:

$$e = 2\pi\rho a\left(1 + \frac{128}{15\sqrt{\pi}}\sqrt{\rho a^3} + o(\sqrt{\rho})\right), \quad \rho \to 0, \tag{1-3}$$

where a, called the *scattering length*, is a property of the pair-interaction potential $\mathcal{V}(x)$ and is defined in (4-8)–(4-12) below. Here, we set both the mass m of the particle and Planck's constant $\hbar$ to 1. This early work was not mathematically rigorous, and it was not until [Lieb and Yngvason 1998] that the validity of the first term $2\pi\rho a$ was proved, and not until [Fournais and Solovej 2019] that the validity of second term was also proved, utilizing upper bounds proved earlier in [Dyson 1957; Yau and Yin 2009].

This timeline gives some idea of the complexity of the problem of directly studying the Bose gas ground state as a many-body problem. The complexity makes it very attractive to try to show that the system (1-1)–(1-2) provides a useful and illuminating route to the computation of the properties of the ground state for a Bose gas. Interest is piqued further by the fact that numerical studies show that the function $e(\rho)$ computed using the system (1-1)–(1-2) is surprisingly accurate for *all* densities, not only low densities, as we discuss later in this paper. Until now, however, there has been no mathematically rigorous study of this system, and even the most basic questions concerning existence and uniqueness of solutions had remained open.

In this paper, we settle some of these basic questions and raise others. It may at first appear surprising that (1-1) poses any serious mathematical challenges. After all, if one replaced the convolution nonlinearity $u * u$ in (1-1) by a power nonlinearity, say u^2, one would have a familiar sort of local elliptic equation:

$$(-\Delta + 4e + \mathcal{V}(x))u(x) = \mathcal{V}(x) + 2e\rho u^2(x). \tag{1-4}$$

However, the convolution nonlinearity in (1-1) makes it nonlocal, and very different from (1-4).

As explained in [Lieb 1963] the solutions of physical interest are integrable and *must* satisfy $u(x) \leqslant 1$ for all x. Our first result is that for integrable solutions of the system (1-1)–(1-2), the upper bound $u \leqslant 1$ implies the lower bound $u \geqslant 0$:

Theorem 1.1 (positivity). *Suppose that $\mathcal{V}$ is nonnegative and integrable and that u is an integrable solution of* (1-1)–(1-2) *such that $u(x) \leqslant 1$ for all x. Then $u(x) \geqslant 0$ for all x, and all such solutions have* *fairly slow decay at infinity in that they satisfy*

$$\int |x| u(x)\, dx = \infty. \tag{1-5}$$

Thus, any physical solutions of (1-1)–(1-2) *must necessarily satisfy the* ***pair*** *of inequalities*

$$0 \leqslant u(x) \leqslant 1 \quad \textit{for all } x. \tag{1-6}$$

This a priori result, which we prove before we take up existence and uniqueness, relies on results [Carlen et al. 2020] obtained in collaboration with Michael Loss on the convolution inequality $f \geqslant f * f$ in $L^1(\mathbb{R}^d)$. While $u(x) \leqslant 1$ is a physical requirement, $u(x) \geqslant 0$ is not; see Section 6 for details.

The converse of Theorem 1.1 also holds, as stated in the following theorem.

Theorem 1.2. *Let* $\mathcal{V} \in L^1(\mathbb{R}^d) \cap L^p(\mathbb{R}^d)$, $p > \max\{d/2, 1\}$, *be nonnegative. If* u *is an integrable solution of* (1-1)–(1-2) *such that* $u(x) \geqslant 0$ *for all* x, *then* $u(x) \leqslant 1$ *for all* x.

Remark. We have thus proved that $u \geqslant 0$ if and only if $u \leqslant 1$. This, in principle, leaves the door open to solutions that are sometimes > 1 and sometimes < 0, though we do not believe such solutions exist.

Before stating our main theorems, we make a few observations.

1A. The system (1-1)–(1-2) is actually equivalent to (1-1) and the constraint

$$\int u(x)\, dx = \frac{1}{\rho}. \tag{1-7}$$

To prove this, consider the operator

$$G_e := [-\Delta + 4e]^{-1}, \tag{1-8}$$

which is given by

$$G_e f = Y_{4e} * f, \tag{1-9}$$

where Y_{4e} is the *Yukawa potential* [Lieb and Loss 2001, Section 6.23], which is nonnegative and satisfies $\int Y_{4e}\, dx = (4e)^{-1}$. When $d = 3$,

$$Y_{4e}(x) = \frac{e^{-2\sqrt{e}|x|}}{4\pi |x|}. \tag{1-10}$$

Equation (1-1) can be rewritten as

$$u(x) = Y_{4e} * (\mathcal{V}(1 - u(x))) + 2e\rho Y_{4e} * u * u. \tag{1-11}$$

Since u and $\mathcal{V}$ are assumed to be integrable, and $u(x)$ is assumed to satisfy (1-6), all terms in (1-11) are integrable, and integrating yields

$$\int u(x)\, dx = \frac{1}{4e} \int \mathcal{V}(x)(1 - u(x))\, dx + \frac{\rho}{2}\left(\int u(x)\, dx\right)^2. \tag{1-12}$$

Thus, for integrable solutions u of (1-1) satisfying (1-6), the constraint (1-2) is equivalent to (1-7).

1B. There is another useful way to write the system (1-1)–(1-2). The damped heat semigroup $e^{-t(-\Delta+4e)}$ is a strongly continuous contraction semigroup on $L^p(\mathbb{R}^d)$, and the domain of its generator is $\mathcal{D}(-\Delta+4e) = W^{2,p}(\mathbb{R}^d)$. By the Sobolev embedding theorem [Lieb and Loss 2001, Theorem 10.2], since $p > d/2$, all functions $f \in \mathcal{D}(-\Delta+4e)$ are continuous and vanish at infinity. Since $\mathcal{V} \geqslant 0$, we know $e^{-t\mathcal{V}}$ is also a strongly continuous contraction semigroup on $L^p(\mathbb{R}^d)$, and since $\mathcal{V} \in L^p(\mathbb{R}^d)$, the domain of its generator, $\mathcal{D}(\mathcal{V})$, contains all bounded functions, and in particular $W^{2,p}(\mathbb{R}^d)$. Writing $\mathcal{V}$ as the sum of a piece with a small norm in $L^p(\mathbb{R}^d)$ and another piece that is bounded, it is easy to see that there are numbers $a, b > 0$ with $a < \frac{1}{2}$ such that for all $f \in W^{2,p}(\mathbb{R}^d)$,

$$\|\mathcal{V} f\|_p \leqslant a\|(-\Delta+4e) f\|_p + b\|f\|_p. \tag{1-13}$$

Then by the Banach space version of the Kato–Rellich theorem [Reed and Simon 1975, p. 244] the operator $-\Delta+4e+\mathcal{V}(x)$ maps $W^{2,p}(\mathbb{R}^d)$ invertibly onto $L^p(\mathbb{R}^d)$. Define K_e to be the inverse operator

$$K_e := [-\Delta+4e+\mathcal{V}(x)]^{-1}. \tag{1-14}$$

By the Trotter product formula, the operator K_e has a positive kernel that we denote by $K_e(x, y)$; in particular, K_e preserves positivity. By the resolvent identity

$$K_e = G_e - G_e \mathcal{V} K_e, \tag{1-15}$$

we conclude that

$$0 \leqslant K_e(x, y) \leqslant G_e(x, y) \tag{1-16}$$

for all x, y. Thus, the operator K_e extends to a bounded operator on $L^1(\mathbb{R}^d)$ and all terms in the equation

$$u(x) = K_e \mathcal{V}(x) + 2e\rho K_e u * u(x) \tag{1-17}$$

are well-defined whenever u is integrable. Moreover, since $\mathcal{V} \in L^p(\mathbb{R}^d)$, and since $u * u \in L^p(\mathbb{R}^d)$ when u is integrable and satisfies (1-6), every integrable solution u of (1-17) that satisfies (1-6) actually belongs to $W^{2,p}(\mathbb{R}^d)$ and satisfies (1-1).

Several simple bounds follow almost immediately from this form of the equation. First of all, since the last term on the right of (1-17) is nonnegative, we have an a priori lower bound on $u(x)$, namely

$$u(x) \geqslant u_1(x) := K_e \mathcal{V}(x). \tag{1-18}$$

Integrating both sides of (1-18), and using (1-7) yields an upper bound on ρ depending only on e, namely, $\rho \leqslant \left(\int K_e \mathcal{V}(x)\, dx\right)^{-1}$. By (1-2) and (1-18),

$$\rho = 2e\left(\int \mathcal{V}(1-u)(x)\, dx\right)^{-1} \geqslant 2e\left(\int \mathcal{V}(1-K_e\mathcal{V})(x)\, dx\right)^{-1}. \tag{1-19}$$

Altogether,

$$2e\left(\int \mathcal{V}(1-K_e\mathcal{V})(x)\, dx\right)^{-1} \leqslant \rho \leqslant \left(\int K_e \mathcal{V}(x)\, dx\right)^{-1}. \tag{1-20}$$

In fact, the left side of (1-20) is equal to half of the right side. To see this observe that $u_1 = K_d\mathcal{V}$ satisfies $(-\Delta + 4e + \mathcal{V})u_1 = \mathcal{V}$, and hence $u_1 = G_e(\mathcal{V}(1-u_1))$. Integrating both sides yields $\int u_1\,dx = (1/(4e))\int \mathcal{V}(1-u_1)\,dx$. By (1-18), we obtain the simpler (albeit less sharp) bounds

$$2e\left(\int \mathcal{V}\,dx\right)^{-1} \leqslant \rho \leqslant 4e\left(\int \mathcal{V}\,dx\right)^{-1}, \tag{1-21}$$

or equivalently

$$\left(\frac{1}{4}\int \mathcal{V}\,dx\right)\rho \leqslant e \leqslant \left(\frac{1}{2}\int \mathcal{V}\,dx\right)\rho. \tag{1-22}$$

In particular, this shows that the system (1-1)–(1-2) does not have a solution for arbitrary values of ρ and e: when either is small, a solution of the type we seek can only exist if the other is correspondingly small, as specified by (1-21) and (1-22). In fact, as is stated in the following theorem, ρ and e are constrained to be related by a functional equation.

Theorem 1.3 (existence and uniqueness). *Let $\mathcal{V} \in L^1(\mathbb{R}^d) \cap L^p(\mathbb{R}^d)$, $p > \max\{d/2, 1\}$, be nonnegative. Then there is a constructively defined continuous function $\rho(e)$ on $(0,\infty)$ such that $\lim_{e\to 0}\rho(e) = 0$ and $\lim_{e\to\infty}\rho(e) = \infty$ and such that for any $e \geqslant 0$ and $\rho = \rho(e)$, the system (1-1) and (1-2) has a unique integrable solution $u(x)$ satisfying $u(x) \leqslant 1$. Moreover, if $\rho \neq \rho(e)$, the system (1-1) and (1-2) has **no** integrable solution $u(x)$ satisfying (1-6).*

Remarks. • We do not assume here that the potential is radially symmetric. However, the uniqueness statement implies that u is radially symmetric whenever $\mathcal{V}$ is radially symmetric.

• The function $\rho(e)$ is the *density function*, which specifies the density as a function of the energy. Thus, our system together with (1-6) constrains the parameters e and ρ to be related by a strict functional relation $\rho = \rho(e)$. In most of the early literature on the Bose gas, ρ is taken as the independent parameter, as suggested by (1-3): One puts N particles in a box of volume N/ρ, and seeks to find the ground state energy per particle, e, as a function of ρ. Our theorem goes in the other direction, with ρ specified as a function of e. We prove that $e \mapsto \rho(e)$ is continuous, and we conjecture that $\rho(e)$ is a strictly monotone increasing function. In that case, the functional relation could be inverted, and we would have a well-defined function $e(\rho)$.

• Since $\lim_{e\to 0}\rho(e) = 0$ and $\lim_{e\to\infty}\rho(e) = \infty$, the continuity of $e \to \rho(e)$ implies that for each $\rho \in (0,\infty)$ there is *at least one* e such that $\rho(e) = \rho$.

Having proved that the solution to the simple equation is unique, our second main result is an asymptotic expression for $e(\rho)$, both for low and for high density.

Theorem 1.4 (asymptotics of the energy for $d = 3$). *Consider the case $d = 3$. Let $\mathcal{V}$ be nonnegative, integrable and square-integrable. Then, for each $\rho > 0$ there is at least one $e > 0$ such that $\rho = \rho(e)$. For any such ρ and e we have the following bounds for low and high density (i.e., small and large ρ). For low density,*

$$e = 2\pi\rho a\left(1 + \frac{128}{15\sqrt{\pi}}\sqrt{\rho a^3} + o(\sqrt{\rho})\right), \tag{1-23}$$

where a is the scattering length of the potential, which is defined in (4-11). For high density, in any dimension $d \geqslant 1$,

$$e = \frac{\rho}{2} \int \mathcal{V}(x)\,dx + o(\rho). \tag{1-24}$$

Remark. For low densities in $d = 3$, the energy e predicted by the simple equation (1-1)–(1-2) is asymptotically equal to the ground state energy of the Bose gas [Lee et al. 1957; Yau and Yin 2009; Fournais and Solovej 2019]. For high densities, when the potential has a nonnegative Fourier transform, the asymptotic formula for the ground state energy of the Bose gas coincides with (1-24) [Lieb 1963, Appendix]. Thus, the simple equation yields the same asymptotes for both low and high densities as the Bose gas does (at least when the potential has a nonnegative Fourier transform, as in the example $\mathcal{V}(x) = e^{-|x|}$ discussed in Section 6B).

Theorem 1.5 (decay of u at infinity). *In all dimensions, provided $\mathcal{V}$ is spherically symmetric with $\int |x|^2 \mathcal{V}\,dx < \infty$ in addition to satisfying the hypotheses imposed in Theorem 1.3, all integrable solutions of (1-1)–(1-2) with $u(x) \leqslant 1$ for all x satisfy*

$$\int |x|u(x)\,dx = \infty \quad \textit{and} \quad \int |x|^r u(x)\,dx < \infty \quad \textit{for all } 0 < r < 1. \tag{1-25}$$

Thus, if $u(x) \sim |x|^{-m}$ for some m, the only possibility is $m = d + 1$. Under stronger assumptions on the potential, this is actually the case. For $d = 3$, if $\mathcal{V}$ is nonnegative, square-integrable, spherically symmetric (that is, $\mathcal{V}(x) = \mathcal{V}(|x|)$), and, for $|x| > R$,

$$\mathcal{V}(|x|) \leqslant Ae^{-B|x|} \tag{1-26}$$

for some $A, B > 0$, then there exists $\alpha > 0$ such that

$$u(x) \underset{|x|\to\infty}{\sim} \frac{\alpha}{|x|^4}. \tag{1-27}$$

Remarks. • This result is consistent with a prediction in [Lee et al. 1957] that the truncated 2-point correlation function in the ground state of the Bose gas decays like $|x|^{-4}$.

• To prove this theorem, we will use analytical properties of the Fourier transform $\widehat{\mathcal{V}}$ of $\mathcal{V}$, which is why we assume that $\mathcal{V}$ decays exponentially at infinity. For potentials with slower decay, it seems that the decay of u should still be $|x|^{-4}$, except if $\mathcal{V}$ itself decays slower than $|x|^{-4}$, in which case u should decay like $\mathcal{V}$.

• It is presumably not too difficult to extend this result to cases with potentials that are not spherically symmetric.

Remark. The simple equation (1-1) is actually an approximation of a richer equation for u [Lieb 1963], which should more accurately depict the Bose gas; see (7-2). Little is known about this richer equation.

The paper is organized as follows. We prove Theorem 1.1 in Section 2, Theorems 1.2 and 1.3 in Section 3, Theorem 1.4 in Section 4, and Theorem 1.5 in Section 5. In Section 6, we explain how the simple equation is related to the Bose gas, and present some numerical evidence that it is very good at predicting the ground state energy. In Section 7 we discuss a few open problems and extensions.

2. Proof of Theorem 1.1

As explained in the Introduction, the solutions of (1-1)–(1-2) that are of physical interest are those that are integrable and satisfy $u(x) \leqslant 1$ for all x. In this section we prove, making no assumptions on the potential $\mathcal{V}$ other than its positivity and integrability, that all such solutions are nonnegative and have slow decay so that $\int |x| u(x)\, dx = \infty$.

Our starting point is the form of (1-1) given in (1-11). For an integrable solution u, define

$$f := 2e\rho Y_{4e} * u. \tag{2-1}$$

If (1-2) is satisfied, then

$$\int f\, dx = \tfrac{1}{2}, \tag{2-2}$$

and (1-11) can be written as

$$u = Y_{4e} * (\mathcal{V}(1-u)) + f * u. \tag{2-3}$$

Lemma 2.1. *Let $u(x)$ be an integrable solution of the system (1-1)–(1-2) such that $u(x) \leqslant 1$ for all x. Let f be defined in terms of u, e and ρ by (2-1). If $f(c) \geqslant 0$ for all x, then $u(x) \geqslant 0$ for all x.*

Proof. Since $Y_{4e} * (\mathcal{V}(1-u(x))) \geqslant 0$, it follows that

$$u_- \leqslant (f * u)_- = (f * u_+ - f * u_-)_- \leqslant f * u_-. \tag{2-4}$$

Integrating, we find $\int u_-\, dx \leqslant \frac{1}{2} \int u_-\, dx$, and this implies $u_- = 0$. □

Proof of Theorem 1.1. Multiply (2-3) through by $2e\rho$, and then convolve both sides with Y_{4e}. The result is $f = 2e\rho Y_{4e} * (Y_{4e} * (\mathcal{V}(1-u))) + f * f$, and since $Y_{4e} * (Y_{4e} * (\mathcal{V}(1-u))) \geqslant 0$, we know f is an integrable solution of

$$f(x) \geqslant f * f(x) \tag{2-5}$$

for all x. It is proved in [Carlen et al. 2020] that all integrable solutions of (2-5) are nonnegative and have integral no greater than $\frac{1}{2}$, and that moreover, (2-2) and (2-3) together imply

$$\int |x| f(x)\, dx = \infty. \tag{2-6}$$

However,

$$\int |x| f(x)\, dx = 2e\rho \int |x| Y_{4e} * u(x)\, dx = 2e\rho \int (Y_{4e} * |x|) u(x)\, dx. \tag{2-7}$$

Then since $\lim_{x\to\infty}(4e|x|^{-1} Y_{4e} * |x|) = 1$, (1-5) follows. □

3. Proof of Theorems 1.2 and 1.3

As was shown in (1-11) and (1-17), there are at least two ways to write (1-1) as a fixed-point equation. As it turns out, only the latter one

$$u(x) = \Phi(u)(x) := K_e(\mathcal{V}(x) + 2e\rho u * u(x)) \tag{3-1}$$

is adapted to solution by iteration, because of its monotonicity properties. Starting with $u_0(x) = 0$, define

$$u_n(x) = \Phi(u_{n-1})(x) \tag{3-2}$$

for $n \geqslant 1$. It is easy to see that for arbitrary $e, \rho \geqslant 0$, this produces a monotone increasing sequence of nonnegative integrable functions. Thus, $u(x) := \lim_{n\to\infty} u_n(x)$ will exist, but it need not be integrable and it need not satisfy (1-2) or (1-6).

To bring (1-2) into the iteration scheme, we take e as the independent parameter, and define a sequence $\{\rho_n\}$ along with the sequence $\{u_n(x)\}$, both depending on e, through

$$u_n(x) = K_e\mathcal{V}(x) + 2e\rho_{n-1}K_e u_{n-1} * u_{n-1}(x), \quad u_0(x) = 0, \tag{3-3}$$

and

$$\rho_n := \frac{2e}{\int (1-u_n(x))\mathcal{V}(x)}. \tag{3-4}$$

Comparing (3-3) to (3-1), note that the analog of Φ now depends on n.

Lemma 3.1. *Let* $\mathcal{V} \in L^1(\mathbb{R}^d) \cap L^p(\mathbb{R}^d)$, $p > \max\{d/2, 1\}$. *Both sequences* $\{\rho_n\}$ *and* $\{u_n\}$ *are well-defined and increasing, and for all* n,

$$\int_{\mathbb{R}^d} u_n\,dx < \frac{1}{2e}\int_{\mathbb{R}^d} \mathcal{V}(1-u_n)\,dx. \tag{3-5}$$

Proof. We proceed by induction. By definition, $u_0 = 0$ and $\rho_0 = 2e\bigl(\int_{\mathbb{R}^d} \mathcal{V}(x)\,dx\bigr)^{-1}$. Also by definition $u_1 = K_e\mathcal{V} \geqslant u_0$ and $\rho_1 = 2e\bigl(\int \mathcal{V}(1-K_e\mathcal{V})\,dx\bigr)^{-1}$. As noted in the discussion between (1-20) and (1-21),

$$2\int_{\mathbb{R}^d} u_1\,dx = \frac{1}{e}\int_{\mathbb{R}^d} \mathcal{V}(1-u_1)\,dx \leqslant \frac{1}{e}\int_{\mathbb{R}^d} \mathcal{V}\,dx. \tag{3-6}$$

Since $t \mapsto t^{-1}$ is monotone decreasing on $(0,\infty)$, this shows that $\rho_1 > \rho_0$, and that (3-5) holds for $n = 1$.

Now suppose that $u_n \geqslant u_{n-1} \geqslant 0$, $\rho_n \geqslant \rho_{n-1} \geqslant 0$, and $\int_{\mathbb{R}^d} u_n\,dx < (1/(2e))\int_{\mathbb{R}^d} \mathcal{V}(1-u_n)$, all of which we have just verified for $n = 1$. Then

$$u_{n+1} = K_e\mathcal{V} + 2e\rho_n K_e u_n * u_n(x) \geqslant K_e\mathcal{V} + 2e\rho_{n-1}K_e u_{n-1} * u_{n-1}(x) = u_n(x), \tag{3-7}$$

and thus

$$\int_{\mathbb{R}^d} \mathcal{V}(1-u_{n+1})\,dx < \int_{\mathbb{R}^d} \mathcal{V}(1-u_n)\,dx. \tag{3-8}$$

Integrating both sides of $u_{n+1} = G_e\mathcal{V}(1-u_{n+1}) + 2e\rho_n G_e u_n * u_n$ yields

$$2\int_{\mathbb{R}^d} u_{n+1}\,dx = \frac{1}{2e}\int_{\mathbb{R}^d} \mathcal{V}(1-u_{n+1}) + \rho_n\Bigl(\int_{\mathbb{R}^d} u_n\,dx\Bigr)^2. \tag{3-9}$$

Then since

$$\int_{\mathbb{R}^d} u_n\,dx < \frac{1}{2e}\int_{\mathbb{R}^d} \mathcal{V}(1-u_n) = \frac{1}{\rho_n},$$

(3-9) implies

$$2\int_{\mathbb{R}^d} u_n\,dx \leqslant \frac{1}{2e}\int_{\mathbb{R}^d} \mathcal{V}(1-u_n) + \int_{\mathbb{R}^d} u_{n-1}\,dx. \tag{3-10}$$

Then because $\int_{\mathbb{R}^d} u_n\,dx < \int_{\mathbb{R}^d} u_{n+1}\,dx$, we have

$$\int_{\mathbb{R}^d} u_{n+1}\,dx < \frac{1}{2e}\int_{v} \mathcal{V}(1-u_{n+1}).$$

This proves (3-5) for $n+1$ and shows that

$$0 \leqslant \frac{1}{2e}\int_{\mathbb{R}^d} \mathcal{V}(1-u_{n+1})\,dx \leqslant \frac{1}{2e}\int_{\mathbb{R}^d} \mathcal{V}(1-u_n)\,dx, \tag{3-11}$$

and then, as before, $\rho_{n+1} \geqslant \rho_n$. □

Lemma 3.2. *Let* $\mathcal{V} \in L^1(\mathbb{R}^d) \cap L^p(\mathbb{R}^d)$, $p > \max\{d/2, 1\}$. *Then for all* n, *the function* $u_n(x)$ *is continuous, vanishing at infinity, and* $0 \leqslant u_n(x) \leqslant 1$.

Proof. First consider $n = 1$. Since $u_n = K_e\mathcal{V}$ with $\mathcal{V} \in L^p(\mathbb{R}^d)$, we have $u_1 \in W^{2,p}(\mathbb{R}^d)$ and

$$\Delta u_1(x) = \mathcal{V}(x)(u_1(x)-1) + 4eu_1(x). \tag{3-12}$$

Since K_e maps $L^p(\mathbb{R}^d)$ into $W^{2,p}(\mathbb{R}^d)$, u_1 is continuous and vanishes at infinity. Let $A := \{x : u_1(x) > 1\}$. Then A is open. If A is nonempty, then u_1 is subharmonic on A, and hence takes on its maximum on the boundary of A. Since u_1 would equal 1 on the boundary, this is impossible, and A is empty. This proves the assertion for $n = 1$.

Now make the inductive hypothesis that $0 \leqslant u_n(x) \leqslant 1$ for all x. Then

$$\|u_n\|_p^p \leqslant \|u_n\|_1 \leqslant \frac{1}{2e}\int_{\mathbb{R}^d} \mathcal{V}\,dx.$$

By Young's inequality, $\|u_n * u_n\|_p \leqslant \|u_n\|_p\|u_1\|_1$, and hence $\mathcal{V} + 2e\rho_n u_n * u_n \in L^p(\mathbb{R}^d)$. Therefore, $u_{n+1} = K_e(\mathcal{V} + 2e\rho_n u_n * u_n) \in W^{2,p}(\mathbb{R}^d)$. It follows as before that u_{n+1} is continuous and vanishing at infinity, and in particular, bounded, and

$$\begin{aligned}\Delta u_{n+1}(x) &= \mathcal{V}(x)(u_n(x)-1) + 4eu_n(x) - 2e\rho_n u_n * u_n\\ &\geqslant \mathcal{V}(x)(u_n(x)-1) + 4eu_n(x) - 2e\rho_n\|u_n\|_1\|u_n\|_\infty\\ &\geqslant \mathcal{V}(x)(u_n(x)-1) + 4eu_n(x) - 2e,\end{aligned}$$

where we have used $\rho_n\|u_n\|_1 \leqslant 1$, which is valid on account of (3-5). Define $A := \{x : u_{n+1}(x) > 1\}$. Then u_{n+1} is subharmonic on A, and maximal on the boundary of A, where $u_n(x)$ would equal 1. This contradiction shows that $\|u_{n+1}\|_\infty \leqslant 1$. □

Lemma 3.3. *Let $\mathcal{V} \in L^1(\mathbb{R}^d) \cap L^p(\mathbb{R}^d)$, $p > \max\{d/2, 1\}$. Now let*

$$u(x) := \lim_{n\to\infty} u_n(x) \quad \textit{and} \quad \rho(e) = \lim_{n\to\infty} \rho_n(e). \tag{3-13}$$

Then both limits exist, $u \in W^{2,p}(\mathbb{R}^d)$ and u satisfies (1-1), (1-2) and (1-6).

Proof. By Lemma 3.1, both limits exist, and by (3-5), $\rho(e) \leqslant \left(\int_{\mathbb{R}^d} K_e \mathcal{V}\, dx\right)^{-1}$. Also by Lemma 3.1, $\int_{\mathbb{R}^d} \leqslant (1/2e) \int_{\mathbb{R}^d} \mathcal{V}(x)\, dx$, u is integrable and $\lim_{n\to\infty} \|u_n - u\|_1 = 0$. Moreover, by Lemma 3.2, $0 \leqslant u \leqslant 1$, and then $\|u\|_p^p \leqslant \|u\|_1$ and $\|u_n - u\|_p^p \leqslant (p+1)\|u_n - u\|_1$. Thus by Young's inequality

$$\|u * u - u_n * u_n\|_p \leqslant \|u_n\|_1 \|u_n - u\|_p + \leqslant \|u\|_1 \|u_n - u\|_p. \tag{3-14}$$

Therefore, $\lim_{n\to\infty}(\mathcal{V} + 2e\rho_n(e) u_n * u_n) = (\mathcal{V} + 2e\rho(e) u * u)$ with convergence in $L^p(\mathbb{R}^d)$. Then $\lim_{n\to\infty} K_e(\mathcal{V} + 2e\rho_n(e) u_n * u_n) = K_e(\mathcal{V} + 2e\rho(e) u * u)$ with convergence in $W^{2,p}(\mathbb{R}^d)$, and, in particular, in $L^p(\mathbb{R}^d)$. It now follows that $u = K_e(\mathcal{V} + 2e\rho(e) u * u)$, and by the dominated convergence theorem, the constraint $\rho = (1/(2e)) \int_{\mathbb{R}^d} \mathcal{V}(1-u)\, dx$ is satisfied. By remarks made above, this means that u satisfies (1-1)–(1-2). □

Lemma 3.4. *For all $e \in (0, \infty)$, the solution u of the system (1-1)–(1-2) that we have constructed by iteration in Lemma 3.3 is the unique nonnegative integrable solution for $\rho = \rho(e)$. Moreover, there does not exist such any such solution when $\rho \neq \rho(e)$.*

Proof. Consider any nonnegative solution integrable $\tilde{u}$, with

$$\tilde{\rho} = \frac{2e}{\int (1 - \tilde{u}(x))\mathcal{V}(x)\, dx}. \tag{3-15}$$

We first show that $\tilde{u} \geqslant u_n$ by induction. We have

$$\tilde{u}(x) - u_n(x) = 2eK_e(\tilde{\rho}\tilde{u} * \tilde{u}(x) - \rho_{n-1} u_{n-1} * u_{n-1}(x)). \tag{3-16}$$

Since $u_0 = 0$, the positivity of $\tilde{u}$ implies the positivity of $\tilde{u}(x) - u_1(x)$. If $\tilde{u} \geqslant u_{n-1}$, then, by (3-4), $\tilde{\rho} \geqslant \rho_{n-1}$, from which $\tilde{u} \geqslant u_n$ follows easily. This proves that both $\tilde{\rho} \geqslant \rho$ and $\tilde{u} \geqslant u$. However, integrating both sides of the latter inequality yields

$$\frac{1}{\tilde{\rho}(e)} = \int \tilde{u}(x)\, dx \geqslant \int u(x)\, dx = \frac{1}{\rho(e)}. \tag{3-17}$$

Since $\tilde{\rho} \geqslant \rho$, equality must hold, and then since $\tilde{u} \geqslant u$, it must be that so $u = \tilde{u}$. □

Lemma 3.5. *The function $\rho(e)$ is continuous on $(0, \infty)$, with*

$$\lim_{e\to 0} \rho(e) = 0, \qquad \lim_{e\to\infty} \rho(e) = \infty. \tag{3-18}$$

In particular, for each $\rho \in (0, \infty)$, there is at least one $e \in (0, \infty)$ such that $\rho = \rho(e)$.

Proof. We now turn to the continuity of $e \to \rho(e)$. For $n \in \mathbb{N}$, define functions $a_n(e)$ and $b_n(e)$ by

$$a_n := \int u_n(x, e)\, dx \quad \text{and} \quad b_n(e) = \frac{1}{2e} \int (1 - u_n(x, e))\mathcal{V}(x)\, dx, \tag{3-19}$$

where we have temporarily made the dependence of u_n on e explicit. Note that $b_n(e) = 1/\rho_n(e)$, and $u_1(x, e) = K_e\mathcal{V}$ is continuous in e (and monotone decreasing) for each x. A simple induction shows that $u_n(x, e)$ is continuous in e for each x. Then since $(1 - u_n(x, e))\mathcal{V}(x) \leqslant \mathcal{V}(x)$, the dominated convergence theorem yields the continuity of $\rho_n(e)$ for each n. Writing our iteration in the equivalent form (as in (1-11))

$$u_n(x, e) = Y_{4e} * (\mathcal{V}(1 - u_n(x, e))) + 2e\frac{1}{b_{n-1}(e)}Y_{4e} * u_{n-1} * u_{n-1}(x, e), \tag{3-20}$$

and integrating, we obtain

$$2a_n(x) = b_n(e) + \frac{1}{b_{n-1}(e)}a^2_{n-1}(e). \tag{3-21}$$

Now an easy induction shows that $a_n(e)$ is continuous for each n. By (3-5), for each n,

$$a_n(e) \leqslant \frac{1}{\rho(e)} \leqslant b_n(e). \tag{3-22}$$

By Lemma 3.1, as n increases to infinity, $a_n(e)$ increases to $1/\rho(e)$, while $b_n(e)$ decreases to $1/\rho(e)$. It remains to show that this convergence is uniform on any compact interval in $(0, \infty)$. By (3-21),

$$\frac{1}{b_n(e)}(a_n(e) - b_n(e))^2 = \frac{a^2_n(e)}{b_n(e)} - (2a_n(e) - b_n(e)) = \frac{a^2_n(e)}{b_n(e)} - \frac{a^2_{n-1}(e)}{b_{n-1}(e)}. \tag{3-23}$$

Sum both sides over $n \in \mathbb{N}$. The sum on the right telescopes, and since, for all e, it holds that $a^2_0/b_0 = 0$ while $\lim_{n\to\infty} a^2_n(e)/b_n(e) = 1/\rho_n(e)$, we have

$$\sum_{n=1}^{\infty} \frac{1}{b_n(e)}(a_n(e) - b_n(e))^2 = \frac{1}{\rho(e)}. \tag{3-24}$$

By the bounds on $b_(e) = 1/\rho_n(e)$ and $\rho(e)$ provided by Lemma 3.1, for all $e > 0$,

$$\sum_{n=1}^{\infty}(a_n(e) - b_n(e))^2 \leqslant \frac{\int \mathcal{V}\,dx}{\int K_e\mathcal{V}\,dx}, \tag{3-25}$$

and on any compact interval $[e_1, e_2]$, the right-hand side is uniformly bounded by C, its value at e_2. Then since the summand on the left is monotone decreasing in n, we obtain for each n that

$$(a_n(e) - b_n(e))^2 \leqslant \frac{C}{n} \tag{3-26}$$

uniformly on $[e_1, e_2]$. This proves the desired uniform convergence, and hence the continuity of $\rho(e)$. The final statement now follows from (1-21). □

Remark. Note that $\|u - u_n\|_1 = 1/\rho - a_n$, and hence by (3-26), $\|u - u_n\|_1 \leqslant Cn^{-1/2}$. In fact, numerically, we find that the rate is significantly faster than this. For example, with $\mathcal{V}(x) = e^{-|x|}$ and $e = 10^{-4}$, $\|u - u_n\|_1$ decays at least as fast as $n^{-3.5}$.

Proof of Theorem 1.2. This theorem follows from Lemmas 3.2, 3.3 and 3.4. □

Proof of Theorem 1.3. Every statement in the theorem has been established in Lemmas 3.1–3.5. □

We close this section by remarking that if $\mathcal{V}$ is radially symmetric, then so is $u_1 = K_e\mathcal{V}$, and then by a simple induction, so is u_n, and hence also the unique solution u provided by Theorem 1.3. This is consistent with the first remark following Theorem 1.3.

4. Asymptotics

In this section, we prove Theorem 1.4. We will first prove the high-density asymptote (1-24), and then proceed to the low-density (1-23).

By Theorem 1.3, for each $\rho > 0$ there exists at least one e such that $\rho(e) = \rho$. If there is more than one, the theorems proved in this section apply to every such solution. Throughout this section, let u_ρ denote the solution provided by Theorem 1.3 and any such choice of e.

4A. *High-density ρ.*

Lemma 4.1 (high-density asymptotics). *If $\mathcal{V}$ is integrable, then as $\rho \to \infty$,*

$$e = \frac{\rho}{2}\left(\int \mathcal{V}(x)\,dx\right)(1+o(1)). \tag{4-1}$$

Remark. From (1-2),

$$e \leqslant \frac{\rho}{2}\int \mathcal{V}(x)\,dx. \tag{4-2}$$

Note that this is not an optimal bound, as follows from (1-20).

Proof. By (1-2), it suffices to prove that

$$\lim_{\rho\to\infty}\int u_\rho(x)\mathcal{V}(x)\,dx = 0. \tag{4-3}$$

Let

$$\chi_\gamma := \{x : \mathcal{V}(x) \geqslant \gamma\} \tag{4-4}$$

and take the decomposition

$$\int u_\rho(x)\mathcal{V}(x)\,dx = \int_{\chi_\gamma} u_\rho(x)\mathcal{V}(x)\,dx + \int_{\mathbb{R}^d\setminus\chi_\gamma} u_\rho(x)\mathcal{V}(x)\,dx, \tag{4-5}$$

which, by (1-7), is bounded as

$$\int u_\rho(x)\mathcal{V}(x)\,dx \leqslant \int_{\chi_\gamma} \mathcal{V}(x)\,dx + \frac{\gamma}{\rho}. \tag{4-6}$$

Since $\mathcal{V}$ is integrable, $\int_{\chi_\gamma} \mathcal{V}(x)\,dx \to 0$ as $\gamma \to \infty$. Therefore,

$$\inf_{\gamma>0}\left(\int_{\chi_\gamma} \mathcal{V}(x)\,dx + \frac{\gamma}{\rho}\right) \xrightarrow[\rho\to\infty]{} 0, \tag{4-7}$$

completing the proof. □

4B. ***Low-density ρ.*** In this section, we only consider the dimension $d = 3$. As before, we suppose that $\mathcal{V} \in L^1(\mathbb{R}^3) \cap L^p(\mathbb{R}^3),\ p > \frac{3}{2}$, and $\mathcal{V} \geqslant 0$.

We first recall the definition of the *scattering length* of the potential $\mathcal{V}$ and relate it to the solution of the system (1-1)–(1-2). The *scattering equation* is defined as

$$-\Delta\varphi(x) = (1-\varphi(x))\mathcal{V}(x), \qquad \lim_{|x|\to\infty} \varphi(x) = 0. \tag{4-8}$$

Note that (4-8) can be written as $(-\Delta + \mathcal{V})\varphi = \mathcal{V}$, and hence the solution is

$$\varphi(x) = \lim_{e\downarrow 0} K_e\mathcal{V}(x) = \lim_{e\downarrow 0} u_1(x, e), \tag{4-9}$$

where u_1 is the first term of the iteration introduced in the previous section. It follows from Lemma 3.2 that $0 \leqslant \varphi(x) \leqslant 1$ for all x.

We now impose a mild localization hypothesis on $\mathcal{V}$: For $R > 0$ define $\mathcal{V}_R(x) = \mathcal{V}(x)$ for $|x| > R$ and otherwise $\mathcal{V}_R(x) = 0$. We require that, for some $q > 1$ and all sufficiently large R,

$$\|\mathcal{V}_R\|_1 < R^{-q} \quad \text{and} \quad \|\mathcal{V}_R\|_p < R^{-q}. \tag{4-10}$$

By the lemma below, $\lim_{|x|\to\infty} |x|\varphi(x)$ exists. The *scattering length a* is defined to be (in dimension $d = 3$).

$$a = \lim_{|x|\to\infty} |x|\varphi(x). \tag{4-11}$$

For more information on the scattering length, see [Lieb and Yngvason 2001, Appendix A].

Lemma 4.2. *Let $\mathcal{V} \in L^1(\mathbb{R}^3) \cap L^p(\mathbb{R}^3),\ p > \frac{3}{2}$, and suppose that the localization condition (4-10) is satisfied. Let φ be the corresponding scattering solution given by (4-9). Then the scattering length $a := \lim_{|x|\to\infty} \varphi(x)$ exists and satisfies*

$$4\pi a = \int \mathcal{V}(x)(1-\varphi(x))\, dx. \tag{4-12}$$

Proof. By the resolvent identity, $\varphi(x) = G * (\mathcal{V}(1-\varphi))(x)$ where $G(x) = 1/(4\pi|x|)$. Since $p > \frac{3}{2}$, $p' < 3$, and it is easy to decompose G into the sum of two pieces $G = G_1 + G_2$, where $G_1 \in L^{p'}(\mathbb{R}^d)$ and $G_2 \in L^4(\mathbb{R}^d)$. Then for all R sufficiently large,

$$0 \leqslant G * (\mathcal{V}_{\mathcal{R}}(1-\varphi))(x) \leqslant (\|G_1\|_{p'} + \|G_2\|_4)R^{-q}. \tag{4-13}$$

For $0 < r < 1$ and $|y| < r|x|$,

$$\frac{1}{1+r} \leqslant \frac{|x|}{|x-y|} \leqslant \frac{1}{1-r}.$$

It follows that for all sufficiently large $|x|$,

$$\frac{1}{1+r}\int_{|y|<r|x|} \mathcal{V}(y)(1-\varphi(y))\, dy + o(1) \leqslant 4\pi|x|\varphi(x) \leqslant \frac{1}{1-r}\int_{|y|<r|x|} \mathcal{V}(y)(1-\varphi(y))\, dx + o(1). \tag{4-14}$$

Taking $|x| \to \infty$ and then $r \to 0$ proves (4-10). □

Remark. The following lemma is valid if the scattering length a is *defined* by (4-12). For this reason, we do not impose the additional condition (4-10) in the statement of Theorem 1.4: Lemma 4.2 reconciles the stated definition with the formula (4-12).

Lemma 4.3 (low-density asymptotics). *If $\mathcal{V}$ is nonnegative and integrable and $d=3$, then*

$$e=2\pi\rho a\left(1+\frac{128}{15\sqrt{\pi}}\sqrt{\rho a^3}+o(\sqrt{\rho})\right). \tag{4-15}$$

Proof. The scheme of the proof is as follows. We first approximate the solution u by w, which is defined as the decaying solution of

$$-\Delta w_\rho(x)=(1-u_\rho(x))\mathcal{V}(x). \tag{4-16}$$

The energy of w_ρ is defined to be

$$e_w:=\frac{\rho}{2}\int(1-w_\rho(x))\mathcal{V}(x)\,dx, \tag{4-17}$$

and, as we will show, it is *close* to e; more precisely,

$$e-e_w=\frac{16\sqrt{2}e^{3/2}}{15\pi^2}\int\mathcal{V}(x)\,dx+o(\rho^{3/2}). \tag{4-18}$$

In addition, (4-16) is quite similar to the scattering equation (4-8). In fact we will show that e_w is *close* to the energy $2\pi\rho a$ of the scattering equation

$$e_w-2\pi\rho a=-\frac{16\sqrt{2}e^{3/2}}{15\pi^2}\int\varphi(x)\mathcal{V}(x)\,dx+o(\rho^{3/2}). \tag{4-19}$$

Summing (4-18) and (4-19), we find

$$e=2\pi\rho a\left(1+\frac{32\sqrt{2}e^{3/2}}{15\pi^2\rho}+o(\sqrt{\rho})\right), \tag{4-20}$$

from which (4-15) follows. We are thus left with proving (4-18) and (4-19).

Proof of (4-18): By (1-2) and (4-17),

$$e-e_w=\frac{\rho}{2}\int(w_\rho(x)-u_\rho(x))\mathcal{V}(x)\,dx. \tag{4-21}$$

We will work in Fourier space

$$\hat{u}_\rho(k):=\int e^{ikx}u_\rho(x)\,dx, \tag{4-22}$$

which satisfies, by (1-1),

$$(k^2+4e)\hat{u}_\rho(k)=\frac{2e}{\rho}S(k)+2e\rho\hat{u}^2(k), \tag{4-23}$$

with

$$S(k):=\frac{\rho}{2e}\int e^{ikx}(1-u_\rho(x))\mathcal{V}(x)\,dx. \tag{4-24}$$

Therefore,

$$\hat{u}_\rho(k) = \frac{1}{\rho}\left(\frac{k^2}{4e} + 1 - \sqrt{\left(\frac{k^2}{4e}+1\right)^2 - S(k)}\right). \tag{4-25}$$

Similarly, the Fourier transform of w_ρ is

$$\hat{w}_\rho(k) := \int e^{ikx} w_\rho(x)\,dx = \frac{2eS(k)}{\rho k^2}. \tag{4-26}$$

Note that, as $|k| \to \infty$, we have $\hat{u} \sim 2eS(k)/(\rho k^2)$, so, while $\hat{u}_\rho$ is not integrable, $\hat{u}_\rho - \hat{w}_\rho$ is. We invert the Fourier transform:

$$u_\rho(x) - w_\rho(x) = \frac{1}{8\pi^3\rho}\int e^{-ikx}\left(\frac{k^2}{4e} + 1 - \sqrt{\left(\frac{k^2}{4e}+1\right)^2 - S(k)} - \frac{2eS(k)}{k^2}\right)dk. \tag{4-27}$$

We change variables to $\tilde{k} := k/(2\sqrt{e})$:

$$u_\rho(x) - w_\rho(x) = \frac{e^{3/2}}{\rho\pi^3}\int e^{-i2\sqrt{e}\tilde{k}x}\left(\tilde{k}^2 + 1 - \sqrt{(\tilde{k}^2+1)^2 - S(2\tilde{k}\sqrt{e})} - \frac{S(2\tilde{k}\sqrt{e})}{2\tilde{k}^2}\right)d\tilde{k}. \tag{4-28}$$

Furthermore,

$$s \mapsto \left|\tilde{k}^2 + 1 - \sqrt{(\tilde{k}^2+1)^2 - s} - \frac{s}{2\tilde{k}^2}\right| \tag{4-29}$$

is monotone increasing. In addition, by (4-24) and (1-1), and using the fact that $u_\rho(x) \leqslant 1$ (see Lemma 3.2) and $\mathcal{V}(x) \geqslant 0$,

$$|S(k)| \leqslant \frac{\rho}{2e}\int |(1-u_\rho(x))\mathcal{V}(x)|\,dx = 1. \tag{4-30}$$

Therefore

$$\left|\tilde{k}^2 + 1 - \sqrt{(\tilde{k}^2+1)^2 - S(2\tilde{k}\sqrt{e})} - \frac{S(2\tilde{k}\sqrt{e})}{2\tilde{k}^2}\right| \leqslant \left|\tilde{k}^2 + 1 - \sqrt{(\tilde{k}^2+1)^2 - 1} - \frac{1}{2\tilde{k}^2}\right|. \tag{4-31}$$

Thus

$$|u_\rho(x) - w_\rho(x)| \leqslant \frac{e^{3/2}}{\rho\pi^3}\int\left|\tilde{k}^2 + 1 - \sqrt{(\tilde{k}^2+1)^2 - 1} - \frac{1}{2\tilde{k}^2}\right|d\tilde{k} = \frac{32\sqrt{2}e^{3/2}}{15\pi^2\rho}. \tag{4-32}$$

By dominated convergence, and using the fact that $S(0) = 1$,

$$\begin{aligned}\lim_{e\to 0}\frac{1}{e^{3/2}}(e - e_w) &= -\lim_{e\to 0}\frac{\rho}{2e^{3/2}}\int (u_\rho(x) - w_\rho(x))\mathcal{V}(x)\,dx \\ &= -\frac{1}{2}\int \mathcal{V}(x)\left(\frac{1}{\pi^3}\int\left(\tilde{k}^2 + 1 - \sqrt{(\tilde{k}^2+1)^2 - 1} - \frac{1}{2\tilde{k}^2}\right)d\tilde{k}\right)dx \\ &= \frac{16\sqrt{2}}{15\pi^2}\int \mathcal{V}(x)\,dx.\end{aligned} \tag{4-33}$$

Using (1-22), this proves (4-18). Incidentally, again by dominated convergence,

$$u_\rho(x) - w_\rho(x) = \frac{e^{3/2}}{\rho\pi^3}\int\left(\tilde{k}^2 + 1 - \sqrt{(\tilde{k}^2+1)^2 - 1} - \frac{1}{2\tilde{k}^2}\right)d\tilde{k} = -\frac{32\sqrt{2}e^{3/2}}{15\pi^2\rho} + \sqrt{\rho}f_\rho(x), \tag{4-34}$$

with

$$0 \leqslant f_\rho(x) \leqslant \frac{32\sqrt{2}e^{3/2}}{15\pi^2\rho}, \quad f_\rho(x) \xrightarrow[\rho\to 0]{} 0, \tag{4-35}$$

pointwise in x.

Proof of (4-19): Let

$$\xi(r) := w_\rho(r) - \varphi(r). \tag{4-36}$$

By (4-16), (4-8) and (1-1),

$$(-\Delta + \mathcal{V}(x))\xi(x) = -(u_\rho(x) - w_\rho(x))\mathcal{V}(x). \tag{4-37}$$

Therefore, by (4-12),

$$e_w - 2\pi\rho a = -\frac{\rho}{2}\int \xi(x)\mathcal{V}(x)\,dx = -\frac{\rho}{2}\int \mathcal{V}(x)(-\Delta+\mathcal{V})^{-1}((u-w)\mathcal{V})(x)\,dx \tag{4-38}$$

and

$$(-\Delta+\mathcal{V})^{-1}\mathcal{V}(x) = \varphi(x), \tag{4-39}$$

so

$$e_w - 2\pi\rho a = -\frac{\rho}{2}\int \varphi(x)(u_\rho(x) - w_\rho(x))\mathcal{V}(x)\,dx. \tag{4-40}$$

By (4-34),

$$e_w - 2\pi\rho a = \frac{16\sqrt{2}e^{3/2}}{15\pi^2}\int \varphi(x)\mathcal{V}(x)\,dx - \frac{\rho^{3/2}}{2}\int \varphi(x)f_\rho(x)\mathcal{V}(x)\,dx. \tag{4-41}$$

Since $x \mapsto f_\rho(x)$ is bounded, we can use dominated convergence to show (4-19). □

5. Decay of u

In this section, we prove Theorem 1.5. Our proof assumes that $\mathcal{V}$ decays exponentially, because we will use analyticity properties of the Fourier transform of the potential $\mathcal{V}$. In particular, the theorem holds if $\mathcal{V}$ has compact support. We expect the result to hold for any potential that decays faster than $|x|^{-4}$. Algebraic decay for u seems natural: by (1-1), $u * u$ must decay at infinity in the same way as u. This is the case if u decays algebraically, but would not be so if, say, it decayed exponentially.

Proof of Theorem 1.5. We begin by proving (1-25) in arbitrary dimension. Recall that the first part has already been proved in Theorem 1.1 without the additional assumption on the potential. For the second part, recall that by the first remark after Theorem 1.3, u is also radial, and hence $\mathcal{V}(1-u)$ is nonnegative and radial. It then follows from the hypotheses on $\mathcal{V}$ that $g := 2e\rho Y_{4e} * Y_{4e} * [\mathcal{V}(1-u)]$ satisfies

$$\int |x|^2 g(x)\,dx < \infty \quad \text{and} \quad \int x g(x)\,dx = 0. \tag{5-1}$$

Then, as explained in Section 2, if $f := 2e\rho Y_{4e} * u$, we have $f - f * f = g \geqslant 0$, and then by [Carlen et al. 2020, Theorem 4], the second part of (1-25) follows. Note that if

$$u(|x|) \underset{|x|\to\infty}{\sim} \frac{\alpha}{|x|^m} \tag{5-2}$$

for some $\alpha > 0$, then the only choice of m that is consistent with (1-25) is $m = d + 1$.

We now specialize to $d=3$, and impose the additional assumption on the potential. Recall that the Fourier transform of u (4-22) satisfies (4-25),

$$\hat{u}(|k|)=\frac{1}{\rho}\left(\frac{k^2}{4e}+1-\sqrt{\left(\frac{k^2}{4e}+1\right)^2-S(|k|)}\right), \tag{5-3}$$

where S was defined in (4-24),

$$S(|k|):=\frac{\rho}{2e}\int e^{ikx}(1-u(|x|))\mathcal{V}(|x|)\,dx. \tag{5-4}$$

We split $\hat{u}$ into

$$\hat{u}(|k|)=\widehat{\mathcal{U}}_1(|k|)+\widehat{\mathcal{U}}_2(|k|), \tag{5-5}$$

with

$$\widehat{\mathcal{U}}_1(|k|):=\frac{2eS(|k|)}{\rho(1+k^2)}, \tag{5-6}$$

so that, taking the large $|k|$ limit in (4-25),

$$\widehat{\mathcal{U}}_2(|k|)=O(|k|^{-4}S^2(|k|)) \tag{5-7}$$

so $\widehat{\mathcal{U}}_2$ is integrable.

5A. ***Decay of*** $\boldsymbol{\mathcal{U}_1}$**.** We first show that

$$\mathcal{U}_1(|x|):=\frac{1}{(2\pi)^3}\int e^{-ikx}\widehat{\mathcal{U}}_1(|k|)\,dk \tag{5-8}$$

decays exponentially in $|x|$. We have

$$\mathcal{U}_1(|x|)=(-\Delta+1)^{-1}(1-u(|x|))\mathcal{V}(|x|)=Y_1*((1-u)\mathcal{V})(|x|), \tag{5-9}$$

with

$$Y_1(|x|):=\frac{e^{-|x|}}{4\pi|x|}. \tag{5-10}$$

Therefore, by (1-26),

$$\mathcal{U}_1(|x|)\leqslant\frac{A}{4\pi}\int_{|y|>R}\frac{e^{-|x-y|-B|y|}}{|x-y|}\,dy+\frac{1}{4\pi}\int_{|y|<R}\frac{e^{-|x-y|}}{|x-y|}\mathcal{V}(|y|)\,dy, \tag{5-11}$$

so, setting $b:=\min(B,1)$,

$$\mathcal{U}_1(|x|)\leqslant\frac{A}{4\pi}\int\frac{e^{-b(|x-y|+|y|)}}{|x-y|}\,dy+\frac{e^{-(|x|-R)}}{4\pi(|x|-R)}\int\mathcal{V}(|y|)\,dy, \tag{5-12}$$

and since

$$\frac{A}{4\pi}\int\frac{e^{-b(|x-y|+|y|)}}{|x-y|}\,dy=\frac{Ae^{-b|x|}}{4b^2}(b|x|+1), \tag{5-13}$$

we have

$$\mathcal{U}_1(|x|)\leqslant\frac{Ae^{-b|x|}}{4b^2}(b|x|+1)+\frac{e^{-(|x|-R)}}{4\pi(|x|-R)}\int\mathcal{V}(|y|)\,dy. \tag{5-14}$$

5B. ***Analyticity of $\mathcal{U}_2$.*** We now turn to

$$\mathcal{U}_2(|x|) := \frac{1}{(2\pi)^3}\int e^{-ikx}\widehat{\mathcal{U}}_2(|k|)\,dk = \frac{1}{4i\pi^2|x|}\sum_{\eta=\pm}\eta\int_0^\infty e^{i\eta\kappa|x|}\kappa\widehat{\mathcal{U}}_2(\kappa)\,d\kappa. \tag{5-15}$$

We start by proving some analytic properties of $\widehat{\mathcal{U}}_2$, which, we recall from (4-25) and (5-5), is

$$\widehat{\mathcal{U}}_2(|k|) = \frac{1}{\rho}\left(\frac{k^2}{4e}+1-\sqrt{\left(\frac{k^2}{4e}+1\right)^2 - S(|k|)} - \frac{2eS(|k|)}{1+k^2}\right). \tag{5-16}$$

5B1. First of all, S is analytic in a strip about the real axis,

$$S(\kappa) = 4\pi\int_0^\infty \operatorname{sinc}(\kappa r)r^2\mathcal{V}(r)(1-u(r))\,dr, \quad \operatorname{sinc}(\xi) := \frac{\sin(\xi)}{\xi}, \tag{5-17}$$

so

$$\partial^n S(\kappa) = 4\pi\int_0^\infty \partial^n \operatorname{sinc}(\kappa r)r^{n+2}\mathcal{V}(r)(1-u(r))\,dr. \tag{5-18}$$

We will show that if $\mathcal{I}m(\kappa) \leqslant \frac{1}{2}B$ (the factor $\frac{1}{2}$ can be improved to any factor that is < 1, but this does not matter here), then there exists $C > 0$ which only depends on A and B such that

$$|\partial^n S(\kappa)| \leqslant n!\,C^n. \tag{5-19}$$

As a consequence, S is analytic in a strip around the real line of height $\frac{1}{2}B$. In particular, if we define the strip

$$H_\tau := \{z : |\mathcal{I}m(z)| \leqslant r^{-\tau},\ \mathcal{R}e(z) > 0\}, \tag{5-20}$$

with $0 < \tau < 1$, and take

$$r > \left(\frac{B}{2}\right)^{-1/\tau}, \tag{5-21}$$

then S is analytic in H_τ.

5B1.2. We now prove (5-19). We first treat the case $|\kappa| \leqslant \frac{1}{2}B$. We have

$$\operatorname{sinc}(\xi) = \sum_{p=0}^\infty \frac{(-1)^p\xi^{2p}}{(2p+1)!}, \tag{5-22}$$

so

$$\partial^n \operatorname{sinc}(\xi) = \sum_{p=\lceil n/2\rceil}^\infty \frac{(-1)^p\xi^{2p-n}}{(2p+1)(2p-n)!}. \tag{5-23}$$

Therefore

$$|\partial^n \operatorname{sinc}(\xi)| \leqslant \sum_{p=\lceil n/2\rceil}^\infty \frac{|\xi|^{2p-n}}{(2p-n)!} \leqslant \cosh(|\xi|). \tag{5-24}$$

Thus

$$|\partial^n S(\kappa)| \leqslant 4\pi\int_0^\infty \cosh(|\kappa|r)r^{n+2}\mathcal{V}(r)(1-u(r))\,dr, \tag{5-25}$$

so, by (1-26),

$$|\partial^n S(\kappa)| \leqslant 4A\pi \int_R^\infty \cosh(|\kappa|r) r^{n+2} e^{-Br}\, dr + 4\pi \int_0^R \cosh(|\kappa|r) r^{n+2} \mathcal{V}(r)\, dr, \tag{5-26}$$

$$|\partial^n S(\kappa)| \leqslant 8A\pi \int_0^\infty r^{n+2} e^{-(B-|\kappa|)r}\, dr + 8\pi e^{|\kappa|R} R^n \int r^2 \mathcal{V}(r)\, dr, \tag{5-27}$$

which, if $|\kappa| \leqslant \frac{1}{2}B$, implies

$$8A\pi \int_0^\infty r^{n+2} e^{-(B-|\kappa|)r}\, dr \leqslant 8A\pi \int_0^\infty r^{n+2} e^{-(B/2)r}\, dr = \frac{2^{n+6}A\pi}{B^{n+3}}(n+2)!\,, \tag{5-28}$$

$$8\pi e^{|\kappa|R} R^{n+2} \int \mathcal{V}(r)\, dr \leqslant 8\pi e^{(B/2)R} R^n \int r^2 \mathcal{V}(r)\, dr, \tag{5-29}$$

which implies (5-19) in this case.

5B1.2. We now turn to $|\kappa| \geqslant \frac{1}{2}B$:

$$\partial^n \operatorname{sinc}(\xi) = \sum_{p=0}^n \binom{n}{p} \partial^p \sin(\xi) \frac{(n-p)!(-1)^{n-p}}{\xi^{n-p+1}} \tag{5-30}$$

so

$$|\partial^n \operatorname{sinc}(\xi)| \leqslant 2e^{\mathcal{I}m(\xi)} \sum_{p=0}^n \frac{n!}{p!} |\xi|^{-(n-p+1)}. \tag{5-31}$$

Therefore,

$$|\partial^n S(\kappa)| \leqslant 8\pi \sum_{p=0}^n \frac{n!}{p!|\kappa|^{n-p+1}} \int_0^\infty e^{\mathcal{I}m(\kappa)r} r^{p+1} \mathcal{V}(r)(1-u(r))\, dr, \tag{5-32}$$

so, by (1-26),

$$|\partial^n S(\kappa)| \leqslant \sigma_1 + \sigma_2, \tag{5-33}$$

with

$$\sigma_1 := 8A\pi \sum_{p=0}^n \frac{n!}{p!\,|\kappa|^{n-p+1}} \int_R^\infty r^{p+1} e^{-(B-\mathcal{I}m(\kappa))r}\, dr, \tag{5-34}$$

$$\sigma_2 := 8\pi \sum_{p=0}^n \frac{n!}{p!\,|\kappa|^{n-p+1}} \int_0^R r^{p+1} e^{\mathcal{I}m(\kappa)r} \mathcal{V}(r)\, dr. \tag{5-35}$$

Furthermore,

$$\sigma_1 = 8A\pi n! \sum_{p=0}^n \frac{p+1}{(B-\mathcal{I}m(\kappa))^{p+2}|\kappa|^{n-p+1}} \tag{5-36}$$

so, as long as $|\kappa| \geqslant \frac{1}{2}B$ and $\mathcal{I}m(\kappa) \leqslant \frac{1}{2}B$,

$$\sigma_1 \leqslant \frac{2^{n+6}A\pi}{B^{n+3}} n! \sum_{p=0}^n (p+1) = \frac{2^{n+5}A\pi}{B^{n+3}}(n+2)!\,. \tag{5-37}$$

In addition,

$$\sigma_2 \leqslant 8\pi \sum_{p=0}^{n} \frac{n!}{p!\,|\kappa|^{n-p+1}} R^{p-1} e^{\mathcal{I}m(\kappa)R} \int_0^R r^2 \mathcal{V}(r)\,dr, \tag{5-38}$$

so

$$\sigma_2 \leqslant 8\pi \sum_{p=0}^{n} \frac{n!\,2^{n-p+1}}{p!\,B^{n-p+1}} R^{p-1} e^{\mathcal{I}m(\kappa)R} \int_0^R r^2 \mathcal{V}(r)\,dr \leqslant \frac{2^{n+4}\pi}{RB^{n+1}} n! e^{RB} \int_0^R r^2 \mathcal{V}(r)\,dr, \tag{5-39}$$

which implies (5-19) in this case.

5B2. We have thus proved that S is analytic in H_τ, which implies that the singularities of $\widehat{\mathcal{U}}_2$ in H_τ all come from the branch points of $\sqrt{F(|k|)}$ with $F(|k|) := (k^2/(4e)+1)^2 - S(|k|)$. For $\kappa \in \mathbb{R}$,

$$|S(\kappa)| \leqslant 1, \tag{5-40}$$

so, for $\kappa \in \mathbb{R}$,

$$F(\kappa) \geqslant \frac{\kappa^2}{2e}. \tag{5-41}$$

Therefore, since F is analytic in a strip around the real axis, there exists an open set containing the real axis in which F has one and only one root, at 0. Thus the only branch point of $\sqrt{F}$ on the real axis is 0. Thus, $\widehat{\mathcal{U}}_2$ is analytic in H_τ.

5C. *Decay of $\mathcal{U}_2$.* We deform the integral to the path

$$\{i\eta y : 0 < y < |x|^{-\tau}\} \cup \{i\eta|x|^{-\tau} + y : y > 0\} \tag{5-42}$$

and find

$$\int_0^\infty e^{i\eta\kappa|x|} \kappa \widehat{\mathcal{U}}_2(\kappa)\,d\kappa = I_1 + I_2, \tag{5-43}$$

with

$$I_1 := -\int_0^{|x|^{-\tau}} e^{-y|x|} y \widehat{\mathcal{U}}_2(i\eta y)\,dy, \tag{5-44}$$

$$I_2 := e^{-|x|^{1-\tau}} \int_0^\infty e^{i\eta y|x|} (i\eta|x|^{-\tau} + y) \widehat{\mathcal{U}}_2(i\eta|x|^{-\tau} + y)\,dy. \tag{5-45}$$

5C1. We first estimate I_1. We expand S:

$$S(\kappa) = 1 - \beta\kappa^2 + O(|\kappa|^4), \tag{5-46}$$

with $\beta > 0$ (since S is analytic and symmetric, and $|S(|k|)| \leqslant 1$). Therefore, $y \mapsto \widehat{\mathcal{U}}_2(iy)$ is $\mathcal{C}^2$ for $y \neq 0$, and

$$\widehat{\mathcal{U}}_2(i\eta y) = \frac{1}{\rho} - \frac{i\eta y}{\rho} \sqrt{\frac{1}{2e} + \beta} + O(y^2). \tag{5-47}$$

Furthermore,

$$-\int_0^{|x|^{-\tau}} e^{-y|x|}y\,dy = -\frac{1}{|x|^2}+\frac{1+|x|^{1-\tau}}{|x|^2}e^{-|x|^{1-\tau}}, \tag{5-48}$$

$$-\int_0^{|x|^{-\tau}} e^{-y|x|}y^2\,dy = -\frac{2}{|x|^3}+\frac{1+|x|^{1-\tau}(2+x^{1-\tau})}{|x|^3}e^{-|x|^{1-\tau}} \tag{5-49}$$

and

$$-\int_0^{|x|^{-\tau}} e^{-y|x|}y^3\,dy = O(|x|^{-4}), \tag{5-50}$$

$$I_1 = -\frac{1}{\rho|x|^2}+\frac{2i\eta}{\rho|x|^3}\sqrt{\frac{1}{2e}+\beta}+O(|x|^{-4}), \tag{5-51}$$

so

$$\frac{1}{4i\pi^2|x|}\sum_{\eta=\pm}\eta I_1 = \frac{1}{\pi^2\rho|x|^4}\sqrt{\frac{1}{2e}+\beta}+O(|x|^{-5}). \tag{5-52}$$

5C2. We now bound I_2. Recall that, for $\kappa\in\mathbb{R}$, we have $|S(\kappa)|\leqslant 1$. Recalling (5-19),

$$|S(\kappa+i\eta|x|^{-\tau})|\leqslant\sum_{n=0}^{\infty}\frac{1}{n!}|\partial^n S(\kappa)|^n|x|^{-n\tau}\leqslant\frac{1}{1-C|x|^{-\tau}}\leqslant 2 \tag{5-53}$$

provided $|x|^\tau > 2C$. Therefore, for large κ,

$$|\widehat{\mathcal{U}}_2(\kappa+i\eta)| = O(\kappa^{-4}), \tag{5-54}$$

so

$$I_2\leqslant C'e^{-|x|^{1-\tau}} \tag{5-55}$$

for some constant $C'>0$.

5C3. Inserting (5-52) and (5-55) into (5-43) and (5-15), we find that

$$\mathcal{U}_2(|x|) = \frac{1}{\pi^2\rho|x|^4}\sqrt{\frac{1}{2e}+\beta}+O(|x|^{-5}), \tag{5-56}$$

which, using (5-14), concludes the proof of the theorem. □

6. Comparison with the Bose gas

6A. ***Sketch of the derivation of the simple equation.*** The simple equation (1-1)–(1-2) was originally derived [Lieb 1963] to approximate the ground state energy E_0 of a repulsive Bose gas, which is a system of N quantum particles interacting via the repulsive potential $\mathcal{V}$. The ground state energy of this system is the lowest eigenvalue of the Hamiltonian operator

$$H_N := -\frac{1}{2}\sum_{i=1}^{N}\Delta_i+\sum_{1\leqslant i<j\leqslant N}\mathcal{V}(x_i-x_j) \tag{6-1}$$

acting on the space of L_2 functions on the torus $\mathbb{T}_V$ of volume V. The corresponding eigenfunction, which we will denote by ψ_N, satisfies

$$H_N\psi_N(x_1,\dots,x_N)=E_0\psi_N(x_1,\dots,x_N),\tag{6-2}$$

with $x_i\in\mathbb{T}_V$. As is well known, by a Perron–Frobenius argument, ψ_N is unique, nonnegative, and hence symmetric under exchanges $x_i\leftrightarrow x_j$ and under translations.

We can write E_0 by integrating both sides of (6-2),

$$E_0=\frac{N(N-1)}{2V}\int g_N^{(2)}(x)\mathcal{V}(x)\,dx,\tag{6-3}$$

with

$$g_N^{(p)}(x_1,\dots,x_p):=\frac{V^j}{\int\psi_N(x_1,\dots,x_N)\,dx_1\cdots dx_N}\int\psi_N(x_1,\dots,x_N)\,dx_{p+1}\cdots dx_N\tag{6-4}$$

and $g_N^{(2)}(x_1,x_2)\equiv g_N^{(2)}(x_1-x_2)$. The computation of E_0 thus reduces to that of $g_N^{(2)}$. Note that the kinetic energy does not appear explicitly in (6-3).

To compute $g_N^{(2)}$, integrate both sides of (6-2) with respect to $x_3,\dots,x_N$. This yields an equation relating $g_N^{(2)}$, $g_N^{(3)}$ and $g_N^{(4)}$. The main approximation made in [Lieb 1963] is to write $g_N^{(3)}$ and $g_N^{(4)}$ as products of $g_N^{(2)}$ factors: roughly,

$$g_N^{(p)}(x_1,\dots,x_p)\approx\prod_{1\leqslant i<j\leqslant p}g_N^{(2)}(x_i-x_j).\tag{6-5}$$

This is a sensible approximation in the case of low-density $\rho=N/V\ll 1$. Indeed, in this regime, one might expect ψ_N to be approximately a *Bijl–Dingle–Jastrow function*,

$$\psi_N(x_1,\dots,x_N)\approx\prod_{1\leqslant i<j\leqslant N}e^{-\phi(x_i-x_j)}\tag{6-6}$$

for some appropriately chosen real function ϕ. Thus, ψ_N is approximated by the *partition function* of a classical statistical mechanical model of particles interacting via the pair-potential ϕ. In this setting, $g_N^{(p)}$ is the *p-point correlation function* of the canonical Gibbs distribution of this model. When (6-5) holds asymptotically as the particles move away from each other (remember, the density is low), the statistical mechanics system is said to satisfy the *clustering property*. There is a long literature on proving the clustering property for a large class of potentials ϕ; see, among many others, [Ruelle 1969; Gallavotti 1999; Pulvirenti and Tsagkarogiannis 2012].

Assuming the clustering property for the potential ϕ, the assumption (6-5) does not seem far fetched. This product structure leads to an equation for $g_N^{(2)}$. At this stage, one takes the thermodynamic limit: $N\to\infty$ and $\rho=N/V$ fixed. There are some subtleties to taking this limit, which are explained in [Lieb 1963]. Defining $u:=1-g_\infty^{(2)}$, the equation for u is [Lieb 1963, (3.29)]. After a few extra reasonable approximations, this equation reduces to (1-1). The equation for the energy (1-2) is simply the $N\to\infty$ limit of (6-3).

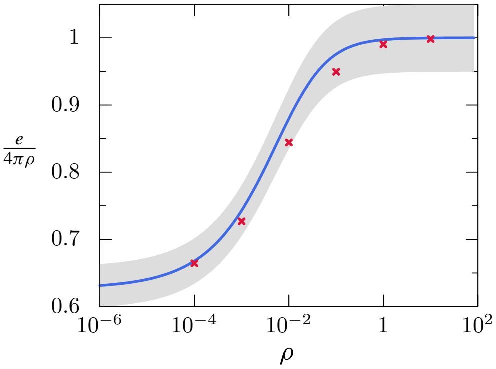


Figure 1. Plot of $e/(4\pi\rho)$ as a function of ρ on a log scale. The potential is $\mathcal{V}(r) = e^{-r}$, in which case the scattering length is $a \approx 1.25$. The solid curve is the energy computed from the simple equation (1-1)–(1-2), and the discrete points are the values of the energy of the Bose gas computed by M. Holzmann using a Monte Carlo algorithm. The gray area corresponds to a 5% error on the value of the energy. At low densities, we recover the Lenz asymptote $e/(4\pi\rho) \sim a/2$ and at high densities, we recover $e/(4\pi\rho) \sim 1$. The difference between the Monte Carlo simulation and the solution of the simple equation is smaller than 5%.

In particular, u is related to the correlation function $g^{(2)}$ of the Bose gas. The condition (1-6) that $u(x) \leqslant 1$ is necessary to ensure that $g^{(2)}(x) \geqslant 0$. However, $u(x) \geqslant 0$ is not a physical requirement, as $g^{(2)}(x)$ could, in principle, be > 1 for some x.

6B. ***Numerical comparison.*** One of the motivations for studying the simple equation is that it provides a simple tool to approximate the ground state energy of the Bose gas. In [Lieb and Liniger 1964], it was found that in one dimension the simple equation gives a value for the energy that differs from the Bose gas ground state energy by at most 69% (a more complete form of the equation yields an even better result with a maximal error of 19%). In one dimension, the difference is larger at high density.

In three dimensions, by Theorem 1.4, the simple equation predicts the correct low-density asymptote as the Bose gas. This is a not so surprising, since the derivation of the simple equation from the ground state equation of the Bose gas sketched above seems somewhat sensible when the density is low. However, when the density is high, at least in the case in which the potential has a nonnegative Fourier transform, the simple equation also yields the same asymptote as the Bose gas. In fact, considering the case

$$\mathcal{V}(x) = e^{-|x|} \tag{6-7}$$

(which has a positive Fourier transform), we compared the ground state energy of the simple equation with values from a Monte Carlo simulation of the Bose gas computed by M. Holzmann (work in preparation), to whom we are most grateful for sharing his unpublished work. The comparison is in Figure 1, in which we found that the maximal error made by the simple equation, over the *entire range of densities*, is 5%! This is a promising result, which we will investigate in more depth and with more rigor in a later publication.

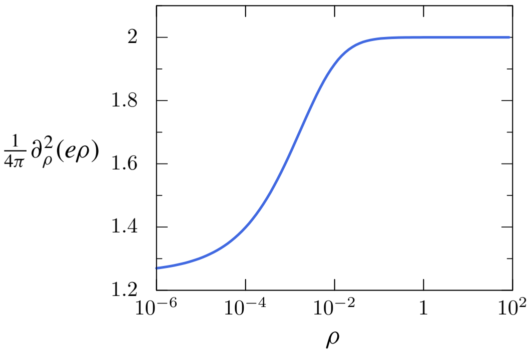


Figure 2. Numerical evaluation of $(1/(4\pi))\partial_\rho^2(\rho e)$ for $\mathcal{V}(r) = e^{-r}$. The asymptotic values are $a \approx 1.25$ for $\rho \to 0$ and 2 for $\rho \to \infty$. This second derivative seems to be clearly positive, so ρe appears to be convex.

7. Open problems and conjectures

7A. ***Monotonicity.*** An important open problem is to show that $e \mapsto \rho(e)$ is an increasing function. If the solution of the simple equation is in any way related to the ground state wave function of the Bose gas, then this should hold: if the density increases, the energy should increase. In addition, it would enable us to prove the uniqueness of the solution of the simple equation with fixed ρ, and might even allow us to generalize our result to potentials with hard core components, as well as to relax the constraint that $\mathcal{V}$ decays exponentially in Theorem 1.5. By running a few numerical computations, it seems clear that $\rho(e)$ should be increasing; see Figure 1. Using a modified iteration in which ρ is fixed, we have proved that $e\rho(e)$ is strictly monotone increasing in e, but the proof that $\rho(e)$ is as well has eluded us thus far.

7B. ***Convexity.*** Another open problem is to prove that $\rho e(\rho)$ *is a convex function*, or, equivalently, that $1/\rho(e)$ is convex. In a physical setting, one expects $\rho e(\rho)$ to be convex. Indeed if $\rho e =: e_{\mathrm{v}}$ were *not convex*, there would exist $\rho_1 < \rho < \rho_2$ such that $(\rho_1 + \rho_2)/2 = \rho$ and $e_{\mathrm{v}}(\rho_1) + e_{\mathrm{v}}(\rho_2) < 2e_{\mathrm{v}}(\rho)$. Furthermore, e_{v} is the energy per unit volume, and, considering a volume V that is split into two equal halves, we find that a configuration in which one half of the volume holds a density ρ_1 of particles and the other holds ρ_2 would have energy

$$\frac{V}{2}(e_{\mathrm{v}}(\rho_1) + e_{\mathrm{v}}(\rho_2)) < V e_{\mathrm{v}}(\rho). \tag{7-1}$$

Therefore, it would pay to have more particles in one half than in the other, which is unstable. Numerically, it seems quite clear that $\rho e(\rho)$ is convex; see Figure 2.

7C. ***Solution of the full equation.*** The simple equation (1-1) is actually a simplified version of an equation that should approximate the Bose gas more accurately [Lieb 1963]:

$$(-\Delta + \mathcal{V}(x))u(x) = \mathcal{V}(x) - \rho(1 - u(x))(2K(x) - \rho L(x)), \tag{7-2}$$

with

$$K(x) := u * S(x), \quad S(x) := (1-u(x))\mathcal{V}(x), \tag{7-3}$$

$$L(x) := \int u(y)u(z-x)\big(1-u(z)-u(y-x)+\tfrac{1}{2}u(z)u(y-x)\big)S(z-y)\,dy\,dz. \tag{7-4}$$

Note that e appears only as the integral of S; see (1-2). While little is known rigorously about this equation, we have been studying it numerically in collaboration with M. Holzmann, and have found it to be remarkably accurate. These results will be detailed in a future publication.

Acknowledgements

Thanks go to Markus Holzmann for his hard work and skill computing the ground state energy of the Bose gas, and for sharing his results with us. Carlen gratefully acknowledges support through NSF grant DMS-174625. Jauslin gratefully acknowledges support through NSF grant DMS-1802170.

References

[Bogolubov 1947] N. Bogolubov, "On the theory of superfluidity", *Izv. Akad. Nauk Ser. Fiz.* **11** (1947), 77–90. In Russian; translated in *Acad. Sci. USSR J. Phys.* **11** (1947), 23–32. MR

[Carlen et al. 2020] E. A. Carlen, I. Jauslin, E. H. Lieb, and M. P. Loss, "On the convolution inequality $f \geqslant f * f$", preprint, 2020. arXiv

[Dyson 1957] F. J. Dyson, "Ground-state energy of a hard-sphere gas", *Phys. Rev.* **106**:1 (1957), 20–26. Zbl

[Fournais and Solovej 2019] S. Fournais and J. P. Solovej, "The energy of dilute Bose gases", preprint, 2019. arXiv

[Gallavotti 1999] G. Gallavotti, *Statistical mechanics*: *a short treatise*, Springer, 1999. MR Zbl

[Lee et al. 1957] T. D. Lee, K. Huang, and C. N. Yang, "Eigenvalues and eigenfunctions of a Bose system of hard spheres and its low-temperature properties", *Phys. Rev.* (2) **106**:6 (1957), 1135–1145. MR Zbl

[Lenz 1929] W. Lenz, "Die Wellenfunktion und Geschwindigkeitsverteilung des entarteten Gases", *Z. Phys.* **56** (1929), 778–789. Zbl

[Lieb 1963] E. H. Lieb, "Simplified approach to the ground-state energy of an imperfect Bose gas", *Phys. Rev.* **130**:6 (1963), 2518–2528.

[Lieb and Liniger 1964] E. H. Lieb and W. Liniger, "Simplified approach to the ground-state energy of an imperfect Bose gas, III: Application to the one-dimensional model", *Phys. Rev.* (2) **134**:2A (1964), 312–315. MR Zbl

[Lieb and Loss 2001] E. H. Lieb and M. Loss, *Analysis*, 2nd ed., Grad. Studies in Math. **14**, Amer. Math. Soc., Providence, RI, 2001. MR Zbl

[Lieb and Yngvason 1998] E. H. Lieb and J. Yngvason, "Ground state energy of the low density Bose gas", *Phys. Rev. Lett.* **80**:12 (1998), 2504–2507. Zbl

[Lieb and Yngvason 2001] E. H. Lieb and J. Yngvason, "The ground state energy of a dilute two-dimensional Bose gas", *J. Statist. Phys.* **103**:3-4 (2001), 509–526. MR Zbl

[Pulvirenti and Tsagkarogiannis 2012] E. Pulvirenti and D. Tsagkarogiannis, "Cluster expansion in the canonical ensemble", *Comm. Math. Phys.* **316**:2 (2012), 289–306. MR Zbl

[Reed and Simon 1975] M. Reed and B. Simon, *Methods of modern mathematical physics, II*: *Fourier analysis, self-adjointness*, Academic Press, New York, 1975. MR Zbl

[Ruelle 1969] D. Ruelle, *Statistical mechanics*: *rigorous results*, Benjamin, New York, 1969. MR Zbl

[Yau and Yin 2009] H.-T. Yau and J. Yin, "The second order upper bound for the ground energy of a Bose gas", *J. Stat. Phys.* **136**:3 (2009), 453–503. MR Zbl

Received 6 Jan 2020. Revised 26 Feb 2020. Accepted 25 Mar 2020.

ERIC A. CARLEN: carlen@rutgers.edu
Department of Mathematics, Rutgers University, Piscataway, NJ, United States

IAN JAUSLIN: ijauslin@princeton.edu
Department of Physics, Princeton University, Princeton, NJ, United States

ELLIOTT H. LIEB: lieb@princeton.edu
Departments of Mathematics and Physics, Princeton University, Princeton, NJ, United States

PURE and APPLIED
ANALYSIS

Vol. 2, No. 3, 2020
https://doi.org/10.2140/paa.2020.2.685

A MAXIMUM PRINCIPLE FOR A FOURTH-ORDER DIRICHLET PROBLEM ON SMOOTH DOMAINS

Inka Schnieders and Guido Sweers

Our main result is that for any bounded smooth domain $\Omega \subset \mathbb{R}^n$ there exists a positive-weight function w and an interval I such that for $\lambda \in I$ and $\Delta^2 u = \lambda w u + f$ in Ω with $u = \frac{\partial}{\partial \nu} u = 0$ on $\partial\Omega$ the following holds: if f is positive, then u is positive. The proofs are based on the construction of an appropriate weight function w with a corresponding strongly positive eigenfunction and on a converse of the Krein–Rutman theorem. For the Dirichlet bilaplace problem above with $\lambda = 0$ the Boggio–Hadamard conjecture from around 1908 claimed that positivity is preserved on convex 2-dimensional domains and was disproved by counterexamples from Duffin and Garabedian some 40 years later. With $w = 1$ not even the first eigenfunction is in general positive. So by adding a certain weight function our result shows a striking difference: not only is a corresponding eigenfunction positive but also a fourth-order "maximum principle" holds for some range of λ.

1. Introduction

Consider for $\Omega \subset \mathbb{R}^n$ a bounded domain with a smooth boundary $\partial\Omega$ and $\lambda \in \mathbb{R}$ the fourth-order Dirichlet problem

$$\begin{cases} (\Delta^2 - \lambda w)u = f & \text{in } \Omega, \\ u = \dfrac{\partial}{\partial \nu} u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1}$$

with weight function $w > 0$. Here ν is the exterior normal on $\partial\Omega$ and Ω is a domain, whenever it is open and connected. For (1) with $\lambda = 0$ and $\Omega = B$, a ball in $\mathbb{R}^n$, Boggio [1905] constructed explicit Green's functions G_B. Since his Green's functions are positive, one finds for any f for which the corresponding solution is well-defined through $u(x) = \int_B G_B(x, y) f(y)\, dy$ that

$$f \geq 0 \quad \text{implies} \quad u \geq 0,$$

and not only for $\lambda = 0$ but even for λ in some interval. By introducing an appropriate weight w that depends on the domain, we derive such kind of positivity-preserving property (PPP) on general domains for some range of λ. For $\lambda = 0$ and $\Omega \subset \mathbb{R}^2$ (1) is called *the clamped plate problem* [Hadamard 1968a].

Concerning that just-mentioned interval for λ, if (1) is positivity-preserving for $\lambda = 0$ on a domain Ω as above, then by a Krein–Rutman theorem, see [Gazzola et al. 2010, page 63], there is a first and simple eigenvalue $\lambda_1 \in \mathbb{R}^+$ for the biharmonic eigenvalue problem. Moreover, $\rho = \lambda_1^{-1}$ is the spectral radius of

MSC2010: primary 35B50; secondary 35J40, 47B65.
Keywords: maximum principle, fourth-order, weighted Dirichlet bilaplace problem, positivity-preserving, positive eigenfunction.

the corresponding solution operator for $\lambda = 0$ and by a Neumann series expansion [Grunau and Sweers 1998, Proposition 4.1] one finds that PPP holds for all $\lambda \in [0, \lambda_1)$. The eigenfunction φ_1 for λ_1 is of fixed sign, and hence can be chosen positive. For $\partial\Omega$ smooth, the function φ_1 is then even strongly positive in the sense that for some $c > 0$

$$\varphi_1(x) \geq c\, d(x, \partial\Omega)^2 \quad \text{for all } x \in \Omega. \tag{2}$$

Here $d(\,\cdot\,, \partial\Omega)$ is the distance to the boundary

$$d(x, \partial\Omega) := \inf_{y \in \partial\Omega} |x - y|. \tag{3}$$

In [Schnieders and Sweers 2020] a converse of the Krein–Rutman theorem is shown for (1) with $w = 1$ on arbitrary smooth and bounded domains Ω:

> *If there exists a simple eigenvalue λ_j to the biharmonic eigenvalue problem with the corresponding eigenfunction strongly positive in the sense of (2), then (1) is positivity-preserving for λ in a left neighbourhood of λ_j.*

Although there are domains besides balls for which there exists an eigenfunction that satisfies (2), see [Sweers 2001], for most domains there is no positive eigenfunction. In this article we overcome that restriction by introducing an appropriate weight function w that is positive. With this w we prove the existence of a simple eigenvalue $\lambda_{j,w}$ and a corresponding eigenfunction $\varphi_{j,w}$ for the weighted eigenvalue problem

$$\begin{cases} \Delta^2\varphi = \lambda w \varphi & \text{in } \Omega, \\ \varphi = \dfrac{\partial}{\partial\nu}\varphi = 0 & \text{on } \partial\Omega, \end{cases} \tag{4}$$

where $\varphi_{j,w}$ is strongly positive as in (2). We will do this for arbitrary bounded smooth domains Ω in any dimension. As a consequence and by arguing as in [Schnieders and Sweers 2020], we find for (1) a positivity-preserving property if λ is in a left neighbourhood of $\lambda_{j,w}$.

Remark 1. Although the positive eigenfunction for most Ω will correspond to the first eigenvalue, [Duffin and Shaffer 1952; Coffman et al. 1979] give an example where such an eigenvalue is the third one. See also [Schnieders and Sweers 2020]. So, we suppose that the eigenvalue $\lambda_{j,w}$ is the j-th eigenvalue, where eigenvalues are counted with their multiplicity. Hence $0 < \lambda_{1,w} \leq \lambda_{2,w} \leq \cdots \leq \lambda_{j,w} \leq \cdots \to \infty$.

The precise statements and main result of the present article are presented in the following theorem and corollary:

Theorem 2. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with $\partial\Omega \in C^{4,\gamma}$ for some $\gamma \in (0, 1)$. Then, there exists a strictly positive-weight function $w \in C^{0,\gamma}(\overline{\Omega})$, meaning $\min_{x \in \overline{\Omega}} w(x) > 0$, such that the eigenvalue problem (4) has the simple eigenvalue $\lambda_{j,w} = 1$ with an eigenfunction $\varphi_{j,w} \in C^{4,\gamma}(\overline{\Omega}) \cap C^1_0(\overline{\Omega})$ satisfying $\varphi_{j,w}(x) \geq d(x, \partial\Omega)^2$ for all $x \in \Omega$.*

The existence of a strictly positive weight w with a strongly positive eigenfunction $\varphi_{j,w}$ is established in Section 2, more precisely in Proposition 9 below. Since the corresponding eigenvalue is not necessarily simple, we have to consider an eventual small perturbation of the weight function. In Section 4 we describe a perturbation procedure so that the slightly changed weight is still positive, the eigenfunction remains strongly positive and the corresponding eigenvalue becomes simple.

Remark 3. The generic simplicity of the spectrum for the clamped plate equation with respect to domain was proved in [Ortega and Zuazua 2000; Pereira 2004]. Other results for generic simplicity under perturbations with respect to the coefficients can be found in [Albert 1975; Teytel 1999].

With the w-variant of the main theorem from [Schnieders and Sweers 2020] we find the following positivity-preserving property for the biharmonic Dirichlet problem in (1).

Corollary 4 (PPP). *Let Ω, w and $\lambda_{j,w} = 1$ be as in Theorem 2. Then there is $\lambda_c < \lambda_{j,w}$ such that for $0 \leq f \in L^2(\Omega)$ with f nontrivial and u the solution of (1):*

(1) *If $\lambda \in [\lambda_c, \lambda_{j,w})$, then $u > 0$ in Ω.*

(2) *If $\lambda \in (\lambda_c, \lambda_{j,w})$, then a Hopf type result holds: there exists $c_{f,\lambda} > 0$ such that*

$$u(x) \geq c_{f,\lambda}\, d(x, \partial\Omega)^2 \quad \textit{for all } x \in \Omega.$$

Proof. With the existence of a simple eigenvalue $\lambda_{j,w} = 1$ with a strongly positive eigenfunction $\varphi_{j,w}$ from Theorem 2 one may continue with the estimates in Theorem 16 and find statement (1) for $\lambda \in [\lambda_{j,w} - C_2/C_3, \lambda_{j,w})$ and (2) for $\lambda \in (\lambda_{j,w} - C_2/C_3, \lambda_{j,w})$. □

Remark 5. As already mentioned, the positivity-preserving property does not hold true for the biharmonic Dirichlet problem on general domains Ω,

$$\begin{cases} \Delta^2 u = f & \text{in } \Omega, \\ u = \dfrac{\partial}{\partial \nu} u = 0 & \text{on } \partial\Omega. \end{cases} \tag{5}$$

Hadamard [1968b] reported on discussions with Boggio and conjectured that at least on convex domains (in $\mathbb{R}^2$) there should be a positivity-preserving property for (5). The first, by now well-known counterexample was established by Duffin [1949], who considered the biharmonic Dirichlet problem on an infinitely long strip. Garabedian [1951] showed that the Green's function changes sign in the case that the underlying domain is a sufficiently eccentric ellipse. For a survey see [Sweers 2001]. An interesting family of domains concerning PPP are the limaçons of Pascal. Hadamard calculated an explicit Green's function for those limaçons in [Hadamard 1968a, Supplement] and, as was shown in [Dall'Acqua and Sweers 2005], those functions are positive only when the limaçon is not too far from the disk. Other known examples with PPP for (5) are based on perturbations of Boggio's results [1905] for balls. See [Gazzola et al. 2010, Chapter 6].

One notices that if (5) is not positivity-preserving for a domain Ω and $\lambda = \lambda_c$, with λ_c as described in Corollary 4, then a Hopf principle fails for the solution to (1). Moreover, for $\lambda < \lambda_c$ one expects some negativity close to the boundary since this is the same phenomenon that appears for the limaçons which are close to the cardioid.

Remark 6. When asked about a physical meaning of the weighted problem, we recall that (5) for $n = 2$ is used to model the deviation u of a thin plate due to a force density f that is clamped at its boundary.

The eigenvalues here correspond to resonances due to exterior induced vibrations and the weight w would be a measure for the stiffness of the plate. This stiffness could be x-dependent, although then the corresponding differential equation should be $\Delta(w^{-1}\Delta u) = \lambda u + f$. A second-order term $b\Delta u$ in the equation also appears when modelling a prestressed plate and fixing the horizontal movements at the boundary. The value of b can have either sign, although for reinforced concrete no engineer would like $b > 0$. If we forget about the third-order terms and compensate the second-order term by prestressing appropriately, the present weight produces a plate that is very stiff near the boundary and rather flexible in the interior.

The structure of the paper is as follows. In Section 2 we introduce a specific weight with which we get a positive eigenfunction with corresponding eigenvalue $\lambda = 1$ for the eigenvalue problem in (4). Next in Section 3, we will describe the adapted setting and adjust and expand the results in [Schnieders and Sweers 2020] to the weighted biharmonic problem (1). Finally in Section 4, we prove that by perturbing the initial weight function slightly, we obtain a simple eigenvalue with a positive eigenfunction.

2. Construction of weight and eigenfunction

In this section we will construct an explicit weight function that guarantees the existence of a positive eigenfunction. To this end we suppose that $\Omega \subset \mathbb{R}^n$ is a bounded domain with $\partial\Omega \in C^{4,\gamma}$ for some $\gamma \in (0, 1)$. We start with one special positive combination u, f for (5). Let $\boldsymbol{e} : \overline{\Omega} \to \mathbb{R}$ be the solution of

$$\begin{cases} -\Delta \boldsymbol{e} = 1 & \text{in } \Omega, \\ \quad \boldsymbol{e} = 0 & \text{on } \partial\Omega. \end{cases}$$

It holds that $\boldsymbol{e} \in C^{4,\gamma}(\overline{\Omega})$; see [Gilbarg and Trudinger 1983, Theorem 6.19]. Using the maximum principle for the laplacian, it follows that $\boldsymbol{e} > 0$ in Ω, and with Hopf's boundary point lemma [Gilbarg and Trudinger 1983, Section 3.2] and the mean value theorem, we obtain constants $c_1, c_2 > 0$ such that

$$c_1 d(x) \le \boldsymbol{e}(x) \le c_2 d(x) \quad \text{for all } x \in \Omega, \tag{6}$$

where we let $d(x) := d(x, \partial\Omega)$ from (3). In [Gilbarg and Trudinger 1983, Lemma 14.16] one finds that $d \in C^{4,\gamma}$ near $\partial\Omega$ follows from $\partial\Omega \in C^{4,\gamma}$.

A direct computation shows

$$\boldsymbol{e}^2 = \frac{\partial}{\partial \nu}\boldsymbol{e}^2 = 0$$

on $\partial\Omega$ and

$$\begin{aligned} \Delta^2 \boldsymbol{e}^2 &= 2(-\Delta)\biggl((-\Delta \boldsymbol{e})\boldsymbol{e} - \sum_{i=1}^n \Bigl(\frac{\partial \boldsymbol{e}}{\partial x_i}\Bigr)^2\biggr) = 2(-\Delta)\biggl(\boldsymbol{e} - \sum_{i=1}^n \Bigl(\frac{\partial \boldsymbol{e}}{\partial x_i}\Bigr)^2\biggr) \\ &= 2 + 4\sum_{i=1}^n \Bigl(\frac{\partial \boldsymbol{e}}{\partial x_i}\frac{\partial \Delta \boldsymbol{e}}{\partial x_i}\Bigr) + 4\sum_{i,j=1}^n \Bigl(\frac{\partial^2 \boldsymbol{e}}{\partial x_i \partial x_j}\Bigr)^2 = 2 + 4\sum_{i,j=1}^n \Bigl(\frac{\partial^2 \boldsymbol{e}}{\partial x_i \partial x_j}\Bigr)^2 =: f. \end{aligned} \tag{7}$$

Note that the function $f \in C^{2,\gamma}(\overline{\Omega})$ is strictly positive on $\overline{\Omega}$.

Example 7. For $\Omega = B_R(0)$ we find

$$\boldsymbol{e}(x) = \frac{R^2 - \|x\|^2}{2n} \quad \text{and} \quad f(x) = 2 + \frac{4}{n}.$$

The main idea of the construction of the weighted problem with positive eigenfunction is the following: If we define $\tilde{w} = f/\boldsymbol{e}^2$, then the function $\boldsymbol{e}^2$ would be a solution to

$$\begin{cases} (-\Delta)^2 \boldsymbol{e}^2 = \tilde{w}\boldsymbol{e}^2 & \text{in } \Omega, \\ \boldsymbol{e}^2 = \dfrac{\partial}{\partial \nu}\boldsymbol{e}^2 = 0 & \text{on } \partial\Omega. \end{cases}$$

Hence $\boldsymbol{e}^2$ is a weighted eigenfunction with corresponding eigenvalue $\lambda = 1$ and there is a constant $c > 0$ such that $\boldsymbol{e}^2(x) \geq c\, d(x)^2$ for all $x \in \Omega$. We notice that f is strictly positive and $\boldsymbol{e}^2$ behaves like $d(x)^2$ near the boundary. So the weight function $\tilde{w}$ is unbounded and especially not Hölder-continuous on Ω. In order to deduce estimates for the Green's function and positivity results we will apply a converse of the Krein–Rutman theorem. In Section 3 we need regularity results from Agmon–Douglis–Nirenberg results, and Hölder-continuity of the weight function is necessary.

So, the combination of the positive functions $\boldsymbol{e}^2$ and f is not directly suitable and we need a combination where both functions grow like $d(x)^2$ near the boundary. In order to achieve this we modify f and consider $f_\varepsilon : \overline{\Omega} \to \mathbb{R}$ defined by

$$f_\varepsilon(x) = \chi_\varepsilon(d(x))^2 f(x), \tag{8}$$

where $\varepsilon > 0$ is small enough and $\chi_\varepsilon \in C^\infty(\mathbb{R};\mathbb{R})$ is an ε-sized mollification of the sign-function. A sketch of χ_ε can be found in Figure 1. For ε small one finds $f_\varepsilon \in C^{2,\gamma}(\overline{\Omega})$ and on $\partial\Omega$ that

$$f_\varepsilon = \frac{\partial}{\partial \nu} f_\varepsilon = 0 \quad \text{and} \quad \frac{\partial^2}{\partial \nu^2} f_\varepsilon > 0.$$

Remark 8. The function χ_ε is constructed with the usual mollifiers $\varphi_\varepsilon : \mathbb{R} \to \mathbb{R}$ with support in $[-\varepsilon, \varepsilon]$ and defined by

$$\varphi_\varepsilon(t) = \frac{1}{\varepsilon}\varphi\left(\frac{t}{\varepsilon}\right)$$

and

$$\varphi(t) = \begin{cases} c_m^{-1}\exp\left(-\dfrac{1}{1-t^2}\right) & \text{for } |t| < 1, \\ 0 & \text{for } |t| \geq 1, \end{cases} \qquad \text{with } c_m = \int_{-1}^{1} \exp\left(-\frac{1}{1-s^2}\right) ds.$$

With $\operatorname{sign}(t) = t/|t|$ for $t \neq 0$ we define the function

$$\chi_\varepsilon(t) = (\varphi_\varepsilon * \operatorname{sign})(t) \quad \text{for } t \in \mathbb{R}.$$

Note that $\chi_\varepsilon \in C^\infty(\mathbb{R})$ satisfies

$$\chi_\varepsilon(0) = 0, \quad \chi_\varepsilon'(0) = \frac{2}{c_m e}\varepsilon^{-1}, \quad \text{and} \quad \chi_\varepsilon(t) = 1 \quad \text{for } t > \varepsilon.$$

Moreover

$$\min\left(\frac{t}{\varepsilon}, 1\right) \leq \chi_\varepsilon(t) \leq \min\left(\frac{2t/(c_m e)}{\varepsilon}, 1\right) \quad \text{for } t \geq 0. \tag{9}$$

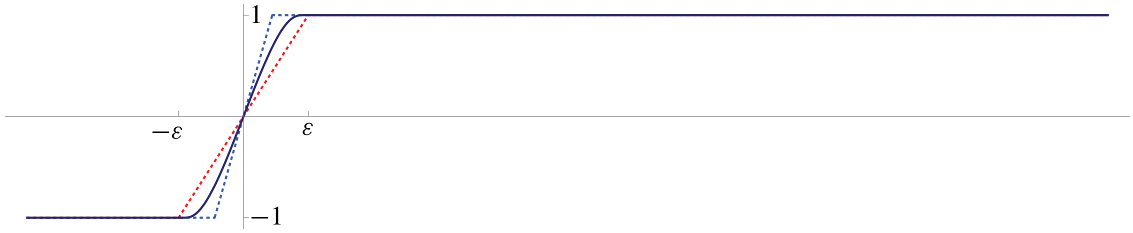


Figure 1. Sketch of χ_ε as mollified sign-function with the estimates from (9).

Letting u_ε be the solution of

$$\begin{cases} \Delta^2 u_\varepsilon = f_\varepsilon & \text{in } \Omega, \\ u_\varepsilon = \dfrac{\partial}{\partial \nu} u_\varepsilon = 0 & \text{on } \partial\Omega, \end{cases} \tag{10}$$

we shall prove with the next proposition that $w_\varepsilon = f_\varepsilon / u_\varepsilon$ is well-defined and that $\varphi = u_\varepsilon$ with $\lambda = 1$ is an appropriate eigenfunction of the eigenvalue problem

$$\begin{cases} \Delta^2 \varphi = \lambda w_\varepsilon \varphi & \text{in } \Omega, \\ \varphi = \dfrac{\partial}{\partial \nu} \varphi = 0 & \text{on } \partial\Omega. \end{cases} \tag{11}$$

Theorem 2 will follow from this result except for the simplicity of the eigenvalue.

Proposition 9. *Let f, f_ε, u_ε be defined in (7), (8) and (10). Then there exists $\varepsilon_0 > 0$ such that for $\varepsilon \in (0, \varepsilon_0)$ the following holds*:

(1) $w_\varepsilon := f_\varepsilon / u_\varepsilon \in C^{0,\gamma}(\bar{\Omega})$ *and* $\min\{w_\varepsilon(x) : x \in \bar{\Omega}\} > 0$.

(2) $\varphi := u_\varepsilon$ *is a strongly positive eigenfunction in the sense of* (2), *with eigenvalue* $\lambda = 1$, *for the weighted eigenvalue problem* (11).

Remark 10. One may guess that generically the eigenvalue $\lambda = 1$ is simple for $\varepsilon \in (0, \varepsilon_0)$. We need, however, that the eigenvalue is simple and not just generically. To obtain this we may fix $\varepsilon = \frac{1}{2}\varepsilon_0$ and proceed by an appropriate perturbation of f_ε for this fixed ε. This is done in Section 4 and yields a simple eigenvalue 1.

Proof. Let $\Omega(\varepsilon) = \{x \in \Omega : d(x) < \varepsilon\}$. One directly checks that for any $p \in [1, \infty)$ it holds that

$$\|f_\varepsilon - f\|_{L^p(\Omega)} \leq \|f\|_{L^\infty(\Omega)} |\Omega(\varepsilon)|^{1/p} \to 0 \quad \text{for } \varepsilon \downarrow 0. \tag{12}$$

Since (12) holds, we find by Agmon–Douglis–Nirenberg [Gazzola et al. 2010, Theorem 2.20] that

$$\|u_\varepsilon - \boldsymbol{e}^2\|_{W^{4,p}(\Omega)} \leq C_{ADN} \|f_\varepsilon - f\|_{L^p(\Omega)} \to 0 \quad \text{for } \varepsilon \downarrow 0. \tag{13}$$

By Sobolev imbedding [Adams and Fournier 2003, Theorem 4.12] and taking $p > n$ it follows that

$$\|u_\varepsilon - \boldsymbol{e}^2\|_{C^3(\bar{\Omega})} \to 0 \quad \text{for } \varepsilon \downarrow 0. \tag{14}$$

Using the mean value theorem, we find

$$\boldsymbol{e}(x)^2 - u_\varepsilon(x) \leq \|u_\varepsilon - \boldsymbol{e}^2\|_{C^2(\bar{\Omega})}\, d(x)^2,$$

and applying (6) we obtain

$$u_\varepsilon(x) \geq \boldsymbol{e}(x)^2 - \|u_\varepsilon - \boldsymbol{e}^2\|_{C^2(\overline{\Omega})}\, d(x)^2 \geq (c_1^2 - \|u_\varepsilon - \boldsymbol{e}^2\|_{C^2(\overline{\Omega})})\, d(x)^2.$$

So there exists $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$ the function u_ε is strongly positive and uniformly in the sense that $\tilde{c}_1 > 0$ exists, not depending on ε, it satisfies

$$u_\varepsilon(x) \geq \tilde{c}_1\, d(x)^2 \quad \text{in } \Omega. \tag{15}$$

Using the mean value theorem, we also find a constant $\tilde{c}_2 > 0$, also independent of ε, such that

$$u_\varepsilon(x) \leq \tilde{c}_2\, d(x)^2 \quad \text{in } \Omega. \tag{16}$$

Hence, we find by (15) an upper bound for f_ε:

$$f_\varepsilon(x) \leq \tilde{c}\varepsilon^{-2} u_\varepsilon(x) \tag{17}$$

for some constant $\tilde{c} > 0$. For $\varepsilon \in (0, \varepsilon_0)$ the function $\varphi_1 = u_\varepsilon$ is a strictly positive eigenfunction of (11) for $\lambda = 1$ and $0 \leq w_\varepsilon = f_\varepsilon / u_\varepsilon$. We also find by applying (7), (9), (16) and (17) that

$$0 < \frac{2}{\tilde{c}_2\, d^2} \min(d^2 \varepsilon^{-2}, 1) \leq \frac{f_\varepsilon}{u_\varepsilon} = w_\varepsilon \leq \tilde{c}\varepsilon^{-2} < \infty,$$

so $\min w_\varepsilon > 0$, and since $u_\varepsilon \in C^3(\overline{\Omega})$ and $f_\varepsilon \in C^{2,\gamma}(\overline{\Omega})$, we obtain $w_\varepsilon \in C^{0,\gamma}(\overline{\Omega})$. □

Remark 11. Note that even if one considers a small perturbation of f_ε, respectively w_ε, one obtains a positive eigenfunction with the eigenvalue $\lambda = 1$. For example, by setting $\tilde{f}_\varepsilon(x) = f_\varepsilon(x) + tq(x)$ for $t \in \mathbb{R}$ with $|t|$ small and $q \in C_c^\infty(\Omega)$, one finds

$$\tilde{u}_\varepsilon(x) = u_\varepsilon + t u_q(x),$$

where u_q is the solution to

$$\begin{cases} \Delta^2 u_q = q & \text{in } \Omega, \\ u_q = \dfrac{\partial}{\partial \nu} u_q = 0 & \text{on } \partial\Omega. \end{cases}$$

Hence, also

$$\tilde{w}_\varepsilon = \frac{f_\varepsilon + tq}{u_\varepsilon + t u_q} = w_\varepsilon + \sum_{k=1}^{\infty} t^k (-1)^k \left(\frac{u_q}{u_\varepsilon}\right)^{k-1} \frac{1}{u_\varepsilon}(u_q w_\varepsilon - q)$$

is a real analytic function of t. Analogously to (12)-(15) it follows that $\tilde{u}_\varepsilon \geq 0$ in Ω for sufficiently small $|t|$. We use this fact in Section 4.

3. Adapting to the weight

In this section we will present the new weighted setting and the results from [Schnieders and Sweers 2020] adjusted to the weighted case. The results in that paper depend strongly on the estimates in [Grunau et al. 2011] for the Green's function for $\lambda = 0$, which use the positive function $H_n : \overline{\Omega} \times \overline{\Omega} \to [0, \infty]$

defined by

$$H_n(x,y)=\begin{cases}(d(x)^2d(y)^2)^{1-n/4}\min\left(1,\dfrac{d(x)^2d(y)^2}{|x-y|^4}\right)^{n/4} & \text{for } 1\le n<4,\\ \log\left(1+\dfrac{d(x)^2d(y)^2}{|x-y|^4}\right) & \text{for } n=4,\\ |x-y|^{4-n}\min\left(1,\dfrac{d(x)^2d(y)^2}{|x-y|^4}\right) & \text{for } n>4.\end{cases}\tag{18}$$

The functions H_n give the asymptotic behaviour of the biharmonic Green's function on bounded smooth domains $\Omega\subset\mathbb{R}^n$ besides a rank-1 perturbation.

Notation 12. Throughout the paper:

(1) Calligraphic $\mathcal{H}_n : L^2(\Omega)\to L^2(\Omega)$ denotes the operator defined by

$$(\mathcal{H}_n f)(x)=\int_\Omega H_n(x,y)f(y)\,dy.$$

(2) For $w\in C^{0,\gamma}(\overline{\Omega})$ a positive weight as in Section 2 we write $\tilde{f} := f/w$ for $f\in L^2(\Omega)$ and we let $G_{\lambda,w}(\,\cdot\,,\,\cdot\,)$ denote the Green's function for

$$\begin{cases}(\Delta^2-\lambda w)u=w\,\tilde{f} & \text{in } \Omega,\\ u=\dfrac{\partial}{\partial\nu}u=0 & \text{on } \partial\Omega;\end{cases}\tag{19}$$

that is, $u(x)=\int_\Omega G_{\lambda,w}(x,y)\tilde{f}(y)\,dy$ solves (19) if defined. By $G_{0,1}$ we mean the Green's function for the biharmonic Dirichlet problem (5) and if we write $G_{\lambda,1}$, we consider the Green's function for (19) without a weight function, i.e., $w\equiv 1$.

(3) For $\mathcal{A},\mathcal{B}: L^2(\Omega)\to L^2(\Omega)$ we write $\mathcal{A}\ge\mathcal{B}$ whenever, for all $f\in L^2(\Omega)$ with $f(x)\ge 0$ a.e., it holds that

$$(\mathcal{A}f)(x)\ge(\mathcal{B}f)(x)\text{ a.e.}\tag{20}$$

If $\mathcal{A},\mathcal{B}$ are defined through kernels, i.e., $(\mathcal{A}f)(x)=\int_\Omega A(x,y)f(x)\,dx$, and these kernels are continuous except maybe for the diagonal $x=y$, then $A(x,y)\ge B(x,y)$ for all $x\ne y\in\Omega$ implies (20).

One finds for $G_{0,1}$ the Green's function for (1) with $\lambda=0$ that for c large enough

$$G_{0,1}(x,y)+cd(x)^2d(y)^2\sim H_n(x,y)\quad\text{for all } x,y\in\Omega.$$

In [Schnieders and Sweers 2020] such estimate was first extended to $G_{\lambda,1}$ for any bounded interval in $\mathbb{R}$ below λ_1. In a second step the asymptotic behaviour of the constant c was studied for $\lambda\uparrow\lambda_1$. We have to adapt these results for the weighted problem and can do so by similar lemmata to those in [Schnieders and Sweers 2020, Sections 4–7].

Remark 13. The weight w does not influence the arguments in the proofs of the results in [Schnieders and Sweers 2020]. The results are consequences of estimates for the Green's function and since there exist two constants $c_{w,1}, c_{w,2} > 0$ such that

$$c_{w,1} \leq w(x) \leq c_{w,2} \quad \text{for all } x \in \Omega,$$

we can follow the steps with only adjusted constants.

We will obtain the two following results, which are variations of [Schnieders and Sweers 2020, Theorems 1, 2]:

Theorem 14. *Suppose that $\Omega \subset \mathbb{R}^n$ with $n \geq 2$ is a bounded domain with $\partial\Omega \in C^{4,\gamma}$ for some $\gamma \in (0, 1)$. Suppose $0 < w \in C^{0,\gamma}(\bar{\Omega})$ and let $\{\lambda_{i,w}\}_{i\in\mathbb{N}^+}$ denote the eigenvalues for (4) and take $M, \delta \in \mathbb{R}^+$. Set*

$$I_{M,\delta} = [-M, M] \setminus \bigcup_{i=1}^{\infty} (\lambda_{i,w} - \delta, \lambda_{i,w} + \delta).$$

Let $G_{\lambda,w}$ be the Green's function for (19). Then there are $c_1, c_2, c_3 > 0$, depending on the domain, M, δ and w, such that for all $\lambda \in I_{M,\delta}$ the following estimate holds:

$$c_1\, H_n(x, y) \leq G_{\lambda,w}(x, y) + c_2\, d(x)^2\, d(y)^2 \leq c_3\, H_n(x, y) \quad \textit{for all } x, y \in \Omega. \tag{21}$$

Remark 15. For $w = 1$ and $\lambda = 0$ this result can be found in [Grunau et al. 2011, Theorem 1]. For $w = 1$ and $\lambda \in I = [-M, \lambda_1 - \delta]$ see [Schnieders and Sweers 2020, Theorem 1].

Theorem 16. *Suppose that $\Omega \subset \mathbb{R}^n$ with $n \geq 2$ is a bounded domain with $\partial\Omega \in C^{4,\gamma}$ for some $\gamma \in (0, 1)$. Let $\delta > 0$. Suppose $0 < w \in C^{0,\gamma}(\bar{\Omega})$ and that $\lambda_{j,w}$ is a simple eigenvalue of (4) with the corresponding eigenfunction $\varphi_{j,w}$ strongly positive as in (2). Suppose $I_\delta = [\lambda_{j,w} - \delta, \lambda_{j,w})$ contains no eigenvalue. Let $G_{\lambda,w}$ be the Green's function for (19). Then there exist $C_1, C_2, C_3 > 0$, depending on Ω, δ and w, such that for all $\lambda \in I_\delta$*

$$G_{\lambda,w}(x, y) \geq C_1\, H_n(x, y) + \left(\frac{C_2}{\lambda_{j,w} - \lambda} - C_3\right)\varphi_{j,w}(x)\varphi_{j,w}(y) \quad \textit{for all } x, y \in \Omega. \tag{22}$$

Proof of Theorems 14 and 16. The proofs use the estimate from [Grunau et al. 2011] just as [Schnieders and Sweers 2020] does. Instead of using a Weyl-type asymptotics for the growth rate of eigenvalues, we exploit here regularity results and Sobolev imbeddings.

- We first recall the standard arguments for existence and the relation with corresponding eigenvalues. Let $L^2_w(\Omega)$ denote the Hilbert space $(L^2(\Omega), \langle\cdot\,,\cdot\rangle_{L^2_w})$,

$$\langle u, v\rangle_{L^2_w(\Omega)} := \int_\Omega u(x)v(x)w(x)\,dx,$$

equivalent with the standard inner product since $w \in C^{0,\gamma}(\bar{\Omega})$ satisfies $w > 0$ on $\bar{\Omega}$.

A weak solution to (1) for $f \in L^2(\Omega)$ is defined by $u \in W^{2,2}_0(\Omega)$ such that

$$\int_\Omega (\Delta u \Delta v - \lambda w u v - \tilde{f} w v)\,dx = 0 \quad \text{for all } v \in W^{2,2}_0(\Omega). \tag{23}$$

We obtain that the standard norm on $W_0^{2,2}(\Omega)$ is equivalent to the norm

$$\|u\| := \|\Delta u\|_{L^2(\Omega)} - \lambda\sqrt{\langle u, u\rangle_{L_w^2(\Omega)}} \quad \text{for any } \lambda \leq 0.$$

Hence, by the Riesz representation theorem there exists a solution $u_{\lambda,w}$ to (23) for every $\tilde{f} \in L_w^2(\Omega)$ and $\lambda \leq 0$. The solution operator $\mathcal{G}_{\lambda,w}$, i.e., $u_{\lambda,w} = \mathcal{G}_{\lambda,w}\tilde{f}$ solves (19), is well-defined on $L_w^2(\Omega)$. Using the results by Agmon–Douglis–Nirenberg [Gazzola et al. 2010, Theorems 2.19, 2.10], we find that

$$\mathcal{G}_{\lambda,w} : L_w^2(\Omega) \to W^{4,2}(\Omega) \cap W_0^{2,2}(\Omega)$$

is an isomorphism for $\lambda \leq 0$.

With $\mathcal{I}$ the compact imbedding from $W^{4,2}(\Omega)$ to $L_w^2(\Omega)$, one finds $\mathcal{I} \circ \mathcal{G}_{0,w}$ is compact and it is the inverse operator of $A_w : D(A_w) \subset L_w^2(\Omega) \to L_w^2(\Omega)$ defined by

$$D(A_w) = W^{4,2}(\Omega) \cap W_0^{2,2}(\Omega) \quad \text{with } A_w = \frac{1}{w}\Delta^2.$$

Since $\mathcal{I} \circ \mathcal{G}_{0,w}$ is compact, the spectrum of A_w is discrete and since A_w is self-adjoint and positive, i.e., $\langle A_w u, u\rangle_{L_w^2(\Omega)} = \langle A_1 u, u\rangle_{L^2(\Omega)} > 0$ for $u \neq 0$, the spectrum consists of countably many real eigenvalues $\{\lambda_{i,w}\}_{i\in\mathbb{N}^+}$, with $0 < \lambda_{1,w} \leq \lambda_{2,w} \leq \cdots \to \infty$ and corresponding eigenfunctions $\{\varphi_{i,w}\}_{i\in\mathbb{N}^+}$. The eigenfunctions can be chosen such that they are orthonormal in the norm induced by $\langle\,\cdot\,,\,\cdot\,\rangle_{L_w^2(\Omega)}$. By the Hilbert–Schmidt theorem we then find a complete orthonormal system of eigenfunctions, still denoted by $\{\varphi_{i,w}\}_{i\in\mathbb{N}}$, and such that for $\lambda \notin \{\lambda_{i,w}\}_{i\in\mathbb{N}^+}$ and $\tilde{f} \in L_w^2(\Omega)$

$$\mathcal{G}_{\lambda,w}\tilde{f} = \sum_{i=1}^{\infty} \frac{1}{\lambda_{i,w} - \lambda} \langle \varphi_{i,w}, \tilde{f}\rangle_{L_w^2(\Omega)}\, \varphi_{i,w}.$$

- Next we recall an asymptotic formula for $\mathcal{G}_{\lambda,w}$ that uses $\mathcal{G}_{0,1}$. If $|\lambda| < \lambda_{1,w}$ and $u_{\lambda,w} = \mathcal{G}_{\lambda,w}\tilde{f}$, then also

$$u_{\lambda,w} = \mathcal{G}_{0,w}(\lambda u_{\lambda,w} + \tilde{f}) = \mathcal{G}_{0,1}(\lambda w u_{\lambda,w} + w\,\tilde{f}),$$

which is equivalent to

$$(\mathcal{I} - \lambda\mathcal{G}_{0,1}(w\,\cdot))u_{\lambda,w} = \mathcal{G}_{0,1}(w\tilde{f}),$$

where $\mathcal{G}_{0,1}$ is the solution operator for (5). For $\lambda \in (-\lambda_{1,w}, \lambda_{1,w})$ we may invert $\mathcal{I} - \lambda\mathcal{G}_{0,1}(w\,\cdot)$ and by using a Neumann series we obtain

$$u_{\lambda,w} = \sum_{k=0}^{\infty} \lambda^k (\mathcal{G}_{0,1}(w\,\cdot))^{k+1}\tilde{f}. \tag{24}$$

We can still find a similar expression when $|\lambda| > \lambda_{1,w}$ when we single out the lower eigenfunctions. Let $\lambda_{m,w}$ be the smallest eigenvalue larger than M and we may use for $\lambda \in (-\lambda_{m,w}, \lambda_{m,w}) \setminus \{\lambda_{i,w}\}_{i<m}$ the expression

$$u_{\lambda,w} = \underbrace{\sum_{i=1}^{m} \frac{1}{\lambda_{i,w} - \lambda}\mathcal{P}_i\tilde{f}}_{I} + \sum_{k=0}^{\infty} \lambda^k (\mathcal{G}_{0,1}(w\,\cdot))^{k+1}\mathcal{P}_{m,+}\tilde{f}, \tag{25}$$

with the following orthogonal projections in $L^2_w(\Omega)$:

$$(\mathcal{P}_i v)(x) := \varphi_{i,w}(x) \int_\Omega \varphi_{i,w}(y)\, v(y)\, w(y)\, dy,$$
$$\mathcal{P}_{i,+} := \mathcal{I} - \mathcal{P}_1 - \cdots - \mathcal{P}_i .$$

We may suppose that $\lambda_{m,w} > M \geq \lambda_{j,w}$.

- In order to estimate I in (25) we will use $\mathcal{D} : L^2(\Omega) \to L^2(\Omega)$, defined by

$$(\mathcal{D} f)(x) := d(x)^2 \int_\Omega f(y) d(y)^2\, dy. \tag{26}$$

With the mean value theorem, we get for every $\varphi \in C^2(\bar{\Omega}) \cap C^1_0(\bar{\Omega})$ and all $x \in \Omega$

$$|\varphi(x)| \leq \|\varphi\|_{C^2(\bar{\Omega})} d(x)^2. \tag{27}$$

Since (27) holds for each eigenfunction $\varphi_{i,w}$ there exists $c_i > 0$ such that

$$-c_i \mathcal{D} \leq \mathcal{P}_i \leq c_i \mathcal{D}. \tag{28}$$

- We split the series on the right of (25) into a finite part with singular behaviour *III* and an infinite remainder *II* that can be estimated by $\mathcal{D}$. The splitting for those λ above is as follows:

$$\sum_{k=0}^{\infty} \lambda^k (\mathcal{G}_{0,1}(w\,\cdot))^{k+1} \mathcal{P}_{m,+} \tilde{f} = \underbrace{\sum_{k=2k_n}^{\infty} \lambda^k \mathcal{G}_{0,w}^{k+1} \mathcal{P}_{m,+} \tilde{f}}_{II} + \underbrace{\sum_{k=0}^{2k_n-1} \lambda^k \mathcal{G}_{0,w}^{k+1} \mathcal{P}_{m,+} \tilde{f}}_{III}, \tag{29}$$

where $k_n = \left[\frac{1}{8}(n+4)\right] + 1$.

- This number k_n is determined as follows. With $\partial\Omega \in C^{4,\gamma}$ the regularity results of Agmon–Douglis–Nirenberg state that for all $p \in (1, \infty)$

$$\mathcal{G}_{0,1} : L^p(\Omega) \to W^{4,p}(\Omega) \cap W^{2,p}_0(\Omega),$$
$$\mathcal{G}_{0,1} : C^{0,\gamma}(\bar{\Omega}) \to C^{4,\gamma}(\bar{\Omega}) \cap C^1_0(\bar{\Omega})$$

are bounded operators. Alternating such a regularity result with a Sobolev imbedding [Adams and Fournier 2003, Theorem 4.12],

$$W^{4,p}(\Omega) \hookrightarrow L^q(\Omega) \qquad \text{for } 4 - \frac{n}{p} > -\frac{n}{q},$$
$$W^{4,p}(\Omega) \hookrightarrow C^{k,\gamma}(\bar{\Omega}) \quad \text{for } 4 - \frac{n}{p} > k + \gamma,$$

one finds after $k_n = \left[\frac{1}{8}(n+4)\right] + 1$ iterations that

$$(\mathcal{G}_{0,1}(w\,\cdot))^{k_n} : L^2_w(\Omega) \to W^{4,p}(\Omega) \cap W^{2,p}_0(\Omega)$$

is bounded for some $p > \frac{1}{2}n$ (and $p \geq 2$): there is $c > 0$ such that

$$\|(\mathcal{G}_{0,1}(w\,\cdot))^{k_n} f\|_{W^{4,p}(\Omega)\cap W^{2,p}_0(\Omega)} \leq c \|f\|_{L^2_w(\Omega)} \quad \text{for all } f \in L^2(\Omega). \tag{30}$$

Since $W^{4,p}(\Omega) \cap W_0^{2,p}(\Omega)$ imbeds in $C^2(\bar{\Omega}) \cap C_0^1(\bar{\Omega})$ for $p > \frac{1}{2}n$ there exists $\tilde{c} > 0$ such that for all $\varphi \in W^{4,p}(\Omega) \cap W_0^{2,p}(\Omega)$

$$\sup_{x\in\Omega}\left|\frac{\varphi(x)}{d(x)^2}\right| \le \|\varphi\|_{C^2(\bar{\Omega})} \le \tilde{c}\|\varphi\|_{W^{4,p}(\Omega)\cap W_0^{2,p}(\Omega)}. \tag{31}$$

By combining (30) and (31) we find for $C = c\tilde{c}$ that

$$|(\mathcal{G}_{0,1}(w\,\cdot))^{k_n} f(x)| \le C\|f\|_{L^2_w(\Omega)} d(x)^2 \quad \text{for all } f \in L^2(\Omega). \tag{32}$$

• This number k_n not only allows the estimate in (32) but also allows us to have a dual estimate by working in Sobolev spaces with a negative coefficient. By duality one finds that also

$$(\mathcal{G}_{0,1}(w\,\cdot)^*)^{k_n} : (W^{4,p}(\Omega) \cap W_0^{2,p}(\Omega))^* \to L^2_w(\Omega)$$

is bounded with k_n as above for some $p > \frac{1}{2}n$ (and $p \ge 2$). Therefore we find a constant c such that for all $g \in (W^{4,p}(\Omega) \cap W_0^{2,p}(\Omega))^*$

$$\|(\mathcal{G}_{0,1}(w\,\cdot)^*)^{k_n} g\|_{L^2_w(\Omega)} \le c\|g\|_{(W^{4,p}(\Omega)\cap W_0^{2,p}(\Omega))^*}. \tag{33}$$

Since $p \ge 2$, one has $L^2(\Omega) \subset (W^{4,p}(\Omega) \cap W_0^{2,p}(\Omega))^*$ in the sense that $f \in L^2(\Omega)$ determines a continuous linear mapping on $W^{4,p}(\Omega) \cap W_0^{2,p}(\Omega)$. Indeed, we have $\langle f, \cdot\rangle_{L^2_w(\Omega)} \in (W^{4,p}(\Omega) \cap W_0^{2,p}(\Omega))^*$ for all $f \in L^2(\Omega)$, see [Adams and Fournier 2003, Paragraph 3.13]. For $f \in L^2(\Omega)$ the symmetry of the kernel implies that $\mathcal{G}_{0,1}(w\,\cdot)^* f = \mathcal{G}_{0,1}(w\,f)$. Moreover, using the imbedding in (31) one obtains for $p > \frac{1}{2}n$

$$\begin{aligned}\|f\|_{(W^{4,p}(\Omega)\cap W_0^{2,p}(\Omega))^*} &:= \sup\left\{\int_\Omega f(x)w(x)\varphi(x)\,dx : \varphi \in W^{4,p}(\Omega) \cap W_0^{2,p}(\Omega) \text{ with } \|\varphi\|_{W^{4,p}(\Omega)} \le 1\right\}\\ &\le \tilde{c}\sup\left\{\int_\Omega f(x)w(x)\varphi(x)\,dx : \varphi \in C^2(\bar{\Omega}) \cap C_0^1(\bar{\Omega}) \text{ with } \|\varphi\|_{C^2(\bar{\Omega})} \le 1\right\}.\end{aligned} \tag{34}$$

With (27) and $c_{w,2}$ as in Remark 13 we find that for all $f \in L^2(\Omega)$ and $\varphi \in C^2(\bar{\Omega}) \cap C_0^1(\bar{\Omega})$ with $\|\varphi\|_{C^2(\bar{\Omega})} \le 1$

$$\int_\Omega f(x)w(x)\varphi(x)\,dx \le \int_\Omega |f(x)|\,w(x)\,|\varphi(x)|\,dx \le c_{w,2}\int_\Omega |f(x)|d(x)^2\,dx. \tag{35}$$

Inequality (34) and (35) imply

$$\|f\|_{(W^{4,p}(\Omega)\cap W_0^{2,p}(\Omega))^*} \le \tilde{c}\,c_{w,2}\int |f(x)|d(x)^2\,dx. \tag{36}$$

By combining (33) and (36) we find a constant $C > 0$ such that

$$\|(\mathcal{G}_{0,1}(w\,\cdot))^{k_n} f\|_{L^2_w(\Omega)} \le C\int |f(x)|d(x)^2\,dx \quad \text{for all } f \in L^2(\Omega). \tag{37}$$

- For part *II* in (29) we write

$$\sum_{k=2k_n}^{\infty} \lambda^k \mathcal{G}_{0,w}^{k+1} \mathcal{P}_{m,+} \tilde{f} = \lambda^{2k_n} \mathcal{G}_{0,w}^{k_n} \left(\sum_{k=0}^{\infty} \lambda^k \mathcal{G}_{0,w}^{k+1} \mathcal{P}_{m,+} \right) \mathcal{G}_{0,w}^{k_n} \tilde{f},$$

with the middle series denoting a bounded operator in $L^2_w(\Omega)$. With (32) and (37) we find that $\tilde{c}_m > 0$ exists with

$$\left| \sum_{k=2k_n}^{\infty} \lambda^k \mathcal{G}_{0,w}^{k+1} \mathcal{P}_{m,+} f(x) \right| \le \tilde{c}_m (\mathcal{D}|f|)(x) \quad \text{for all } f \in L^2(\Omega).$$

This means that there is $C_m > 0$ such that

$$-C_m\, \mathcal{D} \le \sum_{k=2k_n}^{\infty} \lambda^k \mathcal{G}_{0,w}^{k+1} \mathcal{P}_{m,+} \le C_m\, \mathcal{D}. \tag{38}$$

- We are left with an estimate for *III* in (29). We refer to [Schnieders and Sweers 2020, Corollary 9, Lemma 10, 11], from which it follows that for each $k \ge 1$ there are $c_{1,k}, c_{2,k}, c_{3,k}, c_{4,k}, c_{5,k} > 0$ such that

$$c_{1,k} \mathcal{D} \le c_{2,k} \mathcal{H}_n^k \le \mathcal{G}_{0,1}^k + c_{3,k} \mathcal{D} \le c_{4,k} \mathcal{H}_n^k \le c_{5,k} \mathcal{H}_n.$$

Since we consider such estimates for only finitely many terms, the additional factor w only results in adapted constants and we again find

$$c_1 \mathcal{H}_n - c_2 \mathcal{D} \le \sum_{k=0}^{2k_n - 1} \lambda^k \mathcal{G}_{0,w}^{k+1} \mathcal{P}_{m,+} \le c_3 \mathcal{H}_n - c_4 \mathcal{D}. \tag{39}$$

- We may wrap up our estimates to finish the proof for both theorems. If $|\lambda| < \lambda_{m,w}$ and $|\lambda - \lambda_{i,w}| > \delta$ for all $i \le m$, then we may combine (38), (28) and (39) to find through the splitting in (25) and (29) that

$$\mathcal{G}_{\lambda,w} \ge C_{1,m} \mathcal{H}_n - C_{2,m} \mathcal{D}.$$

This shows Theorem 14.

For $\lambda \in [\lambda_{j,w} - \delta, \lambda_{j,w})$ we also single out $\mathcal{P}_j$ in *I* and find for those λ uniform constants $C_0, C_1, C_2 \in \mathbb{R}^+$ such that

$$\mathcal{G}_{\lambda,w} \ge C_0 \frac{1}{\lambda_{j,w} - \lambda} \mathcal{P}_j + C_1 \mathcal{H}_n - C_2 \mathcal{D}. \tag{40}$$

Only here we will use that besides $\varphi_{j,w} \in C^{4,\gamma}(\bar{\Omega}) \cap C_0^1(\bar{\Omega})$ this function $\varphi_{j,w}$ is strongly positive and hence there are $c_1, c_2 > 0$ such that

$$c_1 \mathcal{D} \le \mathcal{P}_j \le c_2 \mathcal{D}. \tag{41}$$

The estimate in (22) follows from (40) and (41) and this completes the proof of Theorem 16. □

4. Simplicity of the eigenvalue

If $\lambda_{w_\varepsilon} = 1$ is a simple eigenvalue of (11), we consider problem (1) with $w = w_\varepsilon$ and find a positivity-preserving property for λ in a small left neighbourhood of λ_{w_ε}. If the multiplicity of $\lambda_{w_\varepsilon} = 1$ is greater than or equal to 2, we will show the simplicity of the eigenvalue after a small perturbation of the weight function w_ε.

The perturbations we consider start from the function f_ε defined in (8).

Definition 17. Let f_ε, u_ε be as in Proposition 9 and Remark 10. For $q \in C_c^\infty(\Omega)$ and $t \in \mathbb{R}$ with $|t|$ small, set

$$w_{tq} = \frac{f_\varepsilon + tq}{u_\varepsilon + t\mathcal{G}_{0,1}(q)}, \tag{42}$$

where $\mathcal{G}_{0,1}$ is the solution operator for (5), and define $A(tq) : W_0^{2,2}(\Omega) \cap W^{4,2}(\Omega) \to L^2(\Omega)$ by

$$A(tq) = \Delta^2 - w_{tq}. \tag{43}$$

We will consider the t-dependent eigenvalue problems

$$\mathcal{A}(tq) : \begin{cases} (\Delta^2 - w_{tq})\varphi = \lambda\varphi & \text{in } \Omega, \\ \varphi = \dfrac{\partial}{\partial\nu}\varphi = 0 & \text{on } \partial\Omega. \end{cases} \tag{44}$$

Remark 18. Note that the multiplicity of the eigenvalue $\lambda_{w_\varepsilon} = 1$ for (11) coincides with the multiplicity of $\lambda = 0$ for (44) with $t = 0$, since $w_0 = w_\varepsilon$.

Assuming that $\lambda = 0$ is an eigenvalue of multiplicity $m \geq 2$ for (44) with $t = 0$, one finds by [Kato 1980, Theorem 3.9, Chapter 7] or [Rellich 1969, pages 76–77] the existence of an interval $(-t_0, t_0) \subset \mathbb{R}$ and m real analytic functions

$$t \mapsto (\lambda_{i,t,q}, \varphi_{i,t,q}) : (-t_0, t_0) \to \mathbb{R} \times C_0^1(\bar\Omega) \cap C^{4,\gamma}(\bar\Omega) \quad \text{for } i \in \{1, \dots, m\},$$

with:

(1) $(\lambda_{i,t,q}, \varphi_{i,t,q})$ are pairs of eigenvalues and eigenfunctions for $\mathcal{A}(tq)$ for all $i \in \{1, \dots, m\}$.

(2) $\{\varphi_{i,0,q}\}_{i=1}^m$ is an orthogonal system and so $\{\varphi_{i,t,q}\}_{i=1}^m$ is independent for $|t|$ small.

(3) $\lambda_{i,0,q} = 0$ for all $i \in \{1, \dots, m\}$.

With our construction we may fix the first one by

$$\varphi_{1,t,q} = u_\varepsilon + t\mathcal{G}_{0,1}(q) \tag{45}$$

and find

$$\lambda_{1,t,q} = 0 \quad \text{for all } t \in (-t_0, t_0).$$

We will show that there exists q_1 such that

$$\lambda'_{k,0,q_1} := \left(\frac{\partial}{\partial t}\lambda_{k,t,q_1}\right)_{t=0} \neq 0$$

for at least one $k \in \{2, \dots, m\}$. In that case one finds for some small positive t_1 that $\lambda_{k,t_1,q_1} \neq 0$ and hence that 0 is an eigenvalue of multiplicity at most $m-1$ for $\mathcal{A}(t_1q_1)$. If the multiplicity of the eigenvalue 0 for $\mathcal{A}(t_1q_1)$ is 1, we are done. Otherwise we repeat our arguments for $\mathcal{A}(t_1q_1 + tq)$. After $k \leq m-1$ steps we have found an eigenvalue problem $\mathcal{A}(t_1q_1 + \cdots + t_kq_k)$ having 0 as a simple eigenvalue. The idea of the proof was inspired by [Albert 1975; Teytel 1999].

Lemma 19. *Suppose that* 0 *is an eigenvalue of multiplicity* $m \geq 2$ *for problem* (44) *with* $t = 0$. *Then there exist* $k \in \{2, \dots, m\}$ *and* $q_1 \in C_c^\infty(\Omega)$ *such that*

$$\Big(\frac{\partial}{\partial t}\lambda_{k,t,q_1}\Big)_{|t=0} \neq 0.$$

Proof. Suppose that $\lambda'_{k,0,q} = 0$ for all $k \in \{2, \dots, m\}$ and $q \in C_c^\infty(\Omega)$. Note that $\lambda'_{1,t,q} = 0$ by construction. Differentiation with respect to t of

$$A(tq)\varphi_{k,t,q} = \lambda_{k,tq}\varphi_{k,t,q} \quad \text{for all } k \in \{1, \dots, m\}$$

yields

$$(A(tq) - \lambda_{k,t,q})\frac{\partial}{\partial t}\varphi_{k,t,q} = \Big(\frac{\partial}{\partial t}w_{tq} + \lambda'_{k,t,q}\Big)\varphi_{k,t,q}$$

and setting $t = 0$, we find using (42), (43) and $\lambda'_{k,0,q} = 0$ that

$$A(0)\Big(\frac{\partial}{\partial t}\varphi_{k,t,q}\Big)_{|t=0} = \frac{1}{u_\varepsilon}(q - w_0\mathcal{G}_{0,1}(q))\varphi_{k,0,q}.$$

Hence, we obtain that $(1/u_\varepsilon)(q - w_0\mathcal{G}_{0,1}(q))\varphi_{k,0,q}$ is in the range of $A(0)$ for all $q \in C_c^\infty(\Omega)$. Since every eigenfunction in $\ker(A(0))$ can be written in the form $\sum_{k=1}^m c_k\varphi_{k,0,q}$ and $A(0)$ is self-adjoint, it follows that

$$\frac{1}{u_\varepsilon}(q - w_0\mathcal{G}_{0,1}(q))\psi_1 \perp \ker(A(0)) \quad \text{for all } \psi_1 \in \ker(A(0)),$$

or in other words

$$\int_\Omega \frac{1}{u_\varepsilon}(q - w_0\mathcal{G}_{0,1}(q))\psi_1\,\psi_2\,dx = 0 \quad \text{for all } \psi_1, \psi_2 \in \ker(A(0)).$$

Since $G_{0,1}(x, y) = G_{0,1}(y, x)$, we obtain

$$\begin{aligned} 0 &= \int_\Omega \frac{1}{u_\varepsilon}(q - w_0\mathcal{G}_{0,1}(q))\psi_1\,\psi_2\,dx \\ &= \int_\Omega \Big(q(x) - w_0(x)\int_\Omega G_{0,1}(x, y)q(y)\,dy\Big)\frac{\psi_1(x)\,\psi_2(x)}{u_\varepsilon(x)}\,dx \\ &= \int_\Omega q(x)\bigg(\frac{\psi_1(x)\,\psi_2(x)}{u_\varepsilon(x)} - \mathcal{G}_{0,1}\Big(w_0\frac{\psi_1\,\psi_2}{u_\varepsilon}\Big)(x)\bigg)\,dx \end{aligned} \tag{46}$$

and we can use the fundamental lemma of calculus of variations to find for all $\psi_1, \psi_2 \in \ker(A(0))$ that

$$\frac{\psi_1(x)\,\psi_2(x)}{u_\varepsilon(x)} - \mathcal{G}_{0,1}\Big(w_0\frac{\psi_1\psi_2}{u_\varepsilon}\Big)(x) = 0.$$

So if ψ_1 and ψ_2 are eigenfunctions of $\mathcal{A}(0)$ with $\lambda = 0$ in (44), then also

$$\tilde{\psi}_{1,2} := \frac{\psi_1\,\psi_2}{u_\varepsilon} \tag{47}$$

is an eigenfunction for $\mathcal{A}(0)$ with $\lambda = 0$. This is obvious for $\psi_1 = u_\varepsilon$, since then $\tilde{\psi}_{1,2} = \psi_2$, but it is not to be expected for all $\psi_1, \psi_2 \in \ker(A(0))$. Indeed, we will show that this cannot be true. Therefore fix some eigenfunction $\psi \in \ker(A(0))\backslash\{0\}$ orthogonal to u_ε.

Let $x_0 \in \Omega$ be a point on a nodal line of ψ. Indeed the existence of the nodal line follows since u_ε is positive with ψ orthogonal. Suppose that $\beta_0 \in [1, \infty]$ is the largest constant such that

$$\lim_{x \to x_0} \frac{\psi(x)}{|x - x_0|^\beta} = 0 \quad \text{for all } \beta < \beta_0.$$

Here $\beta_0 \geq 1$ follows from the fact that ψ is differentiable and $\psi(x_0) = 0$. By repeating (47) we find nonzero eigenfunctions $\{\psi_n\}_{n \in \mathbb{N}}$ defined by

$$\psi_n(x) = \left(\frac{\psi(x)}{u_\varepsilon(x)}\right)^n \psi(x)$$

and $\beta_n = (n+1)\beta_0$ is the largest constant in $[1, \infty]$ such that

$$\lim_{x \to x_0} \frac{\psi_n(x)}{|x - x_0|^\beta} = 0 \quad \text{for all } \beta < \beta_n. \tag{48}$$

Since the multiplicity is m, there is $m_0 \leq m$ such that ψ_{m_0} is a linear combination of the previous ones. Since any such linear combination inherits the behaviour as in (48) of the lowest-order term ψ_n, one finds a contradiction for $\beta_0 < \infty$. Hence ψ_n and also $D^\alpha \psi_n$, with $|\alpha| \leq 4$ and $n \geq |\alpha|$, contain a factor ψ and satisfy

$$\lim_{x \to x_0} \frac{D^\alpha \psi_n(x)}{|x - x_0|^\beta} = 0 \quad \text{for all } \beta \in \mathbb{R}. \tag{49}$$

One finds by the unique continuation theorem of [Shirota 1960] that $\psi_n \equiv 0$ for $n \geq 4$ and hence that $\psi \equiv 0$, a contradiction. So there exists $q_1 \in C_c^\infty(\Omega)$ and $k \in \{1, \dots, m\}$ such that $\lambda'_{k,0,q_1} \neq 0$. □

The previous lemma implies:

Corollary 20. *Let ε be fixed as in Remark 10. Then there is $q^* \in C_c^\infty(\Omega)$ such that*

(1) *$w^* = (f_\varepsilon + q^*)/(u_\varepsilon + \mathcal{G}_{0,1}(q^*)) \in C^{0,\gamma}(\bar{\Omega})$ is strictly positive on $\bar{\Omega}$, and*

(2) *$\varphi = u_\varepsilon + \mathcal{G}_{0,1}(q^*)$ is a strongly positive eigenfunction in the sense of (2) for*

$$\begin{cases} (\Delta^2 - w^*)\varphi = \lambda\varphi & \text{in } \Omega, \\ \varphi = \dfrac{\partial}{\partial \nu}\varphi = 0 & \text{on } \partial\Omega, \end{cases}$$

with simple eigenvalue $\lambda = 0$.

Proof. If the multiplicity of the eigenfunction $\varphi = u_\varepsilon$ for the weight function $w = f_\varepsilon / u_\varepsilon$ is $m \geq 2$ we may proceed as in Lemma 19 and find q_1 such that for $t_1 > 0$ small enough, problem $\mathcal{A}(t_1 q_1)$ contains a positive weight and has a positive eigenfunction φ_{1,t_1,q_1} with eigenvalue 0 of multiplicity at most $m - 1$. Repeating the argument now starting with $\mathcal{A}(t_1 q_1)$ as in (43) and considering $\mathcal{A}_1(tq) = \mathcal{A}(t_1 q_1 + tq)$, we may again reduce the multiplicity. After at most $k \leq m - 1$ steps the multiplicity for $\mathcal{A}(q^*)$ with

$$q^* = t_1 q_1 + t_2 q_2 + \cdots + t_k q_k,$$

with $t_1 \gg t_2 \gg \cdots \gg t_k > 0$, has reduced to 1. □

Using this result, the proof of Theorem 2 is complete.

References

[Adams and Fournier 2003] R. A. Adams and J. J. F. Fournier, *Sobolev spaces*, 2nd ed., Pure and Applied Mathematics (Amsterdam) **140**, Elsevier/Academic Press, Amsterdam, 2003. MR Zbl

[Albert 1975] J. H. Albert, "Genericity of simple eigenvalues for elliptic PDE's", *Proc. Amer. Math. Soc.* **48** (1975), 413–418. MR Zbl

[Boggio 1905] T. Boggio, "Sulle funzioni di Green d'ordine m", *Palermo Rend.* **20** (1905), 97–135. JFM

[Coffman et al. 1979] C. V. Coffman, R. J. Duffin, and D. H. Shaffer, "The fundamental mode of vibration of a clamped annular plate is not of one sign", pp. 267–277 in *Constructive approaches to mathematical models* (Pittsburgh, PA, 1978), Academic Press, New York, 1979. MR Zbl

[Dall'Acqua and Sweers 2005] A. Dall'Acqua and G. Sweers, "The clamped-plate equation for the limaçon", *Ann. Mat. Pura Appl.* (4) **184**:3 (2005), 361–374. MR Zbl

[Duffin 1949] R. J. Duffin, "On a question of Hadamard concerning super-biharmonic functions", *J. Math. Physics* **27** (1949), 253–258. MR

[Duffin and Shaffer 1952] R. J. Duffin and D. H. Shaffer, "On the modes of vibration of a ring-shaped plate", p. 652 in "The summer meeting in East Lansing", *Bull. Amer. Math. Soc.*, **28**:6 (1952), 612–669.

[Garabedian 1951] P. R. Garabedian, "A partial differential equation arising in conformal mapping", *Pacific J. Math.* **1** (1951), 485–524. MR Zbl

[Gazzola et al. 2010] F. Gazzola, H.-C. Grunau, and G. Sweers, *Polyharmonic boundary value problems*, Lecture Notes in Mathematics **1991**, Springer, 2010. MR Zbl

[Gilbarg and Trudinger 1983] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, 2nd ed., Grundlehren der Mathematischen Wissenschaften **224**, Springer, 1983. MR Zbl

[Grunau and Sweers 1998] H.-C. Grunau and G. Sweers, "The maximum principle and positive principal eigenfunctions for polyharmonic equations", pp. 163–182 in *Reaction diffusion systems* (Trieste, 1995), edited by G. Caristi and E. Mitidieri, Lecture Notes in Pure and Appl. Math. **194**, Dekker, New York, 1998. MR Zbl

[Grunau et al. 2011] H.-C. Grunau, F. Robert, and G. Sweers, "Optimal estimates from below for biharmonic Green functions", *Proc. Amer. Math. Soc.* **139**:6 (2011), 2151–2161. MR Zbl

[Hadamard 1968a] J. Hadamard, "Mémoire sur le problème d'analyse relatif à l'équilibrie des plaques élastiques encastrées", pp. 515–641 in *Œuvres de Jacques Hadamard, Tome II*, Éditions du Centre National de la Recherche Scientifique, Paris, 1968.

[Hadamard 1968b] J. Hadamard, *Sur certain cas intéressants du problème biharmonic*, Éditions du Centre National de la Recherche Scientifique, Paris, 1968.

[Kato 1980] T. Kato, *Perturbation theory for linear operators*, Grundlehren Math. Wissenschaften **132**, Springer, 1980. Zbl

[Ortega and Zuazua 2000] J. H. Ortega and E. Zuazua, "Generic simplicity of the spectrum and stabilization for a plate equation", *SIAM J. Control Optim.* **39**:5 (2000), 1585–1614. Addendum in **42**:5 (2003), 1905–1910. MR Zbl

[Pereira 2004] M. C. Pereira, "Generic simplicity of eigenvalues for a Dirichlet problem of the bilaplacian operator", *Electron. J. Differential Equations* **2004** (2004), art. id. 114. MR Zbl

[Rellich 1969] F. Rellich, *Perturbation theory of eigenvalue problems*, Gordon and Breach Science Publishers, New York, 1969. MR Zbl

[Schnieders and Sweers 2020] I. Schnieders and G. Sweers, "A biharmonic converse to Krein–Rutman: a maximum principle near a positive eigenfunction", *Positivity* **24**:3 (2020), 677–710. MR Zbl

[Shirota 1960] T. Shirota, "A remark on the unique continuation theorem for certain fourth order elliptic equations", *Proc. Japan Acad.* **36** (1960), 571–573. MR Zbl

[Sweers 2001] G. Sweers, "When is the first eigenfunction for the clamped plate equation of fixed sign?", pp. 285–296 in *Proceedings of the USA-Chile Workshop on Nonlinear Analysis* (Viña del Mar-Valparaiso, 2000), edited by R. Manasevich and P. Rabinowitz, Electron. J. Differ. Equ. Conf. **6**, Southwest Texas State Univ., San Marcos, TX, 2001. MR Zbl

[Teytel 1999] M. Teytel, "How rare are multiple eigenvalues?", *Comm. Pure Appl. Math.* **52**:8 (1999), 917–934. MR Zbl

Received 13 Jan 2020. Revised 7 May 2020. Accepted 3 Jul 2020.

INKA SCHNIEDERS: ischnied@math.uni-koeln.de
Department Mathematik/Informatik, Universität zu Köln, Köln, Germany

GUIDO SWEERS: gsweers@math.uni-koeln.de
Department Mathematik/Informatik, Universität zu Köln, Köln, Germany

PURE and APPLIED ANALYSIS

Vol. 2, No. 3, 2020

https://doi.org/10.2140/paa.2020.2.703

STRUCTURAL ANALYSIS OF AN L-INFINITY VARIATIONAL PROBLEM AND RELATIONS TO DISTANCE FUNCTIONS

LEON BUNGERT, YURY KOROLEV AND MARTIN BURGER

We analyse the functional $\mathcal{J}(u) = \|\nabla u\|_\infty$ defined on Lipschitz functions with homogeneous Dirichlet boundary conditions. Our analysis is performed directly on the functional without the need to approximate with smooth p-norms. We prove that its ground states coincide with multiples of the distance function to the boundary of the domain. Furthermore, we compute the L^2-subdifferential of $\mathcal{J}$ and characterize the distance function as the unique nonnegative eigenfunction of the subdifferential operator. We also study properties of general eigenfunctions, in particular their nodal sets. Furthermore, we prove that the distance function can be computed as the asymptotic profile of the gradient flow of $\mathcal{J}$ and construct analytic solutions of fast marching type. In addition, we give a geometric characterization of the extreme points of the unit ball of $\mathcal{J}$.

Finally, we transfer many of these results to a discrete version of the functional defined on a finite weighted graph. Here, we analyze properties of distance functions on graphs and their gradients. The main difference between the continuum and discrete setting is that the distance function is not the unique nonnegative eigenfunction on a graph.

1. Introduction

1A. ***Eigenvalue problems associated to Rayleigh quotients.*** Eigenvalue problems are a very old tool in mathematics with a long list of theoretical and practical applications. In particular, nonlinear eigenvalue problems have become increasingly popular in the last decades due to their challenging mathematical properties and their wide range of theoretical and practical applications. A special class of nonlinear

MSC2010: primary 35P30, 47J10, 47J35, 49R05; secondary 05C12, 26A16.

Keywords: distance functions, nonlinear eigenfunctions, extreme points, gradient flows, weighted graphs.

eigenvalue problems are those which arise from a variational principle, like the minimization of a Rayleigh quotient

$$\frac{J(u)}{H(u)} \to \min, \tag{1-1}$$

where J and H typically are convex functionals which share the same homogeneity. In this abstract setting the eigenvalue problem is often defined by

$$\lambda \partial H(u) \cap \partial J(u) \neq \varnothing, \tag{1-2}$$

where $\lambda = J(u)/H(u)$ denotes the eigenvalue and ∂ stands for the subdifferential. For smooth J and H this is exactly the condition for being a critical point of the Rayleigh quotient. Elements actually minimizing the Rayleigh quotient, and thus having the lowest possible eigenvalue, are referred to as ground states. Obviously, due to the homogeneity of J and H ground states are invariant under multiplication with a scalar. By choosing

$$J(u) = \int_\Omega |\nabla u|^p \,\mathrm{d}x, \quad H(u) = \int_\Omega |u|^p \,\mathrm{d}x, \tag{1-3}$$

one obtains the eigenvalue problem of the p-Laplacian

$$\lambda |u|^{p-2} u = -\operatorname{div}(|\nabla u|^{p-2} \nabla u), \tag{1-4}$$

which has to be complemented with suitable boundary conditions, and is a very well-studied nonlinear eigenvalue problem; see, for instance, [Binding et al. 2006; Kawohl and Lindqvist 2006; Barles 1988; Lê 2006; Kawohl and Novaga 2008]. Interesting but challenging limit cases are $p \to 1$ and $p \to \infty$ since in these cases functionals J and H are nonsmooth and not strictly convex. In particular, this means that there can exist linearly independent ground states. For more details about the 1-Laplacian eigenvalue problem we refer to [Kawohl and Schuricht 2007]; explicit solutions can be found in [Bellettini et al. 2005; Alter et al. 2005]. The infinity-Laplacian eigenvalue equation takes the form

$$0 = \begin{cases} \min(|\nabla u| - \lambda u, -\Delta_\infty u), & u > 0, \\ -\Delta_\infty u, & u = 0, \\ \max(-|\nabla u| - \lambda u, -\Delta_\infty u), & u < 0, \end{cases} \tag{1-5}$$

which has to be understood in the viscosity sense. Typically, the problem is complemented with homogeneous Dirichlet conditions. We refer to [Juutinen et al. 1999; 2001; Yu 2007] for more details. Positive solutions of (1-5) on a domain Ω are called infinity ground states and indeed they minimize the Rayleigh quotient

$$u \mapsto \frac{\|\nabla u\|_\infty}{\|u\|_\infty} \tag{1-6}$$

among all functions $u \in W^{1,\infty}(\Omega)$ that vanish on the boundary $\partial\Omega$. However, minimizers of (1-6) are far from being unique up to scalar multiplication. In particular, the distance function $x \mapsto \operatorname{dist}(x, \partial\Omega)$ is always a minimizer of (1-6) but not necessarily a solution of (1-5). Furthermore, solutions of (1-5) are not unique [Hynd et al. 2013]. The infinity-Laplacian eigenvalue problem falls under the scope of L^∞-variational problems, which have been an active field of research, with the main contributions being due to Aronsson [2004]. One big challenge with these problems is that the involved subdifferentials lie in a space of measures and not in a function space.

1B. ***Structure of regularizers.*** From an application point of view, eigenvalue problems of the form (1-2) are interesting since they allow one to study the structural properties of the functional J, if it is interpreted as regularization functional. For instance, in the case of $J:\mathcal{H}\to\mathbb{R}\cup\{\infty\}$ being defined on a Hilbert space $\mathcal{H}$, and $H(\,\cdot\,)=\|\cdot\|_{\mathcal{H}}$ coinciding with its norm, it holds that eigenfunctions f are precisely the separated variable solutions to the gradient flow

$$\begin{cases} u'(t)+\partial J(u(t))\ni 0,\\ u(0)=f.\end{cases}\tag{1-7}$$

In this case the solution of (1-7) has the form $u(t)=a(t)f$, where function $a(t)$ depends on the homogeneity of J; see [Bungert and Burger 2020; Bungert et al. 2019a; Burger et al. 2016a; Cohen and Gilboa 2020]. If J is one-homogeneous and f is an eigenfunction, then this separated variable solution also solves the variational regularization problem

$$\tfrac{1}{2}\|u-f\|_{\mathcal{H}}^2+tJ(u).\tag{1-8}$$

Recent results for general homogeneous functionals [Bungert and Burger 2020; Bungert et al. 2019a] showed that also for general data f, the gradient flow (1-7) behaves like a separated variable solution asymptotically. Under some conditions it was shown that asymptotic profiles of (1-7) are eigenfunctions, meaning

$$\lim_{t\to\infty}\frac{u(t)}{\|u(t)\|_{\mathcal{H}}}=w,\qquad \lim_{t\to\infty}\frac{J(u(t))}{\|u(t)\|_{\mathcal{H}}}=\lambda,\qquad \lambda w\in\partial J(w).\tag{1-9}$$

Subsuming these results, one can say that eigenfunctions to some extent describe which structures are preserved by regularization methods like (1-7) or (1-8). For example, in the case of J being the total variation, it is well known that a large class of eigenfunctions are given by so-called calibrable sets [Alter et al. 2005], which provides an explanation of the staircasing effect in total variation regularization [Burger and Osher 2013]. Furthermore, the study of regularizers through their eigenfunction has sparked applications in image processing, as for instance in [Gilboa 2014; Benning et al. 2017].

An alternative way to study structural properties of regularizers is through the extreme points of their unit ball, where the extreme points of a convex set C in a vector space are given by

$$\operatorname{extr}(C):=\{u\in C:\text{there do not exist } v\neq w\in C,\ \lambda\in(0,1)\text{ such that } u=\lambda v+(1-\lambda)w\}.\tag{1-10}$$

So-called representer theorems study qualitative properties of solutions to the optimization problems

$$u^*\in\operatorname*{argmin}_{u\in\mathcal{X}} J(u):Au=f,\tag{1-11a}$$

$$u^*\in\operatorname*{argmin}_{u\in\mathcal{X}} F(Au)+J(u),\tag{1-11b}$$

where $\mathcal{X}$ is a Banach space and $A:\mathcal{X}\to\mathcal{H}$ is a linear operator mapping into a *finite-dimensional* Hilbert space. The functionals J and F are convex regularization and data fitting functionals, respectively. Recent results [Bredies and Carioni 2020; Boyer et al. 2019; Unser 2019] show that in this case there exists a minimizer u^* of (1-11) which can essentially be expressed as finite linear combination of extreme points in the unit ball of J, meaning

$$u^* = n + \sum_{i=1}^{k} c_i u_i, \tag{1-12}$$

where $n \in \mathcal{N}(J)$ denotes an element in the null-space of J, (c_i) are real numbers, and $(u_i) \subset \operatorname{extr}(B_J)$ are extreme points of the unit ball $B_J = \{u \in \mathcal{X} : J(u) \leq 1\}$. Typically, extreme points have interesting geometric properties which they hand down to minimizers of (1-11). If J equals the total variation of a function, for instance, extreme points are given by characteristic functions of so-called simple sets [Bredies and Carioni 2020], which gives yet another explanation for the staircasing phenomenon.

1C. ***Set-up and outline of this paper.*** Let $\Omega \subset \mathbb{R}^n$ be an open and bounded domain and for $1 \leq p \leq \infty$ we let $\|\cdot\|_p$ denote the Lebesgue p-norms of functions or vector fields. We define the function space

$$W_0^{1,\infty}(\Omega) := \{u \in W^{1,\infty}(\Omega) : u = 0 \text{ on } \partial\Omega\}, \tag{1-13}$$

which consists of all Lipschitz continuous functions, vanishing on $\partial\Omega$. In this paper we study the functional

$$\mathcal{J}(u) = \begin{cases} \|\nabla u\|_\infty, & u \in W_0^{1,\infty}(\Omega), \\ +\infty, & u \in L^2(\Omega) \setminus W_0^{1,\infty}(\Omega), \end{cases} \tag{1-14}$$

which coincides with the Lipschitz constant if $u \in W_0^{1,\infty}(\Omega)$. We would like to understand its structure in terms of eigenfunctions and extreme points.

Remark 1.1. Although the space $W^{1,\infty}(\Omega)$ only coincides with the Lipschitz functions on Ω if Ω is at least quasiconvex [Heinonen 2005], for the space $W_0^{1,\infty}(\Omega)$ this is always true. Furthermore, $\mathcal{J}(u)$ equals the Lipschitz constant of $u \in W_0^{1,\infty}(\Omega)$. This is due to the fact that functions in $W_0^{1,\infty}(\Omega)$ can be extended by zero to lie in $W^{1,\infty}(\mathbb{R}^n)$, which coincides with the space of all Lipschitz functions due to the convexity of $\mathbb{R}^n$.

Although $\mathcal{J}$ is defined on $L^2(\Omega)$ and hence admits standard Hilbert space subdifferential calculus, it comes with many of the challenges and properties of a pure L^∞-variational problem. The associated Rayleigh quotient is

$$u \mapsto \frac{\mathcal{J}(u)}{\|u\|_2} = \frac{\|\nabla u\|_\infty}{\|u\|_2} \tag{1-15}$$

and admits an easier treatment than the "pure" L^∞ Rayleigh quotient (1-6) due to the presence of the L^2-norm in the denominator. In particular, (1-15) has essentially a unique minimizer, given by the distance function to the boundary of the domain. Note that a similar functional has been studied in [Burger et al. 2016b] and a Rayleigh quotient of mixed L^∞-L^2-type was considered in [Barron and Jensen 2005]. While in the first work the analysis is limited to the one-dimensional case, and in the second work the authors approximate the L^∞-norm with smooth p-norms, our subdifferential techniques work in arbitrary dimension and without approximation. The abstract eigenvalue problem (1-2) associated to $\mathcal{J}$ becomes

$$\lambda \frac{u}{\|u\|_2} \in \partial \mathcal{J}(u). \tag{1-16}$$

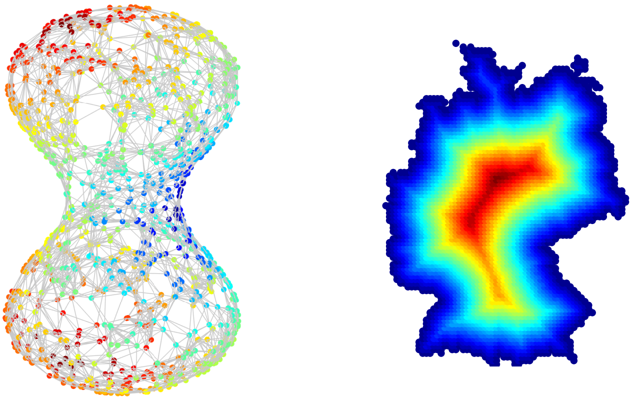


Figure 1. Left: distance function to a point on a discretized manifold. Right: distance function to the boundary of a grid graph.

We also consider a discrete variant of $\mathcal{J}$ defined on a finite weighted graph and transfer most of our continuous results to the discrete setting. Naturally, due to the finite-dimensional character of graphs, the proofs simplify a lot. However, the nonlocal nature of graphs makes the results interesting, nevertheless. In particular, the ground state of this functional is also given by the distance function with respect to the weighted graph distance. From an applied point of view, this interpretation as nonlinear eigenfunction opens the doors for new computational methods for the distance function on graphs. Traditional approaches to compute distance functions on graphs or grids typically rely on level set methods or schemes to solve the eikonal equation $|\nabla u| = 1$; see for instance [Mémoli and Sapiro 2001; Desquesnes et al. 2010; 2013]. Although this paper is mainly of theoretical nature, in Figure 1 we show some distance functions on graphs which were computed using asymptotic profiles of gradient flows in the sense of (1-9); see also [Bungert et al. 2019a; 2019b; Bungert and Burger 2020] for theory and computational results for the 1-Laplacian on graphs, respectively.

This paper is organized as follows. In Section 2 we analyze spectral properties of the functional $\mathcal{J}$. We characterize ground states as distance functions and compute the L^2-subdifferential of in Sections 2A and 2B, respectively. Subsequently, in Section 2C we study the geometrical properties of eigenfunctions. In particular, we prove that under a regularity condition, the nodal set of eigenfunctions has zero Lebesgue measure. Next, in Section 3 we construct an explicit solution to the gradient flow and variational regularization problem of $\mathcal{J}$ which converges to the distance function and possesses level sets that move parallelly to the boundary of the domain. In Section 4 we give a characterization of the extreme points of the unit ball, which gives intuition on the geometrical structure of optimization problems involving $\mathcal{J}$. In Section 5 we transfer most of these results to finite weighted graphs. We prove that ground states are distance functions in Section 5A and study some properties of graph distance functions. In Section 5B we finally collect the graph versions of our results from Sections 2 and 4, hereby skipping most of the proofs since they are elementary, given the proofs in the continuous setting.

We would like to conclude with a remark on how to read this paper. For those readers who are primarily interested in graphs, it is possible to only read Section 5 since it is self-contained in its presentation. Similarly, readers interested mainly in the continuous setting are welcome to only read Section 2 since the results in the graph setting are somewhat similar.

2. Spectral properties

2A. ***Ground states.*** In this section we will investigate the ground states of $\mathcal{J}$, i.e., minimizers of the nonlinear Rayleigh quotient

$$u^* \in \operatorname*{argmin}_{u\in W_0^{1,\infty}(\Omega)} \frac{\mathcal{J}(u)}{\|u\|_2}. \tag{2-1}$$

We prove that — up to multiplicative constants — they coincide with the distance function of the boundary $\partial\Omega$ of the domain which is defined as

$$d(x) := \operatorname{dist}(x, \partial\Omega) := \inf_{y\in\partial\Omega} |x-y|. \tag{2-2}$$

Note that this in particular implies that ground states are unique up to scaling, which is often referred to as simplicity. Indeed, our statement is slightly more general since it holds for minimizers of

$$u^* \in \operatorname*{argmin}_{u\in W_0^{1,\infty}(\Omega)} \frac{\mathcal{J}(u)}{\|u\|_p}, \quad 1 \le p < \infty, \tag{2-3}$$

where (2-1) is a special case when choosing $p = 2$.

Theorem 2.1 (ground states are distance functions). *All solutions u^* to (2-3) are multiples of the distance function to $\partial\Omega$, given by (2-2).*

Proof. By homogeneity, the solutions to (2-3) are given by multiples of the solutions to

$$\hat{u} \in \operatorname{argmax}\{\|u\|_p : \mathcal{J}(u) = 1\} = \operatorname{argmax}\{\|u\|_p : |\nabla u| \le 1 \text{ a.e. in } \Omega,\ u|_{\partial\Omega} = 0\}.$$

From [Zagatti 2014] we infer that — up to global sign — $\hat{u}$ coincides with the unique viscosity solution of the eikonal equation which is given by the distance function (2-2). □

Hence, we have characterized the distance function to the boundary of a set in $\mathbb{R}^n$ — whose properties are well known and have been investigated for decades already — as solution to a nonlinear eigenvalue problem associated to the nonlinear and multivalued operator $\partial\mathcal{J}$. As already mentioned in the Introduction, it is important to notice the difference between our model and infinity Laplacian ground states (see [Juutinen et al. 1999; Barron et al. 2008] for an overview), which are defined as positive viscosity solutions to

$$\min\{|\nabla u| - \Lambda_\infty u, -\Delta_\infty u\} = 0, \tag{2-4}$$

where Δ_∞ denotes the infinity Laplacian. Here, the eigenvalue Λ_∞ is given by

$$\Lambda_\infty := \min_{u\in W_0^{1,\infty}(\Omega)} \frac{\|\nabla u\|_\infty}{\|u\|_\infty} = \frac{1}{\max_{x\in\Omega} \operatorname{dist}(x, \partial\Omega)} \tag{2-5}$$

and every infinity ground state realizes the minimum. However, also the distance function is a minimizer but *not* an infinity ground state, in general [Juutinen et al. 2001], which means that there are minimizers of (2-3) for $p=\infty$ which are not multiples of the distance function.

2B. ***Subdifferential.*** In the following we would like to characterize the L^2-subdifferential of the functional $\mathcal{J}$, which is given by

$$\partial\mathcal{J}(u)=\{\zeta\in L^2(\Omega):\langle\zeta,v\rangle\le\mathcal{J}(v)\text{ for all }v\in L^2(\Omega),\ \langle\zeta,u\rangle=\mathcal{J}(u)\},\quad u\in L^2(\Omega), \tag{2-6}$$

since $\mathcal{J}$ is absolutely one-homogeneous; see [Benning and Burger 2013; Burger et al. 2016a; 2019a; Bungert and Burger 2020], for instance. Note that the L^2-subdifferential of the functionals

$$\mathcal{J}_p(u)=\|\nabla u\|_p,\quad 1<p<\infty, \tag{2-7}$$

is single-valued for $u\in W_0^{1,p}(\Omega)\setminus\{0\}$ and given by

$$\partial\mathcal{J}_p(u)=-\mathcal{J}_p(u)^{1-p}\Delta_p u, \tag{2-8}$$

where $\Delta_p u:=\operatorname{div}(|\nabla u|^{p-2}\nabla u)$ denotes the p-Laplacian. Hence, one could think that by sending $p\to\infty$ one obtains an expression for the subdifferential of $\mathcal{J}$ which involves the ∞-Laplacian. This, however, turns out not to be the case since the competing limits in (2-8) lead to a loss of regularity, as we will see below.

To formulate the subdifferential we define the space

$$H(\operatorname{div};\Omega):=\{q\in L^2(\Omega):\operatorname{div}q\in L^2(\Omega)\} \tag{2-9}$$

of all L^2-vector-fields whose distributional divergence is square-integrable. The space $H(\operatorname{div};\Omega)$ is a Hilbert space when equipped with the inner product

$$\langle q,r\rangle_{H(\operatorname{div};\Omega)}=\int_\Omega[q\cdot r+(\operatorname{div}q)(\operatorname{div}r)]\,dx. \tag{2-10}$$

Remark 2.2. It is well known that vector fields in $H(\operatorname{div};\Omega)$ posses a normal trace and furthermore the space $C^\infty(\overline{\Omega},\mathbb{R}^n)$ of smooth vector fields is dense in $H(\operatorname{div};\Omega)$; see for instance [Girault and Raviart 1986, Chapter 1].

Using that $W_0^{1,\infty}(\Omega)\subset H_0^1(\Omega)$ one obtains the following integration by parts formula, which we will use throughout this work without further references.

Proposition 2.3 (integration by parts). *Let $q\in H(\operatorname{div};\Omega)$ and $u\in W_0^{1,\infty}(\Omega)$. Then it holds*

$$\int_\Omega-(\operatorname{div}q)u\,dx=\int_\Omega q\cdot\nabla u\,dx. \tag{2-11}$$

The following closed subspace of $H(\operatorname{div};\Omega)$ — which consists of all gradient fields with L^2-divergence — will be of great importance:

$$G_0^1(\Omega):=\{\nabla\varphi:\varphi\in H_0^1(\Omega),\ \Delta\varphi\in L^2(\Omega)\}. \tag{2-12}$$

For details on this space, such as Helmholtz decompositions, we refer to [Auchmuty 2006]. Finally, we also introduce the space of vector-valued Radon measures $\mathcal{M}(\Omega, \mathbb{R}^n)$, equipped with the total variation norm $\|\mu\|_{\mathcal{M}(\Omega,\mathbb{R}^n)} := |\mu|(\Omega)$, and the closed subspace

$$\mathcal{N}(\mathrm{div}; \Omega) := \{r \in \mathcal{M}(\Omega, \mathbb{R}^n) : \mathrm{div}\, r = 0\} \tag{2-13}$$

of solenoidal measures. The divergence is understood in the distributional sense, meaning that

$$\int_\Omega \nabla\varphi \cdot \mathrm{d}r = 0 \quad \text{for all } r \in \mathcal{N}(\mathrm{div}; \Omega),\ \varphi \in C_c^\infty(\Omega). \tag{2-14}$$

In order to characterize the subdifferential of $\mathcal{J}$, it is useful to express the functional by duality as

$$\mathcal{J}(u) = \sup\left\{\int_\Omega -(\mathrm{div}\, q)\, u \,\mathrm{d}x : q \in C^\infty(\overline{\Omega}, \mathbb{R}^n),\ \|q\|_1 \le 1\right\}. \tag{2-15}$$

Using this representation we obtain an integral characterization of the subdifferential $\partial\mathcal{J}$ as divergences of sums of regular functions and divergence-free measures. The proof is similar to the characterization of the subdifferential of the total variation in [Bredies and Holler 2016] and can be found in Appendix A.

Proposition 2.4 (integral characterization of the subdifferential). *For $u \in L^2(\Omega)$ it holds*

$$\partial\mathcal{J}(u) = \left\{-\mathrm{div}\, q : q = g + r,\ g \in G_0^1(\Omega),\ r \in \mathcal{N}(\mathrm{div}; \Omega),\ \int_\Omega -(\mathrm{div}\, q)\, u \,\mathrm{d}x = \mathcal{J}(u),\ |q|(\Omega) \le 1\right\}. \tag{2-16}$$

Definition 2.5 (calibrations). Any measure $q \in \mathcal{M}(\Omega, \mathbb{R}^n)$ such that $-\mathrm{div}\, q \in \partial\mathcal{J}(u)$ is called calibration of u.

Remark 2.6 (one space dimension). If $\Omega \subset \mathbb{R}$ is an open interval then $\mathcal{N}(\mathrm{div}; \Omega)$ coincides with constant functions. Hence, in this case calibrations q such that $-\mathrm{div}\, q = -q' \in \partial\mathcal{J}(u)$ are always $H(\mathrm{div})$-functions since the measure part is just a constant.

Having the integral characterization from Proposition 2.4 at hand, we are now interested in explicit forms of calibrations q such that $-\mathrm{div}\, q \in \partial\mathcal{J}(u)$. In the following we fix $0 \ne u \in W_0^{1,\infty}(\Omega)$ and use the notation

$$L := \mathcal{J}(u) < \infty. \tag{2-17}$$

Furthermore, we define the subset of Ω where ∇u attains its maximal modulus as

$$\Omega_{\max} := \{x \in \Omega : |\nabla u(x)| = L\}, \tag{2-18}$$

a set being defined up to a Lebesgue null-set. If we assume for a moment that the calibration q is in $H(\mathrm{div}; \Omega)$, then integrating by parts in (2-16) according to Proposition 2.3 yields

$$\mathcal{J}(u) = \int_\Omega q \cdot \nabla u \,\mathrm{d}x, \tag{2-19}$$

which suggests that a possible calibration is given by

$$q(x) := \begin{cases} \dfrac{\nabla u(x)}{L} \dfrac{1}{|\Omega_{\max}|}, & x \in \Omega_{\max}, \\ 0, & \text{else}. \end{cases} \tag{2-20}$$

However, it is obvious from such a choice of q that $\operatorname{div} q \notin L^2(\Omega)$, in general. As already mentioned, an alternative attempt to characterize the subdifferential of $\mathcal{J}$ could be to send p to infinity in (2-8). However, it is straightforward to see that one formally gets

$$\mathcal{J}_p(u)^{1-p}|\nabla u|^{p-2}\nabla u \to q, \quad p \to \infty,$$

where q is again given by (2-20). Hence, also this approach fails to describe the subdifferential of $\mathcal{J}$. Another difficulty comes through the set $\Omega_{\max}$, given by (2-18), which cannot be expected to have any regularity, as the following example shows.

Example 2.7 (structure of $\Omega_{\max}$). In this example we would like to highlight that the structure of the set $\Omega_{\max}$ defined in (2-18) can be highly degenerate. To this end let $\Omega = (0, 1)$ and $F \subset \Omega$ be the middle-fourth, fat Smith–Volterra–Cantor set, which is a closed set with empty interior and positive measure $|F| = \frac{1}{2}$. Furthermore, we set $u(x) = \operatorname{dist}(x, F)$. Then it is straightforward that $\Omega_{\max} = \Omega \setminus F$ is an open set and $\overline{\Omega}_{\max} = \Omega$. In particular, the topological boundary $\partial\Omega_{\max}$ coincides with F and has positive Lebesgue measure. Nevertheless, u has nonempty subdifferential, as we will see.

From (2-19) we can derive yet another regular calibration, given by

$$q(x) = f(x)\nabla u(x), \tag{2-21}$$

where $f(x) \geq 0$, $\operatorname{supp}(f) \subset \overline{\Omega}_{\max}$ and $\|f\|_1 = 1/L$. Expanding $\operatorname{div} q$ yields

$$\operatorname{div} q = \nabla f \cdot \nabla u + f\Delta u, \tag{2-22}$$

where Δu denotes the distributional Laplacian of u. Hence in order to satisfy $\operatorname{div} q \in L^2(\Omega)$, the function f has to be $H^1(\Omega)$ and satisfy $f = 0$, where Δu is singular. The following examples illustrate that this can be achieved very frequently.

Example 2.8 (measure Laplacians). Let us assume that $u \in W_0^{1,\infty}(\Omega)$ is such that Δu is represented by a finite Radon measure. In this case it holds that $|\Delta u| \ll \mathcal{H}^{n-1}$ according to [Chen et al. 2009, Lemma 2.25]. Since $f \in H^1(\Omega)$ can be defined in the sense of traces on $(n-1)$-dimensional sets, one can find a calibration of the form $q = f\nabla u$, where f vanishes on the support of Δu.

Example 2.9 ($\Omega_{\max}$ with nonempty interior). Let $u \in W_0^{1,\infty}(\Omega)$ such that $\Omega_{\max}$ has nonempty interior. Then one can easily find a smooth nonnegative function f supported on some subset of $\Omega_{\max}$ with integral $1/L$. In particular, $q = f\nabla u$ will be a calibration.

An important property of calibrations of the form (2-21) with a suitable function f is that q is not a measure but an $H(\operatorname{div})$-function in this case. In fact, having such regular calibrations is equivalent to having the form (2-21) as the following proposition shows.

Proposition 2.10 (pointwise characterization of regular calibrations). *Let $0 \neq u \in \operatorname{dom}(\mathcal{J})$ and $q \in H(\operatorname{div}; \Omega)$ with $\|q\|_1 = 1$. It holds that $-\operatorname{div} q \in \partial\mathcal{J}(u)$ if and only if $q = 0$ almost everywhere in $\Omega \setminus \Omega_{\max}$, and $q \cdot \nabla u = |q|\,|\nabla u|$ almost everywhere in Ω.*

Proof. Let us show first that $-\operatorname{div} q \in \partial\mathcal{J}(u)$ for q as above. Again we use the notation $\mathcal{J}(u) = L$. Using the assumptions we compute

$$\begin{aligned} L &\geq \int_\Omega q \cdot \nabla u \,\mathrm{d}x = \int_\Omega |q||\nabla u| \,\mathrm{d}x = \int_{\Omega_{\max}} |q||\nabla u| \,\mathrm{d}x \\ &= L \int_{\Omega_{\max}} |q| \,\mathrm{d}x = L. \end{aligned}$$

Hence, equality holds and we infer

$$\int_\Omega -\operatorname{div} q \, u \,\mathrm{d}x = \int_\Omega q \cdot \nabla u \,\mathrm{d}x = L,$$

which shows $-\operatorname{div} q \in \partial\mathcal{J}(u)$ according to (2-16).

Conversely, let us assume that we have $-\operatorname{div} q \in \partial\mathcal{J}(u)$. First, we show that $q = 0$ holds a.e. in $\Omega \setminus \Omega_{\max}$. For any $\varepsilon > 0$ we define the measurable set

$$\Omega_\varepsilon := \{x \in \Omega : |\nabla u(x)| \leq L - \varepsilon\}$$

and compute using (2-19):

$$\begin{aligned} L = \mathcal{J}(u) = \int_\Omega q \cdot \nabla u \,\mathrm{d}x &= \int_{\Omega_\varepsilon} q \cdot \nabla u \,\mathrm{d}x + \int_{\Omega\setminus\Omega_\varepsilon} q \cdot \nabla u \,\mathrm{d}x \\ &\leq (L - \varepsilon) \int_{\Omega_\varepsilon} |q| \,\mathrm{d}x + L \int_{\Omega\setminus\Omega_\varepsilon} |q| \,\mathrm{d}x \\ &= L - \varepsilon \int_{\Omega_\varepsilon} |q| \,\mathrm{d}x. \end{aligned}$$

This inequality implies that $q = 0$ a.e. on Ω_ε and letting $\varepsilon \searrow 0$ we obtain from the continuity of the Lebesgue measure on nested sets that $q = 0$ a.e. on $\Omega \setminus \Omega_{\max}$.

Now we show that q is parallel to ∇u. To this end we redefine the set

$$\Omega_\varepsilon := \{x \in \Omega : q(x) \cdot \nabla u(x) \leq (1 - \varepsilon)|q(x)||\nabla u(x)|,\ |q(x)||\nabla u(x)| \geq \varepsilon\}$$

for $\varepsilon > 0$ and obtain with a computation similar to that above that

$$L \leq L - \varepsilon \int_{\Omega_\varepsilon} |q||\nabla u| \,\mathrm{d}x,$$

which implies

$$0 = \int_{\Omega_\varepsilon} |q||\nabla u| \,\mathrm{d}x \geq |\Omega_\varepsilon|\varepsilon.$$

This is only possible if $|\Omega_\varepsilon| = 0$ and since the sets Ω_ε are also nested we again infer from the continuity of the Lebesgue measure that

$$\begin{aligned} 0 = \Big|\bigcup_{\varepsilon>0} \Omega_\varepsilon\Big| &= \big|\{x \in \Omega : q(x) \cdot \nabla u(x) < |q(x)||\nabla u(x)|,\ |q(x)||\nabla u(x)| > 0\}\big| \\ &= \big|\Omega \setminus \{x \in \Omega : q(x) \cdot \nabla u(x) = |q(x)||\nabla u(x)|\}\big|, \end{aligned}$$

which shows that q and ∇u are parallel a.e. in Ω. □

2C. ***Eigenfunctions.*** In this section we would like to study geometrical properties of eigenfunctions associated to functional $\mathcal{J}$, meaning functions $u \in W_0^{1,\infty}(\Omega)$ that satisfy

$$\lambda u \in \partial\mathcal{J}(u) \tag{2-23}$$

for some $\lambda > 0$. In particular, we study their nodal sets

$$N(u) = \{x \in \Omega : u(x) = 0\} \tag{2-24}$$

and the set $\Omega_{\max}$ as defined in (2-18). To this end, for the first two statements we assume the regularity condition that the eigenfunctions u under consideration possess a $H(\mathrm{div})$-calibration q, i.e.,

$$\lambda u = -\operatorname{div} q, \qquad q \in H(\mathrm{div}; \Omega), \ \|q\|_1 = 1, \tag{2-25}$$

which makes Proposition 2.10 applicable. Remember that the existence of $H(\mathrm{div})$-calibrations is ensured in many cases (see Remark 2.6, Examples 2.8, 2.9). Note that the nodal set $N(u)$ is closed due to continuity of u. There are only a few results in the literature which deal with nodal sets of p-Laplacian-type eigenfunctions for $p \neq 2$. In particular, it is not even known whether they have nonempty interior. Even if one assumes them to have empty interior, one can only prove lower bounds for their Hausdorff measure, meaning that nodal sets can in principle be very irregular; see [Weih-Wadman 2019; Kawohl and Horák 2017]. For the infinity-Laplacian there do not seem to be any results on the geometry of nodal sets. Also in our slightly different scenario (2-25), where the operator is $\partial\mathcal{J}$, we cannot fully answer the question. However, we can show that $N(u)$ has zero Lebesgue measure if the eigenfunction is sufficiently regular. Furthermore, we prove that the interior of the nodal set coincides with the complement of $\overline{\Omega}_{\max}$, which informally means that at each point an eigenfunction is either zero or it has maximal gradient.

Proposition 2.11. *Let u satisfy (2-25). Then it holds that*

$$\Omega \setminus \overline{\Omega}_{\max} = \operatorname{inn}(N(u)). \tag{2-26}$$

Furthermore, the set $S := \{x \in \Omega_{\max} : q(x) = 0\}$ has empty interior.

Proof. To avoid trivialities we assume $u \neq 0$, which means $\lambda > 0$. We use the abbreviation $\Omega_0 := \Omega \setminus \overline{\Omega}_{\max}$. Since Ω_0 is open, for any $x_0 \in \Omega_0$ there is $r > 0$ small enough such that $B_r(x_0) \subset \Omega_0$. Hence, it holds

$$\lambda \int_{B_r(x_0)} u^2 \,\mathrm{d}x = -\int_{B_r(x_0)} u \operatorname{div} q \,\mathrm{d}x = \int_{B_r(x_0)} q \cdot \nabla u \,\mathrm{d}x - \int_{\partial B_r(x_0)} u\, q \cdot \nu \,\mathrm{d}\mathcal{H}^{n-1}(x) = 0,$$

since $q = 0$ a.e. in $\Omega \setminus \Omega_{\max} \supset \Omega_0$ according to Proposition 2.10. This implies $u = 0$ on $B_r(x_0)$ and hence $B_r(x_0) \subset \operatorname{inn}(N(u))$. Since x_0 was arbitrary we obtain $\Omega_0 \subset \operatorname{inn}(N(u))$. For the converse inclusion we take $x_0 \in \operatorname{inn}(N(u))$ and $r > 0$ such that $B_r(x_0) \subset \operatorname{inn}(N(u))$. Then it holds $u = 0$ and $\nabla u = 0$ on $B_r(x_0)$, which implies $\operatorname{inn}(N(u)) \subset \operatorname{inn}(\Omega \setminus \Omega_{\max}) = \Omega \setminus \overline{\Omega}_{\max} = \Omega_0$.

For the second claim, we assume that there is $x_0 \in \Omega_{\max}$ and $r > 0$ such that $B_r(x_0) \subset S$. Then u cannot be constant on $B_r(x_0)$ since otherwise $|\nabla u| = 0$ would hold on $B_r(x_0)$ which contradicts being a subset of S. Hence, using that $\int_{B_r(x_0)} u(x)^2 \,\mathrm{d}x > 0$ and doing precisely the same computation as above, we obtain a contradiction. □

Using this statement we can easily assert that the set $\Omega_{\max}$ has nonempty interior and hence cannot be too degenerate.

Corollary 2.12. *Let u satisfy (2-25). Then $\Omega_{\max}$ has nonempty interior.*

Proof. From Proposition 2.11 we know that $u = 0$ on $\Omega \setminus \overline{\Omega}_{\max}$. If we assume that $\Omega_{\max}$ has empty interior, this implies that $\overline{\Omega}_{\max} = \Omega_{\max}$ and hence $u = 0$ on $\Omega \setminus \Omega_{\max}$. Now u is a continuous function which implies that $u = 0$ on $\overline{\Omega \setminus \Omega_{\max}} = \Omega$, which is a contradiction. □

Proposition 2.13 (nodal set of eigenfunctions with regularity). *Let u satisfy (2-25) and assume that $\{u \neq 0\}$ has a Lipschitz boundary. Then it holds $|N(u)| = 0$.*

Proof. If the nodal set has empty interior it holds $N(u) = \partial\{u \neq 0\}$, which means that $|N(u)| = 0$ since it coincides with a Lipschitz boundary. Hence we just have to deal with the case that $N(u)$ has nonempty interior. We write $\lambda u = -\operatorname{div} q$ with some calibration $q \in H(\operatorname{div}; \Omega)$. Without loss of generality, let us fix a point x_0 in $\partial\{u > 0\} \cap N(u)$ and for $\varepsilon > 0$ we consider $B_\varepsilon^+(x_0) = B_\varepsilon(x_0) \cap \{u > 0\}$. We choose x_0 and $\varepsilon > 0$ such that $B_\varepsilon(x_0) \cap \{u < 0\} = \varnothing$. This is possible due to the continuity of u. From the characterization of the subdifferential in Proposition 2.10 we know that $q = 0$ a.e. in $N(u)$ and since $N(u)$ has nonempty interior, q has vanishing normal trace on $\partial\{u > 0\} \cap B_\varepsilon(x_0)$. This implies

$$0 < \int_{B_\varepsilon^+(x_0)} \lambda u \,\mathrm{d}x = -\int_{B_\varepsilon^+(x_0)} \operatorname{div} q \,\mathrm{d}x = -\int_{\partial B_\varepsilon(x_0) \cap \{u>0\}} q \cdot \nu \,\mathrm{d}x.$$

Now since q is parallel to ∇u for small enough $\varepsilon > 0$ it holds that $q \cdot \nu \geq 0$, which is a contradiction. Hence, $N(u)$ has zero Lebesgue measure. □

Next we show that every nonnegative eigenfunction coincides with a ground state, i.e., is a multiple of the distance function to $\partial\Omega$. Note that this result *does not* require the regularity condition (2-25) but follows from a simple comparison argument.

Proposition 2.14 (uniqueness of nonnegative eigenfunction). *Any nonnegative eigenfunction $u \neq 0$ of $\partial\mathcal{J}$, satisfying $\lambda u \in \partial\mathcal{J}(u)$, is a ground state.*

Proof. Let us assume that we have a nonnegative eigenfunction $u \neq 0$ on Ω which is not a ground state. We can normalize in such a way that $\mathcal{J}(u) = 1$. Furthermore, we let d denote the distance function which is the unique ground state with $\mathcal{J}(d) = 1$ according to Theorem 2.1. Then from [Zagatti 2014] we know that $u \leq d$ holds pointwise almost everywhere in Ω. Much as before we define the set

$$\Omega_\varepsilon := \{x \in \Omega : d(x) > u(x) + \varepsilon,\ u(x) > \varepsilon\}.$$

Since u is an eigenfunction it holds $\lambda\langle u, v\rangle \leq \mathcal{J}(v)$ for all $v \in L^2(\Omega)$, where $\lambda = 1/\|u\|_2^2$. Testing this with $v = d$, using the definition of Ω_ε and the fact that $d \geq u$, we obtain

$$\begin{aligned}
\|u\|_2^2 \geq \langle u, d\rangle &\geq \int_{\Omega_\varepsilon} u(x)(u(x) + \varepsilon) \,\mathrm{d}x + \int_{\Omega\setminus\Omega_\varepsilon} u(x) d(x) \,\mathrm{d}x \\
&\geq \int_\Omega u(x)^2 \,\mathrm{d}x + \varepsilon \int_{\Omega_\varepsilon} u(x) \,\mathrm{d}x \\
&\geq \|u\|_2^2 + \varepsilon^2 |\Omega_\varepsilon|,
\end{aligned}$$

which tells us that $|\Omega_\varepsilon| = 0$. Letting ε tend to zero we infer as before that almost everywhere in Ω it holds $u = d$ or $u = 0$. Since, however both u and d are continuous functions and by assumption $u \neq 0$, we find that $u = d$ holds almost everywhere in Ω. □

Using this uniqueness of nonnegative eigenfunctions together with the results in [Bungert and Burger 2020] we obtain the result that the gradient flow of $\mathcal{J}$ asymptotically converges to the distance function.

Theorem 2.15 (asymptotic profiles). *Let $u(t)$ be the solution of the gradient flow (1-7) with respect to $\mathcal{J}$ and datum $f \geq 0$. Denote the finite extinction time of the flow by T. Then $u(t)/\|u(t)\|_2$ converges strongly in $L^2(\Omega)$ to a multiple of the distance function as $t \nearrow T$.*

Proof. Since $\operatorname{dom}(\mathcal{J}) = W_0^{1,\infty}(\Omega)$ is compactly embedded in $L^2(\Omega)$ we infer from [Bungert and Burger 2020, Theorem 2.5] that $u(t)/\|u(t)\|_2$ has a subsequence which strongly converges to an eigenfunction. Now [Bungert and Burger 2020, Theorem 2.6] implies that the whole sequence converges to a nonnegative eigenfunction. From Proposition 2.14 and Theorem 2.1 we conclude that this eigenfunction has to be a multiple of the distance function. □

Example 2.16 (distance function of the $(n-1)$-sphere). In this example we study the distance function d of the $(n-1)$-sphere $S_{n-1} := \{x \in \mathbb{R}^n : |x| = 1\}$, where we choose $\Omega = B_1(0)$. We already know from Theorem 2.1 that the distance function is an eigenfunction, i.e., $\lambda d = -\operatorname{div} q$, where $\lambda = \mathcal{J}(d)/\|d\|_2^2 = 1/\|d\|_2^2$ and $\|q\|_1 \leq 1$. Furthermore, since q is parallel to ∇u, we can write q as $q = f\nabla u$ with $f \geq 0$. In the following we would like to examine the function f. We claim that in spherical coordinates it holds

$$f(r) = \lambda\left(\frac{r}{n} - \frac{r^2}{n+1}\right).$$

The radial component of the gradient of $d(r) = 1 - r$ is given by $\nabla_r d = d'(r) = -1$ and there is no angular component. Hence, we obtain that the radial component of the calibration vector field $q = f\nabla d$ is given by

$$q_r(r) = \lambda_n\left(\frac{r^2}{n+1} - \frac{r}{n}\right),$$

which implies

$$\begin{aligned} -\operatorname{div}(f(r)\nabla d(r)) &= -\frac{1}{r^{n-1}}\frac{\mathrm{d}}{\mathrm{d}r}(r^{n-1}q_r(r)) \\ &= \lambda\frac{1}{r^{n-1}}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^n}{n} - \frac{r^{n+1}}{n+1}\right) \\ &= \lambda(1-r) = \lambda d(r). \end{aligned}$$

Furthermore, it is straightforward to check that $\|q\|_1 = 1$. Note that the qualitative behavior of f changes with the dimension $n \in \mathbb{N}$. In particular, $f(r)$ attains its maximum for $r = (n+1)/(2n)$, which tends to $\frac{1}{2}$ as the dimension grows. Furthermore, f has roots at $r = 0$ and $r = (n+1)/n$, which tends to 1 from above. Furthermore, the value of $f(1)$ diverges.

Example 2.17 (a basis of one-dimensional eigenfunctions). In this example we construct a set of eigenfunctions on the interval $\Omega = [-1, 1]$ which constitutes a Riesz basis of $L^2(\Omega)$. They disintegrate into

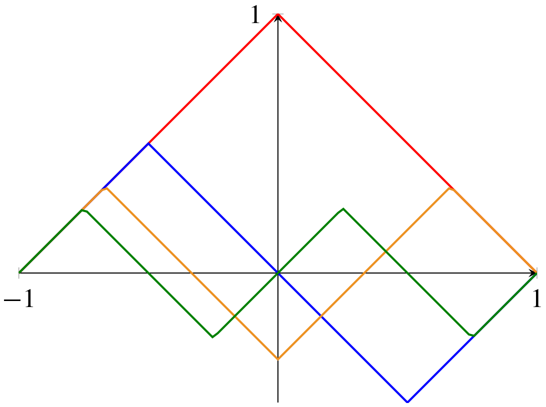


Figure 2. First four eigenfunctions with increasing number of oscillations

odd and even functions with respect to the center of the interval and can be constructed by simple gluing principles. We start with the odd ones, which we denote by $(v_n)_{n\in\mathbb{N}}$. Let $\Omega=\bigcup_{k=1}^{2n}\Omega_k$ be a decomposition of Ω into $2n$ intervals of length $1/n$ such that $\Omega_k\leq\Omega_{k+1}$ holds for all $k=1,\dots,2n-1$. Letting d_k denote the distance function of Ω_k we set

$$u_n|_{\Omega_k}(x)=(-1)^{k+1}d_k(x).$$

Note that all functions u_n satisfy $u_n(0)=0$ and $u(-x)=-u(x)$. Furthermore, it is worth noting that the functions (u_n) form an orthogonal set. This follows directly from the fact that u_n consists of equally many positive and negative distance functions. The eigenvalues of u_n can be easily computed and are given by

$$R(u_n)=\frac{1}{\|u_n\|_2}=\sqrt{\tfrac{3}{2}}\,2n.$$

The even eigenfunctions (v_n) are generated similarly. Here we divide the interval Ω into $2n-1$ intervals Ω_k of length $2/(2n-1)$ such that $\Omega=\bigcup_{k=1}^{2n-1}\Omega_k$ and $\Omega_k\leq\Omega_{k+1}$ holds for all $k=1,\dots,2n-2$. Letting d_k again denote the distance function of Ω_k we set

$$v_n|_{\Omega_k}(x)=(-1)^{k+1}d_k(x).$$

All functions v_n satisfy $v_n(-x)=v_n(x)$ and, in particular, v_1 coincides with the distance function of Ω which is even and a ground state. Note that functions (v_n) are *not* mutually orthogonal. Their eigenvalues are given by

$$R(v_n)=\frac{1}{\|v_n\|_2}=\sqrt{\tfrac{3}{2}}(2n-1).$$

Figure 2 shows the first four eigenfunctions $\{v_1,u_1,v_2,u_2\}$ sorted by eigenvalue. Note that — up to the factor $\sqrt{\frac{3}{2}}$ — the eigenvalues of u_n and v_n precisely count the numbers of peaks or oscillations.

The fact that $\{u_n,v_n:n\in\mathbb{N}\}$ is a Riesz basis of $L^2(\Omega)$ was proven in [Binding et al. 2006].

3. Explicit solution of gradient flow and variational problem

We already know from Theorem 2.15 that the solution of the gradient flow (1-7) with respect to $\mathcal{J}$ asymptotically behaves like the distance function of the domain. In the following, we prove that for sufficiently regular domains and constant initialization, one can compute the solution of the gradient flow analytically. In addition, this solution also solves the variational regularization problem (1-8) associated to $\mathcal{J}$. Notably, this solution exhibits an interesting behavior of its level sets, which is reminiscent of the fast marching algorithm or other level set approaches; see [Sethian 1996; Sussman 1994]. Before we construct these analytic solutions we start with some definitions regarding the kind of domains we consider.

Definition 3.1 (inner parallel body). Let $\Omega \subset \mathbb{R}^n$ be an open set and let $d(x) := \operatorname{dist}(x, \partial\Omega)$ denote the distance function to $\partial\Omega$. Then

$$\Omega_\tau := \{x \in \Omega : d(x) \geq \tau\} \tag{3-1}$$

is called the inner parallel body of Ω with distance $\tau > 0$.

Definition 3.2 (perimeter bound for inner parallel body). We say that Ω admits a perimeter bound for its inner parallel bodies if there are $\tilde{r} > 0$ and $0 < \tilde{\tau} \leq \tilde{r}$ such that

$$P(\Omega_\tau) \geq P(\Omega)\left(1 - \frac{\tau}{\tilde{r}}\right)^{n-1} \quad \text{for all } 0 \leq \tau \leq \tilde{\tau}. \tag{3-2}$$

Example 3.3 (convex domains). According to [Larson 2016] convex domains $\Omega \subset \mathbb{R}^n$ always fulfill a perimeter bound like (3-2) with $\tilde{r} = \tilde{\tau} = r$, where $r = \max_{x\in\Omega} \operatorname{dist}(x, \partial\Omega)$ denotes the in-radius of Ω. Furthermore, if Ω is homothetic to its *form body* then (3-2) becomes an equality. This is the case, for instance, if Ω is a ball or a polytope whose faces are tangential to the largest ball which can be inscribed in Ω.

Example 3.4 (L-shaped domain). Let us consider an L-shaped domain with equal width and height given by $L > 0$ and thickness $\delta \in (0, L)$. For instance, one could set $\Omega := [0, L]^2 \setminus [0, L-\delta]^2 \subset \mathbb{R}^2$. We are interested in whether Ω admits the perimeter bound (3-2). To this end we notice that the perimeter of Ω is given by $P(\Omega) = 4L$ and the perimeter of Ω_τ for $0 \leq \tau \leq \min(L-\delta, \delta/2)$ can be computed as

$$\begin{aligned} P(\Omega_\tau) &= 2(L-2\tau) + 2(\delta-2\tau) + 2(L-\delta-\tau) + \tfrac{1}{4}2\tau\pi \\ &= 4L\left(1 - \tau\frac{20-\pi}{8L}\right) = P(\Omega)\left(1 - \frac{\tau}{\tilde{r}}\right), \end{aligned}$$

where $\tilde{r} = 8L/(20-\pi)$. The number $\tilde{\tau}$ is given by $\tilde{\tau} = \min(L-\delta, \delta/2)$ and satisfies $\tilde{\tau} < \tilde{r}$. Hence, the L-shape admits the perimeter bound (3-2).

Before we turn to the main theorem of this section, which constructs the explicit solution, we have to study the properties of a geometric integral which will appear in the proof.

Lemma 3.5. *Let $\Omega \subset \mathbb{R}^n$ be a domain, $d(x) := \operatorname{dist}(x, \partial\Omega)$ denote the distance function to $\partial\Omega$, and $r := \max_{x\in\Omega} d(x)$ the in-radius of Ω. Then for $k \in \mathbb{N}$, we define the function*

$$I_k(g) := \int_{\Omega\setminus\Omega_{rg}} d(x)^k \,\mathrm{d}x, \qquad 0 \leq g \leq 1. \tag{3-3}$$

- *For all $k \in \mathbb{N}$ it holds that $I_k(0) = 0$ and I_k is monotonously increasing and differentiable with*
$$I_k'(g) = P(\Omega_{rg}) r^{k+1} g^k \quad \textit{for all } 0 < g < 1. \tag{3-4}$$
- *If Ω admits the perimeter bound (3-2) for its inner parallel body, then the function I_2 admits the following estimate for all $0 \le g \le \tilde{\tau}/r$:*
$$I_2(g) \ge \frac{\tilde{r}^3 P(\Omega)}{n} \left\{ \frac{2}{(n+1)(n+2)} \left[1 - \left(1 - \frac{rg}{\tilde{r}}\right)^{n+2} \right] - \frac{2}{n+1} \left(1 - \frac{rg}{\tilde{r}}\right)^{n+1} \frac{rg}{\tilde{r}} - \left(\frac{rg}{\tilde{r}}\right)^2 \left(1 - \frac{rg}{\tilde{r}}\right)^n \right\}. \tag{3-5}$$

Proof. It is trivial that $I_k(0) = 0$ and I_k is monotonously increasing. For showing (3-4) we let $\tilde{g} < g$ and compute using the coarea formula
$$I_k(g) - I_k(\tilde{g}) = \int_{S_{r\tilde{g},rg}} d(x)^k \, \mathrm{d}x = \int_{r\tilde{g}}^{rg} P(\Omega_t) t^k \, \mathrm{d}t.$$

Consequently, we obtain
$$I_k'(g) = \lim_{\tilde{g} \to g} \frac{I_k(g) - I_k(\tilde{g})}{g - \tilde{g}} = r \lim_{\tilde{g} \to g} \frac{1}{rg - r\tilde{g}} \int_{r\tilde{g}}^{rg} P(\Omega_t) t^k \, \mathrm{d}t = r P(\Omega_{rg}) (rg)^k = P(\Omega_{rg}) r^{k+1} g^k.$$

To evaluate $I_2(g)$ we make use of the layer-cake formula, which states that the integral of a nonnegative function $h : \Omega \to \mathbb{R}$ can be computed as
$$\int_\Omega h(x) \, \mathrm{d}x = \int_0^\infty |\{x \in \Omega : h(x) > t\}| \, \mathrm{d}t. \tag{3-6}$$

Let us first estimate the Lebesgue measure of the strip $S_{s,t} := \Omega_s \setminus \Omega_t$, where $s < t$. By using the coarea formula and the perimeter bound (3-2) it holds, for $0 \le s \le t < \tilde{\tau}$,
$$|S_{s,t}| = \int_s^t P(\Omega_\tau) \, \mathrm{d}\tau \ge P(\Omega) \int_s^t \left(1 - \frac{\tau}{\tilde{r}}\right)^{n-1} \mathrm{d}\tau = \frac{\tilde{r} P(\Omega)}{n} \left[\left(1 - \frac{s}{\tilde{r}}\right)^n - \left(1 - \frac{t}{\tilde{r}}\right)^n \right]. \tag{3-7}$$

Letting $h_g(x) := d(x)^2 \chi_{\Omega \setminus \Omega_{rg}}$ for $0 \le g \le \tilde{\tau}/r$ we infer from (3-6) and (3-7)
$$\begin{aligned} I_2(g) &= \int_\Omega h_g(x) \, \mathrm{d}x \\ &= \int_0^{(rg)^2} |\{x \in \Omega : t < h_g(x) < (rg)^2\}| \, \mathrm{d}t \\ &= \int_0^{(rg)^2} |S_{\sqrt{t},rg}| \, \mathrm{d}t \\ &\ge \frac{\tilde{r} P(\Omega)}{n} \int_0^{(rg)^2} \left(1 - \frac{\sqrt{t}}{\tilde{r}}\right)^n - \left(1 - \frac{rg}{\tilde{r}}\right)^n \mathrm{d}t \\ &= \frac{\tilde{r}^3 P(\Omega)}{n} \left\{ \frac{2}{(n+1)(n+2)} \left[1 - \left(1 - \frac{rg}{\tilde{r}}\right)^{n+2} \right] - \frac{2}{n+1} \left(1 - \frac{rg}{\tilde{r}}\right)^{n+1} \frac{rg}{\tilde{r}} - \left(\frac{rg}{\tilde{r}}\right)^2 \left(1 - \frac{rg}{\tilde{r}}\right)^n \right\}, \end{aligned}$$
where we used elementary integration for the last equality. This shows (3-5). □

Theorem 3.6. *Under the conditions of Lemma 3.5 there is $t_* > 0$ such that the initial value problem*

$$\begin{cases} g'(t) = \dfrac{g(t)^2}{I_2(g(t))}, & t > 0, \\ g(0) = 0, \end{cases} \tag{3-8}$$

where I_2 is given by (3-3) for $k = 2$, has a solution for $t \in [0, t_*]$. Furthermore,

$$u(t,x) = \begin{cases} \min\Bigl(\dfrac{1}{g(t)} d(x), r\Bigr), & 0 \le t < t_*, \\ \dfrac{1}{\|d\|_2^2}(\|d\|_2^2 + t_* - t)_+ d(x), & t \ge t_*, \end{cases} \tag{3-9}$$

solves the gradient flow (1-7) with respect to $\mathcal{J}$ and datum $f \equiv r$.

Proof. Note that since d is an eigenfunction of $\partial\mathcal{J}$, it is known that the dynamics for $t \ge t_*$ will linearly shrink the eigenfunction until extinction; see [Bungert et al. 2019a; Burger et al. 2016a], for instance. Hence, we will focus on the initial dynamics and first show that the initial value problem (3-9) has a solution $g(t)$ which persists long enough such that $g(t_*) = 1$ for some $t_* > 0$. Afterwards, we will show that (3-9) solves the gradient flow.

Step 1: First we study the fine behavior of the lower bound in (3-5) as $g \searrow 0$. To this end, one notes that the derivative of the right-hand side in (3-5) with respect to g is given by

$$C\Bigl(\frac{rg}{\tilde{r}}\Bigr)^2\Bigl(1 - \frac{rg}{\tilde{r}}\Bigr)^{n-1}$$

with a positive constant $C = C(n, \Omega) > 0$, which by L'Hôpital's rule shows that

$$\liminf_{g \searrow 0} \frac{I_2(g)}{g^3} > 0.$$

In particular, for the ODE $g'(t) = g(t)^2/I_2(g(t))$ this implies that for small times $t > 0$ the right-hand side is dominated by $1/g(t)$. The fact that the problem

$$\phi'(t) = \frac{1}{\phi(t)}, \qquad \phi(0) = 0,$$

has a solution (namely $\phi(t) = \sqrt{2t}$) implies existence of a solution to (3-8) for small times. Analogously, due to the fact that $I_2(g)$ is bounded from above by the value $I_2(1)$ according to Lemma 3.5, the right-hand side in (3-8) is bounded from below by $g(t)^2/I_2(1)$. Hence, if we fix $t_0 > 0$ in the existence interval of g, it holds for all $t \ge t_0$ in the existence interval that $g(t) \ge \phi(t - t_0)$, where ϕ solves

$$\phi'(t) = \frac{\phi(t)^2}{I_2(1)}, \qquad \phi(0) = g(t_0) > 0.$$

This problem has the blow-up solution

$$\phi(t) = \frac{g(t_0) I_2(1)}{I_2(1) - g(t_0)t}$$

and hence we infer the existence of $t_* > 0$ such that $g(t_*) = 1$.

Step 2: It remains to be shown that (3-9) solves the gradient flow. Obviously, it holds $u(0,x)=r=f(x)$ for all $x\in\Omega$ since $g(0)=0$. Furthermore, we can compute that

$$\partial_t u(t,x)=-\frac{1}{2}\frac{g'(t)}{g(t)^2}d(x)[1-\operatorname{sgn}(d(x)-rg(t))],$$

which yields that for all $0<t<t_*$ we have

$$\langle -\partial_t u(t),u(t)\rangle=\frac{g'(t)}{g(t)^3}\underbrace{\int_{\Omega\setminus\Omega_{rg(t)}}d(x)^2\,\mathrm{d}x}_{=:I_2(g(t))}=\frac{1}{g(t)}=\mathcal{J}(u(t)),$$

using that g solves (3-8). Hence, we have shown $\langle -\partial_t u(t),u(t)\rangle=\mathcal{J}(u(t))$ and it remains to be shown that $\langle -\partial_t u(t),v\rangle\le\mathcal{J}(v)$ holds for all $v\in W_0^{1,\infty}(\Omega)$. We compute using that $g(t)$ solves (3-8):

$$\langle -\partial_t u(t),v\rangle=\frac{g'(t)}{g(t)^2}\int_{\Omega\setminus\Omega_{rg(t)}}d(x)v(x)\,\mathrm{d}x=\frac{1}{I_2(g(t))}\int_{\Omega\setminus\Omega_{rg(t)}}d(x)v(x)\,\mathrm{d}x.$$

For any $x\in\Omega$ we choose $y=y_x\in\partial\Omega$ such that $|x-y_x|=\min_{y\in\partial\Omega}|x-y|=d(x)$. Then using the Lipschitz continuity of v (see Remark 1.1) and $v(y_x)=0$, we obtain

$$|v(x)|=|v(x)-v(x_y)|\le\mathcal{J}(v)d(x).$$

Putting things together we can finish the proof by calculating

$$\langle -\partial_t u(t),v\rangle\le\frac{1}{I_2(g(t))}\int_{\Omega\setminus\Omega_{rg(t)}}d(x)|v(x)|\,\mathrm{d}x\le\frac{\mathcal{J}(v)}{I_2(g(t))}\int_{\Omega\setminus\Omega_{rg(t)}}d(x)^2\,\mathrm{d}x=\mathcal{J}(v),$$

which yields that $-\partial_t u(t)\in\partial\mathcal{J}(u(t))$. □

Corollary 3.7 (motion of level sets). *Under the conditions of Theorem 3.6 the level sets*

$$\Gamma_c(t)=\{x\in\Omega:u(x)=c\}$$

of $u(t)$ at level $c\ge 0$ and time $0\le t\le t_*$ are given by

$$\Gamma_c(t)=\{x\in\Omega:d(x)=cg(t)\},\quad 0\le c<r, \tag{3-10a}$$

$$\Gamma_r(t)=\{x\in\Omega:d(x)\ge rg(t)\}. \tag{3-10b}$$

This means that the level sets are inner parallel sets of $\partial\Omega$ moving with a velocity that is proportional to both the level and function $g'(t)\approx 1/\sqrt{t}$ for small t.

Remark 3.8 (comparison to level set methods). A traditional way to compute distance functions was proposed in [Sussman 1994] and uses the PDE

$$\begin{cases}u(0,x)=f(x), & x\in\mathbb{R}^n,\\ \partial_t u(t,x)+\operatorname{sgn}(f(x))(|\nabla u(t,x)|-1)=0, & (t,x)\in(0,\infty)\times\mathbb{R}^n,\end{cases} \tag{3-11}$$

where the initial datum f fulfills $f > 0$ in Ω, $f < 0$ in $\mathbb{R}^n \setminus \Omega$, and $f = 0$ in $\partial\Omega$. The steady state of this equation solves the eikonal equation $|\nabla u| = 1$ and coincides with the signed distance function of Ω. Similarly, in [Lee et al. 2017] the authors use the PDE

$$\partial_t u(t,x) + |\nabla u(t,x)| = 0 \tag{3-12}$$

for a redistancing procedure that converges to the signed distance function as well. It is straightforward to see that points $x(t)$ in the level sets of the solutions of (3-11) move with the velocity

$$\dot{x}(t) = \operatorname{sgn}(f(x(t)))\frac{|\nabla u(t,x(t))| - 1}{|\nabla u(t,x(t))|}\frac{\nabla u(t,x(t))}{|\nabla u(t,x(t))|}. \tag{3-13}$$

In particular, for regions where the gradient is very steep the level sets of (3-11) move with unit velocity whereas the level sets (3-10) of our gradient flow solution move with velocity $\approx 1/\sqrt{t}$ for small times.

Example 3.9 (one-dimensional interval). Let us consider the gradient flow (1-7) with datum $f := 1$ on the domain $\Omega := (-1, 1)$. Then the solution is given by

$$u(t,x) = \begin{cases} \min\left(\frac{1}{\sqrt{3t}}(1-|x|), 1\right), & 0 \le t < \frac{1}{3}, \\ \frac{3}{2}(1-t)_+(1-|x|), & t \ge \frac{1}{3}. \end{cases} \tag{3-14}$$

Example 3.10 (two-dimensional disk). We study the case $\Omega = B_1(0) \subset \mathbb{R}^2$ where $r = 1$. From Example 3.3 we know that (3-5) is in fact an equality since Ω is a ball and thus it holds

$$I_2(g) = \tfrac{\pi}{6} g^3(4 - 3g).$$

Hence the initial value problem (3-8) becomes

$$g'(t) = \frac{g(t)^2}{I_2(g(t))} = \frac{6}{\pi}\frac{1}{g(t)}\frac{1}{4 - 3g(t)}, \qquad g(0) = 0. \tag{3-15}$$

In Figure 3 we plot a numerical approximation for g. In particular, we see that for small times $t > 0$ the function $g(t)$ is proportional to the square root of t, whereas these dynamics change for larger times, as it can be expected from (3-15).

Next, we prove that the analytic solution (3-9) also solves the variational regularization problem (1-8).

Theorem 3.11 (variational problem). *Under the conditions of Theorem 3.6 it holds that (3-9) is the unique solution of*

$$\min_{u \in W_0^{1,\infty}(\Omega)} \tfrac{1}{2}\|u - f\|_2^2 + t\|\nabla u\|_\infty, \tag{3-16}$$

where $f \equiv r$.

Proof. The optimality condition for problem (3-16) is given by $(f - u(t))/t \in \partial J(u(t))$, which is sufficient for optimality due to the convexity of (3-16). We first show that $(f - u(t))/\tilde{t} \in \partial\mathcal{J}(u(t))$, where

$$\tilde{t} := rI_1(g(t)) - \frac{1}{g(t)} I_2(g(t)), \tag{3-17}$$

and the functions I_k for $k \in \{1, 2\}$ are given by (3-3). In a second step we show that $\tilde{t} = t$.

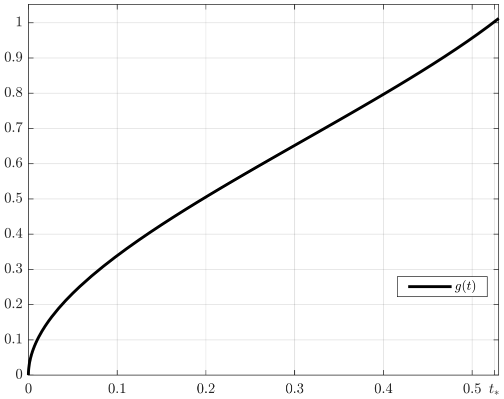


Figure 3. $g(t)$ for the unit circle.

Step 1: By the definition of $\tilde{t}$ and the functions I_k it holds

$$\begin{aligned}\left\langle \frac{f-u(t)}{\tilde{t}}, u(t)\right\rangle &= \frac{1}{\tilde{t}}\int_{\Omega\setminus\Omega_{rg(t)}}\left(r-\frac{d(x)}{g(t)}\right)\frac{d(x)}{g(t)}\,\mathrm{d}x\\ &= \frac{1}{\tilde{t}}\left(\frac{r}{g(t)}I_1(g(t))-\frac{1}{g(t)^2}I_2(g(t))\right)\\ &= \frac{1}{g(t)} = \mathcal{J}(u(t)).\end{aligned}$$

Furthermore, for any $v\in W_0^{1,\infty}(\Omega)$ one computes

$$\left\langle \frac{f-u(t)}{\tilde{t}}, v\right\rangle = \frac{1}{\tilde{t}}\int_{\Omega\setminus\Omega_{rg(t)}}\left(r-\frac{d(x)}{g(t)}\right)v(x)\,\mathrm{d}x \le \mathcal{J}(v),$$

where we used Lipschitz continuity of v just as in the proof of Theorem 3.6. Hence, we have established $(f-u(t))/\tilde{t}\in\partial\mathcal{J}(u(t))$.

Step 2: To show $\tilde{t}=t$ we use the chain rule and (3-4) from Lemma 3.5 for $k\in\{1,2\}$ to obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{t} &= rg'(t)I_1'(g(t))+\frac{g'(t)}{g(t)^2}I_2(g(t))-\frac{g'(t)}{g(t)}I_2'(g(t))\\ &= rg'(t)P(\Omega_{rg(t)})r^2g(t)+\frac{g'(t)}{g(t)^2}I_2(g(t))-\frac{g'(t)}{g(t)}P(\Omega_{rg(t)})r^3g(t)^2\\ &= \frac{g'(t)}{g(t)^2}I_2(g(t))=1,\end{aligned}$$

where the last equality holds since $g(t)$ solves the ODE (3-8). Furthermore, using L'Hôpital's rule and (3-4) it holds

$$\lim_{t\searrow 0}\tilde{t} = \lim_{t\searrow 0}\left[rI_1(g(t))-\frac{1}{g(t)}I_2(g(t))\right] = -\lim_{t\searrow 0}\frac{I_2(g(t))}{g(t)} = -\lim_{t\searrow 0}I_2'(g(t))=0,$$

which finally implies that $\tilde{t}=t$. □

4. Extreme points

In this section we aim to characterize extreme points of the unit ball $B_{\mathcal{J}}$ of $\mathcal{J}$, which is given by

$$B_{\mathcal{J}} := \{u \in L^2(\Omega) : \mathcal{J}(u) \leq 1\} \tag{4-1}$$

and is a convex set and closed set in $L^2(\Omega)$. For a general convex set C, its extreme points are defined as

$$\operatorname{extr}(C) := \{u \in C : \text{there do not exist } v \neq w \in C,\ \lambda \in (0, 1) \text{ such that } u = \lambda v + (1 - \lambda)w\}, \tag{4-2}$$

meaning the extreme points of C are precisely those points which cannot be expressed through a nontrivial convex combination of other points in C.

The set of extreme points of the unit ball of a similar functional has already been studied in [Farmer 1994; Smarzewski 1997]. There the authors considered the Lipschitz seminorm of functions on a metric space which have a prescribed value in one point. Our situation is more complicated since we prescribe a value on the whole boundary of Ω.

The following theorem characterizes the extreme points of $B_{\mathcal{J}}$ analogously to the results in [Farmer 1994]. In a nutshell, a function in $B_{\mathcal{J}}$ is extreme if and only if for almost every point in the domain there exists a path from the point to the boundary of the domain such that the gradient of the function has unit modulus along this path. To this end one introduces the quantity

$$\varepsilon^u_{x,z} := \inf\{\varepsilon > 0 : |x_{i-1} - x_i| - \varepsilon_i \leq |u(x_{i-1}) - u(x_i)|\}, \tag{4-3}$$

where the infimum is computed over all finite sequences of nonnegative numbers $(\varepsilon_i)_{i=1,\dots,n}$ fulfilling $\sum_{i=1}^n \varepsilon_i \leq \varepsilon$, and points $(x_i)_{i=0,\dots,n}$ with $x_0 = z,\ x_1, \dots,\ x_n = x$.

Loosely speaking, $\varepsilon^u_{x,z}$ measures the deviation of the gradient norm from 1, while moving on a path from x to the boundary point z. The following theorem states that if the infimum of (4-3) over all boundary points z is zero, u is an extreme function. We postpone the proof to Appendix B since it is a lengthy generalization of the proof in [Farmer 1994].

Theorem 4.1 (characterization of extreme points). *It holds that $u \in \operatorname{extr}(B_{\mathcal{J}})$ if and only if for almost all $x \in \Omega$ it holds*

$$\inf_{z \in \partial\Omega} \varepsilon^u_{x,z} = 0, \tag{4-4}$$

where $\varepsilon^u_{x,z}$ is given by (4-3).

In the following proposition we sandwich the set of extreme points between two other interesting sets, namely those functions whose gradient has modulus 1 everywhere except for a set with zero measure or nonempty interior, respectively.

Proposition 4.2 (sandwiching extreme points). *It holds that*

$$\{u \in B_{\mathcal{J}} : |\Omega \setminus \Omega_{\max}| = 0\} \subset \operatorname{extr}(B_{\mathcal{J}}) \subset \{u \in B_{\mathcal{J}} : \operatorname{inn}(\Omega \setminus \Omega_{\max}) = \varnothing\}. \tag{4-5}$$

Proof. For the first inclusion we take $u \in W^{1,\infty}_0(\Omega)$ with $|\nabla u| = 1$ almost everywhere, and assume that there are $v \neq w \in B_{\mathcal{J}}$ and $\lambda \in (0, 1)$ such that $u = \lambda v + (1 - \lambda)w$. Defining the set $\Omega_\varepsilon = \{x \in \Omega : |\nabla v(x)| \leq 1 - \varepsilon\}$

for $\varepsilon > 0$, we obtain

$$1 = |\nabla u(x)| \le \lambda|\nabla v(x)| + (1-\lambda)|\nabla w(x)| \le \lambda(1-\varepsilon) + (1-\lambda) = 1 - \lambda\varepsilon \quad \text{for a.e. } x \in \Omega_\varepsilon.$$

Since $\lambda > 0$, this implies that $|\Omega_\varepsilon| = 0$ and hence $|\nabla v| = 1$ almost everywhere in Ω. Applying the same argument to w shows that $|\nabla w| = 1$ holds almost everywhere, as well. Using the Cauchy–Schwarz inequality, we can compute for almost every $x \in \Omega$

$$\begin{aligned} 1 &= |\nabla u(x)|^2 \\ &= \lambda^2|\nabla v(x)|^2 + (1-\lambda)^2|\nabla w(x)|^2 + 2\lambda(1-\lambda)\nabla v(x)\cdot\nabla w(x) \\ &\le \lambda^2 + (1-\lambda)^2 + 2\lambda(1-\lambda) \\ &= 1. \end{aligned}$$

Since $|\nabla v| = 1 = |\nabla w|$, equality has to hold for Cauchy–Schwarz which implies that $\nabla v(x) = c\nabla w(x)$ for some $c \ge 0$. Using that $|\nabla v| = 1 = |\nabla w|$ implies $c = 1$ and hence $\nabla v = \nabla w$ almost everywhere in Ω. Therefore, $v - w$ is constant in Ω and from $v, w = 0$ on $\partial\Omega$ we infer that $v = w$, a contradiction.

For the second inclusion we take some $u \in \operatorname{extr}(B_{\mathcal{J}})$ and — again aiming for a contradiction — we assume that $\Omega \setminus \Omega_{\max}$ has nonempty interior. In this case we set

$$v_\pm(x) := \begin{cases} u, & x \in \Omega_{\max}, \\ u \pm \phi, & x \in \Omega \setminus \Omega_{\max}. \end{cases} \tag{4-6}$$

with a function $\phi \ne 0$ to be specified. Obviously, it holds $v_+ \ne v_-$ since $|\Omega \setminus \Omega_{\max}| > 0$ and furthermore $u = v_+/2 + v_-/2$. If we can choose ϕ in such a way that $\mathcal{J}(v_\pm) \le 1$, we have reached the desired contradiction. Since $\Omega \setminus \Omega_{\max}$ has nonempty interior there is $\varepsilon > 0$ and a set $\Omega_\varepsilon \subset \Omega \setminus \Omega_{\max}$ with nonempty interior such that $|\nabla u| \le 1 - \varepsilon$ almost everywhere on Ω_ε. If we define

$$\phi(x) = \begin{cases} \varepsilon \operatorname{dist}(x, \partial\Omega_\varepsilon), & x \in \Omega_\varepsilon, \\ 0, & \text{else}, \end{cases} \tag{4-7}$$

we infer that $|\nabla v_\pm(x)| = 1$ for $x \in \Omega_{\max}$ and $|\nabla v_\pm(x)| \le (1-\varepsilon) + \varepsilon = 1$ for $x \in \Omega \setminus \Omega_{\max}$. Hence, it holds $\mathcal{J}(v_\pm) \le 1$ which means $v_\pm \in B_{\mathcal{J}}$. Finally, $\phi \ne 0$ holds since Ω_ε has nonempty interior and therefore does not coincide with its boundary. This is a contradiction and we can conclude. □

Corollary 4.3 (distance function is extreme point). *Since $\Omega_{\max} = \Omega$ for the distance function to $\partial\Omega$, we obtain that the distance function is an extreme point.*

Remark 4.4. In general, both inclusions in Proposition 4.2 are proper. The second inclusion is proper even in one dimension, as Example 4.5 below shows. In general, also the first inclusion is proper since in [Rolewicz 1986] the author constructs an extremal function $u : [0, 1]^2 \to \mathbb{R}$ with $\|\nabla u\|_\infty = 1$ whose gradient is supported on a set with arbitrarily small positive measure. This function can be slightly modified to vanish on the boundary of Ω and hence provides a valid counterexample. The construction involves the distance function of a fat Cantor set, which we have already investigated in Example 2.7, and relies on a connectedness argument. However, in one space dimension one can prove that the first inclusion is indeed an equality.

Before we prove that the first inclusion in Proposition 4.2 is an equality in one dimension, we give an example to show the second inclusion is proper. To this end we show that the distance function to a fat Cantor set is no extreme point.

Example 4.5 (distance function to Smith–Volterra–Cantor set). As in Example 2.7 we let $u(x) = \operatorname{dist}(x, F)$ denote the distance function of the fat Smith–Volterra–Cantor set $F \subset \Omega$ with $\Omega = [0, 1]$. Trivially, since $\Omega \setminus \Omega_{\max} = F$, it holds that

$$u \in \{u \in W_0^{1,\infty}(\Omega) : \mathcal{J}(u) = 1 \wedge \operatorname{inn}(\Omega \setminus \Omega_{\max}) = \varnothing\}$$

but we will show that $u \notin \operatorname{extr}(B_{\mathcal{J}})$. To this end, let $f = u'$, which is defined almost everywhere and satisfies $\|f\|_\infty = 1$. We define

$$g_\pm(x) := \begin{cases} f(x), & x \notin F, \\ \pm 1, & x \in F \cap \left[0, \frac{1}{2}\right], \\ \mp 1, & x \in F \cap \left[\frac{1}{2}, 1\right], \end{cases} \tag{4-8}$$

and observe that $g_+ \neq g_-$ since F has positive measure. Next, we define the functions for almost every $x \in \Omega$

$$\tilde{f}(x) := \tfrac{1}{2} g_+(x) + \tfrac{1}{2} g_-(x) = \begin{cases} f(x), & x \notin F, \\ 0, & x \in F, \end{cases} \tag{4-9}$$

$$\tilde{u}(x) := \int_0^x \tilde{f}(t)\, \mathrm{d}t. \tag{4-10}$$

Using the definition of the function $\tilde{f}$ and the fact $\int_a^b f(t)\, \mathrm{d}t = 0$ for every maximally chosen interval $(a, b) \subset \Omega \setminus F$, it is easy to see that $\tilde{u} = u$ holds almost everywhere in Ω. In particular, this also implies that $\tilde{f} = f$ almost everywhere. Finally, we can express u as $u = v_+/2 + v_-/2$, where

$$v_\pm(x) := \int_0^x g_\pm(t)\, \mathrm{d}t$$

satisfy $v_\pm \in W_0^{1,\infty}(\Omega)$ and hence $\mathcal{J}(v_\pm) = \|v_\pm'\|_\infty = \|g_\pm\|_\infty = 1$. This shows that u is not an extreme point.

The construction of this example carries over to the general case and allows us to prove that the first inclusion in Proposition 4.2 is an equality in one space dimension. Note that for Lipschitz continuous functions with one prescribed value in the interval the following was already proved in [Rolewicz 1984]. However, since we demand zero boundary conditions on both boundary points, the proof changes.

Proposition 4.6 (extreme points in one space dimension). *Let $\Omega \subset \mathbb{R}$ be an interval. Then it holds*

$$\operatorname{extr}(B_{\mathcal{J}}) = \{u \in W_0^{1,\infty}(\Omega) : |\Omega \setminus \Omega_{\max}| = 0\}. \tag{4-11}$$

Proof. We just have to show the inclusion "$\subset$". Assume that we have a function $u \in W_0^{1,\infty}(\Omega)$ such that $|\Omega_0| > 0$ where $\Omega_0 := \Omega \setminus \Omega_{\max}$. Without loss of generality we assume that $\Omega = [0, 1]$. We let $f = u'$ denote its derivative and since $|\Omega_0| > 0$ there is some $\varepsilon \in (0, 1]$ such that set $\Omega_\varepsilon := \{x \in \Omega : |f(x)| \leq 1 - \varepsilon\}$

has positive measure. We define

$$g_\pm(x) = \begin{cases} f(x), & x \in \Omega \setminus \Omega_\varepsilon, \\ f(x) \pm \varepsilon, & x \in \Omega_\varepsilon^1, \\ f(x) \mp \varepsilon, & x \in \Omega_\varepsilon^2, \end{cases} \tag{4-12}$$

where the sets Ω_ε^k for $k = 1, 2$ satisfy $\Omega_\varepsilon = \Omega_\varepsilon^1 \dot\cup \Omega_\varepsilon^2$ and are chosen in such a way that $\int_0^1 g_\pm(t)\,\mathrm{d}t = 0$. The construction works as follows. For $\alpha \in [0, 1]$ we define the continuous function

$$h(\alpha) = \int_{\Omega\setminus\Omega_\varepsilon} f(t)\,\mathrm{d}t + \int_{\Omega_\varepsilon\cap[0,\alpha]} f(t) + \varepsilon\,\mathrm{d}t + \int_{\Omega_\varepsilon\cap[\alpha,1]} f(t) - \varepsilon\,\mathrm{d}t.$$

Since u vanishes on the boundary of Ω, its derivative f has zero mean. Hence, we find that

$$h(0) = \int_\Omega f(t)\,\mathrm{d}t - \varepsilon|\Omega_\varepsilon| = -\varepsilon|\Omega_\varepsilon| < 0,$$
$$h(1) = \int_\Omega f(t)\,\mathrm{d}t + \varepsilon|\Omega_\varepsilon| = \varepsilon|\Omega_\varepsilon| > 0.$$

Thus, by the intermediate value theorem for continuous functions, there has to be $\tilde\alpha \in (0, 1)$ such that $h(\tilde\alpha) = 0$. Setting $\Omega_\varepsilon^1 := \Omega_\varepsilon \cap [0, \tilde\alpha]$ and $\Omega_\varepsilon^2 := \Omega_\varepsilon \cap (\tilde\alpha, 1]$, we see from (4-12) that $h(\tilde\alpha) = 0$ is equivalent to $\int_0^1 g_\pm(t)\,\mathrm{d}t = 0$.

It is obvious that $g_+ \neq g_-$ and $\|g_\pm\|_\infty = 1$. Furthermore, it holds $f = g_+/2 + g_-/2$, which means that we decompose $u = v_+/2 + v_-/2$, where $v_\pm = \int_0^x g_\pm(t)\,\mathrm{d}t$ satisfy $\|v_\pm'\|_\infty = \|g_\pm\|_\infty = 1$ and have zero boundary conditions due to $\int_0^1 g_\pm(t)\,\mathrm{d}t = 0$. Hence it holds $\mathcal{J}(v_\pm) = 1$ and we can conclude. □

5. Extension to finite weighted graphs

In this section we analyse a discrete version of functional $\mathcal{J}$ within the framework of finite weighted graphs. This requires equipping the graph with suitable differential operators and function space structures, according to [Elmoataz et al. 2015]. The main appeal of differential calculus on graphs is certainly that it allows for complicated topologies, and generalizes standard finite difference approximations on grids. Furthermore, graphs do not necessarily have to be interpreted as approximations of physical domains, but can also model images, networks, and databases.

After introducing notation and important quantities related to finite weighted graphs, we analyse the functional $\mathcal{J}_w$, given in (5-13) below. In more detail, we study its ground states, characterize its subdifferential and extreme points and investigate some properties of eigenfunctions. One of the main results is Theorem 5.1 below, which states that ground states are distance functions, just as in the continuous case. In general, many results carry over from the continuous case directly, which is why we omit most proofs.

A finite weighted graph G is a triple $G = (V, E, w)$, consisting of a finite set of vertices V, an edge set $E \subset V \times V$, and a weight function $w : E \to \mathbb{R}_{\geq 0}$. The notation $x \sim y$ for $x, y \in V$ indicates that $(x, y) \in E$. In the following, we assume the symmetry conditions

$$x \sim y \iff y \sim x, \tag{5-1}$$

$$w(x, y) = w(y, x) \quad \text{if } x \sim y. \tag{5-2}$$

Furthermore, we assume that the graph is connected, which means that for any two vertices $x, y \in V$ there are edges $(x_0, x_1), (x_1, x_2), \dots, (x_{n-1}, x_n) \in E$ such that $x_0 = x$ and $x_n = y$. On the graph we can define vertex functions $\mathcal{H}(V) = \{u : V \to \mathbb{R}\}$ and edge functions $\mathcal{H}(E) = \{q : E \to \mathbb{R}\}$ which can be viewed as real Hilbert spaces with the inner products

$$\langle u, v\rangle = \sum_{x\in V} u(x)v(x), \qquad u, v \in \mathcal{H}(V), \tag{5-3}$$

$$\langle q, p\rangle = \sum_{x\sim y} q(x, y)\, p(x, y), \quad q, p \in \mathcal{H}(E). \tag{5-4}$$

If an edge function $q \in \mathcal{H}(E)$ satisfies $q(x, y) = -q(y, x)$ for all $x, y \in V$ we call q antisymmetric. Next, we define the weighted gradient ∇_w of a vertex function $u \in \mathcal{H}(V)$ evaluated on an edge $(x, y) \in E$ as

$$(\nabla_w u)(x, y) := w(x, y)^{1/2}(u(y) - u(x)), \tag{5-5}$$

which makes $\nabla_w u : E \to \mathbb{R}$ an antisymmetric edge function. Obviously, $\nabla_w : \mathcal{H}(V) \to \mathcal{H}(E)$ is a linear operator and its adjoint is given by $\nabla_w^* = -\operatorname{div}_w$, where

$$(\operatorname{div}_w q)(x) := \sum_{y:x\sim y} w(x, y)^{1/2}(q(y, x) - q(x, y)) \tag{5-6}$$

denotes the weighted divergence of an edge function $q \in \mathcal{H}(E)$ evaluated in $x_i \in V$. This implies the validity of the integration by parts formula

$$\langle q, \nabla_w u\rangle = -\langle \operatorname{div}_w q, u\rangle \quad \text{for all } u \in \mathcal{H}(V),\ q \in \mathcal{H}(E). \tag{5-7}$$

Furthermore, we define the one-sided gradient

$$(\nabla_w^- u)(x, y) := w(x, y)^{1/2}(u(y) - u(x))_-, \tag{5-8}$$

where $(x)_- := -\min(x, 0)$, and introduce p-norms on $\mathcal{H}(V)$ and $\mathcal{H}(E)$ by setting

$$\|u\|_p = \Big(\sum_{x\in V} |u(x)|^p\Big)^{1/p}, \qquad 1 \le p < \infty, \tag{5-9}$$

$$\|u\|_\infty = \max_{x\in V} |u(x)|, \tag{5-10}$$

$$\|q\|_p = \Big(\sum_{x\sim y} |q(x, y)|^p\Big)^{1/p}, \quad 1 \le p < \infty, \tag{5-11}$$

$$\|q\|_\infty = \max_{x\sim y} |q(x, y)|. \tag{5-12}$$

Next we take a subset of the vertex set $\Gamma \subset V$ which we identify with a Dirichlet boundary, and consider the subspace $\mathcal{H}_0(V) = \{u \in \mathcal{H}(V) : u(x) = 0 \text{ for all } x \in \Gamma\}$ of all vertex functions which vanish on Γ. Analogous to (1-14), we define the functional

$$\mathcal{J}_w(u) = \begin{cases} \|\nabla_w u\|_\infty, & u \in \mathcal{H}_0(V), \\ +\infty, & \text{else.} \end{cases} \tag{5-13}$$

Note that also $\mathcal{J}_w$ is a convex and absolutely one-homogeneous functional on a Hilbert space. The aim of the following section is to analyse $\mathcal{J}_w$ and show results analogous to those we have seen in Section 2.

5A. ***Ground states and properties of the distance function.*** First we will study ground states of $\mathcal{J}_w$, i.e., functions $u^* \in \mathcal{H}_0(V)$ such that

$$u^* \in \operatorname*{argmin}_{u \in \mathcal{H}_0(V)} \frac{\mathcal{J}_w(u)}{\|u\|_2}. \tag{5-14}$$

Since ground states are invariant under multiplication with scalars, we can again replace the problem with

$$u^* \in \operatorname*{argmax}_{\substack{u \in \mathcal{H}_0(V) \\ |\nabla_w u| \le 1}} \|u\|_2. \tag{5-15}$$

Theorem 5.1 (ground states are distance functions). *Up to global sign, the unique solution of* (5-15) *is given by*

$$u^*(x) = d(x) := \min_{y \in \Gamma} d_w(x, y), \qquad x \in V, \tag{5-16}$$

where

$$d_w(x, y) := \min\left\{ \sum_{i=1}^{n} w(x_{i-1}, x_i)^{-1/2} : n \in \mathbb{N},\ x_0 \sim \cdots \sim x_n,\ x_0 = x, x_n = y \right\} \tag{5-17}$$

denotes the graph distance of $x, y \in V$.

Proof. Since $d_w(\,\cdot\,, \cdot\,)$ is a distance and hence fulfills the triangle inequality it is standard to check that (5-16) is 1-Lipschitz and hence admissible in (5-15). To show that (5-16) indeed solves (5-15) we note that by possibly replacing u^* with $|u^*|$ one can restrict the maximization to nonnegative functions. From there it is straightforward to see that $u(x) \le d(x)$ for all $x \in V$, which implies that (5-16) solves (5-15). □

Note that on graphs the distance function, and hence the solution of (5-15), does typically not fulfill $|(\nabla_w d)(x, y)| = 1$ for all $(x, y) \in E$, as the following simple example shows.

Example 5.2 (distance function with vanishing gradient). We consider the graph $G = (V, E)$ with vertices $V = \{x_0, x_1, x_2, x_3\}$ and edges $E = \{(x_0, x_1), (x_1, x_2), (x_2, x_3)\}$. The weights are assumed to be 1 and we take $\Gamma = \{x_0, x_3\}$. Using compact tuple notation, the distance function is given by

$$d = (0, 1, 1, 0)$$

and obviously it holds $(\nabla_w d)(x_1, x_2) = 0$.

Of course, the fact that $|(\nabla_w d)(x, y)| \neq 1$ in general is due to the fact that $(\nabla_w d)(x, y)$ can only be interpreted as a directional derivative and not as a full gradient. However, we have the following theorem.

Proposition 5.3 (properties of the distance function). *For all* $x \in V$ *and* $y \sim x$ *the distance function* d *to* Γ *satisfies*

$$|(\nabla_w d)(x, y)| \begin{cases} = 1 & \textit{if } y \in \mathrm{SP}(x, \Gamma) \textit{ or } x \in \mathrm{SP}(y, \Gamma), \\ < 1 & \textit{else}, \end{cases} \tag{5-18}$$

where

$$\mathrm{SP}(x,\Gamma) := \left\{ x_0 \sim \cdots \sim x_n,\ x_0 = x,\ x_n \in \Gamma,\ d(x) = \sum_{i=1}^{n} w(x_{i-1}, x_i)^{-1/2} \right\} \tag{5-19}$$

denotes the set of all shortest paths from x to Γ.

Proof. Let $x \in V$ and $y \sim x$ be a neighboring node. If $y \in \mathrm{SP}(x,\Gamma)$, then $x \sim y \sim x_1 \sim \cdots \sim x_n$ with $x_n \in \Gamma$ is a shortest path for x and $y \sim x_1 \sim \cdots \sim x_n$ is a shortest path for y. Consequently, $d(x)$ and $d(y)$ differ by the value $d_w(x, y) = w(x, y)^{-1/2}$, which means $|(\nabla_w d)(x, y)| = 1$. If $x \in \mathrm{SP}(y,\Gamma)$, the same holds true by interchanging the roles of x and y.

In the case that x and y do not lie on a common shortest path, it holds

$$\begin{aligned} d(y) &< d(x) + w(x, y)^{-1/2}, \\ d(x) &< d(y) + w(x, y)^{-1/2}, \end{aligned}$$

and hence $|d(y) - d(x)| < w(x, y)^{-1/2}$, which implies $|(\nabla_w d)(x, y)| < 1$. □

For a nonweighted graph, meaning that all weights are 1, we can obtain a more precise characterization of the directional derivatives of the distance function. Furthermore, we show that the 1-norm of the one-sided gradient $\nabla_w^- d$ as in a point $x \in V$ counts the number of optimal paths from x to Γ.

Corollary 5.4 (unitary weights). *Assume that $w(x, y) = 1$ for all $x \sim y$. Then for all $x \in V$ and $y \sim x$ it holds*

$$|(\nabla_w d)(x, y)| = \begin{cases} 1 & \textit{if } y \in \mathrm{SP}(x,\Gamma) \textit{ or } x \in \mathrm{SP}(y,\Gamma), \\ 0 & \textit{else}. \end{cases} \tag{5-20}$$

Furthermore, it holds

$$\sum_{x \sim y} |(\nabla_w^-) d(x, y)| = \#\{y : y \in \mathrm{SP}(x,\Gamma)\}. \tag{5-21}$$

Proof. The first statement follows from Proposition 5.3, observing that $1 > |(\nabla_w d)(x, y)| = |d(y) - d(x)|$ implies $d(x) = d(y)$ since d takes only integer values. For the second statement we note that the one-sided gradient $(\nabla_w^- d)(x, y)$ equals zero if $x \in \mathrm{SP}(y,\Gamma)$ since in this case $d(y) > d(x)$. Hence, it holds

$$|(\nabla_w^- d)(x, y)| = \begin{cases} 1 & \text{if } y \in \mathrm{SP}(x,\Gamma), \\ 0 & \text{else}. \end{cases} \tag{5-22}$$

which directly implies (5-21). □

5B. ***Subdifferential and eigenfunctions.*** After having characterized the ground state of $\mathcal{J}_w$ as distance function and having studied its geometric properties, we proceed with the characterization of the subdifferential $\partial\mathcal{J}_w$ and study properties of eigenfunctions.

In the following, we fix a function $u \in \mathcal{H}_0(V)$, and define the set of edges where the gradient of u attains its maximal modulus as

$$E_{\max} = \{(x, y) \in E : |(\nabla_w u)(x, y)| = \mathcal{J}_w(u)\}. \tag{5-23}$$

Note that $E_{\max}$ is never empty due to the finite-dimensional nature of all quantities involved. The following proposition characterizes the subdifferential of $\mathcal{J}_w$ analogously to Proposition 2.10.

Proposition 5.5 (characterization of the subdifferential). *Let $u \in \mathcal{H}_0(V) \setminus \{0\}$ and let $E_{\max}$ be given by (5-23). Then it holds*

$$\partial\mathcal{J}_w(u) = \{-\operatorname{div}_w q : q \in \mathcal{H}(E),\ \|q\|_1 = 1,\ q(x,y) = 0 \text{ for all } (x,y) \in E \setminus E_{\max},\\ q(x,y)(\nabla_w u)(x,y) = |q(x,y)||(\nabla_w u)(x,y)| \text{ for all } (x,y) \in E_{\max}\}.$$

Next we study extreme points of the unit ball $B_{\mathcal{J}_w}$ of $\mathcal{J}_w$, given by

$$B_{\mathcal{J}_w} = \{u \in \mathcal{H}(V) : \mathcal{J}_w(u) \le 1\}. \tag{5-24}$$

Next we turn to the study of eigenfunctions of $\partial\mathcal{J}_w$. We should first remark that $\lambda u \in \partial\mathcal{J}_w(u)$ is not a good definition for eigenfunctions due to the Dirichlet conditions on Γ. This means that in general, one cannot find $u \in \mathcal{H}_0(V)$ and $q \in \mathcal{H}(E)$ such that $\lambda u = -\operatorname{div}_w q$. This is illustrated in the following example.

Example 5.6. Let $V = \{x_0, x_1, x_2\}$, $E = \{(x_0, x_1), (x_1, x_2)\}$, and assume all weights are 1. We set $\Gamma = \{x_0, x_2\}$. Then, trivially, the distance function $d = (0, 1, 0)$ is an eigenfunction. If we assume that $\lambda u = -\operatorname{div}_w q \in \partial\mathcal{J}_w(u)$ then $d(x_0) = 0$ implies $q(x_0, x_1) = q(x_1, x_0)$ by definition of the divergence operator. The characterization of the subdifferential in Proposition 5.5 then tells us that $q(x_0, x_1) = 0 = q(x_1, x_0)$ since q has to be parallel to $(\nabla_w d)(x_0, x_1) = 1$ and $(\nabla_w d)(x_1, x_0) = -1$. The same holds for $q(x_1, x_2)$ and hence $q = 0$ which contradicts $-\operatorname{div} q = \lambda d$.

Definition 5.7 (eigenfunctions of $\partial\mathcal{J}_w$). We call $u \in \mathcal{H}_0(V)$ an eigenfunction of $\partial\mathcal{J}_w$ if there exist $\lambda > 0$ and $q \in H(E)$ with $-\operatorname{div} q \in \partial J(u)$ such that

$$\langle \lambda u, v\rangle = \langle -\operatorname{div}_w q, v\rangle \quad \text{for all } v \in \mathcal{H}_0(V). \tag{5-25}$$

This is equivalent to $\lambda u(x) = -\operatorname{div}_w q(x)$ for all $x \in V \setminus \Gamma$.

The next example shows that nonnegative eigenfunctions of $\partial\mathcal{J}_w$ are not unique, as opposed to the continuum case where Proposition 2.14 asserted that every nonnegative eigenfunction is a ground state.

Example 5.8 (multiple nonnegative eigenfunctions). We return to the graph from Example 5.2. The functional $\mathcal{J}_w$ can be explicitly expressed as

$$\mathcal{J}_w(u) = \max(|u_1|, |u_2|, |u_1 - u_2|),$$

where $u_i := u(x_i)$ for $i = 1, 2$. The unit ball and dual unit ball of $\mathcal{J}_w$ are depicted in Figure 4. Following [Bungert et al. 2019a], eigenvectors are precisely all multiples of vectors in the dual unit ball whose orthogonal hyperplane is tangent to the boundary. Here they correspond to all multiples of the four vertex functions having the values

$$\left(0, \tfrac{1}{2}, \tfrac{1}{2}, 0\right), \quad (0, 1, 0, 0), \quad (0, 0, 1, 0), \quad (0, -1, 1, 0).$$

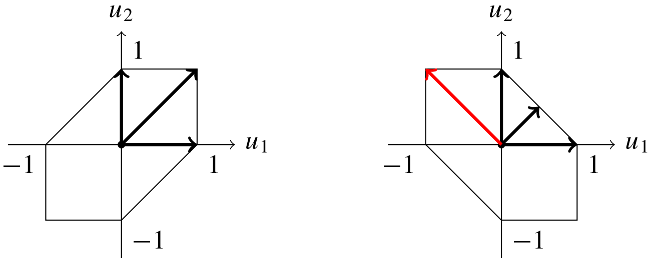


Figure 4. Primal and dual unit balls of $\mathcal{J}_w$ with all extreme points and eigenvectors (up to scalar multiplication).

Note that the first three are also extreme points of the primal unit ball (up to scalar multiplication), whereas the fourth one, marked in red, is not. Furthermore, the first three eigenfunctions are all nonnegative.

We have just seen that nonnegative eigenfunctions do in general not coincide with a ground state, as is the case in the continuum. However, thanks to the following proposition, whose proof works just as in the continuous case of Proposition 2.14, *positive* eigenfunctions are unique.

Proposition 5.9 (positive eigenfunctions). *Let $u \in \mathcal{H}_0(V)$ be a nonnegative eigenfunction with $\mathcal{J}_w(u) = 1$ and let d denote the distance function to Γ. Then for every $x \in V$ it holds $u(x) = d(x)$ or $u(x) = 0$. Consequently, any eigenfunction which is positive in $V \setminus \Gamma$ coincides with a ground state.*

5C. ***Extreme points.*** As in the continuous case of Section 4, the main motivation for studying extreme points is representer theorems. They assert certain optimization problems involving $\mathcal{J}_w$ admit a solution which is a linear combination of extreme points. As before, we obtain a characterization of extreme points which is based on the existence of paths from every vertex to the boundary Γ such that all directional derivatives are one along this path.

Theorem 5.10 (characterization of extreme points). *It holds that*

$$\operatorname{extr}(B_{\mathcal{J}_w}) = \Big\{u \in \mathcal{H}_0(V) : \textit{for all } x \in V \textit{ there exist } x_0 \sim \cdots \sim x_n \textit{ with } x_0 = x,\ \textit{and } x_n \in \Gamma,\\ |(\nabla_w u)(x_{i-1}, x_i)| = 1 \textit{ for all } i = 1, \dots, n\Big\}.$$

However, as opposed to the continuous case, even one-dimensional extreme functions do not necessarily have constant modulus of the gradient, as the following example shows.

Example 5.11. We return to Example 5.2 with the distance function $d(x) = (0, 1, 1, 0)$, which fulfills $\nabla_w d(x_1, x_2) = 0$. Nevertheless, it obviously is an extreme point taking the paths $x_1 \sim x_0$ and $x_2 \sim x_3$. If one however adds a node x_4 with $x_3 \sim x_4$ and sets $u(x) = (0, 1, 1, 0, 0)$ this is not extreme anymore, since there is no path from x_3 to x_0 or x_4 along which $\nabla_w u$ has modulus 1.

Appendix A: Proof of Proposition 2.4

Before we proceed to the proof of the theorem, we need a straightforward approximation lemma for Lipschitz functions.

Lemma A.1. *Let $v \in W_0^{1,\infty}(\Omega)$. Then there exists a sequence $(v_n) \subset C_c^\infty(\Omega)$ such that*

- $\|\nabla v_n\|_\infty \le \|\nabla v\|_\infty$,
- $\|v - v_n\|_\infty \to 0$ *as* $n \to \infty$.

Proof. First, we approximate v with compactly supported functions $(w_n) \subset C_c^{0,1}(\Omega)$. To this end, set

$$w_n^\pm(x) = \min\Big(v^\pm(x) - \frac{1}{n}, 0\Big),$$

where $v^\pm$ denote the positive and negative parts of v. If we define $w_n := w_n^+ - w_n^-$, it holds

$$\|v - w_n\|_\infty \le \frac{1}{n} \to 0, \quad n \to \infty,$$

and $\|\nabla w_n\|_\infty \le \|\nabla v\|_\infty$. Furthermore, all w_n are compactly supported. To see this one notes that

$$|v(x)| \le \mathcal{J}(v)\,\mathrm{dist}(x, \partial\Omega),$$

which implies that $w_n = 0$ for all $x \in \Omega$ such that $\mathrm{dist}(x, \partial\Omega) \le 1/(\mathcal{J}(v)n)$. Now let $\varepsilon = 1/(2n)$ and define mollifications $v_n := w_n * \varphi_\varepsilon$. Then it holds $\|\nabla v_n\|_\infty \le \|\nabla w_n\|_\infty \le \|\nabla v\|_\infty$ and

$$\|v - v_n\|_\infty \le \|v - w_n\|_\infty + \|v_n - w_n\|_\infty.$$

The first term on the right-hand side can be bounded by $1/n$ as shown above. For the second term we notice

$$|v_n(x) - w_n(x)| \le \int_\Omega |\varphi_\varepsilon(y)|w_n(x-y) - w_n(x)|\,\mathrm{d}y \le \|\nabla w_n\|_\infty \frac{1}{2n} \le \|\nabla v\|_\infty \frac{1}{2n} \quad \text{for all } x \in \Omega.$$

Hence, both terms converge to zero and we can conclude. □

Proof of Proposition 2.4. We follow the argumentation of [Bredies and Holler 2016, Proposition 7], which deals with the subdifferential of the total variation. Defining the set

$$C := \{-\operatorname{div} q : q \in C^\infty(\overline{\Omega}, \mathbb{R}^n),\ \|q\|_1 \le 1\},$$

it holds $\mathcal{J}(u) = \chi_C^*(u)$, where χ denotes the characteristic function of a set and * denotes the convex conjugate. Hence, it holds

$$\mathcal{J}^*(\zeta) = \chi_C^{**}(\zeta) = \chi_{\overline{C}}(\zeta)$$

and by (2-6) one gets that $\zeta \in \partial\mathcal{J}(u)$ if and only if $\zeta \in \overline{C}$ and $\langle \zeta, u\rangle = \mathcal{J}(u)$.

Therefore, we just have to find the L^2-closure of C and we claim

$$\overline{C} = \{-\operatorname{div} q : g = g + r,\ g \in G_0^1(\Omega),\ r \in \mathcal{N}(\mathrm{div};\Omega),\ |q|(\Omega) \le 1\} =: K.$$

Inclusion $K \subset \overline{C}$: For this inclusion it is enough to show that for any $q \in \mathcal{M}(\Omega, \mathbb{R}^n)$ with $-\operatorname{div} q \in K$ it holds

$$\int_\Omega -(\operatorname{div} q)v\,\mathrm{d}x \le \mathcal{J}(v) \quad \text{for all } v \in L^2(\Omega)$$

since this implies $\chi_{\bar{C}}(-\operatorname{div} q) = \mathcal{J}^*(-\operatorname{div} q) = 0$ and hence $-\operatorname{div} q \in \bar{C}$. Indeed, it suffices to check the inequality for $v \in W_0^{1,\infty}(\Omega)$. By Lemma A.1, we can find a sequence of functions $(v_n) \subset C_c^\infty(\Omega)$ such that $\|\nabla v_n\|_\infty \le \|\nabla v\|_\infty$ and $\|v_n - v\|_\infty \to 0$ as $n \to \infty$. This implies

$$\int_\Omega -(\operatorname{div} q) v \,\mathrm{d}x = \lim_{n\to\infty} \int_\Omega -(\operatorname{div} q) v_n \,\mathrm{d}x = \lim_{n\to\infty} \int_\Omega \nabla v_n \cdot \mathrm{d}q \,\mathrm{d}x \le |q|(\Omega) \|\nabla v_n\|_\infty \le \mathcal{J}(v).$$

Inclusion $\bar{C} \subset K$: To prove the converse inclusion it suffices to show that K is closed in $L^2(\Omega)$ since $C \subset K$ is obviously correct. Let $(q_n) \subset \mathcal{M}(\Omega, \mathbb{R}^n)$ be a sequence of measures such that $q_n = g_n + r_n$, with $(g_n) \subset G_0^1(\Omega)$, $(r_n) \subset \mathcal{N}(\operatorname{div}; \Omega)$. Furthermore, assume that $|q_n|(\Omega) \le 1$ and $-\operatorname{div} q_n \to \mu$ strongly in $L^2(\Omega)$. From [Auchmuty 2006, (1.2)] we infer that $\|g_n\|_2$ is uniformly bounded and hence, up to a subsequence, g_n converges weakly in $L^2(\Omega)$ to some $g \in L^2(\Omega)$. By the closedness of $G_0^1(\Omega)$ we infer that $g \in G_0^1(\Omega)$. We first show that $\mu = -\operatorname{div} g$. To this end, we use the convergences $g_n \rightharpoonup g$ and $\operatorname{div} q_n \to \mu$ together with the fact that $\operatorname{div} g_n = \operatorname{div} q_n$ to compute

$$\langle g, \nabla\varphi\rangle = \lim_{n\to\infty} \langle g_n, \nabla\varphi\rangle = -\lim_{n\to\infty} \langle \operatorname{div} g_n, \varphi\rangle = -\lim_{n\to\infty} \langle \operatorname{div} q_n, \varphi\rangle = \langle \mu, \varphi\rangle \quad \text{for all } \varphi \in C_c^\infty(\Omega),$$

which shows $\mu = -\operatorname{div} g$. Since $|q_n|(\Omega) \le 1$, by the sequential Banach–Alaoglu theorem there exists a measure $q \in \mathcal{M}(\Omega, \mathbb{R}^n)$ such that, up to a subsequence, it holds $q_n \rightharpoonup q$. The lower semicontinuity of the total variation implies $|q|(\Omega) \le 1$. Furthermore, $g_n \rightharpoonup g$ implies that in fact $r_n \rightharpoonup r := q - g$. By the closedness of $\mathcal{N}(\operatorname{div}; \Omega)$, we infer $r \in \mathcal{N}(\operatorname{div}; \Omega)$. Hence, we have shown that $\mu = -\operatorname{div} q \in K$, as desired. □

Appendix B: Proof of Theorem 4.1

In order to prove the theorem, we first need the following lemma, which states a triangle inequality for the map $x \mapsto \varepsilon^u_{x,z}$, given by (4-3).

Lemma B.1. *Let $u \in B_{\mathcal{J}}$, $x, y \in \Omega$, and $z \in \partial\Omega$. Then it holds*

$$\varepsilon^u_{y,z} \le \varepsilon^u_{x,z} + |x - y| - |u(x) - u(y)|.$$

Proof. We denote by $(\varepsilon^n)_{n\in\mathbb{N}}$ a minimizing sequence for $\varepsilon^u_{x,z}$, i.e., $\lim_{n\to\infty} \varepsilon^n = \varepsilon^u_{x,z}$. This means that for each $n \in \mathbb{N}$ there exists a path of n points $x_0^n = z$, $x_1^n, \dots, x_n^n = x$ connecting z and x, and nonnegative numbers $(\varepsilon_i)_{i=1,\dots,n}$ such that

$$|x_{i-1} - x_i| - \varepsilon_i \le |u(x_{i-1}) - u(x_i)|, \quad i = 1, \dots, n,$$
$$\sum_{i=1}^n \varepsilon_i \le \varepsilon^n.$$

Now we define the path of $n + 1$ points

$$y_i = \begin{cases} x_i^n, & i = 0, \dots, n, \\ y, & i = n + 1, \end{cases}$$

which connects z and y, set

$$\varepsilon_{n+1} = |x - y| - |u(x) - u(y)| \ge 0,$$

and observe that this constellation is admissible for the minimization that defines $\varepsilon^u_{y,z}$ since

$$|x_{i-1}-x_i|-\varepsilon_i \le |u(x_{i-1})-u(x_i)|, \quad i=1,\dots,n+1,$$
$$\sum_{i=0}^{n+1}\varepsilon_i \le \varepsilon^n+|x-y|-|u(x)-u(y)|.$$

Hence it holds

$$\varepsilon^u_{x,y}\le\varepsilon^n+|x-y|-|u(x)-u(y)|$$

and letting n tend to infinity we obtain the desired inequality. □

Now we can proceed to the proof of the theorem.

Proof of Theorem 4.1. The proof works similarly to [Farmer 1994] with the main difference being that there the point $z=0$ is fixed. Since this causes nontrivial modifications, we present the full proof for completeness.

We start with the implication "$\Leftarrow$": to this end, we assume that (4-4) holds for almost all $x\in\Omega$. Since $\varepsilon^u_{x,z}$ depends continuously on $z\in\partial\Omega$ and $\partial\Omega$ is compact, we infer that for almost all $x\in\Omega$ there exists $z\in\partial\Omega$ with $\varepsilon^u_{x,z}=0$. Aiming for a contradiction we assume $u=v/2+w/2$ with two functions $v,w\in B_{\mathcal{J}}$. Since $\varepsilon^u_{x,z}=0$, for any $\varepsilon>0$ we can find finite sequences of points $(x_i)_{i=0,\dots,n}$ and numbers $(\varepsilon_i)_{i=1,\dots,n}$ satisfying the restrictions such that

$$|x_{i-1}-x_i|-\varepsilon_i\le|u(x_{i-1})-u(x_i)| \quad \text{for all } i=1,\dots,n.$$

Without loss of generality we assume that $u(x_{i-1})-u(x_i)\ge 0$. Using also $u=v/2+w/2$ we infer

$$\begin{aligned}-\varepsilon_i&=|x_{i-1}-x_i|-\varepsilon_i-|x_{i-1}-x_i|\\&\le|u(x_{i-1})-u(x_i)|-|v(x_{i-1})-v(x_i)|\\&\le u(x_{i-1})-u(x_i)-(v(x_{i-1})-v(x_i))\\&=w(x_{i-1})-w(x_i)-(u(x_{i-1})-u(x_i))\\&\le|x_{i-1}-x_i|-(\varepsilon_i-|x_{i-1}-x_i|)\\&=\varepsilon_i,\end{aligned}$$

which means

$$|u(x_{i-1})-u(x_i)-(v(x_{i-1})-v(x_i))|\le\varepsilon_i \quad \text{for all } i=1,\dots,n.$$

Iterating this estimate, we obtain

$$\begin{aligned}|u(x)-v(x)|&=|u(x_n)-v(x_n)|\\&=|u(x_n)-u(x_{n-1})-(v(x_n)-v(x_{n-1}))+u(x_{n-1})-v(x_{n-1})|\\&\le\varepsilon_n+|u(x_{n-1})-v(x_{n-1})|\\&\le\cdots\\&\le\sum_{i=1}^{n}\varepsilon_i+|u(x_0)-v(x_0)|\le\varepsilon,\end{aligned}$$

where we used that $x_0 = z \in \partial\Omega$ and hence $u(x_0) = v(x_0) = 0$ there. Since this estimate holds for all $\varepsilon > 0$ and almost all $x \in \Omega$, we infer $u = v$ and hence also $u = w$ in almost everywhere in Ω, which means that u is extreme.

For the converse implication "⇒" we assume that there exists a set $A \subset \Omega$ of positive measure such that it holds $\hat{\varepsilon}_x := \inf_{z\in\partial\Omega} \varepsilon^u_{x,z} > 0$ for almost all $x \in A$. We define the functions

$$v_\pm(x) = \begin{cases} u(x) \pm \hat{\varepsilon}_x, & x \in A, \\ u(x), & x \in \Omega \setminus A, \end{cases}$$

which obviously satisfy $v_+ \neq v_-$ and $v_+/2 + v_-/2 = u$. It remains to show that $v_\pm \in B_{\mathcal{J}}$ to obtain that u is not extreme. We consider v_+ only since the considerations for v_- are identical. We just have to show that $|v_+(x) - v_-(y)| \leq |x - y|$ for all $x, y \in \Omega$. For $x, y \in \Omega \setminus A$ this is clear and hence we first assume that $x \in \Omega \setminus A$ and $y \in A$. In this case it holds

$$|v_+(x) - v_+(y)| = |u(x) - u(y) - \hat{\varepsilon}_y| \leq |u(x) - u(y)| + \hat{\varepsilon}_y.$$

Since $\hat{\varepsilon}_x = 0$, by the assumption $x \in \Omega \setminus A$ we can choose $z_0 \in \partial\Omega$ such that $\varepsilon^u_{x,z_0} = 0$. By the definition of $\hat{\varepsilon}_y$ and the triangle inequality from Lemma B.1 we obtain

$$\hat{\varepsilon}_y \leq \varepsilon^u_{y,z_0} \leq \underbrace{\varepsilon^u_{x,z_0}}_{=0} + |x - y| - |u(x) - u(y)|,$$

which yields

$$|v_+(x) - v_-(y)| \leq |x - y|.$$

Assume now that $x, y \in A$ in which case it holds

$$|v_+(x) - v_+(y)| = |u(x) - u(y) + \hat{\varepsilon}_x - \hat{\varepsilon}_y| \leq |u(x) - u(y)| + |\hat{\varepsilon}_x - \hat{\varepsilon}_y|.$$

Now we choose elements $z_x, z_y \in \partial\Omega$ such that $\hat{\varepsilon}_x = \varepsilon^u_{x,z_x}$ and $\hat{\varepsilon}_y = \varepsilon^u_{y,z_y}$. By using the triangle inequality from Lemma B.1 for $z \in \{z_x, z_y\}$ we obtain

$$|u(x) - u(y)| \leq |x - y| + \tfrac{1}{2}(\varepsilon_{x,z_x} + \varepsilon_{x,z_y}) - \tfrac{1}{2}(\varepsilon_{y,z_x} + \varepsilon_{y,z_y}).$$

After possibly exchanging the roles of x and y we can assume that the right-hand side is less than or equal to $|x - y|$, which concludes the proof. □

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 777826 (NoMADS).

Yury Korolev acknowledges support of the Royal Society (Newton International Fellowship NF170045), the EPSRC (Fellowship EP/V003615/1), the Cantab Capital Institute for the Mathematics of Information and the National Physical Laboratory.

References

[Alter et al. 2005] F. Alter, V. Caselles, and A. Chambolle, "A characterization of convex calibrable sets in $\mathbb{R}^N$", *Math. Ann.* **332**:2 (2005), 329–366. MR Zbl

[Aronsson et al. 2004] G. Aronsson, M. G. Crandall, and P. Juutinen, "A tour of the theory of absolutely minimizing functions", *Bull. Amer. Math. Soc.* (*N.S.*) **41**:4 (2004), 439–505. MR Zbl

[Auchmuty 2006] G. Auchmuty, "Divergence L^2-coercivity inequalities", *Numer. Funct. Anal. Optim.* **27**:5-6 (2006), 499–515. MR Zbl

[Barles 1988] G. Barles, "Remarks on uniqueness results of the first eigenvalue of the p-Laplacian", *Ann. Fac. Sci. Toulouse Math.* (5) **9**:1 (1988), 65–75. MR Zbl

[Barron and Jensen 2005] E. N. Barron and R. Jensen, "Minimizing the L^∞ norm of the gradient with an energy constraint", *Comm. Partial Differential Equations* **30**:10-12 (2005), 1741–1772. MR Zbl

[Barron et al. 2008] E. N. Barron, L. C. Evans, and R. Jensen, "The infinity Laplacian, Aronsson's equation and their generalizations", *Trans. Amer. Math. Soc.* **360**:1 (2008), 77–101. MR Zbl

[Bellettini et al. 2005] G. Bellettini, V. Caselles, and M. Novaga, "Explicit solutions of the eigenvalue problem $-\mathrm{div}\big(\frac{Du}{|Du|}\big) = u$ in $\mathbb{R}^2$", *SIAM J. Math. Anal.* **36**:4 (2005), 1095–1129. MR Zbl

[Benning and Burger 2013] M. Benning and M. Burger, "Ground states and singular vectors of convex variational regularization methods", *Methods Appl. Anal.* **20**:4 (2013), 295–334. MR Zbl

[Benning et al. 2017] M. Benning, M. Möller, R. Z. Nossek, M. Burger, D. Cremers, G. Gilboa, and C.-B. Schönlieb, "Nonlinear spectral image fusion", pp. 41–53 in *Scale space and variational methods in computer vision* (Kolding, Denmark, 2017), edited by F. Lauze et al., Lecture Notes in Computer Science **10302**, Springer, 2017.

[Binding et al. 2006] P. Binding, L. Boulton, J. Čepička, P. Drábek, and P. Girg, "Basis properties of eigenfunctions of the p-Laplacian", *Proc. Amer. Math. Soc.* **134**:12 (2006), 3487–3494. MR Zbl

[Boyer et al. 2019] C. Boyer, A. Chambolle, Y. De Castro, V. Duval, F. de Gournay, and P. Weiss, "On representer theorems and convex regularization", *SIAM J. Optim.* **29**:2 (2019), 1260–1281. MR Zbl

[Bredies and Carioni 2020] K. Bredies and M. Carioni, "Sparsity of solutions for variational inverse problems with finite-dimensional data", *Calc. Var. Partial Differential Equations* **59**:1 (2020), art. id. 14. MR Zbl

[Bredies and Holler 2016] K. Bredies and M. Holler, "A pointwise characterization of the subdifferential of the total variation functional", preprint, 2016. arXiv

[Bungert and Burger 2020] L. Bungert and M. Burger, "Asymptotic profiles of nonlinear homogeneous evolution equations of gradient flow type", *J. Evol. Equ.* **20**:3 (2020), 1061–1092. MR Zbl

[Bungert et al. 2019a] L. Bungert, M. Burger, A. Chambolle, and M. Novaga, "Nonlinear spectral decompositions by gradient flows of one-homogeneous functionals", preprint, 2019. To appear in *Anal. PDE*. arXiv

[Bungert et al. 2019b] L. Bungert, M. Burger, and D. Tenbrinck, "Computing nonlinear eigenfunctions via gradient flow extinction", pp. 291–302 in *Scale space and variational methods in computer vision* (Hofgeismar, Germany, 2019), edited by J. Lellmann et al., Lecture Notes in Computer Science **11603**, Springer, 2019.

[Burger and Osher 2013] M. Burger and S. Osher, "A guide to the TV zoo", pp. 1–70 in *Level set and PDE based reconstruction methods in imaging*, edited by M. Burger and S. Osher, Lecture Notes in Math. **2090**, Springer, 2013. MR Zbl

[Burger et al. 2016a] M. Burger, G. Gilboa, M. Moeller, L. Eckardt, and D. Cremers, "Spectral decompositions using one-homogeneous functionals", *SIAM J. Imaging Sci.* **9**:3 (2016), 1374–1408. MR Zbl

[Burger et al. 2016b] M. Burger, K. Papafitsoros, E. Papoutsellis, and C.-B. Schönlieb, "Infimal convolution regularisation functionals of BV and L^p spaces: the case $p = \infty$", pp. 169–179 in *System modeling and optimization* (Sophia Antipolis, France, 2015), edited by L. Bociu et al., IFIP Adv. Info. Comm. Tech. **494**, 2016. Zbl

[Chen et al. 2009] G.-Q. Chen, M. Torres, and W. P. Ziemer, "Gauss–Green theorem for weakly differentiable vector fields, sets of finite perimeter, and balance laws", *Comm. Pure Appl. Math.* **62**:2 (2009), 242–304. MR Zbl

[Cohen and Gilboa 2020] I. Cohen and G. Gilboa, "Introducing the p-Laplacian spectra", *Signal Processing* **167** (2020), art. id.107281.

[Desquesnes et al. 2010] X. Desquesnes, A. Elmoataz, O. Lézoray, and V.-T. Ta, "Efficient algorithms for image and high dimensional data processing using eikonal equation on graphs", pp. 647–658 in *Advances in Visual Computing* (Las Vegas, NV 2010), edited by G. Bebis et al., Lecture Notes in Computer Science **6454**, 2010.

[Desquesnes et al. 2013] X. Desquesnes, A. Elmoataz, and O. Lézoray, "Eikonal equation adaptation on weighted graphs: fast geometric diffusion process for local and non-local image and data processing", *J. Math. Imaging Vision* **46**:2 (2013), 238–257. MR Zbl

[Elmoataz et al. 2015] A. Elmoataz, M. Toutain, and D. Tenbrinck, "On the p-Laplacian and ∞-Laplacian on graphs with applications in image and data processing", *SIAM J. Imaging Sci.* **8**:4 (2015), 2412–2451. MR Zbl

[Farmer 1994] J. D. Farmer, "Extreme points of the unit ball of the space of Lipschitz functions", *Proc. Amer. Math. Soc.* **121**:3 (1994), 807–813. MR Zbl

[Gilboa 2014] G. Gilboa, "A total variation spectral framework for scale and texture analysis", *SIAM J. Imaging Sci.* **7**:4 (2014), 1937–1961. MR Zbl

[Girault and Raviart 1986] V. Girault and P.-A. Raviart, *Finite element methods for Navier–Stokes equations: theory and algorithms*, Springer Series in Computational Mathematics **5**, Springer, 1986. MR Zbl

[Heinonen 2005] J. Heinonen, *Lectures on Lipschitz analysis*, Report. University of Jyväskylä Department of Mathematics and Statistics **100**, University of Jyväskylä, 2005. MR Zbl

[Hynd et al. 2013] R. Hynd, C. K. Smart, and Y. Yu, "Nonuniqueness of infinity ground states", *Calc. Var. Partial Differential Equations* **48**:3-4 (2013), 545–554. MR Zbl

[Juutinen et al. 1999] P. Juutinen, P. Lindqvist, and J. J. Manfredi, "The ∞-eigenvalue problem", *Arch. Ration. Mech. Anal.* **148**:2 (1999), 89–105. MR Zbl

[Juutinen et al. 2001] P. Juutinen, P. Lindqvist, and J. J. Manfredi, "The infinity Laplacian: examples and observations", pp. 207–217 in *Papers on analysis*, edited by J. Heinonen et al., Rep. Univ. Jyväskylä Dep. Math. Stat. **83**, Univ. Jyväskylä, 2001. MR Zbl

[Kawohl and Horák 2017] B. Kawohl and J. Horák, "On the geometry of the p-Laplacian operator", *Discrete Contin. Dyn. Syst. Ser. S* **10**:4 (2017), 799–813. MR Zbl

[Kawohl and Lindqvist 2006] B. Kawohl and P. Lindqvist, "Positive eigenfunctions for the p-Laplace operator revisited", *Analysis* (*Munich*) **26**:4 (2006), 545–550. MR Zbl

[Kawohl and Novaga 2008] B. Kawohl and M. Novaga, "The p-Laplace eigenvalue problem as $p \to 1$ and Cheeger sets in a Finsler metric", *J. Convex Anal.* **15**:3 (2008), 623–634. MR Zbl

[Kawohl and Schuricht 2007] B. Kawohl and F. Schuricht, "Dirichlet problems for the 1-Laplace operator, including the eigenvalue problem", *Commun. Contemp. Math.* **9**:4 (2007), 515–543. MR Zbl

[Larson 2016] S. Larson, "A bound for the perimeter of inner parallel bodies", *J. Funct. Anal.* **271**:3 (2016), 610–619. MR Zbl

[Lê 2006] A. Lê, "Eigenvalue problems for the p-Laplacian", *Nonlinear Anal.* **64**:5 (2006), 1057–1099. MR Zbl

[Lee et al. 2017] B. Lee, J. Darbon, S. Osher, and M. Kang, "Revisiting the redistancing problem using the Hopf–Lax formula", *J. Comput. Phys.* **330** (2017), 268–281. MR Zbl

[Mémoli and Sapiro 2001] F. Mémoli and G. Sapiro, "Fast computation of weighted distance functions and geodesics on implicit hyper-surfaces", *J. Comput. Phys.* **173**:2 (2001), 730–764. MR Zbl

[Rolewicz 1984] S. Rolewicz, "On optimal observability of Lipschitz systems", pp. 152–158 in *Selected topics in operations research and mathematical economics* (Karlsruhe, 1983), edited by G. Hammer and D. Pallaschke, Lecture Notes in Econom. and Math. Systems **226**, Springer, 1984. MR Zbl

[Rolewicz 1986] S. Rolewicz, "On extremal points of the unit ball in the Banach space of Lipschitz continuous functions", *J. Austral. Math. Soc. Ser. A* **41**:1 (1986), 95–98. MR Zbl

[Sethian 1996] J. A. Sethian, "A fast marching level set method for monotonically advancing fronts", *Proc. Nat. Acad. Sci. U.S.A.* **93**:4 (1996), 1591–1595. MR Zbl

[Smarzewski 1997] R. Smarzewski, "Extreme points of unit balls in Lipschitz function spaces", *Proc. Amer. Math. Soc.* **125**:5 (1997), 1391–1397. MR Zbl

[Sussman 1994] M. Sussman, *A level set approach for computing solutions to incompressible two-phase flow*, Ph.D. thesis, University of California, Los Angeles, 1994, available at https://www.proquest.com/docview/304083099. MR Zbl

[Unser 2019] M. Unser, "A unifying representer theorem for inverse problems and machine learning", preprint, 2019. arXiv

[Weih-Wadman 2019] I. Weih-Wadman, "Notes on Cheeger estimates and nodal sets of the p-Laplacian", course project, 2019, available at https://www.math.mcgill.ca/gantumur/math581w19/pLaplaceNotes.pdf.

[Yu 2007] Y. Yu, "Some properties of the ground states of the infinity Laplacian", *Indiana Univ. Math. J.* **56**:2 (2007), 947–964. MR Zbl

[Zagatti 2014] S. Zagatti, "Maximal generalized solution of eikonal equation", *J. Differential Equations* **257**:1 (2014), 231–263. MR Zbl

Received 21 Jan 2020. Revised 13 May 2020. Accepted 3 Jul 2020.

LEON BUNGERT: leon.bungert@fau.de
Department Mathematik, Universität Erlangen-Nürnberg, Erlangen, Germany

YURY KOROLEV: y.korolev@damtp.cam.ac.uk
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom

MARTIN BURGER: martin.burger@fau.de
Department Mathematik, Universität Erlangen-Nürnberg, Erlangen, Germany

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the submission page.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles are usually in English or French, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not refer to bibliography keys. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and a Mathematics Subject Classification for the article, and, for each author, affiliation (if appropriate) and email address.

Format. Authors are encouraged to use LaTeX and the standard amsart class, but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should normally be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages — Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc. — allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with as many details as you can about how your graphics were generated.

Bundle your figure files into a single archive (using zip, tar, rar or other format of your choice) and upload on the link you been provided at acceptance time. Each figure should be captioned and numbered so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript with all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables.

White Space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

PURE and APPLIED ANALYSIS

vol. 2 no. 3 2020