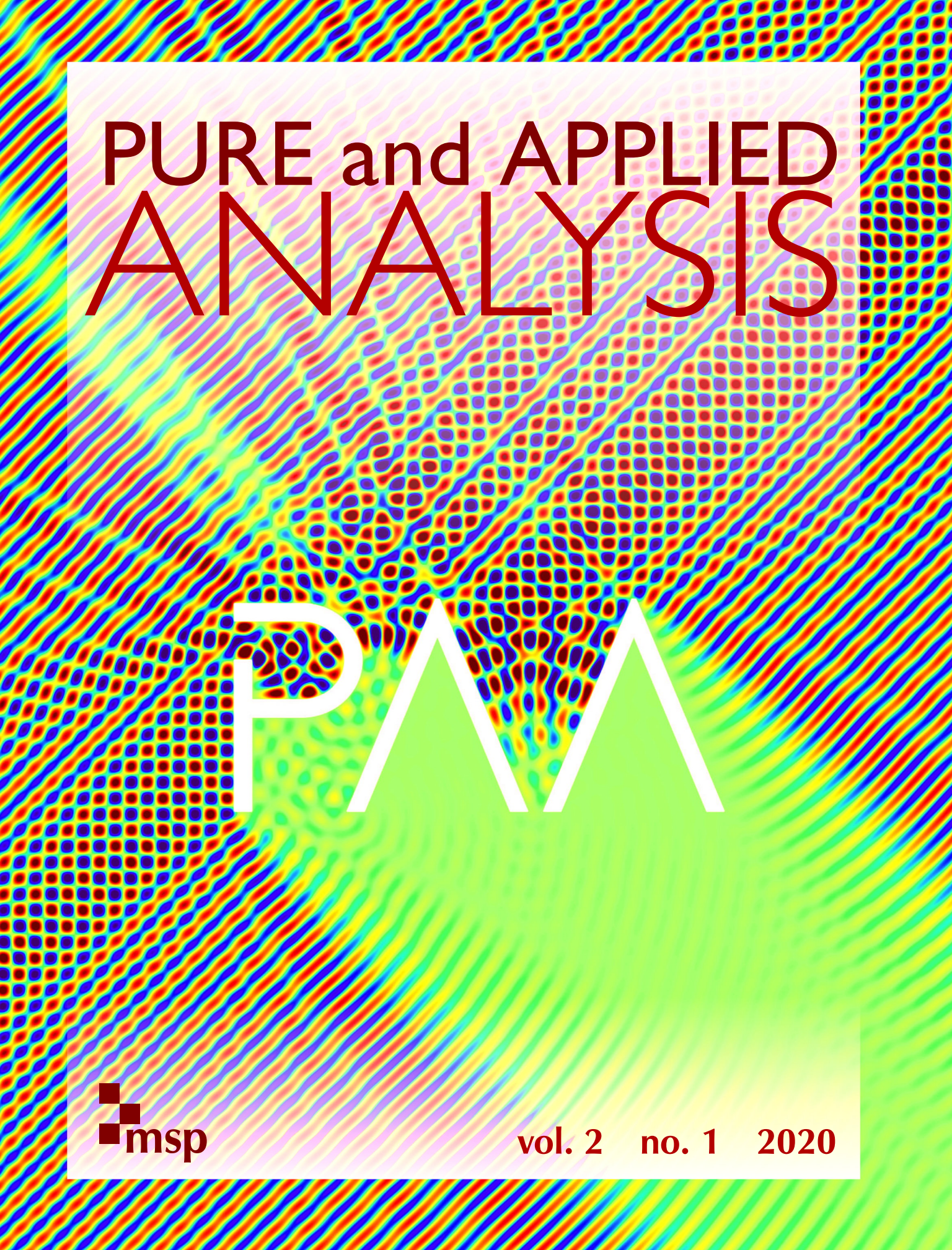

vol. 2 no. 1 2020

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
http://msp.org/

PURE and APPLIED ANALYSIS

Vol. 2, No. 1, 2020
dx.doi.org/10.2140/paa.2020.2.1

msp

EMERGENCE OF NONTRIVIAL MINIMIZERS FOR THE THREE-DIMENSIONAL OHTA–KAWASAKI ENERGY

HANS KNÜPFER, CYRILL B. MURATOV AND MATTEO NOVAGA

This paper is concerned with the diffuse interface Ohta–Kawasaki energy in three space dimensions, in a periodic setting, in the parameter regime corresponding to the onset of nontrivial minimizers. We identify the scaling in which a sharp transition from asymptotically trivial to nontrivial minimizers takes place as the small parameter characterizing the width of the interfaces between the two phases goes to zero, while the volume fraction of the minority phases vanishes at an appropriate rate. The value of the threshold is shown to be related to the optimal binding energy solution of Gamow's liquid drop model of the atomic nucleus. Beyond the threshold the average volume fraction of the minority phase is demonstrated to grow linearly with the distance to the threshold. In addition to these results, we establish a number of properties of the minimizers of the sharp interface screened Ohta–Kawasaki energy in the considered parameter regime. We also establish rather tight upper and lower bounds on the value of the transition threshold.

1. Introduction and main results

The Ohta–Kawasaki energy is a prototypical energy functional in the studies of spatially modulated phases that appear as a result of the competition of short-range attractive and long-range repulsive forces in physical systems of very different nature. Although it was originally introduced in the context of microphase separation in diblock copolymer melts [Ohta and Kawasaki 1986], Ohta–Kawasaki energy is relevant to a wide range of both soft and hard condensed matter systems (for a discussion of the specific physical systems see, e.g., [Muratov 2002]), as well as to dense nuclear matter at the other extreme of energy and spatial scales [Lattimer et al. 1985; Maruyama et al. 2005]. From the mathematical point of view, the Ohta–Kawasaki energy functional, together with the closely related Thomas–Fermi–Dirac–von Weizsäcker energy [Heisenberg 1934; von Weizsäcker 1935; Lieb 1981; Le Bris and Lions 2005], serves as a paradigm for energy-driven pattern-forming systems with competing interactions [Choksi et al. 2017], which is why the associated variational problem has received an increasing amount of attention in recent years [Choksi 2001; Choksi et al. 2009; Spadaro 2009; Muratov 2010; Choksi and Peletier 2011; Goldman et al. 2013; Goldman et al. 2014; Lu and Otto 2014].

In the macroscopic setting, one considers the Ohta–Kawasaki energy functional for configurations defined on a sufficiently large box with periodic boundary conditions; i.e., for $u \in H^1(\mathbb{T}_\ell)$ one sets

$$\mathcal{E}_\varepsilon(u) := \int_{\mathbb{T}_\ell} \left(\tfrac{1}{2}\varepsilon^2|\nabla u|^2 + \tfrac{1}{4}(1-u^2)^2 + \tfrac{1}{2}(u-\bar{u}_\varepsilon)(-\Delta)^{-1}(u-\bar{u}_\varepsilon)\right) dx, \tag{1-1}$$

MSC2010: primary 34A34, 35A15, 35J50; secondary 49Q20, 51P05.
Keywords: energy driven pattern formation, Gamma convergence.

where $\mathbb{T}_\ell$ is the flat d-dimensional torus with side length $\ell > 0$. Furthermore, $\varepsilon > 0$ is the parameter characterizing interfacial thickness and assumed to be sufficiently small. Also, the parameter $\bar{u}_\varepsilon \in (-1, 1)$ denotes the constant background charge density. In the sequel, we will investigate the limit $\varepsilon \to 0$, assuming that u_ε depends on ε suitably. We note that the physically most relevant dimension is $d = 3$. In order for the last term in (1-1) to be well-defined, the definition in (1-1) needs to be supplemented with the "charge neutrality" constraint:

$$\frac{1}{\ell^d} \int_{\mathbb{T}_\ell} u \, dx = \bar{u}_\varepsilon. \tag{1-2}$$

One then wishes to characterize global energy minimizers of the energy in (2-1) for all ℓ sufficiently large. These global energy minimizers are expected to determine the ground states of the corresponding physical system in a macroscopically large sample.

It is widely believed that as the value of ℓ is increased with all other parameters fixed, the global energy minimizer of $\mathcal{E}_\varepsilon$ should be either constant or spatially periodic, with period approaching a constant independent of ℓ as $\ell \to \infty$. Proving such a *crystallization* result would be one of the main challenges in the theory of energy-driven pattern formation and is currently out of reach (for a recent review, see [Blanc and Lewin 2015]), except for the case $d = 1$, $\bar{u}_\varepsilon \in (-1, 1)$ fixed and $\varepsilon > 0$ sufficiently small [Müller 1993; Ren and Wei 2003; Chen and Oshita 2005] (for some results in that direction in higher dimensions, see [Shirokoff et al. 2015; Morini and Sternberg 2014; Daneri and Runa 2018]). On the other hand, it is known that for $\bar{u}_\varepsilon \in (-1, 1)$ fixed, global energy minimizers are not constant as soon as $\varepsilon \ll 1$ and $\ell \gtrsim 1$ [Choksi 2001; Muratov 2010]. This is in contrast with the case $\bar{u}_\varepsilon \notin (-1, 1)$, for which by direct inspection $u = \bar{u}_\varepsilon$ is the unique global minimizer of the energy. Thus, a transition from the trivial minimizer $u = \bar{u}_\varepsilon$ to a nontrivial, spatially nonuniform minimizer of $\mathcal{E}_\varepsilon$ must occur for $\varepsilon \ll 1$ and $\ell \gtrsim 1$ fixed as the value of $\bar{u}_\varepsilon$ increases from $\bar{u}_\varepsilon = -1$ towards $\bar{u}_\varepsilon = 0$ (in view of the symmetry exhibited by the energy when changing $u \to -u$, it is sufficient to consider only the case $\bar{u}_\varepsilon \le 0$). In fact, for $d \ge 2$ nontrivial minimizers emerge at some $\bar{u}_\varepsilon = -1 + \delta_\varepsilon$ with $\varepsilon \lesssim \delta_\varepsilon \lesssim \varepsilon^{2/3} |\ln \varepsilon|^{1/3}$ [Choksi et al. 2009; Muratov 2010], while for $d = 1$ they emerge for some $\varepsilon \lesssim \delta_\varepsilon \lesssim \varepsilon^{1/2}$ [Choksi et al. 2009; Muratov 2002]. The nature of the transition towards nontrivial minimizers is quite delicate and at present not well understood.

In the absence of general results for nontrivial minimizers of $\mathcal{E}_\varepsilon$ for $\varepsilon \lesssim 1$, $\bar{u}_\varepsilon \in (-1, 1)$ and $\ell \gg 1$ in $d \ge 2$, one can consider different asymptotic regimes that admit further analytical characterization. One such regime was analyzed in [Goldman et al. 2013], where the behavior of the minimizers of $\mathcal{E}_\varepsilon$ was studied in the limit $\varepsilon \to 0$ for $\bar{u}_\varepsilon = -1 + \lambda \varepsilon^{2/3} |\ln \varepsilon|^{1/3}$, with $\lambda > 0$ and $\ell > 0$ fixed, in the case $d = 2$. In this regime, nontrivial minimizers are expected to consist of well-separated "droplets" of the minority phase, i.e., regions where $u \simeq +1$ surrounded by the sea of the majority phase where $u \simeq -1$, separated by narrow domain walls of thickness $\sim \varepsilon$. It was found that there exists an explicit critical value of $\lambda = \lambda_c > 0$ such that the minimizers of $\mathcal{E}_\varepsilon$ are nontrivial for all $\lambda > \lambda_c$, while for $\lambda \le \lambda_c$ the minimizers are "asymptotically trivial", namely, that the energy of the minimizers converges to that of $u = \bar{u}_\varepsilon$, and the minimizer converges to $u = \bar{u}_\varepsilon$ in a certain sense as $\varepsilon \to 0$. Moreover, the threshold value λ_c corresponding to the onset of nontrivial minimizers was found to be independent of ℓ, suggesting

that the transition should persist to the macroscopic limit $\ell \to \infty$ with $\varepsilon \ll 1$ and $\bar{u}_\varepsilon$ fixed (i.e., when commuting the order of the $\varepsilon \to 0$ and $\ell \to \infty$ limits). The obtained nontrivial minimizers exhibit a kind of a homogenization limit, with mass distributing uniformly on average throughout the domain. Furthermore, by performing a two-scale expansion of the energy, one can make more precise conclusions about the detailed properties of the minimizers and, in particular, formulate a variational problem in the whole space that determines the placement of the connected components of the minimizers in terms of the so-called renormalized energy, whose minimizers are conjectured to concentrate on the vertices of a hexagonal lattice [Goldman et al. 2014].

Here, we would like to understand how the transition to nontrivial minimizers happens when $\varepsilon \to 0$ and $\ell \gtrsim 1$ in the physical three-dimensional case. Therefore, from now on we fix $d = 3$ throughout the rest of the paper. Once again, in this regime the minimizers are expected to exhibit a two-phase character, with the minority phase occupying a small fraction of space. To this end, we define

$$\bar{u}_\varepsilon := -1 + \lambda \varepsilon^{2/3}, \tag{1-3}$$

where $\lambda > 0$ is fixed. Our main result is the following theorem.

Theorem 1.1. *Let $\ell > 0$ and $\lambda > 0$, and let $\mathcal{E}_\varepsilon$ be defined in* (1-1) *with $\bar{u}_\varepsilon$ given by* (1-3)*. Then, there exists a universal constant $\lambda_c > 0$ such that if u_ε is a minimizer of $\mathcal{E}_\varepsilon(u)$ among all $u \in H^1(\mathbb{T}_\ell)$ satisfying* (1-2)*, and $\mu_\varepsilon \in \mathcal{M}^+(\mathbb{T}_\ell)$ is such that $d\mu_\varepsilon(x) = \frac{1}{2}\varepsilon^{-2/3}(1 + \operatorname{sgn} u_\varepsilon(x))\,dx$, we have as $\varepsilon \to 0$:*

(i) $\mu_\varepsilon \rightharpoonup 0$ *in* $\mathcal{M}(\mathbb{T}_\ell)$ *if* $\lambda \le \lambda_c$.

(ii) $\mu_\varepsilon \rightharpoonup \bar{\mu}$ *in* $\mathcal{M}(\mathbb{T}_\ell)$, *where* $d\bar{\mu} = \frac{1}{2}(\lambda - \lambda_c)\,dx$ *if* $\lambda > \lambda_c$.

(iii) $\varepsilon^{-4/3}\mathcal{E}_\varepsilon(u_\varepsilon) \to \min\{\lambda^2\ell^3, \lambda_c(2\lambda - \lambda_c)\ell^3\}$.

Thus, the onset of nontrivial minimizers in three space dimensions occurs sooner in terms of $0 < 1 + \bar{u}_\varepsilon \ll 1$ than the corresponding transition in two dimensions. In particular, cylindrical morphologies obtained by trivially extending the two-dimensional minimizers into the third dimension are no longer global energy minimizers. One would, therefore, expect that the emergent nontrivial minimizers consist of a collection of well-separated small droplets of the minority phase surrounded by the sea of the majority phase. Furthermore, the size and the distance between the droplets should scale differently (see [Knüpfer et al. 2016]) from those in two dimensions. Furthermore, in contrast to the latter [Goldman et al. 2013; Goldman et al. 2014] we may no longer conclude that the droplets are nearly spherical.

Concerning the threshold λ_c, its precise value may be expressed in terms of several characteristics of the double-well potential appearing in (1-1) and the solution of the nonlocal isoperimetric problem in the whole space that goes back to [Gamow 1930]; see also [Choksi et al. 2017]. Namely, defining

$$\sigma := \frac{2\sqrt{2}}{3}, \quad \kappa := \frac{1}{\sqrt{2}}, \tag{1-4}$$

which are the *interfacial energy* and the *screening parameter*, respectively, we have (for details, see the following sections)

$$\lambda_c = 2^{-1/3}\sigma^{2/3}\kappa^2 f^*, \tag{1-5}$$

where $f^* > 0$ is the minimum energy per unit volume for Gamow's liquid drop model:

$$f^* := \inf_{u \in \mathrm{BV}(\mathbb{R}^3;\{0,1\})} \frac{\int_{\mathbb{R}^3} |\nabla u|\, dx + \frac{1}{8\pi} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} u(x)u(y)/|x-y|\, dx\, dy}{\int_{\mathbb{R}^3} u\, dx}. \tag{1-6}$$

Note that the minimization in (1-6) is equivalent to minimizing the perimeter plus the Coulombic self-energy over all Borel sets, per unit volume.

Although the precise value of f^* is currently unknown, the characterization in (1-5) allows us to obtain a quantitative estimate for the threshold λ_c in Theorem 1.1, using balls as competitors and a recent quantitative nonexistence result for the Gamow's liquid drop model [Frank et al. 2016].

Theorem 1.2. *With the notation of Theorem 1.1, we have*

$$\frac{3}{4\sqrt[3]{2}} \le \lambda_c \le \frac{3}{2\sqrt[3]{5}}. \tag{1-7}$$

Numerically, the bound in (1-7) appears to be fairly tight: $0.5952 < \lambda_c < 0.8773$, with the lower bound to within 33% of the value of the upper bound. Note that if the conjecture that the minimizers of Gamow's liquid drop model are balls is true, then the upper bound in (1-7) should in fact yield equality.

Our proof relies on our previous results obtained for the three-dimensional sharp interface version of the Ohta–Kawasaki energy [Knüpfer et al. 2016]. Together with the approach from [Muratov 2010, Section 4], the result in Theorem 1.1 is obtained along the lines of the arguments in [Goldman et al. 2013], suitably adapted from the two-dimensional to the three-dimensional case. Note, however, that since the sharp interface energy studied by Knüpfer et al. does not include the effect of *charge screening*, their results cannot be directly combined with those of [Muratov 2010]. In fact, there is no transition from trivial to nontrivial minimizers in the unscreened sharp interface energy. Therefore, as a first step towards the proof one needs to adapt the results of Knüpfer et al. to the case of screened sharp interface energy and obtain an asymptotic characterization of its minimizers as $\varepsilon \to 0$.

As in [Knüpfer et al. 2016], we separate the nonlocal energy into the near-field and far-field contributions, with screening appearing explicitly in the latter. At the same time, the self-interaction energy of the droplets turns out to be still well-approximated by that of Gamow's liquid drop model. Combining the far-field with the near-field contributions to the energy then allows us to establish a Γ-convergence result for the screened sharp interface energy to an energy functional which is quadratic in the limit charge density, with the notion of convergence being the weak* convergence of measures (see, for instance, [Knüpfer et al. 2016, Appendix A]). Along the way, we establish uniform estimates for the connected components of the minimizers similar to those in [Knüpfer et al. 2016], which, in turn, allows us to characterize nonexistence of nontrivial minimizers of the screened sharp interface energy for $\lambda < \lambda_c$ and ε sufficiently small.

Once the Γ-convergence result is established for the screened sharp interface energy, we proceed as in [Goldman et al. 2013] by introducing a piecewise-constant charge density associated with the admissible configurations for the diffuse interface energy that eliminates the small deviations of the charge density from their equilibrium values ± 1 for the double-well potential (for a more detailed explanation of the need of such a step, see the beginning of Section 2.2 in [Goldman et al. 2013]). We then adapt the arguments of

[loc. cit., Section 6] to obtain the corresponding Γ-convergence result for the diffuse interface energy to the same quadratic functional in the limit charge density as for the screened sharp interface energy. Finally, explicitly minimizing the limit energy we obtain the main result of our paper contained in Theorem 1.1. Furthermore, we relate the value of the threshold λ_c with the optimal energy per unit mass for Gamow's liquid drop model. In addition, we use recent results in [Frank and Lieb 2015; Frank et al. 2016] characterizing the minimizers of the latter problem to obtain sharp quantitative bounds on the value of the threshold.

To summarize, our paper provides an extension of various recent results for the diffuse interface Ohta–Kawasaki energy to the case of a macroscopic three-dimensional domain, establishing a sharp transition from trivial to nontrivial minimizers in the asymptotic limit of vanishingly thin interfaces. Most of the techniques used in our proofs are adaptations of those that appeared in the earlier studies of this problem in different settings. The main novelty of our results, however, is the way these arguments are combined to yield a nontrivial scaling for the transition to nontrivial minimizers and the limit energy functional for the three-dimensional Ohta–Kawasaki energy. To our knowledge, this is the first sharp asymptotic result for this energy in the regime of strong compositional asymmetry and large number of droplets (for the case of finitely many droplets, see [Choksi and Peletier 2011]). We note that the present lack of knowledge about the minimizers of Gamow's liquid drop model prevents us from going to the next order in a two-scale Γ-expansion to describe local interactions of droplets via a "renormalized energy" [Rougerie and Serfaty 2016]. In particular, it is not known at present whether the minimizer per unit mass exists only for a unique value of the mass (this would be true if minimizers were balls). Thus, further insights into the solution of Gamow's model would be needed to carry out the program realized for the two-dimensional Ohta–Kawasaki energy in [Goldman et al. 2014].

Our paper is organized as follows. In Section 2, we introduce the different energies appearing in our study and state a number of results related to each of the associated variational problems. Also in this section, we prove Theorem 2.2, which gives a quantitative lower bound for the self-interaction energy per unit mass for Gamow's liquid drop model. Then, in Section 3 we state the Γ-convergence result for the sharp interface energy in Theorem 3.2, followed by a proof. Also in Section 3, we provide some further results about the connected components of minimizers of the screened sharp interface energy, see Theorem 3.4 and Corollary 3.5. Finally, in Section 4 we state and prove the corresponding Γ-convergence result for the diffuse interface energy; see Theorem 4.1. The results in Theorems 1.1 and 1.2 are then obtained as simple corollaries of the above theorems.

2. Setting

We now introduce the basic notation used throughout the rest of the paper, together with the assumptions and some technical results.

The diffuse interface energy. We begin by generalizing the diffuse energy functional in (1-1) to one involving an arbitrary symmetric double-well potential $W(u)$:

$$\mathcal{E}_\varepsilon(u) := \int_{\mathbb{T}_\ell} \left(\tfrac{1}{2}\varepsilon^2|\nabla u|^2 + W(u) + \tfrac{1}{2}(u-\bar{u}_\varepsilon)(-\Delta)^{-1}(u-\bar{u}_\varepsilon)\right) dx, \tag{2-1}$$

with $W(u)$ satisfying [Muratov 2010]

(i) $W \in C^2(\mathbb{R})$, $W(u) = W(-u)$, and $W \geq 0$,

(ii) $W(+1) = W(-1) = 0$ and $W''(+1) = W''(-1) > 0$,

(iii) $W''(|u|)$ is monotonically increasing for $|u| \geq 1$, $\lim_{|u|\to\infty} W''(u) = +\infty$, and $|W'(u)| \leq C(1+|u|^q)$ for some $C > 0$ and $1 < q < 5$.

The bounds on the exponent q in condition (iii) are the same as in [Muratov 2010]. In particular, the upper bound $q < 5$ guarantees that minimizers of $\mathcal{E}_\varepsilon$ are bounded and are therefore classical solutions of the Euler–Lagrange equation (2-5). Boundedness of minimizers is also needed to establish the estimate (4-11) in the proof of Theorem 4.1 below.

The energy $\mathcal{E}_\varepsilon$ is well-defined and bounded on the admissible class

$$\mathcal{A}_\varepsilon := \left\{ u \in H^1(\mathbb{T}_\ell) : \frac{1}{\ell^3} \int_{\mathbb{T}_\ell} u \, dx = \bar{u}_\varepsilon \right\}, \tag{2-2}$$

with the nonlocal term interpreted, as usual, with the help of the Green's function $G_0(x)$ solving

$$-\Delta G_0(x) = \delta(x) - \ell^{-3} \quad \text{in } \mathcal{D}'(\mathbb{T}_\ell) \qquad \text{and} \qquad \int_{\mathbb{T}_\ell} G_0(x)\, dx = 0. \tag{2-3}$$

Explicitly, the energy takes the form

$$\mathcal{E}_\varepsilon(u) = \int_{\mathbb{T}_\ell} \Big(\frac{1}{2}\varepsilon^2 |\nabla u|^2 + W(u) \Big)\, dx + \frac{1}{2} \int_{\mathbb{T}_\ell} \int_{\mathbb{T}_\ell} (u(x) - \bar{u}_\varepsilon) G_0(x-y) (u(y) - \bar{u}_\varepsilon)\, dx\, dy, \tag{2-4}$$

noting that the last term in the right-hand side is well-defined by Young's inequality.

Under the above assumptions, every critical point $u \in \mathcal{A}_\varepsilon$ of $\mathcal{E}_\varepsilon$ weakly solves the Euler–Lagrange equation, which can be written as (see [Muratov 2010, Section 4])

$$-\varepsilon^2 \Delta u + W'(u) + v = \Lambda, \quad -\Delta v = u - \bar{u}_\varepsilon, \tag{2-5}$$

where $v \in H^3(\mathbb{T}_\ell)$ is a zero-average solution of the second equation in (2-5) and $\Lambda \in \mathbb{R}$ is the Lagrange multiplier satisfying

$$\Lambda = \frac{1}{\ell^3} \int_{\mathbb{T}_\ell} W'(u)\, dx, \tag{2-6}$$

as can be seen by integrating the first equation in (2-5) over $\mathbb{T}_\ell$. In particular, we have

$$v(x) = \int_{\mathbb{T}_\ell} G_0(x-y)(u(y) - \bar{u}_\varepsilon)\, dy, \tag{2-7}$$

and $u, v \in C^\infty(\mathbb{T}_\ell)$ are classical solutions of (2-5) [Muratov 2010, Section 4]. We note that, by the direct method of calculus of variations, minimizers of $\mathcal{E}_\varepsilon$ are easily seen to exist for all choices of the parameters.

The sharp interface energy with screening. For $\varepsilon \ll 1$, minimizers of $\mathcal{E}_\varepsilon$ are expected to consist of functions which take values close to ± 1, except for narrow transition regions of width of order ε [Muratov 2010]. As usual, we define the energy of an optimal one-dimensional transition layer connecting $u = \pm 1$

[Modica 1987]:

$$\sigma := \int_{-1}^{1} \sqrt{2W(s)}\, ds > 0. \tag{2-8}$$

We also define the so-called screening parameter

$$\kappa := \frac{1}{\sqrt{W''(1)}} > 0. \tag{2-9}$$

This parameter measures the stiffness of the double-well potential near the wells. Physically, it is related to the effect of charge screening appearing in the sharp interface version of the energy $\mathcal{E}_\varepsilon$, introduced in the sequel (as in the Debye–Hückel theory [Landau and Lifshitz 1980; Muratov 2002; Muratov 2010]). With some obvious modifications, the results of [Muratov 2010, Section 4] apply to $\mathcal{E}_\varepsilon$ defined in (2-1), with the corresponding sharp interface energy E_ε defined as

$$E_\varepsilon(u) := \frac{\varepsilon\sigma}{2} \int_{\mathbb{T}_\ell} |\nabla u|\, dx + \frac{1}{2} \int_{\mathbb{T}_\ell} (u - \bar{u}_\varepsilon)(-\Delta + \kappa^2)^{-1}(u - \bar{u}_\varepsilon)\, dx, \tag{2-10}$$

where u belongs to the admissible class

$$\mathcal{A} := \mathrm{BV}(\mathbb{T}_\ell; \{-1, 1\}). \tag{2-11}$$

Specifically, in the considered scaling regime we have the following relation between the two energies (see the following sections):

$$\frac{\min_{u\in\mathcal{A}_\varepsilon} \mathcal{E}_\varepsilon(u)}{\min_{u\in\mathcal{A}} E_\varepsilon(u)} \to 1 \quad \text{as } \varepsilon \to 0. \tag{2-12}$$

Notice that the neutrality constraint in (1-2) is no longer present in the case of the sharp interface energy.

The energy in (2-10) may be rewritten with the help of the Green's function as

$$E_\varepsilon(u) := \frac{\varepsilon\sigma}{2} \int_{\mathbb{T}_\ell} |\nabla u|\, dx + \frac{1}{2} \int_{\mathbb{T}_\ell} \int_{\mathbb{T}_\ell} (u(x) - \bar{u}_\varepsilon) G(x - y)(u(y) - \bar{u}_\varepsilon)\, dx\, dy, \tag{2-13}$$

where G solves

$$-\Delta G(x) + \kappa^2 G(x) = \delta(x) \quad \text{in } \mathcal{D}'(\mathbb{T}_\ell). \tag{2-14}$$

Notice that G has an explicit representation

$$G(x) = \frac{1}{4\pi} \sum_{\boldsymbol{n}\in\mathbb{Z}^3} \frac{e^{-\kappa|x-\boldsymbol{n}\ell|}}{|x - \boldsymbol{n}\ell|}. \tag{2-15}$$

In particular, we have

$$G(x) \simeq \frac{1}{4\pi|x|} \quad |x| \ll 1, \qquad G(x) \geq c \quad \text{for all } x \in \mathbb{T}_\ell, \tag{2-16}$$

for some $c > 0$ depending on κ and ℓ. Also, integrating (2-14) we get

$$\int_{\mathbb{T}_\ell} G(x)\, dx = \kappa^{-2}. \tag{2-17}$$

The latter allows us to rewrite the energy E_ε in an equivalent form in terms of $\chi \in \mathrm{BV}(\mathbb{T}_\ell; \{0,1\})$, where

$$\chi(x) := \frac{1+u(x)}{2}, \quad x \in \mathbb{T}_\ell, \tag{2-18}$$

as

$$E_\varepsilon(u) = \frac{\varepsilon^{4/3}\lambda^2\ell^3}{2\kappa^2} + \varepsilon\sigma \int_{\mathbb{T}_\ell} |\nabla\chi|\,dx - \frac{2\varepsilon^{2/3}\lambda}{\kappa^2}\int_{\mathbb{T}_\ell}\chi\,dx + 2\int_{\mathbb{T}_\ell}\int_{\mathbb{T}_\ell} G(x-y)\chi(x)\chi(y)\,dx\,dy, \tag{2-19}$$

where we also used (1-2).

We now introduce a version of the energy E_ε written in terms of the rescaling

$$\tilde\chi(x) := \chi\left(\frac{\ell x}{\ell_\varepsilon}\right), \quad x \in \mathbb{T}_{\ell_\varepsilon}, \quad \ell_\varepsilon := \left(\frac{4}{\sigma\varepsilon}\right)^{1/3}\ell. \tag{2-20}$$

With this definition we have $\tilde\chi \in \widetilde{\mathcal{A}}_{\ell_\varepsilon}$, where

$$\widetilde{\mathcal{A}}_{\ell_\varepsilon} := \mathrm{BV}(\mathbb{T}_{\ell_\varepsilon}; \{0,1\}), \tag{2-21}$$

for every $\chi \in \mathcal{A}$, and $E_\varepsilon(\chi) = \widetilde{E}_{\ell_\varepsilon}(\tilde\chi)$, with

$$\widetilde{E}_{\ell_\varepsilon}(\tilde\chi) := \frac{\varepsilon^{4/3}\lambda^2\ell^3}{2\kappa^2} - \frac{\varepsilon^{5/3}\sigma\lambda}{2\kappa^2}\int_{\mathbb{T}_{\ell_\varepsilon}}\tilde\chi\,dx + \left(\frac{\varepsilon^{5/3}\sigma^{5/3}}{4^{2/3}}\right)\left[\int_{\mathbb{T}_{\ell_\varepsilon}} |\nabla\tilde\chi|\,dx + \frac{1}{2}\int_{\mathbb{T}_{\ell_\varepsilon}}\tilde\chi(-\Delta + 4^{-2/3}\varepsilon^{2/3}\sigma^{2/3}\kappa^2)^{-1}\tilde\chi\,dx\right]. \tag{2-22}$$

Introducing G_ε, which solves

$$-\Delta G_\varepsilon(x) + 4^{-2/3}\kappa^2\varepsilon^{2/3}\sigma^{2/3}G_\varepsilon(x) = \delta(x) \quad \text{in } \mathbb{T}_{\ell_\varepsilon}, \tag{2-23}$$

we can then express the energy $\widetilde{E}_{\ell_\varepsilon}$ as

$$\widetilde{E}_{\ell_\varepsilon}(\tilde\chi) := \frac{\varepsilon^{4/3}\lambda^2\ell^3}{2\kappa^2} - \frac{\varepsilon^{5/3}\sigma\lambda}{2\kappa^2}\int_{\mathbb{T}_{\ell_\varepsilon}}\tilde\chi\,dx + \left(\frac{\varepsilon^{5/3}\sigma^{5/3}}{4^{2/3}}\right)\left[\int_{\mathbb{T}_{\ell_\varepsilon}} |\nabla\tilde\chi|\,dx + \frac{1}{2}\int_{\mathbb{T}_{\ell_\varepsilon}} G_\varepsilon(x-y)\tilde\chi(x)\tilde\chi(y)\,dx\,dy\right]. \tag{2-24}$$

Note that, as in (2-15), we have the following representation for G_ε:

$$G_\varepsilon(x) = \frac{1}{4\pi}\sum_{\boldsymbol{n}\in\mathbb{Z}^3}\frac{e^{-4^{-1/3}\varepsilon^{1/3}\sigma^{1/3}\kappa|x-\boldsymbol{n}\ell_\varepsilon|}}{|x-\boldsymbol{n}\ell_\varepsilon|}. \tag{2-25}$$

The whole space energy. As was shown by us in [Knüpfer et al. 2016], in the absence of screening, i.e., with $\kappa = 0$ and $u \in \mathcal{A}$ also satisfying (1-2), the asymptotic behavior of the minimizers of E_ε in (2-10) with $\bar u_\varepsilon$ satisfying (1-3) can be expressed in terms of those for the energy defined on the whole of $\mathbb{R}^3$:

$$\widetilde{E}_\infty(\tilde\chi) := \int_{\mathbb{R}^3}|\nabla\tilde\chi|\,dx + \frac{1}{8\pi}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\tilde\chi(x)\tilde\chi(y)}{|x-y|}\,dx\,dy, \tag{2-26}$$

which is well-defined in the admissible class

$$\widetilde{\mathcal{A}}_\infty := \mathrm{BV}(\mathbb{R}^3; \{0, 1\}). \tag{2-27}$$

In particular, the optimal self-energy per unit volume of the minority phase is

$$f^* := \inf_{\tilde{\chi} \in \widetilde{\mathcal{A}}_\infty} \frac{\widetilde{E}_\infty(\tilde{\chi})}{\int_{\mathbb{R}^3} \tilde{\chi}\, dx}. \tag{2-28}$$

Note that within the nuclear physics context, this is precisely the dimensionless form of the celebrated Gamow's liquid drop model of the atomic nucleus [Gamow 1930] (for a recent mathematical overview, see [Choksi et al. 2017]). In particular, the value of f^* corresponds to the energy per nucleon in the tightest bound nucleus.

The relationship between E_ε and $\widetilde{E}_\infty$ can be seen formally by passing to the limit $\varepsilon \to 0$ in (2-24) with $\tilde{\chi}$ taken to be the characteristic function of a fixed bounded set restricted to $\mathbb{T}_{\ell_\varepsilon}$. Then we have

$$\left(\frac{4^{2/3}}{\varepsilon^{5/3}\sigma^{5/3}}\right)\left(\widetilde{E}_{\ell_\varepsilon}(\tilde{\chi}) - \frac{\varepsilon^{4/3}\lambda^2\ell^3}{2\kappa^2}\right) \to -\frac{\lambda f^*}{\lambda_c}\int_{\mathbb{R}^3} \tilde{\chi}\, dx + \widetilde{E}_\infty(\tilde{\chi}), \tag{2-29}$$

where $\tilde{\chi}$ was extended by zero to the whole of $\mathbb{R}^3$ and λ_c is defined in (1-5).

The following result was recently established about minimizers of the problem in the whole space [Knüpfer et al. 2016; Frank et al. 2016].

Theorem 2.1. *There exists a bounded, connected open set $F^* \subset \mathbb{R}^3$ with smooth boundary such that*

$$f^* = \frac{\widetilde{E}_\infty(\tilde{\chi}_{F^*})}{|F^*|}, \tag{2-30}$$

where χ_{F^*} is the characteristic function of the set F^*.

It has been conjectured that the minimizer of $\widetilde{E}_\infty$ with fixed mass is given by a ball whenever such a minimizer exists [Choksi and Peletier 2011]. Therefore, taking a ball of radius R as a test function in (2-28) and optimizing in R, one obtains an estimate

$$f^* \le 3^{5/3} \cdot 2^{-2/3} \cdot 5^{-1/3}. \tag{2-31}$$

The conjecture above would imply that the inequality in (2-31) is in fact an equality. Proving such a result is a difficult hard analysis problem that currently appears to be out of reach. Nevertheless, we can establish a first quantitative lower bound for the value of f^*, using equipartition of energy of F^* established in [Frank and Lieb 2015] and a quantitative upper bound on $|F^*|$ obtained in [Frank et al. 2016]. Note that the resulting lower bound equals about 67% of the upper bound in (2-31). This is one of the main results of the present paper.

Theorem 2.2. *We have*

$$f^* \ge \frac{3^{5/3}}{4}. \tag{2-32}$$

Proof. Let F^* be a minimizer from Theorem 2.1, and write

$$f^* = \frac{P(F^*) + V(F^*)}{|F^*|}, \tag{2-33}$$

where $P(F^*)$ is the perimeter of F^* and $V(F^*)$ is the Coulombic self-energy of F^*. By the result from [Frank and Lieb 2015], the energy exhibits equipartition of energy in the sense that

$$V(F^*) = \tfrac{1}{2} P(F^*), \tag{2-34}$$

which can be easily seen by considering the sets λF^* as competitors for f^* and taking advantage of the homogeneity of P and V with respect to dilations. Thus, we have

$$f^* = \frac{3P(F^*)}{2|F^*|}. \tag{2-35}$$

Therefore, applying the isoperimetric inequality yields

$$f^* \geq \left(\frac{243\pi}{2|F^*|}\right)^{1/3}. \tag{2-36}$$

The proof is then concluded by recalling the quantitative upper bound $|F^*| \leq 32\pi$ from [Frank et al. 2016]. □

The limit energy. For $\mu \in \mathcal{M}^+(\mathbb{T}_\ell) \cap H^{-1}(\mathbb{T}_\ell)$, define

$$E_0(\mu) := \frac{\lambda^2 \ell^3}{2\kappa^2} - \frac{2}{\kappa^2}(\lambda - \lambda_c) \int_{\mathbb{T}_\ell} d\mu + 2 \int_{\mathbb{T}_\ell} \int_{\mathbb{T}_\ell} G(x-y)\, d\mu(x)\, d\mu(y). \tag{2-37}$$

Note that $\mu \in \mathcal{M}^+(\mathbb{T}_\ell) \cap H^{-1}(\mathbb{T}_\ell)$ implies that μ is a nonnegative Radon measure with bounded Coulombic energy:

$$\int_{\mathbb{T}_\ell} \int_{\mathbb{T}_\ell} G(x-y)\, d\mu(x)\, d\mu(y) < \infty, \tag{2-38}$$

where G is the screened Coulombic kernel from (2-15). The converse is also true; i.e., a positive Radon measure with bounded Coulombic energy defines a bounded linear functional on $H^1(\mathbb{T}_\ell)$. This fact is a consequence of the following lemma, whose proof is a straightforward adaptation of the proof of [Goldman et al. 2013, Lemma 3.2] in two dimensions (see [Knüpfer et al. 2016, Appendix A]). In particular, it allows us to extend the definition of E_0 to arbitrary positive Radon measures on $\mathbb{T}_\ell$, with $E_0(\mu) < +\infty$ if and only if $\mu \in H^{-1}(\mathbb{T}_\ell)$.

Lemma 2.3. *Let $\mu \in \mathcal{M}^+(\mathbb{T}_\ell)$ and let* (2-38) *hold. Then*:

(i) $\mu \in H^{-1}(\mathbb{T}_\ell)$, *in the sense that it can be extended to a bounded linear functional over $H^1(\mathbb{T}_\ell)$.*

(ii) *If*

$$v(x) := \int_{\mathbb{T}_\ell} G(x-y)\, d\mu(y), \tag{2-39}$$

then $v \in H^1(\mathbb{T}_\ell)$. Furthermore, v solves

$$-\Delta v + \kappa^2 v = \mu \tag{2-40}$$

weakly in $H^1(\mathbb{T}_\ell)$, and

$$\nabla v(x) = \int_{\mathbb{T}_\ell} \nabla G(x-y)\, d\mu(y), \tag{2-41}$$

in the sense of distributions.

(ii) *If v is as in* (ii), *we have $\kappa^2 \int_{\mathbb{T}_\ell} v\, dx = \int_{\mathbb{T}_\ell} d\mu$ and*

$$\int_{\mathbb{T}_\ell}\int_{\mathbb{T}_\ell} G(x-y)\, d\mu(x)\, d\mu(y) = \int_{\mathbb{T}_\ell} (|\nabla v|^2 + \kappa^2 v^2)\, dx. \tag{2-42}$$

According to Lemma 2.3, the energy E_0 may be equivalently rewritten in terms of the associated potential v in (2-39) as

$$E_0(\mu) = \frac{\lambda^2 \ell^3}{2\kappa^2} - 2(\lambda - \lambda_c)\int_{\mathbb{T}_\ell} v\, dx + 2\int_{\mathbb{T}_\ell} (|\nabla v|^2 + \kappa^2 v^2)\, dx, \tag{2-43}$$

and minimizing $E_0(\mu)$ over $\mu \in \mathcal{M}^+(\mathbb{T}_\ell) \cap H^{-1}(\mathbb{T}_\ell)$ is the same as minimizing the right-hand side of (2-43) with respect to all $v \in H^1(\mathbb{T}_\ell)$ such that $v \geq 0$ in $\mathbb{T}_\ell$ and $-\Delta v + \kappa^2 v \in \mathcal{M}^+(\mathbb{T}_\ell)$. By inspection, the latter is minimized by $v = \bar{v}$, where

$$\bar{v} = \begin{cases} 0, & \lambda \leq \lambda_c, \\ (1/(2\kappa^2))(\lambda - \lambda_c), & \lambda > \lambda_c. \end{cases} \tag{2-44}$$

In terms of the measures, we can state this result as follows:

Proposition 2.4. *The energy $E_0(\mu)$ is minimized by a unique measure $\bar{\mu}$ among all $\mu \in \mathcal{M}^+(\mathbb{T}_\ell) \cap H^{-1}(\mathbb{T}_\ell)$, with $\bar{\mu} = 0$ for all $\lambda \leq \lambda_c$, and $d\bar{\mu} = \frac{1}{2}(\lambda - \lambda_c)dx$ for $\lambda > \lambda_c$, respectively. Moreover, we have*

$$E_0(\bar{\mu}) = \begin{cases} \lambda^2 \ell^3/(2\kappa^2), & \lambda \leq \lambda_c, \\ \lambda_c(2\lambda - \lambda_c)\ell^3/(2\kappa^2), & \lambda > \lambda_c, \end{cases} \tag{2-45}$$

and (2-40) *is solved by $v(x) = \bar{v}$.*

3. Sharp interface energy E_ε

We now consider the sharp-interface functional E_ε defined in (2-10) in the limit $\varepsilon \to 0$ with $\bar{u}_\varepsilon$ given by (1-3) and positive σ, λ, κ, ℓ fixed. For a given sequence $(u_\varepsilon) \in \mathcal{A}$, we introduce a measure μ_ε that is continuous with respect to the Lebesgue measure on $\mathbb{T}_\ell$ and whose density is an appropriately rescaled characteristic function of the minority phase:

$$d\mu_\varepsilon(x) := \tfrac{1}{2}\varepsilon^{-2/3}(1 + u_\varepsilon(x))dx. \tag{3-1}$$

Note that by definition the measure μ_ε is nonnegative. We also introduce the potential v_ε via

$$-\Delta v_\varepsilon + \kappa^2 v_\varepsilon = \mu_\varepsilon \quad \text{in } \mathbb{T}_\ell. \tag{3-2}$$

Our first result establishes compactness of sequences with bounded energy after a suitable rescaling.

Theorem 3.1 (equicoercivity). *Let $(u_\varepsilon) \in \mathcal{A}$ be such that*

$$\limsup_{\varepsilon\to 0} \varepsilon^{-4/3} E_\varepsilon(u_\varepsilon) < +\infty, \tag{3-3}$$

and let μ_ε and v_ε be defined in (3-1) *and* (3-2)*, respectively. Then, up to extraction of a subsequence, we have*

$$\mu_\varepsilon \rightharpoonup \mu \quad \text{in } \mathcal{M}(\mathbb{T}_\ell), \qquad v_\varepsilon \rightharpoonup v \quad \text{in } H^1(\mathbb{T}_\ell), \tag{3-4}$$

as $\varepsilon \to 0$, for some $\mu \in \mathcal{M}^+(\mathbb{T}_\ell) \cap H^{-1}(\mathbb{T}_\ell)$ and $v \in H^1(\mathbb{T}_\ell)$ satisfying

$$-\Delta v + \kappa^2 v = \mu \quad \text{in } \mathbb{T}_\ell. \tag{3-5}$$

Proof. Inserting (3-1) into (2-10) and dropping the perimeter term, following the argument of [Muratov 2010] we arrive at (see also (2-19))

$$E_\varepsilon(u_\varepsilon) \geq \frac{\lambda^2 \ell^3}{2\kappa^2} - \frac{2\lambda}{\kappa^2} \int_{\mathbb{T}_\ell} d\mu_\varepsilon(x) + 2 \int_{\mathbb{T}_\ell} \int_{\mathbb{T}_\ell} G(x-y)\, d\mu_\varepsilon(x)\, d\mu_\varepsilon(y), \tag{3-6}$$

where we used (2-17) and (1-3) and took into account the translational invariance of the problem in $\mathbb{T}_\ell$. By (2-16) we get

$$E_\varepsilon(u_\varepsilon) \geq -\frac{2\lambda}{\kappa^2} \mu_\varepsilon(\mathbb{T}_\ell) + 2c\mu_\varepsilon^2(\mathbb{T}_\ell), \tag{3-7}$$

where we again recall that μ_ε is nonnegative by definition. It then follows that

$$\mu_\varepsilon(\mathbb{T}_\ell) < C \tag{3-8}$$

for some constant $C > 0$ independent of ε, which implies that $\mu_\varepsilon \rightharpoonup \mu$ up to a subsequence by the Banach–Alaoglu theorem (see [Brezis 2011, Theorem 3.16]). The considerations above together with Lemma 2.3(iii) and (3-3) show that

$$\int_{\mathbb{T}_\ell} (|\nabla v_\varepsilon|^2 + \kappa^2 v_\varepsilon^2)\, dx = \int_{\mathbb{T}_\ell} \int_{\mathbb{T}_\ell} G(x-y)\, d\mu_\varepsilon(x)\, d\mu_\varepsilon(y) < C, \tag{3-9}$$

and upon extraction of a further subsequence by the Banach–Alaoglu theorem we get $v_\varepsilon \rightharpoonup v$ in $H^1(\mathbb{T}_\ell)$. Finally, (3-5) follows by passing to the limit in (3-2). □

We now proceed to the main result of this section which establishes the Γ-limit of the screened sharp interface energy, similar to its two-dimensional analog in [Goldman et al. 2013, Theorem 1] (see also [Knüpfer et al. 2016, Theorem 3.3] for the unscreened case).

Theorem 3.2 (Γ-convergence of E_ε). *As $\varepsilon \to 0$ we have*

$$\varepsilon^{-4/3} E_\varepsilon \xrightarrow{\Gamma} E_0, \tag{3-10}$$

with respect to the weak convergence of measures. More precisely, we have:

(i) *Lower bound*: *Suppose that* $(u_\varepsilon) \in \mathcal{A}$ *and let* μ_ε *be defined as in* (3-1), *and suppose that*

$$\mu_\varepsilon \rightharpoonup \mu \quad \text{in } \mathcal{M}(\mathbb{T}_\ell), \tag{3-11}$$

as $\varepsilon \to 0$, *for some* $\mu \in \mathcal{M}_\ell{}^+(\mathbb{T}_\ell)$. *Then*

$$\liminf_{\varepsilon\to 0} \varepsilon^{-4/3} E_\varepsilon(u_\varepsilon) \geq E_0(\mu). \tag{3-12}$$

(ii) *Upper bound*: *Given* $\mu \in \mathcal{M}^+(\mathbb{T}_\ell)$, *there exists* $(u_\varepsilon) \in \mathcal{A}$ *such that for the corresponding* μ_ε *as in* (3-1) *we have*

$$\mu_\varepsilon \rightharpoonup \mu \quad \text{in } \mathcal{M}(\mathbb{T}_\ell), \tag{3-13}$$

as $\varepsilon \to 0$, *and*

$$\limsup_{\varepsilon\to 0} \varepsilon^{-4/3} E_\varepsilon(u_\varepsilon) \leq E_0(\mu). \tag{3-14}$$

Proof. Assume first that $\mu \in \mathcal{M}^+(\mathbb{T}_\ell) \cap H^{-1}(\mathbb{T}_\ell)$, so that $E_0(\mu) < +\infty$. As in the proof of Propositions 5.1 and 5.2 in [Knüpfer et al. 2016], we separate the contributions of the near-field and far-field interaction; i.e., for $0 < \rho \leq \frac{1}{4}$ we write

$$G_\rho(x) = \eta_\rho(x) G(x), \quad H_\rho(x) := G(x) - G_\rho(x), \tag{3-15}$$

where $\eta_\rho(x)$ is a smooth cutoff function depending on $|x|$ which is monotonically increasing from 0 to 1 as $|x|$ goes from 0 to ρ, with $\eta_\rho(x) = 0$ for all $|x| < \frac{1}{2}\rho$ and $\eta_\rho(x) = 1$ for all $|x| > \rho$. With the help of (2-19), for any $u_\varepsilon \in \mathcal{A}$ we decompose the energy as $E_\varepsilon = E_\varepsilon^{(1)} + E_\varepsilon^{(2)}$, where

$$\begin{aligned} \varepsilon^{-4/3} E_\varepsilon^{(1)}(u_\varepsilon) &= \frac{\lambda^2 \ell^3}{2\kappa^2} - \frac{2\lambda}{\kappa^2} \int_{\mathbb{T}_\ell} d\mu_\varepsilon(x) + 2 \int_{\mathbb{T}_\ell} \int_{\mathbb{T}_\ell} G_\rho(x-y)\, d\mu_\varepsilon(x)\, d\mu_\varepsilon(y), \\ \varepsilon^{-4/3} E_\varepsilon^{(2)}(u_\varepsilon) &= \varepsilon^{-1/3}\sigma \int_{\mathbb{T}_\ell} |\nabla \chi_\varepsilon|\, dx + 2\varepsilon^{-4/3} \int_{\mathbb{T}_\ell} \int_{\mathbb{T}_\ell} H_\rho(x-y)\chi_\varepsilon(x)\chi_\varepsilon(y)\, dx\, dy, \end{aligned} \tag{3-16}$$

where χ_ε is as in (2-18) with u replaced with u_ε. The term $E_\varepsilon^{(1)}$ is continuous with respect to the weak convergence of measures; hence

$$\int_{\mathbb{T}_\ell} \int_{\mathbb{T}_\ell} G_\rho(x-y)\, d\mu_\varepsilon(x)\, d\mu_\varepsilon(y) \to \int_{\mathbb{T}_\ell} \int_{\mathbb{T}_\ell} G_\rho(x-y)\, d\mu(x)\, d\mu(y) \quad \text{as } \varepsilon \to 0. \tag{3-17}$$

The proof of the lower bound for $E_\varepsilon^{(2)}$ follows with similar arguments as in the proof of [Knüpfer et al. 2016, Theorem 3.3]. After the rescaling in (2-20), one can write

$$\varepsilon^{-4/3} E_\varepsilon^{(2)}(u_\varepsilon) = \left(\frac{\varepsilon^{1/3}\sigma^{5/3}}{4^{2/3}}\right)\left[\int_{\mathbb{T}_{\ell_\varepsilon}} |\nabla \tilde{\chi}_\varepsilon|\, dx + \frac{1}{2} \int_{\mathbb{T}_{\ell_\varepsilon}} \int_{\mathbb{T}_{\ell_\varepsilon}} \widetilde{H}_\rho^\varepsilon(x-y)\tilde{\chi}_\varepsilon(x)\tilde{\chi}_\varepsilon(y)\, dx\, dy\right], \tag{3-18}$$

where $\tilde{\chi}_\varepsilon(x) := \chi_\varepsilon(x\ell/\ell_\varepsilon)$ and

$$\widetilde{H}_\rho^\varepsilon(x) := (1 - \eta_\rho(x\ell/\ell_\varepsilon)) G_\varepsilon(x). \tag{3-19}$$

Observe that by (2-25) and monotonicity of $\eta_\rho(x)$ in $|x|$ we have

$$\widetilde{H}_\rho^\varepsilon(x) \geq (1-\rho)\Gamma_{\rho_0}^{\#}(x),$$

where $\Gamma^{\#}_{\rho_0}(x) := (1 - \eta_{\rho_0}(x))\Gamma^{\#}(x)$ and $\Gamma^{\#}(x) := 1/(4\pi|x|)$ is the restriction of the Newton potential on the torus, for any $\rho_0 > 0$ and all ε small enough depending only on κ, σ and ρ_0. The rest of the proof of the lower bound, as well as the proof of the upper bound, follows exactly as in [Knüpfer et al. 2016].

Finally, if $\mu \notin H^{-1}(\mathbb{T}_\ell)$, then $E_0(\mu) = +\infty$ and the upper bound is trivial, while the lower bound follows via a contradiction argument from the compactness result established in Theorem 3.1. □

As a direct consequence of Theorems 3.1 and 3.2, we have the following characterization of the minimizers of the sharp interface energy in the limit $\varepsilon \to 0$.

Corollary 3.3. *Let $(u_\varepsilon) \in \mathcal{A}$ be minimizers of E_ε. Let μ_ε be defined in* (3-1) *and let v_ε be the solution of* (3-2)*. Then as $\varepsilon \to 0$, we have*

$$\mu_\varepsilon \rightharpoonup \bar{\mu} \quad \text{in } \mathcal{M}(\mathbb{T}_\ell), \qquad v_\varepsilon \rightharpoonup \bar{v} \quad \text{in } H^1(\mathbb{T}_\ell), \tag{3-20}$$

where $\bar{\mu}$ and $\bar{v}$ are as in Proposition 2.4.

We note that for $\lambda \gg \lambda_c$ the minimum energy per unit volume for minimizers in Corollary 3.3 approaches asymptotically that of the unscreened sharp interface energy studied in [Knüpfer et al. 2016], indicating that the presence of an additional screening does not affect the limit behavior of the energy at higher densities than those appearing in (1-3). We would thus expect that the same result would still hold for the sharp interface energy even for $1 + \bar{u}_\varepsilon = o(1)$ as $\varepsilon \to 0$, consistently with a recent result for the sharp interface energy without screening [Emmert et al. 2018].

We conclude by proving an analog of [Knüpfer et al. 2016, Theorem 3.6] that provides uniform bounds on the diameter of the connected components of minimizers of E_ε as $\varepsilon \to 0$, and convergence of most of the connected components to minimizers of Gamow's model per unit mass.

Theorem 3.4 (minimizers: droplet structure). *For $\lambda > 0$, let $(u_\varepsilon) \in \mathcal{A}$ be regular representatives of minimizers of E_ε, and assume that the sets $\{u_\varepsilon = +1\}$ are nonempty for ε sufficiently small. Let N_ε be the number of connected components of the set $\{u_\varepsilon = +1\}$, let $\chi_{\varepsilon,k} \in \mathrm{BV}(\mathbb{R}^3; \{0, 1\})$ be the characteristic function of the k-th connected component of the support of the periodic extension of $\{u_\varepsilon = +1\}$ to the whole of $\mathbb{R}^3$ modulo translations in $\mathbb{Z}^3$, and let $x_{\varepsilon,k} \in \mathrm{supp}(\chi_{\varepsilon,k})$. Then there exists $\varepsilon_0 > 0$ such that the following properties hold*:

(i) *There exist constants $C, c > 0$ depending only on σ, κ, λ and ℓ such that for all $\varepsilon \le \varepsilon_0$ we have*

$$0 < v_\varepsilon \le C \quad \text{and} \quad \int_{\mathbb{R}^3} \chi_{\varepsilon,k}\, dx \ge c\varepsilon, \tag{3-21}$$

where v_ε solves (3-2)*. Moreover we have*

$$\mathrm{supp}(\chi_{\varepsilon,k}) \subseteq B_{C\varepsilon^{1/3}}(x_{\varepsilon,k}). \tag{3-22}$$

(ii) *If $\lambda > \lambda_c$, where λ_c is given by* (1-5)*, there exist constants $C, c > 0$ as above such that for all $\varepsilon \le \varepsilon_0$ we have*

$$c(\lambda - \lambda_c)\varepsilon^{-1/3} \le N_\varepsilon \le C(\lambda - \lambda_c)\varepsilon^{-1/3}. \tag{3-23}$$

Moreover, there exists $\widetilde{N}_\varepsilon \le N_\varepsilon$ *with* $\widetilde{N}_\varepsilon/N_\varepsilon \to 1$ *as* $\varepsilon \to 0$ *and a subsequence* $\varepsilon_n \to 0$ *such that for every* $k_n \le \widetilde{N}_{\varepsilon_n}$ *the following holds*: *after possibly relabeling the connected components, we have*

$$\tilde{\chi}_n \to \tilde{\chi} \quad \text{in } L^1(\mathbb{R}^3), \tag{3-24}$$

where $\tilde{\chi}_n(x) := \chi_{\varepsilon_n,k_n}(\varepsilon_n^{1/3}(x + x_{\varepsilon_n,k_n}))$, *and* $\tilde{\chi} \in \widetilde{\mathcal{A}}_\infty$ *is a minimizer of the right-hand side of* (2-28).

Proof. The proof can be obtained as in [Knüpfer et al. 2016, Theorem 3.6], with some simplifications due to the absence of a volume constraint. We outline the necessary modifications below. As stated above, the constants in the estimates below depend on σ, κ, λ and ℓ, and may change from line to line.

For $\tilde{u}_\varepsilon(x) := u_\varepsilon(\ell x/\ell_\varepsilon)$, we define $F \subset \mathbb{T}_{\ell_\varepsilon}$ to be the set $\{\tilde{u}_\varepsilon = +1\}$, which by our assumption is nonempty for ε sufficiently small. Then we can write $v_\varepsilon(x) = (\sigma/4)^{2/3} v_F(\ell x/\ell_\varepsilon)$, where $v_F(x) := \int_{\mathbb{T}_{\ell_\varepsilon}} G_\varepsilon(x-y)\chi_F(y)\,dy$, and G_ε is defined in (2-25). The first step in the proof is to obtain an L^∞-bound on the potential v_F analogous to the one in [Knüpfer et al. 2016, Lemma 6.3]:

$$0 < v_F \le C\varepsilon^{-2/9}. \tag{3-25}$$

Observe that by strict positivity of G_ε we clearly have $v_F > 0$. On the other hand, the upper bound follows exactly as in [Knüpfer et al. 2016, Lemma 6.3], due to the fact that $G_\varepsilon(x) \le C/|x|$ for some $C > 0$, since in view of (2-15) we have

$$G_\varepsilon(x) = \frac{\ell}{\ell_\varepsilon} G\left(\frac{\ell x}{\ell_\varepsilon}\right) = \frac{1}{4\pi} \sum_{\boldsymbol{n}\in\mathbb{Z}^3} \frac{e^{-\kappa\ell|(x/\ell_\varepsilon)-\boldsymbol{n}\ell|}}{|x-\boldsymbol{n}\ell|}. \tag{3-26}$$

Next, we need to estimate the gradient of v_F pointwise in terms v_F itself, as in [Knüpfer et al. 2016, Lemma 6.5], which relies on [Knüpfer et al. 2016, equation (6.15)]. It is easy to see that the latter estimate still holds in the present setting, with the constants depending on κ and ℓ. The proof then follows as in [Knüpfer et al. 2016], with a few simplifications due to positivity of G. Also, since v_F satisfies

$$-\Delta v_F + \left(\tfrac{1}{4}\varepsilon\sigma\right)^{2/3}\kappa^2 v_F = \left(\tfrac{1}{4}\sigma\right)^{2/3}\chi_F \quad \text{in } \mathbb{T}_{\ell_\varepsilon}, \tag{3-27}$$

by the positivity of v_F we have that v_F is subharmonic outside $\overline{F}$. Thus, v_F attains its global maximum in $\mathbb{T}_{\ell_\varepsilon}$ for some $\bar{x} \in \overline{F}$, and the analog of [Knüpfer et al. 2016, equation (6.19)] holds true:

$$v_F(x) \ge \tfrac{3}{4} v_F(\bar{x}) - C \quad \text{for all } x \in B_r(\bar{x}), \tag{3-28}$$

for some $C > 0$ and $r > 0$.

Proceeding as in [Knüpfer et al. 2016, Lemma 6.7 and Proposition 6.2], we establish a lower density estimate for F: given $x_0 \in \overline{F}$ and letting F_0 be the connected component of F containing x_0, we have

$$|F_0 \cap B_r(x_0)| \ge cr^3 \quad \text{for all } r \le C\min(1, \|v_F\|_\infty^{-1}) \le C\varepsilon^{2/9}, \tag{3-29}$$

for some $c, C > 0$, where the last inequality follows from (3-25). The assertion in (3-21) then follows as in [Knüpfer et al. 2016, Theorem 6.9] from (3-25), (3-28) and (3-29). The idea of the proof in [Knüpfer et al. 2016] is to find a suitable competitor F' which is obtained by cutting from F a ball of radius independent of ε, centered at the point where the potential v_F attains its maximum. Compared to [Knüpfer et al. 2016],

the proof here is simpler since we don't have a volume constraint, so that we can allow competitors with smaller volume than F. Arguing by contradiction, if the maximum of v_F is large, then necessarily the density of F in the ball has to be small, otherwise the energy of F' would be less than the energy of F. However, this contradicts the density estimate in (3-29). Finally, exactly as in [Knüpfer et al. 2016, Lemma 6.11], the bound on the potential and the density estimate (3-29) also imply the diameter bound

$$\operatorname{diam}(F_0) \le C \tag{3-30}$$

for some constant $C > 0$, which gives (3-22). This concludes the proof of part (i).

The proof of part (ii) follows as in the proof of [Knüpfer et al. 2016, Theorem 3.6], with the exception that the estimate on N_ε in (3-23) now follows from (3-21) and the fact that, recalling Corollary 3.3,

$$\lim_{\varepsilon\to 0} \int_{\mathbb{T}_\ell} d\mu_\varepsilon = \bar{\mu}(\mathbb{T}_\ell) = \tfrac{1}{2}(\lambda - \lambda_c), \tag{3-31}$$

where $\bar{\mu}$ is as in Proposition 2.4. □

The results obtained in Theorem 3.4 allow us to establish a sharp transition from trivial to nontrivial minimizers at the level of the sharp interface energy near $\lambda = \lambda_c$ for all $\varepsilon \ll 1$.

Corollary 3.5. *There exists $\varepsilon_0 = \varepsilon_0(\sigma, \kappa, \lambda, \ell) > 0$ such that if λ_c is given by* (1-5), *then*:

(i) *For any $\lambda < \lambda_c$ and $\varepsilon < \varepsilon_0$ we have that $u = -1$ is the unique minimizer of E_ε in $\mathcal{A}$.*

(ii) *For $\lambda > \lambda_c$ and $\varepsilon < \varepsilon_0$ we have that $u = -1$ is not a minimizer of E_ε in $\mathcal{A}$.*

Proof. Since the statement in (ii) follows immediately from Corollary 3.3, we only need to demonstrate (i). The strategy is analogous to the one used in the proof of [Muratov 2010, Proposition 3.2]. For $\lambda < \lambda_c$, let u_ε be a minimizer of E_ε over $\mathcal{A}$, and assume, by contradiction, that $u_\varepsilon \ne -1$ for a sequence of $\varepsilon \to 0$. Let $\chi_{\varepsilon,k}$ be as in Theorem 3.4. By (2-15) we have

$$E_\varepsilon(u_\varepsilon) \ge \frac{\varepsilon^{4/3}\lambda^2\ell^3}{2\kappa^2} + \sum_{k=1}^{N_\varepsilon}\bigg(\varepsilon\sigma\int_{\mathbb{R}^3} |\nabla\chi_{\varepsilon,k}|\,dx - \frac{2\varepsilon^{2/3}\lambda}{\kappa^2}\int_{\mathbb{R}^3}\chi_{\varepsilon,k}\,dx + \frac{1}{2\pi}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{e^{-\kappa|x-y|}}{|x-y|}\chi_{\varepsilon,k}(x)\chi_{\varepsilon,k}(y)\,dx\,dy\bigg). \tag{3-32}$$

At the same time, since by Theorem 3.4 the diameter of the support of $\chi_{\varepsilon,k}$ is bounded above by $C\varepsilon^{1/3}$, for every $\delta > 0$ we have $e^{-\kappa|x-y|} \ge 1-\delta$ for all ε sufficiently small and all $x, y \in \operatorname{supp}(\chi_{\varepsilon,k})$. Introducing $\tilde{\chi}_{\varepsilon,k}(x) := \chi_{\varepsilon,k}(\ell_\varepsilon x/\ell)$ as in (2-20), we can then write

$$\begin{aligned} E_\varepsilon(u_\varepsilon) &\ge \frac{\varepsilon^{4/3}\lambda^2\ell^3}{2\kappa^2} - \frac{\varepsilon^{5/3}\sigma\lambda}{2\kappa^2}\sum_{k=1}^{N_\varepsilon}\int_{\mathbb{R}^3}\tilde{\chi}_{\varepsilon,k}\,dx \\ &\quad + \frac{\varepsilon^{5/3}\sigma^{5/3}}{4^{2/3}}\sum_{k=1}^{N_\varepsilon}\bigg(\int_{\mathbb{R}^3}|\nabla\tilde{\chi}_{\varepsilon,k}|\,dx + \frac{1-\delta}{8\pi}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\tilde{\chi}_{\varepsilon,k}(x)\tilde{\chi}_{\varepsilon,k}(y)}{|x-y|}\,dx\,dy\bigg) \qquad (3\text{-}33) \\ &\ge \frac{\varepsilon^{4/3}\lambda^2\ell^3}{2\kappa^2} - \frac{\varepsilon^{5/3}\sigma\lambda}{2\kappa^2}\sum_{k=1}^{N_\varepsilon}\int_{\mathbb{R}^3}\tilde{\chi}_{\varepsilon,k}\,dx + \frac{\varepsilon^{5/3}\sigma^{5/3}(1-\delta)}{4^{2/3}}\sum_{k=1}^{N_\varepsilon}\widetilde{E}_\infty(\tilde{\chi}_{\varepsilon,k}). \end{aligned}$$

Now we substitute the definitions of f^* and λ_c in (2-28) and (1-5), respectively, into (3-33). This yields

$$E_\varepsilon(u_\varepsilon) \geq \frac{\varepsilon^{4/3}\lambda^2\ell^3}{2\kappa^2} + \frac{\varepsilon^{5/3}\sigma((1-\delta)\lambda_c - \lambda)}{2\kappa^2} \sum_{k=1}^{N_\varepsilon} \int_{\mathbb{R}^3} \tilde{\chi}_{\varepsilon,k}\, dx. \tag{3-34}$$

In particular, for $\lambda < \lambda_c$ one can choose δ small enough so that $E_\varepsilon(u_\varepsilon) > \varepsilon^{4/3}\lambda^2\ell^3/(2\kappa^2) = E_\varepsilon(-1)$ for all ε sufficiently small, contradicting minimality of u_ε. □

4. Diffuse interface energy $\mathcal{E}_\varepsilon$

We now consider the diffuse-interface functional $\mathcal{E}_\varepsilon$ defined in (1-1) in the limit $\varepsilon \to 0$ with, as before, $\bar{u}_\varepsilon$ given by (1-3) and positive σ, λ, κ, ℓ fixed.

Let

$$d\mu_\varepsilon^0(x) := \tfrac{1}{2}\varepsilon^{-2/3}(1 + u_\varepsilon^0(x))\, dx, \tag{4-1}$$

where

$$u_\varepsilon^0(x) := \begin{cases} +1 & \text{if } u_\varepsilon(x) > 0, \\ -1 & \text{if } u_\varepsilon(x) \leq 0, \end{cases} \tag{4-2}$$

and let v_ε^0 satisfy

$$-\Delta v_\varepsilon^0 + \kappa^2 v_\varepsilon^0 = \mu_\varepsilon^0 \quad \text{in } \mathbb{T}_\ell. \tag{4-3}$$

With this notation, we are now in the position to state the main technical result of this paper.

Theorem 4.1 (equicoercivity and Γ-convergence of $\mathcal{E}_\varepsilon$). *For $\lambda > 0$ and $\ell > 0$, let $\mathcal{E}_\varepsilon$ be defined by* (2-1) *with W satisfying the assumptions of Section 2, let $\bar{u}_\varepsilon$ given by* (1-3), *and let σ and κ be given by* (2-8) *and* (2-9), *respectively. Then, as $\varepsilon \to 0$ we have*

$$\varepsilon^{-4/3}\mathcal{E}_\varepsilon \xrightarrow{\Gamma} E_0(\mu), \tag{4-4}$$

where $\mu \in \mathcal{M}^+(\mathbb{T}_\ell) \cap H^{-1}(\mathbb{T}_\ell)$. More precisely, we have:

(i) *Compactness and lower bound*: *Let $(u_\varepsilon) \in \mathcal{A}_\varepsilon$ be such that $\limsup_{\varepsilon\to 0} \|u_\varepsilon\|_{L^\infty(\mathbb{T}_\ell)} \leq 1$ and*

$$\limsup_{\varepsilon\to 0} \varepsilon^{-4/3}\mathcal{E}_\varepsilon(u_\varepsilon) < +\infty. \tag{4-5}$$

Then, up to extraction of a subsequence, we have

$$\mu_\varepsilon^0 \rightharpoonup \mu \quad \text{in } \mathcal{M}(\mathbb{T}_\ell), \qquad v_\varepsilon^0 \rightharpoonup v \quad \text{in } H^1(\mathbb{T}_\ell), \tag{4-6}$$

as $\varepsilon \to 0$, where $\mu \in \mathcal{M}^+(\mathbb{T}_\ell) \cap H^{-1}(\mathbb{T}_\ell)$ and $v \in H^1(\mathbb{T}_\ell)$ satisfy

$$-\Delta v + \kappa^2 v = \mu \quad \text{in } \mathbb{T}_\ell. \tag{4-7}$$

Moreover, we have

$$\liminf_{\varepsilon\to 0} \varepsilon^{-4/3}\mathcal{E}_\varepsilon(u_\varepsilon) \geq E_0(\mu). \tag{4-8}$$

(ii) *Upper bound*: *Given* $\mu \in \mathcal{M}^+(\mathbb{T}_\ell) \cap H^{-1}(\mathbb{T}_\ell)$ *and* $v \in H^1(\mathbb{T}_\ell)$ *solving* (4-7), *there exist* $(u_\varepsilon) \in \mathcal{A}_\varepsilon$ *such that for the corresponding* μ^0_ε, v^0_ε *as in* (4-1) *and* (4-3) *we have*

$$\mu^0_\varepsilon \rightharpoonup \mu \quad \text{in } \mathcal{M}(\mathbb{T}_\ell), \qquad v^0_\varepsilon \rightharpoonup v \quad \text{in } H^1(\mathbb{T}_\ell), \tag{4-9}$$

as $\varepsilon \to 0$, *and*

$$\limsup_{\varepsilon\to 0} \varepsilon^{-4/3}\mathcal{E}_\varepsilon(u_\varepsilon) \le E_0(\mu). \tag{4-10}$$

Proof. As in [Muratov 2010], the basic strategy is to relate the minimization problem for $\mathcal{E}_\varepsilon$ to that for E_ε and apply the results in Theorems 3.1 and 3.2. The proof relies on the fact, first observed in [Muratov 2010], that the energy $\mathcal{E}_\varepsilon$ is asymptotically equivalent to E_ε in the following sense: for any $\delta > 0$ and $u_\varepsilon \in \mathcal{A}_\varepsilon$ satisfying some mild technical conditions (see below) there is $\tilde{u}_\varepsilon \in \mathcal{A}$ such that

$$E_\varepsilon[\tilde{u}_\varepsilon] \le (1+\delta)\mathcal{E}_\varepsilon(u_\varepsilon) \tag{4-11}$$

for all $\varepsilon \ll 1$, and, conversely, for any $\tilde{u}_\varepsilon \in \mathcal{A}$, again, satisfying some mild technical conditions, there is $u_\varepsilon \in \mathcal{A}_\varepsilon$ such that

$$\mathcal{E}_\varepsilon[u_\varepsilon] \le (1+\delta)E_\varepsilon(\tilde{u}_\varepsilon) \tag{4-12}$$

for all $\varepsilon \ll 1$. The proof proceeds as in the two-dimensional case [Goldman et al. 2013, Theorem 1], with modifications appropriate to three space dimensions. We outline the key differences below.

For (4-11) to hold, we need to verify the assumptions of [Muratov 2010, Proposition 4.2], which are equivalent to checking that $\|u_\varepsilon\|_\infty \to 1$, $\mathcal{E}_\varepsilon(u_\varepsilon) \to 0$ and $\|v_\varepsilon\|_\infty \to 0$ as $\varepsilon \to 0$, where $v_\varepsilon(x) := \int_{\mathbb{T}_\ell} G_0(x-y)(u_\varepsilon(y) - \bar{u}_\varepsilon)\,dy$. The first and second conditions are clearly satisfied by the assumptions of the theorem. To check the third condition, we note that the nonlocal part of the energy may be written in terms of v_ε as

$$\frac{1}{2}\int_{\mathbb{T}_\ell}\int_{\mathbb{T}_\ell} (u_\varepsilon(x) - \bar{u}_\varepsilon)G_0(x-y)(u_\varepsilon(y)-\bar{u}_\varepsilon)\,dx\,dy = \frac{1}{2}\int_{\mathbb{T}_\ell} |\nabla v_\varepsilon|^2\,dx. \tag{4-13}$$

Since $\int_{\mathbb{T}_\ell} v_\varepsilon(x)\,dx = 0$ we have by Poincaré's inequality that the right-hand side of (4-13) is bounded below by a multiple of $\|v_\varepsilon\|_2^2$. In turn, the latter is bounded below by a multiple of $\|v_\varepsilon\|_\infty^5$, in view of the fact that, by elliptic regularity [Gilbarg and Trudinger 1983], we have $\|\nabla v_\varepsilon\|_\infty \le C$ for some $C > 0$ depending only on ℓ, for all $\varepsilon \ll 1$. Therefore, from $\mathcal{E}_\varepsilon(u_\varepsilon) \to 0$ we also obtain that $\|v_\varepsilon\|_\infty \to 0$ as $\varepsilon \to 0$. Thus, by (4-11) $\tilde{u}_\varepsilon$ satisfies the assumptions of Theorem 3.1, and so there exists $\mu \in \mathcal{M}(\mathbb{T}_\ell)$ such that, upon extraction of subsequences, $\mu_\varepsilon \rightharpoonup \mu$ in $\mathcal{M}(\mathbb{T}_\ell)$, where the measure μ_ε is defined by (3-1) with u_ε replaced by $\tilde{u}_\varepsilon$. For those subsequences, Theorem 3.2 holds true for μ_ε as well.

Now, from the construction of $\tilde{u}_\varepsilon$ in the proof of [Muratov 2010, Lemma 4.1] we know that $\tilde{u}_\varepsilon(x) = u^0_\varepsilon(x)$ for all $x \in \mathbb{T}_\ell$ such that $|u_\varepsilon(x)| > 1 - \delta^2$. Hence from the bound on $\mathcal{E}_\varepsilon(u_\varepsilon)$ and the assumptions on W we get that $\|\tilde{u}_\varepsilon - u^0_\varepsilon\|_1 \le C\varepsilon^{4/3}\delta^{-4}$ for some $C > 0$ and all $\varepsilon \ll 1$. This implies that $\mu^0_\varepsilon \rightharpoonup \mu$ in $\mathcal{M}(\mathbb{T}_\ell)$ as well $\varepsilon \to 0$. Together with the conclusions of Theorems 3.1 and 3.2, this gives the compactness and the lower bound statement of Theorem 4.1, in view of the arbitrariness of δ.

For (4-12) to hold, we need to verify the assumptions of [Muratov 2010, Proposition 4.3] on $\tilde{u}_\varepsilon \in \mathcal{A}$, namely, that the connected components of the support of $\{\tilde{u}_\varepsilon = +1\}$ are smooth and at least ε^α apart for

some $\alpha \in [0, 1)$ and have boundaries whose curvature is bounded by $\varepsilon^{-\alpha}$, and that $\|\tilde{v}_\varepsilon\|_\infty \to 0$ as $\varepsilon \to 0$, where $\tilde{v}_\varepsilon(x) := \int_{\mathbb{T}_\ell} G(x-y)(\tilde{u}_\varepsilon(y) - \bar{u}_\varepsilon)\, dy$. Clearly the first two assumptions hold true for the recovery sequence in the proof of Theorem 3.2 with any $\alpha \in \left(\frac{1}{3}, 1\right)$, provided that $\varepsilon \ll 1$. The third assumption is satisfied for all $\varepsilon \ll 1$, in view of the fact that the nonlocal part of the sharp interface energy can be written as

$$\frac{1}{2}\int_{\mathbb{T}_\ell}\int_{\mathbb{T}_\ell} (\tilde{u}_\varepsilon(x) - \bar{u}_\varepsilon) G_0(x-y)(\tilde{u}_\varepsilon(y) - \bar{u}_\varepsilon)\, dx\, dy = \frac{1}{2}\int_{\mathbb{T}_\ell} (|\nabla \tilde{v}_\varepsilon|^2 + \kappa^2 \tilde{v}_\varepsilon^2)\, dx, \tag{4-14}$$

and the desired estimate follows from $E_\varepsilon(\tilde{u}_\varepsilon) \to 0$ just like in the case of the diffuse interface energy. Thus, the proof of the upper bound is concluded by taking the functions u_ε appearing in (4-12), associated with the recovery sequence $(\tilde{u}_\varepsilon)$ from Theorem 3.2, once again, in view of arbitrariness of δ. □

Similarly to the sharp interface energy, as a direct consequence of Theorem 4.1 we have the following asymptotic characterization of the minimizers of the diffuse interface energy.

Corollary 4.2. *Under the assumptions of Theorem 4.1, let $(u_\varepsilon) \in \mathcal{A}_\varepsilon$ be minimizers of $\mathcal{E}_\varepsilon$. Let μ_ε^0 be defined in* (4-1) *and v_ε^0 be the solution of* (4-3)*. Then as $\varepsilon \to 0$, we have*

$$\mu_\varepsilon^0 \rightharpoonup \bar{\mu} \quad \text{in } \mathcal{M}(\mathbb{T}_\ell), \qquad v_\varepsilon^0 \rightharpoonup \bar{v} \quad \text{in } H^1(\mathbb{T}_\ell), \tag{4-15}$$

where $\bar{\mu}$ and $\bar{v}$ are as in Proposition 2.4.

We emphasize that the limit behavior of the minimal energy obtained in Corollary 4.2 *differs* from that of the unscreened sharp interface energy one would naively associate with $\mathcal{E}_\varepsilon$. In particular, the minimal energy exhibits a threshold behavior, contrary to that of the minimizers of the unscreened sharp interface energy studied in [Knüpfer et al. 2016; Emmert et al. 2018].

Proof of Theorems 1.1 and 1.2. The statement of Theorem 1.1 is simply the restatement of Theorem 4.1 that does not specify the precise values of the constants appearing there. In turn, the statement of Theorem 1.2 uses the explicit values of σ and κ given by (1-4) for (1-1), together with the bounds on f^* obtained in (2-31) and (2-32). □

Acknowledgements

Knüpfer was supported by the DFG via grant #392124319. Muratov was supported, in part, by the NSF via grants DMS-1313687 and DMS-1614948. Novaga was partially supported by INdAM-GNAMPA, and by the University of Pisa Project PRA 2017 “Problemi di ottimizzazione e di evoluzione in ambito variazionale”.

References

[Blanc and Lewin 2015] X. Blanc and M. Lewin, “The crystallization conjecture: a review”, *EMS Surv. Math. Sci.* **2**:2 (2015), 225–306. MR Zbl

[Brezis 2011] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Springer, 2011. MR Zbl

[Chen and Oshita 2005] X. Chen and Y. Oshita, “Periodicity and uniqueness of global minimizers of an energy functional containing a long-range interaction”, *SIAM J. Math. Anal.* **37**:4 (2005), 1299–1332. MR Zbl

[Choksi 2001] R. Choksi, "Scaling laws in microphase separation of diblock copolymers", *J. Nonlinear Sci.* **11**:3 (2001), 223–236. MR Zbl

[Choksi and Peletier 2011] R. Choksi and M. A. Peletier, "Small volume-fraction limit of the diblock copolymer problem, II: Diffuse-interface functional", *SIAM J. Math. Anal.* **43**:2 (2011), 739–763. MR Zbl

[Choksi et al. 2009] R. Choksi, M. A. Peletier, and J. F. Williams, "On the phase diagram for microphase separation of diblock copolymers: an approach via a nonlocal Cahn–Hilliard functional", *SIAM J. Appl. Math.* **69**:6 (2009), 1712–1738. MR Zbl

[Choksi et al. 2017] R. Choksi, C. B. Muratov, and I. Topaloglu, "An old problem resurfaces nonlocally: Gamow's liquid drops inspire today's research and applications", *Notices Amer. Math. Soc.* **64**:11 (2017), 1275–1283. MR Zbl

[Daneri and Runa 2018] S. Daneri and E. Runa, "Pattern formation for a local/nonlocal interaction functional arising in colloidal systems", preprint, 2018. arXiv

[Emmert et al. 2018] L. Emmert, R. L. Frank, and T. König, "Liquid drop model for nuclear matter in the dilute limit", preprint, 2018. arXiv

[Frank and Lieb 2015] R. L. Frank and E. H. Lieb, "A compactness lemma and its application to the existence of minimizers for the liquid drop model", *SIAM J. Math. Anal.* **47**:6 (2015), 4436–4450. MR Zbl

[Frank et al. 2016] R. L. Frank, R. Killip, and P. T. Nam, "Nonexistence of large nuclei in the liquid drop model", *Lett. Math. Phys.* **106**:8 (2016), 1033–1036. MR

[Gamow 1930] G. Gamow, "Mass defect curve and nuclear constitution", *Proc. Royal Soc. A* **126**:803 (1930), 632–644. 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

[Goldman et al. 2013] D. Goldman, C. B. Muratov, and S. Serfaty, "The Γ-limit of the two-dimensional Ohta–Kawasaki energy, I: Droplet density", *Arch. Ration. Mech. Anal.* **210**:2 (2013), 581–613. MR Zbl

[Goldman et al. 2014] D. Goldman, C. B. Muratov, and S. Serfaty, "The Γ-limit of the two-dimensional Ohta–Kawasaki energy: droplet arrangement via the renormalized energy", *Arch. Ration. Mech. Anal.* **212**:2 (2014), 445–501. MR Zbl

[Heisenberg 1934] M. W. Heisenberg, "Considérations théoriques générales sur la structure du noyau", pp. 289–335 in *Structure et propriétés des noyaux atomiques: rapports et discussions du Septiéme Conseil de Physique* (Bruxelles, 1933), Gauthier-Villars, Paris, 1934.

[Knüpfer et al. 2016] H. Knüpfer, C. B. Muratov, and M. Novaga, "Low density phases in a uniformly charged liquid", *Comm. Math. Phys.* **345**:1 (2016), 141–183. MR Zbl

[Landau and Lifshitz 1980] L. D. Landau and E. M. Lifshitz, *Course of theoretical physics, Vol. 5: Statistical physics*, 3rd ed., Pergamon Press, Oxford, 1980.

[Lattimer et al. 1985] J. M. Lattimer, C. J. Pethick, D. G. Ravenhall, and D. Q. Lamb, "Physical properties of hot, dense matter: the general case", *Nucl. Phys. A* **432**:3 (1985), 646–742.

[Le Bris and Lions 2005] C. Le Bris and P.-L. Lions, "From atoms to crystals: a mathematical journey", *Bull. Amer. Math. Soc. (N.S.)* **42**:3 (2005), 291–363. MR Zbl

[Lieb 1981] E. H. Lieb, "Thomas–Fermi and related theories of atoms and molecules", *Rev. Modern Phys.* **53**:4 (1981), 603–641. MR Zbl

[Lu and Otto 2014] J. Lu and F. Otto, "Nonexistence of a minimizer for Thomas–Fermi–Dirac–von Weizsäcker model", *Comm. Pure Appl. Math.* **67**:10 (2014), 1605–1617. MR Zbl

[Maruyama et al. 2005] T. Maruyama, T. Tatsumi, D. N. Voskresensky, T. Tanigawa, and S. Chiba, "Nuclear "pasta" structures and the charge screening effect", *Phys. Rev. C* **72**:1 (2005), art. id. 015802.

[Modica 1987] L. Modica, "The gradient theory of phase transitions and the minimal interface criterion", *Arch. Rational Mech. Anal.* **98**:2 (1987), 123–142. MR Zbl

[Morini and Sternberg 2014] M. Morini and P. Sternberg, "Cascade of minimizers for a nonlocal isoperimetric problem in thin domains", *SIAM J. Math. Anal.* **46**:3 (2014), 2033–2051. MR Zbl

[Müller 1993] S. Müller, "Singular perturbations as a selection criterion for periodic minimizing sequences", *Calc. Var. Partial Differential Equations* **1**:2 (1993), 169–204. MR Zbl

[Muratov 2002] C. B. Muratov, "Theory of domain patterns in systems with long-range interactions of Coulomb type", *Phys. Rev. E* (3) **66**:6 (2002), art. id. 066108. MR

[Muratov 2010] C. B. Muratov, "Droplet phases in non-local Ginzburg–Landau models with Coulomb repulsion in two dimensions", *Comm. Math. Phys.* **299**:1 (2010), 45–87. MR Zbl

[Ohta and Kawasaki 1986] T. Ohta and K. Kawasaki, "Equilibrium morphology of block copolymer melts", *Macromolecues* **19**:10 (1986), 2621–2632.

[Ren and Wei 2003] X. Ren and J. Wei, "On energy minimizers of the diblock copolymer problem", *Interfaces Free Bound.* **5**:2 (2003), 193–238. MR Zbl

[Rougerie and Serfaty 2016] N. Rougerie and S. Serfaty, "Higher-dimensional Coulomb gases and renormalized energy functionals", *Comm. Pure Appl. Math.* **69**:3 (2016), 519–605. MR Zbl

[Shirokoff et al. 2015] D. Shirokoff, R. Choksi, and J.-C. Nave, "Sufficient conditions for global minimality of metastable states in a class of non-convex functionals: a simple approach via quadratic lower bounds", *J. Nonlinear Sci.* **25**:3 (2015), 539–582. MR Zbl

[Spadaro 2009] E. N. Spadaro, "Uniform energy and density distribution: diblock copolymers' functional", *Interfaces Free Bound.* **11**:3 (2009), 447–474. MR Zbl

[von Weizsäcker 1935] C. F. von Weizsäcker, "Zur Theorie der Kernmassen", *Z. Phys. A* **96**:7-8 (1935), 431–458.

Received 6 Dec 2018. Revised 8 Jun 2019. Accepted 16 Aug 2019.

HANS KNÜPFER: hans.knuepfer@math.uni-heidelberg.de
Institut für Angewandte Mathematik, Universität Heidelberg, Heidelberg, Germany

CYRILL B. MURATOV: muratov@njit.edu
Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ, United States

MATTEO NOVAGA: matteo.novaga@unipi.it
Dipartimento di Matematica, Università di Pisa, Pisa, Italy

PURE and APPLIED ANALYSIS

Vol. 2, No. 1, 2020
dx.doi.org/10.2140/paa.2020.2.23

msp

MAXIMAL L^2-REGULARITY IN NONLINEAR GRADIENT SYSTEMS AND PERTURBATIONS OF SUBLINEAR GROWTH

WOLFGANG ARENDT AND DANIEL HAUER

The nonlinear semigroup generated by the subdifferential of a convex lower semicontinuous function φ has a smoothing effect, discovered by Haïm Brezis, which implies maximal regularity for the evolution equation. We use this and Schaefer's fixed point theorem to solve the evolution equation perturbed by a Nemytskii operator of sublinear growth. For this, we need that the sublevel sets of φ are not only closed, but even compact. We apply our results to the p-Laplacian and also to the Dirichlet-to-Neumann operator with respect to p-harmonic functions.

1. Introduction

Let H be a real Hilbert space, $\varphi : H \to (-\infty, +\infty]$ a proper, convex, lower semicontinuous function, $A = \partial\varphi$ the subdifferential of φ, and $D(\varphi) := \{u \in H \mid \varphi(u) < +\infty\}$ the effective domain of φ (see Section 2 for more details). Then A is a maximal monotone (in general, multivalued) operator on H for which the following remarkable well-posedness result holds.

Theorem 1.1 [Brezis 1971]. *Let $u_0 \in \overline{D(\varphi)}$ and $f \in L^2(0, T; H)$. Then, there exists a unique $u \in H^1_{\mathrm{loc}}((0, T]; H) \cap C([0, T]; H)$ such that*

$$\begin{cases} \dot{u}(t) + Au(t) \ni f(t) & \text{a.e. on } (0, T), \\ u(0) = u_0. \end{cases} \tag{1-1}$$

If $u \in D(\varphi)$ then $\dot{u} \in L^2(0, T; H)$.

Our aim in this article is to establish existence of solutions of a perturbed version of (1-1) and to show that these solutions have the same regularity result as in Theorem 1.1. We fix $T > 0$, and denote by $\mathcal{H}$ the space $L^2(0, T; H)$ and by $\|\cdot\|_{\mathcal{H}}$ the norm $\|\cdot\|_{L^2(0,T;H)}$. Then for $f \in \mathcal{H}$ and $u_0 \in H$, we call a function $u : [0, T] \to H$ a (*strong*) *solution* of (1-1) if $u \in H^1_{\mathrm{loc}}((0, T]; H) \cap C([0, T]; H)$, $u(0) = u_0$ and for a.e. $t \in (0, T)$ we have $u(t) \in D(A)$ and $f(t) - \dot{u}(t) \in Au(t)$.

Now, let $G : \mathcal{H} \to \mathcal{H}$ be a continuous mapping satisfying the *sublinear* growth condition

$$\|Gv(t)\|_H \le L\,\|v(t)\|_H + b(t) \quad \text{a.e. on } (0, T) \text{ and for all } v \in \mathcal{H}, \tag{1-2}$$

Hauer is very grateful for the warm hospitality received during his visits at the University of Ulm.

MSC2010: 35K92, 35K58, 47H20, 47H10.

Keywords: nonlinear semigroups, subdifferential, Schaefer's fixed point theorem, existence, smoothing effect, perturbation, compact sublevel sets.

for some L, $b \in L^2(0,T)$ satisfying $b(t) \geq 0$ for a.e. $t \in (0,T)$. Here we let $Gv(t) := (G(v))(t)$ to use less heavy notation. Then we study the evolution problem

$$\begin{cases} \dot{u}(t) + Au(t) \ni Gu(t) & \text{a.e. on } (0,T), \\ u(0) = u_0. \end{cases} \tag{1-3}$$

Note that $Gu \in \mathcal{H}$. Thus, the inclusion in (1-3) means that $Gu(t) - \dot{u}(t) \in Au(t)$ a.e. on $(0,T)$.

For proving existence of solutions to (1-3), we will use a compactness argument in form of Schaefer's fixed point theorem (see Theorem 2.1 in Section 2). Recall that lower semicontinuity of φ is equivalent to saying that the sublevel sets $E_c := \{u \in H \mid \varphi(u) \leq c\},\ c \in \mathbb{R}$, are closed. We will assume more, namely, compactness of the sublevel sets E_c. In fact, we need this assumption only for the shifted function φ_ω given by $\varphi_\omega(u) = \varphi(u) + \frac{1}{2}\omega\|u\|_H^2$, $u \in H$, which is important for applications. Then our main result says the following.

Theorem 1.2. *Let $\varphi : H \to (-\infty, +\infty]$ be a proper function such that for some $\omega \geq 0$, φ_ω is convex and has compact sublevel sets. Let $A = \partial\varphi$ and $G : \mathcal{H} \to \mathcal{H}$ be a continuous mapping satisfying* (1-2). *Then for every $u_0 \in \overline{D(\varphi)}$ and $f \in \mathcal{H}$, there exists $u \in H^1_{\mathrm{loc}}((0,T]; H) \cap C([0,T]; H)$ solving* (1-3). *In particular, if $u_0 \in D(\varphi)$, then $u \in H^1(0,T;H)$.*

We show in Example 3.3 that the solution is not unique in general. Further, we have the following regularity result for the composition $\varphi \circ u$ and a uniform estimate.

Remark 1.3. Suppose, the hypotheses of Theorem 1.2 hold. Then every solution u of (1-3) satisfies

$$\varphi \circ u \in W^{1,1}_{\mathrm{loc}}((0,T]) \cap L^1(0,T)$$

and

$$\|u(t)\|_H \leq (\|u_0\|_H^2 + \|b\|^2_{L^2(0,T)})^{1/2}\, e^{(2L+1+2\omega)/2\,t} \quad \text{for all } t \in [0,T]. \tag{1-4}$$

As application, we consider $H = L^2(\Omega)$ and G a Nemytskii operator. The operator A may be the p-Laplacian ($1 \leq p < +\infty$) with possibly lower-order terms and equipped with some boundary conditions (Dirichlet, Neumann, or Robin, see [Coulhon and Hauer 2016]) or a p-version of the Dirichlet-to-Neumann operator considered recently in [Hauer 2015] and via the abstract theory of j-elliptic functions (see [Arendt and ter Elst 2011; 2012; Chill et al. 2016]).

2. Preliminaries

In this section, we define the precise setting used throughout this paper and explain our main tools: Schaefer's fixed point theorem and Brezis' L^2-maximal regularity result for semiconvex functions.

We begin by recalling that a mapping $\mathcal{T}$ defined on a Banach space X is called *compact* if $\mathcal{T}$ maps bounded sets into relatively compact sets.

Theorem 2.1 (Schaefer's fixed point theorem [1955]). *Let X be a Banach space and $\mathcal{T} : X \to X$ be continuous and compact. Assume that the "Schaefer set"*

$$\mathcal{S} := \{u \in X \mid \text{there exists } \lambda \in [0,1] \text{ such that } u = \lambda\mathcal{T}u\}$$

is bounded in X. Then $\mathcal{T}$ has a fixed point.

This result is a special case of *Leray–Schauder* degree theory, but Schaefer [1955] gave a most elegant proof, which also is valid in locally convex spaces; see also [Arendt and Chill 2010; Evans 2010, §9.2.2].

Given a function $\varphi : H \to (-\infty, +\infty]$, we call the set $D(\varphi) := \{u \in H \mid \varphi(u) < +\infty\}$ the *effective domain* of φ, and φ is said to be *proper* if $D(\varphi)$ is nonempty. Further, we say that φ is *lower semicontinuous* if for every $c \in \mathbb{R}$ the sublevel set

$$E_c := \{u \in D(\varphi) \mid \varphi(u) \le c\}$$

is closed in H, and φ is *semiconvex* if there exists an $\omega \in \mathbb{R}$ such that the shifted function $\varphi_\omega : H \to (-\infty, +\infty]$ defined by

$$\varphi_\omega(u) := \varphi(u) + \tfrac{1}{2}\omega\|u\|_H^2, \quad u \in H,$$

is convex. Then, $\varphi_{\hat\omega}$ is convex for all $\hat w \ge \omega$, and φ_ω is lower semicontinuous if and only if φ is lower semicontinuous.

Given a function $\varphi : H \to (-\infty, +\infty]$, its *subdifferential* $A = \partial\varphi$ is defined by

$$\partial\varphi = \left\{(u, h) \in H \times H \;\middle|\; \liminf_{t\downarrow 0} \frac{\varphi(u + tv) - \varphi(u)}{t} \ge (h, v)_H \text{ for all } v \in D(\varphi)\right\},$$

which, if φ_ω is convex, reduces to

$$\partial\varphi = \{(u, h) \in H \times H \mid \varphi_\omega(u + v) - \varphi_\omega(u) \ge (h + \omega u, v)_H \text{ for all } v \in D(\varphi)\}.$$

It is standard to identify a (possibly multivalued) operator A on H with its graph and for every $u \in H$, one sets $Au := \{v \in H \mid (u, v) \in A\}$ and calls $D(A) := \{u \in H \mid Au \ne \varnothing\}$ the *domain of* A and $\mathrm{Rg}(A) := \bigcup_{u\in D(A)} Au$ the *range of* A.

Now, suppose $\varphi : H \to (-\infty, +\infty]$ is proper, lower semicontinuous, and semiconvex; more precisely, let us fix $\omega \in \mathbb{R}$ such that φ_ω is convex. Then the subdifferential $\partial\varphi_\omega$ of φ_ω is a simple perturbation of $\partial\varphi$, namely $\partial\varphi_\omega = \partial\varphi + \omega I$. For this reason, Brezis' well-posedness result (Theorem 1.1) remains true; see [Brezis 1973, Proposition 3.12]. In addition, it is not difficult to verify that each solution of (1-1) satisfies (2-2) and the estimates (2-3)–(2-6) below. For later use, we summarize these results in one theorem.

Theorem 2.2 (Brezis' L^2-maximal regularity for semiconvex φ). *Let $u_0 \in \overline{D(\varphi)}$ and $f \in \mathcal{H}$. Then, there exists a unique $u \in H^1_{\mathrm{loc}}((0, T]; H) \cap C([0, T]; H)$ satisfying*

$$\begin{cases} \dot u(t) + Au(t) \ni f(t) & \text{a.e. on } (0, T), \\ u(0) = u_0. \end{cases} \tag{2-1}$$

Moreover,

$$\varphi \circ u \in W^{1,1}_{\mathrm{loc}}((0, T]) \cap L^1(0, T), \tag{2-2}$$

$$\|u(t)\|_H \le \left(\|u_0\|_H^2 + \int_0^T \|f(s)\|_H^2 \,\mathrm{d}s\right)^{1/2} e^{(1+2\omega)/2t} \quad \text{for every } t \in (0, T], \tag{2-3}$$

$$\int_0^T \varphi(u(s))\,\mathrm{d}s \le \tfrac{1}{2}\|f\|_{\mathcal{H}}^2 + \tfrac{1}{2}(1 + \omega)\|u\|_{\mathcal{H}}^2 + \tfrac{1}{2}\|u_0\|_H^2, \tag{2-4}$$

$$t\varphi(u(t)) \le \int_0^T \varphi(u(s))\,\mathrm{d}s + \tfrac{1}{2}\|\sqrt{\cdot}\, f\|_{\mathcal{H}}^2 \quad \text{for every } t \in (0, T], \tag{2-5}$$

$$\|\sqrt{\cdot}\,\dot{u}\|_{\mathcal{H}}^2 \le 2\int_0^T \varphi(u(t))\,\mathrm{d}t + \|\sqrt{\cdot}\,f\|_{\mathcal{H}}^2. \tag{2-6}$$

Finally, if $u_0 \in D(\varphi)$, *then* $u \in H^1(0,T;H)$.

Remark 2.3 (maximal L^2-regularity). If $u_0 \in H$ such that $\varphi(u_0)$ is finite, then Theorem 1.1 (or Theorem 2.2) says that for every $f \in L^2(0,T;H)$, the unique solution u of (1-1) has its time derivative $\dot{u} \in L^2(0,T;H)$ and hence by the differential inclusion

$$\dot{u}(t) + Au(t) \ni f(t) \quad \text{a.e. on } (0,T), \tag{2-7}$$

and also $Au \in L^2(0,T;H)$. In other words, for $f \in L^2(0,T;H)$, $\dot{u}$ and $Au \in L^2(0,T;H)$ admit the maximal possible regularity. For this reason, we call this property *maximal L^2-regularity*, as it is customary for generators of holomorphic semigroups on Hilbert spaces; see [Arendt 2004] for a survey on this subject.

Given $\omega \in \mathbb{R}$, we say that the shifted function $\varphi_\omega : H \to (-\infty, +\infty]$ has *compact sublevel sets* if

$$E_{\omega,c} := \{u \in D(\varphi) \mid \varphi_\omega(u) \le c\} \quad \text{is compact in } H \text{ for every } c \in \mathbb{R}. \tag{2-8}$$

Remark 2.4. We emphasize that condition (2-8) does not imply that φ has compact sublevel sets. This becomes more clear if one considers as φ the function associated with the negative *Neumann p-Laplacian* $-\Delta_p^N$ on a bounded, open subset Ω of $\mathbb{R}^d$ with a Lipschitz boundary $\partial\Omega$. For $\max\{1, 2d/(d+2)\} < p < \infty$, $d \ge 1$, let $V = W^{1,p}(\Omega)$, $H = L^2(\Omega)$, and $\varphi : H \to (-\infty, +\infty]$ be given by

$$\varphi(u) := \begin{cases} \frac{1}{p}\int_\Omega |\nabla u|^p \,\mathrm{d}x & \text{if } u \in V, \\ +\infty & \text{if } u \in H \setminus V \end{cases} \tag{2-9}$$

for every $u \in H$. Then, for every $c > 0$, the sublevel set $E_{0,c}$ of φ contains the sequence $(u_n)_{n\ge 0}$ of constant functions $u_n \equiv n$, which does not admit any convergent subsequence in H. On the other hand, for every $\omega > 0$ and $c > 0$, the sublevel set $E_{\omega,c}$ is a bounded set in V and by Rellich–Kondrachov compactness, $V \hookrightarrow H$ by a compact embedding. Thus, for every $\omega > 0$ and $c > 0$, the sublevel set $E_{\omega,c}$ is compact in $L^2(\Omega)$.

3. An example and nonuniqueness

The main example of perturbations G allowed in Theorem 1.2 are Nemytskii operators on the space $\mathcal{H} = L^2(0,T;L^2(\Omega))$. Let $\Omega \subseteq \mathbb{R}^d$ be open and $g : (0,T)\times\Omega\times\mathbb{R} \to \mathbb{R}$ be a *Carathéodory function*, that is,

- $g(\,\cdot\,,\cdot,v) : (0,T)\times\Omega \to \mathbb{R}$ is measurable for all $v \in \mathbb{R}$,
- $g(t,x,\cdot\,) : \mathbb{R} \to \mathbb{R}$ is continuous for a.e. $(t,x) \in (0,T)\times\Omega$.

Assume furthermore g has *sublinear growth*; that is, there exist $L \ge 0$ and $b \in L^2(0,T;L^2(\Omega))$ such that

$$|g(t,x,v)| \le L\,|v| + b(t,x) \quad \text{for all } v \in \mathbb{R}, \text{ a.e. } (t,x) \in (0,T)\times\Omega. \tag{3-1}$$

Proposition 3.1. *Let* $\mathcal{H} = L^2(0,T;L^2(\Omega))$. *Then, the relation*

$$Gv(t,x) := g(t,x,v(t,x)) \quad \textit{for a.e. } (t,x) \in (0,T)\times\Omega \textit{ and every } v \in \mathcal{H}, \tag{3-2}$$

defines a continuous operator $G : \mathcal{H} \to \mathcal{H}$ *of sublinear growth* (1-2).

The proof of Proposition 3.1 is standard (see [Zeidler 1990, Proposition 26.7]) if one uses that $f_n \to f$ in $\mathcal{H}$ if and only if each subsequence of $(f_n)_{n\geq 1}$ has a dominated subsequence converging to f a.e. (which is well known from the completeness proof of L^2).

For illustrating the theory developed in this paper, we consider the following standard example: the *Dirichlet p-Laplacian* perturbed by a lower-order term.

Example 3.2. Let Ω be an open, bounded subset of $\mathbb{R}^d$, $d \geq 1$, $H = L^2(\Omega)$, and for $2d/(d+2) \leq p < \infty$, let $V = W_0^{1,p}(\Omega)$ be the closure of $C_c^1(\Omega)$ equipped with respect to the norm $\|u\|_V := \|\nabla u\|_{L^p(\Omega;\mathbb{R}^d)}$. Then, one has that V is continuously embedded into H (see [Brezis 2011, Theorem 9.16]); we write for this $V \hookrightarrow H$.

Further, let $f = \beta + f_1$ be the sum of a maximal monotone graph β of $\mathbb{R}$ satisfying $(0,0) \in \beta$ and a *Lipschitz–Carathéodory function* $f_1 : \Omega \times \mathbb{R} \to \mathbb{R}$ satisfying $f_1(x,0) = 0$; that is, for a.e. $x \in \Omega$, the function $f_1(x,\cdot)$ is Lipschitz continuous (with constant $\omega > 0$) uniformly for a.e. $x \in \Omega$, and $f_1(\cdot, u)$ is measurable on Ω for every $u \in \mathbb{R}$. Then, there is a proper, convex and lower semicontinuous function $j : \mathbb{R} \to (-\infty, +\infty]$ satisfying $j(0) = 0$ and $\partial j = \beta$ in $\mathbb{R}$; see [Barbu 2010, Example 1, p. 53]. We set

$$F_1(u) = \int_0^{u(x)} f_1(\cdot, s)\,\mathrm{d}s, \quad \varphi_2(u) := \begin{cases} \int_\Omega j(u(x))\,\mathrm{d}x & \text{if } j(u) \in L^1(\Omega), \\ +\infty & \text{if otherwise,} \end{cases}$$
$$F(u) = \varphi_2(u) + \int_\Omega F_1(u(x))\,dx \tag{3-3}$$

for every $u \in H$. Further, let $\varphi_1 : H \to (-\infty, +\infty]$ be given by

$$\varphi_1(u) = \begin{cases} \frac{1}{p}\int_\Omega |\nabla u|^p\,\mathrm{d}x + \int_\Omega F_1(u)\,\mathrm{d}x & \text{if } u \in V, \\ +\infty & \text{if } u \in H \setminus V \end{cases}$$

for every $u \in H$. Then the domain $D(\varphi_1)$ of φ_1 is V. The function φ_1 is lower semicontinuous on H and is proper, $\varphi_{1,\omega}$ is convex, and for every $u \in V$, φ_1 is Gâteaux-differentiable with

$$D_v\varphi_1(u) = \lim_{t\to 0+} \frac{\varphi_1(u+tv) - \varphi_1(u)}{t} = \int_\Omega |\nabla u|^{p-2}\nabla u \nabla v + f_1(x,u)\,v\,\mathrm{d}x$$

for every $v \in V$. Since V is dense in H, the subdifferential operator $\partial\varphi_1$ is a single-valued operator on H with domain

$$D(\partial\varphi_1) = \big\{u \in V \;\big|\; \text{there exists } h \in H \text{ such that } D_v\varphi_1(u) = \textstyle\int_\Omega hv\,\mathrm{d}x \text{ for all } v \in V\big\},$$
$$\partial\varphi_1(u) = h = -\Delta_p u + f_1(x,u) \quad \text{in } \mathcal{D}'(\Omega).$$

The operator $\partial\varphi_1$ is the negative *Dirichlet p-Laplacian* $-\Delta_p^D$ *on* Ω *with a Lipschitz continuous lower-order term* f_1. Next, we add the function φ_2 given by (3-3) to φ_1. For this, note that φ_2 is proper (since for $u_0 \equiv 0$, we have $\varphi_2(u_0) = 0$) with $\mathrm{int}(D(\varphi_2)) \neq \varnothing$, convex (since j is convex), and lower semicontinuous on H. Thus, the function $\varphi : H \to (-\infty, +\infty]$, given by

$$\varphi(u) = \varphi_1(u) + \varphi_2(u) \quad \text{for every } u \in H, \tag{3-4}$$

is convex, lower semicontinuous, and proper with domain $D(\varphi) = \{u \in V \mid j(u) \in L^1(\Omega)\}$, and the operator $A = \partial\varphi$ is given by

$$D(A) = \big\{u \in D(\varphi) \;\big|\; \text{there exists } h \in H \text{ such that } D_v\varphi(u) = \textstyle\int_\Omega hv\,\mathrm{d}x \text{ for all } v \in D(\varphi)\big\},$$
$$Au = h = -\Delta_p u + \beta(u) + f_1(x,u).$$

Here, we note that

$$\overline{D(A)} = \overline{D(\varphi)} = \{u \in H \mid j(u(x)) \in \overline{D(\beta)} \text{ for a.e. } x \in \Omega\}.$$

Due to Theorem 2.1, for every $u_0 \in \overline{D(\varphi)}$ and $f \in \mathcal{H}$, there is a unique solution $u \in H^1_{\mathrm{loc}}((0,T];H) \cap C([0,T];H)$ of the parabolic boundary-value problem

$$\begin{cases} \partial_t u(t) - \Delta_p u(t) + \beta(u(t)) + f_1(\,\cdot\,, u(t)) \ni f(t) & \text{on } (0,T)\times\Omega,\\ u(t) = 0 & \text{on } (0,T)\times\partial\Omega,\\ u(0) = u_0 & \text{on } \Omega. \end{cases}$$

Here, we write $\partial_t u(t)$ instead of $\dot{u}(t)$ since we rewrote the abstract Cauchy problem (1-1) as an explicit parabolic partial differential equation.

If $\max\{1, 2d/(d+2)\} < p < \infty$, then for the Lipschitz constant ω of f_1, we have φ_ω is convex, and for every $c > 0$ the sublevel set $E_{\omega,c}$ is compact in $L^2(\Omega)$. Furthermore, let $g : (0,T)\times\Omega\times\mathbb{R} \to \mathbb{R}$ be a Carathéodory function with sublinear growth and $u_0 \in \overline{D(\varphi)}$. Then, there is at least one solution $u \in H^1_{\mathrm{loc}}((0,T];H) \cap C([0,T];H)$ of the parabolic boundary-value problem

$$\begin{cases} \partial_t u(t,\cdot) - \Delta_p u(t,\cdot) + \beta(u(t,\cdot)) + f_1(\,\cdot\,, u(t,\cdot)) \ni g(t,\cdot,u(t,\cdot)) & \text{on } (0,T)\times\Omega,\\ u(t,\cdot) = 0 & \text{on } (0,T)\times\partial\Omega,\\ u(0,\cdot) = u_0 & \text{on } \Omega. \end{cases}$$

In general, the solutions u to the Cauchy problem (1-3) are not unique. We give an example.

Example 3.3 (nonuniqueness). Let $g(u) = \sqrt{|u|},\ u \in \mathbb{R}$, and Ω be an open and bounded subset of $\mathbb{R}^d$, $d \geq 1$, with a Lipschitz boundary $\partial\Omega$. Then, there are $L, b > 0$ such that $\hat{g}$ satisfies

$$|g(u)| \leq L\,|u| + b \quad \text{for every } u \in \mathbb{R}.$$

Thus, for $H = L^2(\Omega)$ and $\mathcal{H} = L^2((0,T)\times\Omega)$, the associated Nemytskii operator $G : \mathcal{H} \to \mathcal{H}$ defined by (3-2) satisfies the sublinear growth condition (1-2).

Further, for $\max\{1, 2d/(d+2)\} < p < +\infty$, let $\varphi : L^2(\Omega) \to (-\infty, +\infty]$ be the energy function (2-9) associated with the negative Neumann p-Laplacian $-\Delta_p^N$ on Ω. Then, by Theorem 1.2, for every $u_0 \in L^2(\Omega)$ and every $T > 0$, there is a solution $u \in H^1_{\mathrm{loc}}((0,T];L^2(\Omega)) \cap C([0,T];L^2(\Omega))$ of

$$\begin{cases} \partial_t u(t,\cdot) - \Delta_p^N u(t,\cdot) = \sqrt{|u|(t,\cdot)} & \text{in } (0,T)\times\Omega,\\ |\nabla u(t,\cdot)|^{p-2} D_\nu u(t,\cdot) = 0 & \text{on } (0,T)\times\partial\Omega,\\ u(0) = u_0 & \text{on } \Omega. \end{cases} \tag{3-5}$$

Here, $|\nabla u|^{p-2} D_\nu u$ denotes the (weak) conormal derivative of u on $\partial\Omega$; see [Coulhon and Hauer 2016].

Now, for the initial value $u_0 \equiv 0$ on Ω, the constant zero function $u \equiv 0$ is certainly a solution of (3-5). For constructing a nontrivial solution of (3-5) with initial value $u_0 \equiv 0$, let $w \in C^1[0, T]$ be a nontrivial solution of the classical ordinary differential equation

$$w' = \sqrt{|w|} \quad \text{on } (0, T),\ w(0) = 0, \tag{3-6}$$

For instance, one nontrivial solution is $w(t) = \frac{1}{4}t^2$. Since for every constant $c \in \mathbb{R}$ we have $-\Delta_p^N(c\mathbb{1}_\Omega) = 0$, the function $u(t) := w(t)$ is another nontrivial solution of (3-5) with initial value $u_0 \equiv 0$.

4. Proof of the main result

We now give the proof of Theorem 1.2. After possibly replacing φ by a translation, we may always assume without loss of generality that $0 \in D(\partial\varphi_\omega)$ and φ_ω attains a minimum at 0 with $\varphi_\omega(0) = 0$; for further details see [Barbu 2010, p. 159]. By the convexity of φ_ω, this implies $(0, 0) \in \omega I_H + A$, that is,

$$(h + \omega u, u)_H \geq 0 \quad \text{for all } (u, h) \in A. \tag{4-1}$$

For the proof of Theorem 1.2, we need some auxiliary results. The first concerns continuity and is standard; see [Bénilan et al. ca. 1990, (6.5), p. 87] or [Barbu 2010, (4.2), p. 128].

Lemma 4.1. *Let $f_1, f_2 \in \mathcal{H}$ and $u_1, u_2 \in H^1(0, T; H)$ such that*

$$\dot{u}_1 + Au_1 \ni f_1 \quad \text{on } (0, T),$$
$$\dot{u}_2 + Au_2 \ni f_2 \quad \text{on } (0, T).$$

Then,

$$\|u_1(t) - u_2(t)\|_H \leq e^{\omega t}\|u_1(0) - u_2(0)\|_H + \int_0^t e^{\omega(t-s)}\|f_1(s) - f_2(s)\|_H \, \mathrm{d}s \tag{4-2}$$

for every $t \in [0, T]$.

Next, we establish the compactness of the *solution operator* P associated with evolution problem (1-1). We recall that the closure $\overline{D(\varphi)}$ in H of the effective domain of a semiconvex function φ is a convex subset of H.

Lemma 4.2. *Let $P : \overline{D(\varphi)} \times \mathcal{H} \to \mathcal{H}$ be the mapping defined by*

$$P(u_0, f) = \text{solution } u \text{ of (1-1)} \quad \text{for every } u_0 \in \overline{D(\varphi)} \text{ and } f \in \mathcal{H}.$$

Then, P is continuous and compact.

Proof. (a) By Lemma 4.1, the map P is continuous from $\overline{D(\varphi)} \times \mathcal{H}$ to $\mathcal{H}$.

(b) We show that P is compact. Let $(u_n^{(0)})_{n\geq 1} \subseteq \overline{D(\varphi)}$ and $(f_n)_{n\geq 1} \subseteq \mathcal{H}$ such that $\|u_n^{(0)}\|_H + \|f_n\|_{\mathcal{H}} \leq c$ and $u_n = P(u_n^{(0)}, f_n)$ for every $n \geq 1$. Then, by (2-3), (2-4) and by (2-6), for every $\delta \in (0, T)$, there is a $c_\delta > 0$ such that

$$\sup_{n\geq 1}\|u_n\|_{H^1(\delta,T;H)} \leq c_\delta.$$

Since $H^1(\delta, T; H) \hookrightarrow C^{1/2}([\delta, T]; H)$, the sequence $(u_n)_{n\geq 1}$ is equicontinuous on $[\delta, T]$ for each $0 < \delta < T$. Choose a countable dense subset $D := \{t_m \mid m \in \mathbb{N}\}$ of $(0, T]$. Let $m \geq 1$. Then by (2-5),

$$\sup_{n\geq 1} \varphi(u_n(t_m)) \quad \text{is finite}$$

and since by (2-3), $(u_n(t_m))_{n\geq 1}$ is bounded in H, there is a $c' > 0$ such that $(u_n(t_m))_{n\geq 1}$ is in the sublevel set $E_{\omega,c'}$. Thus and by the assumption (2-8), $(u_n(t_m))_{n\geq 1}$ has a convergent subsequence in H. By Cantor's diagonalization argument, we find a subsequence $(u_{n_k})_{k\geq 1}$ of $(u_n)_{n\geq 1}$ such that

$$\lim_{k\to+\infty} u_{n_k}(t_m) \quad \text{exists in } H \text{ for all } m \in \mathbb{N}.$$

It follows from the equicontinuity of $(u_{n_k})_{k\geq 1}$ that u_{n_k} converges in $C([\delta, T]; H)$ for all $\delta \in (0, T]$. In particular, $(u_{n_k}(t))_{k\geq 1}$ converges in H for every $t \in (0, T)$ and by (2-3), $(u_{n_k})_{k\geq 1}$ is uniformly bounded in $L^\infty(0, T; H)$. Thus, it follows from Lebesgue's dominated convergence theorem that $u_{n_k} = P(u_{n_k}^{(0)}, f_{n_k})$ converges in $\mathcal{H}$. □

Remark 4.3. In the previous proof, we have actually shown that P is compact from $\overline{D(\varphi)} \times \mathcal{H}$ into the Fréchet space $C((0, T]; H)$.

With these preliminaries, we can now give the proof of our main result. Here, we were inspired by the linear case [Arendt and Chill 2010].

Proof of Theorem 1.2. First, let $u_0 \in \overline{D(\varphi)}$.

For $v \in \mathcal{H}$, one has $Gv \in \mathcal{H}$ and so, by Brezis' maximal L^2-regularity result (Theorem 2.2), there is a unique solution $u \in H^1_{\mathrm{loc}}((0, T]; H) \cap C([0, T]; H)$ of the evolution problem

$$\begin{cases} \dot{u}(t) + Au(t) \ni Gv(t) & \text{a.e. on } (0, T), \\ u(0) = u_0. \end{cases}$$

Let $\mathcal{T}v := P(u_0, Gv)$. Then by the continuity and linear growth of G and since $P(u_0, \cdot) : \mathcal{H} \to \mathcal{H}$ is continuous and compact (Lemma 4.2), the mapping $\mathcal{T} : \mathcal{H} \to \mathcal{H}$ is continuous and compact.

(a) We consider the Schaefer set

$$\mathcal{S} := \{u \in \mathcal{H} \mid \text{there exists } \lambda \in [0, 1] \text{ such that } u = \lambda \mathcal{T} u\}.$$

We show that $\mathcal{S}$ is bounded in $\mathcal{H}$. Let $u \in \mathcal{S}$. We may assume that $\lambda \in (0, 1]$; otherwise, $u \equiv 0$. Then, $u \in H^1_{\mathrm{loc}}((0, T]; H) \cap C([0, T]; H)$ and

$$\begin{cases} \dot{u}/\lambda + A(u/\lambda) \ni Gu & \text{on } (0, T), \\ u(0) = u_0. \end{cases}$$

It follows from (4-1) that

$$\left(-\frac{\dot{u}}{\lambda}(t) + Gu(t) + \omega\frac{u}{\lambda}(t), \frac{u}{\lambda}\right)_H \geq 0 \quad \text{for a.e. } t \in (0, T).$$

Thus and by (1-2),

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\tfrac{1}{2}\|u(t)\|_H^2 &= (\dot{u}(t), u(t))_H = (\dot{u}(t) - \lambda Gu(t) - \omega\lambda u(t), u(t))_H + (\lambda Gu(t) + \omega\lambda u(t), u(t))_H \\ &\leq (\lambda Gu(t) + \omega\lambda u(t), u(t))_H \\ &\leq \lambda(\|Gu(t)\|_H \|u(t)\|_H + \omega \|u(t)\|_H^2) \\ &\leq \lambda(L \|u(t)\|_H^2 + b(t) \|u(t)\|_H + \omega \|u(t)\|_H^2) \\ &\leq (2L + 1 + 2\omega)\, \tfrac{1}{2}\|u(t)\|_H^2 + \tfrac{1}{2}b^2(t) \end{aligned}$$

for a.e. $t \in (0, T)$. It follows from Gronwall's lemma that (1-4) holds for every $t \in [0, T]$. Thus, $\mathcal{S}$ is bounded in $\mathcal{H}$. Now, Schaefer's fixed point theorem implies that there exists $u \in \mathcal{H}$ such that $u = \mathcal{T}u$; that is, $u \in H^1_{\mathrm{loc}}((0, T]; H) \cap C([0, T]; H)$ is a solution of the evolution problem (1-3).

(b) Let $u_0 \in D(\varphi)$. Then, by the first part of this proof, there is a solution $u \in H^1_{\mathrm{loc}}((0, T]; H) \cap C([0, T]; H)$ of the evolution problem (1-3). However, by Brezis' maximal regularity result applied to $f = Gu \in \mathcal{H}$, it follows that $u \in H^1(0, T; H)$. □

5. Application to j-elliptic functions

In Examples 3.2 and 3.3, V is a Banach space injected in H. Recently, in Chill, Hauer and Kennedy [Chill et al. 2016] extended results of [Arendt and ter Elst 2011; 2012] to a nonlinear framework of *j-elliptic functions* $\varphi : V \to (-\infty, +\infty]$ generating a quasimaximal monotone operator $\partial_j\varphi$ on H, where $j : V \to H$ is just a linear operator which is not necessarily injective. This enabled the authors of [Chill et al. 2016] to show that several coupled parabolic-elliptic systems can be realized as a gradient system in a Hilbert space H and to extend the linear variational theory of the Dirichlet-to-Neumann operator to the nonlinear p-Laplace operator; see also [Belhachmi and Chill 2015; 2018] for further applications and extensions of this theory.

The aim of this section is to illustrate that Theorem 1.2 of Section 3 can also be applied to the framework of j-elliptic functions.

Let us briefly recall some basic notions and facts about j-elliptic functions from [Chill et al. 2016]. Let V be a real locally convex topological vector space and $j : V \to H$ be a linear operator which is merely weak-to-weak continuous (and, in general, not injective). Given a function $\varphi : V \to (-\infty, +\infty]$, the *$j$-subdifferential* is the operator

$$\partial_j\varphi := \left\{ (u, f) \in H \times H \;\middle|\; \begin{array}{c} \text{there exists } \hat{u} \in D(\varphi) \text{ such that } j(\hat{u}) = u \text{ and for every } \hat{v} \in V \\ \liminf_{t \searrow 0} (\varphi(\hat{u} + t\hat{v}) - \varphi(\hat{u}))/t \geq (f, j(\hat{v}))_H \end{array} \right\}.$$

The function φ is called *j-semiconvex* if there exists $\omega \in \mathbb{R}$ such that the "shifted" function $\varphi_\omega : V \to (-\infty, +\infty]$ given by

$$\varphi(\hat{u}) + \tfrac{1}{2}\omega \, \|j(\hat{u})\|_H^2 \quad \text{for every } \hat{u} \in V$$

is convex. If $V = H$ and $j = I_H$, then j-semiconvex functions φ are the *semiconvex* ones (see Section 1). The function φ is called *j-elliptic* if there exists $\omega \geq 0$ such that φ_ω is convex and for every $c \in \mathbb{R}$, the sublevel sets $\{\hat{u} \in V \mid \varphi_\omega(u) \leq c\}$ are relatively weakly compact. Finally, we say that the function φ is *lower semicontinuous* if the sublevel sets $\{\varphi \leq c\}$ are closed in the topology of V for every $c \in \mathbb{R}$. It was highlighted in [Chill et al. 2016, Lemma 2.2] that:

(a) If φ is j-semiconvex, then there is an $\omega \in \mathbb{R}$ such that

$$\partial_j\varphi = \left\{ (u, f) \in H \times H \;\middle|\; \begin{array}{c} \text{there exists } \hat{u} \in D(\varphi) \text{ such that } j(\hat{u}) = u \text{ and for every } \hat{v} \in V \\ \varphi_\omega(\hat{u} + \hat{v}) - \varphi_\omega(\hat{u}) \geq (f + \omega j(\hat{u}), j(\hat{v}))_H \end{array} \right\}.$$

(b) If φ is Gâteaux differentiable with directional derivative $D_{\hat{v}}\varphi$, ($\hat{v} \in V$), then

$$\partial_j\varphi = \left\{ (u, f) \in H \times H \;\middle|\; \begin{array}{c} \text{there exists } \hat{u} \in D(\varphi) \text{ such that } j(\hat{u}) = u \text{ and for every } \hat{v} \in V \\ D_{\hat{v}}\varphi(\hat{u}) = (f, j(\hat{v}))_H \end{array} \right\}.$$

The main result in [Chill et al. 2016] is that the j-subdifferential $\partial_j\varphi$ of a j-elliptic function φ is already a classical subdifferential. More precisely, the following holds.

Theorem 5.1 [Chill et al. 2016, Corollary 2.7]. *Let $\varphi : V \to (-\infty, +\infty]$ be proper, lower semicontinuous, and j-elliptic. Then there is a proper, lower semicontinuous, semiconvex function $\varphi^H : H \to (-\infty, +\infty]$ such that $\partial_j\varphi = \partial\varphi^H$. The function φ^H is unique up to an additive constant.*

Thus the operator $A = \partial_j\varphi$ has the properties of maximal regularity we used before. The following result gives a description of φ^H in the convex case and will be important for our intentions in this paper.

Theorem 5.2 [Chill et al. 2016, Theorem 2.9]. *Assume that $\varphi : V \to (-\infty, +\infty]$ is convex, proper, lower semicontinuous and j-elliptic, and let $\varphi^H : H \to (-\infty, +\infty]$ be the function from Theorem 5.1. Then, there is a constant $c \in \mathbb{R}$ such that*

$$\varphi^H(u) = c + \inf_{\hat{u} \in j^{-1}(\{u\})} \varphi(\hat{u}) \quad \textit{for every } u \in H,$$

with effective domain $D(\varphi^H) = j(D(\varphi))$.

For our perturbation result, we need the compactness of the sublevel sets of φ^H. With the help of Theorem 5.2 we can establish a criterion in terms of the given φ for this property.

Lemma 5.3. *Let $\varphi : V \to (-\infty, +\infty]$ be proper, lower semicontinuous j-semiconvex, and j-elliptic. Assume that*

$$j : V \to H \textit{ maps weakly relatively compact sets of } V \textit{ into relatively norm-compact sets of } H. \tag{5-1}$$

Then there is an $\omega \geq 0$ such that for every $c \in \mathbb{R}$ the sublevel set

$$E_{\omega,c} = \{u \in H \mid \varphi_\omega^H(u) \leq c\} \quad \textit{is compact in } H.$$

Remark 5.4. If V is a normed space, then by the Eberlein–Šmulian theorem hypothesis (5-1) is equivalent to j *maps weakly convergent sequences in V to norm convergent sequences in H*. This in turn is equivalent to j being compact if V is reflexive.

Proof of Lemma 5.3. By hypothesis, there is an $\omega \geq 0$ such that φ_ω is convex, lower semicontinuous, and for every $c \in \mathbb{R}$, the sublevel sets $\{\hat{u} \in V \mid \varphi_\omega(u) \leq c\}$ are weakly relatively compact and closed. By Theorem 5.1, there is a lower semicontinuous, proper function $\varphi^H : H \to (-\infty, +\infty]$ such that φ_ω^H is convex and $\partial\varphi_\omega^H = \partial_j\varphi_\omega$. Applying Theorem 5.2 to φ_ω and φ_ω^H, we have

$$\varphi_\omega^H(u) = d + \inf_{\hat{u} \in j^{-1}(\{u\})} \varphi_\omega(\hat{u}) \quad \text{for every } u \in H \tag{5-2}$$

and some constant $d \in \mathbb{R}$. For $c \in \mathbb{R}$, let $(u_n)_{n\geq 1}$ be an arbitrary sequence in $E_{\omega,c}$. By (5-2), for every $n \in \mathbb{N}$, there is a $\hat{u}_n \in j^{-1}(\{u_n\})$ such that

$$d + \varphi_\omega(\hat{u}_n) \leq c + 1.$$

By hypothesis, all sublevel sets of φ_ω are weakly relatively compact in V. Thus, by our hypothesis, the image under j is relatively compact in H. Consequently, there are a subsequence $(u_{n_l})_{l\geq 1}$ of $(u_n)_{n\geq 1}$ and a $u \in H$ such that $u_{n_l} = j(\hat{u}_{n_l}) \to u$ in H as $l \to +\infty$. Since $\varphi_\omega^H(u_{n_l}) \leq c$ and since φ^H is lower semicontinuous, it follows that $\varphi^H(u) \leq c$. This shows that $E_{\omega,c}$ is compact. □

Now, applying Lemma 5.3 to Theorem 1.2, we can state the following existence theorem.

Theorem 5.5. *Let $\varphi : V \to (-\infty, +\infty]$ be proper, lower semicontinuous j-semiconvex, and j-elliptic. Assume that the mapping j satisfies* (5-1) *and let $G : \mathcal{H} \to \mathcal{H}$ be a continuous mapping of sublinear growth* (1-2)*. Then, for $A = \partial_j\varphi$ the nonlinear evolution problem* (1-3) *admits for every $u_0 \in \overline{j(D(\varphi))}$ and $f \in \mathcal{H}$ at least one solution $u \in H^1_{\mathrm{loc}}((0,T]; H) \cap C([0,T]; H)$. In particular, $\varphi \circ u$ belongs to $W^{1,1}_{\mathrm{loc}}((0,T]) \cap L^1(0,T)$ and inequality* (1-4) *holds. If $u_0 \in j(D(\varphi))$, then problem* (1-3) *has a solution $u \in H^1(0,T; H)$.*

We complete this section by considering the following evolution problem involving the *Dirichlet-to-Neumann operator* associated with the p-Laplacian; see [Hauer 2015; Chill et al. 2016].

Example 5.6. Let Ω be a bounded domain with a Lipschitz continuous boundary $\partial\Omega$. Then, for $2d/(d+1) < p < +\infty$, the trace operator $\mathrm{Tr} : W^{1,p}(\Omega) \to L^2(\partial\Omega)$ is a completely continuous operator (see [Nečas 1967, Théorème 6.2] for the case $p < d$; the other cases $p = d$ and $p > d$ can be deduced from Conséquences 6.2 and 6.3 of the same work). Now, we take

$$V = W^{1,p}(\Omega), \quad H = L^2(\partial\Omega), \quad \text{and} \quad j = \mathrm{Tr}\,.$$

Then, j is a linear bounded mapping satisfying hypothesis (5-1). In fact, j is a prototype of a noninjective mapping. Furthermore, let $\varphi : V \to \mathbb{R}$ be the function given by

$$\varphi(\hat{u}) = \frac{1}{p}\int_\Omega |\nabla\hat{u}|^p \,\mathrm{d}x \quad \text{for every } \hat{u} \in V.$$

Then, φ is continuously differentiable on V and convex. Thus, the Tr-subdifferential operator $\partial_{\mathrm{Tr}}\varphi$ is given by

$$\partial_{\mathrm{Tr}}\varphi = \left\{ (u, f) \in H \times H \;\middle|\; \begin{array}{c} \text{there exists } \hat{u} \in V \text{ such that } \mathrm{Tr}(\hat{u}) = u \text{ and for every } \hat{v} \in V \\ \int_\Omega |\nabla\hat{u}|^{p-2}\nabla\hat{u}\nabla\hat{v}\,\mathrm{d}x = (f, j(\hat{v}))_H \end{array} \right\}.$$

Moreover, by [Hauer 2015, inequality (20)], for any $\omega > 0$, the shifted function φ_ω has bounded sublevel sets in V. Since V is reflexive, every sublevel set of φ_ω is weakly compact in V. In addition, by Lemma 2.1 of the same work, $j(D(\varphi))$ is dense in H.

Now, let $g : (0,T)\times\Omega\times\mathbb{R} \to \mathbb{R}$ be a Carathéodory function with sublinear growth. Then by Theorem 5.5, for every $u_0 \in L^2(\partial\Omega)$, there is at least one solution $u \in H^1_{\mathrm{loc}}((0,T]; L^2(\partial\Omega)) \cap C([0,T]; L^2(\partial\Omega))$ of the elliptic-parabolic boundary-value problem

$$\begin{cases} -\Delta_p\hat{u}(t,\cdot) = 0 & \text{on } (0,T)\times\Omega, \\ \partial_t u(t,\cdot) + |\nabla u(t,\cdot)|^{p-2}\frac{\partial}{\partial\nu}u(t,\cdot) = g(t,\cdot,u(t,\cdot)) & \text{on } (0,T)\times\partial\Omega, \\ u(t,\cdot) = \hat{u}(t,\cdot) & \text{on } (0,T)\times\partial\Omega, \\ u(0,\cdot) = u_0 & \text{on } \partial\Omega. \end{cases}$$

References

[Arendt 2004] W. Arendt, "Semigroups and evolution equations: functional calculus, regularity and kernel estimates", pp. 1–85 in *Evolutionary equations, Vol. I*, edited by C. M. Dafermos and E. Feireisl, North-Holland, Amsterdam, 2004. MR Zbl

[Arendt and Chill 2010] W. Arendt and R. Chill, "Global existence for quasilinear diffusion equations in isotropic nondivergence form", *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (5) **9**:3 (2010), 523–539. MR Zbl

[Arendt and ter Elst 2011] W. Arendt and A. F. M. ter Elst, "The Dirichlet-to-Neumann operator on rough domains", *J. Differential Equations* **251**:8 (2011), 2100–2124. MR Zbl

[Arendt and ter Elst 2012] W. Arendt and A. F. M. ter Elst, "Sectorial forms and degenerate differential operators", *J. Operator Theory* **67**:1 (2012), 33–72. MR Zbl

[Barbu 2010] V. Barbu, *Nonlinear differential equations of monotone types in Banach spaces*, Springer, 2010. MR Zbl

[Belhachmi and Chill 2015] Z. Belhachmi and R. Chill, "Application of the j-subgradient in a problem of electropermeabilization", *J. Elliptic Parabol. Equ.* **1** (2015), 13–29. MR Zbl

[Belhachmi and Chill 2018] Z. Belhachmi and R. Chill, "The bidomain problem as a gradient system", 2018. arXiv

[Bénilan et al. ca. 1990] P. Bénilan, M. G. Crandall, and A. Pazy, "Evolution problems governed by accretive operators", unpublished manuscript, ca. 1990.

[Brezis 1971] H. Brézis, "Propriétés régularisantes de certains semi-groupes non linéaires", *Israel J. Math.* **9** (1971), 513–534. MR Zbl

[Brezis 1973] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland Mathematics Studies **5**, North-Holland, Amsterdam, 1973. MR Zbl

[Brezis 2011] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Springer, 2011. MR Zbl

[Chill et al. 2016] R. Chill, D. Hauer, and J. Kennedy, "Nonlinear semigroups generated by j-elliptic functionals", *J. Math. Pures Appl.* (9) **105**:3 (2016), 415–450. MR Zbl

[Coulhon and Hauer 2016] T. Coulhon and D. Hauer, "Regularisation effects of nonlinear semigroups", preprint, 2016. To appear in *BCAM Springer Briefs*. arXiv

[Evans 2010] L. C. Evans, *Partial differential equations*, 2nd ed., Graduate Studies in Mathematics **19**, American Mathematical Society, Providence, RI, 2010. MR Zbl

[Hauer 2015] D. Hauer, "The p-Dirichlet-to-Neumann operator with applications to elliptic and parabolic problems", *J. Differential Equations* **259**:8 (2015), 3615–3655. MR Zbl

[Nečas 1967] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson et Cie, Paris, 1967. MR Zbl

[Schaefer 1955] H. Schaefer, "Über die Methode der a priori-Schranken", *Math. Ann.* **129** (1955), 415–416. MR Zbl

[Zeidler 1990] E. Zeidler, *Nonlinear functional analysis and its applications, II/B: Nonlinear monotone operators*, Springer, 1990. MR Zbl

Received 14 Feb 2019. Revised 3 Aug 2019. Accepted 9 Sep 2019.

WOLFGANG ARENDT: wolfgang.arendt@uni-ulm.de
Institute of Applied Analysis, University of Ulm, Ulm, Germany

DANIEL HAUER: daniel.hauer@sydney.edu.au
School of Mathematics and Statistics, The University of Sydney, Sydney, Australia

PURE and APPLIED
ANALYSIS

Vol. 2, No. 1, 2020
dx.doi.org/10.2140/paa.2020.2.35

msp

THE LOCAL DENSITY APPROXIMATION IN DENSITY FUNCTIONAL THEORY

MATHIEU LEWIN, ELLIOTT H. LIEB AND ROBERT SEIRINGER

We give the first mathematically rigorous justification of the local density approximation in density functional theory. We provide a quantitative estimate on the difference between the grand-canonical Levy–Lieb energy of a given density (the lowest possible energy of all quantum states having this density) and the integral over the uniform electron gas energy of this density. The error involves gradient terms and justifies the use of the local density approximation in the situation where the density is very flat on sufficiently large regions in space.

1. Introduction

Density functional theory (DFT) [Dreizler and Gross 1990; Parr and Weitao 1994; Engel and Dreizler 2011; Burke and Wagner 2013; Pribram-Jones et al. 2015] is the most efficient approximation of the many-body Schrödinger equation for electrons. It is used in several areas of physics and chemistry and its success in predicting the electronic properties of atoms, molecules and materials is unprecedented. Among the many functionals that have been developed over the years [Mardirossian and Head-Gordon 2017], the *local density approximation* (LDA) is the standard and simplest scheme [Hohenberg and Kohn 1964; Kohn and Sham 1965; Dreizler and Gross 1990; Parr and Weitao 1994; Perdew and Kurth 2003]. It is not as accurate as its successors involving gradient corrections, but it is considered as "*the mother of all approximations*" [Perdew and Schmidt 2001] and it is still one of the methods of choice in solid state physics.

In the orbital-free formulation of density functional theory [Levy 1979; Lieb 1983], the local density approximation consists in replacing the full ground state energy by a local functional as follows:

$$F_{\mathrm{LL}}(\rho) \approx \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy + \int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho(x))\,dx. \tag{1}$$

MSC2010: 35Q40, 81V55, 82B10.
Keywords: Schrödinger operators, statistical mechanics, density functional theory, uniform electron gas.

Here ρ is the given one-particle density of the system and $F_{\mathrm{LL}}(\rho)$ is the *Levy–Lieb functional* [Levy 1979; Lieb 1983], the main object of interest in DFT. This is the lowest possible Schrödinger energy of all quantum states having the prescribed density ρ. The first term on the right side is called the *direct* or *Hartree term*. It is the classical electrostatic interaction energy of the density ρ and it is the only nonlocal term in the LDA. The second term is the energy of the *uniform electron gas* (*UEG*) [Dreizler and Gross 1990; Parr and Weitao 1994; Giuliani and Vignale 2005; Lewin et al. 2018], containing all of the kinetic energy and the exchange-correlation energy in our convention. That is, $e_{\mathrm{UEG}}(\rho_0)$ is the ground state energy per unit volume of the infinite electron gas with the prescribed constant density ρ_0 over the whole space (from which the direct term has been dropped). The rationale for the approximation (1) is to assume that the density is almost constant locally (in little boxes of volume dx), and to replace the local energy per unit volume by that of the infinite gas at that density $\rho(x)$.[1]

Our goal in this paper is to justify the approximation (1) in the appropriate regime where ρ is flat in sufficiently large regions of $\mathbb{R}^3$. We will prove the following quantitative estimate:

$$\left| F_{\mathrm{LL}}(\rho) - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy - \int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho(x))\,dx \right| \leqslant \varepsilon \int_{\mathbb{R}^3} (\rho(x)+\rho(x)^2)\,dx + \frac{C(1+\varepsilon)}{\varepsilon}\int_{\mathbb{R}^3} |\nabla\sqrt{\rho}(x)|^2\,dx + \frac{C}{\varepsilon^{4p-1}}\int_{\mathbb{R}^3} |\nabla\rho^\theta(x)|^p\,dx \quad (2)$$

for all $\varepsilon > 0$, where $F_{\mathrm{LL}}(\rho)$ is the grand-canonical version of the Levy–Lieb functional. The parameters $p > 3$ and $0 < \theta < 1$ should satisfy some conditions which will be explained below. For instance, $p = 4$ and $\theta = \frac{1}{2}$ is allowed. After optimizing over ε, this justifies the LDA when the two gradient terms are much smaller than the local term:

$$\int_{\mathbb{R}^3} |\nabla\sqrt{\rho}(x)|^2\,dx \ll \int_{\mathbb{R}^3} (\rho(x)+\rho(x)^2)\,dx,$$
$$\int_{\mathbb{R}^3} |\nabla\rho^\theta(x)|^p\,dx \ll \int_{\mathbb{R}^3} (\rho(x)+\rho(x)^2)\,dx.$$

For instance for a rescaled density in the form

$$\rho_N(x) := \rho(N^{-\frac{1}{3}}x),$$

with $\int_{\mathbb{R}^3}\rho = 1$, we obtain after taking $\varepsilon = N^{-1/12}$

$$\left| F_{\mathrm{LL}}(\rho_N) - \frac{N^{\frac{5}{3}}}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy - N\int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho(x))\,dx \right| \leqslant C N^{\frac{11}{12}}.$$

The bound (2) is, to our knowledge, the first estimate of this kind on the fundamental functional F_{LL}. Although it should be possible to extract a definite value of the constant C from our proof, it is probably very large and we have not tried to do it. The factor $1/\varepsilon^{4p-1}$ is also quite large and it is an open problem

[1] It is often more convenient to fix the densities $\rho^\uparrow(x)$ and $\rho^\downarrow(x)$ of, respectively, spin-up and spin-down electrons instead of the total density $\rho(x) = \rho^\uparrow(x) + \rho^\downarrow(x)$. All our results apply similarly to this situation, as explained below in Remark 6.

to improve it. We hope that our work will stimulate more results on the functional $F_{\rm LL}$ in the regime of slowly varying densities.

In physics and chemistry, the *exchange-correlation energy* is defined by subtracting a kinetic energy term $T(\rho)$ from $F_{\rm LL}(\rho)$. In this paper we also derive a bound on $T(\rho)$ which, when combined with (2), provides a bound on the exchange-correlation energy similar to (2). This is explained below in Remark 5.

In the next section we provide the precise mathematical definition of $F_{\rm LL}$ and $e_{\rm UEG}$, and we state our main theorem containing the estimate (2). In Section 3 we review some known a priori estimates on $F_{\rm LL}$ and prove a new upper bound on the kinetic energy. Section 4 contains the proof of our main results. Finally, in the Appendix we discuss a similar bound in the classical case where the kinetic energy is dropped, extending thereby our previous result in [Lewin et al. 2018].

2. Main result

2.1. ***The grand-canonical Levy–Lieb functional.*** Let us consider a density $\rho \in L^1(\mathbb{R}^3, \mathbb{R}_+)$ such that $\sqrt{\rho} \in H^1(\mathbb{R}^3)$. Naturally we should assume in addition that $\int_{\mathbb{R}^3} \rho = N$ is an integer, but here we will work in the grand-canonical ensemble where this is not needed. The *grand-canonical Levy–Lieb functional* [Levy 1979; Lieb 1983; Lewin et al. 2018] is defined by

$$F_{\rm LL}(\rho) := \inf_{\substack{\Gamma_n = \Gamma_n^* \geqslant 0 \\ \sum_{n=0}^\infty \operatorname{Tr}(\Gamma_n) = 1 \\ \sum_{n=1}^\infty \rho_{\Gamma_n} = \rho}} \left\{ \sum_{n=1}^\infty \operatorname{Tr}_{\mathfrak{H}^n} \left(-\sum_{j=1}^n \Delta_{x_j} + \sum_{1 \leqslant j < k \leqslant n} \frac{1}{|x_j - x_k|} \right) \Gamma_n \right\}. \tag{3}$$

Here

$$\mathfrak{H}^n := L^2_a((\mathbb{R}^3 \times \{1, \dots, q\})^n, \mathbb{C})$$

is the n-particle space of antisymmetric square-integrable functions on $(\mathbb{R}^3 \times \{1, \dots, q\})^n$, with q spin states (for electrons $q = 2$). The family of operators $\Gamma = \{\Gamma_n\}_{n \geqslant 0}$ forms a grand-canonical mixed quantum state, that is, a state over the fermionic Fock space (commuting with the particle number operator). The density of each Γ_n is defined by

$$\rho_{\Gamma_n}(x) = n \sum_{\substack{\sigma_1, \dots, \sigma_n \\ \in \{1, \dots, q\}}} \int_{\mathbb{R}^{3(n-1)}} \Gamma_n(x, \sigma_1, x_2, \dots, x_n, \sigma_n; x, \sigma_1, x_2, \dots, x_n, \sigma_n)\, dx_2 \cdots dx_n,$$

where $\Gamma_n(x_1, \dots, \sigma_n; x_1', \dots, \sigma_n')$ is the kernel of the trace-class operator Γ_n. This kernel is such that

$$\begin{aligned} &\Gamma_n(x_{\tau(1)}, \sigma_{\tau(1)}, \dots, x_{\tau(N)}, \sigma_{\tau(N)}; x_1', \sigma_1', \dots, x_N', \sigma_N') \\ &\qquad = \Gamma_n(x_1, \sigma_1, \dots, x_N, \sigma_N; x'_{\tau(1)}, \sigma'_{\tau(1)}, \dots, x'_{\tau(N)}, \sigma'_{\tau(N)}) \\ &\qquad = \varepsilon(\tau)\, \Gamma_n(x_1, \sigma_1, \dots, x_N, \sigma_N; x_1', \sigma_1', \dots, x_N', \sigma_N') \end{aligned}$$

for every permutation $\tau \in \mathfrak{S}^N$ with signature $\varepsilon(\tau) \in \{\pm 1\}$.

If $\int_{\mathbb{R}^3} \rho = N \in \mathbb{N}$ and we restrict ourselves to mixed states Γ where only Γ_N is nonzero, we obtain Lieb's functional [1983]. If we further assume that $\Gamma_N = |\Psi\rangle\langle\Psi|$ is a rank-one projection, then we

find the original Levy [1979] or Hohenberg–Kohn functional [1964]. It is well known [Lieb 1983] that working with mixed states has several advantages, in particular we obtain a convex function of ρ.

The grand-canonical version (3) is less popular but still important physically.[2] It is also a convex function of ρ. The fact that we can appeal to states with an arbitrary number n of particles (but still a fixed average number $\int_{\mathbb{R}^3}\rho$) will considerably simplify several technical parts of our study. We expect that our main result (Theorem 3 below) holds the same for the canonical functionals, for which the energy is minimized over mixed or pure states with N particles.

It is useful to subtract the direct term from F_{LL}, hence to consider the energy

$$E(\rho) := F_{\mathrm{LL}}(\rho) - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy. \tag{4}$$

The (grand-canonical) *exchange-correlation energy* is defined by $E_{\mathrm{xc}}(\rho) := E(\rho) - T(\rho)$, where

$$T(\rho) := \inf_{\substack{\Gamma_n=\Gamma_n^*\geqslant 0\\ \sum_{n=0}^\infty \mathrm{Tr}(\Gamma_n)=1\\ \sum_{n=1}^\infty \rho_{\Gamma_n}=\rho}} \left\{\sum_{n=1}^\infty \mathrm{Tr}_{\mathfrak{H}^n}\left(-\sum_{j=1}^n \Delta_{x_j}\right)\Gamma_n\right\} = \inf_{\substack{0\leqslant\gamma=\gamma^*\leqslant 1\\ \rho_\gamma=\rho}} \mathrm{Tr}(-\Delta)\gamma \tag{5}$$

is the lowest possible kinetic energy. We will study the functional $T(\rho)$ in Section 3.2 below.

2.2. ***The uniform electron gas.*** In [Lewin et al. 2018, Section 5], we have defined the uniform electron gas, which is obtained in the limit when ρ approaches a constant function in the whole space. This is believed to be the same as the ground state energy of jellium, where the density is not necessarily constant but the electrons instead evolve in a constant background [Lieb and Narnhofer 1975]. This has recently been proved in the classical case in [Lewin et al. 2019; Cotar and Petrache 2019a] and the same is expected in the quantum case.

The following result is a slight improvement of [Lewin et al. 2018, Theorem 5.1].

Theorem 1 (quantum uniform electron gas). *Let $\rho_0 > 0$. Let $\{\Omega_N\}\subset\mathbb{R}^3$ be a sequence of bounded connected domains with $|\Omega_N|\to\infty$ such that Ω_N has a uniformly regular boundary in the sense that*

$$|\partial\Omega_N + B_r| \leqslant C r|\Omega_N|^{\frac{2}{3}} \quad \text{for all } r\leqslant |\Omega_N|^{\frac{1}{3}}/C,$$

for some constant $C>0$. Let $\delta_N>0$ be any sequence such that $\delta_N/|\Omega_N|^{1/3}\to 0$ and $\delta_N|\Omega_N|^{1/3}\to\infty$. Let $\chi\in L^1(\mathbb{R}^3)$ be a radial nonnegative function of compact support such that $\int_{\mathbb{R}^3}\chi = 1$ and $\int_{\mathbb{R}^3}|\nabla\sqrt{\chi}|^2<\infty$. Set $\chi_\delta(x)=\delta^{-3}\chi(x/\delta)$. Then the following thermodynamic limit exists:

$$\lim_{N\to\infty}\frac{E(\rho_0\mathbb{1}_{\Omega_N}*\chi_{\delta_N})}{|\Omega_N|} = e_{\mathrm{UEG}}(\rho_0), \tag{6}$$

where the function e_{UEG} is independent of the sequence $\{\Omega_N\}$, of δ_N, and of χ.

For more properties of the UEG energy e_{UEG} we refer to [Lewin et al. 2018]. In Theorem 5.1 of that paper we rather optimized over the values of ρ in the transition region around $\partial\Omega_N$. We were able to

[2]The grand-canonical functional F_{LL} is the weak-$*$ lower semicontinuous closure of the canonical Lieb functional [Lewin 2011]. Hence it appears naturally in situations where some particles can be lost, e.g., in scattering processes.

prove the simple limit (6) only when Ω_N is a tetrahedron. Using an upper bound on $E(\rho)$ that will be derived later in Proposition 13, we are now able to treat more reasonable limits in the form of (6). The proof of Theorem 1 is provided in Section 4.4 below.

The function χ_{δ_N} is used to regularize the function $\rho_0 \mathbb{1}_{\Omega_N}$, which cannot be the density of a quantum state since its square root is not in $H^1(\mathbb{R}^3)$ [Lieb 1983]. The first condition $\delta_N/|\Omega_N|^{1/3} \to 0$ implies that the smearing happens in a neighborhood of the boundary $\partial\Omega_N$ which has a negligible volume compared to $|\Omega_N|$. The second condition $\delta_N|\Omega_N|^{1/3} \to \infty$ ensures that the kinetic energy in the transition region stays negligible in the thermodynamic limit.

Remark 2. The same result holds under the weaker condition that Ω_N has an η-regular boundary, which means that $|\partial\Omega_N + B_r| \leqslant C|\Omega_N|\eta(r|\Omega_N|^{-1/3})$ for all $r \leqslant |\Omega_N|^{1/3}/C$, with $\eta(t) \to 0$ when $t \to 0^+$. The condition $\delta_N|\Omega_N|^{1/3} \to \infty$ is then replaced by $\delta_N^{-2}\eta(\delta_N|\Omega_N|^{-1/3}) \to 0$.

2.3. ***The local density approximation.*** We are now able to state our main result.

Theorem 3 (local density approximation). *Let $p > 3$ and $0 < \theta < 1$ such that*

$$2 \leqslant p\theta \leqslant 1 + \tfrac{1}{2}p. \tag{7}$$

There exists a constant $C = C(p, \theta, q)$ such that

$$\left| E(\rho) - \int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho(x))\,dx \right| \leqslant \varepsilon \int_{\mathbb{R}^3} (\rho(x) + \rho(x)^2)\,dx + \frac{C(1+\varepsilon)}{\varepsilon} \int_{\mathbb{R}^3} |\nabla\sqrt{\rho}(x)|^2\,dx + \frac{C}{\varepsilon^{4p-1}} \int_{\mathbb{R}^3} |\nabla\rho^\theta(x)|^p\,dx \tag{8}$$

for every $\varepsilon > 0$ and every nonnegative density $\rho \in L^1(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$ such that $\nabla\sqrt{\rho} \in L^2(\mathbb{R}^3)$ and $\nabla\rho^\theta \in L^p(\mathbb{R}^3)$.

The constant $C = C(p, \theta, q)$ in our estimate (8) depends on the number of spin states q ($q = 2$ for electrons), in addition to the parameters p and θ. It diverges when $p \to 3^+$. If $p \to 3^+$ then we can take $\theta \to \frac{5}{6}^-$. Our estimate therefore applies to densities ρ with compact support, which vanish at the boundary of their support like $\delta(x)^a$ with $a > \frac{4}{5}$, where $\delta(x) = \mathrm{d}(x, \partial\rho^{-1}(\{0\}))$. In particular, densities which vanish linearly are allowed. Our proof allows one to consider more singular densities, that is, to relax the constraint that $\theta p \leqslant 1 + \frac{1}{2}p$, but then the power of ε deteriorates.

Our estimate (8) is certainly not optimal and it is an interesting challenge to improve it. We conjecture that a similar inequality holds with $\rho + \rho^2$ replaced by $\rho^{4/3} + \rho^{5/3}$, which have the scaling of the Coulomb and kinetic energies, respectively. The higher power ρ^2 arises from the trial state used in our upper bound (Proposition 13) and it is used to control some errors appearing when merging quantum systems with overlapping supports. This is explained in Section 4.1 below. Finally, the last gradient term in (8) is used to control local variations of ρ in L^∞. One could expect gradient errors involving only $\int_{\mathbb{R}^3} |\nabla\sqrt{\rho}|^2$ and $\int_{\mathbb{R}^3} |\nabla\rho^{1/3}|^2$ which are believed to arise in the gradient expansion of the uniform electron gas for, respectively, the Coulomb and kinetic energies.

One interesting case is when the density is given by a fixed function ρ with $\int_{\mathbb{R}^3} \rho = 1$, which is rescaled in the manner

$$\rho_N(x) = \rho(N^{-\frac{1}{3}}x).$$

After taking $\varepsilon = N^{-1/12}$ in (8) we obtain the simple bound

$$\left| E(\rho_N) - N \int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho(x))\, dx \right| \leqslant C N^{\frac{5}{12}} \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}(x)|^2\, dx + C N^{\frac{11}{12}} \int_{\mathbb{R}^3} (\rho(x) + \rho(x)^2 + |\nabla \rho^\theta(x)|^p)\, dx. \quad (9)$$

It is conjectured [Kirzhnits 1957; Langreth and Perdew 1980; Langreth and Mehl 1983; Levy and Perdew 1993; Engel and Dreizler 2011] that the next order in the expansion of $E(\rho_N)$ should involve the gradient correction to the kinetic energy

$$\int_{\mathbb{R}^3} |\nabla \sqrt{\rho_N}(x)|^2\, dx = N^{\frac{1}{3}} \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}(x)|^2\, dx$$

and the gradient correction to the Coulomb energy

$$\int_{\mathbb{R}^3} |\nabla \rho_N^{\frac{1}{3}}(x)|^2\, dx = N^{\frac{1}{3}} \int_{\mathbb{R}^3} |\nabla \rho^{\frac{1}{3}}(x)|^2\, dx.$$

In particular, the next order should be proportional to $N^{1/3}$. It remains an open problem to establish this rigorously.

In the classical case where the kinetic energy is neglected, the limit of $E_{\mathrm{cl}}(\rho_N)/N$ was found in our previous work [Lewin et al. 2018], but without a quantitative estimate on the remainder. We can give an estimate similar to (8) in the classical case, with a lower power of ε in front of the gradient term. This is just a slight adaptation of the proof in [Lewin et al. 2018], which is much easier than the quantum case. The argument is explained for completeness in the Appendix.

In the classical case the limit for $E_{\mathrm{cl}}(\rho_N)/N$ was later extended to Riesz interactions $|x|^{-s}$ and other dimensions $d \geqslant 1$ in [Cotar and Petrache 2019b]. Although our result (8) in the quantum case can probably be extended to other Riesz interactions by using ideas from [Fefferman 1985; Hughes 1985; Gregg 1989; Cotar and Petrache 2019b], we only consider here the physically relevant three-dimensional Coulomb case for shortness.

Remark 4 (canonical case). We expect an inequality similar to (8) for the (mixed) canonical version of $E(\rho)$, where $\int_{\mathbb{R}^3} \rho = N \in \mathbb{N}$ and $\Gamma_n = 0$ for $n \neq N$. However, our proof does not adapt in an obvious way to this case.

Remark 5 (exchange-correlation energy). In physics and chemistry, the LDA is usually expressed in terms of the *exchange-correlation energy*. In the grand-canonical setting it is defined by

$$E_{\mathrm{xc}}(\rho) = E(\rho) - T(\rho),$$

with $T(\rho)$ the lowest possible kinetic energy (5). The functional $T(\rho)$ is studied in Section 3 below, where it is proved that

$$\left| T(\rho) - q^{-\frac{2}{3}} c_{\mathrm{TF}} \int_{\mathbb{R}^3} \rho(x)^{\frac{5}{3}}\, dx \right| \leqslant \varepsilon q^{-\frac{2}{3}} \int_{\mathbb{R}^3} \rho(x)^{\frac{5}{3}}\, dx + \frac{C}{\varepsilon^{\frac{13}{3}}} \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}(x)|^2\, dx, \tag{10}$$

with $c_{\mathrm{TF}} = \frac{1}{5} 3^{5/3} 4^{1/3} \pi^{4/3}$ the Thomas–Fermi constant. The lower bound was derived by Nam [2018] and we prove the missing upper bound (with a better power of ε) in Theorem 8 below. Actually, by following our proof of Theorem 3 (simply discarding the Coulomb interaction) we can also prove a lower bound on $T(\rho)$, with an error similar to the right side of (8) but with a smaller power of ε in front of $|\nabla \rho^\theta|^p$. This provides the following estimate on the exchange-correlation energy:

$$\begin{aligned} &\left| E_{\mathrm{xc}}(\rho) - \int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho(x))\, dx + q^{-\frac{2}{3}} c_{\mathrm{TF}} \int_{\mathbb{R}^3} \rho(x)^{\frac{5}{3}}\, dx \right| \\ &\qquad \leqslant \varepsilon \int_{\mathbb{R}^3} (\rho(x) + \rho(x)^2)\, dx + \frac{C(1+\varepsilon)}{\varepsilon} \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}(x)|^2\, dx + \frac{C}{\varepsilon^{4p-1}} \int_{\mathbb{R}^3} |\nabla \rho^\theta(x)|^p\, dx. \end{aligned} \tag{11}$$

For a rescaled density $\rho_N(x) = \rho(x/N^{1/3})$ we obtain the same rate of convergence $N^{11/12}$ as in (9).

Remark 6 (local spin density approximation). In practice, it is often convenient to not fix the total density but, rather, the density of each spin component

$$\rho_\sigma(x) = \sum_{n \geqslant 1} n \sum_{\substack{\sigma_2, \dots, \sigma_n \\ \in \{1, \dots, q\}}} \int_{\mathbb{R}^{3(n-1)}} \Gamma_n(x, \sigma, x_2, \dots, x_n, \sigma_n; x, \sigma, x_2, \dots, x_n, \sigma_n)\, dx_2 \cdots dx_n$$

for $\sigma \in \{1, \dots, q\}$. Much as in Theorem 1 one can define the corresponding spin-polarized UEG energy $e_{\mathrm{UEG}}(\rho_1, \dots, \rho_q)$ of the uniform electron gas where the electrons of spin σ are assumed to have the constant density ρ_σ. By following the arguments in this paper, one can then prove an estimate similar to (8):

$$\begin{aligned} &\left| E(\rho_1, \dots, \rho_q) - \int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho_1(x), \dots, \rho_q(x))\, dx \right| \\ &\qquad \leqslant \varepsilon \int_{\mathbb{R}^3} (\rho(x) + \rho(x)^2)\, dx + \frac{C(1+\varepsilon)}{\varepsilon} \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}(x)|^2\, dx + \frac{C}{\varepsilon^{4p-1}} \int_{\mathbb{R}^3} |\nabla \rho^\theta(x)|^p\, dx. \end{aligned} \tag{12}$$

It is only for simplicity of notation that we work with the total density $\rho = \sum_{\sigma=1}^q \rho_\sigma$.

3. A priori estimates on $T(\rho)$ and $E(\rho)$

Lower bounds on $E(\rho)$ in (4) are well known and will be recalled below. Upper bounds are somewhat difficult to derive due to the constraint that the quantum states considered need to have the exact given density ρ. In this section we prove an upper bound on the best kinetic energy and use it to derive an upper bound on $E(\rho)$. Because our bounds are of independent interest we work in this section in any dimension $d \geqslant 1$. First we quickly recall the known lower bounds.

3.1. ***Known lower bounds.*** We recall that the Lieb–Thirring inequality [1975; 1976], see also [Lieb and Seiringer 2010], states that there exists a positive constant $c_{\mathrm{LT}} = c_{\mathrm{LT}}(d) > 0$ such that

$$\mathrm{Tr}(-\Delta)\gamma \geqslant q^{-\frac{2}{d}} c_{\mathrm{LT}} \int_{\mathbb{R}^d} \rho_\gamma(x)^{1+\frac{2}{d}}\, dx \tag{13}$$

for every self-adjoint operator γ on $L^2(\mathbb{R}^d, \mathbb{C}^q)$ such that $0 \leqslant \gamma \leqslant 1$. The best constant c_{LT} is unknown but has been conjectured to be the semiclassical constant

$$c_{\mathrm{TF}} = \frac{4\pi^2 d}{(d+2)} \left(\frac{d}{|\mathbb{S}^{d-1}|} \right)^{\frac{2}{d}} \tag{14}$$

in dimension $d \geqslant 3$. Nam [2018] has proved that

$$\mathrm{Tr}(-\Delta)\gamma \geqslant q^{-\frac{2}{d}} c_{\mathrm{TF}}(1-\varepsilon) \int_{\mathbb{R}^d} \rho_\gamma(x)^{1+\frac{2}{d}}\, dx - \frac{\kappa}{\varepsilon^{3+\frac{4}{d}}} \int_{\mathbb{R}^d} |\nabla \sqrt{\rho_\gamma}(x)|^2\, dx \tag{15}$$

for every $\varepsilon > 0$ and some constant $\kappa = \kappa(d)$, in all space dimensions $d \geqslant 1$.

We also recall the Hoffmann-Ostenhof inequality [Hoffmann-Ostenhof and Hoffmann-Ostenhof 1977], which states that

$$\mathrm{Tr}(-\Delta)\gamma \geqslant \int_{\mathbb{R}^d} |\nabla \sqrt{\rho_\gamma}(x)|^2\, dx \tag{16}$$

and always imposes that $\sqrt{\rho} \in H^1(\mathbb{R}^d)$. The inequality (16) does not require the fermionic constraint $0 \leqslant \gamma \leqslant 1$.

The Lieb–Oxford inequality [1981], see also [Lieb 1979; Chan and Handy 1999; Lieb and Seiringer 2010], states that the total Coulomb energy is bounded from below by

$$\sum_{n=1}^{\infty} \mathrm{Tr}_{\mathfrak{H}^n} \left(\sum_{1 \leqslant j < k \leqslant n} \frac{1}{|x_j - x_k|} \right) \Gamma_n \geqslant \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\, dx\, dy - 1.64 \int_{\mathbb{R}^3} \rho(x)^{\frac{4}{3}}\, dx, \tag{17}$$

where $\Gamma = \{\Gamma_n\}$ is a grand-canonical quantum state satisfying the conditions in (4). Inspired by [Benguria et al. 2012], this bound was recently generalized in [Lewin and Lieb 2015] to

$$\begin{aligned} &\sum_{n=1}^{\infty} \mathrm{Tr}_{\mathfrak{H}^n} \left(\sum_{1 \leqslant j < k \leqslant n} \frac{1}{|x_j - x_k|} \right) \Gamma_n \\ &\qquad \geqslant \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\, dx\, dy - \left(\frac{3}{5} \left(\frac{9\pi}{2} \right)^{\frac{1}{3}} + \varepsilon \right) \int_{\mathbb{R}^3} \rho(x)^{\frac{4}{3}}\, dx - \frac{0.001206}{\varepsilon^3} \int_{\mathbb{R}^3} |\nabla \rho(x)|\, dx, \end{aligned} \tag{18}$$

but the constant $\frac{3}{5}\left(\frac{9\pi}{2}\right)^{1/3} \simeq 1.4508$ is not expected to be the optimal Lieb–Oxford constant.

Using (13) together with (17), we obtain the following.

Corollary 7 (lower bound on $E(\rho)$). *We have*

$$E(\rho) \geqslant q^{-\frac{2}{3}} c_{\mathrm{LT}} \int_{\mathbb{R}^3} \rho(x)^{\frac{5}{3}}\, dx - 1.64 \int_{\mathbb{R}^3} \rho(x)^{\frac{4}{3}}\, dx \tag{19}$$

for every $\rho \geqslant 0$ such that $\sqrt{\rho} \in H^1(\mathbb{R}^d)$.

The constants can be improved at the expense of adding gradient corrections, using (15) and (18).

3.2. ***Upper bound on the best kinetic energy.*** Let us recall that the lowest possible kinetic energy of a fixed density $\rho \in L^1(\mathbb{R}^d, \mathbb{R}_+)$ with $\sqrt{\rho} \in H^1(\mathbb{R}^d)$ reads

$$T(\rho) := \min_{\substack{0 \leqslant \gamma = \gamma^* \leqslant 1 \\ \rho_\gamma = \rho}} \operatorname{Tr}(-\Delta)\gamma.$$

In [Lewin et al. 2018], we have shown that, for $\rho \leqslant 1$,

$$T(\rho) \leqslant \int_{\mathbb{R}^d} |\nabla \sqrt{\rho}|^2 + q^{-\frac{2}{d}} c_{\mathrm{TF}} \int_{\mathbb{R}^d} \rho$$

by using the trial state

$$\gamma = \sqrt{\rho(x)}\, \mathbb{1}\left(-\Delta \leqslant \frac{d+2}{d} c_{\mathrm{TF}}\, q^{-\frac{2}{d}}\right) \sqrt{\rho(x)}.$$

Our goal in this section is to prove a similar bound without the assumption that $\rho \leqslant 1$. Coherent states [Lieb 1981a] can usually give good bounds on the kinetic energy but they do not preserve the density. The main difficulty here is to construct a state having the exact given density ρ. The next result says that the semiclassical approximation to the kinetic energy is an upper bound to the exact $T(\rho)$, up to some gradient corrections.

Theorem 8 (upper bound on the best kinetic energy). *There are two constants $\kappa_1, \kappa_2 > 0$ depending only on the space dimension d such that*

$$T(\rho) \leqslant q^{-\frac{2}{d}} c_{\mathrm{TF}} (1 + \kappa_1 \varepsilon) \int_{\mathbb{R}^d} \rho(x)^{1+\frac{2}{d}}\, dx + \frac{\kappa_2 (1+\sqrt{\varepsilon})^2}{\varepsilon} \int_{\mathbb{R}^d} |\nabla \sqrt{\rho}(x)|^2\, dx \tag{20}$$

for every $\rho \in L^1(\mathbb{R}^d, \mathbb{R}_+)$ with $\sqrt{\rho} \in H^1(\mathbb{R}^d)$ and every $\varepsilon > 0$, where c_{TF} is the Thomas–Fermi constant (14).

Note that the gradient correction in (20) has a better behavior in ε than in Nam's lower bound (15).

In dimension $d = 1$, March and Young [1958, equation (9)] have given the proof of a better estimate without the parameter ε:

$$T(\rho) \leqslant q^{-2} c_{\mathrm{TF}} \int_{\mathbb{R}} \rho(x)^3\, dx + \int_{\mathbb{R}} |(\sqrt{\rho})'(x)|^2\, dx.$$

In the same paper they also state a result in three dimensions (for a constant $c > c_{\mathrm{TF}}$) but the proof has a mistake. This was mentioned as a conjecture in [Lieb 1983, Section 5.B]. Our result (20) can therefore be seen as a solution to the March–Young problem. We conjecture that a similar bound holds without the parameter ε in dimensions $d = 2, 3$ as well.

Remark 9 (explicit constants in three dimensions). In dimension $d = 3$ one can take

$$\kappa_1 = 1, \quad \kappa_2 = 48 \quad \text{in (20)}.$$

These constants are not optimal and they are only displayed for concreteness. Our proof allows us to slightly improve the constants under the assumption that ε is small enough. For instance, for $\varepsilon \leqslant 1$, we have the better inequality

$$T(\rho) \leqslant q^{-\frac{2}{3}} c_{\mathrm{TF}} \left(1 + \frac{\varepsilon}{15}\right) \int_{\mathbb{R}^3} \rho(x)^{\frac{5}{3}}\, dx + \frac{19}{\varepsilon} \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}(x)|^2\, dx. \tag{21}$$

Proof of Theorem 8. For simplicity we only write the proof in the no-spin case $q = 1$. Recall that the free Fermi sea

$$P_r = \mathbb{1}\left(-\Delta \leqslant \frac{d+2}{d} c_{\mathrm{TF}}\, r^{\frac{2}{d}}\right)$$

has the constant density $\rho_{P_r} = r$ and the constant kinetic energy density $c_{\mathrm{TF}}\, r^{1+2/d}$. In particular, if we take

$$\gamma = \sqrt{f(x)}\, \mathbb{1}\left(-\Delta \leqslant \frac{d+2}{d} c_{\mathrm{TF}}\, r^{\frac{2}{d}}\right) \sqrt{f(x)}$$

for some $f \geqslant 0$, we obtain

$$\rho_\gamma = r f(x), \quad \mathrm{Tr}(-\Delta)\gamma = r \int_{\mathbb{R}^d} |\nabla \sqrt{f}(x)|^2\, dx + c_{\mathrm{TF}}\, r^{1+\frac{2}{d}} \int_{\mathbb{R}^d} f(x)\, dx. \tag{22}$$

See for instance [Lewin et al. 2018, Section 5] for details. In addition, we have in the sense of operators $0 \leqslant \gamma \leqslant f(x)$; hence γ is a fermionic one-particle density matrix under the additional condition that $f \leqslant 1$.

Let now $\eta \geqslant 0$ be a smooth nonnegative function such that

$$\int_0^\infty \eta(t)\, dt = 1, \quad \int_0^\infty \eta(t)\, \frac{dt}{t} \leqslant 1. \tag{23}$$

Using the smooth layer cake principle

$$\rho(x) = \int_0^\infty \eta\left(\frac{t}{\rho(x)}\right) dt$$

we introduce the trial state

$$\gamma = \int_0^\infty \sqrt{\eta\left(\frac{t}{\rho(x)}\right)}\, \mathbb{1}\left(-\Delta \leqslant \frac{d+2}{d} c_{\mathrm{TF}}\, t^{\frac{2}{d}}\right) \sqrt{\eta\left(\frac{t}{\rho(x)}\right)}\, \frac{dt}{t}.$$

In the sense of operators, we have

$$0 \leqslant \gamma \leqslant \int_0^\infty \eta\left(\frac{t}{\rho(x)}\right) \frac{dt}{t} = \int_0^\infty \eta(t)\, \frac{dt}{t} \leqslant 1.$$

In addition, γ has the required density

$$\rho_\gamma(x) = \int_0^\infty \eta\left(\frac{t}{\rho(x)}\right) dt = \rho(x).$$

Hence γ is admissible and it can be used to get an upper bound on $T(\rho)$. From (22), its kinetic energy is

$$\mathrm{Tr}(-\Delta)\gamma = \int_{\mathbb{R}^d} dx \int_0^\infty dt \left| \nabla_x \sqrt{\eta\left(\frac{t}{\rho(x)}\right)} \right|^2 + c_{\mathrm{TF}} \int_{\mathbb{R}^d} \rho(x)^{1+\frac{2}{d}}\, dx \int_0^\infty \eta(t) t^{\frac{2}{d}}\, dt.$$

Note that

$$\begin{aligned}
\int_{\mathbb{R}^d} dx \int_0^\infty dt \left| \nabla_x \sqrt{\eta\left(\frac{t}{\rho(x)}\right)} \right|^2 &= \int_{\mathbb{R}^d} dx \int_0^\infty dt\, \frac{|\nabla_x \eta(t/\rho(x))|^2}{4\eta(t/\rho(x))} \\
&= \int_{\mathbb{R}^d} dx\, \frac{|\nabla\rho(x)|^2}{4\rho(x)^4} \int_0^\infty t^2\, dt\, \frac{|\eta'(t/\rho(x))|^2}{\eta(t/\rho(x))} \\
&= \int_{\mathbb{R}^d} |\nabla\sqrt{\rho}(x)|^2\, dx \int_0^\infty \frac{t^2\eta'(t)^2}{\eta(t)}\, dt.
\end{aligned}$$

Hence we have proved that

$$\mathrm{Tr}(-\Delta)\gamma = \int_{\mathbb{R}^d} |\nabla\sqrt{\rho}(x)|^2\, dx \int_0^\infty \frac{t^2\eta'(t)^2}{\eta(t)}\, dt + c_{\mathrm{TF}} \int_{\mathbb{R}^d} \rho(x)^{1+\frac{2}{d}}\, dx \int_0^\infty \eta(t) t^{\frac{2}{d}}\, dt$$

for all $\eta \geqslant 0$ satisfying the two constraints (23). The smallest constant we can get in front of the $\rho^{1+2/d}$ term is c_{TF}, by concentrating η at the point $t = 1$, but this makes the other term blow up. If we fix

$$\int_0^\infty \eta(t) t^{\frac{2}{d}}\, dt = 1 + \varepsilon$$

then the best constant we can get in front of the gradient term is given by the variational problem

$$C(\varepsilon) := \inf_{\substack{\eta\geqslant 0 \\ \int_0^\infty \eta = 1 \\ \int_0^\infty \eta/t \leqslant 1 \\ \int_0^\infty t^{2/d}\eta \leqslant 1+\varepsilon}} \int_0^\infty \frac{t^2\eta'(t)^2}{\eta(t)}\, dt.$$

We claim that $C(\varepsilon) \leqslant \text{const.}(1+\varepsilon^{-1})$ for ε small enough, which we prove by an appropriate choice of η.

Let us first take, for instance,

$$\eta_\varepsilon(t) = \frac{3}{2\varepsilon^3}(t-1)^2 \mathbb{1}(1 \leqslant t \leqslant 1+\varepsilon) + \frac{3}{2\varepsilon^3}(1+2\varepsilon-t)^2 \mathbb{1}(1+\varepsilon \leqslant t \leqslant 1+2\varepsilon). \tag{24}$$

Then $\int \eta_\varepsilon = 1$ and $\int \eta_\varepsilon(t)/t\, dt \leqslant 1$ since η_ε is supported on $[1,\infty)$. Using the simple bounds

$$\begin{aligned}
&\int_0^\infty \eta(t) t^{\frac{2}{d}}\, dt \leqslant (1+2\varepsilon)^{2/d}, \\
&\int_0^\infty \frac{t^2\eta'(t)^2}{\eta(t)}\, dt \leqslant (1+2\varepsilon)^2 \int_0^\infty \frac{\eta'(t)^2}{\eta(t)}\, dt = \frac{12(1+2\varepsilon)^2}{\varepsilon^2},
\end{aligned}$$

we obtain $\big($after changing ε into $\frac{1}{2}\varepsilon\big)$

$$T(\rho) \leqslant c_{\mathrm{TF}}(1+\varepsilon)^{\frac{2}{d}} \int_{\mathbb{R}^d} \rho(x)^{1+\frac{2}{d}}\, dx + \frac{48(1+\varepsilon)^2}{\varepsilon^2} \int_{\mathbb{R}^d} |\nabla\sqrt{\rho}(x)|^2\, dx \quad \text{for all } \varepsilon > 0. \tag{25}$$

The behavior of the correction in front of the semiclassical term c_{TF} is not optimal for small ε. It can be replaced by $1+\kappa_1\varepsilon^2$ for $\varepsilon \leqslant 1$. To see this we slightly translate the function (24) to the left by an amount $-\varepsilon b$ and introduce

$$\eta_{\varepsilon,b}(t) = \frac{3}{2\varepsilon^3}(t-1+\varepsilon b)^2 \mathbb{1}(1-\varepsilon b \leqslant t \leqslant 1+(1-b)\varepsilon) + \frac{3}{2\varepsilon^3}(1+(2-b)\varepsilon - t)^2 \mathbb{1}(1-\varepsilon b+\varepsilon \leqslant t \leqslant 1+(2-b)\varepsilon). \quad (26)$$

Then we have

$$\int_0^\infty \eta_{\varepsilon,b}(t)\frac{dt}{t} = 1+(b-1)\varepsilon + \left(\tfrac{11}{10}-2b+b^2\right)\varepsilon^2 + \left(-\tfrac{13}{10}+\tfrac{33}{10}\varepsilon - 3b^2+b^3\right)\varepsilon^3 + O(\varepsilon^4),$$

$$\int_0^\infty \eta_{\varepsilon,b}(t) t^{\frac{2}{d}}\, dt = 1 - \frac{2}{d}(b-1)\varepsilon + \frac{1}{10d^2}(22-40b+20b^2-11d+20bd-10b^2d)\varepsilon^2 + O(\varepsilon^3).$$

The unique b_ε such that $\int_0^\infty \eta_{\varepsilon,b_\varepsilon}(t)\frac{dt}{t} = 1$ satisfies

$$b_\varepsilon = 1 - \tfrac{1}{10}\varepsilon - \tfrac{3}{350}\varepsilon^3 + O(\varepsilon^4)$$

and for this b_ε we have

$$\int_0^\infty \eta_{\varepsilon,b_\varepsilon}(t) t^{\frac{2}{d}}\, dt = 1 + \frac{2+d}{10d^2}\varepsilon^2 + O(\varepsilon^4),$$

$$\int_0^\infty \frac{t^2 \eta'_{\varepsilon,b_\varepsilon}(t)^2}{\eta_{\varepsilon,b_\varepsilon}(t)}\, dt = \frac{12}{\varepsilon^2} + O(1).$$

This is how we can get (21) for ε small enough (after replacing ε^2 by ε). □

Remark 10. In the three-dimensional case we can take for instance $b = 1 - \frac{1}{10}\varepsilon - \frac{4}{350}\varepsilon^3$. One can then verify that

$$\int_0^\infty \eta_{\varepsilon,b}(t)\frac{dt}{t} \leqslant 1, \quad \int_0^\infty \eta_{\varepsilon,b}(t) t^{\frac{2}{3}}\, dt \leqslant 1 + \frac{\varepsilon^2}{15},$$

and

$$\int_0^\infty \frac{t^2 \eta'_{\varepsilon,b_\varepsilon}(t)^2}{\eta_{\varepsilon,b_\varepsilon}(t)}\, dt \leqslant \frac{19}{\varepsilon^2}$$

for all $\varepsilon \leqslant 1$. Hence

$$T(\rho) \leqslant c_{\mathrm{TF}}\left(1+\frac{\varepsilon^2}{15}\right)\int_{\mathbb{R}^3} \rho(x)^{\frac{5}{3}}\, dx + \frac{19}{\varepsilon^2}\int_{\mathbb{R}^3} |\nabla\sqrt{\rho}(x)|^2\, dx \quad \text{for all } \varepsilon \leqslant 1. \quad (27)$$

Combining with (25), we find the estimate (20) for $\kappa_1 = 1$ and $\kappa_2 = 48$.

3.3. ***Upper bound on $E(\rho)$.*** It is well known that any fermionic one-particle density matrix γ (i.e., an operator satisfying $0 \leqslant \gamma = \gamma^* \leqslant 1$) is representable by a quasifree state Γ_γ in Fock space [Bach et al. 1994]. The two-particle density matrix of such a state is given by Wick's formula

$$\Gamma_\gamma^{(2)}(x_1,\sigma_1,x_2,\sigma_2;y_1,\sigma_1',y_2,\sigma_2') = \gamma(x_1,\sigma_1;y_1,\sigma_1')\gamma(x_2,\sigma_2;y_2,\sigma_2') - \gamma(x_1,\sigma_1;x_2,\sigma_2)\gamma(y_1,\sigma_1';y_2,\sigma_2'). \quad (28)$$

In particular, the corresponding interaction energy with pair potential w is

$$\frac{1}{2}\iint_{\mathbb{R}^d\times\mathbb{R}^d} w(x-y)\bigg(\rho(x)\rho(y)-\sum_{\sigma,\sigma'=1}^{q}|\gamma(x,\sigma;y,\sigma')|^2\bigg)\,dx\,dy.$$

From this we immediately obtain the following.

Corollary 11 (upper bound on $E(\rho)$ in dimension $d\geqslant 1$). *We have*

$$E(\rho)\leqslant q^{-\frac{2}{d}}c_{\mathrm{TF}}(1+\kappa_1\varepsilon)\int_{\mathbb{R}^d}\rho(x)^{1+\frac{2}{d}}\,dx+\frac{\kappa_2(1+\sqrt{\varepsilon})^2}{\varepsilon}\int_{\mathbb{R}^d}|\nabla\sqrt{\rho}(x)|^2\,dx \tag{29}$$

for every $\rho\geqslant 0$ such that $\sqrt{\rho}\in H^1(\mathbb{R}^d)$.

This is for the grand-canonical version (4) of the Levy–Lieb functional which is the object of concern in this paper. It was proved in [Lieb 1981b] that any fermionic γ with integer trace $N=\mathrm{Tr}(\gamma)$ is also the one-particle density matrix of an N-particle mixed state Γ on the fermionic space $\mathfrak{H}^N$, such that the corresponding two-particle density matrix satisfies

$$\Gamma^{(2)}\leqslant\Gamma^{(2)}_\gamma$$

in the sense of operators. From the positivity of the Coulomb potential we deduce immediately the following result for mixed canonical states.

Corollary 12 (upper bound in the mixed canonical case). *Let $\rho\in L^1(\mathbb{R}^d,\mathbb{R}_+)$ be such that $\int_{\mathbb{R}^d}\rho=N\in\mathbb{N}$, and $\sqrt{\rho}\in H^1(\mathbb{R}^d)$. Then there exists a mixed state Γ on the fermionic space $\bigwedge_1^N L^2(\mathbb{R}^d\times\{1,\dots,q\})$ such that $\rho_\Gamma=\rho$ and*

$$\mathrm{Tr}\bigg(\sum_{j=1}^N-\Delta_{x_j}+\sum_{1\leqslant j<k\leqslant N}\frac{1}{|x_j-x_k|}\bigg)\Gamma \leqslant q^{-\frac{2}{d}}c_{\mathrm{LT}}(1+\kappa_1\varepsilon)\int_{\mathbb{R}^d}\rho(x)^{1+\frac{2}{d}}\,dx+\frac{\kappa_2(1+\sqrt{\varepsilon})^2}{\varepsilon}\int_{\mathbb{R}^d}|\nabla\sqrt{\rho}(x)|^2\,dx.$$

4. Proof of Theorems 1 and 3

Our proof is divided into several steps. The first is to show that the energy is essentially local, that is, to prove that

$$E(\rho)\approx\sum_k E(\rho\chi_k), \tag{30}$$

where $\{\chi_k\}$ is a smooth partition of unity, $\sum_k\chi_k=1$. The precise statement of (30) will involve upper and lower bounds, as well as an average over the translations, rotations and dilations of the partition itself. The lower bound was indeed already shown in [Lewin et al. 2018] using the Graf–Schenker inequality [1995]. The upper bound is the main new ingredient of our proof. The two bounds are derived in Sections 4.1 and 4.2. This will allow us to provide a rather simple proof of Theorem 1 in Section 4.4, using the convergence for tetrahedra which will be studied in Section 4.3.

In Section 4.5 we will estimate the deviation of the energy when we replace $\rho\chi_k$ by a constant function, say $\rho(x_k)\chi_k$ for some x_k in the support of χ_k. If ρ is essentially constant in the corresponding region, the error will be small, but if ρ is not constant we bound the energy using some gradient terms, utilizing our upper bound (29).

After showing the Lipschitz regularity of e_{UEG} in Section 4.6, we will be able to conclude the proof of Theorem 3 in Section 4.7. We replace $E(\rho(x_k)\chi_k)$ by $\int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho(x))\chi_k(x)\,dx$ when ρ is large enough on $\mathrm{supp}(\chi_k)$, using some quantitative estimates derived for tetrahedra in Section 4.3.

In the rest of the paper we call C a generic constant which can sometimes change from line to line, but which only depends on q (the number of spin states) and p, θ, the two parameters appearing in the statement of Theorem 3.

4.1. ***Upper bound in terms of local densities.*** Our main goal here is to give an upper bound on the energy $E(\rho)$ by splitting ρ into a sum of local densities. In the classical case, we have the exact subadditivity property (see [Lewin et al. 2018, Lemma 2.5])

$$E_{\mathrm{cl}}(\rho_1+\rho_2) \leqslant E_{\mathrm{cl}}(\rho_1)+E_{\mathrm{cl}}(\rho_2),$$

which considerably simplifies the analysis and was one of the main tools of our previous work [Lewin et al. 2018]. In particular we immediately find an upper bound in the form

$$E_{\mathrm{cl}}(\rho) \leqslant \sum_k E_{\mathrm{cl}}(\rho\chi_k)$$

for a partition of unity χ_k. In the quantum case this is not as easy. The first difficulty is that we cannot cut sharply and have to use a smooth partition of unity. This has the consequence that neighboring local densities overlap. But then, for two densities ρ_1 and ρ_2 with overlapping support, it is not obvious how to relate $E(\rho_1+\rho_2)$ with $E(\rho_1)$ and $E(\rho_2)$. This is due to the fermionic nature of the electrons which puts a very strong constraint on trial states. If we take two trial quantum states for ρ_1 and ρ_2, we cannot simply take their tensor product and use it as a trial state for $\rho_1+\rho_2$. The tensor product does not have the fermionic symmetry, and if we antisymmetrize it, the density is not equal to $\rho_1+\rho_2$ anymore.

For this reason, we will use an incomplete partition of unity with holes, in order to make sure that the local quantum states are not overlapping. Since we want to get the exact density, holes are however in principle not allowed. Instead of filling the holes with electrons, our idea is to rather average over all the possible rotations, translations and dilations of the partition of unity, which will make the holes disappear in average. All the arguments of this section apply the same to a tiling made of cubes, but for a better matching with the lower bound we will consider a tiling made of tetrahedra. Our lower bound relies on the Graf–Schenker inequality [1995] which requires the use of tetrahedra.

Let us consider the unit cube $C_1 = \left(-\frac{1}{2},\frac{1}{2}\right)^3$, which is the union of 24 disjoint identical tetrahedra $\mathbf{\Delta}_1,\dots,\mathbf{\Delta}_{24}$, all of volume $\frac{1}{24}$. Since the cube can be repeated in the whole space, we obtain a tiling of $\mathbb{R}^3$ with tetrahedra:

$$\mathbb{R}^3 = \bigcup_{z\in\mathbb{Z}^3}\bigcup_{j=1}^{24}(\bar{\mathbf{\Delta}}_j+z) \tag{31}$$

and the corresponding partition of unity

$$\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}\mathbb{1}_{\ell\mathbf{\Delta}_j}(x-\ell z)\equiv 1 \quad \text{for a.e. } x\in\mathbb{R}^3,$$

for any fixed size $\ell>0$ of the tiles. Any $\mathbf{\Delta}_j$ can be written as $\mathbf{\Delta}_j=\mu_j\mathbf{\Delta}$, where $\mathbf{\Delta}$ is a reference tetrahedron with 0 as its center of mass. Here $\mu_j=(z_j,R_j)\in C_1\times \mathrm{SO}(3)$ is an appropriate translation and rotation, which acts as $\mu_j x=R_j x-z_j$; hence $\mathbf{\Delta}_j=R_j\mathbf{\Delta}-z_j$. In each tetrahedron we now place the regularized characteristic function

$$\chi_{\ell,\delta,j}:=\frac{1}{(1-\delta/\ell)^3}\mathbb{1}_{\ell\mu_j(1-\delta/\ell)\mathbf{\Delta}}*\eta_\delta. \tag{32}$$

Here $\eta_\delta(x)=(10/\delta)^3\eta_1(10x/\delta)$, where η_1 is a fixed C_c^∞ nonnegative radial function with support in the unit ball and such that $\int_{\mathbb{R}^3}\eta_1=1$. Assuming that $\frac{1}{2}\delta\leqslant\ell$, the function $\chi_{\ell,\delta,j}$ has its support well inside $\ell\mathbf{\Delta}_j$, at a distance proportional to δ from its boundary. The prefactor has been chosen to ensure that

$$\int_{\mathbb{R}^3}\chi_{\ell,\delta,j}=\tfrac{1}{24}\ell^3=|\ell\mathbf{\Delta}|.$$

The function

$$\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}\chi_{\ell,\delta,j}(x-\ell z)$$

is equal to $(1-\delta/\ell)^{-3}>1$ inside the tiles but vanishes in a neighborhood of the boundary of the tiles. It is the incomplete partition of unity which we have mentioned above. We obtain a partition of unity after averaging over the translations of the tiling:

$$\frac{1}{\ell^3}\int_{C_\ell}\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}\chi_{\ell,\delta,j}(x-\ell z-\tau)\,d\tau=1 \quad \text{for a.e. } x\in\mathbb{R}^3. \tag{33}$$

Here $C_\ell=\left(-\frac{1}{2}\ell,\frac{1}{2}\ell\right)^3=\ell C_1$ is the cube of side length ℓ. This is because, for any $f\in L^1(\mathbb{R}^3)$,

$$\begin{aligned}\int_{C_\ell}\sum_{z\in\mathbb{Z}^3}f(x-\ell z-\tau)\,d\tau&=\int_{C_\ell}\sum_{z\in\ell\mathbb{Z}^3}f(x-z-\tau)\,d\tau\\&=\int_{\mathbb{R}^3}f(x-\tau)\,d\tau=\int_{\mathbb{R}^3}f(\tau)\,d\tau.\end{aligned}$$

The main result of this section is the following upper bound.

Proposition 13 (upper bound in terms of local densities). *There exists a universal constant C such that for any $\sqrt{\rho}\in H^1(\mathbb{R}^3)$, any $0<\delta<\frac{1}{2}\ell$, and any $0<\alpha<\frac{1}{2}$,*

$$\begin{aligned}E(\rho)\leqslant\left(\int_{1-\alpha}^{1+\alpha}\frac{ds}{s^4}\right)^{-1}\int_{1-\alpha}^{1+\alpha}\frac{dt}{t^4}\int_{\mathrm{SO}(3)}dR\int_{C_{t\ell}}\frac{d\tau}{(t\ell)^3}\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}E(\chi_{t\ell,t\delta,j}(R\cdot-t\ell z-\tau)\rho)\\+C\delta^2\log(\alpha^{-1})\int_{\mathbb{R}^3}\rho^2.\end{aligned} \tag{34}$$

In particular, we can find $\ell' \in (\ell(1-\alpha), \ell(1+\alpha))$, $\delta' \in (\delta(1-\alpha), \delta(1+\alpha))$ *and an isometry* $(\tau, R) \in \mathbb{R}^3 \times \mathrm{SO}(3)$ *such that*

$$E(\rho) \leqslant \sum_{z\in\mathbb{Z}^3} \sum_{j=1}^{24} E(\chi_{\ell',\delta',j}(R\cdot - \ell' z - \tau)\rho) + C\delta^2 \log(\alpha^{-1}) \int_{\mathbb{R}^3} \rho^2. \tag{35}$$

The right side of (34) involves our incomplete partition of unity with holes, which is rotated (with the rotation R), translated (with the translation τ) and dilated (with the dilation parameter t). The error is small only when δ (the size of the holes) is small. However we cannot take $\delta = 0$ since that would make the gradient of the densities $\chi_{t\ell,t\delta,j}(R\cdot - t\ell z - \tau)\rho$ blow up. Nevertheless, the statement is that the energy decouples and the holes can be neglected, at the expense of an error of the order $\delta^2 \int_{\mathbb{R}^3} \rho^2$. In (34) we use dilations for purely technical reasons, in order to better control error terms.

Proof of Proposition 13. Using (33), we write our density ρ as

$$\rho(x) = \frac{1}{\ell^3} \int_{C_\ell} \sum_{z\in\mathbb{Z}^3} \sum_{j=1}^{24} \chi_{\ell,\delta,j}(x - \ell z - \tau)\rho(x)\, d\tau. \tag{36}$$

For every fixed $\tau \in C_\ell$, we can construct a grand-canonical trial state Γ_τ having the density

$$\rho_{\Gamma_\tau}(x) = \sum_{z\in\mathbb{Z}^3} \sum_{j=1}^{24} \chi_{\ell,\delta,j}(x - \ell z - \tau)\, \rho(x).$$

For this we pick $\Gamma_\tau = \bigotimes_{j=1}^{24} \bigotimes_{z\in\mathbb{Z}^3} \Gamma_{\tau,z,j}$, where each $\Gamma_{\tau,z,j}$ has the density

$$\rho_{\Gamma_{\tau,z,j}}(x) = \chi_{\ell,\delta,j}(x - \ell z - \tau)\rho(x)$$

and minimizes the corresponding energy $E(\chi_{\ell,\delta,j}(\cdot - \ell z - \tau)\rho)$. Since the quantum states $\Gamma_{\tau,z,j}$ have disjoint supports, we can antisymmetrize the state Γ_τ in the standard manner. We denote by $\Gamma_{\tau,a}$ the antisymmetrized state. The energy of $\Gamma_{\tau,a}$ is equal to that of Γ_τ and so is its density $\rho_{\Gamma_{\tau,a}} = \rho_{\Gamma_\tau}$. Finally we take as trial state

$$\Gamma = \frac{1}{\ell^3} \int_{C_\ell} \Gamma_{\tau,a}\, d\tau,$$

which satisfies by construction that $\rho_\Gamma = \rho$. We find the upper bound

$$\begin{aligned} E(\rho) &\leqslant \frac{1}{\ell^3} \int_{C_\ell} E(\rho_{\Gamma_\tau})\, d\tau + \frac{1}{\ell^3} \int_{C_\ell} D(\rho_{\Gamma_\tau})\, d\tau - D(\rho) \\ &= \frac{1}{\ell^3} \int_{C_\ell} \sum_{z\in\mathbb{Z}^3} \sum_{j=1}^{24} E(\chi_{\ell,\delta,j}(\cdot - \ell z - \tau)\rho)\, d\tau + \frac{1}{\ell^3} \int_{C_\ell} D(\rho_{\Gamma_\tau} - \rho)\, d\tau. \end{aligned} \tag{37}$$

Here we employ the usual notation

$$D(\rho) := \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\, dx\, dy. \tag{38}$$

In the second line of (37) we have used that the energy of a tensor product of states of disjoint supports is the sum of the energies of the pieces [Lewin et al. 2018]. This is because the cross terms in the direct energy exactly cancel with the many-particle interactions of different states in the tensor product.

The error term in (37) is solely due to the nonlinearity of the direct term and it may be rewritten as

$$\frac{1}{\ell^3}\int_{C_\ell} D(\rho_{\Gamma_\tau}-\rho)\,d\tau = \frac{1}{\ell^3}\int_{C_\ell} D\bigg(\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}(\mathbb{1}_{\ell\mathbf{\Delta}_j}-\chi_{\ell,\delta,j})(\cdot-\ell z-\tau)\rho\bigg)\,d\tau.$$

For every real-valued $(\ell\mathbb{Z}^3)$-periodic function f, we have

$$\begin{aligned}\frac{1}{\ell^3}\int_{C_\ell} D(f(\cdot-\tau)\rho)\,d\tau &= \frac{1}{2}\sum_{k\in(2\pi/\ell)\mathbb{Z}^3}\left|\frac{1}{\ell^3}\int_{C_\ell} f(z)e^{-ik\cdot z}\,dz\right|^2 \iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{e^{ik\cdot(x-y)}}{|x-y|}\rho(x)\,\rho(y)\,dx\,dy\\ &= 2\pi\sum_{k\in(2\pi/\ell)\mathbb{Z}^3}\left|\frac{1}{\ell^3}\int_{C_\ell} f(z)e^{-ik\cdot z}\,dz\right|^2\int_{\mathbb{R}^3}\frac{|\hat\rho(p)|^2}{|p-k|^2}\,dp.\end{aligned}$$

Hence we obtain

$$\frac{1}{\ell^3}\int_{C_\ell} D(\rho_{\Gamma_\tau}-\rho)\,d\tau = 2\pi\sum_{k\in2\pi\mathbb{Z}^3}\left|\int_{C_1} f_{\delta/\ell}(z)e^{-ik\cdot z}\,dz\right|^2\int_{\mathbb{R}^3}\frac{|\hat\rho(p)|^2}{|p-k/\ell|^2}\,dp,$$

with

$$f_\varepsilon(x) = \sum_{j=1}^{24}\bigg(\mathbb{1}_{\mu_j\mathbf{\Delta}} - \frac{1}{(1-\varepsilon)^3}\mathbb{1}_{\mu_j(1-\varepsilon)\mathbf{\Delta}} * \eta_\varepsilon\bigg) \tag{39}$$

for $\varepsilon=\delta/\ell$. We have

$$\int_{C_1} f_{\delta/\ell}(z)\,dz = 0.$$

Since all the functions appearing in the sum on the right side of (39) are supported in the unit cube, we also obtain, for $k\in2\pi\mathbb{Z}^3\setminus\{0\}$,

$$\int_{C_1} f_{\delta/\ell}(z)e^{-ik\cdot z}\,dz = -\frac{(2\pi)^3}{(1-\varepsilon)^3}\sum_{j=1}^{24}\widehat{\mathbb{1}_{\mu_j(1-\varepsilon)\mathbf{\Delta}}}(k)\hat\eta_1(\varepsilon k). \tag{40}$$

This results in the final formula for the error term

$$\frac{1}{\ell^3}\int_{C_\ell} D(\rho_{\Gamma_\tau}-\rho)\,d\tau = (2\pi)^7\sum_{\substack{k\in2\pi\mathbb{Z}^3\\ k\neq0}}|\hat\eta_1(\varepsilon k)|^2\left|\frac{1}{(1-\varepsilon)^3}\sum_{j=1}^{24}\widehat{\mathbb{1}_{\mu_j(1-\varepsilon)\mathbf{\Delta}}}(k)\right|^2\int_{\mathbb{R}^3}\frac{|\hat\rho(p)|^2}{|p-k/\ell|^2}\,dp. \tag{41}$$

In order to control the denominator $|p-k/\ell|^2$, we are going to average our calculation over all the rotations of the tiling. We also replace ℓ and δ by, respectively, $t\ell$ and $t\delta$ and we average over $t\in(1-\alpha,1+\alpha)$ with a weight t^{-4}. Rotating the tiling is the same as rotating ρ. In addition, $\varepsilon=\delta/\ell$ is

independent of t. Hence we are left with estimating

$$\left(\int_{1-\alpha}^{1+\alpha}\frac{dt}{t^4}\right)^{-1}\int_{1-\alpha}^{1+\alpha}\frac{dt}{t^4}\int_{\mathrm{SO}(3)}dR\frac{1}{|p-Rk/(t\ell)|^2}=\frac{\ell^2}{4\pi|k|^2}\frac{3(1-\alpha^2)^3}{2\alpha(3+\alpha^2)}\int_{\frac{1}{1+\alpha}}^{\frac{1}{1-\alpha}}r^2\,dr\int_{S^2}d\omega\frac{1}{|p'-\omega r|^2}$$
$$=\frac{\ell^2}{4\pi|k|^2}\frac{3(1-\alpha^2)^3}{2\alpha(3+\alpha^2)}\mathbb{1}_{A_\alpha}*\frac{1}{|\cdot|^2}(p'),$$

with $p'=p\ell/|k|$ and where A_α is the annulus

$$A_\alpha=\left\{\frac{1}{1+\alpha}\leqslant|x|\leqslant\frac{1}{1-\alpha}\right\}.$$

We will use the following estimate:

Lemma 14. *We have*

$$\left\|\mathbb{1}_{A_\alpha}*\frac{1}{|\cdot|^2}\right\|_{L^\infty}\leqslant C\alpha\log(\alpha^{-1}) \tag{42}$$

for all $\alpha\leqslant\frac{1}{2}$.

The proof of (42) is a simple computation which is provided at the very end of the proof. Using (42) we obtain

$$\left(\int_{1-\alpha}^{1+\alpha}\frac{dt}{t^4}\right)^{-1}\int_{1-\alpha}^{1+\alpha}\frac{dt}{t^4}\int_{\mathrm{SO}(3)}dR\frac{1}{(t\ell)^3}\int_{C_{t\ell}}D(\rho_{\Gamma_\tau,R}-\rho)\,d\tau$$
$$\leqslant C\log(\alpha^{-1})\left(\sum_{\substack{k\in2\pi\mathbb{Z}^3\\k\neq0}}\frac{\ell^2}{|k|^2}\left|\hat\eta_1(\varepsilon k)\right|^2\left|\frac{1}{(1-\varepsilon)^3}\sum_{j=1}^{24}\widehat{\mathbb{1}_{\mu_j(1-\varepsilon)\mathbf{\Delta}}}(k)\right|^2\right)\int_{\mathbb{R}^3}\rho^2, \tag{43}$$

and it remains to estimate the sum in the parenthesis. For this we have to bound $\sum_{j=1}^{24}\widehat{\mathbb{1}_{\mu_j(1-\varepsilon)\mathbf{\Delta}}}(k)$.

Lemma 15 (Fourier transform of the reduced tetrahedra). *We have*

$$\left|\frac{1}{(1-\varepsilon)^3}\sum_{j=1}^{24}\widehat{\mathbb{1}_{\mu_j(1-\varepsilon)\mathbf{\Delta}}}(k)\right|^2\leqslant C\left(\varepsilon^4+\varepsilon^2|k|^2\left|\int_{C_1}\left(x-\sum_{j=1}^{24}z_j\mathbb{1}_{\mathbf{\Delta}_j}\right)e^{-ik\cdot x}\,dx\right|^2\right) \tag{44}$$

for all $0<\varepsilon<\frac{1}{2}$ *and all* $k\in2\pi\mathbb{Z}^3\setminus\{0\}$.

Proof. We recall that $\mathbf{\Delta}_j=R_j\mathbf{\Delta}_1-z_j$ with $\mu_j=(z_j,R_j)\in C_1\times\mathrm{SO}(3)$. We have

$$(1-\varepsilon)^{-3}\widehat{\mathbb{1}_{\mu_j(1-\varepsilon)\mathbf{\Delta}}}(k)=(2\pi)^{-\frac{3}{2}}(1-\varepsilon)^{-3}\int_{\mu_j(1-\varepsilon)\mathbf{\Delta}}e^{-ik\cdot x}\,dx$$
$$=(2\pi)^{-\frac{3}{2}}\int_{\mathbf{\Delta}}e^{-ik\cdot(R_j(1-\varepsilon)x-z_j)}\,dx$$
$$=(2\pi)^{-\frac{3}{2}}\int_{\mu_j\mathbf{\Delta}}e^{-ik\cdot x+i\varepsilon k\cdot(x-z_j)}\,dx.$$

Since $k \in 2\pi\mathbb{Z}^3 \setminus \{0\}$, the integral vanishes at $\varepsilon = 0$ after summing over j. Inserting the derivative at $\varepsilon = 0$ yields

$$(1-\varepsilon)^{-3} \sum_{j=1}^{24} \widehat{\mathbb{1}_{\mu_j}}_{(1-\varepsilon)\mathbf{\Delta}}(k) = i(2\pi)^{-\frac{3}{2}} \varepsilon k \cdot \int_{C_1} \left(x - \sum_{j=1}^{24} z_j \mathbb{1}_{\mathbf{\Delta}_j} \right) e^{-ik\cdot x}\, dx$$
$$- \varepsilon^2 \sum_{j=1}^{24} \int_0^1 (1-s)\, ds \int_{\mathbf{\Delta}_j} (k\cdot(x-z_j))^2 e^{-ik\cdot x + i\varepsilon s k\cdot(x-z_j)}\, dx.$$

We claim the second term is uniformly bounded with respect to k. Indeed, one integration by parts gives

$$\int_0^1 (1-s)\, ds \int_{\mathbf{\Delta}_j} (k\cdot(x-z_j))^2 e^{-ik\cdot x + i\varepsilon s k\cdot(x-z_j)}\, dx$$
$$= i \int_0^1 \frac{1-s}{1-\varepsilon s}\, ds \int_{\mathbf{\Delta}_j} k\cdot(x-z_j)\,(x-z_j)\cdot \nabla_x e^{-ik\cdot x + i\varepsilon s k\cdot(x-z_j)}\, dx$$
$$= i \int_0^1 \frac{1-s}{1-\varepsilon s}\, ds \int_{\partial\mathbf{\Delta}_j} k\cdot(x-z_j)\,(x-z_j)\cdot n_j(x) e^{-ik\cdot x + i\varepsilon s k\cdot(x-z_j)}\, dx$$
$$- 4i \int_0^1 \frac{1-s}{1-\varepsilon s}\, ds \int_{\mathbf{\Delta}_j} k\cdot(x-z_j) e^{-ik\cdot x + i\varepsilon s k\cdot(x-z_j)}\, dx. \quad (45)$$

Here $n_j(x)$ is the normalized vector perpendicular to $\partial\mathbf{\Delta}_j$ pointing outwards. Integrating once more in the same manner (involving the edges of the faces of $\partial\mathbf{\Delta}_j$ for the first term), we see that (45) is bounded uniformly in k; hence we obtain (44). □

Inserting (44) in (43), we obtain the two error terms

$$\varepsilon^2 \sum_{k\in 2\pi\mathbb{Z}^3 k\neq 0} |\hat{\eta}_1(\varepsilon k)|^2 \left| \int_{C_1} \left(x - \sum_{j=1}^{24} z_j \mathbb{1}_{\mathbf{\Delta}_j} \right) e^{-ik\cdot x}\, dx \right|^2 \leqslant \varepsilon^2 (2\pi)^3 \left\| x - \sum_{j=1}^{24} z_j \mathbb{1}_{\mathbf{\Delta}_j} \right\|^2_{L^2(C_1)} = C\varepsilon^2$$

(using here that $\|\hat{\eta}_1\|_{L^\infty} \leqslant (2\pi)^{3/2}$) and

$$\varepsilon^4 \sum_{\substack{k\in 2\pi\mathbb{Z}^3 \\ k\neq 0}} \frac{|\hat{\eta}_1(\varepsilon k)|^2}{|k|^2} \underset{\varepsilon\to 0}{\sim} \varepsilon^3 \int_{\mathbb{R}^3} \frac{|\hat{\eta}_1(k)|^2}{|k|^2}\, dk.$$

Recalling that $\varepsilon = \delta/\ell$, our final estimate on the averaged error is proportional to

$$\delta^2 \log(\alpha^{-1}) \left(1 + \frac{\delta}{\ell} \right) \int_{\mathbb{R}^3} \rho^2.$$

In order to conclude the proof of Proposition 13, it remains to provide the following:

Proof of Lemma 14. We have

$$\mathbb{1}_{A_\alpha} * \frac{1}{|\cdot|^2}(x) = 2\pi \int_{\frac{1}{1+\alpha}}^{\frac{1}{1-\alpha}} r^2\, dr \int_0^\pi \sin(\varphi)\, d\varphi \frac{1}{r^2 + |x|^2 - 2r|x|\cos\varphi}$$
$$= \frac{\pi}{|x|} \int_{\frac{1}{1+\alpha}}^{\frac{1}{1-\alpha}} \log\left(\frac{r+|x|}{|r-|x||} \right) r\, dr = \pi|x| \int_{\frac{1}{|x|(1+\alpha)}}^{\frac{1}{|x|(1-\alpha)}} \log\left(\frac{r+1}{|r-1|} \right) r\, dr.$$

For $|x| \leqslant \frac{1}{5}$ and $0 < \alpha < \frac{1}{2}$, the integrand is bounded on the corresponding interval and we obtain

$$\mathbb{1}_{A_\alpha} * \frac{1}{|\cdot|^2}(x) \leqslant C\alpha.$$

Similarly, for $|x| \geqslant 4$ and $0 < \alpha < \frac{1}{2}$ the integrand can be estimated by r^2, which gives again

$$\mathbb{1}_{A_\alpha} * \frac{1}{|\cdot|^2}(x) \leqslant C\,\frac{\alpha}{|x|^2} \leqslant C\alpha.$$

Finally, for $\frac{1}{5} \leqslant |x| \leqslant 4$, we have

$$\mathbb{1}_{A_\alpha} * \frac{1}{|\cdot|^2}(x) \leqslant C\alpha + C\int_{\frac{1}{|x|(1+\alpha)}}^{\frac{1}{|x|(1-\alpha)}} |\log|r-1||\,dr.$$

The last integral is over an interval of length

$$\frac{1}{|x|(1-\alpha)} - \frac{1}{|x|(1+\alpha)} = \frac{2\alpha}{|x|(1-\alpha^2)} \leqslant \frac{40}{3}\alpha.$$

The integral is maximum when the interval is placed at the divergence point $r = 1$. So we have

$$\int_{\frac{1}{|x|(1+\alpha)}}^{\frac{1}{|x|(1-\alpha)}} |\log|r-1||\,dr \leqslant \int_{\max\left(0,1-\frac{40}{3}\alpha\right)}^{1+\frac{40}{3}\alpha} |\log|r-1||\,dr \leqslant C\alpha\log(\alpha^{-1}). \qquad \square$$

This concludes the proof of Proposition 13. $\square$

4.2. ***Lower bound in terms of local densities.*** Next we turn to the lower bound. We are going to use the same tiling made of tetrahedra, with the difference that we do not insert any hole. Similarly to (32), we introduce

$$\xi_{\ell,\delta,j} := \mathbb{1}_{\ell\mu_j\mathbf{\Delta}} * \eta_\delta, \tag{46}$$

which forms a smooth partition of unity, without holes,

$$\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24} \xi_{\ell,\delta,j}(x-\ell z) = 1.$$

Proposition 16 (lower bound in terms of local densities). *There exists a universal constant C such that for any $\sqrt{\rho} \in H^1(\mathbb{R}^3)$ and any $\delta > 0$ with $0 < \delta/\ell < 1/C$, we have*

$$E(\rho) \geqslant \frac{1-C\delta/\ell}{\ell^3}\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}\int_{\mathrm{SO}(3)}\int_{C_\ell} E(\xi_{\ell,\delta,j}(R\cdot-\ell z-\tau)\rho)\,dR\,d\tau - \frac{C}{\ell}\int_{\mathbb{R}^3}((1+\delta^{-1})\rho+\delta^3\rho^2). \tag{47}$$

In particular, we can find an isometry $(\tau, R) \in \mathbb{R}^3 \times \mathrm{SO}(3)$ such that

$$E(\rho) \geqslant \left(1-\frac{C\delta}{\ell}\right)\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24} E(\xi_{\ell,\delta,j}(R\cdot-\ell z-\tau)\rho) - \frac{C}{\ell}\int_{\mathbb{R}^3}((1+\delta^{-1})\rho+\delta^3\rho^2). \tag{48}$$

Proof. For a state $\Gamma = \bigoplus_{n\geqslant 0} \Gamma_n$ on Fock space (commuting with the particle number operator) and an interaction potential w, we introduce the simplified notation

$$\mathcal{C}_w(\Gamma) := \sum_{n\geqslant 2} \mathrm{Tr}_{\mathfrak{H}^n}\Big(\sum_{1\leqslant j<k\leqslant n} w(x_j - x_k)\Big)\Gamma_n, \tag{49}$$

$$D_w(\rho) := \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} w(x-y)\rho(x)\rho(y)\,dx\,dy. \tag{50}$$

For the Coulomb potential $w(x) = |x|^{-1}$ we simply use the notation $\mathcal{C}(\Gamma)$ and $D(\rho)$. For the kinetic energy, we write

$$\mathcal{T}(\Gamma) := \sum_{n\geqslant 1} \mathrm{Tr}_{\mathfrak{H}^n}\Big(-\sum_{j=1}^n \Delta_{x_j}\Big)\Gamma_n \tag{51}$$

and finally denote by

$$\mathcal{E}(\Gamma) := \mathcal{T}(\Gamma) + \mathcal{C}(\Gamma) - D(\rho_\Gamma)$$

the total energy, with the direct term subtracted.

The proof uses the well-known fact that, for any interaction potential w and any state Γ on Fock space, we have

$$\mathcal{C}_w(\rho) - D_w(\rho) \geqslant \begin{cases} -\frac{1}{2}w(0)\int_{\mathbb{R}^3}\rho_\Gamma & \text{when } \hat{w}\in L^1(\mathbb{R}^d) \text{ and } \hat{w}\geqslant 0,\\ -\frac{1}{2}\big(\int_{\mathbb{R}^3} w\big)\int_{\mathbb{R}^3}(\rho_\Gamma)^2 & \text{when } w\in L^1(\mathbb{R}^d) \text{ and } w\geqslant 0.\end{cases} \tag{52}$$

For the first bound see for instance [Lewin et al. 2018, equation (4.8)]. The second bound uses only that $\mathcal{C}_w(\Gamma)\geqslant 0$ and $D_w(\rho)\leqslant \frac{1}{2}\|w\|_{L^1}\|\rho\|_{L^2}^2$, by Young's inequality.

We are now ready to prove (48). The smeared Graf–Schenker inequality from [1995, Lemma 6] states that the potential

$$\tilde{w}_\ell(x) = \frac{1}{|x|} - \left(1-\frac{\kappa\delta}{\ell}\right)\frac{\tilde{h}_{\ell,\delta}(x)}{|x|} - \frac{\kappa\delta^2}{\ell}\frac{\tilde{h}_{\ell,\delta}(0)}{|x|(\delta+|x|)}$$

has a positive Fourier transform for all $\ell > \kappa\delta$, with $\tilde{w}_\ell(0) = -\kappa\ell^{-1}\tilde{h}_{\ell,\delta}(0)$, where

$$\begin{aligned}\tilde{h}_{\ell,\delta}(x-y) &= \frac{1}{|\ell\mathbf{\Delta}|}\int_{\mathrm{SO}(3)} (\mathbb{1}_{\ell\mathbf{\Delta}} * \eta_\delta) * (\mathbb{1}_{-\ell\mathbf{\Delta}} * \eta_\delta)(Rx - Ry)\,dR\\ &= \frac{1}{|\ell\mathbf{\Delta}|}\int_{\mathrm{SO}(3)}\int_{\mathbb{R}^3} (\mathbb{1}_{R^{-1}\ell\mathbf{\Delta}+z} * \eta_\delta)(y)(\mathbb{1}_{R^{-1}\ell\mathbf{\Delta}+z} * \eta_\delta)(x)\,dz\,dR.\end{aligned}$$

Here $\mathbf{\Delta}$ is a tetrahedron and $\kappa > 0$ is a large enough constant. In addition, we have from [Lewin et al. 2018, Proof of Lemma 5.5] that the potential

$$\widetilde{W}_\delta(x) = \frac{1}{|x|(\delta+|x|)} - \frac{e^{-\frac{\sqrt{2}}{\delta}|x|}}{\delta|x|}$$

is positive and has positive Fourier transform, with $\widetilde{W}_\delta(0) = \sqrt{2}\delta^{-2}$.

Arguing exactly as in [Lewin et al. 2018, Lemma 5.5] using (52), we find that for any fermionic grand-canonical mixed state $\Gamma = \bigoplus_{n\geqslant 0} \Gamma_n$ with density ρ_Γ we have

$$\begin{aligned}\mathcal{C}(\Gamma)-D(\rho_\Gamma)&\\ \geqslant \frac{1-\kappa\delta/\ell}{\ell^3}\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}\int_{\mathrm{SO}(3)}\int_{C_\ell}&\left\{\mathcal{C}(\Gamma_{|\sqrt{\xi_{\ell,\delta,j}(R\cdot-\ell z-\tau)}})-D(\xi_{\ell,\delta,j}(R\cdot-\ell z-\tau)\rho_\Gamma)\right\}dR\,d\tau\\ &-\frac{C}{\ell}\int_{\mathbb{R}^3}\rho_\Gamma-\frac{C\delta^3}{\ell}\int_{\mathbb{R}^3}(\rho_\Gamma)^2. \quad (53)\end{aligned}$$

Here $\Gamma_{|f}$ is the geometrically f-localized state on Fock space [Dereziński and Gérard 1999; Hainzl et al. 2009; Lewin 2011], that is, the unique state which has the k-particle reduced density matrices $f^{\otimes k}\Gamma^{(k)}f^{\otimes k}$. The last term proportional to δ^3/ℓ comes from the L^1 norm of $\ell^{-1}\delta e^{-(\sqrt{2}/\delta)|x|}|x|^{-1}$.

For the kinetic energy we use the IMS formula as in [Graf and Schenker 1995; Hainzl et al. 2009] and [Lewin et al. 2018, Lemma 5.6], which yields an error in the form

$$\frac{N}{\ell^3}\int_{\mathbb{R}^3}|\nabla\sqrt{\xi_{\ell,\delta,j}}|^2 = O\left(\frac{N}{\ell\delta}\right),$$

where $N=\int_{R^3}\rho_\Gamma$. For the total energy we obtain

$$\begin{aligned}\mathcal{E}(\Gamma)\geqslant\frac{1-\kappa\delta/\ell}{\ell^3}\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}\int_{\mathrm{SO}(3)}\int_{C_\ell}&\mathcal{E}(\Gamma_{|\sqrt{\xi_{\ell,\delta,j}(R\cdot-\ell z-\tau)}})\,dR\,d\tau\\ &-\frac{C}{\ell}\left(1+\frac{1}{\delta}\right)\int_{\mathbb{R}^3}\rho_\Gamma-\frac{C\delta^3}{\ell}\int_{\mathbb{R}^3}(\rho_\Gamma)^2, \quad (54)\end{aligned}$$

which yields the result. □

4.3. ***A convergence rate for tetrahedra.*** In this section we study the convergence of the energy per unit volume for tetrahedra and find a convergence rate. We introduce the energy per unit volume of a tetrahedron at constant density $\rho_0>0$

$$e_{\mathbf{\Delta}}(\rho_0,\ell,\delta) := |\ell\mathbf{\Delta}|^{-1}E(\rho_0\,\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta), \quad (55)$$

where $\eta_\delta(x)=(10/\delta)^3\eta_1(10x/\delta)$ with η_1 a fixed C_c^∞ nonnegative radial function with support in the unit ball and such that $\int_{\mathbb{R}^3}\eta_1=1$. We prove the following:

Proposition 17 (thermodynamic limit for tetrahedra). *For every fixed $\rho_0>0$, we have*

$$\lim_{\substack{\delta/\ell\to 0\\ \delta^3/\ell\to 0\\ \ell\delta\to\infty}} e_{\mathbf{\Delta}}(\rho_0,\ell,\delta) = e_{\mathrm{UEG}}(\rho_0). \quad (56)$$

For $\delta\leqslant\ell/C$ and $0<\alpha<\frac{1}{2}$, we have the upper bound

$$e_{\mathbf{\Delta}}(\rho_0,\ell,\delta)\leqslant e_{\mathrm{UEG}}(\rho_0)+C\frac{\rho_0}{\ell}(1+\delta^{-1}+\delta^3\rho_0+\delta\rho_0^{\frac{2}{3}}) \quad (57)$$

and the averaged lower bound

$$\left(\int_{1-\alpha}^{1+\alpha}\frac{ds}{s^4}\right)^{-1}\int_{1-\alpha}^{1+\alpha}e_{\mathbf{\Delta}}(\rho_0,t\ell,t\delta)\,\frac{dt}{t^4}\geqslant e_{\mathrm{UEG}}(\rho_0)-C\delta^2\rho_0^2\log(\alpha^{-1}). \tag{58}$$

If in addition $\rho_0^{1/3}\ell\geqslant C$, *we have the pointwise lower bound*

$$e_{\mathbf{\Delta}}(\rho_0,\ell,\delta)\geqslant e_{\mathrm{UEG}}(\rho_0)-C\delta\frac{\rho_0^{\frac{5}{3}}+\rho_0^{\frac{4}{3}}}{\ell}-C\frac{\rho_0^{\frac{23}{15}}+\rho_0^{\frac{18}{15}}}{\ell^{\frac{2}{5}}}. \tag{59}$$

The constant C *only depends on the chosen regularizing function* η_1. *It is independent of* ρ_0, ℓ, δ, α.

We will later see that the condition $\delta^3/\ell\to 0$ is actually not needed in the limit (56). It is an interesting problem to replace the error term in the lower bound (59) by an error similar to the upper bound (57). Note that the error term in (58) goes to zero only when $\delta\to 0$, whereas (59) does not require $\delta\to 0$.

Proof. For fixed $\rho_0>0$ and $\delta>0$, the existence of the limit (56) for $\ell\to\infty$ was proved in [Lewin et al. 2018], using a lower bound similar to (48).

We consider a large tetrahedron $\ell'\mathbf{\Delta}$, smeared at a scale δ' and a tiling of smaller tetrahedra of size $\ell\ll\ell'$, smeared at scale δ. Applying our lower bound (47), we find

$$\begin{aligned}&e_{\mathbf{\Delta}}(\rho_0,\ell',\delta')\\&\quad\geqslant\frac{1-C\frac{\delta}{\ell}}{|\ell'\mathbf{\Delta}|\ell^3}\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}\int_{C_\ell}\int_{\mathrm{SO}(3)}E(\rho_0\,(\mathbb{1}_{R(\ell\mu_j\mathbf{\Delta}-\ell z-\tau)}*\eta_\delta)\,(\mathbb{1}_{\ell'\mathbf{\Delta}}*\eta_{\delta'}))\,dR\,d\tau-\frac{C\rho_0}{\ell}(1+\delta^{-1}+\delta^3\rho_0),\end{aligned}$$

where for the error term we have used that

$$\int_{\mathbb{R}^3}(\rho_0\mathbb{1}_{\ell'\mathbf{\Delta}}*\eta_{\delta'})^2=\rho_0^2\int_{\mathbb{R}^3}(\mathbb{1}_{\ell'\mathbf{\Delta}}*\eta_{\delta'})^2\leqslant\rho_0^2\int_{\mathbb{R}^3}\mathbb{1}_{\ell'\mathbf{\Delta}}*\eta_{\delta'}=\rho_0^2|\ell'\mathbf{\Delta}|.$$

For all the tetrahedra such that $R(\ell\mu_j\mathbf{\Delta}-\ell z-\tau)+B_{\delta/10}\subset(\ell'-\delta')\mathbf{\Delta}$, we obtain exactly $|\ell\mathbf{\Delta}|\,e_{\mathbf{\Delta}}(\rho_0,\ell,\delta)$ in the integral. The other tetrahedra are at a distance proportional to $\ell+\delta+\delta'$ from the boundary of $\ell'\mathbf{\Delta}$. Hence, using our lower bound (19) on the energy, they give rise to an error term of the order $\rho_0^{4/3}(\ell+\delta+\delta')/\ell'$. We obtain

$$e_{\mathbf{\Delta}}(\rho_0,\ell',\delta')\geqslant\left(1-C\sigma\frac{\delta}{\ell}-C\sigma\frac{\ell+\delta+\delta'}{\ell'}\right)e_{\mathbf{\Delta}}(\rho_0,\ell,\delta)-C\rho_0^{\frac{4}{3}}\frac{\ell+\delta+\delta'}{\ell'}-\frac{C\rho_0}{\ell}(1+\delta^{-1}+\delta^3\rho_0). \tag{60}$$

Here $\sigma=1$ if $e_{\mathbf{\Delta}}(\rho_0,\ell,\delta)\geqslant 0$ and $\sigma=0$ otherwise.

After taking the limit $\ell'\to\infty$ at fixed ℓ, δ, δ', ρ_0, we obtain

$$e_{\mathrm{UEG}}(\rho_0)\geqslant\left(1-C\sigma\frac{\delta}{\ell}\right)e_{\mathbf{\Delta}}(\rho_0,\ell,\delta)-\frac{C\rho_0}{\ell}(1+\delta^{-1}+\delta^3\rho_0).$$

It follows from our upper bound (29) (see also [Lewin et al. 2018, Remark 5.4]) that

$$e_{\mathrm{UEG}}(\rho_0)\leqslant q^{-\frac{2}{3}}c_{\mathrm{TF}}\,\rho_0^{\frac{5}{3}}$$

for all $\rho_0 > 0$. Hence after dividing by $1 - C\sigma\delta/\ell$ we have shown the claimed upper bound

$$e_{\mathbf{\Delta}}(\rho_0, \ell, \delta) \leqslant e_{\mathrm{UEG}}(\rho_0) + C\frac{\rho_0}{\ell}(1 + \delta^{-1} + \delta^3\rho_0 + \delta\sigma\rho_0^{\frac{2}{3}}).$$

We may use exactly the same argument using our upper bound (34) in place of (48), and we obtain the lower bound (58).

Next we replace ℓ by $t\ell$ and δ by $t\delta$ in our lower bound (60) on the energy of the large simplex of size ℓ', and average over $t \in \left(\frac{1}{2}, \frac{3}{2}\right)$ with the measure t^{-4}. We then insert our lower bound (58) and, after collecting the different error terms, we obtain

$$e_{\mathbf{\Delta}}(\rho_0, \ell', \delta') \geqslant e_{\mathrm{UEG}}(\rho_0) - C\frac{\ell + \delta + \delta'}{\ell'}(\rho_0^{\frac{5}{3}} + \rho_0^{\frac{4}{3}}) - C\frac{\rho_0}{\ell}(1 + \delta^{-1} + \delta^3\rho_0 + \delta\rho_0^{\frac{2}{3}}) - C\delta^2\rho_0^2.$$

It is natural to choose $\delta = \ell^{-1/3}(\rho_0)^{-4/9}$, which provides the estimate

$$e_{\mathbf{\Delta}}(\rho_0, \ell', \delta') \geqslant e_{\mathrm{UEG}}(\rho_0) - C\frac{\ell + \ell^{-\frac{1}{3}}(\rho_0)^{-\frac{4}{9}} + \delta'}{\ell'}(\rho_0^{\frac{5}{3}} + \rho_0^{\frac{4}{3}}) - C\frac{\rho_0^{\frac{13}{9}} + \rho_0^{\frac{10}{9}}}{\ell^{\frac{2}{3}}} - C\frac{\rho_0^{\frac{11}{9}}}{\ell^{\frac{4}{3}}} - C\frac{\rho_0^{\frac{2}{3}}}{\ell^2}$$

and then $\ell = (\ell')^{3/5}\rho_0^{-2/15}$, which gives

$$e_{\mathbf{\Delta}}(\rho_0, \ell', \delta') \geqslant e_{\mathrm{UEG}}(\rho_0) - C\frac{\delta'}{\ell'}(\rho_0^{\frac{5}{3}} + \rho_0^{\frac{4}{3}}) - C\frac{\rho_0^{\frac{23}{15}} + \rho_0^{\frac{18}{15}}}{(\ell')^{\frac{2}{5}}}$$

under the assumption that $\ell'(\rho_0)^{\frac{1}{3}} \geqslant C$. This is exactly (59). The two bounds (57) and (59) give the limit (56). □

4.4. ***Proof of Theorem 1.*** Let Ω_N be a sequence of domains as in the statement, that is, such that $|\Omega_N| \to \infty$ and $|\partial\Omega_N + B_r| \leqslant Cr|\Omega_N|^{2/3}$ for all $r \leqslant |\Omega_N|^{1/3}/C$. Assume also that $\delta_N|\Omega_N|^{-1/3} \to 0$.

By following the proof of (60) we see that a similar inequality holds with the large tetrahedron $\ell'\mathbf{\Delta}$ replaced by Ω_N. This gives

$$\frac{E(\rho_0 \mathbb{1}_{\Omega_N} * \eta_{\delta_N})}{|\Omega_N|} \geqslant \left(1 - C\sigma\frac{\delta}{\ell} - C\sigma\frac{\ell + \delta + \delta_N}{|\Omega_N|^{\frac{1}{3}}}\right)e_{\mathbf{\Delta}}(\rho_0, \ell, \delta) - C\rho_0^{\frac{4}{3}}\frac{\ell + \delta + \delta_N}{|\Omega_N|^{\frac{1}{3}}} - \frac{C\rho_0}{\ell}(1 + \delta^{-1} + \delta^3\rho_0).$$

Under the sole condition that $\delta_N|\Omega_N|^{-1/3} \to 0$, the right side tends to $e_{\mathrm{UEG}}(\rho_0)$ if we take for instance δ fixed and $\ell = |\Omega_N|^{1/6}$.

We then use the upper bound (34) with $\alpha = \frac{1}{2}$, as well as the fact that

$$E(\rho_0 \chi_{t\ell, t\delta, j}(R \cdot - t\ell z - \tau)\mathbb{1}_{\Omega_N} * \eta_{\delta_N}) \leqslant C\rho_0\ell^3(\rho_0^{\frac{2}{3}} + (\ell\delta)^{-1}) + C\rho_0 \int_{\mathbb{R}^3} \chi_{t\ell, t\delta, j}(R \cdot - t\ell z - \tau)|\nabla\sqrt{\mathbb{1}_{\Omega_N} * \eta_{\delta_N}}|^2$$

by (29), for the tetrahedra close to the boundary. We find

$$\frac{E(\rho_0 \mathbb{1}_{\Omega_N} * \eta_{\delta_N})}{|\Omega_N|} \leqslant \left(1 + C\sigma\frac{\ell + \delta + \delta_N}{|\Omega_N|^{\frac{1}{3}}}\right)\left(\int_{\frac{1}{2}}^{\frac{3}{2}} \frac{ds}{s^4}\right)^{-1} \int_{\frac{1}{2}}^{\frac{3}{2}} e_{\mathbf{\Delta}}(\rho_0, t\ell, t\delta)\frac{dt}{t^4} + C\frac{\ell + \delta + \delta_N}{|\Omega_N|^{\frac{1}{3}}}\rho_0(\rho_0^{\frac{2}{3}} + (\ell\delta)^{-1}) + \frac{C\rho_0}{\delta_N|\Omega_N|^{\frac{1}{3}}} + C\rho_0^2\delta^2. \quad (61)$$

We have used here that

$$\frac{1}{|\Omega_N|}\int_{\mathbb{R}^3}|\nabla\sqrt{\mathbb{1}_{\Omega_N}*\eta_{\delta_N}}|^2\leqslant\frac{C}{\delta_N|\Omega_N|^{\frac{1}{3}}}.$$

Under the additional assumption that $\delta_N|\Omega_N|^{1/3}\to\infty$, we may choose for instance $\ell=|\Omega_N|^{1/6}$ and $\delta=|\Omega_N|^{-1/12}$, which yields the result. □

4.5. ***Replacing the local density by a constant density.*** The goal of this section is to provide estimates on the variation of the energy in a (smeared) tetrahedron, when we replace the local density by a constant, chosen to be either the minimum or the maximum of the density in the tetrahedron.

Proposition 18 (replacing ρ by a constant locally). *Let $p>3$ and $0<\theta<1$ such that*

$$2\leqslant p\theta\leqslant 1+\tfrac{1}{2}p. \tag{62}$$

There exists a constant $C=C(p,\theta,q)$ such that, for $\ell\geqslant C$ and $\delta\leqslant\ell/C$, we have

$$\begin{aligned}E(\rho\,(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta))\leqslant E(\underline{\rho}\,(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta))+C\varepsilon\int_{\mathbb{R}^3}(\rho+\rho^2)(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)+C\int_{\mathbb{R}^3}\rho|\nabla\sqrt{\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta}|^2\\ +\frac{C}{\varepsilon}\int_{\mathbb{R}^3}|\nabla\sqrt{\rho}|^2(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)+C\left(\frac{\ell^{2p}}{\varepsilon^{p-1}}+\frac{\ell^p}{\varepsilon^{\frac{5}{4}p-1}}\right)\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p\end{aligned} \tag{63}$$

and

$$\begin{aligned}E(\rho\,(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta))\geqslant E(\bar{\rho}\,(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta))-C\varepsilon\ell^3(\bar{\rho}+\bar{\rho}^2)-\frac{C\ell^2}{\delta}\bar{\rho}\\ -\frac{C}{\varepsilon}\int_{\mathbb{R}^3}|\nabla\sqrt{\rho}|^2(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)-C\left(\frac{\ell^{2p}}{\varepsilon^{p-1}}+\frac{\ell^p}{\varepsilon^{\frac{5}{4}p-1}}\right)\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p\end{aligned} \tag{64}$$

for all $0<\varepsilon\leqslant\frac{1}{2}$, where

$$\underline{\rho}=\min_{x\in\mathrm{supp}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)}\rho(x),\quad\bar{\rho}=\max_{x\in\mathrm{supp}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)}\rho(x)$$

are respectively the minimum and maximum value of ρ on the support of $\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta$.

Under the assumption that $\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p$ is finite, the density ρ is continuous on $\overline{\ell\mathbf{\Delta}+B_\delta}$, so that $\underline{\rho}$ and $\bar{\rho}$ are well-defined.

We have already discussed in the beginning of Section 4.1 the difficulty of deriving a subadditivity-type estimate relating $E(\rho_1+\rho_2)$ to $E(\rho_1)$ and $E(\rho_2)$. The following lemma provides a rather rough inequality, which however will be sufficient for our purposes.

Lemma 19 (rough subadditivity estimate). *Let $\rho_1,\rho_2\in L^1(\mathbb{R}^3,\mathbb{R}_+)$ be two densities such that $\sqrt{\rho_1},\sqrt{\rho_2}\in H^1(\mathbb{R}^3)$. Then*

$$E(\rho_1+\rho_2)\leqslant E(\rho_1)+C\varepsilon\int_{\mathbb{R}^3}(\rho_1^{\frac{5}{3}}+\rho_1^{\frac{4}{3}})+C\varepsilon^{-\frac{2}{3}}\int_{\mathbb{R}^3}\rho_2^{\frac{5}{3}}+C\int_{\mathbb{R}^3}|\nabla\sqrt{\rho_2+\varepsilon\rho_1}|^2+\frac{1-\varepsilon}{\varepsilon}D(\rho_2) \tag{65}$$

for all $0<\varepsilon\leqslant 1$.

Here we have in mind that ρ_2 is small compared to ρ_1 and we estimate $E(\rho_1+\rho_2)$ in terms of $E(\rho_1)$ plus some error terms. The worse error in the estimate (65) is $D(\rho_2)/\varepsilon$, because it grows much faster than the volume. Later we will only use (65) locally and this bad term will not be too large. But it will be responsible for the large power of ε in front of the gradient correction in our main estimate (8). We conjecture that there is an inequality similar to (65) without the term $D(\rho_2)/\varepsilon$.

Note that we can estimate

$$\int_{\mathbb{R}^3} |\nabla\sqrt{\rho_2+\varepsilon\rho_1}|^2 \leqslant \int_{\mathbb{R}^3} |\nabla\sqrt{\rho_2}|^2 + \varepsilon\int_{\mathbb{R}^3} |\nabla\sqrt{\rho_1}|^2$$

by the convexity of $\rho\mapsto|\nabla\sqrt{\rho}|^2$.

Proof of Lemma 19. Fix an $\varepsilon\in(0,1]$ and consider two optimal states Γ_1 and Γ_2 in Fock space, for ρ_1 and $\rho_2/\varepsilon+\rho_1$, respectively. Then

$$\Gamma := (1-\varepsilon)\Gamma_1 + \varepsilon\Gamma_2$$

is a proper quantum state which has the density

$$\rho_\Gamma = (1-\varepsilon)\rho_1 + \varepsilon\left(\frac{\rho_2}{\varepsilon}+\rho_1\right) = \rho_1+\rho_2.$$

Inserting this trial state and using (29) for $E(\rho_2/\varepsilon+\rho_1)$, we deduce that

$$\begin{aligned} E(\rho_1+\rho_2) \leqslant (1-\varepsilon)E(\rho_1) + \frac{C}{\varepsilon^{\frac{2}{3}}}\int_{\mathbb{R}^3}\rho_2^{\frac{5}{3}} + C\int_{\mathbb{R}^3}|\nabla\sqrt{\rho_2+\varepsilon\rho_1}|^2 \\ + C\varepsilon\int_{\mathbb{R}^3}\rho_1^{\frac{5}{3}} - D(\rho_1+\rho_2) + (1-\varepsilon)D(\rho_1) + \varepsilon D(\rho_1+\rho_2/\varepsilon). \end{aligned}$$

We have

$$-D(\rho_1+\rho_2)+(1-\varepsilon)D(\rho_1)+\varepsilon D(\rho_1+\rho_2/\varepsilon) = \frac{1-\varepsilon}{\varepsilon}D(\rho_2).$$

By the Lieb–Oxford inequality $E(\rho_1)\geqslant -C\int_{\mathbb{R}^3}\rho_1^{4/3}$, and the result follows. □

We are now able to provide the following:

Proof of Proposition 18. We write $\rho=\underline{\rho}+(\rho-\underline{\rho})$ and apply (65). We obtain

$$\begin{aligned} E(\rho\,(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)) \leqslant E(\underline{\rho}\,(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)) + C\varepsilon\int_{\mathbb{R}^3}(\rho^{\frac{5}{3}}+\rho^{\frac{4}{3}})(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta) \\ + \frac{C}{\varepsilon^{\frac{2}{3}}}\int_{\mathbb{R}^3}(\rho-\underline{\rho})^{\frac{5}{3}}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta) + \frac{1}{\varepsilon}D((\rho-\underline{\rho})(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)) \\ + C\int_{\mathbb{R}^3}\left|\nabla\sqrt{(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)(\rho-(1-\varepsilon)\underline{\rho})}\right|^2. \end{aligned}$$

In the first line we have used that $\underline{\rho}\leqslant\rho$ on the support of $\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta$ and that $\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta\leqslant 1$. First we can bound $\rho^{4/3}+\rho^{5/3}$ by $\rho+\rho^2$. Next, using

$$|\nabla\sqrt{fg}|^2 = \frac{|\nabla(fg)|^2}{4fg} \leqslant \frac{f|\nabla g|^2}{2g}+\frac{g|\nabla f|^2}{2f},$$

and $\nabla(\rho-(1-\varepsilon)\underline{\rho})=\nabla\rho=2\sqrt{\rho}\nabla\sqrt{\rho}$, we can bound the gradient term pointwise by

$$\left|\nabla\sqrt{(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)(\rho-(1-\varepsilon)\underline{\rho})}\right|^2\leqslant(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)\frac{2\rho\left|\nabla\sqrt{\rho}\right|^2}{\rho-(1-\varepsilon)\underline{\rho}}+2\rho\left|\nabla\sqrt{\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta}\right|^2.$$

Since $\rho\geqslant\underline{\rho}$, we have $\varepsilon\rho\leqslant\rho-(1-\varepsilon)\underline{\rho}$ and hence

$$\frac{\rho}{\rho-(1-\varepsilon)\underline{\rho}}\leqslant\frac{1}{\varepsilon}.$$

This gives the estimate on the gradient term

$$\int_{\mathbb{R}^3}\left|\nabla\sqrt{(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)(\rho-(1-\varepsilon)\underline{\rho})}\right|^2\leqslant\frac{2}{\varepsilon}\int_{\mathbb{R}^3}|\nabla\sqrt{\rho}|^2(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)+2\int_{\mathbb{R}^3}\rho\left|\nabla\sqrt{\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta}\right|^2.$$

Next we estimate the terms involving $\rho-\underline{\rho}$ in terms of the gradient of ρ^θ. We use the Sobolev inequality in the bounded set $\ell\mathbf{\Delta}+B_\delta$

$$\|u\|^p_{L^\infty(\overline{\ell\mathbf{\Delta}+B_\delta})}\leqslant C\ell^{p-3}\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla u(x)|^p\,dx \tag{66}$$

for $p>3$ and every continuous u which vanishes at least at one point in $\ell\mathbf{\Delta}+B_\delta$ (we always assume $\delta\leqslant\ell/C$ so that $\ell\mathbf{\Delta}+B_\delta$ is included in a ball of radius proportional to ℓ). By the Hardy–Littlewood–Sobolev inequality, this gives

$$\begin{aligned}D((\rho-\underline{\rho})(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta))&\leqslant C\|(\rho-\underline{\rho})(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)\|^2_{L^{6/5}}\\&\leqslant C\|\rho^\theta-\underline{\rho}^\theta\|^2_{L^\infty(\ell\mathbf{\Delta}+B_\delta)}\left(\int_{\mathbb{R}^3}\rho^{\frac{6}{5}(1-\theta)}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)\right)^{\frac{5}{3}}\\&\leqslant C\left(\ell^{2p}\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p\right)^{\frac{2}{p}}\left(\int_{\mathbb{R}^3}\rho^{\frac{2p}{p-2}(1-\theta)}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)\right)^{1-\frac{2}{p}}\\&\leqslant C\varepsilon\left(\frac{\ell^{2p}}{\varepsilon^{p-1}}\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p+\varepsilon\int_{\mathbb{R}^3}\rho^{\frac{2p}{p-2}(1-\theta)}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)\right).\end{aligned} \tag{67}$$

In the second estimate we have used that

$$\rho-\underline{\rho}\leqslant C(\rho^\theta-\underline{\rho}^\theta)\rho^{1-\theta}$$

since $\theta\leqslant1$. In the third estimate we have used Hölder's inequality to obtain an integral to the power $1-2/p$. This yields some power of ℓ which has been taken into account in the first factor. In order to bound $\rho^{(1-\theta)(2p)/(p-2)}$ by $\rho+\rho^2$ we need that

$$1\leqslant\frac{2p}{p-2}(1-\theta)\leqslant2,$$

which is equivalent to our assumption (62).

Similarly, we can bound the other error term as follows:

$$\begin{aligned}\int_{\mathbb{R}^3}(\rho-\underline{\rho})^{\frac{5}{3}}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta) &\leqslant C\,\|\rho^\theta-\underline{\rho}^\theta\|_{L^\infty(\ell\mathbf{\Delta}+B_\delta)}^{\frac{5}{3}a}\int_{\mathbb{R}^3}\rho^{\frac{5}{3}(1-\theta a)}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)\\ &\leqslant C\left(\ell^p\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p\right)^{\frac{5a}{3p}}\left(\int_{\mathbb{R}^3}\rho^{\frac{5p}{3p-5a}(1-\theta a)}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)\right)^{1-\frac{5a}{3p}}\\ &\leqslant C\,\varepsilon^{\frac{2}{3}}\left(\frac{\ell^p}{\varepsilon^{\frac{p}{a}-1}}\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p+\varepsilon\int_{\mathbb{R}^3}\rho^{\frac{5p}{3p-5a}(1-\theta a)}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)\right),\end{aligned}\tag{68}$$

where $0<a\leqslant 1$ is a parameter to be chosen. As before we need the condition

$$1\leqslant\frac{5p}{3p-5a}(1-\theta a)\leqslant 2$$

in order to bound the last term by $\rho+\rho^2$. This is equivalent to

$$2-\frac{p}{5a}\leqslant\theta p\leqslant 1+\frac{2p}{5a},$$

where the left inequality is always satisfied under our assumption (62). If we choose $a=1$ then the upper bound on $p\theta$ is stronger than (62). Hence we rather choose $a=\frac{4}{5}$ and obtain (63).

The argument for (64) is similar. This time we write $\bar\rho=\rho+(\bar\rho-\rho)$ and obtain from (65)

$$\begin{aligned}E(\bar\rho\,(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta))\leqslant E(\rho\,(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta))&+C\varepsilon\int_{\mathbb{R}^3}(\rho^{\frac{5}{3}}+\rho^{\frac{4}{3}})(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)\\ &+\frac{C}{\varepsilon^{\frac{2}{3}}}\int_{\mathbb{R}^3}(\bar\rho-\rho)^{\frac{5}{3}}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)+\frac{1}{\varepsilon}D((\bar\rho-\rho)(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta))\\ &+C\int_{\mathbb{R}^3}\left|\nabla\sqrt{(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)(\bar\rho-(1-\varepsilon)\rho)}\right|^2.\end{aligned}$$

The gradient term can be bounded above by

$$\begin{aligned}\int_{\mathbb{R}^3}\left|\nabla\sqrt{(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)(\bar\rho-(1-\varepsilon)\rho)}\right|^2&\leqslant\frac{2}{\varepsilon}\int_{\mathbb{R}^3}|\nabla\sqrt{\rho}|^2(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)+2\bar\rho\int_{\mathbb{R}^3}\left|\nabla\sqrt{\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta}\right|^2\\ &\leqslant\frac{2}{\varepsilon}\int_{\mathbb{R}^3}|\nabla\sqrt{\rho}|^2(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)+\frac{C\ell^2}{\delta}\bar\rho.\end{aligned}$$

The other terms are estimated as before, using that $\bar\rho-\rho\leqslant\bar\rho$. □

4.6. ***Lipschitz regularity of*** $e_{\mathbf{UEG}}$**.** In this section we prove that the UEG energy e_{UEG} is locally Lipschitz. The main result is the following.

Proposition 20 (Lipschitz regularity of e_{UEG}). *There exists a universal constant C so that*

$$e_{\mathrm{UEG}}(\rho)-C(\rho^{\frac{1}{3}}+\rho^{\frac{2}{3}})\rho'\leqslant e_{\mathrm{UEG}}(\rho-\rho')\leqslant e_{\mathrm{UEG}}(\rho)+C\rho'\rho^{\frac{1}{3}}\tag{69}$$

for every $0\leqslant\rho'\leqslant\rho$. In particular, we have

$$|e_{\mathrm{UEG}}(\rho_1)-e_{\mathrm{UEG}}(\rho_2)|\leqslant C(\max(\rho_1,\rho_2)^{\frac{1}{3}}+\max(\rho_1,\rho_2)^{\frac{2}{3}})|\rho_1-\rho_2|.\tag{70}$$

Proof. By scaling we have

$$E(\alpha\rho(\alpha^{\frac{1}{3}}\cdot)) = \min_{\Gamma}\{\alpha^{\frac{2}{3}}\mathcal{T}(\Gamma) + \alpha^{\frac{1}{3}}(\mathcal{C}(\Gamma) - D(\rho_\Gamma))\} \leqslant \alpha^{\frac{1}{3}}E(\rho)$$

for $\alpha \leqslant 1$. This proves that

$$\frac{E(\alpha\rho_0 \mathbb{1}_{\ell/\alpha^{1/3}\mathbf{\Delta}} * \chi_{\delta/\alpha^{1/3}})}{\ell^3|\mathbf{\Delta}|} = \frac{e_{\mathbf{\Delta}}(\alpha\rho_0, \ell/\alpha^{\frac{1}{3}}, \delta/\alpha^{\frac{1}{3}})}{\alpha} \leqslant \alpha^{\frac{1}{3}}\, e_{\mathbf{\Delta}}(\rho_0, \ell, \delta).$$

Passing to the limit using Proposition 17, we find

$$e_{\mathrm{UEG}}(\alpha\rho_0) \leqslant \alpha^{\frac{4}{3}} e_{\mathrm{UEG}}(\rho_0)$$

for every $0 \leqslant \alpha = 1 - \varepsilon \leqslant 1$; hence

$$e_{\mathrm{UEG}}((1-\varepsilon)\rho_0) \leqslant (1-\varepsilon)^{\frac{4}{3}} e_{\mathrm{UEG}}(\rho_0) \leqslant e_{\mathrm{UEG}}(\rho_0) + C\,\varepsilon e_{\mathrm{UEG}}(\rho_0)_-.$$

Here we have used the notation $x_- = \max(-x, 0)$ for the negative part. Using that $e_{\mathrm{UEG}}(\rho_0) \geqslant -c_{\mathrm{LO}}\rho_0^{4/3}$ by the Lieb–Oxford inequality (with $c_{\mathrm{LO}} \leqslant 1.64$), we obtain

$$e_{\mathrm{UEG}}((1-\varepsilon)\rho_0) \leqslant e_{\mathrm{UEG}}(\rho_0) + C\,\varepsilon\rho_0^{\frac{4}{3}}$$

for all $0 \leqslant \varepsilon \leqslant 1$. This proves the upper bound in (69).

Similarly, we can write (still for $0 \leqslant \alpha \leqslant 1$)

$$\begin{aligned} E(\alpha\rho(\alpha^{\frac{1}{3}}\cdot)) + c_{\mathrm{LO}}\alpha^{\frac{1}{3}}\int_{\mathbb{R}^3}\rho^{\frac{4}{3}} &= \min_{\Gamma}\left\{\alpha^{\frac{2}{3}}\mathcal{T}(\Gamma) + \alpha^{\frac{1}{3}}\left(\mathcal{C}(\Gamma) - D(\rho_\Gamma) + c_{\mathrm{LO}}\int_{\mathbb{R}^3}\rho^{\frac{4}{3}}\right)\right\} \\ &\geqslant \alpha^{\frac{2}{3}}\left(E(\rho) + c_{\mathrm{LO}}\int_{\mathbb{R}^3}\rho^{\frac{4}{3}}\right). \end{aligned}$$

This gives as before

$$e_{\mathrm{UEG}}(\alpha\rho_0) + \alpha^{\frac{4}{3}}c_{\mathrm{LO}}\rho_0^{\frac{4}{3}} \geqslant \alpha^{\frac{5}{3}}(e_{\mathrm{UEG}}(\rho_0) + c_{\mathrm{LO}}\rho_0^{\frac{4}{3}}).$$

Using this time $e_{\mathrm{UEG}}(\rho_0) \leqslant C\rho_0^{5/3}$, we obtain

$$e_{\mathrm{UEG}}((1-\varepsilon)\rho_0) \geqslant e_{\mathrm{UEG}}(\rho_0) - C\,\varepsilon(\rho_0^{\frac{4}{3}} + \rho_0^{\frac{5}{3}})$$

for all $0 \leqslant \varepsilon \leqslant 1$. □

4.7. ***Proof of Theorem 3.*** We have derived all the estimates we need to prove the main inequality (8) in Theorem 3.

Let $\rho \in L^1(\mathbb{R}^3, \mathbb{R}_+) \cap L^2(\mathbb{R}^3, \mathbb{R}_+)$ be any density so that $\nabla\sqrt{\rho} \in L^2(\mathbb{R}^3)$ and $\nabla\rho^\theta \in L^p(\mathbb{R}^3)$. First we recall from our upper bound (29) and the lower bound (19) that

$$|E(\rho)| \leqslant c_{\mathrm{TF}}\, q^{-\frac{2}{3}}(1+\varepsilon)\int_{\mathbb{R}^3}\rho^{\frac{5}{3}} + c_{\mathrm{LO}}\int_{\mathbb{R}^3}\rho^{\frac{4}{3}} + \frac{C(1+\varepsilon)}{\varepsilon}\int_{\mathbb{R}^3}|\nabla\sqrt{\rho}|^2.$$

Similarly, we have

$$|e_{\mathrm{UEG}}(\rho)| \leqslant c_{\mathrm{TF}}\, q^{-\frac{2}{3}}\rho^{\frac{5}{3}} + c_{\mathrm{LO}}\rho^{\frac{4}{3}}. \tag{71}$$

In particular, the inequality (8) is obvious for large ε and we only have to consider small ε.

In our upper bound (34) and our lower bound (47), the worse coefficient involving ℓ and δ in front of $\rho+\rho^2$ is $\delta^2+1/(\ell\delta)$. This suggests to take

$$\delta=\sqrt{\varepsilon},\quad \ell=\varepsilon^{-\frac{3}{2}}, \tag{72}$$

which we do for the rest of the proof. In fact, in our proof we will replace ℓ and δ by $t\ell$ and $t\delta$ and average over $t\in\left[\frac{1}{2},\frac{3}{2}\right]$.

Step 1. Upper bound. Let us first take $\frac{1}{4}\leqslant\varepsilon^{3/2}\ell\leqslant 2$ and $\frac{1}{4}\leqslant\delta\varepsilon^{-1/2}\leqslant 2$ and derive an upper bound on $E(\rho(1_{\ell\mathbf{\Delta}}*\eta_\delta))$. We recall that $\mathbf{\Delta}$ is a tetrahedron of volume $\frac{1}{24}$ as described in Section 4.1 and that $\eta_\delta(x)=(10/\delta)^3\eta_1(10x/\delta)$ with η_1 a fixed C_c^∞ nonnegative radial function with support in the unit ball and such that $\int_{\mathbb{R}^3}\eta_1=1$. We denote by

$$\underline{\rho}:=\min_{\mathrm{supp}(1_{\ell\mathbf{\Delta}}*\eta_\delta)}\rho\quad\text{and}\quad\bar{\rho}:=\min_{\mathrm{supp}(1_{\ell\mathbf{\Delta}}*\eta_\delta)}\rho$$

the maximal and minimal values of ρ on the support of $1_{\ell\mathbf{\Delta}}*\eta_\delta$, as in Proposition 18. We use the upper bound (63) from Proposition 18 which quantifies the error made when replacing $E(\rho(1_{\ell\mathbf{\Delta}}*\eta_\delta))$ by $E(\underline{\rho}(1_{\ell\mathbf{\Delta}}*\eta_\delta))$. For the latter we then use our estimate (57) in Proposition 17 on the energy of a smeared tetrahedron. With our choice (72) of ℓ and δ in terms of ε, this leads to

$$\begin{aligned}E(\rho\,(1_{\ell\mathbf{\Delta}}*\eta_\delta))\leqslant e_{\mathrm{UEG}}(\underline{\rho})\int_{\mathbb{R}^3}1_{\ell\mathbf{\Delta}}*\eta_\delta+C\varepsilon\int_{\mathbb{R}^3}(\rho+\rho^2)(1_{\ell\mathbf{\Delta}}*\eta_\delta)\\ +C\int_{\mathbb{R}^3}\rho\big|\nabla\sqrt{1_{\ell\mathbf{\Delta}}*\eta_\delta}\big|^2+\frac{C}{\varepsilon}\int_{\mathbb{R}^3}|\nabla\sqrt{\rho}|^2(1_{\ell\mathbf{\Delta}}*\eta_\delta)+\frac{C}{\varepsilon^{4p-1}}\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p.\end{aligned} \tag{73}$$

In the first line we have bounded the error terms in (57) by

$$\varepsilon^{\frac{3}{2}}\rho(1+\varepsilon^{-\frac{1}{2}}+\varepsilon^{\frac{3}{2}}\rho+\sqrt{\varepsilon}\rho^{\frac{2}{3}})\leqslant C\varepsilon(\rho+\rho^2)$$

in order to simplify our final bound. In the support of $1_{\ell\mathbf{\Delta}}*\eta_\delta$ we have by (69)

$$e_{\mathrm{UEG}}(\underline{\rho})\leqslant e_{\mathrm{UEG}}(\rho(x))+C(\rho(x)-\underline{\rho})\rho(x)^{\frac{1}{3}};$$

hence

$$e_{\mathrm{UEG}}(\underline{\rho})\int_{\mathbb{R}^3}1_{\ell\mathbf{\Delta}}*\eta_\delta\leqslant\int_{\mathbb{R}^3}e_{\mathrm{UEG}}(\rho)\,(1_{\ell\mathbf{\Delta}}*\eta_\delta)+C\int_{\mathbb{R}^3}(\rho-\underline{\rho})\rho^{\frac{1}{3}}(1_{\ell\mathbf{\Delta}}*\eta_\delta).$$

Similarly as we did for (68), we can bound for $0<a\leqslant 1$

$$\begin{aligned}\int_{\mathbb{R}^3}(\rho-\underline{\rho})\rho^{\frac{1}{3}}(1_{\ell\mathbf{\Delta}}*\eta_\delta)&\leqslant C\,\|\rho^\theta-\underline{\rho}^\theta\|^a_{L^\infty(\ell\mathbf{\Delta}+B_\delta)}\int_{\mathbb{R}^3}\rho^{\frac{4}{3}-\theta a}(1_{\ell\mathbf{\Delta}}*\eta_\delta)\\ &\leqslant C\left(\varepsilon^{-\frac{3}{2}p}\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p\right)^{\frac{a}{p}}\left(\int_{\mathbb{R}^3}\rho^{\frac{4-3\theta a}{3}\frac{p}{p-a}}(1_{\ell\mathbf{\Delta}}*\eta_\delta)\right)^{1-\frac{a}{p}}\\ &\leqslant C\left(\frac{1}{\varepsilon^{\frac{3}{2}p+\frac{p}{a}-1}}\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p+\varepsilon\int_{\mathbb{R}^3}\rho^{\frac{4-3\theta a}{3}\frac{p}{p-a}}(1_{\ell\mathbf{\Delta}}*\eta_\delta)\right).\end{aligned} \tag{74}$$

Again we need

$$1 \leqslant \frac{4-3\theta a}{3}\frac{p}{p-a} \leqslant 2,$$

which is equivalent to

$$2-\frac{2p}{3a} \leqslant \theta p \leqslant 1+\frac{p}{3a},$$

where the left side is automatically satisfied under our main assumption (7). In order to get an error controlled by the other gradient terms, we need

$$\frac{3}{2}p+\frac{p}{a}-1 \leqslant 4p-1,$$

which requires $a \geqslant \frac{2}{5}$. Taking $a=\frac{2}{3}$ provides the smallest power of ε. Collecting our estimates, we have proved the following upper bound on the energy in a tetrahedron:

$$E(\rho(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)) \leqslant \int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho(x))\,(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)\,dx + C\varepsilon\int_{\mathbb{R}^3}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)(\rho+\rho^2) + \frac{C}{\varepsilon^{4p-1}}\int_{\ell\mathbf{\Delta}+B_\delta}|\nabla\rho^\theta|^p + \frac{C}{\varepsilon}\int_{\mathbb{R}^3}(\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta)|\nabla\sqrt{\rho}|^2 + C\int_{\mathbb{R}^3}\rho\big|\nabla\sqrt{\mathbb{1}_{\ell\mathbf{\Delta}}*\eta_\delta}\big|^2. \quad (75)$$

Here we have considered a tetrahedron placed at the origin for simplicity, but we of course get a similar inequality for any tetrahedron, by translating and rotating ρ.

Next we recall our upper bound (34) on the total energy $E(\rho)$

$$E(\rho) \leqslant \left(\int_{\frac{1}{2}}^{\frac{3}{2}}\frac{ds}{s^4}\right)^{-1}\int_{\frac{1}{2}}^{\frac{3}{2}}\frac{dt}{t^4}\int_{\mathrm{SO}(3)}dR\int_{C_{t\ell}}\frac{d\tau}{(t\ell)^3}\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}E(\chi_{t\ell,t\delta,j}(R\cdot-t\ell z-\tau)\rho)+C\varepsilon\int_{\mathbb{R}^3}\rho^2, \quad (76)$$

with $\chi_{\ell,\delta,j} := (1-\varepsilon^2)^{-3}\mathbb{1}_{\ell\mu_j(1-\varepsilon^2)\mathbf{\Delta}} * \eta_\delta$. We also recall from Section 4.1 that $\delta/\ell = (t\delta)/(t\ell) = \varepsilon^2$. Inserting (75) into (76) and using the fact (33) that $\chi_{t\ell,t\delta,j}$ forms a partition of unity after averaging over translations and rotations, we obtain

$$E(\rho) \leqslant (1-\varepsilon^2)^3\int_{\mathbb{R}^3} e_{\mathrm{UEG}}((1-\varepsilon^2)^{-3}\rho(x))\,dx + C\varepsilon\int_{\mathbb{R}^3}(\rho+\rho^2) + \frac{C}{\varepsilon}\int_{\mathbb{R}^3}|\nabla\sqrt{\rho}|^2 + \frac{C}{\varepsilon^{4p-1}}\int_{\mathbb{R}^3}|\nabla\rho^\theta|^p. \quad (77)$$

Note that when we sum over the tiling, the sets $t\ell\mu_j\mathbf{\Delta}+B_{t\delta}$ have finitely many intersections, which just results in a bigger constant in front of $|\nabla\rho^\theta|^p$. We have also used that

$$\int_{\mathrm{SO}(3)}dR\int_{C_{t\ell}}\frac{d\tau}{(t\ell)^3}\sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24}\int_{\mathbb{R}^3}\rho\big|\nabla\sqrt{\chi_{t\ell,t\delta,j}(R\cdot-\ell z-\tau)}\big|^2 = \frac{24}{(t\ell)^3}\int_{\mathbb{R}^3}\rho\int_{\mathbb{R}^3}\big|\nabla\sqrt{\chi_{t\ell,t\delta,j}(R\cdot-\ell z-\tau)}\big|^2 \leqslant \frac{C}{\ell\delta}\int_{\mathbb{R}^3}\rho = C\varepsilon\int_{\mathbb{R}^3}\rho.$$

From Proposition 20 and (71), we have

$$(1-\varepsilon^2)^3\int_{\mathbb{R}^3}e_{\mathrm{UEG}}((1-\varepsilon^2)^{-3}\rho) \leqslant \int_{\mathbb{R}^3}e_{\mathrm{UEG}}(\rho)+C\varepsilon^2\int_{\mathbb{R}^3}(\rho^{\frac{4}{3}}+\rho^{\frac{5}{3}});$$

hence we obtain the desired upper bound

$$E(\rho) \leqslant \int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho) + \varepsilon \int_{\mathbb{R}^3} (\rho + \rho^2) + \frac{C}{\varepsilon} \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2 + \frac{C}{\varepsilon^{4p-1}} \int_{\mathbb{R}^3} |\nabla \rho^\theta|^p \tag{78}$$

for ε small enough.

Step 2. Lower bound. The lower bound is slightly more tedious since all our lower estimates involve $\bar\rho$ which can in general not be bounded by ρ. We shall argue as follows. First we average our lower bound (47) over t. This gives

$$E(\rho) \geqslant \left(\int_{\frac12}^{\frac32} \frac{ds}{s^4}\right)^{-1} \int_{\frac12}^{\frac32} \frac{dt}{t^4} \frac{1 - C\varepsilon}{(t\ell)^3} \sum_{z\in\mathbb{Z}^3} \sum_{j=1}^{24} \int_{\mathrm{SO}(3)} \int_{C_{t\ell}} E(\xi_{t\ell,t\delta,j}(R\cdot - t\ell z - \tau)\rho)\, dR\, d\tau \\ - C\varepsilon \int_{\mathbb{R}^3} (\rho + \varepsilon^2 \rho^2). \tag{79}$$

We recall that here $\xi_{\ell,\delta,j} = \mathbb{1}_{\ell\mu_j \mathbf{\Delta}} * \eta_\delta$; see Section 4.2. In order to use the same argument as for the upper bound, we are going to prove the estimate

$$\left(\int_{\frac12}^{\frac32} \frac{ds}{s^4}\right)^{-1} \int_{\frac12}^{\frac32} \frac{dt}{t^4(t\ell)^3} \Bigg\{ E(\rho(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta})) \\ - \int_{\mathbb{R}^3} e_{\mathrm{UEG}}(\rho(x))\, (\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta})\, dx + C \int_{\mathbb{R}^3} \rho |\nabla \sqrt{\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}}|^2 \\ + C\varepsilon \int_{\mathbb{R}^3} (\rho + \rho^2)(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}) + \frac{C}{\varepsilon} \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2 (\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}) \Bigg\} \\ \geqslant -\frac{C}{\varepsilon^{4p-1}} \int_{2\ell\mathbf{\Delta} + B_{2\delta}} |\nabla \rho^\theta|^p. \tag{80}$$

That the last integral is over the larger set $2\ell\mathbf{\Delta} + B_{2\delta}$ will only affect the multiplicative constant C. Inserting (80) into (79) gives a bound as in (78) but in the opposite direction. This concludes the proof of the theorem and it therefore only remains to prove (80).

With an abuse of notation we consider the minimal and maximal values over the larger set $2(\ell\mathbf{\Delta} + B_\delta)$,

$$\underline{\rho} := \min_{2\ell\mathbf{\Delta} + B_{2\delta}} \rho \quad \text{and} \quad \bar\rho := \min_{2\ell\mathbf{\Delta} + B_{2\delta}} \rho \tag{81}$$

instead of the corresponding definitions on the smaller set $\ell\mathbf{\Delta} + B_\delta$. First we again recall that, by (29) and (19), we have

$$\left| E(\rho(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta})) - \int_{\mathbb{R}^3} (\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}) e_{\mathrm{UEG}}(\rho(x))\, dx \right| \\ \leqslant C \int_{\mathbb{R}^3} (\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta})(\rho^{\frac43} + \rho^{\frac53}) + C \int_{\mathbb{R}^3} (\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}) |\nabla \sqrt{\rho}|^2 + C \int_{\mathbb{R}^3} \rho \left|\nabla \sqrt{\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}}\right|^2.$$

Hence there is nothing to prove when

$$\int_{\mathbb{R}^3} (\rho^{\frac43} + \rho^{\frac53})(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}) \leqslant C\varepsilon \int_{\mathbb{R}^3} (\rho + \rho^2)(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}) + \frac{1}{\varepsilon^{4p-1}} \int_{2\ell\mathbf{\Delta} + B_{2\delta}} |\nabla \rho^\theta|^p.$$

This is the case if $\bar\rho^{1/3} \leqslant C\varepsilon$, for instance. Hence we may assume in the following that $\bar\rho \geqslant C\varepsilon^3$ and that

$$\int_{\mathbb{R}^3} (\rho^{\frac{4}{3}} + \rho^{\frac{5}{3}})(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}) \geqslant \frac{1}{\varepsilon^{4p-1}} \int_{2\ell\mathbf{\Delta}+B_{2\delta}} |\nabla\rho^\theta|^p.$$

By (66), this implies

$$\frac{\ell^3}{\varepsilon^{3p\theta-4}}\bar\rho^{p\theta} \geqslant \frac{C}{\ell^{p-3}\varepsilon^{4p-1}}(\bar\rho^\theta - \underline{\rho}^\theta)^p,$$

that is,

$$\bar\rho^\theta - \underline{\rho}^\theta \leqslant C\varepsilon^{\frac{3}{p}(1+\frac{5p}{6}-\theta p)}\bar\rho^\theta.$$

Under our assumption (7) on p and θ the exponent is positive; hence we deduce that for ε small enough

$$\bar\rho \leqslant C\underline{\rho} \leqslant C\rho(x)$$

on $2\ell\mathbf{\Delta} + B_{2\delta}$. With this additional information we can use our previous estimates.

By arguing exactly as in the proof of Proposition 18 with $\bar\rho$ the maximum over $2\ell\mathbf{\Delta} + B_{2\delta}$ instead of the support of $\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}$, we get the estimate similar to (64)

$$\begin{aligned} &E(\rho\,(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta})) \\ &\geqslant E(\bar\rho\,(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta})) - C\varepsilon\ell^3(\bar\rho + \bar\rho^2) - \frac{C}{\varepsilon}\int_{\mathbb{R}^3} |\nabla\sqrt{\bar\rho}|^2(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}) - \frac{C}{\varepsilon^{4p-1}}\int_{2\ell\mathbf{\Delta}+B_{2\delta}} |\nabla\rho^\theta|^p. \end{aligned} \quad (82)$$

From the fact that $\bar\rho \leqslant C\rho$, the second term on the right side can be bounded by

$$C\varepsilon\int_{\mathbb{R}^3} (\rho + \rho^2)(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}).$$

Then we average over t and use our lower estimate (58) on the averaged energy of a tetrahedron. This gives

$$\left(\int_{\frac{1}{2}}^{\frac{3}{2}} \frac{ds}{s^4}\right)^{-1} \int_{\frac{1}{2}}^{\frac{3}{2}} \frac{dt}{t^4} \frac{E(\bar\rho\,(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}))}{(t\ell)^3|\mathbf{\Delta}|} \geqslant e_{\mathrm{UEG}}(\bar\rho) - C\varepsilon\bar\rho^2.$$

The last term can again be bounded by

$$\varepsilon(t\ell)^{-3}\int_{\mathbb{R}^3} \rho^2(\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta})$$

and included into the average over t. Finally, using (69) and $\bar\rho \leqslant C\rho$, we infer that

$$e_{\mathrm{UEG}}(\bar\rho) \geqslant e_{\mathrm{UEG}}(\rho(x)) - C(\bar\rho - \rho(x))\rho(x)^{\frac{1}{3}}$$

on the support of $\mathbb{1}_{t\ell\mathbf{\Delta}} * \eta_{t\delta}$. To conclude the proof of (80) we can proceed in the same way as for the upper bound (75). This concludes the proof of Theorem 3. □

Appendix: Classical case

In the classical case where the kinetic energy is neglected, the grand-canonical energy functional is defined [Lewin et al. 2018] by

$$E_{\rm cl}(\rho) := \inf_{\substack{\sum_{n=0}^\infty \mathbb{P}_n(\mathbb{R}^{3n})=1 \\ \sum_{n=1}^\infty \rho_{\mathbb{P}_n}=\rho}} \sum_{n=1}^\infty \int_{(\mathbb{R}^3)^n} \sum_{1\leqslant j<k\leqslant n} \frac{1}{|x_j-x_k|}\, d\mathbb{P}_n(x_1,\dots,x_n) - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy, \quad (83)$$

where each $\mathbb{P}_n$ is a symmetric probability measure on $(\mathbb{R}^3)^n$ with density

$$\rho_{\mathbb{P}_n}(x) = n\int_{\mathbb{R}^{3(n-1)}} d\mathbb{P}_n(x,x_2,\dots,x_n).$$

This classical energy (83) is obtained from the quantum energy in the limit

$$\lim_{\alpha\to 0} \alpha^{-\frac{4}{3}} E(\alpha^3\rho(\alpha\cdot)) = E_{\rm cl}(\rho);$$

see [Cotar et al. 2013; 2018; Bindini and De Pascale 2017; Lewin 2018]. When $\int_{\mathbb{R}^3}\rho = N\in\mathbb{N}$, the canonical version of $E_{\rm cl}$ reads

$$E_{\rm cl}^{\rm can}(\rho) := \inf_{\rho_{\mathbb{P}}=\rho} \int_{(\mathbb{R}^3)^N} \sum_{1\leqslant j<k\leqslant N} \frac{1}{|x_j-x_k|}\, d\mathbb{P}(x_1,\dots,x_N) - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy. \quad (84)$$

In [Lewin et al. 2018] we have shown that for $\rho_N(x)=\rho_1(N^{-1/3}x)$ with $\int_{\mathbb{R}^3}\rho_1=1$,

$$\lim_{N\to\infty}\frac{E_{\rm cl}^{\rm can}(\rho_N)}{N} = \lim_{N\to\infty}\frac{E_{\rm cl}(\rho_N)}{N} = c_{\rm UEG}\int_{\mathbb{R}^3}\rho(x)^{\frac{4}{3}}\,dx, \quad (85)$$

where

$$c_{\rm UEG} = \lim_{\rho\to 0^+}\frac{e_{\rm UEG}(\rho)}{\rho^{\frac{4}{3}}} < 0$$

is the energy per unit volume of the classical uniform electron gas at density 1. In this appendix we quickly explain how to derive the following quantitative estimate on the convergence rate in (85).

Theorem 21 (estimate in the (grand-canonical) classical case). *Let $p>3$ and $0<\theta<1$ such that $\theta p\geqslant\frac{4}{3}$. There exists a universal constant $C=C(p,\theta)$ such that*

$$\left|E_{\rm cl}(\rho) - c_{\rm UEG}\int_{\mathbb{R}^3}\rho(x)^{\frac{4}{3}}\,dx\right| \leqslant \varepsilon\int_{\mathbb{R}^3}(\rho(x)+\rho(x)^{\frac{4}{3}})\,dx + \frac{C}{\varepsilon^b}\int_{\mathbb{R}^3}|\nabla\rho^\theta(x)|^p\,dx, \quad (86)$$

with

$$b = \max\{2p-1,(1+3\theta)p-4\},$$

for every $\varepsilon>0$ and every nonnegative density $\rho\in L^1(\mathbb{R}^3)\cap L^{4/3}(\mathbb{R}^3)$ such that $\nabla\rho^\theta\in L^p(\mathbb{R}^3)$.

Under the condition that $\theta p \leqslant 1 + \frac{1}{3}p$, which is slightly more restrictive than in the quantum case (7), we get the much smaller power $2p-1$ of ε in front of the gradient term. Then, after optimizing (86) in ε, we obtain the quantitative estimate

$$\left| E_{\mathrm{cl}}(\rho_N) - N\, c_{\mathrm{UEG}} \int_{\mathbb{R}^3} \rho(x)^{\frac{4}{3}}\, dx \right| \leqslant C N^{\frac{5}{6}} \left(\int_{\mathbb{R}^3} (\rho(x) + \rho(x)^{\frac{4}{3}})\, dx \right)^{1-\frac{1}{2p}} \left(\int_{\mathbb{R}^3} |\nabla \rho^\theta(x)|^p\, dx \right)^{\frac{1}{2p}} \quad (87)$$

for every $\rho_N(x) = \rho_1(N^{-1/3}x)$ and for $\frac{4}{3} \leqslant \theta p \leqslant 1 + \frac{1}{3}p$. The rate $N^{5/6}$ is better than the $N^{11/12}$ obtained in the quantum case, but still far from the expected rate $N^{1/3}$.

Proof. The estimate (86) follows from the Lieb–Oxford inequality (17) (without the gradient term) when ε is large, so we only have to consider the case where ε is small.

We use again the tiling (31) and, as in the proof in [Lewin et al. 2018], the upper bound

$$E_{\mathrm{cl}}(\rho) \leqslant \sum_{z\in\mathbb{Z}^3} \sum_{j=1}^{24} E_{\mathrm{cl}}(\mathbb{1}_{\ell\mu_j \mathbf{\Delta} + \ell z}\rho) \leqslant \sum_{z\in\mathbb{Z}^3} \sum_{j=1}^{24} E_{\mathrm{cl}}(\mathbb{1}_{\ell\mu_j \mathbf{\Delta} + \ell z}\underline{\rho}), \quad (88)$$

where $\underline{\rho} = \min_{\ell\mu_j \mathbf{\Delta} + \ell z} \rho$. The inequality (88) is a consequence of the subadditivity and the negativity of E_{cl}. Now it follows from [Lewin et al. 2018, Corollary 3.4] and from the Graf–Schenker inequality as in Section 4.3 that in a tetrahedron

$$c_{\mathrm{UEG}} \rho_0^{\frac{4}{3}} \leqslant \frac{E_{\mathrm{cl}}(\rho_0 \mathbb{1}_{\ell\mathbf{\Delta}})}{\ell^3 |\mathbf{\Delta}|} \leqslant c_{\mathrm{UEG}} \rho_0^{\frac{4}{3}} + \frac{C\rho_0}{\ell}.$$

This provides the upper bound

$$E_{\mathrm{cl}}(\rho) \leqslant c_{\mathrm{UEG}} \int_{\mathbb{R}^3} \rho^{\frac{4}{3}} + |c_{\mathrm{UEG}}| \sum_{z\in\mathbb{Z}^3} \sum_{j=1}^{24} \int_{\ell\mu_j \mathbf{\Delta} + \ell z} (\rho^{\frac{4}{3}} - \underline{\rho}^{\frac{4}{3}}) + \frac{C}{\ell} \int_{\mathbb{R}^3} \rho.$$

For the rest of the argument we use the notation $\varepsilon = 1/\ell$. In each tetrahedron we can follow the argument in (74) and estimate

$$\begin{aligned}
\int_{\mathbb{R}^3} (\rho^{\frac{4}{3}} - \underline{\rho}^{\frac{4}{3}})(\mathbb{1}_{\ell\mathbf{\Delta}} * \eta_\delta) &\leqslant C \|\rho^\theta - \underline{\rho}^\theta\|^a_{L^\infty(\ell\mathbf{\Delta} + B_\delta)} \int_{\mathbb{R}^3} \rho^{\frac{4}{3} - \theta a} (\mathbb{1}_{\ell\mathbf{\Delta}} * \eta_\delta) \\
&\leqslant C \left(\frac{1}{\varepsilon^p} \int_{\ell\mathbf{\Delta} + B_\delta} |\nabla \rho^\theta|^p \right)^{\frac{a}{p}} \left(\int_{\mathbb{R}^3} \rho^{\frac{4-3\theta a}{3} \frac{p}{p-a}} (\mathbb{1}_{\ell\mathbf{\Delta}} * \eta_\delta) \right)^{1-\frac{a}{p}} \\
&\leqslant C \left(\frac{1}{\varepsilon^{p+\frac{p}{a}-1}} \int_{\ell\mathbf{\Delta} + B_\delta} |\nabla \rho^\theta|^p + \varepsilon \int_{\mathbb{R}^3} \rho^{\frac{4-3\theta a}{3} \frac{p}{p-a}} (\mathbb{1}_{\ell\mathbf{\Delta}} * \eta_\delta) \right),
\end{aligned}$$

with $0 < a \leqslant 1$. In order to estimate the second term by $\rho + \rho^{4/3}$, we need that

$$1 \leqslant \frac{4-3\theta a}{3} \frac{p}{p-a} \leqslant \frac{4}{3},$$

which is equivalent to

$$\frac{4}{3} \leqslant \theta p \leqslant 1 + \frac{p}{3a}$$

and which we assume for the rest of the proof.

We finally turn to the lower bound. Up to an appropriate rotation and translation of the tiling (or, equivalently, of the density ρ), the Graf–Schenker inequality gives the following lower bound [Lewin et al. 2018, p. 100]:

$$E_{\mathrm{cl}}(\rho) \geqslant \sum_{z\in\mathbb{Z}^3}\sum_{j=1}^{24} E_{\mathrm{cl}}(\mathbb{1}_{\ell\mu_j\mathbf{\Delta}+\ell z}\rho). \tag{89}$$

We recall that

$$E_{\mathrm{cl}}(\mathbb{1}_{\ell\mu_j\mathbf{\Delta}+\ell z}\rho) \geqslant -c_{\mathrm{LO}}\int_{\ell\mu_j\mathbf{\Delta}+\ell z}\rho^{\frac{4}{3}}$$

by the Lieb–Oxford inequality (17). We have nothing to prove in any tetrahedron $\ell\mu_j\mathbf{\Delta}+\ell z$ such that

$$c_{\mathrm{LO}}\int_{\ell\mu_j\mathbf{\Delta}+\ell z}\rho^{\frac{4}{3}} \leqslant \varepsilon\int_{\ell\mu_j\mathbf{\Delta}+\ell z}(\rho+\rho^{\frac{4}{3}}) + \frac{A}{\varepsilon^{p+\frac{p}{a}-1}}\int_{\ell\mu_j\mathbf{\Delta}+\ell z}|\nabla\rho^\theta|^p.$$

Here A is a large constant to be chosen later. This is in particular the case when $\bar\rho^{1/3} = \max_{\ell\mu_j\mathbf{\Delta}+\ell z}\rho \leqslant \varepsilon/c_{\mathrm{LO}}$. So we may assume that

$$c_{\mathrm{LO}}\int_{\ell\mu_j\mathbf{\Delta}+\ell z}\rho^{\frac{4}{3}} > \varepsilon\int_{\ell\mu_j\mathbf{\Delta}+\ell z}(\rho+\rho^{\frac{4}{3}}) + \frac{A}{\varepsilon^{p+\frac{p}{a}-1}}\int_{\ell\mu_j\mathbf{\Delta}+\ell z}|\nabla\rho^\theta|^p$$

and that $\bar\rho > (\varepsilon/c_{\mathrm{LO}})^3$. This implies

$$\bar\rho^\theta - \underline{\rho}^\theta \leqslant \frac{C\varepsilon^{\frac{1}{a}-\frac{1}{p}}}{A^{\frac{1}{p}}}\bar\rho^{\frac{4}{3p}} \leqslant \frac{C\varepsilon^{\frac{3}{p}(1+\frac{p}{3a}-\theta p)}}{A^{\frac{1}{p}}}\bar\rho^\theta.$$

The power of ε is nonnegative when, again,

$$\theta p \leqslant 1+\frac{p}{3a}.$$

For A large enough (or ε small enough) this gives $\bar\rho \leqslant C\underline{\rho} \leqslant C\rho$ in the simplex $\ell\mu_j\mathbf{\Delta}+\ell z$. The rest of the argument is then exactly the same as for the upper bound.

As a conclusion we obtain the bound (86) with the error term

$$\frac{C}{\varepsilon^{p+\frac{p}{a}-1}}\int_{\ell\mu_j\mathbf{\Delta}+\ell z}|\nabla\rho^\theta|^p$$

and the restrictions that $p>3,\ 0<\theta\leqslant 1,\ 0<a\leqslant 1$ and

$$\frac{4}{3} \leqslant \theta p \leqslant 1+\frac{p}{3a}.$$

In order to minimize the power of ε we want to take a as large as possible; that is,

$$a = \min\left(1, \frac{p}{3(\theta p-1)}\right).$$

For $\theta p \leqslant 1+\frac{1}{3}p$ we take $a=1$, whereas for $\theta p > 1+\frac{1}{3}p$ we choose the other value and get the stated inequality (86). □

Remark 22 (canonical case). As was mentioned in (85), in [Lewin et al. 2018] we could also handle the canonical case. We would easily obtain a quantitative estimate on the canonical energy $E_{\mathrm{cl}}^{\mathrm{can}}(\rho)$ if we knew the speed of convergence of $\ell^{-3}E_{\mathrm{cl}}^{\mathrm{can}}(\rho_0 \mathbb{1}_{\ell\mathbf{\Delta}})$ to its limit $c_{\mathrm{UEG}}|\mathbf{\Delta}|$. Unfortunately, the argument used in [Lewin et al. 2018, Lemma 3.2] to prove that the limit coincides with the grand-canonical one does not seem to produce a quantitative bound.

Acknowledgments

The authors thank the Institut Henri Poincaré for its hospitality. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreements AQUAMS No. 694227 of Seiringer and MDFT No. 725528 of Lewin).

References

[Bach et al. 1994] V. Bach, E. H. Lieb, and J. P. Solovej, "Generalized Hartree–Fock theory and the Hubbard model", *J. Statist. Phys.* **76**:1-2 (1994), 3–89. MR Zbl

[Benguria et al. 2012] R. D. Benguria, G. A. Bley, and M. Loss, "A new estimate on the indirect Coulomb energy", *Int. J. Quantum Chem.* **112**:6 (2012), 1579–1584.

[Bindini and De Pascale 2017] U. Bindini and L. De Pascale, "Optimal transport with Coulomb cost and the semiclassical limit of density functional theory", *J. Éc. polytech. Math.* **4** (2017), 909–934. MR Zbl

[Burke and Wagner 2013] K. Burke and L. O. Wagner, "DFT in a nutshell", *Int. J. Quantum Chem.* **113**:2 (2013), 96–101.

[Chan and Handy 1999] G. K.-L. Chan and N. C. Handy, "Optimized Lieb-Oxford bound for the exchange-correlation energy", *Phys. Rev. A* **59**:4 (1999), 3075–3077.

[Cotar and Petrache 2019a] C. Cotar and M. Petrache, "Equality of the jellium and uniform electron gas next-order asymptotic terms for Coulomb and Riesz potentials", preprint, 2019. arXiv 1707.07664v5

[Cotar and Petrache 2019b] C. Cotar and M. Petrache, "Next-order asymptotic expansion for N-marginal optimal transport with Coulomb and Riesz costs", *Adv. Math.* **344** (2019), 137–233. MR Zbl

[Cotar et al. 2013] C. Cotar, G. Friesecke, and C. Klüppelberg, "Density functional theory and optimal transportation with Coulomb cost", *Comm. Pure Appl. Math.* **66**:4 (2013), 548–599. MR Zbl

[Cotar et al. 2018] C. Cotar, G. Friesecke, and C. Klüppelberg, "Smoothing of transport plans with fixed marginals and rigorous semiclassical limit of the Hohenberg–Kohn functional", *Arch. Ration. Mech. Anal.* **228**:3 (2018), 891–922. MR Zbl

[Dereziński and Gérard 1999] J. Dereziński and C. Gérard, "Asymptotic completeness in quantum field theory: massive Pauli–Fierz Hamiltonians", *Rev. Math. Phys.* **11**:4 (1999), 383–450. MR Zbl

[Dreizler and Gross 1990] R. M. Dreizler and E. K. U. Gross, *Density functional theory: an approach to the quantum many-body problem*, Springer, 1990. Zbl

[Engel and Dreizler 2011] E. Engel and R. M. Dreizler, *Density functional theory: an advanced course*, Springer, 2011. MR Zbl

[Fefferman 1985] C. Fefferman, "The thermodynamic limit for a crystal", *Comm. Math. Phys.* **98**:3 (1985), 289–311. MR Zbl

[Giuliani and Vignale 2005] G. Giuliani and G. Vignale, *Quantum theory of the electron liquid*, Cambridge University Press, 2005.

[Graf and Schenker 1995] G. M. Graf and D. Schenker, "On the molecular limit of Coulomb gases", *Comm. Math. Phys.* **174**:1 (1995), 215–227. MR Zbl

[Gregg 1989] J. N. Gregg, Jr., "The existence of the thermodynamic limit in Coulomb-like systems", *Comm. Math. Phys.* **123**:2 (1989), 255–276. MR Zbl

[Hainzl et al. 2009] C. Hainzl, M. Lewin, and J. P. Solovej, "The thermodynamic limit of quantum Coulomb systems, II: Applications", *Adv. Math.* **221**:2 (2009), 488–546. 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

[Hohenberg and Kohn 1964] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas", *Phys. Rev.* (2) **136** (1964), B864–B871. MR

[Hughes 1985] W. Hughes, "Thermodynamics for Coulomb systems: a problem at vanishing particle densities", *J. Statist. Phys.* **41**:5-6 (1985), 975–1013. MR Zbl

[Kirzhnits 1957] D. A. Kirzhnits, "Quantum corrections to the Thomas–Fermi equation", *Soviet Physics. JETP* **5** (1957), 64–71. In Russian. MR Zbl

[Kohn and Sham 1965] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects", *Phys. Rev.* (2) **140** (1965), A1133–A1138. MR

[Langreth and Mehl 1983] D. C. Langreth and M. J. Mehl, "Beyond the local-density approximation in calculations of ground-state electronic properties", *Phys. Rev. B* **28**:4 (1983), 1809–1834.

[Langreth and Perdew 1980] D. C. Langreth and J. P. Perdew, "Theory of nonuniform electronic systems, I: Analysis of the gradient approximation and a generalization that works", *Phys. Rev. B* **21**:12 (1980), 5469–5493.

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

[Levy and Perdew 1993] M. Levy and J. P. Perdew, "Tight bound and convexity constraint on the exchange-correlation-energy functional in the low-density limit, and other formal tests of generalized-gradient approximations", *Phys. Rev. B* **48**:16 (1993), 11638–11645.

[Lewin 2011] M. Lewin, "Geometric methods for nonlinear many-body quantum systems", *J. Funct. Anal.* **260**:12 (2011), 3535–3595. MR Zbl

[Lewin 2018] M. Lewin, "Semi-classical limit of the Levy–Lieb functional in density functional theory", *C. R. Math. Acad. Sci. Paris* **356**:4 (2018), 449–455. MR Zbl

[Lewin and Lieb 2015] M. Lewin and E. H. Lieb, "Improved Lieb–Oxford exchange-correlation inequality with a gradient correction", *Phys. Rev. A* (3) **91**:2 (2015), art. id. 022507. MR

[Lewin et al. 2018] M. Lewin, E. H. Lieb, and R. Seiringer, "Statistical mechanics of the uniform electron gas", *J. Éc. polytech. Math.* **5** (2018), 79–116. MR Zbl

[Lewin et al. 2019] M. Lewin, E. H. Lieb, and R. Seiringer, "Floating Wigner crystal with no boundary charge fluctuations", *Phys. Rev. B* **100**:3 (2019), art. id. 035127.

[Lieb 1979] E. H. Lieb, "A lower bound for Coulomb energies", *Phys. Lett. A* **70**:5-6 (1979), 444–446. MR

[Lieb 1981a] E. H. Lieb, "Thomas–Fermi and related theories of atoms and molecules", *Rev. Modern Phys.* **53**:4 (1981), 603–641. MR Zbl

[Lieb 1981b] E. H. Lieb, "Variational principle for many-fermion systems", *Phys. Rev. Lett.* **46**:7 (1981), 457–459. MR

[Lieb 1983] E. H. Lieb, "Density functionals for Coulomb systems", *Int. J. Quantum Chem.* **24**:3 (1983), 243–277.

[Lieb and Narnhofer 1975] E. H. Lieb and H. Narnhofer, "The thermodynamic limit for jellium", *J. Statist. Phys.* **12** (1975), 291–310. MR Zbl

[Lieb and Oxford 1981] E. H. Lieb and S. Oxford, "Improved lower bound on the indirect Coulomb energy", *Int. J. Quantum Chem.* **19**:3 (1981), 427–439.

[Lieb and Seiringer 2010] E. H. Lieb and R. Seiringer, *The stability of matter in quantum mechanics*, Cambridge University Press, 2010. MR Zbl

[Lieb and Thirring 1975] E. H. Lieb and W. E. Thirring, "Bound for the kinetic energy of fermions which proves the stability of matter", *Phys. Rev. Lett.* **35**:11 (1975), 687–689.

[Lieb and Thirring 1976] E. H. Lieb and W. Thirring, "Inequalities for the moments of the eigenvalues of the Schrödinger Hamiltonian and their relation to Sobolev inequalities", pp. 269–303 in *Studies in mathematical physics: essays in honor of Valentine Bargmann*, edited by E. H. Lieb, Princeton University Press, 1976. Zbl

[March and Young 1958] N. H. March and W. H. Young, "Variational methods based on the density matrix", *Proc. Phys. Soc.* **72**:2 (1958), 182–192.

[Mardirossian and Head-Gordon 2017] N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals", *Molecular Phys.* **115**:19 (2017), 2315–2372.

[Nam 2018] P. T. Nam, "Lieb–Thirring inequality with semiclassical constant and gradient error term", *J. Funct. Anal.* **274**:6 (2018), 1739–1746. MR Zbl

[Parr and Weitao 1994] R. G. Parr and Y. Weitao, *Density-functional theory of atoms and molecules*, Oxford University Press, 1994.

[Perdew and Kurth 2003] J. P. Perdew and S. Kurth, "Density functionals for non-relativistic Coulomb systems in the new century", pp. 1–55 in *A primer in density functional theory* (Caramulo Mountains, Portual, 2001), edited by C. Fiolhais et al., Springer, 2003. Zbl

[Perdew and Schmidt 2001] J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy", *AIP Conf. Proc.* **577**:1 (2001), 1–20.

[Pribram-Jones et al. 2015] A. Pribram-Jones, D. A. Gross, and K. Burke, "DFT: a theory full of holes?", *Ann. Rev. Phys. Chem.* **66** (2015), 283–304.

Received 3 Apr 2019. Revised 15 Jul 2019. Accepted 20 Aug 2019.

MATHIEU LEWIN: mathieu.lewin@math.cnrs.fr
CNRS & CEREMADE, Université Paris-Dauphine, PSL University, Paris, France

ELLIOTT H. LIEB: lieb@princeton.edu
Departments of Mathematics and Physics, Princeton University, Princeton, NJ, United States

ROBERT SEIRINGER: robert.seiringer@ist.ac.at
Institute of Science and Technology Austria, Klosterneuburg, Austria

PURE and APPLIED ANALYSIS

Vol. 2, No. 1, 2020
dx.doi.org/10.2140/paa.2020.2.75

msp

SPARSE BOUNDS FOR THE DISCRETE SPHERICAL MAXIMAL FUNCTIONS

ROBERT KESLER, MICHAEL T. LACEY AND DARÍO MENA

We prove sparse bounds for the spherical maximal operator of Magyar, Stein and Wainger. The bounds are conjecturally sharp, and contain an endpoint estimate. The new method of proof is inspired by ones by Bourgain and Ionescu, is very efficient, and has not been used in the proof of sparse bounds before. The Hardy–Littlewood circle method is used to decompose the multiplier into major and minor arc components. The efficiency arises as one only needs a single estimate on each element of the decomposition.

1. Introduction

Let $\mathcal{A}_\lambda f = d\sigma_\lambda * f$, where $d\sigma_\lambda$ is a uniform unit mass spherical measure on a sphere of radius λ in $\mathbb{R}^d$ for $d \geq 3$. Set the Stein spherical maximal operator to be

$$\mathcal{A} f(x) = \sup_{\lambda > 0} \mathcal{A}_\lambda f,$$

where f is a nonnegative compactly supported and bounded function. We are interested in a sparse bound for the maximal function. In the continuous case, this estimate holds, and is sharp, up to the boundary.

Theorem 1.1 [Lacey 2017]. *Let $d \geq 3$ and set $\boldsymbol{R}_d$ to be the polygon with vertices*

$$R_0 = \Big(\frac{d-1}{d}, \frac{1}{d}\Big), \quad R_1 = \Big(\frac{d-1}{d}, \frac{d-1}{d}\Big), \quad R_2 = \Big(\frac{d^2-d}{d^2+1}, \frac{d^2-d+2}{d^2+1}\Big), \quad R_3 = (0,1).$$

(See Figure 1.) *Then, for all $\big(\frac{1}{p}, \frac{1}{q}\big)$ in the interior of $\boldsymbol{R}_d$, we have the sparse bound $\|\mathcal{A}\|_{p,q} < \infty$.*

We set notation for the sparse bounds. Call a collection of cubes $\mathcal{S}$ in $\mathbb{R}^n$ *sparse* if there are sets $\{E_S : S \in \mathcal{S}\}$ which are pairwise disjoint and satisfy $E_S \subset S$ and $|E_S| > \frac{1}{4}|S|$. For any cube Q and $1 \leq r < \infty$, set $\langle f \rangle_{Q,r}^r = |Q|^{-1} \int_Q |f|^r \, dx$. Then the (r,s)-sparse form $\Lambda_{\mathcal{S},r,s} = \Lambda_{r,s}$ indexed by the sparse collection $\mathcal{S}$ is

$$\Lambda_{\mathcal{S},r,s}(f,g) = \sum_{S \in \mathcal{S}} |S| \langle f \rangle_{S,r} \langle g \rangle_{S,s}.$$

For a sublinear operator T, we set $\|T\|_{r,s}$ to be the best constant C in the inequality

$$\langle Tf, g \rangle < C \sup_{\mathcal{S}} \Lambda_{\mathcal{S},r,s}(f,g).$$

We use the same notation for sublinear operators T acting on functions defined on $\mathbb{Z}^d$.

Lacey was supported in part by grant from the US National Science Foundation, DMS-1600693 and the Australian Research Council ARC DP160100153. Mena was supported by project 821-B8-287, CIMPA, Escuela de Matemática, UCR.
MSC2010: 11K70, 42B25.
Keywords: sparse bounds, spherical averages, discrete, spherical maximal function.

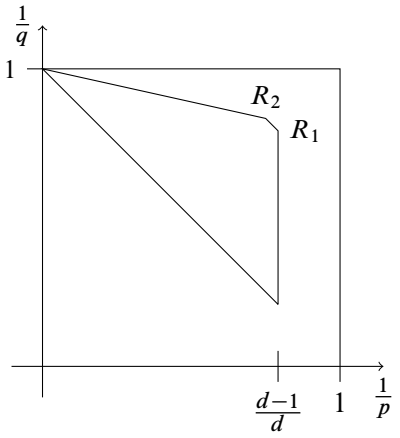

Figure 1. Sparse bounds hold for points $\left(\frac{1}{p}, \frac{1}{q}\right)$ in the interior of the four-sided region $\boldsymbol{R}_d$. The points R_1 and R_2 are as given in Theorem 1.1.

The theorem above refines the well-known L^p-improving properties for the local maximal function $\sup_{1\le\lambda\le 2} \mathcal{A}_\lambda * f$, proved in [Schlag 1997; Schlag and Sogge 1997]; also see [Lee 2003]. The theorem above has as immediate corollaries (a) vector-valued inequalities, and (b) weighted consequences. Both sets of consequences are the strongest known. The method of proof uses the L^p-improving inequalities for the spherical maximal function. That is, the proof is, in some sense, standard, although only recently discovered, and yields the best known information about the mapping properties of the spherical maximal function.

We turn to the setting of discrete spherical averages. Provided λ^2 is an integer, and the dimension d satisfies $d \ge 5$, we can define

$$A_\lambda f(x) = \lambda^{2-d} \sum_{n\in\mathbb{Z}^d : |n|=\lambda} f(x-n)$$

for functions $f \in \ell^2(\mathbb{Z}^d)$. We restrict attention to the case of $d \ge 5$ as in that case $|\{n \in \mathbb{Z}^d : |n| = \lambda\}| \simeq \lambda^{d-2}$ for all $\lambda^2 \in \mathbb{N}$. Let $Af = \sup_\lambda A_\lambda f$, where we will always understand that $\lambda^2 \in \mathbb{N}$. This is the maximal function of [Magyar 1997; Magyar, Stein, and Wainger 2002]. The following theorem is the best known extension of the sparse bounds for the continuous spherical maximal function to the discrete setting.

Theorem 1.2. *Let $\boldsymbol{Z}_d$ be the polygon with vertices*

$$Z_j = \frac{d-4}{d-2} R_j + \frac{2}{d-2}\left(\frac{1}{2}, \frac{1}{2}\right), \quad j = 0, 1, 2, \tag{1.3}$$

and $Z_3 = (0, 1)$. (See Figure 2.) *There holds*:

(1) *For all $\left(\frac{1}{p}, \frac{1}{q}\right)$ in the interior $\boldsymbol{Z}_d$, we have the sparse bound $\|Af\|_{p,q} < \infty$.*

(2) *When $f = \mathbf{1}_F$ and $g = \mathbf{1}_G$,*

$$\langle A\mathbf{1}_F, \mathbf{1}_G\rangle \lesssim \sup_{\mathcal{S}} \Lambda_{\mathcal{S}, d/(d-2), d/(d-2)}(\mathbf{1}_F, \mathbf{1}_G). \tag{1.4}$$

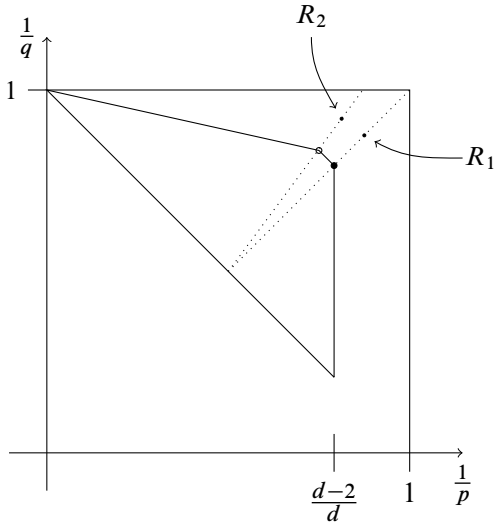

Figure 2. Sparse bounds for the discrete spherical maximal function hold for points $\big(\frac{1}{p}, \frac{1}{q}\big)$ in the interior of the four-sided figure above. The dotted lines pass through the points $\big(\frac{1}{2}, \frac{1}{2}\big)$ and the points R_1 and R_2 of Figure 1. Circles along these lines are the points Z_1 and Z_2. The restricted weak-type sparse bound (1.4) holds at the filled in circle, $Z_1 = \big(\frac{d}{d-2}, \frac{d}{d-2}\big)$.

By direct computation,

$$Z_0 = \left(\frac{d-2}{d}, \frac{2}{d}\right), \quad Z_1 = \left(\frac{d-2}{d}, \frac{d-2}{d}\right), \quad Z_2 = \left(\frac{d^3-4d^2+4d+1}{d^3-2d^2+d-2}, \frac{d^3-4d^2+6d-7}{d^3-2d^2+d-2}\right). \tag{1.5}$$

The sparse bound near the point $\big(\frac{d-2}{d}, \frac{2}{d}\big)$ implies the maximal inequality of [Magyar, Stein, and Wainger 2002], namely that $A : \ell^p(\mathbb{Z}^d) \to \ell^p(\mathbb{Z}^d)$ for $p > \frac{d}{d-2}$. The sparse bound (1.4) requires that both functions be indicator sets, and so is of restricted weak type. It implies the restricted weak-type inequality of [Ionescu 2004]. These inequalities imply a wide range of weighted and vector-valued inequalities, all of which are new. See the applications of the sparse bound in the continuous case in [Lacey 2017].

The discrete spherical maximal function ℓ^p-bounds were established in [Magyar, Stein, and Wainger 2002], with an endpoint restricted weak-type estimate proved in [Ionescu 2004]. The discrete ℓ^p-improving inequalities have only recently been investigated. The case of a fixed radius was addressed, independently, in [Hughes 2018; Kesler and Lacey 2018]. Spherical maximal functions, restricting to lacunary and superlacunary cases, require different techniques [Hughes 2019; Cook 2018; Kesler, Lacey, and Mena 2019]. Robert Kesler [2018a; 2018b] established sparse bounds for the discrete case. This paper extends and simplifies those arguments.

It is very tempting to conjecture that our sparse bounds form the sharp range, up to the endpoints. One would expect that certain kinds of natural examples would demonstrate this. But examples are much harder to come by in the discrete setting. We return to this in Section 5.

The argument in this paper is elegant, especially if one restricts attention to the endpoint estimate (1.4), and much simpler than the arguments in [Kesler 2018a; 2018b]. It proceeds by decomposing the maximal function into a series of terms, guided by the Hardy–Littlewood circle method decomposition developed in [Magyar, Stein, and Wainger 2002]. The decomposition has many parts, as indicated in Figures 3 and 4. But, for each part of the decomposition, we need only one estimate, either an ℓ^2-estimate, or an endpoint estimate. Roughly speaking, one uses either a "high-frequency" ℓ^2-estimate, or a "low-frequency" inequality, in which one compares to smoother averages. Interestingly, the notion of "smoother averages" varies. The argument of Ionescu combined with the sparse perspective yields a powerful inequality. Notation and conventions will be established in this section, and used throughout the paper.

2. Proof of the sparse bounds inside the polygon Z_d

A sparse bound is typically proved by recursion. So, the main step is to prove the recursive statement. To do this, we fix a large dyadic cube E and functions $f = \mathbf{1}_F$ and $g = \mathbf{1}_G$ supported on E. We say that $\tau : E \to \{1, \dots, \ell E\}$ is an *admissible stopping time* if for any subcube $Q \subset E$ with $\langle f \rangle_Q > C \langle f \rangle_E$, for some large constant C to be chosen later, we have $\min_{x \in Q} \tau(x) > \ell Q$.

Lemma 2.1. *Let $\left(\frac{1}{p}, \frac{1}{q}\right)$ be in the interior of $\boldsymbol{Z}_d$. For any dyadic cube, functions $f = \mathbf{1}_F$ and $g = \mathbf{1}_G$ supported on E, and any admissible stopping time τ, there holds*

$$|E|^{-1} \langle A_\tau f, g \rangle \lesssim \langle f \rangle_E^{1/p} \langle g \rangle_E^{1/q}. \tag{2.2}$$

We complete the proof of the main result.

Proof of Theorem 1.2(1). We can assume that there is a fixed dyadic cube E so that $f = \mathbf{1}_F$ is supported on the cube $3E$, and $g = \mathbf{1}_G$ is supported on E. Let $\mathcal{Q}_E$ be the maximal dyadic subcubes of E for which $\langle f \rangle_{3Q} > C \langle f \rangle_{3E}$ for a large constant C. Observe that we have, for an appropriate choice of admissible $\tau(x)$,

$$\Big\langle \sup_{\lambda \le \ell(E)} A_\lambda f, g \Big\rangle \le \langle A_\tau f, g \rangle + \sum_{Q \in \mathcal{Q}_E} \Big\langle \sup_{\lambda \le \ell(Q)} A_\lambda (f \mathbf{1}_{3Q}), g \mathbf{1}_Q \Big\rangle.$$

The first term is controlled by (2.2). For an appropriate constant $C \simeq 3^d$, we have

$$\sum_{Q \in \mathcal{Q}_E} |Q| \le \tfrac{1}{4} |E|.$$

We can clearly recurse on the second term above to construct our sparse bound. This proves a sparse bound for all indicator functions in the interior of $\boldsymbol{Z}_d$.

Sparse bounds for indicator functions in an open set self-improve to sparse bounds for functions. We give the details in the last section; see Lemma 4.1. □

We use the corresponding recursive inequality for spherical averages on $\mathbb{R}^d$. Recall that $\mathcal{A}_\lambda$ is the continuous spherical average.

Lemma 2.3 [Lacey 2017, Lemma 3.4]. *Let $\big(\frac{1}{\bar p}, \frac{1}{\bar q}\big)$ be in the interior of $\boldsymbol{R}_d$. For any dyadic cube E, functions $\phi = \mathbf{1}_F$ and $\gamma = \mathbf{1}_G$ supported on E, and any admissible stopping time τ, there holds*

$$|E|^{-1}\langle \mathcal{A}_\tau \phi, \gamma\rangle \lesssim \langle \phi\rangle_E^{1/\bar p}\langle \gamma\rangle_E^{1/\bar q}. \tag{2.4}$$

We turn to the proof of Lemma 2.1. The restriction to indicator functions will allow us to use interpolation arguments, even though our setting has stopping times, and hence is nonlinear. Let L be the line through $\big(\frac12, \frac12\big)$ and $\big(\frac1p, \frac1q\big)$. Then, let $\big(\frac{1}{\bar p}, \frac{1}{\bar q}\big)$ be a point on L that is in the interior of $\boldsymbol{R}_d$, and very close to the boundary. (The dashed lines in Figure 2 are examples of the lines L we are discussing here.)

This is the point: Fix $\big(\frac{1}{\bar p}, \frac{1}{\bar q}\big) \in \boldsymbol{R}_d$. For all sufficiently small $0 < \epsilon < 1$ so that $\big(\frac{1}{\bar p+\epsilon}, \frac{1}{\bar q+\epsilon}\big) \in \boldsymbol{R}_d$ and integers $N \in \mathbb{N}$, we can write $A_\tau f \le M_1 + M_2$, where

$$|E|^{-1}\langle M_1, g\rangle \lesssim N^{1+\epsilon}\langle f\rangle_E^{1/(\bar p+\epsilon)}\langle g\rangle_E^{1/(\bar q+\epsilon)}, \tag{2.5}$$

$$|E|^{-1}\langle M_2, g\rangle \lesssim N^{d\epsilon+(4-d)/2}\langle f\rangle_E^{1/2}\langle g\rangle_E^{1/2}. \tag{2.6}$$

Implied constants depend upon $\bar p, \bar q$ and ϵ, but we do not track the dependence. Once this is proved, one has

$$|E|^{-1}\langle A_\tau f, g\rangle \lesssim N^{1+\epsilon}\langle f\rangle_E^{1/(\bar p+\epsilon)}\langle g\rangle_E^{1/(\bar q+\epsilon)} + N^{d\epsilon+(4-d)/2}\langle f\rangle_E^{1/2}\langle g\rangle_E^{1/2}.$$

Choosing N to minimize the right-hand side and letting $\big(\frac{1}{\bar p}, \frac{1}{\bar q}\big)$ and $0 < \epsilon < 1$ vary completes the proof. Indeed, ignoring ϵ's, we see that the value of p is given by

$$\frac1p = \frac{1}{\bar p} + \frac{2}{d-2}\Big(\frac12 - \frac{1}{\bar p}\Big) = \frac{2}{d-2}\cdot\frac12 + \frac{d-4}{d-2}\cdot\frac{1}{\bar p}.$$

Compare this to our description of the extreme points of $\boldsymbol{Z}_d$ in (1.3). Thus, our lemma follows.

In proving (2.5) and (2.6), it suffices, given $0 < \epsilon < 1$, to prove the statement for sufficiently large N. We will do so for $N > N_0$, for a sufficiently large choice of $N_0 > 0$. Indeed, we find it necessary to use an absorption argument. We show that

$$A_\tau f \le M_1 + M_2 + \tfrac12 A_\tau f, \tag{2.7}$$

where M_1 and M_2 are as in (2.5) and (2.6).

The reader can consult Figure 3 for a guide to the argument. For a technical reason, we assume that $F \subset (2\mathbb{Z}^d) + \delta_f$ and $G \subset (2\mathbb{Z}^d) + \delta_g$. Here, $\delta_f, \delta_g \in \{0,1\}^d$. This can be assumed without loss of generality.

Small values of τ. The terms M_1 and M_2 have several components. The first contribution to M_1 is the term $M_{1,1} = \mathbf{1}_{\tau \le N^{1+\epsilon}} A_\tau f$. Our verification that $M_{1,1}$ satisfies (2.5) is our first application of the $\mathbb{R}^d$ inequality (2.4).

We need functions on $\mathbb{R}^d$. Take $\phi(x) = \sum_{n\in\mathbb{Z}^d} \mathbf{1}_F(n)\mathbf{1}_{n+[-1,1)^d}(x)$, and define γ similarly. By the reduction we made above, these are indicator functions. Moreover, if τ is an admissible stopping time for f, then it is for ϕ as well. The inequality (2.4) holds for these two functions on $\mathbb{R}^d$. Then, notice that we can compare the discrete and continuous spherical averages as follows:

$$A_\tau f(x) \lesssim \tau \mathcal{A}_\tau \phi(x). \tag{2.8}$$

Therefore, if we require that $\tau \le N^{1+\epsilon}$, we see that (2.4) implies that $M_{1,1}$ satisfies (2.5).

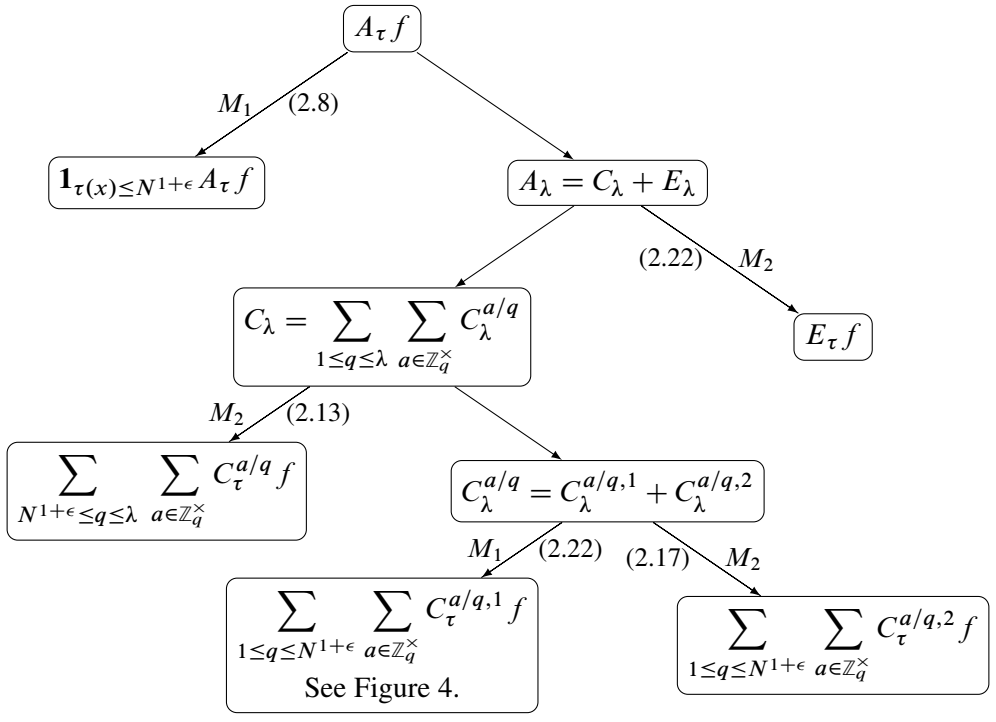

Figure 3. The flow of the proof of (2.2). The nodes of the tree indicate the different elements of the decomposition, and a label on an arrow shows to which of M_1 or M_2 that term contributes. Above, λ represents a fixed choice of radius, and $\tau = \tau(x)$ an admissible choice of radius. For space considerations, several terms of the form $e_q(-\lambda^2 a)$ have been omitted; compare to (2.10).

The decomposition. Below, we assume that $\tau > N^{1+\epsilon}$ pointwise. At this point, we need a decomposition of $A_\lambda f$ into a family of multipliers. We recall this from [Magyar, Stein, and Wainger 2002]. Upper case letters denote a convolution operator, and lower case letters denote the corresponding multiplier. Let $e(x) = e^{2\pi i x}$ and for integers q, $e_q(x) = e\big(\frac{x}{q}\big)$.

$$A_\lambda f = C_\lambda f + E_\lambda f, \tag{2.9}$$

$$C_\lambda f = \sum_{1\le q\le\lambda} \sum_{a\in\mathbb{Z}_q^\times} e_q(-\lambda^2 a) C_\lambda^{a/q} f, \tag{2.10}$$

$$c_\lambda^{a/q}(\xi) = \widehat{C}_\lambda^{a/q}(\xi) = \sum_{\ell\in\mathbb{Z}_q^d} G\left(\frac{a}{q},\ell\right) \widetilde{\psi}_q\left(\xi - \frac{\ell}{q}\right) \widetilde{d\sigma}_\lambda\left(\xi - \frac{\ell}{q}\right), \tag{2.11}$$

$$G\left(\frac{a}{q},\ell\right) = q^{-d} \sum_{n\in\mathbb{Z}_q^d} e_q(|n|^2 a + n\cdot\ell).$$

The term $G\big(\frac{a}{q},\ell\big)$ is a normalized Gauss sum. In (2.10), the sum over $a\in\mathbb{Z}_q^\times$ means that $(a,q)=1$. In (2.11), the hat indicates the Fourier transform on $\mathbb{Z}^d$, and the notation conflates the operator $C_\lambda^{a/q}$ and

the kernel. All our operators are convolution operators or maximal operators formed in the same way. The function ψ is a Schwartz function on $\mathbb{R}^d$ which satisfies

$$\mathbf{1}_{[-1/2,1/2]}(|\xi|) \le \widetilde{\psi}(\xi) \le \mathbf{1}_{[-1,1]}(|\xi|).$$

Above, $\tilde{f}$ denotes the Fourier transform of f on $\mathbb{R}^d$, and $\widetilde{\psi}_q(\xi) = \widetilde{\psi}(q\xi)$. The Fourier transform on $\mathbb{R}^d$ of $d\sigma_\lambda$ is $\widetilde{d\sigma}_\lambda$. Finally, we will use the notation λ for describing multipliers and so on, and using τ especially when obtaining estimates. In this way, many supremums will be suppressed from the notation.

The error term E_λ. The first contribution to M_2 is $M_{2,1} = |E_\tau f|$. The inequality below is from [Magyar, Stein, and Wainger 2002, Proposition 4.1], and it implies that $M_{2,1}$ satisfies (2.6) since $\tau > N^{1+\epsilon}$:

$$\Big\| \sup_{\Lambda \le \lambda \le 2\Lambda} |E_\lambda \cdot| \Big\|_{2\to 2} \lesssim \Lambda^{(4-d)/2}, \quad \Lambda \ge 1. \tag{2.12}$$

Large denominators. The second contribution to M_2 is

$$M_{2,2} = \Big| \sum_{N^{1+\epsilon} \le q \le \tau} e_q(-\lambda^2 a) C_\tau^{a/q} f \Big|. \tag{2.13}$$

The estimate below is a result of [Magyar, Stein, and Wainger 2002, Proposition 3.1], and it verifies that $M_{2,2}$ satisfies (2.6). We need only sum it over $1 \le a \le q$, and $q > N^{1+\epsilon}$:

$$\| \sup_{\lambda > q} |C_\lambda^{a/q} f| \|_2 \lesssim q^{-d/2} \|f\|_2. \tag{2.14}$$

Small denominators: a secondary decomposition. It remains to bound the small denominator case, namely

$$\sum_{1 \le q \le N^{1+\epsilon}} \sum_{a \in \mathbb{Z}_q^\times} e_q(-\lambda^2 a) C_\tau^{a/q} f,$$

with further contributions to M_1 and M_2. Write $C_\tau^{a/q} = C_\tau^{a/q,1} + C_\tau^{a/q,2}$, with this understanding. For an integer $1 \le Q \le \frac{1}{2}N$, and $Q \le q < 2Q$, define

$$\widehat{C}_\lambda^{a/q,1}(\xi) = \sum_{\ell \in \mathbb{Z}^d} G(a,\ell,q) \widetilde{\psi}_q\Big(\xi - \frac{\ell}{q}\Big) \widetilde{\psi}_{\lambda Q/N}\Big(\xi - \frac{\ell}{q}\Big) \widetilde{d\sigma}_\lambda\Big(\xi - \frac{\ell}{q}\Big). \tag{2.15}$$

Above, we have adjusted the cutoff around each point $\frac{\ell}{q} \in \mathbb{T}^d$.

Small denominators: the ℓ^2-part. We complete the construction of the term in M_2, (2.6), by showing that

$$\Big\| \sup_{N^{1+\epsilon} \le \lambda \le \ell(E)} |C_\lambda^{a/q,2} f| \Big\|_2 \lesssim q^{-1} N^{-(d-2)/2} \|f\|_2. \tag{2.16}$$

It follows that

$$\Big\| \sum_{1 \le q \le N^{1+\epsilon}} \sum_{a \in \mathbb{Z}_q^\times} e_q(-\lambda^2 a) C_\tau^{a/q,2} f \Big\|_2 \lesssim N^{-(d-4)/2+\epsilon} \|f\|_2. \tag{2.17}$$

This is the third and final contribution to M_2. We remark that the proof detailed below is a quantitative variant of the proof of Magyar, Stein and Wainger's inequality (2.14). The inequality (2.16) is [Ionescu 2004, (2.14)], but we include details here.

Let m be a smooth function supported on $\left[-\frac{1}{2}, \frac{1}{2}\right]^d$, and let T_m be the corresponding multiplier operator, either on $\mathbb{Z}^d$ or $\mathbb{R}^d$, with the notation indicating in which setting we are considering the multiplier.

This is a factorization argument from [Magyar, Stein, and Wainger 2002, p. 200]. Using the notation of (2.23) to define $M_{\psi_{Q/2}}$ below, for $Q \le q \le 2Q$, we have

$$\begin{aligned}\widehat{C}_\lambda^{a/q,2}(\xi) &= \widehat{M}_{\psi_{Q/2},q}(\xi) \cdot \sum_{\ell \in \mathbb{Z}^d} \widetilde{\psi}_q\left(\xi - \frac{\ell}{q}\right)\left(1 - \widetilde{\psi}_{\lambda Q/N}\left(\xi - \frac{\ell}{q}\right)\right)\widetilde{d\sigma}_\lambda\left(\xi - \frac{\ell}{q}\right)\\ &:= \widehat{M}_{\psi_{q/2},q}(\xi) \cdot \widehat{C}_\lambda^{a/q,3}(\xi).\end{aligned}$$

That is, the operator in question factors as $C_\lambda^{a/q,2} = C_\lambda^{a/q,3} \circ M_{\psi_{Q/2},q}$. Notice that by the Gauss sum estimate, we have

$$\|M_{\psi_{Q/2},q}\|_{2\to 2} \lesssim Q^{-d/2}. \tag{2.18}$$

In controlling the supremum, we need only consider the supremum over $C_\lambda^{a/q,3}$. The transference lemma [Magyar, Stein, and Wainger 2002, Corollary 2.1] allows us to estimate this supremum on $L^2(\mathbb{R}^d)$. We have

$$\|C_\tau^{a/q,3}\|_{\ell^2(\mathbb{Z}^d)\to\ell^2(\mathbb{Z}^d)} \lesssim \big\|\sup_{\lambda>0}|\Psi_{\lambda,q} \cdot |\big\|_{L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)}, \tag{2.19}$$

where $\widetilde{\Psi}_{\lambda,q} = \widetilde{\psi}_q(1-\widetilde{\psi}_{\lambda Q/N})\,\widetilde{d\sigma}_\lambda$. To estimate this last norm on $L^2(\mathbb{R}^d)$, we use this lemma of Bourgain:

Lemma 2.20 [Bourgain 1985, Proposition 2]. *Let m be a smooth function on $\mathbb{R}^d$. We have*

$$\big\|\sup_{r>0}|T_{m(r\,\cdot)} \cdot |\big\|_2 \lesssim \sum_{j\in\mathbb{Z}} \alpha_j^{1/2}(\alpha_j^{1/2} + \beta_j^{1/2}),$$

where

$$\alpha_j = \|\mathbf{1}_{2^j \le |\xi| \le 2^{j+1}} m(\xi)\|_\infty \quad \textit{and} \quad \beta_j = \|\mathbf{1}_{2^j \le |\xi| \le 2^{j+1}} \nabla m(\xi) \cdot \xi\|_\infty. \tag{2.21}$$

We bound the right side of (2.19). Composition with T_{ψ_q} is uniformly bounded on L^2. The multiplier in question is then $m(\xi) = (1 - \widetilde{\psi}_{Q/N})(\xi)\widetilde{d\sigma}_1(\xi)$. This is identically zero for $|\xi| \lesssim \frac{N}{Q}$. That means that for the terms in (2.21), we need only consider $2^j \gtrsim \frac{N}{Q} \ge 100$. Recall the standard stationary phase estimate

$$|\nabla\widetilde{d\sigma}_1(\xi)| + |\widetilde{d\sigma}_1(\xi)| \lesssim |\xi|^{-(d-1)/2}.$$

Hence, the bound for our multiplier is

$$\big\|\sup_{\lambda>0}|\Psi_{\lambda,q} \cdot |\big\|_{L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)} \lesssim \sum_{j\,:\,2^j \ge N/Q} 2^{-j(d-1)/2} \cdot 2^{-j(d-3)/2} \lesssim \left(\frac{Q}{N}\right)^{(d-2)/2}.$$

This estimate combined with (2.18) and (2.19) completes the proof of (2.16).

Small denominators: the sparse part. Recalling the notation from (2.15), we turn to

$$M_{1,2} f = \sum_{1\le q\le N} \sum_{a\in \mathbb{Z}_q^\times} e_q(-\lambda^2 a) C_\tau^{a/q,1} f$$

and show that this term is as in (2.7). This is the term in which the absorbing term $\frac{1}{2} A_\tau f$ in (2.7) arises. It is also the core of the proof.

Define

$$M_{1,Q} f = \sum_{Q\le q<2Q} \sum_{a\in \mathbb{Z}_q^\times} e_q(-\lambda^2 a) C_\tau^{a/q,1} f, \quad 1 \le Q \le \tfrac{1}{2} N,$$

Above, and below, we will treat $M_{1,Q}$ as an operator, as we have yet to tease out some of its additional properties. The main estimates to prove are

$$\begin{aligned} M_{1,Q} f &\le \overline{M}_{1,Q} f + N^{-\epsilon} A_\tau f, \\ |E|^{-1} \langle \overline{M}_{1,Q} f, g\rangle &\lesssim N^{1+\epsilon/2} \langle f\rangle_E^{1/(\bar p+\epsilon)} \langle g\rangle_E^{1/(\bar q+\epsilon)}. \end{aligned} \tag{2.22}$$

These are summed over dyadic $1 \le Q \le \frac{1}{2} N$ to complete the proof of the absorption inequality (2.7). This step requires that N be sufficiently large, $N > \kappa^{1/\epsilon}$, but that is sufficient for our purposes.

We need the estimate (2.4) on $\mathbb{R}^d$. We also need kernel estimates for the operators $M_{1,Q}$, and for that we require this preparation, which has been noted before [Magyar, Stein, and Wainger 2002; Ionescu 2004, p. 1415]. For a function ζ with $\tilde\zeta$ supported on $[-1,1]^d$, define a family of Fourier multipliers by

$$\widehat{M}_{\zeta,q}(\xi) = \sum_{\ell\in \mathbb{Z}_q^d} G\Big(\frac{a}{q}, \ell\Big) \tilde\zeta\Big(\xi - \frac{\ell}{q}\Big).$$

By inspection, the Gauss sum map $\ell \mapsto G\big(\frac{a}{q}, \ell\big)$ is the Fourier transform of $e_q(|x|^2 a)$ as a function on $\mathbb{Z}_q^d$. From this, and a routine computation, it follows that

$$M_{\zeta,q}(x) = e_q(a|x|^2)\zeta(x). \tag{2.23}$$

(Here we identify the kernel of the convolution operator, and the operator itself.)

It follows that the kernel of $M_{1,Q} f$ is

$$\begin{aligned} M_{1,Q}(n) &= \psi_{\tau Q/N} * d\sigma_\tau(n) \sum_{Q\le q\le 2Q} \sum_{a\in \mathbb{Z}_q^\times} e_q(a(|n|^2 - \lambda^2)) \\ &= P_{Q,\tau}(n) \cdot \mathsf{C}_Q(|n|^2 - \lambda^2). \end{aligned} \tag{2.24}$$

Note that $P_{Q,\tau}$ is a maximal average over annuli of outer radius τ, and width about $\tau \frac{Q}{N} < \tau$. The second term above is related to Ramanujan sums, defined by

$$\mathsf{c}_q(m) = \sum_{a\in \mathbb{Z}_q^\times} e_q(am), \quad m \in \mathbb{Z},$$

so that in (2.24), $\mathsf{C}_Q = \sum_{Q\le q\le 2Q} \mathsf{c}_q$. Ramanujan sums satisfy very good cancellation properties. The properties we will need are summarized in the next result.

Lemma 2.25. *The following estimates hold for any $k \in \mathbb{N}$, and $\epsilon > 0$:*

(1) *For any Q and n, $|\mathsf{C}_Q(n)| \le Q^2$.*

(2) *There holds*

$$\max_{0<m\le Q^k} |\mathsf{C}_Q(m)| \lesssim Q^{1+\epsilon}. \tag{2.26}$$

(3) *For $M > Q^k$,*

$$\left[\frac{1}{M}\sum_{m\le M} |C_Q(m)|^k\right]^{1/k} \lesssim Q^{1+\epsilon}. \tag{2.27}$$

The implied constants depend upon k and ϵ.

Proof. The first estimate is trivial, but we include it for the sake of clarity. Note that $C_Q(0) \simeq Q^2$, which will arise in the absorption argument below.

An argument for the second inequality (2.26) begins with the inequality $|\mathsf{c}_q(m)| \le (q, m)$ for $m > 0$. This can be checked by inspection if q is a power of a prime. The general case follows as both sides are multiplicative functions.

Then, of course we have, for any $1 \le d \le q$,

$$\sum_{k:dk\le Q} d \le Q.$$

It follows that

$$\sum_{q\le Q}(q, m) \le Q \sum_{d\le q\,:\,d\,|\,m} 1 = Q\delta(m; Q),$$

where $\delta(m; Q)$ is the number of divisors of m that are less than or equal to Q. But, $m \le Q^k$, so by a well-known logarithmic-type estimate for the divisor function, we have (2.26).

The third property is harder. It is due to [Bourgain 1993]. There are proofs in [Kesler, Lacey, and Mena 2019, Lemma 2.13] and [Kesler 2018b, Lemma 5]. □

A tertiary decomposition. The preparations are finished. It remains to prove (2.22), and this argument is indicated in Figure 4. There are three cases, namely,

(1) $Q < N^{1/2}$,

(2) $NQ^{k_1-1} < \tau$, where $k_1 = k_1(\bar p, \bar q)$,

(3) $N^{1/2} \le Q$ and $\tau \le NQ^{k_1-1}$, implying $N < \tau < Q^{k_2}$, where $k_2 = k_2(\bar p, \bar q)$.

We treat these cases in order, with the core case being the last one.

The first and easiest case concerns $Q < N^{1/2}$. Using the trivial bound $|\mathsf{c}_q(n)| \le q$ and (2.24) we have

$$|M_{1,Q} f| < N \cdot P_{Q,\tau} f.$$

It then follows from the continuous sparse bound (2.4) that we have

$$|E|^{-1}\langle M_{1,Q} f, g\rangle \lesssim N \langle f\rangle_E^{1/\bar p}\langle g\rangle^{1/\bar q}. \tag{2.28}$$

This is as required in (2.22).

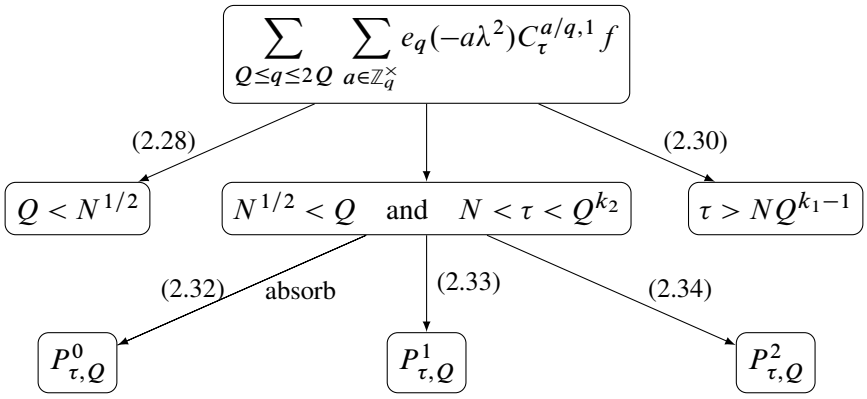

Figure 4. The flow of the proof of (2.22). The integers k_1 and k_2 are large and functions of ϵ, $\bar{p}$ and $\bar{q}$. The first level of the decomposition is motivated by the estimates for Ramanujan sums in Lemma 2.25. The second level of the diagram is associated with the Ramanujan estimate (2.26). It requires a further decomposition of the kernel $P_{\tau,Q}$ in (2.24).

The second case we restrict to the case that $\tau > NQ^{k_1-1}$ for a sufficiently large integer k_1 that is a function of $(\bar{p},\bar{q})$. This case does not have an absorbing term.

Dominate, using Hölder's inequality with ℓ^{k_1}-$\ell^{k_1'}$ duality,

$$\begin{aligned} |M_{1,Q} f(n)| &\lesssim [P_{Q,\tau} * f(n)]^{1/k_1'} \Big[\sum_{x\in\mathbb{Z}^d} |C_Q(|x|^2-\lambda^2)|^{k_1} P_{Q,\tau}(x) \Big]^{1/k_1} \\ &\lesssim Q^{1+\epsilon/2} P_{Q,\tau} * f(n)^{1/k_1'}. \end{aligned} \tag{2.29}$$

Recall that f is an indicator function. Notice that we are using a bound on the Ramanujan sums that follows from (2.27). Recall that $P_{Q,\tau}$ is an average over an annulus around the sphere of radius λ, of width $\tau\frac{Q}{N}$. In particular, the width is greater than Q^{k_1} by the assumption that $\tau > NQ^{k_1-1}$.

But, then, we are free to conclude our statement; since $\big(\frac{1}{\bar{p}},\frac{1}{\bar{q}}\big)$ are in the interior of $\boldsymbol{R}_d$, we have, for k_1 sufficiently large, so that k_1' is sufficiently close to 1, as required in (2.22),

$$\begin{aligned} |E|^{-1}\langle |M_1 f|, g\rangle &\lesssim Q^{1+\epsilon} |E|^{-1} \langle [P_{Q,\tau} * f]^{1/k_1'}, g\rangle \\ &\lesssim N [|E|^{-1} \langle P_{Q,\tau} * f, g\rangle]^{1/k_1'} \\ &\lesssim \langle f\rangle_E^{1/(\bar{p}k_1')} \langle g\rangle_E^{1/(\bar{q}k_1')}. \end{aligned} \tag{2.30}$$

Here, we use (2.29) and then the real-variable inequality (2.4), which we can do if k_1 is sufficiently large, so that $(k_1'\bar{p}, k_1'\bar{q}) \in \boldsymbol{R}_d$. This is our second application of (2.4).

We turn to third case of $N < \tau < Q^{k_2}$. A final, fourth decomposition of $P_{\tau,Q}$ is needed, and the absorption argument appears. Let $\mathbb{S}_\lambda = \{n \in \mathbb{Z}^d : |n| = \lambda\}$ be the integer sphere of radius λ, and set

$$P_{\tau,Q} = \sum_{j=0}^{2} P^j_{\tau,Q},$$

where

$$\begin{aligned} P^0_{\tau,Q}(n) &= P_{\tau,Q}(n)\mathbf{1}_{\mathbb{S}_\tau}(n), \\ P^1_{\tau,Q}(n) &= P_{\tau,Q}(n)\mathbf{1}_{0<\operatorname{dist}(n,\mathbb{S}_\tau)<\tau Q^{1+\epsilon}/N}, \end{aligned} \tag{2.31}$$

and $P^2_{\tau,Q}$ is then defined. The term $P^0_{\tau,Q}$ is a multiple of the average over the integer sphere of radius λ, and the term $P^1_{\tau,Q}$ is just that part of $P_{\tau,Q}$ that is close to, but not equal to, the sphere of radius $\mathbb{S}_\tau$.

Let us detail the absorbing term. Note that $\mathsf{C}_Q(0) \simeq Q^2$. Indeed, we have to single out this case as there is no cancellation in the Ramanujan sum when the argument is zero. Using the definition (2.31), we have

$$\begin{aligned} |P^0_{\tau,Q}(n) \cdot \mathsf{C}_Q(|n|^2 - \tau^2)| &\lesssim Q^2 \frac{N}{Q\tau^d} \mathbf{1}_{\mathbb{S}_\tau}(n) \\ &\lesssim \frac{NQ}{\tau^2} \cdot \tau^{2-d} \mathbf{1}_{\mathbb{S}_\tau}(n) \lesssim N^{-2\epsilon} \cdot \tau^{2-d} \mathbf{1}_{\mathbb{S}_\tau}(n). \end{aligned} \tag{2.32}$$

This is as required in (2.7) and (2.22).

The inequality (2.26) on the Ramanujan sums applies in the analysis of $P^1_{\tau,Q}$, due to our assumptions $Q < \tau < Q^{k_2}$. It shows that

$$|P^1_{Q,\tau}(n) \cdot \mathsf{C}_Q(|n|^2 - \tau^2)| \lesssim Q^{1+\epsilon} \frac{N}{\tau^d Q} \mathbf{1}_{0<\operatorname{dist}(n,\mathbb{S}_\tau)<\tau Q^{1+\epsilon}/N}.$$

Keeping normalizations in mind, it follows from our real-variable sparse inequality (2.4) that

$$|E|^{-1} \langle P^1_{Q,\tau} * f, g \rangle \lesssim Q^{1+2\epsilon} \langle f \rangle^{1/\bar{p}}_E \langle g \rangle^{1/\bar{q}}_E. \tag{2.33}$$

This is as required in (2.22).

The last term is $P^2_{Q,\tau}$. As noted, $|\mathsf{C}_q(m)| \le Q^2$. The condition $\tau < Q^{k_2}$, and simple Schwartz tail considerations then show that

$$|P^2_{Q,\tau}(n) \cdot \mathsf{C}_Q(|n|^2 - \tau^2)| \lesssim \frac{Q}{\tau^d} \left[1 + \frac{|n|}{\tau}\right]^{-2d}. \tag{2.34}$$

Since τ is an admissible stopping time, it follows that

$$|P^1_{Q,\tau} * f| \lesssim \langle f \rangle_E.$$

This completes the proof of (2.22).

3. The endpoint sparse bound

We need this definition. Given a cube E, we say that collection $\mathcal{Q}_E$ of subcubes $Q \subset E$ are *presparse* if the cubes $\{\frac{1}{3}Q : Q \in \mathcal{Q}_E\}$ are pairwise disjoint. Associated to a presparse collection $\mathcal{Q}_E$ is a family of stopping times. We say that τ is $\mathcal{Q}_E$ admissible (or just admissible) if

$$\ell(E) \ge \tau_{\mathcal{Q}_E}(x) = \tau(x) \ge \max\{1, \ell(Q)\mathbf{1}_{Q/3} : Q \in \mathcal{Q}_E\}, \quad x \in E.$$

The relevant lemma is this:

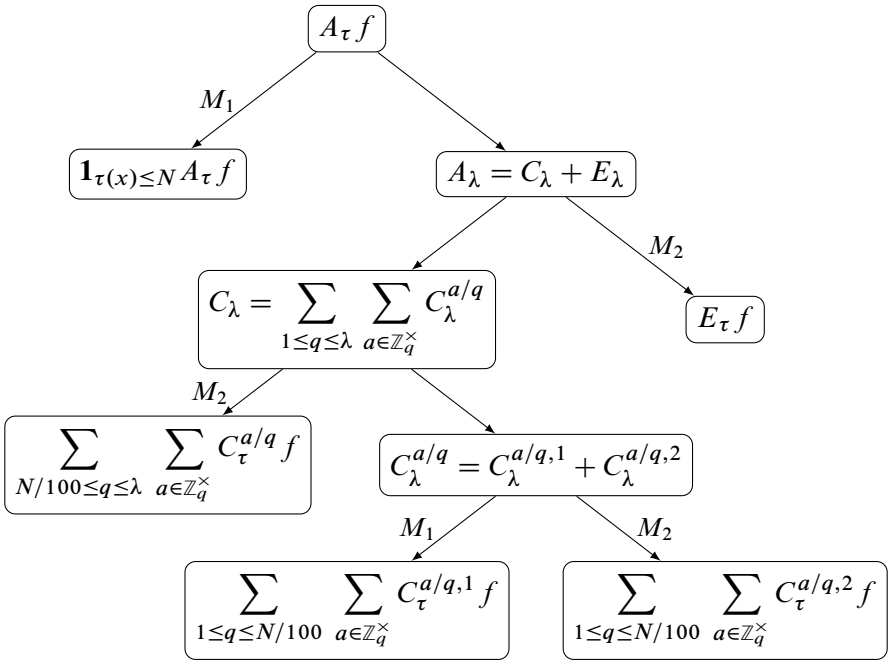

Figure 5. The flow of the proof of (3.3)–(3.5). The notation is similar to Figure 3.

Lemma 3.1. *For $f = \mathbf{1}_F$ supported on the cube $3E$, there is a presparse collection $\mathcal{Q}_E$ so that for all $\mathcal{Q}_E$-admissible $\tau = \tau(x)$, and all $g = \mathbf{1}_G$ supported on E, we have*

$$\langle A_\tau f, g\rangle \lesssim \langle f\rangle_{3E,d/(d-2)} \langle g\rangle_{E,d/(d-2)} |E|. \tag{3.2}$$

The lemma follows from this: for integers $N > 1$, there is a decomposition

$$A_\tau f \le M_1 + M_2, \tag{3.3}$$

$$\langle M_1, g\rangle \lesssim N^2 \langle f\rangle_{3E,1} \langle g\rangle_{E,1} |E|, \tag{3.4}$$

$$\langle M_2, g\rangle \lesssim N^{-(d-4)/2} \langle f\rangle_{3E,2} \langle g\rangle_{E,2} |E|. \tag{3.5}$$

Recalling that $f = \mathbf{1}_F$ and $g = \mathbf{1}_G$, the right sides above are comparable for $N \simeq [\langle f\rangle_{3E,1} \langle g\rangle_{E,1}]^{-1/d}$, and this proves (3.2).

It remains to prove (3.3)–(3.5). Our proof will be much shorter because we do not need to compare to the very rough continuous spherical averages, but to averages over balls. (In particular, we will not need any subtle facts about Ramanujan sums.) A guide to the argument is in Figure 5. The decomposition has several elements. The first begins with the trivial bound $A_\lambda f(x) \lesssim \lambda^2 B_\lambda * f(x)$, where B_λ is the average of over a ball of radius λ. Our first contribution to M_1 is

$$M_{1,1} = \mathbf{1}_{\tau(x)\le 100N} A_\tau f,$$

which is pointwise bounded by $CN^2\langle f\rangle_{3E,1}$, by choice of $\mathcal{Q}_E$. Thus (3.4) holds for this term. Below, we are free to assume that $\tau \geq 100N$.

Recall the decomposition of $A_\lambda f$, beginning with (2.9). Our first contribution to M_2 is $M_{2,1} = |E_\tau f|$. This satisfies (3.5) by (2.12).

It remains to bound $C_\tau f$ as defined in (2.10). This requires further contributions to M_1 and M_2. Apply (2.14) to see that this term obeys (3.5):

$$M_{2,2} = \sum_{N/100 \leq q \leq \lambda} \sum_{a \in \mathbb{Z}_q^\times} |C_\tau^{a/q} f|.$$

The remaining terms are

$$\sum_{1 \leq q \leq N/100} \sum_{a \in \mathbb{Z}_q^\times} e_q(-\lambda^2 a) C_\tau^{a/q} f.$$

Control will consist of an additional contribution to M_1 and M_2.

Write $C_\lambda^{a/q} = C_\lambda^{a/q,1} + C_\lambda^{a/q,2}$, where a different cut-off in frequency is inserted:

$$\begin{aligned}\widehat{C}_\lambda^{a/q,1}(\xi) &= \sum_{\ell \in \mathbb{Z}^d} G\Big(\frac{a}{q}, \ell\Big) \widetilde{\psi}_q\Big(\xi - \frac{\ell}{q}\Big) \widetilde{\psi}_{\lambda q/N}\Big(\xi - \frac{\ell}{q}\Big) \widetilde{d\sigma}_\lambda\Big(\xi - \frac{\ell}{q}\Big) \\ &= \sum_{\ell \in \mathbb{Z}^d} G\Big(\frac{a}{q}, \ell\Big) \widetilde{\psi}_{\lambda q/N}\Big(\xi - \frac{\ell}{q}\Big) \widetilde{d\sigma}_\lambda\Big(\xi - \frac{\ell}{q}\Big).\end{aligned}$$

This follows from the definition of ψ in (2.11). This is slightly different from (2.15); in particular, the term below satisfies (3.5), just as in the previous section:

$$M_{2,3} = \Big| \sum_{1 \leq q \leq N/100} \sum_{a \in \mathbb{Z}_q^\times} e_q(-\lambda^2 a) C_\tau^{a/q,2} f \Big|.$$

We claim that

$$\Big\| \sum_{1 \leq q \leq N/100} \sum_{a \in \mathbb{Z}_q^\times} e_q(-\lambda^2 a) C_\tau^{a/q,1} f \Big\|_\infty \lesssim N^2 \langle f\rangle_{E,1}. \tag{3.6}$$

That is, the term on the left is our second and final contribution to M_1, as in (3.4).

Recalling (2.23) and (2.24), we have

$$|C_\lambda^{a/q,1}(n)| \lesssim d\sigma_\lambda * \psi_{\lambda q/N}(n)$$

The convolution is with $d\sigma_\lambda$ and $\psi_{\lambda q/N}$, which is a bump function of integral 1, supported on scale $\frac{N}{\lambda q}$, which is much smaller than λ. As a consequence, we have

$$|C_\lambda^{a/q,1}(n)| \lesssim \frac{N}{q} \cdot \lambda^{-d} \Big[1 + \frac{|n|}{\lambda}\Big]^{-3d}.$$

This is summed over $1 \leq a < q \leq \frac{1}{100} N$. And, one appeals to the admissibility of the stopping time τ to complete the proof of (3.6).

4. Interpolation of sparse bounds

We show that if a sublinear operator satisfies an open range of sparse bounds for indicator sets, then they improve to sparse bounds for functions.

Lemma 4.1. *Suppose that a sublinear operator T satisfies the bound below for $1 < p, q < \infty$. For a fixed function f and all $|g| \le \mathbf{1}_G$, there is a sparse collection $\mathcal{S}$ so that*

$$|\langle Tf, g\rangle| \lesssim \Lambda_{\mathcal{S},p,q}(f, \mathbf{1}_G).$$

Then,

$$|\langle Tf, g\rangle| \lesssim \sup_{\mathcal{S}} \Lambda_{\mathcal{S},p,q}(f, g), \quad q < r < \infty.$$

Proof. In this proof, we will work on $\mathbb{R}^d$, with the same proof working on $\mathbb{Z}^d$. We will also assume that (1) T is a positive operator, and (b) all cubes are dyadic. The general case is not much harder than these considerations. Let f be a fixed function, and let g be a bounded compactly supported function. We will show that there is a sparse collection $\mathcal{S}$ so that

$$\langle Tf, g\rangle \lesssim \sum_{Q\in\mathcal{S}} \langle f\rangle_Q^{1/p} \langle g\rangle_{Q,q,1} |Q|, \tag{4.2}$$

where $\langle g\rangle_{Q,q,1} = \|g\mathbf{1}_Q\|_{L^{q,1}(dx/|Q|)}$ is the Lorentz space with normalized measure. Since $\langle g\rangle_{Q,q,1} < \langle g\rangle_{Q,r}$ for $1 < r < q$, this completes the proof of the lemma.

The argument for (4.2) is a level set argument. Thus, write $g \le \sum_{k\in\mathbb{Z}} 2^k \mathbf{1}_{G_k}$ for disjoint sets G_k. Apply the assumed sparse bound for indicators for each pair of sets (F, G_k). We get a sequence of sparse sets $\mathcal{S}_k$ so that

$$\langle Tf, g\rangle \lesssim \sum_{k\in\mathbb{Z}} \Lambda_{\mathcal{S}_k,p,q}(f, \mathbf{1}_{G_k}). \tag{4.3}$$

Let $\mathcal{F}$ be a sequence of stopping cubes for the averages $\langle f\rangle_{Q,p}$. That is, we choose $\mathcal{F}$ so that for any dyadic cube Q there is a dyadic cube in $\mathcal{F}$ that contains Q. And setting Q^a to be the minimal such cube in $\mathcal{F}$, we have $\langle f\rangle_{Q,p} \lesssim \langle f\rangle_{Q^a,p}$. The sum in (4.3) is organized according to $\mathcal{F}$. Below, we take $q < \alpha < \infty$, very close to q. For each $P \in \mathcal{F}$ we have

$$\begin{aligned}\Gamma(P) &= \sum_{k\in\mathbb{Z}} 2^{-k} \sum_{\substack{Q\in\mathcal{S}_k\\ Q^a=P}} \langle \mathbf{1}_{G_k}\rangle_Q^{1/q} |Q| \lesssim \sum_{k\in\mathbb{Z}} 2^{-k} |P|^{1/\alpha'} \Big[\sum_{\substack{Q\in\mathcal{S}_k\\ Q^a=P}} \langle \mathbf{1}_{G_k}\rangle_Q^{\alpha/q} \Big]^{1/\alpha}\\ &\lesssim |P| \sum_{k\in\mathbb{Z}} 2^{-k} \langle \mathbf{1}_{G_k}\rangle_P^{1/q} = \langle g\rangle_{L^{q,1}(P)} |P|.\end{aligned} \tag{4.4}$$

Here, we have used Hölder's inequality, sparseness of the collections $\mathcal{S}_k$ and the Carleson embedding inequality. The last inequality is one way to define the $L^{q,1}$ norm.

And, so we have

$$(4.3) \lesssim \sum_{P\in\mathcal{F}} \langle f\rangle_{P,p} \Gamma(P).$$

The bound in (4.4) then implies (4.2). □

Concerning our Theorem 1.2. For the first assertion, we have proved a restricted weak-type inequality for all $\left(\frac{1}{p}, \frac{1}{q}\right)$ in the interior of $\boldsymbol{Z}_d$. We see that we can then replace the indicator functions in both coordinates by functions. That is, we have $\langle A\rangle_{p,q} < \infty$ for all $\left(\frac{1}{p}, \frac{1}{q}\right)$ in the interior of $\boldsymbol{Z}_d$.

Concerning the second assertion, we have the restricted weak-type inequality at $\left(\frac{d}{d-2}, \frac{d}{d-2}\right)$. We see from the theorem above that we have

$$\langle Af, g\rangle \lesssim \sup_{\mathcal{S}} \sum_{Q\in\mathcal{S}} \langle f\rangle_{Q,d/(d-2),1} \langle g\rangle_{Q,q} |Q|, \quad q > \frac{d}{d-2}.$$

5. Counterexamples

We can show the following proposition, showing necessary conditions on (p, q) for the sparse bound to hold. The gap between the sufficient conditions for a sparse bound and these necessary conditions is illustrated in Figure 6.

Proposition 5.1. *If the sparse bound $\|A\|_{p,q} < \infty$ holds, we have*

$$\frac{1}{p_1} \le \frac{d-2}{d}, \quad \frac{2}{p} + \frac{d-1}{q} \le d. \tag{5.2}$$

Proof. If a (p, q)-sparse bound held we would conclude that $A : \ell^p \to \ell^{p,\infty}$. Taking $f = \mathbf{1}_0$, note that $Af(x) \simeq (1 + |x|)^{2-d}$. The latter function is $\ell^{d/(d-2),\infty}$. Hence, $p_1 \ge \frac{d}{d-2}$ is necessary.

For the second inequality in (5.2), set $\mathbb{S}_\lambda = \{n \in \mathbb{Z}^d : |n| = \lambda\}$. We recall that for $d \ge 4$, and any odd choice of $\lambda^2 \in \mathbb{N}$, we have $|\mathbb{S}_\lambda| \simeq \lambda^{d-2}$. For an odd choice of $\lambda^2 \in \mathbb{N}$, let $f = \mathbf{1}_{\mathbb{S}_\lambda}$, and consider $G = \left\{Af > \frac{c}{\lambda}\right\}$. This is the set of $x \in \mathbb{Z}^d$ for which the two spheres $\mathbb{S}_\lambda$ and $x + \mathbb{S}_\mu$, for some choice of $\mu \simeq \lambda$, have about the expected size. Here, necessarily $G \subset E$, a cube centered at the origin of side length about λ.

We claim that $|G| \gtrsim \lambda$. And observe that $Af(x) \gtrsim \lambda^{-1}$ for $x \in G$. Moreover, from the assumed (p, q)-sparse bound,

$$\lambda^{-1} \langle \mathbf{1}_G\rangle_E \lesssim \lambda^{-d} \langle Af, g\rangle \lesssim \langle f\rangle_E^{1/p} \langle \mathbf{1}_G\rangle_E^{1/q} \lesssim \lambda^{-2/p} \langle \mathbf{1}_G\rangle_E^{1/q}. \tag{5.3}$$

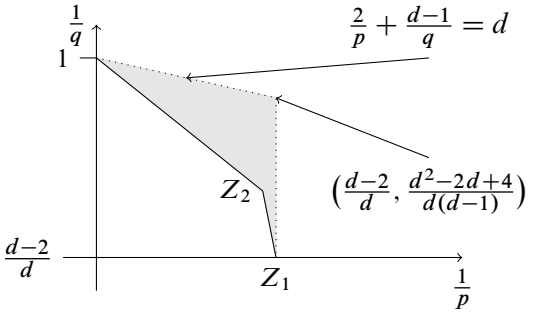

Figure 6. The horizontal line is set at $\frac{1}{q} = \frac{d-2}{d}$ for reasons of clarity. Sparse bounds hold below the solid line from $(0, 1)$, to Z_2 to $Z_1 = \left(\frac{d-2}{d}, \frac{d-2}{d}\right)$. (Recall that Z_2 is defined in (1.5).) They cannot hold to the right of Z_1, nor above the dotted line. The gray area is unresolved.

For this pair of functions, it is easy to see that the maximal sparse form is the expression on the right. Our lower bound on the size of G proves the proposition.

Note that since $d \geq 5$, and λ^2 is odd, there are about λ^{d-3} choices of vectors $y = (0, y_2, \dots, y_d) \in \mathbb{S}_\lambda$. (We insist on λ^2 being odd to capture the case of $d = 5$ here.) Then, if $|x_1| \leq \frac{1}{2}\lambda$, note that

$$\|(0, y_2, \dots, y_d) - (x_1, 0, \dots, 0)\| = \sqrt{\lambda^2 + x_1^2} = \lambda'.$$

From that, it follows that

$$A_{\lambda'} f(x_1, 0, \dots, 0) \simeq \lambda^{-1}.$$

This shows that $|G| \gtrsim \lambda$. □

Observe that the $\left(\frac{d}{d-2}, \frac{d}{d-2}\right)$-sparse bound and (5.3) imply that $|G| \lesssim \lambda^{(d+4)/2}$. Our lower bound is certainly not sharp. What is the correct size of G?

References

[Bourgain 1985] J. Bourgain, "Estimations de certaines fonctions maximales", *C. R. Acad. Sci. Paris Sér. I Math.* **301**:10 (1985), 499–502. MR

[Bourgain 1993] J. Bourgain, "Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations, I: Schrödinger equations", *Geom. Funct. Anal.* **3**:2 (1993), 107–156. MR Zbl

[Cook 2018] B. Cook, "A note on discrete spherical averages over sparse sequences", 2018. To appear in *Proc. Amer. Math. Soc.* arXiv

[Hughes 2018] K. Hughes, "ℓ^p-improving for discrete spherical averages", preprint, 2018. arXiv

[Hughes 2019] K. Hughes, "The discrete spherical averages over a family of sparse sequences", *J. Anal. Math.* (online publication July 2019).

[Ionescu 2004] A. D. Ionescu, "An endpoint estimate for the discrete spherical maximal function", *Proc. Amer. Math. Soc.* **132**:5 (2004), 1411–1417. MR Zbl

[Kesler 2018a] R. Kesler, "$\ell^p(\mathbb{Z}^d)$-improving properties and sparse bounds for discrete spherical maximal averages", 2018. To appear in *J. Anal. Math.* arXiv

[Kesler 2018b] R. Kesler, "$\ell^p(\mathbb{Z}^d)$-improving properties and sparse bounds for discrete spherical maximal means, revisited", preprint, 2018. arXiv

[Kesler and Lacey 2018] R. Kesler and M. T. Lacey, "ℓ^p-improving inequalities for discrete spherical averages", preprint, 2018. arXiv

[Kesler, Lacey, and Mena 2019] R. Kesler, M. T. Lacey, and D. Mena Arias, "Lacunary discrete spherical maximal functions", *New York J. Math.* **25** (2019), 541–557.

[Lacey 2017] M. T. Lacey, "Sparse bounds for spherical maximal functions", 2017. To appear in *J. Anal. Math.* arXiv

[Lee 2003] S. Lee, "Endpoint estimates for the circular maximal function", *Proc. Amer. Math. Soc.* **131**:5 (2003), 1433–1442. MR Zbl

[Magyar 1997] A. Magyar, "L^p-bounds for spherical maximal operators on $\mathbb{Z}^n$", *Rev. Mat. Iberoam.* **13**:2 (1997), 307–317. MR Zbl

[Magyar, Stein, and Wainger 2002] A. Magyar, E. M. Stein, and S. Wainger, "Discrete analogues in harmonic analysis: spherical averages", *Ann. of Math.* (2) **155**:1 (2002), 189–208. MR Zbl

[Schlag 1997] W. Schlag, "A generalization of Bourgain's circular maximal theorem", *J. Amer. Math. Soc.* **10**:1 (1997), 103–122. MR Zbl

[Schlag and Sogge 1997] W. Schlag and C. D. Sogge, "Local smoothing estimates related to the circular maximal theorem", *Math. Res. Lett.* **4**:1 (1997), 1–15. MR Zbl

Received 8 Apr 2019. Accepted 5 Jul 2019.

ROBERT KESLER: robertmkesler@gmail.com
School of Mathematics, Georgia Institute of Technology, Atlanta, GA, United States

MICHAEL T. LACEY: lacey@math.gatech.edu
School of Mathematics, Georgia Institute of Technology, Atlanta, GA, United States

DARÍO MENA: dario.menaarias@ucr.ac.cr
Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica

PURE and APPLIED
ANALYSIS

Vol. 2, No. 1, 2020
dx.doi.org/10.2140/paa.2020.2.93

msp

CHARACTERIZATION BY OBSERVABILITY INEQUALITIES OF CONTROLLABILITY AND STABILIZATION PROPERTIES

EMMANUEL TRÉLAT, GENGSHENG WANG AND YASHAN XU

Given a linear control system in a Hilbert space with a bounded control operator, we establish a characterization of exponential stabilizability in terms of an observability inequality. Such dual characterizations are well known for exact (null) controllability. Our approach exploits classical Fenchel duality arguments and, in turn, leads to characterizations in terms of observability inequalities of approximate null controllability and of α-null controllability. We comment on the relationships among those various concepts, at the light of the observability inequalities that characterize them.

1. Context and main result on stabilizability

Framework. Let X and U be Hilbert spaces. We consider the linear control system

$$\dot{y}(t) = Ay(t) + Bu(t), \quad t \geqslant 0, \tag{1}$$

where $A : D(A) \to X$ is a linear operator generating a C_0 semigroup $(S(t))_{t\geqslant 0}$ on X and $B \in L(U, X)$ is a control operator.

Given an initial state $y_0 \in X$ and a control $u \in L^2_{\mathrm{loc}}(0, +\infty; U)$, the unique solution $y(t) = y(t; y_0, u)$ $(t \geqslant 0)$ to (1), associated with u and the initial condition $y(0) = y_0$, satisfies

$$y(T; y_0, u) = S(T)y_0 + L_T u \quad \text{for all } T > 0, \tag{2}$$

with $L_T \in L(L^2(0, T; U), X)$ defined by $L_T u = \int_0^T S(T-t)Bu(t)\,dt$. Note that $y \in C^0([0, +\infty), X) \cap H^1_{\mathrm{loc}}(0, +\infty; X_{-1})$, with $X_{-1} = D(A^*)'$, the dual of $D(A^*)$ with respect to the pivot space X; see, e.g., [Engel and Nagel 2000; Tucsnak and Weiss 2009]. We recall that the dual mapping $L_T^* \in L(X, L^2(0, T; U))$ is given by $(L_T^*\psi)(t) = B^*S(T-t)^*\psi$ for every $\psi \in X$.

Throughout the paper we identify X (resp., U) with its dual X' (resp., U'). We denote by $\|\cdot\|_X$ and $\langle\cdot,\cdot\rangle_X$ (resp. $\|\cdot\|_U$ and $\langle\cdot,\cdot\rangle_U$) the Hilbert norm and scalar product in X (resp., in U).

It is well known that, for the control system (1), exact (null) controllability is equivalent by duality to an observability inequality. In the existing results (e.g., heat, wave, Schrödinger equations), such inequalities are instrumental to establish controllability properties; see the textbooks [Curtain and Zwart 1995; Lasiecka and Triggiani 2000a; Lions 1988; Staffans 2005; Tucsnak and Weiss 2009; Zabczyk 1995]. Additionally, exponential stabilizability, meaning that there exists a feedback operator K such that $A + BK$ generates an exponentially stable semigroup, is characterized in the existing literature in

MSC2010: 93B05, 93B07, 93C20.
Keywords: observability inequality, stabilizability, controllability.

terms of infinite-horizon linear quadratic optimal control and algebraic Riccati theory (see the previous references). But, to our knowledge, a dual characterization of exponential stabilizability in terms of an observability inequality is not known.

Stabilizability. The control system (1) is exponentially stabilizable if there exists a feedback operator $K \in L(X, U)$ such that the operator $A + BK$, of domain $D(A + BK) = D(A)$, generates an exponentially stable C_0 semigroup $(S_K(t))_{t\geqslant 0}$, i.e., there exists $M \geqslant 1$ and $\omega < 0$ such that

$$\|S_K(t)\|_{L(X)} \leqslant M\mathrm{e}^{\omega t} \quad \text{for all } t \geqslant 0. \tag{3}$$

The infimum ω_K of all possible real numbers ω such that (3) is satisfied for some $M \geqslant 1$ is the growth bound of the semigroup $(S_K(t))_{t\geqslant 0}$ and is given, see [Engel and Nagel 2000; Pazy 1983], by

$$\omega_K = \inf_{t>0} \frac{1}{t} \ln \|S_K(t)\|_{L(X)} = \lim_{t\to+\infty} \frac{1}{t} \ln \|S_K(t)\|_{L(X)}.$$

Exponential stabilizability means that there exists $K \in L(X, U)$ such that $\omega_K < 0$.

When the control system (1) is exponentially stabilizable, the best stabilization decay rate is defined by

$$\omega^* = \inf\{\omega_K \mid K \in L(X, U) \text{ such that } (S_K(t))_{t\geqslant 0} \text{ is exponentially stable}\}. \tag{4}$$

When $\omega^* = -\infty$, the control system (1) is said to be *completely stabilizable*: this means that stabilization can be achieved at any decay rate. We also speak of *rapid stabilization*.

Main result. Hereafter, we set

$$\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)} = \left(\int_0^T \|B^*S(T-t)^*\psi\|_U^2\, dt\right)^{1/2}.$$

Given $\alpha \geqslant 0$, $T > 0$ and $y_0 \in X$, we define

$$\mu^T_{y_0,\alpha} = \inf\{C \geqslant 0 \mid \langle\psi, S(T)y_0\rangle_X - \alpha\|y_0\|_X\|\psi\|_X \leqslant C\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)} \text{ for all } \psi \in X\}, \tag{5}$$

with the convention that $\inf\varnothing = +\infty$. Actually, when the set in (5) is not empty, the infimum is reached (see Section 4). By definition, we have

$$\begin{aligned}
&\mu^T_{y_0,\alpha} \in [0, +\infty],\\
&\mu^T_{y_0,\alpha_2} \leqslant \mu^T_{y_0,\alpha_1} \quad \text{if } \alpha_1 \leqslant \alpha_2,\\
&\mu^T_{y_0,\alpha} = 0 \quad \text{if } \alpha \geqslant \|S(T)\|_{L(X)},\\
&\mu^T_{\lambda y_0,\alpha} = \lambda\mu^T_{y_0,\alpha} \quad \text{for every } \lambda > 0
\end{aligned}$$

(and thus $\mu^T_{y_0,\alpha} = \mu^T_{y_0/\|y_0\|_X,\alpha}\|y_0\|_X$). This homogeneity property leads us to define

$$\mu^T_\alpha = \sup_{\|y_0\|_X=1} \mu^T_{y_0,\alpha}. \tag{6}$$

We claim that

$$\mu^T_\alpha = \inf\{C \geqslant 0 \mid \|S(T)^*\psi\|_X - \alpha\|\psi\|_X \leqslant C\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)} \text{ for all } \psi \in X\} \tag{7}$$

(see Lemma 29 in Section 4.2) and we have as well

$$\begin{aligned} &\mu_\alpha^T \in [0, +\infty], \\ &\mu_{\alpha_2}^T \leqslant \mu_{\alpha_1}^T \quad \text{if } \alpha_1 \leqslant \alpha_2, \\ &\mu_\alpha^T = 0 \qquad \text{if } \alpha \geqslant \|S(T)\|_{L(X)}. \end{aligned}$$

Theorem 1. *The following items are equivalent*:

(i) *The control system* (1) *is exponentially stabilizable.*

(ii) *For every* $\alpha \in (0, 1)$, *there exists* $T > 0$ *such that* $\mu_\alpha^T < +\infty$.

(iii) *There exist* $\alpha \in (0, 1)$ *and* $T > 0$ *such that* $\mu_\alpha^T < +\infty$.

(iv) *For every* $\alpha \in (0, 1)$, *there exists* $T > 0$ *such that the control system* (1) *is cost-uniformly* α*-null controllable*[1] *in time* T; *i.e.*, *there exists* $C = C(\alpha, T) \geqslant 0$ *such that*, *for every* $y_0 \in X$, *there exists* $u \in L^2(0, T; U)$ *such that*

$$\|y(T; y_0, u)\|_X \leqslant \alpha \|y_0\|_X \quad \text{and} \quad \|u\|_{L^2(0,T;U)} \leqslant C\|y_0\|_X. \tag{8}$$

(v) *There exist* $\alpha \in (0, 1)$ *and* $T > 0$ *such that the control system* (1) *is cost-uniformly* α*-null controllable in time* T.

(vi) *For every* $\alpha \in (0, 1)$, *there exist* $T > 0$ *and* $C \geqslant 0$ *such that*

$$\|S(T)^*\psi\|_X \leqslant C\|B^*S(T - \cdot)^*\psi\|_{L^2(0,T;U)} + \alpha\|\psi\|_X \quad \text{for all } \psi \in X. \tag{9}$$

(vii) *There exist* $\alpha \in (0, 1)$, $T > 0$ *and* $C \geqslant 0$ *such that inequality* (9) *is satisfied.*

When one of these items is satisfied, *the smallest possible constant* C *in* (8) *and in the observability inequality* (9) *is* $C = \mu_\alpha^T$; *moreover*, *for every* $\alpha \in (0, 1)$, *the real number* $T > 0$ *in* (ii), (iii) *and* (iv) *above can be taken to be the same.*

Furthermore, *the best stabilization decay rate defined by* (4) *is*

$$\omega^* = \inf\left\{ \frac{\ln \alpha}{T} \;\middle|\; (\alpha, T) \in \mathcal{A} \right\} = \lim_{\alpha \to 0+} \inf\left\{ \frac{\ln \alpha}{T} \;\middle|\; T \in \mathcal{T}(\alpha) \right\}, \tag{10}$$

where

$$\begin{aligned} \mathcal{A} &= \{(\alpha, T) \in (0, 1) \times (0, +\infty) \mid \mu_\alpha^T < +\infty\}, \\ \mathcal{T}(\alpha) &= \{T > 0 \mid \mu_\alpha^T < +\infty\} \quad \text{for all } \alpha \in (0, 1). \end{aligned} \tag{11}$$

Theorem 1 will be proved in Section 4.

Remark 2. If the semigroup $(S(t))_{t \geqslant 0}$ is exponentially stable then the observability inequality (9) is obviously satisfied with $B = 0$. Indeed, the semigroup $(S(t))_{t \geqslant 0}$ is exponentially stable if and only if there exists $T > 0$ such that $\|S(T)\|_{L(X)} < 1$. This is in accordance with the fact that, in this case, no control is required to stabilize the control system.

[1] This definition and some characterizations will be given with more details in Section 2.1.

When the semigroup is not exponentially stable, the observability inequality (9) can be seen as a weakened version of the observability inequality corresponding to exact null controllability (see (14) in Remark 8), by adding the term $\alpha\|\psi\|_X$ at the right-hand side for some $\alpha \in (0,1)$: this appears as a kind of compromise between the lack of exponential stability of $(S(t))_{t\geqslant 0}$ and the feedback action needed to exponentially stabilize the control system (1).

Remark 3. The equality (10) is comparable with the results on the stability rate given in [Engel and Nagel 2000, Chapter 4, Proposition 2.2].

By the second item of Theorem 1, if the control system (1) is exponentially stabilizable, then

$$\{\alpha \mid (\alpha, T) \in \mathcal{A} \text{ for some } T > 0\} = (0,1).$$

A number of comments, in relation to other controllability concepts, are provided in Sections 2 and 3. It can already be noted that, in the weak observability inequality (9) which characterizes the stabilizability property, the coefficient α satisfies $0 < \alpha < 1$. The limit case $\alpha = 1$ is critical. Also, it is interesting to underline that, in some sense, the constant μ_α^T quantifies the stabilizability property.

The proof of Theorem 1 follows a series of easy arguments essentially exploiting Fenchel duality. In turn, these arguments allow us to obtain characterizations, in terms of observability inequalities, of the concepts of α-null controllability and of approximate null controllability, that we gather in the next section.

2. Several results on null and approximate controllability

2.1. *α-null controllability.*

Let $T > 0$ and $\alpha \geqslant 0$ be arbitrary.

Definition 4. Given some $y_0 \in X$, the control system (1) is *α-null controllable from y_0 in time T* if there exists $u \in L^2(0,T;U)$ such that $\|y(T;y_0,u)\|_X \leqslant \alpha\|y_0\|_X$, i.e., $y(T;y_0,u) \in \alpha\|y_0\|_X\mathcal{B}$, where $\mathcal{B}$ is the closed unit ball in X.

The control system (1) is *α-null controllable in time T* if, for every $y_0 \in X$, the system is α-null controllable from y_0 in time T.

The control system (1) is *cost-uniformly α-null controllable in time T* if there exists $C = C(\alpha, T) \geqslant 0$ such that, for every $y_0 \in X$, there exists $u \in L^2(0,T;U)$ such that $\|y(T;y_0,u)\| \leqslant \alpha\|y_0\|_X$ and $\|u\|_{L^2(0,T;U)} \leqslant C\|y_0\|_X$.

Note that, for $\alpha = 0$, the notion of 0-null controllability coincides with the usual notion of exact null controllability. We have the following results.

Proposition 5. *Let $T > 0$, $\alpha \geqslant 0$ and $y_0 \in X$ be arbitrary. The following items are equivalent*:

(i) *The control system* (1) *is α-null controllable from y_0 in time $T > 0$.*

(ii) *We have $\mu_{y_0,\alpha}^T < +\infty$.*

(iii) *There exists $C \geqslant 0$ such that*

$$\langle \psi, S(T)y_0\rangle_X \leqslant C\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)} + \alpha\|y_0\|_X\|\psi\|_X \quad \text{for all } \psi \in X. \tag{12}$$

When one of these items is satisfied, the smallest possible constant C in the observability inequality (12) *is $C = \mu^T_{y_0,\alpha}$. Moreover, $\mu^T_{y_0,\alpha}$ is the minimal L^2 norm of the control required to steer the control system* (1) *from y_0 to the target set $\alpha\|y_0\|_X\mathcal{B}$.*

Proposition 6. *Let $T > 0$ and $\alpha \geqslant 0$ be arbitrary. The following items are equivalent*:

(i) *The control system* (1) *is cost-uniformly α-null controllable in time $T > 0$.*

(ii) *We have $\mu^T_\alpha < +\infty$.*

(iii) *There exists $C \geqslant 0$ such that*

$$\|S(T)^*\psi\|_X \leqslant C\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)} + \alpha\|\psi\|_X \quad \textit{for all } \psi \in X. \tag{13}$$

When one of these items is satisfied, the smallest possible constant C in the observability inequality (13) *is $C = \mu^T_\alpha$.*

Propositions 5 and 6 will be proved in Section 4.

Remark 7. The control system (1) is exponentially stabilizable if and only if there exists (or, for every) $\alpha \in (0,1)$, there exists $T > 0$ such that the control system (1) is cost-uniformly α-null controllable in time T (i.e., $\mu^T_\alpha < +\infty$).

Remark 8. As said above, for $\alpha = 0$ we recover the usual notion of exact null controllability. Recall that the control system (1) is exactly null controllable in time $T > 0$ if, for every $y_0 \in X$, there exists $u \in L^2(0,T;U)$ such that $y(T;y_0,u) = 0$. This is equivalent to $\mathrm{Ran}(S(T)) \subset \mathrm{Ran}(L_T)$ (see (2)), and also, by duality, to the observability inequality

$$\|S(T)^*\psi\|_X \leqslant \mu^T_0\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)} \quad \text{for all } \psi \in X, \tag{14}$$

which is (13) with $\alpha = 0$.

Remark 9. If the control system (1) is exactly null controllable in time T then $\mu^T_\alpha < +\infty$ for every $\alpha > 0$, and μ^T_α has a limit as $\alpha \to 0^+$, denoted by μ^T_0 (these facts are easily seen by considering the optimal controls $\bar{u}_{y_0,\alpha}$). In particular, we have the observability inequality

$$\langle\psi, S(T)y_0\rangle_X \leqslant \mu^T_0\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)}$$

for every $\psi \in X$ and every $y_0 \in X$ of norm 1. Taking $y_0 = S(T)^*\psi/\|S(T)^*\psi\|_X$, we recover the observability inequality (14). Actually

$$\text{exact null controllable in time } T \quad \Longleftrightarrow \quad \mu^T_0 = \lim_{\alpha\to 0^+}\mu^T_\alpha < +\infty.$$

Remark 10. We claim that, for $\alpha = 0$,

$$\text{cost-uniformly 0-null controllable in time } T \quad \Longleftrightarrow \quad \text{0-null controllable in time } T,$$

but for $\alpha > 0$ we only have

$$\text{cost-uniformly } \alpha\text{-null controllable in time } T \quad \begin{matrix}\Rightarrow \\ \not\Leftarrow\end{matrix} \quad \alpha\text{-null controllable in time } T.$$

Indeed, the first claim is a classical fact of the HUM theory (see [Lions 1988]; see also [Boyer 2013, Proposition 1.19] where the uniform boundedness principle is used to get the result).

In the second claim, the converse is wrong because it may happen that $\mu^T_{y_0,\alpha} < +\infty$ for every $y_0 \in X$, while $\mu^T_\alpha = \sup_{\|y_0\|_X=1} \mu^T_{y_0,\alpha} = +\infty$: see an example in Section 3.2.1. See also Remark 19.

Remark 11. Let $\alpha \geqslant 0$ and $T > 0$ be fixed. Summing up, we have seen that

- α-null controllable from y_0 in time $T \iff \mu^T_{y_0,\alpha} < +\infty$,
- α-null controllable in time $T \iff$ for all $y_0 \in X$, $\mu^T_{y_0,\alpha} < +\infty$,
- cost-uniformly α-null controllable in time $T \iff \mu^T_\alpha < +\infty$,

and that none of these properties are equivalent when $\alpha > 0$. We have also seen that

the system is exponentially stabilizable

$\iff \forall \alpha \in (0,1)$, $\exists T > 0$ such that the system is cost-uniform α-null controllable in time T

$\iff \forall \alpha \in (0,1)$, $\exists T > 0$ such that $\mu^T_\alpha < +\infty$

$\iff \exists \alpha \in (0,1)$, $\exists T > 0$ such that the system is cost-uniform α-null controllable in time T

$\iff \exists \alpha \in (0,1)$, $\exists T > 0$ such that $\mu^T_\alpha < +\infty$.

Remark 12. The constant $\mu^T_{y_0,\alpha}$ quantifies the α-null controllability property: actually, as established in the proof in Section 4.1, when $\mu^T_{y_0,\alpha} < +\infty$ we have

$$\mu^T_{y_0,\alpha} = \|\bar{u}_{y_0,\alpha}\|_{L^2(0,T;U)},$$

where $\bar{u}_{y_0,\alpha}$ is the (unique) control of minimal L^2 norm steering in time T the control system (1) from y_0 to the ball $\alpha\|y_0\|_X\mathcal{B}$.

2.2. ***Approximate controllability.*** Let $T > 0$ be arbitrary.

Definition 13. Given some $y_0 \in X$, the control system (1) is *approximately null controllable from y_0 in time T* if, for every $\alpha > 0$, the system is α-null controllable from y_0 in time T. Equivalently, for every $\varepsilon > 0$ there exists $u \in L^2(0,T;U)$ such that $\|y(T;y_0,u)\|_X \leqslant \varepsilon$.

The control system (1) is *approximately null controllable in time T* if, for every $y_0 \in X$, it is approximately null controllable from y_0 in time T.

Proposition 14. *Let $y_0 \in X$ be arbitrary. The following items are equivalent*:

(i) *The control system* (1) *is approximately null controllable from y_0 in time $T > 0$.*

(ii) *For every $\alpha > 0$, we have $\mu^T_{y_0,\alpha} < +\infty$.*

(iii) *For every $\alpha > 0$, there exists $C \geqslant 0$ such that*

$$\langle \psi, S(T)y_0\rangle_X \leqslant C\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)} + \alpha\|y_0\|_X\|\psi\|_X \quad \text{for all } \psi \in X. \tag{15}$$

(iv) *$B^*S(T-t)^*\psi = 0$ for all $t \in [0,T]$ implies $\langle \psi, S(T)y_0\rangle = 0$.*

When one of these items is satisfied, the smallest possible constant C in the observability inequality (15) *is $C = \mu^T_{y_0,\alpha}$.*

Equivalently, one can replace "every $\alpha > 0$" with "every $\alpha \in (0, \|S(T)\|_{L(X)})$" in Proposition 14. This proposition will be proved in Section 4.

Remark 15. This proposition is not new and can be found, in a slightly different form, in [Boyer 2013].

Proposition 14(iv) (unique continuation property) was suggested by the referee, for which we are grateful. The proof of the equivalence of (iii) and (iv) follows [Boyer 2013, Proposition 1.17]: (iii) obviously implies (iv) by taking $\alpha \to 0$. That (iv) implies (iii) is proved by contradiction: If the inequality (15) does not hold, then there exists $\alpha > 0$ such that, for every $n \in \mathbb{N}^*$, there exists $\psi_n \in X$ such that $\langle \psi_n, S(T)y_0\rangle_X = 1 > n\|B^*S(T-\cdot)^*\psi_n\|_{L^2(0,T;U)} + \alpha\|y_0\|_X\|\psi_n\|_X$. Hence ψ_n is bounded and hence, up to some subsequence, it converges weakly to some $\psi \in X$, which must satisfy $\langle \psi, S(T)y_0\rangle_X = 1$ (and thus $\psi \neq 0$) and also $B^*S(T-\cdot)^*\psi = 0$ on $[0, T]$, whence $\psi = 0$, which raises a contradiction.

It is interesting to note that the unique continuation property (Proposition 14(iv)) is a qualitative way of expressing approximate controllability, while the observability inequality (15) is a quantitative way of expressing it: the constant $\mu^T_{y_0,\alpha}$ quantifies approximate controllability — it gives an account for the "quality" of the approximate controllability property.

Remark 16. The control system (1) is approximately null controllable in time $T > 0$ if and only if $\mathrm{Ran}(S(T))$ (or its closure) is contained in the closure of $\mathrm{Ran}(L_T)$, or, by duality, given any $\psi \in X$, if $B^*S(T-t)^*\psi = 0$ for every $t \in [0, T]$ then $S(T)^*\psi = 0$; see [Zabczyk 1995, Theorem 2.1, page 207].

Remark 17. Until now, we have spoken only of approximate null controllability. Let us comment about the more usual concept of approximate controllability.

The control system (1) is *approximately controllable in time* $T > 0$ if, for every $\varepsilon > 0$, for all $y_0, y_1 \in X$, there exists $u \in L^2(0, T; U)$ such that $\|y(T; y_0, u) - y_1\|_X \leqslant \varepsilon$.

Equivalently, $\mathrm{Ran}(L_T)$ is dense in X, or, by duality, L_T^* is injective, which means that, given any $\psi \in X$, if $B^*S(T-t)^*\psi = 0$ for every $t \in [0, T]$ then $\psi = 0$ (unique continuation property; compare with the one given in Remark 16 and with the one given in Proposition 14).

It is interesting to note the following result:

> *Let $T > 0$ be arbitrary. Assume that $S(T)^*$ is injective, i.e., that* $\mathrm{Ran}(S(T))$ *is dense in X. Then approximate controllability in time T is equivalent to approximate null controllability in time T.*

The assumption that $S(T)^*$ is injective is satisfied when $(S(t))_{t\geqslant 0}$ is either a group (obviously) or an analytic C_0 semigroup.

To prove the latter fact, by taking the adjoint, let us prove that if $(S(t))_{t\geqslant 0}$ is an analytic semigroup then for every $T \geqslant 0$ the operator $S(T)$ is injective: Let $T \geqslant 0$ and $x \in X$ be such that $S(T)x = 0$. Then $S(t)x = S(t-T)S(T)x = 0$ for every $t \geqslant T$, and by analyticity we infer that $S(t)x = 0$ for every $t \geqslant 0$, whence $x = 0$.

In contrast, when the semigroup is neither a group nor analytic, $S(T)$ may fail to be injective: for instance, the left-shift semigroup on the positive half-line is such that, for every y_0, there exists $T = T(y_0)$ such that $S(T)y_0 = 0$.

Remark 18. By the same approach, we obtain as well the following characterization of approximate controllability by an observability inequality:

Define $\mu^T_{y_0,y_1,\alpha} \in [0,+\infty]$ similarly to $\mu^T_{y_0,\alpha}$, replacing the term $S(T)y_0$ with $S(T)y_0 - y_1$. The control system (1) *is approximately controllable in time $T>0$ if and only if, for all $y_0, y_1 \in X$ and for every $\alpha > 0$, there exists $C \geqslant 0$ such that*

$$\langle \psi, S(T)y_0 - y_1\rangle_X \leqslant C\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)} + \alpha\|y_0\|_X\|\psi\|_X \quad \text{for all } \psi \in X.$$

When $\mu^T_{y_0,y_1,\alpha} < +\infty$, it is the smallest constant C in the observability inequality above.

This statement underlines in an unusual way the difference between approximate controllability and approximate null controllability (as said in Remark 17, both notions coincide if $S(T)^*$ is injective, or equivalently if $\mathrm{Ran}(S(T))$ is dense in X).

Similar statements can be given as well for α-controllability to some target point y_1. We do not give details.

Remark 19. There is no relationship in general between approximate controllability and exponential stabilizability:

• There exist control systems that are exponentially stabilizable but that are not approximately null controllable in any time T.

For instance, take $B=0$ and take A generating an exponentially stable semigroup, i.e., $\|S(t)\|_{L(X)} \leqslant Me^{-\beta t}$ for some $M \geqslant 1$ and $\beta > 0$. Then a minimal time $T_\alpha = (1/\beta)\ln(M/\alpha)$ is at least required to realize α-null controllability, which cannot be bounded uniformly with respect to every $\alpha > 0$ arbitrarily small.

• There exist control systems that are approximately controllable in some time T but that are not exponentially stabilizable (see [Curtain and Zwart 1995, Example 5.2.2, page 228; Pritchard and Zabczyk 1981, Example 3.16; Triggiani 1975; Zabczyk 1995, Theorem 3.3(ii), page 227], which all give the same example; see also the example provided in Section 3.2.1, already commented on in Remark 10).

We mention however [Badra and Takahashi 2014, Theorem 1.6], which states that, when $(S(t))_{t\geqslant 0}$ is analytic, under some spectral assumptions, approximate controllability is equivalent to exponential stabilizability and to a Fattorini–Hautus criterion.[2]

Remark 20. Given $T>0$, recall that $\mu^T_\alpha = \mu^T_{y_0,\alpha} = 0$ when $\alpha \geqslant \|S(T)\|_{L(X)}$. When $\alpha \to 0^+$, μ^T_α and $\mu^T_{y_0,\alpha}$ may tend to $+\infty$. Assuming that μ^T_α is uniformly bounded with respect $\alpha \in (0, \|S(T)\|_{L(X)}]$, approximate null controllability in time T is equivalent to exact null controllability in time T.

3. Further comments

3.1. ***Null controllability implies stabilizability.*** Inspecting the observability inequalities (14) and (9), we recover the well known fact that if the control system (1) is exactly null controllable in some time T then it is exponentially stabilizable; see, e.g., [Zabczyk 1995, Theorem 3.3, page 227]. The converse is wrong in general: as said in Remark 9, the control system (1) is exactly null controllable in time T if and only if

[2] We thank Guillaume Olive for having indicated this reference to us.

$\mu_0^T = \lim_{\alpha\to 0^+} \mu_\alpha^T < +\infty$; it may happen that when the system is exponentially stabilizable, as $\alpha \to 0^+$, the infimum of times T such that $\mu_\alpha^T < +\infty$ tends to $+\infty$.

Complete stabilizability. Complete stabilizability means that, given any $\omega \in \mathbb{R}$, one can find a feedback $K \in L(X, U)$ such that $\omega_K < \omega$, or equivalently, that the best stabilization rate given by (10) is $-\infty$. According to the expression (10), we have complete stabilizability if either, for a given $\alpha \in (0, 1)$, we have $\mu_\alpha^T < +\infty$ for every $T > 0$ arbitrarily small, or, for a given $T > 0$, we have $\mu_\alpha^T < +\infty$ for every $\alpha > 0$ arbitrarily small (which is equivalent, by Remark 9, to exact null controllability in time T). Several remarks on complete stabilizability are in order.

Proposition 21. *If the control system* (1) *is exactly null controllable in some time $T > 0$ then it is completely stabilizable.*

This result is proved in Appendix A.1. It also follows by Remark 9: since μ_α^T remains uniformly bounded as $\alpha \to 0$ for some $T > 0$ fixed, we see from (10) that $\omega^* = -\infty$.

Proposition 22. *When $(S(t))_{t\geqslant 0}$ is a group, the following properties are equivalent*:

(i) *Exact controllability in some time T.*

(ii) *Exact null controllability in some time T.*

(iii) *Complete stabilizability.*

This result is already known; see [Slemrod 1974; Urquiza 2005; Zabczyk 1995, Theorem 3.4, page 229]. We provide in Appendix A.2 a proof of it that uses Theorem 1.

The strategy developed in [Komornik 1997] (applying also, to some extent, to unbounded admissible control operators) consists of taking $K_\lambda = -B^* C_\lambda^{-1}$, where C_λ is defined by

$$C_\lambda = \int_0^{T+1/(2\lambda)} f_\lambda(t) S(-t) B B^* S(-t)^*\, dt$$

(variant of the Gramian operator), with $\lambda > 0$ arbitrary,

$$f_\lambda(t) = \begin{cases} e^{-2\lambda t} & \text{if } t \in [0, T], \\ 2\lambda e^{-2\lambda T}(T + 1/(2\lambda) - t) & \text{if } t \in [T, T + 1/(2\lambda)]. \end{cases}$$

The function $V(y) = \langle y, C_\lambda^{-1} y\rangle$ is a Lyapunov function (as noticed in [Coron 2007]), and the feedback K_λ yields exponential stability with rate $-\lambda$. We also refer to [Coron and Lü 2014; Coron and Trélat 2004; Phung et al. 2017; Russell 1978] for issues on rapid stabilization.

Let us assume that A is skew-adjoint, which is equivalent, by the Stone theorem, to the fact that A generates a unitary group $(S(t))_{t\in\mathbb{R}}$; see [Engel and Nagel 2000]. Then $\|S(T)^*\psi\|_X = \|\psi\|_X$ for every $\psi \in X$ and therefore the observability inequality (9) characterizing exponential stabilizability is equivalent to the observability inequality characterizing exact controllability and can be achieved, for a given $T > 0$, with arbitrarily small values of $\alpha > 0$, which implies that $\omega^* = -\infty$. Therefore we recover a result of [Liu 1997]:

Proposition 23. *When A is skew-adjoint, the following properties are equivalent*:

(i) *exact controllability in time T*,

(ii) *exact null controllability in time T*,

(iii) *exponential stabilizability*,

(iv) *complete stabilizability.*

3.2. *Examples.*

3.2.1. *First example.* We take $X = U = \ell^2(\mathbb{N}, \mathbb{R})$, A the infinite-dimensional identity matrix, and B the diagonal infinite-dimensional matrix $B = \operatorname{diag}\big(1, \frac{1}{2}, \frac{1}{3}, \ldots\big)$. We claim that, with this choice:

- For every $T > 0$ and every $\alpha \in (0, 1)$, the control system (1) is α-null controllable in time T; i.e.,

$$\mu^T_{y_0,\alpha} < +\infty \quad \text{for all } y_0 \in X, \text{ for all } \alpha \in (0, 1), \text{ for all } T > 0.$$

Hence, the control system (1) is approximately null controllable in any time $T > 0$.

- The control system (1) is not exponentially stabilizable; i.e.,

$$\mu^T_\alpha = +\infty \quad \text{for all } \alpha \in (0, 1), \text{ for all } T > 0.$$

By Remark 7, the control system (1) is not cost-uniformly α-null controllable in time T.

Let us first prove the α-null controllability property. Let $(e_n)_{n\in\mathbb{N}^*}$ be the canonical base of X. For every $n \in \mathbb{N}^*$, we denote by P_n the orthogonal projection of X onto $\operatorname{Span}(e_1, \ldots, e_n)$. Let $y_0 \in X$ be arbitrary. Taking $n \in \mathbb{N}^*$ large enough such that $\|y_0 - P_n y_0\|_X \leqslant \alpha e^{-T} \|y_0\|_X$, and taking the control $u = 0$, we have

$$\|y(T; (\mathrm{id} - P_n) y_0, 0)\|_X \leqslant \alpha \|y_0\|_X. \tag{16}$$

By the Duhamel formula, for every $u \in L^2(0, T; U)$ we have

$$y(T; y_0, u) = y(T; P_n y_0, u) + y(T; (\mathrm{id} - P_n) y_0, 0). \tag{17}$$

Note that, taking u such that $u(t) \in \operatorname{Ran}(P_n)$ for almost every $t \in [0, T]$, we have $y(t; P_n y_0, u) \in \operatorname{Ran}(P_n)$ for every $t \in [0, T]$. Since the control system in $\mathbb{R}^n$ given by $\dot{x}(t) = A_n x(t) + B_n u(t)$, with A_n the identity matrix and $B_n = \operatorname{diag}\big(1, \frac{1}{2}, \ldots, \frac{1}{n}\big)$, is controllable (it satisfies the Kalman condition), there exists $u \in L^2(0, T; U)$ such that $y(T; P_n y_0, u) = 0$. Using (17) and (16), it follows that $\|y(T; y_0, u)\|_X \leqslant \alpha \|y_0\|_X$; i.e., we have proved that the system is α-null controllable in time T.

Let us now prove that the system is not exponentially stabilizable. By contradiction, let us assume that there exist $K \in L(X, U)$, $M > 0$ and $\rho > 0$ such that $\|S_K(t)\| \leqslant M e^{-\rho t}$ for every $t \geqslant 0$ (we use the notation of (3)). Denoting by $a_n(t)$ the n-th component of $y_K(t; y_0) = S_K(t) y_0$, we have

$$\dot{a}_n(t) = a_n(t) + \langle e_n, BK y_K(t; y_0)\rangle = a_n(t) + \langle Be_n, K y_K(t; y_0)\rangle.$$

Hence,

$$\begin{aligned}|a_n(t)| &= \left| e^t a_n(0) + \int_0^t e^{t-s} \langle Be_n, Ky_K(s; y_0)\rangle \, ds \right| \\ &\geqslant e^t |a_n(0)| - \int_0^t e^{t-s} \|Be_n\| \|K\| \|y_K(s; y_0)\| \, ds \\ &\geqslant e^t |a_n(0)| - \int_0^t e^{t-s} \frac{1}{n} \|K\| M e^{-\rho s} \|y_0\| \, ds \geqslant e^t \left(|a_n(0)| - \frac{M\|K\|}{(1+\rho)n} \|y_0\| \right). \end{aligned} \tag{18}$$

Take $m \in \mathbb{N}$ satisfying

$$\frac{M\|K\|}{(1+\rho)m} \leqslant \frac{1}{2}$$

and take $y_0 = e_m$. Then $a_m(0) = \|y_0\| = 1$. It follows from (18) with $n = m$ that

$$\|a_m(t)\| \geq e^t \left(1 - \frac{M\|K\|}{(1+\rho)m} \right) \geq \frac{1}{2} e^t.$$

Since $\lim_{m\to+\infty} |a_m(t)| = +\infty$, we obtain that $\|y_K(t; y_0)\| \to +\infty$, which is a contradiction.

3.2.2. *Second example.* Consider the coupled control system with Dirichlet boundary conditions

$$\begin{aligned} \partial_{tt} z &= \triangle z + w + u, \\ \partial_t w &= \triangle w + w + \frac{2}{\pi} \sin x \int_0^\pi (\partial_t z(t,s) + u(t,s)) \sin s \, ds, \\ z(t,0) = z(t,\pi) &= w(t,0) = w(0,\pi) = 0, \end{aligned} \tag{19}$$

where $z = z(t,x)$, $w = w(t,x)$, and $u = u(t,x)$ for $t \geqslant 0$ and $x \in (0,\pi)$. We take initial data $z(0,\cdot) = z_0 \in H_0^1(0,\pi)$, $\partial_t z(0,\cdot) = z_1 \in L^2(0,\pi)$, and $w(0,\cdot) = w_0 \in L^2(0,\pi)$.

To write (19) in the form (1), we define $X = H_0^1(0,\pi) \times L^2(0,\pi) \times L^2(0,\pi)$, $U = L^2(0,\pi)$, and

$$y = \begin{pmatrix} z \\ \partial_t z \\ w \end{pmatrix}, \quad A = \begin{pmatrix} 0 & \mathrm{id} & 0 \\ \triangle & 0 & \mathrm{id} \\ 0 & P & \triangle + \mathrm{id} \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \\ P \end{pmatrix},$$

where $D(A) = (H^2(0,\pi) \cap H_0^1(0,\pi)) \times H_0^1(0,\pi) \times (H^2(0,\pi) \cap H_0^1(0,\pi))$ and P is the selfadjoint bounded operator defined on $L^2(0,\pi)$ by

$$(Pv)(x) = \frac{2}{\pi} \sin x \int_0^\pi v(s) \sin s \, ds \quad \text{for all } v \in L^2(0,\pi).$$

Note that $B \in L(X,U)$ is a bounded control operator. We claim that:

- With $u = 0$, the control system (19) is not asymptotically stable.
- The control system (19) satisfies the observability inequality (9) of Theorem 1, and hence it is stabilizable.
- The control system (19) is not exactly null controllable in any time $T > 0$. In particular, it is not completely stabilizable.

The fact that, with $u = 0$, the control system (19) is not asymptotically stable, can be seen from the fact that $(z, \partial_t z, w) = (\sin x, 0, \sin x)$ is a steady-state solution to (19) with $u = 0$.

Let us prove that the control system (19) is not exactly null controllable in any time $T > 0$. By contradiction, if there were some $T > 0$ such that the system (19) is exactly null controllable, then by taking the initial data $(z_0, z_1, w_0) = (0, 0, \sin 2x)$, we could find a control $u \in L^2(0, T; L^2(0, \pi))$ steering the system (19) from (z_0, z_1, w_0) to zero in time T. However, setting $a(t) = \int_0^\pi w(t, x) \sin(2x)\, dx$, from the second equation of (19), we obtain

$$\dot{a}(t) = \langle w(t, x), (\triangle + \mathrm{id}) \sin(2x) \rangle_{L^2(0,\pi)} = -3\langle w(t, x), \sin(2x) \rangle_{L^2(0,\pi)} = -3a(t).$$

Since $a(0) = \frac{\pi}{2} \neq 0$, we must have $a(T) \neq 0$, which raises a contradiction.

Let us finally prove that the control system (19) is stabilizable. By Theorem 1, it suffices to establish:

There exist $\alpha \in (0, 1)$, $T > 0$, and $C > 0$ such that

$$\|S(T)^*\Psi\|_X^2 \leqslant C\|B^*S(T - \cdot)^*\Psi\|_{L^2(0,\pi;U)}^2 + \alpha^2\|\Psi\|_X^2 \quad \textit{for all } \Psi \in X. \tag{20}$$

To prove (20), we first observe that $D(A^*) = D(A)$ and

$$A^* = \begin{pmatrix} 0 & -\mathrm{id} & 0 \\ -\triangle & 0 & P \\ 0 & \mathrm{id} & \triangle + \mathrm{id} \end{pmatrix}, \quad B^* \begin{pmatrix} \phi \\ \psi \\ \xi \end{pmatrix} = \psi + P\xi.$$

We define the adjoint system

$$\begin{aligned} \partial_t \phi &= \psi, \\ \partial_t \psi &= \triangle\phi - P\xi, \\ \partial_t \xi &= -\psi - \triangle\xi - \xi, \end{aligned} \tag{21}$$

with $(\phi(T), \psi(T), \xi(T)) = (\phi_T, \psi_T, \xi_T) \in D(A^*)$. Inequality (20) is then equivalent to

$$\left\| \begin{pmatrix} \phi(0) \\ \psi(0) \\ \xi(0) \end{pmatrix} \right\|_X^2 \leqslant C \int_0^\pi \|\psi(t) + P\xi(t)\|^2\, dt + \alpha^2 \left\| \begin{pmatrix} \phi(T) \\ \psi(T) \\ \xi(T) \end{pmatrix} \right\|_X^2 \quad \text{for all } (\phi_T, \psi_T, \xi_T) \in D(A^*). \tag{22}$$

Let us establish (22) for $T = \pi$, $\alpha = \sqrt{2}e^{-3T} < 1$, and for some $C > 0$ which will be given later. To this end, we arbitrarily fix $(\phi_T, \psi_T, \xi_T) \in D(A^*)$. We write

$$\begin{pmatrix} \phi_T \\ \psi_T \\ \xi_T \end{pmatrix} = \begin{pmatrix} P\phi_T \\ P\psi_T \\ P\xi_T \end{pmatrix} + \begin{pmatrix} (\mathrm{id}-P)\phi_T \\ (\mathrm{id}-P)\psi_T \\ (\mathrm{id}-P)\xi_T \end{pmatrix} = \begin{pmatrix} \phi_{T,1} \\ \psi_{T,1} \\ \xi_{T,1} \end{pmatrix} + \begin{pmatrix} \phi_{T,2} \\ \psi_{T,2} \\ \xi_{T,2} \end{pmatrix}.$$

Denote respectively by

$$\begin{pmatrix} \phi(\cdot) \\ \psi(\cdot) \\ \xi(\cdot) \end{pmatrix}, \quad \begin{pmatrix} \phi_1(\cdot) \\ \psi_1(\cdot) \\ \xi_1(\cdot) \end{pmatrix}, \quad \begin{pmatrix} \phi_2(\cdot) \\ \psi_2(\cdot) \\ \xi_2(\cdot) \end{pmatrix}$$

the solutions to (21) with final data (at time T)

$$\begin{pmatrix}\phi_T\\ \psi_T\\ \xi_T\end{pmatrix},\quad \begin{pmatrix}\phi_{T,1}\\ \psi_{T,1}\\ \xi_{T,1}\end{pmatrix},\quad \begin{pmatrix}\phi_{T,2}\\ \psi_{T,2}\\ \xi_{T,2}\end{pmatrix}.$$

Then we have

$$\begin{pmatrix}\phi\\ \psi\\ \xi\end{pmatrix}=\begin{pmatrix}\phi_1\\ \psi_1\\ \xi_1\end{pmatrix}+\begin{pmatrix}\phi_2\\ \psi_2\\ \xi_2\end{pmatrix}.$$

Since $(\phi_1(t,\cdot),\psi_1(t,\cdot),\xi_1(t,\cdot))\in \operatorname{Span}(\sin x)$ and since the pair

$$\left(\begin{pmatrix}0&1&0\\-1&0&-1\\0&-1&0\end{pmatrix},(0,1,1)\right)$$

satisfies the (observability) Kalman condition, there exists $C_1>0$ independent of the final data such that

$$\left\|\begin{pmatrix}\phi_1(0)\\ \psi_1(0)\\ \xi_1(0)\end{pmatrix}\right\|_X^2\leqslant C_1\int_0^\pi \|\psi_1(t,\cdot)+\xi_1(t,\cdot)\|^2\,dt=C_1\int_0^\pi \|\psi_1(t,\cdot)+P\xi_1(t,\cdot)\|^2\,dt. \tag{23}$$

Now, since $(\phi_2(t,\cdot),\psi_2(t,\cdot),\xi_2(t,\cdot))\in(\operatorname{Span}(\sin x))^\perp$ and $P\xi_2(t,\cdot)=0$, it follows from (21) that

$$\begin{aligned}\partial_t\phi_2=\psi_2,\qquad \partial_t\psi_2=\triangle\phi_2,\\ \phi_2(T)=\phi_{T,2},\quad \psi_2(T)=\psi_{T,2}.\end{aligned} \tag{24}$$

From (24) and by observability of the wave equation (which is true because $T>\pi$), there exists $C_2>0$ independent of the final data such that

$$\left\|\begin{pmatrix}\phi_2(0)\\ \psi_2(0)\end{pmatrix}\right\|_{H_0^1\times L^2}^2\leqslant C_2\int_0^\pi \|\psi_2(t,\cdot)\|^2\,dt=C_2\int_0^\pi \|\psi_2(t,\cdot)+P\xi_2(t,\cdot)\|^2\,dt. \tag{25}$$

Additionally, it follows from the third equation of (21) that

$$\xi_2(0)=e^{T(\triangle+\mathrm{id})}\xi_2(T)+\int_0^T e^{t(\triangle+\mathrm{id})}\psi_2(t,\cdot)\,dt,$$

which implies

$$\begin{aligned}\|\xi_2(0)\|^2&\leqslant\left(e^{-3T}\|\xi_2(T)\|+\int_0^T\|\psi_2(t,\cdot)\|\,dt\right)^2\\ &\leqslant 2e^{-6T}\|\xi_2(T)\|^2+2\left(\int_0^T\|\psi_2(t,\cdot)\|\,dt\right)^2\\ &\leqslant 2e^{-6T}\|\xi_2(T)\|^2+2T\int_0^\pi\|\psi_2(t,\cdot)+P\xi_2(t,\cdot)\|^2\,dt.\end{aligned} \tag{26}$$

From (25) and (26), we infer that

$$\left\| \begin{pmatrix} \phi_2(0) \\ \psi_2(0) \\ \xi_2(0) \end{pmatrix} \right\|_X^2 \leqslant (C_2 + 2T) \int_0^\pi \|\psi_2(t,\cdot) + P\xi_2(t,\cdot)\|^2 \, dt + 2e^{-6T} \|\xi_2(T)\|^2$$
$$\leqslant (C_2 + 2T) \int_0^\pi \|\psi_2(t,\cdot) + P\xi_2(t,\cdot)\|^2 \, dt + 2e^{-6T} \left\| \begin{pmatrix} \phi_2(T) \\ \psi_2(T) \\ \xi_2(T) \end{pmatrix} \right\|_X^2. \tag{27}$$

The desired inequality (22) follows from (23) and (27), with $T = \pi$, $\alpha = \sqrt{2}e^{-3T} < 1$ and $C = \max(C_1, C_2 + 2T)$.

3.3. ***Extension to unbounded admissible control operators.*** Throughout the paper we have assumed that the control operator B is bounded, i.e., is linear continuous with values in X, that is, $B \in L(U, X)$. This assumption covers the case of internal controls, but not, in general, of boundary controls. For the latter case, we speak of *unbounded control operators*, which are operators B that are not continuous from U to X but are continuous from U to some larger space $X_{-\alpha}$ which can be defined by extrapolation (scale of Hilbert spaces; see [Engel and Nagel 2000; Staffans 2005; Tucsnak and Weiss 2009]). Given such a control operator $B \in L(U, X_{-\alpha})$ with $\alpha > 0$, the range of the operator L_T may then fail to be contained in X. We say that B is *admissible* when $\mathrm{Ran}(L_T) \subset X$ for some (and thus for all) $T > 0$; see [Tucsnak and Weiss 2009]. Note that X_{-1} is isomorphic to $D(A^*)$ (with X as a pivot space) and that if B is admissible then $B \in L(U, X_{-1/2})$.

For an admissible control operator, since $\mathrm{Ran}(L_T) \subset X$, all arguments of Section 4.1 (Fenchel duality) remain valid.[3] One should anyway avoid using $\mathcal{G}_T^{1/2}$ (where $\mathcal{G}_T := \int_0^T S(T-t)BB^*S(T-t)^* \, dt$ is the usual Gramian operator when B is bounded) in this more general context and replace $\|\mathcal{G}_T^{1/2}\psi\|_X$ with $\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)}$ everywhere throughout the proof. Therefore:

Proposition 24. *Propositions 5, 6 and 14 remain true in the more general context where the control operator B is admissible (and may be unbounded).*

When the control operator B is not admissible, the question of knowing whether Propositions 5, 6 and 14 may be extended in some way is open.

Now, concerning the exponential stabilizability result, the only critical fact is at the end of the proof of Lemma 31 where we invoke the Riccati theory: indeed this theory is well established in the general case only for admissible control operators and analytic semigroups (see Remark 32 at the end of the proof in Section 4.2). Therefore:

Theorem 25. *Theorem 1 is true, without any change, in the more general context where the control operator B is admissible (and may be unbounded) and the semigroup $(S(t))_{t\geqslant 0}$ is analytic.*

For instance, this situation covers the case of heat equations in a C^2 bounded open subset $\Omega \subset \mathbb{R}^n$, with $X = L^2(\Omega)$, with Neumann control at the boundary of Ω (for which, at best, $B \in L(U, X_{-1/4-\varepsilon})$

[3] When the control operator is not admissible, we have $\mathrm{Ran}(L_T) \subset X_{-\alpha}$ for some $\alpha > 0$ and then the closed unit ball $\mathcal{B}$ should be the one in $X_{-\alpha}$, while the considered controllability concepts are in the space X.

for every $\varepsilon > 0$), but not with Dirichlet control at the boundary (for which, at best, $B \in L(U, X_{-3/4-\varepsilon})$ for every $\varepsilon > 0$). We refer to [Lasiecka and Triggiani 2000a] for details.

Actually, as indicated to us by Marius Tucsnak, we do not need to use the Riccati operator but only the fact that the finite-cost property (also called optimizability) implies stabilizability, which is true (as well as the converse implication) for a bounded control operator.

Let us recall that the control system (1) is *optimizable* (or, enjoys the *finite-cost property*) if, for every $y_0 \in X$, there exists $u \in L^2(0, +\infty; U)$ such that $y(\cdot; y_0, u) \in L^2(0, +\infty; X)$.

The argument of the proof of Lemma 31 can easily be extended (see Remark 32) to the case of an admissible control operator B (which may be unbounded), and we obtain the following result:

Theorem 26. *For an admissible control operator B (which may be unbounded), the following items are equivalent*:

(i) *The control system* (1) *is optimizable.*

(ii) *For every $\alpha \in (0, 1)$ there exists $T > 0$ such that $\mu_\alpha^T < +\infty$.*

(iii) *There exist $\alpha \in (0, 1)$ and $T > 0$ such that $\mu_\alpha^T < +\infty$.*

(iv) *For every $\alpha \in (0, 1)$, there exist $T > 0$ and $C \geqslant 0$ such that, for every $y_0 \in X$, there exists $u \in L^2(0, T; U)$ such that*

$$\|y(T; y_0, u)\|_X \leqslant \alpha \|y_0\|_X \quad \text{and} \quad \|u\|_{L^2(0,T;U)} \leqslant C\|y_0\|_X. \tag{28}$$

(v) *There exist $\alpha \in (0, 1)$, $T > 0$ and $C \geqslant 0$ such that, for every $y_0 \in X$, there exists $u \in L^2(0, T; U)$ such that inequality* (28) *is satisfied.*

(vi) *For every $\alpha \in (0, 1)$ (equivalently, there exists $\alpha \in (0, 1)$), there exist $T > 0$ and $C \geqslant 0$ such that*

$$\|S(T)^*\psi\|_X \leqslant C\|B^*S(T - \cdot)^*\psi\|_{L^2(0,T;U)} + \alpha\|\psi\|_X \quad \text{for all } \psi \in X. \tag{29}$$

(vii) *There exist $\alpha \in (0, 1)$, $T > 0$ and $C > 0$ such that inequality* (29) *is satisfied.*

When one of these items is satisfied, the smallest possible constant C in (8) *and in the observability inequality* (9) *is $C = \mu_\alpha^T$; moreover, for every $\alpha \in (0, 1)$, the real number $T > 0$ in* (ii), (iii) *and* (iv) *above can be taken to be the same.*

By inspecting, in light of Remark 32, the proof of Lemma 31 and in particular (34), we note that, for an admissible control operator B, any of the items of Theorem 26 is equivalent to the following fact:

> *For every $\alpha \in (0, 1)$, there exist $T > 0$ and $M = M(\alpha, T) > 0$ such that, for every $y_0 \in X$, there exists $u \in L^2(0, +\infty; U)$ such that $\|y(t; y_0, u)\|_X \leqslant M\|y_0\|_X e^{((\ln \alpha)/T)t}$ for every $t \geqslant 0$.*

This is only an *open-loop* stabilizability property (or *asymptotic null controllability*), in the sense that the control u, depending on y_0, is not determined by a feedback. The usual concept of stabilizability underlies the existence of a feedback.

Hence, we should now discuss how the concepts of optimizability and of stabilizability are related to each other. As said above, optimizability is equivalent to exponential stabilizability (defined in (3))

when the control operator B is bounded. For unbounded admissible control operators, the concept of exponential stabilizability is more difficult to define.

The definition of exponential stabilizability given in [Weiss and Rebarber 2000] is the following. Let $B \in L(U, X_{-1})$. The control system (1) is said to be exponentially stabilizable if B is admissible and if there exists $K \in L(D(A), U)$ such that $\mathrm{id} - K_\Lambda(\lambda\,\mathrm{id} - A)^{-1}B$ is boundedly invertible for $\lambda > 0$ large enough (and the inverse is uniformly bounded) and such that the operator $A + BK_\Lambda$, of domain $D(A + BK_\Lambda) = \{y \in D(K_\Lambda) \mid (A + BK_\Lambda)y \in X\}$, generates on X an exponentially stable C_0 semigroup $(S_{K_\Lambda}(t))_{t\geqslant 0}$. Here, $K_\Lambda \in L(D(K_\Lambda), U)$ is the Λ-extension (or Yosida extension) of K, defined by $K_\Lambda y = \lim_{\lambda\to+\infty} K\lambda(\lambda\,\mathrm{id} - A)^{-1}y$ for every $y \in D(K_\Lambda)$. The domain $D(K_\Lambda)$ of K_Λ, which is the set of all $y \in X$ such that the limit (in X) exists, can be equipped with a norm making it a Banach space, and we have $D(A) \subset D(K_\Lambda) \subset X$ with continuous embeddings, and, for every $x \in X$, we have $S(t)x \in D(K_\Lambda)$ and $S_{K_\Lambda}(t)x \in D(K_\Lambda)$ for almost every $t \geqslant 0$. Moreover, we have $S_{K_\Lambda}(t) = S(t) + \int_0^t S(t-s)BK_\Lambda S_{K_\Lambda}(s)\,ds$ for every $t \geqslant 0$, and the control operator B is admissible for the semigroup $(S_{K_\Lambda}(t))_{t\geqslant 0}$.

Exponential stabilizability in the above sense implies optimizability. Indeed, given any $y_0 \in X$, take $u(t) = K_\Lambda S_{K_\Lambda}(t)y_0$, which is in $L^2(0, +\infty; U)$, and note that $y(t; y_0, u) = S_{K_\Lambda}(t)y_0 \in L^2(0, +\infty; X)$.

The converse statement is more involved: it is true but only in a weaker sense. Assume that the control system (1) is optimizable. It is proved in [Weiss and Rebarber 2000, Propositions 3.2, 3.3 and 3.4] (see also [Flandoli et al. 1988; Staffans 2005; Zwart 1996]) that, for every $y_0 \in X$, there exists a unique $\hat{u} \in L^2(0, +\infty; U)$ minimizing the functional

$$J(u, y_0) = \int_0^{+\infty} (\|u(t)\|_U^2 + \|y(t; y_0, u)\|_X^2)\,dt$$

over all possible $u \in L^2(0, +\infty; U)$. Moreover, we have $\hat{u}(t) = \widehat{K}\widehat{S}(t)y_0$ for every $y_0 \in D(\widehat{A})$ and almost every $t \geqslant 0$, where:

- $(\widehat{S}(t))_{t\geqslant 0}$ is an exponentially stable C_0 semigroup on X, of infinitesimal generator $\widehat{A} : D(\widehat{A}) \to X$, such that $\widehat{S}(t)y_0 = S(t)y_0 + \int_0^t S(t-s)B\hat{u}(s)\,ds$ for every $y_0 \in X$ and every $t \geqslant 0$.
- The feedback operator $\widehat{K} \in L(D(\widehat{A}), U)$ is defined by $\widehat{K} = -B^*P$, where $P \in L(X)$ is a positive symmetric definite operator mapping $D(\widehat{A})$ to $D(A^*)$, such that $J(\hat{u}, y_0) = \langle Py_0, y_0\rangle_X$ for every $y_0 \in X$. Note that P satisfies a Riccati equation on $D(\widehat{A})$ and possibly also another one on $D(A)$; see [Weiss and Zwart 1998].
- We have $\widehat{A} = A + B\widehat{K}$ on $D(\widehat{A})$, where A in this formula stands for the extension $A : X \to X_{-1}$ of the original infinitesimal operator $A : D(A) \to X$. Moreover, considering the Λ-extension $\widehat{K}_\Lambda = \lim_{\lambda\to+\infty} \widehat{K}\lambda(\lambda\,\mathrm{id} - A)^{-1}$ of $\widehat{K}$, we have $\hat{u}(t) = \widehat{K}_\Lambda\widehat{S}(t)y_0$ for every $y_0 \in X$ and almost every $t \geqslant 0$.

In contrast to the above definition of exponential stabilizability, however, it is not known whether the control operator B is admissible or not for the semigroup $(\widehat{S}(t))_{t\geqslant 0}$.

We may then define the best stabilization decay rate much as in (4), but, with the above results, we do not know if (10) remains true. We leave this problem as an open question.

3.4. ***Extension to Banach spaces.*** Throughout the paper we have assumed that the state space X and the control space U are Hilbert spaces. All our results can be extended without difficulty to the case where X and U are reflexive Banach spaces. One has to be careful to replace, in all observability inequalities, the scalar product $\langle \cdot, \cdot \rangle_X$ with the duality bracket $\langle \cdot, \cdot \rangle_{X',X}$.

3.5. ***Open problems.*** We end the section with several open issues.

Systems with an observation. Throughout the paper we have focused on the control system (1), without observation. We could add to the system an observation $z(t) = Cy(t)$, where $C \in L(X, Y)$ is an observation operator, with Y another Hilbert space. Corresponding notions of controllability and of stabilizability are classically defined as well. Extending our main results to that context is an interesting issue.

Discretization problems. The observability inequality (9) may certainly be exploited to recover results on uniform semidiscretizations (or full discretizations). The problem is the following. Consider a spatial semidiscrete model of (1), written as

$$\dot{y}_N = A_N y_N + B_N u_N, \tag{30}$$

with $y_N(t) \in \mathbb{R}^N$ (see [Boyer 2013; Labbé and Trélat 2006; Lasiecka and Triggiani 2000a; Lasiecka and Triggiani 2000b] for a general framework on discretization issues). Assuming that the control system (1) is exponentially stabilizable (equivalently, that the observability inequality (9) is satisfied), how can one ensure that the family of control systems (30) is *uniformly* exponentially stabilizable? Uniform means here that, for each N, there exists a feedback matrix K_N such that $\| \exp(t(A_N + B_N K_N)) \| \leqslant Me^{\omega t}$ for some $M \geqslant 1$ and some $\omega < 0$ that are *uniform* with respect to N.

This problem has been much studied in the literature with various approaches. In [Banks and Ito 1997; Lasiecka and Triggiani 2000a; Liu and Zheng 1999], the convergence of the Riccati matrix P_N (corresponding to (30)) to the Riccati operator P (corresponding to (1)) is proved in the general parabolic case, even for unbounded control operators, that is, when $A : D(A) \to X$ generates an analytic semigroup, $B \in L(U, D(A^*)')$, and (A, B) is exponentially stabilizable. Uniform exponential stability is also proved under uniform Huang–Prüss conditions in [Liu and Zheng 1999], allowing them to obtain convergence of Riccati operators for second-order systems $\ddot{y} + Ay = Bu$, with $A : D(A) \to X$ positive selfadjoint with compact inverse and B bounded control operator. General conservative equations are treated in [Ervedoza and Zuazua 2009] where it is proved that adding a viscosity term in the numerical scheme helps to recover a uniform exponential decay, provided a uniform observability inequality holds true for the corresponding conservative equation (see also [Alabau-Boussouira and Ammari 2011; Alabau-Boussouira et al. 2017; Trélat 2018] for equations with nonlinear damping). Of course, given a specific equation, the difficulty is to establish the uniform observability inequality. This is a difficult issue, investigated in some particular cases (see [Zuazua 2005] for a survey).

Anyway, the observability inequality (9) is written in the semidiscrete case as

$$\|S_N(T)^* \psi_N\| \leqslant C \left(\int_0^T \|B_N^* S(T-t)^* \psi_N\|^2 \, dt \right)^{1/2} + \alpha \|\psi_N\|,$$

with $\alpha \in (0,1)$ and $C > 0$ uniform with respect to N. Such a uniform inequality is likely to be true in many cases. For instance it is nicely shown in [Zuazua 2004, Section 5] that approximate boundary controllability for one-dimensional waves is satisfied, in a uniform way, for finite-difference semidiscretization schemes. This is shown by using a functional similar to the one (31) used here (see also [Boyer 2013; Labbé and Trélat 2006] for α-null controllability in the parabolic case where α is a decreasing function of the discretization parameter, and see [Boyer and Le Rousseau 2014; Boyer et al. 2019] for more recent results in the semilinear case).

But such considerations go beyond the scope of the present paper. We think that the observability inequalities derived in our main results may be used, at least to recover some known results on uniform convergence, and maybe to establish new ones.

Hautus test. There exist many results with variants of the Hautus test. For instance, when $(S(t))_{t\geqslant 0}$ is a normal C_0 group, the Hautus test property, there exists $C > 0$ such that

$$\|(\lambda \,\mathrm{id} - A^*)\psi\|_X^2 + |\operatorname{Re}(\lambda)|^2 \|B^*\psi\|_X^2 \geqslant C|\operatorname{Re}(\lambda)|^2\|\psi\|_X^2 \quad \text{for all } \lambda \in \mathbb{C}_-, \text{ for all } \psi \in D(A^*),$$

where $\mathbb{C}_-$ is the open left complex half-plane, is sufficient to ensure exponential stabilizability; see [Jacob and Zwart 2009]. It is interesting to investigate the question of how Hautus tests are related to the observability inequalities derived in our paper (see also [Weiss and Rebarber 2000, Proposition 3.5] for an extension of the Hautus test for stabilizability).

Polynomial stabilizability. We have provided in Theorem 1 a characterization of exponential stabilizability. The C_0 semigroup $(S(t))_{t\geqslant 0}$ is said to be polynomially stable when there exist constants $\gamma, \delta > 0$ such that $\|S(t)(A - \beta \,\mathrm{id})^{-\gamma}\|_{L(X)} \leqslant M t^{-\delta}$ for every $t \geqslant 1$, for some $M > 0$ and some $\beta \in \rho(A)$ (see [Alabau-Boussouira 2002; 2003; Jacob and Schnaubelt 2007], where polynomial stability is compared with observability; see also [Borichev and Tomilov 2010]). Finding a dual characterization of polynomial stabilizability in terms of an observability inequality is an open issue, which may be related to the previous question on Hautus tests.

Shape optimization. Let $T > 0$ and $\alpha \geqslant 0$ be arbitrary. The observability inequality (13) characterizes cost-uniform α-null controllability in time T, and we have seen that, when $\alpha \in (0,1)$, it also characterizes exponential stabilizability. The best constant in the inequality (13) is μ_α^T, which, as already said, quantifies the α-null controllability or the stabilizability property. One can then address the problem of optimizing this constant, by choosing an adequate control operator B in a certain class. Let us give an example of such a problem.

The Dirichlet wave equation in a C^2 bounded open subset $\Omega \subset \mathbb{R}^n$ with internal control

$$\partial_{tt} y = \triangle y + \chi_O u, \quad y_{|\partial\Omega} = 0,$$

where O is a measurable subset of Ω, is well known to be exponentially stabilizable as soon as O is an open subset satisfying the geometric control condition; see [Bardos et al. 1992; Le Rousseau et al. 2017]. Here, the control operator is $B = \chi_O$, and the constant μ_α^T depends on O. Following [Privat et al. 2015;

2016], fixing some $L \in (0,1)$ and defining $\mathcal{U}_L$ as the set of all measurable subsets O of Ω of Lebesgue measure $|O| = L|\Omega|$, one can consider the problem

$$\inf_{O \in \mathcal{U}_L} \mu_\alpha^T(O),$$

i.e., the problem of searching the best possible subdomain that minimizes $\mu_\alpha^T(O)$ over all possible measurable subdomains O of Ω of Lebesgue measure $L|\Omega|$. We mention [Privat and Trélat 2015] for a similar problem consisting of maximizing the exponential decay rate.

Such shape optimization issues may also be raised in the above-mentioned context of polynomial stabilizability. We are not aware of any existing results in this direction.

4. Proofs

We now provide proofs of Propositions 5, 6 and 14 and of Theorem 1, following Fenchel duality arguments. In turn, we establish intermediate results which may be of interest themselves for other purposes.

We recall that the Gramian operator $\mathcal{G}_T \in L(X)$ is the symmetric positive semidefinite operator defined by

$$\mathcal{G}_T = \int_0^T S(T-t)BB^*S(T-t)^*\,dt.$$

We note that $\langle \mathcal{G}_T\psi, \psi\rangle_X = \|\mathcal{G}_T^{1/2}\psi\|_X^2 = \int_0^T \|B^*S(T-t)^*\psi\|_U^2\,dt$.

4.1. ***Fenchel duality arguments.*** We start our analysis by considering the α-null controllability problem or the approximate null controllability problem for the control system (1). What is written hereafter in this first subsection essentially follows the classical analysis by Fenchel duality done in [Lions 1992] (see also [Glowinski et al. 2008]), but is written in a more general framework.

Let $\alpha \geqslant 0$, let $T > 0$ and let $y_0 \in X$ be arbitrary. Let $\mathcal{B}$ be the closed unit ball in X. We consider the α-null controllability problem from y_0 in time T, i.e., the problem of steering the control system (1) from y_0 to $\alpha\|y_0\|_X\mathcal{B}$ in time T, meaning that

$$\|y(T; y_0, u)\|_X \leqslant \alpha\|y_0\|_X.$$

When the control system (1) is α-null controllable from y_0 in time T, we search the control of minimal L^2 norm (which is unique by strict convexity). We set

$$S_{y_0,\alpha}^T = \inf\big\{\tfrac{1}{2}\|u\|^2_{L^2(0,T;U)} \;\big|\; y(T; y_0, u) \in \alpha\|y_0\|_X\mathcal{B}\big\},$$

with the convention that $S_{y_0,\alpha}^T = +\infty$ whenever $\alpha\|y_0\|_X\mathcal{B}$ is not reachable from y_0 in time T.

The following lemma is obvious.

Lemma 27. *Let $y_0 \in X$ and $T > 0$ be arbitrary*:

- *Given some $\alpha \geqslant 0$, the control system* (1) *is α-null controllable from y_0 in time T if and only if $S_{y_0,\alpha}^T < +\infty$.*

- *The control system* (1) *is approximately null controllable from* y_0 *in time* T *if and only if* $S^T_{y_0,\alpha} < +\infty$ *for every* $\alpha > 0$.

4.1.1. *Application of Fenchel duality.* Following [Lions 1992], we define the convex and lower semicontinuous functions $F : L^2(0,T;U) \to [0,+\infty)$ and $G : X \to [0,+\infty]$ by

$$F(u) = \tfrac{1}{2}\|u\|^2_{L^2(0,T;U)}$$

and

$$G(\phi) = \begin{cases} 0 & \text{if } \phi \in -S(T)y_0 + \alpha\|y_0\|_X\mathcal{B}, \\ +\infty & \text{otherwise,} \end{cases}$$

and we note that

$$S^T_{y_0,\alpha} = \inf_{u\in L^2(0,T;U)} (F(u) + G(L_T u)).$$

The Fenchel conjugates $F^* : L^2(0,T;U) \to [0,+\infty)$ and $G^* : X \to [0,+\infty]$ are given by

$$F^*(v) = \sup_{u\in L^2(0,T;U)} (\langle v,u\rangle_X - F(u)) = \tfrac{1}{2}\|v\|^2_{L^2(0,T;U)} = F(v)$$

and

$$G^*(\psi) = \sup_{\phi\in X} (\langle \phi,\psi\rangle_X - G(\phi)) = \sup_{\phi\in -S(T)y_0+\alpha\|y_0\|_X\mathcal{B}} \langle \phi,\psi\rangle_X = -\langle \psi, S(T)y_0\rangle_X + \alpha\|y_0\|_X\|\psi\|_X,$$

where the latter equality is obtained by applying the Cauchy–Schwarz inequality.

Noting that $L_T \operatorname{dom}(F) = \operatorname{Ran}(L_T)$ intersects the set of points at which G is continuous when (1) is either α-null controllable in time T (with $\alpha > 0$ fixed) or approximately null controllable in time T, we infer from the Fenchel–Rockafellar duality theorem, see [Fenchel 1949; Ekeland and Témam 1999; Rockafellar 1967], that

$$S^T_{y_0,\alpha} = -\inf_{\phi\in X} (F^*(L_T^*\phi) + G^*(-\phi)) = -\inf_{\psi\in X} (F^*(L_T^*\psi) + G^*(\psi)) = -\inf_{\psi\in X} J^T_{y_0,\alpha}(\psi),$$

where we have set

$$J^T_{y_0,\alpha}(\psi) = F^*(L_T^*\psi) + G^*(\psi) = \tfrac{1}{2}\langle \mathcal{G}_T\psi,\psi\rangle_X - \langle \psi, S(T)y_0\rangle_X + \alpha\|y_0\|_X\|\psi\|_X \tag{31}$$

for every $\psi \in X$. Note that $J^T_{y_0,\alpha}$ is differentiable except at $\psi = 0$.

This result is still valid for $\alpha = 0$ but is not a consequence of the Fenchel–Rockafellar duality theorem (because we have used in a critical way that $\alpha > 0$): for $\alpha = 0$ this is the usual procedure in the Hilbert uniqueness method; see [Glowinski et al. 2008; Lions 1988].

4.1.2. *Computation of the minimizer.* As said above, when $S^T_{y_0,\alpha} < +\infty$, there is a unique minimizer $\bar{u}_{y_0,\alpha} \in L^2(0,T;U)$ and there is also a unique minimizer $\bar{\psi}_{y_0,\alpha} \in X$ of $J^T_{y_0,\alpha}$ (this follows from (32) below), and

$$S^T_{y_0,\alpha} = \tfrac{1}{2}\|\bar{u}_{y_0,\alpha}\|^2_{L^2(0,T;U)} = -J^T_{y_0,\alpha}(\bar{\psi}_{y_0,\alpha}).$$

We have either $\bar\psi_{y_0,\alpha}=0$ and then $\bar u_{y_0,\alpha}=0$ and $S^T_{y_0,\alpha}=0$, or $\bar\psi_{y_0,\alpha}\neq 0$ and then $\nabla J^T_{y_0,\alpha}(\bar\psi_{y_0,\alpha})=0$, which gives

$$\mathcal{G}_T\bar\psi_{y_0,\alpha}-S(T)y_0+\alpha\|y_0\|_X\frac{\bar\psi_{y_0,\alpha}}{\|\bar\psi_{y_0,\alpha}\|_X}=0. \tag{32}$$

Given any $\psi\in X$, we set $\psi=r\sigma$ with $r=\|\psi\|_X$ and $\sigma\in X$ of norm 1 (polar coordinates). For the minimizer $\bar\psi_{y_0,\alpha}=\bar r_{y_0,\alpha}\bar\sigma_{y_0,\alpha}$, we infer from (32) that

$$\bar r_{y_0,\alpha}=\frac{\langle S(T)y_0,\bar\sigma_{y_0,\alpha}\rangle_X-\alpha\|y_0\|_X}{\langle\mathcal{G}_T\bar\sigma_{y_0,\alpha},\bar\sigma_{y_0,\alpha}\rangle_X},\quad \bar\sigma_{y_0,\alpha}=(\bar r_{y_0,\alpha}\mathcal{G}_T+\alpha\|y_0\|_X)^{-1}S(T)y_0.$$

Here, we used the facts that $\langle\mathcal{G}_T\bar\sigma_{y_0,\alpha},\bar\sigma_{y_0,\alpha}\rangle_X\neq 0$ (which follows from (32)) and that $\bar u_{y_0,\alpha}\neq 0$. Note that, necessarily, $\langle S(T)y_0,\bar\sigma_{y_0,\alpha}\rangle_X-\alpha\|y_0\|_X\geqslant 0$.

Note also that, in the Fenchel duality argument, the optimal control $\bar u_{y_0,\alpha}$ is given as a function of $\bar\psi_{y_0,\alpha}$ by

$$\bar u_{y_0,\alpha}(t)=(L^*_T\bar\psi_{y_0,\alpha})(t)=B^*S(T-t)^*\bar\psi_{y_0,\alpha}.$$

Until that step, there is nothing new with respect to the existing literature. Up to our knowledge, the novelty is in the next step, with a simple remark leading to an observability inequality.

4.1.3. *An alternative optimization problem.* Following the above arguments, we first note that

$$J^T_{y_0,\alpha}(\psi)=J^T_{y_0,\alpha}(r\sigma)=\tfrac12 r^2\langle\mathcal{G}_T\sigma,\sigma\rangle_X-r(\langle\sigma,S(T)y_0\rangle_X-\alpha\|y_0\|_X)$$

and that $J^T_{y_0,\alpha}(0)=0$, and given $a\geqslant 0$ and $b\in\mathbb{R}$, we have

$$\inf_{r>0}(ar^2-br)=\begin{cases}0 & \text{if } b\leqslant 0,\\ -\dfrac{b^2}{4a}\ (-\infty \text{ if } a=0), & \text{if } b>0, \text{ reached at } r=\dfrac{b}{2a}.\end{cases}$$

Hence for any fixed $\sigma\in X$,

$$\inf_{r>0}J^T_{y_0,\alpha}(r\sigma)=\begin{cases}0 & \text{if } \langle\sigma,S(T)y_0\rangle_X-\alpha\|y_0\|_X\leqslant 0,\\ -\dfrac12\dfrac{(\langle\sigma,S(T)y_0\rangle_X-\alpha\|y_0\|_X)^2}{\langle\mathcal{G}_T\sigma,\sigma\rangle_X} & \text{if } \mathcal{G}_T\sigma\neq 0 \text{ and } \langle\sigma,S(T)y_0\rangle_X-\alpha\|y_0\|_X>0,\\ -\infty & \text{if } \mathcal{G}_T\sigma=0 \text{ and } \langle\sigma,S(T)y_0\rangle_X-\alpha\|y_0\|_X>0.\end{cases}$$

Additionally, by the definition (5) of $\mu^T_{y_0,\alpha}$, we have

$$\mu^T_{y_0,\alpha}=\sup_{\psi\in X}\begin{cases}0 & \text{if } \langle\psi,S(T)y_0\rangle_X-\alpha\|y_0\|_X\|\psi\|_X\leqslant 0,\\ \dfrac{\langle\psi,S(T)y_0\rangle_X-\alpha\|y_0\|_X\|\psi\|_X}{\|\mathcal{G}_T^{1/2}\psi\|_X} & \text{if } \mathcal{G}_T\psi\neq 0 \text{ and } \langle\psi,S(T)y_0\rangle_X-\alpha\|y_0\|_X\|\psi\|_X>0,\\ +\infty & \text{if } \mathcal{G}_T\psi=0 \text{ and } \langle\psi,S(T)y_0\rangle_X-\alpha\|y_0\|_X\|\psi\|_X>0.\end{cases}$$

It follows that

$$S^T_{y_0,\alpha}=-\inf_{\psi\in X}J^T_{y_0,\alpha}(\psi)=-\inf_{\|\sigma\|_X=1}\inf_{r>0}J^T_{y_0,\alpha}(r\sigma)=\tfrac12(\mu^T_{y_0,\alpha})^2.$$

Note that, when $0 < \mu^T_{y_0,\alpha} < +\infty$, we have

$$\mu^T_{y_0,\alpha} = \|\bar{u}_{y_0,\alpha}\|_{L^2(0,T;U)} = \frac{\langle \bar{\psi}_{y_0,\alpha}, S(T)y_0\rangle_X - \alpha\|y_0\|_X\|\bar{\psi}_{y_0,\alpha}\|_X}{\|\mathcal{G}_T^{1/2}\bar{\psi}_{y_0,\alpha}\|_X}.$$

Remark 28. According to the above discussion, if $\mu^T_{y_0,\alpha} < +\infty$, then $\mu^T_{y_0,\alpha}$ is the smallest constant $C \in [0, +\infty)$ such that there exists $u \in L^2(0,T;U)$ such that $\|y(T;y_0,u)\|_X \leqslant \alpha\|y_0\|_X$ and $\|u\|_{L^2(0,T;U)} \leqslant C$. When $C = \mu^T_{y_0,\alpha}$, one has $u = \bar{u}_{y_0,\alpha}$.

4.1.4. *Proofs of Propositions 5 and 14.* Propositions 5 and 14 follow, using Lemma 27 and the fact that $S^T_{y_0,\alpha} = \frac{1}{2}(\mu^T_{y_0,\alpha})^2$.

4.2. *Fenchel dualization of the exponential stabilization property.*

4.2.1. *Proof of* (7).

Lemma 29. *Given $\alpha > 0$ and $T > 0$, μ^T_α is the smallest possible $C \in [0, +\infty]$ such that the following weak observability inequality is satisfied*:

$$\|S(T)^*\psi\|_X - \alpha\|\psi\|_X \leqslant C\|\mathcal{G}_T^{1/2}\psi\|_X \quad \textit{for all } \psi \in X,$$

and this is independent of the sign of $\|S(T)^*\psi\|_X - \alpha\|\psi\|_X$. Moreover, when $\mu^T_\alpha < +\infty$, it is the smallest constant $C \in [0, +\infty)$ such that the above weak observability inequality holds.

Proof. Using (5), we have

$$\mu^T_{y_0,\alpha} = \sup_{\psi\in X} F^T_\alpha(y_0,\psi), \tag{33}$$

where

$$F^T_\alpha(y_0,\psi) = \begin{cases} \max\Big(\dfrac{\langle\psi, S(T)y_0\rangle_X - \alpha\|y_0\|_X\|\psi\|_X}{\|\mathcal{G}_T^{1/2}\psi\|_X},\ 0\Big) & \text{if } (y_0,\psi) \in D_1, \\ +\infty & \text{if } (y_0,\psi) \in D_2, \\ 0 & \text{if } (y_0,\psi) \in D_3, \end{cases}$$

with

$$\begin{aligned} D_1 &= \{(y_0,\psi) \in X\times X \mid \mathcal{G}_T\psi \neq 0\}, \\ D_2 &= \{(y_0,\psi) \in X\times X \mid \mathcal{G}_T\psi = 0 \text{ and } \langle\psi, S(T)y_0\rangle_X - \alpha\|y_0\|_X\|\psi\|_X > 0\}, \\ D_3 &= \{(y_0,\psi) \in X\times X \mid \mathcal{G}_T\psi = 0 \text{ and } \langle\psi, S(T)y_0\rangle_X - \alpha\|y_0\|_X\|\psi\|_X \leqslant 0\}. \end{aligned}$$

Since

$$\sup_{\|y_0\|_X=1}\ \sup_{\psi\in X} F^T_\alpha(y_0,\psi) = \sup_{\psi\in X}\ \sup_{\|y_0\|_X=1} F^T_\alpha(y_0,\psi)$$

and

$$\sup_{\|y_0\|_X=1} F^T_\alpha(y_0,\psi) = \begin{cases} \max\Big(\dfrac{\|S(T)^*\psi\|_X - \alpha\|\psi\|_X}{\|\mathcal{G}_T^{1/2}\psi\|_X},\ 0\Big) & \text{if } \mathcal{G}_T\psi \neq 0, \\ +\infty & \text{if } \mathcal{G}_T\psi = 0,\ \exists\, y_0 \text{ such that } (y_0,\psi) \in D_2, \\ 0 & \text{if } \mathcal{G}_T\psi = 0,\ \nexists\, y_0 \text{ such that } (y_0,\psi) \in D_2, \end{cases}$$

we derive the desired result from (6) and (33). □

4.2.2. *Another interpretation of* μ_α^T. We now give another interpretation of μ_α^T, useful for addressing exponential stabilizability.

Lemma 30. *Let* $\alpha > 0$ *and* $T > 0$ *be such that* $\mu_\alpha^T < +\infty$. *Then* μ_α^T *is the smallest constant* $C \geqslant 0$ *such that, for every* $y_0 \in X$, *there exists* $u \in L^2(0,T;U)$ *such that* $\|y(T;y_0,u)\|_X \leqslant \alpha\|y_0\|_X$ *and* $\|u\|_{L^2(0,T;U)} \leqslant C\|y_0\|_X$.

Proof. The result follows from the fact that $S_{y_0,\alpha}^T = \frac{1}{2}(\mu_{y_0,\alpha}^T)^2 = \frac{1}{2}\|\bar{u}_{y_0,\alpha}\|^2_{L^2(0,T;U)}$, that $\mu_{y_0,\alpha}^T = \mu_{y_0/\|y_0\|_X,\alpha}^T\|y_0\|_X \leqslant \mu_\alpha^T\|y_0\|_X$ and from Remark 28. □

Lemma 30 is closely related to exponential stabilizability when $\alpha < 1$. We indeed have the following result (an easy consequence of well-known results; however, we provide a proof).

Lemma 31. *The control system* (1) *is exponentially stabilizable if and only if, for every* $\alpha \in (0,1)$ (*equivalently, there exists* $\alpha \in (0,1)$), *there exist* $T > 0$ *and* $C > 0$ *such that, for every* $y_0 \in X$, *there exists* $u \in L^2(0,T;U)$ *such that* $\|y(T;y_0,u)\|_X \leqslant \alpha\|y_0\|_X$ *and* $\|u\|_{L^2(0,T;U)} \leqslant C\|y_0\|_X$.

When this is satisfied, the best stabilization decay rate ω^* *is the infimum of* $(\ln\alpha)/T$ *over all possible couples* (T,α) *for which the above inequalities are satisfied for some constant* $C > 0$.

Proof. Assume that the control system (1) is exponentially stabilizable. Then there exists $K \in L(X,U)$ such that $\|S_K(t)\|_{L(X)} \leqslant Me^{\omega_K t}$ for every $t \geqslant 0$, with $M \geqslant 1$ and $\omega_K < 0$. Let $y_0 \in X$. We set $u(t) = KS_K(t)y_0$ and $y(t) = S_K(t)y_0$. We have

$$\|y(T;y_0,u)\|_X = \|S_K(T)y_0\|_X \leqslant Me^{\omega_K T}\|y_0\|_X = \alpha\|y_0\|_X$$

for $T = (1/\omega_K)\ln(\alpha/M)$ and we compute

$$\|u\|_{L^2(0,T;U)} \leqslant \frac{\|K\|_{L(X,U)}M}{\sqrt{-2\omega_K}}\|y_0\|_X,$$

whence the result.

Conversely, we proceed by iteration. For the initial condition y_0, there exists a control $u_0 \in L^2(0,T;U)$ such that $\|y(T;y_0,u_0)\|_X \leqslant \alpha\|y_0\|_X$ and $\|u_0\|_{L^2(0,T;U)} \leqslant C\|y_0\|_X$. We set $y_1 = y(T;y_0,u_0)$ and we repeat the argument for this new initial condition y_1, and then we iterate, obtaining that $\|y_{j+1}\|_X = \|y((j+1)T;y_j,u_j)\|_X \leqslant \alpha\|y_j\|_X$ and $\|u_j\|_{L^2(0,T;U)} \leqslant C\|y_j\|_X$. The control u defined as the concatenation of the controls $u_j \in L^2(jT,(j+1)T;U)$ generates the trajectory $y(\cdot\,;y_0,u)$ satisfying, at time kT, $\|y(kT;y_0,u)\|_X \leqslant \alpha^k\|y_0\|_X$, and

$$\|u\|^2_{L^2(0,+\infty;U)} \leqslant \sum_{j=0}^{+\infty} C^2\|y(jT;y_0,u)\|_X^2 \leqslant C^2\sum_{j=0}^{+\infty}\alpha^{2j}\|y_0\|_X^2 = \frac{C^2}{1-\alpha^2}\|y_0\|_X^2.$$

Let us prove that $\|y(t;y_0,u)\|_X$ decreases exponentially. The argument is standard. Taking $t \geqslant 0$ such that $kT \leqslant t < (k+1)T$ for some $k \in \mathbb{N}$, we have

$$y(t;y_0,u) = S(t-kT)y(kT) + \int_{kT}^t S(t-s)Bu_k(s)\,ds.$$

The semigroup $(S(t))_{t\geqslant 0}$ satisfies $\|S(t)\|_{L(X)} \leqslant M$ for every $t \in [0, T]$ for some $M \geqslant 1$. Therefore

$$\|y(t; y_0, u)\|_X \leqslant M(1 + \|B\|_{L(U,X)}\sqrt{T})C\|y_k\|_X \leqslant \frac{M}{\alpha}(1 + \|B\|_{L(U,X)}\sqrt{T})C\alpha^{k+1}\|y_0\|_X$$

and since $\alpha^{k+1} = e^{(k+1)T((\ln\alpha)/T)} \leqslant e^{((\ln\alpha)/T)t}$ we infer that

$$\|y(t; y_0, u)\|_X \leqslant \frac{M}{\alpha}(1 + \|B\|_{L(U,X)}\sqrt{T})C\|y_0\|_X e^{((\ln\alpha)/T)t} \quad \text{for all } t > 0. \tag{34}$$

We have therefore found a control $u \in L^2(0, +\infty; U)$ such that

$$\int_0^{+\infty} (\|u(t)\|_U^2 + \|y(t; y_0, u)\|_X^2)\, dt < +\infty. \tag{35}$$

Hence, by the classical Riccati theory, see [Zabczyk 1995, Theorem 4.3, page 240], the control system (1) is exponentially stabilizable.

The first equality in (10) concerning the best stabilization rate is obvious by inspecting the above argument, and in particular (34). □

Remark 32. To prove Theorems 25 and 26 in Section 3.3, we note that all arguments in the above proof still work (before the final step where Riccati theory is invoked) for an unbounded admissible control operator B, because the operators $L_t = \int_0^t S(t-s)Bu(s)\, ds$ are bounded in X and we can write

$$\begin{aligned}\left\|\int_{kT}^t S(t-s)Bu_k(s)\, ds\right\|_X &= \left\|\int_0^{t-kT} S(t-kT-s)Bu_k(s+kT)\, ds\right\|_X \\ &= \|L_{t-kT}u_k(kT+\cdot)\|_X \leqslant C\|u_k\|_{L^2(kT,(k+1)T;U)}\end{aligned}$$

and the rest is unchanged: we find u satisfying (35), i.e., the finite-cost condition (optimizability: see Section 3.3).

In other words, we obtain that, for an admissible control operator B, the control system (1) is optimizable if and only if, for every $\alpha \in (0, 1)$ (equivalently, there exists $\alpha \in (0, 1)$), there exist $T > 0$ and $C > 0$ such that, for every $y_0 \in X$, there exists $u \in L^2(0, T; U)$ such that $\|y(T; y_0, u)\|_X \leqslant \alpha\|y_0\|_X$ and $\|u\|_{L^2(0,T;U)} \leqslant C\|y_0\|_X$.

4.2.3. *Proofs of Proposition 6 and Theorem 1.* Proposition 6 follows from Lemmas 29 and 30. We note that in Lemma 29 (resp., in Lemma 30), for every $\alpha \in (0, 1)$, the time T in items (ii) and (iv) (resp., items (ii) and (iii)) of Theorem 1 can be taken to be the same. Hence, Theorem 1 follows from Proposition 6 and Lemma 31, except the second equality in (10), which we prove next.

4.3. ***Proof of* (10).** To establish the second equality of (10), one way would consist of modifying the statement of Lemma 31 and to observe that, in this lemma, one can moreover choose $T \geqslant 1$. Anyway, we provide hereafter another argument of proof, which is, we believe, of independent interest.

The proof goes in two steps.

Step 1: We prove that if $(\alpha, T) \in \mathcal{A}$*, then* $(\alpha^n, nT) \in \mathcal{A}$ *for any* $n \in \mathbb{N}^*$.

Let $(\alpha, T) \in \mathcal{A}$. By (11), the fact that $\mu_\alpha^T < +\infty$ is equivalent to exponential stabilizability (see the two first items of Theorem 1). Using the equivalence of items (ii) and (iv) of the theorem and the fact that for every $\alpha \in (0,1)$ the time T in those items can be taken to be the same, there exists $C \geqslant 0$ such that

$$\|S(T)^*\psi\|_X \leqslant C\|B^*S(T-\cdot)^*\psi\|_{L^2(0,T;U)} + \alpha\|\psi\|_X \quad \text{for all } \psi \in X. \tag{36}$$

For every $n \in \mathbb{N}^*$, we set $\widehat{T}_n = nT$, $\hat{\alpha}_n = \alpha^n$ and $\widehat{C}_n^2 = \sum_{j=0}^{n-1} C^2\alpha^{2j}$. We claim that

$$\|S(\widehat{T}_n)^*\xi\|_X \leqslant \widehat{C}_n\|B^*S(\widehat{T}_n-\cdot)^*\xi\|_{L^2(0,\widehat{T}_n;U)} + \hat{\alpha}_n\|\xi\|_X \quad \text{for all } \xi \in X. \tag{37}$$

Indeed, for an arbitrarily fixed $\xi \in X$, we use (36) with ψ equal, respectively, to ξ, $S(T)^*\xi$, …, $(S(T)^*)^{n-1}\xi$, and we find that

$$\begin{aligned}
\|S(\widehat{T}_n)^*\xi\|_X &= \|S(T)^*(S(T)^*)^{n-1}\xi\|_X \\
&\leqslant C\|B^*S(T-\cdot)^*(S(T)^*)^{n-1}\xi\|_{L^2(0,T;U)} + \alpha\|(S(T)^*)^{n-1}\xi\|_X \\
&\leqslant C\|B^*S(T-\cdot)^*(S(T)^*)^{n-1}\xi\|_{L^2(0,T;U)} \\
&\qquad + C\alpha\|B^*S(T-\cdot)^*(S(T)^*)^{n-2}\xi\|_{L^2(0,T;U)} + \alpha^2\|(S(T)^*)^{n-2}\xi\|_X \\
&\ \ \vdots \\
&\leqslant \sum_{j=0}^{n-1} C\alpha^j\|B^*S(T-\cdot)^*(S(T)^*)^{n-j-1}\xi\|_{L^2(0,T;U)} + \alpha^n\|\xi\|_X.
\end{aligned}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\sum_{j=0}^{n-1} C\alpha^j\|B^*S(T-\cdot)^*(S(T)^*)^{n-j-1}\xi\|_{L^2(0,T;U)} \\
&\qquad \leqslant \left(\sum_{j=0}^{n-1} C^2\alpha^{2j}\sum_{j=0}^{n-1}\|B^*S(T-\cdot)^*(S(T)^*)^{n-j-1}\xi\|^2_{L^2(0,T;U)}\right)^{1/2} \\
&\qquad = \left(\sum_{j=0}^{n-1} C^2\alpha^{2j}\right)^{1/2}\|B^*S(nT-\cdot)^*\xi\|_{L^2(0,nT;U)} = \widehat{C}_n\|B^*S(\widehat{T}_n-\cdot)^*\xi\|_{L^2(0,\widehat{T}_n;U)}.
\end{aligned}$$

The previous inequality leads to (37). Hence $(\alpha^n, nT) \in \mathcal{A}$ for every $n \in \mathbb{N}^*$.

Step 2: We prove the second equality in (10).

Take an arbitrary $(\alpha_0, T_0) \in \mathcal{A}$. By Step 1, we have $(\alpha_0^n, nT_0) \in \mathcal{A}$. Additionally, for every $\alpha \in (0, \alpha_0)$, there exists $n = n(\alpha) \in \mathbb{N}^*$ such that $\alpha_0^n \leqslant \alpha < \alpha_0^{n-1}$. Hence $\mu_\alpha^{nT_0} \leqslant \mu_{\alpha_0^n}^{nT_0} < +\infty$ and thus $(\alpha, nT_0) \in \mathcal{A}$. Therefore, for every $\alpha \in (0, \alpha_0)$, we have

$$\inf\left\{\frac{\ln\alpha}{T} \;\middle|\; T \in \mathcal{T}(\alpha)\right\} \leqslant \frac{\ln\alpha}{nT_0} \leqslant \frac{\ln\alpha_0^{n-1}}{nT_0} = \frac{n-1}{n}\frac{\ln\alpha_0}{T_0}$$

and hence

$$\limsup_{\alpha\to 0+}\inf\left\{\frac{\ln\alpha}{T} \;\middle|\; T \in \mathcal{T}(\alpha)\right\} \leqslant \frac{\ln\alpha_0}{T_0}. \tag{38}$$

Since (α_0, T_0) was taken arbitrarily in $\mathcal{A}$, we infer from (38) that

$$\limsup_{\alpha\to 0+} \inf\left\{ \frac{\ln\alpha}{T} \;\middle|\; (\alpha,T)\in\mathcal{A}\right\} \leqslant \inf\left\{ \frac{\ln\alpha}{T} \;\middle|\; (\alpha,T)\in\mathcal{A}\right\}. \tag{39}$$

On the other hand, one can easily check that

$$\inf\left\{ \frac{\ln\alpha}{T} \;\middle|\; (\alpha,T)\in\mathcal{A}\right\} \leqslant \liminf_{\alpha\to 0+} \inf\left\{ \frac{\ln\alpha}{T} \;\middle|\; T\in\mathcal{T}(\alpha)\right\}. \tag{40}$$

Finally, from (39) and (40), it follows that

$$\inf\left\{ \frac{\ln\alpha}{T} \;\middle|\; (\alpha,T)\in\mathcal{A}\right\} = \lim_{\alpha\to 0+} \inf\left\{ \frac{\ln\alpha}{T} \;\middle|\; T\in\mathcal{T}(\alpha)\right\},$$

which leads to the second equality of (10).

Appendix

A.1. ***Exact null controllability implies complete stabilizability (proof of Proposition 21).*** Let us prove Proposition 21, i.e., if the control system (1) is exactly null controllable in some time $T > 0$ then it is completely stabilizable.

Proof of Proposition 21. We first note that (A, B) is exactly controllable in time T if and only if $(A + \omega\,\mathrm{id}, B)$ is exactly controllable in time T for any $\omega \in \mathbb{R}$. This follows straightforwardly by using the equivalence in terms of the observability inequality and the fact that $S_{A+\omega\,\mathrm{id}}(t) = e^{\omega t} S_A(t)$ (with obvious notation).

Now, let $\omega > 0$ be arbitrary. Since $(A + \omega\,\mathrm{id}, B)$ is exactly controllable in time T, there exists $K_\omega \in L(X, U)$ such that $A+\omega\,\mathrm{id}+BK_\omega$ generates the exponentially stable semigroup $S_{A+\omega\,\mathrm{id}+BK_\omega}(t) = e^{\omega t} S_{A+BK_\omega}(t)$, and thus $\|S_{A+BK_\omega}(t)\|_{L(X)} \leqslant Me^{-\omega t}$ for some $M \geqslant 1$. The result follows. □

Surprisingly, we have not found this result explicitly stated in the existing literature (except, in a rather indirect way, in [Curtain and Zwart 1995, Exercise 6.18, page 312], as kindly indicated to us by Guillaume Olive). What can usually be found is that exact null controllability in some time T implies exponential stabilizability (see, e.g., [Zabczyk 1995]) and that, when $(S(t))_{t\geqslant 0}$ is a group, exact null controllability in some time T implies complete stabilizability (see [Slemrod 1974]) and the converse is true (see [Zabczyk 1995]).

A.2. ***The case of a group (proof of Proposition 22, using Theorem 1).*** Let us prove Proposition 22, i.e., when $(S(t))_{t\geqslant 0}$ is a group, we have equivalence of exact controllability in some time T, exact null controllability in some time T, complete stabilizability.

Proof of Proposition 22.. Since $(S(t))_{t\geqslant 0}$ is a group, we have (i) $\Leftrightarrow$ (ii). By Proposition 21, we have (ii) $\Rightarrow$ (iii).

Let us now prove that (iii) $\Rightarrow$ (ii), by using Theorem 1. Since $(S(t))_{t\geqslant 0}$ is a group, there exists $M \geqslant 1$ and $\omega \geqslant 0$ such that $\|S(t)\|_{L(X)} \leqslant Me^{\omega|t|}$ for every $t \in \mathbb{R}$; hence

$$\|\psi\|_X \leqslant \|S(-T)\|_{L(X)} \|S(T)^*\psi\|_X \leqslant Me^{\omega T}\|S(T)^*\psi\|_X \quad \text{for all } \psi \in X, \text{ for all } T > 0. \tag{41}$$

We set $\alpha_n = 1/n$, for every $n \in \mathbb{N}^*$. Since the control system (1) is completely stabilizable, by using the equivalence of items (i) and (ii) in Theorem 1, for every $n \in \mathbb{N}^*$ there exists $T_n > 0$ such that $\mu_{\alpha_n}^{T_n} < +\infty$ and such that, setting $\beta_n = (\ln n)/T_n$, we have $\beta_n \to +\infty$ as $n \to +\infty$. Hence

$$\lim_{n\to+\infty} \frac{Me^{\omega T_n}}{n} = M \lim_{n\to+\infty} \frac{e^{\omega(\ln n)/\beta_n}}{n} = M \lim_{n\to+\infty} \frac{(e^{\ln n})^{\omega/\beta_n}}{n} = M \lim_{n\to+\infty} n^{\omega/\beta_n - 1} = 0$$

and therefore there exists $N \in \mathbb{N}^*$ such that

$$\frac{Me^{\omega T_N}}{\alpha_N} \leqslant \frac{1}{2}. \tag{42}$$

Now, for every $n \in \mathbb{N}^*$, since $\mu_{\alpha_n}^{T_n} < +\infty$, using the equivalence of items (ii) and (iv) of Theorem 1 and the fact that for every $\alpha \in (0,1)$, the time T in those items can be taken to be the same, there exists $C_n > 0$ such that

$$\|S(T_n)^*\psi\|_X \leqslant C_n \|B^* S(T_n - \cdot)^*\psi\|_{L^2(0,T_n;U)} + \frac{1}{n}\|\psi\|_X \quad \text{for all } \psi \in X. \tag{43}$$

We infer from (41) and (43) that

$$\|S(T_n)^*\psi\|_X \leqslant C_n \|B^* S(T_n - \cdot)^*\psi\|_{L^2(0,T_n;U)} + \frac{Me^{\omega T_n}}{n}\|S(T_n)^*\psi\|_X \quad \text{for all } \psi \in X,$$

and then, taking $n = N$ and using (42), we find that

$$\|S(T_N)^*\psi\|_X \leqslant 2C_N \|B^* S(T_N - \cdot)^*\psi\|_{L^2(0,T_N;U)} \quad \text{for all } \psi \in X,$$

which implies that the control system (1) is exactly null controllable in time T_N. □

Acknowledgement

This work was partially funded by the Natural Science Foundation of China 11871166, 11631004.

References

[Alabau-Boussouira 2002] F. Alabau-Boussouira, "Indirect boundary stabilization of weakly coupled hyperbolic systems", *SIAM J. Control Optim.* **41**:2 (2002), 511–541. MR Zbl

[Alabau-Boussouira 2003] F. Alabau-Boussouira, "A two-level energy method for indirect boundary observability and controllability of weakly coupled hyperbolic systems", *SIAM J. Control Optim.* **42**:3 (2003), 871–906. MR Zbl

[Alabau-Boussouira and Ammari 2011] F. Alabau-Boussouira and K. Ammari, "Sharp energy estimates for nonlinearly locally damped PDEs via observability for the associated undamped system", *J. Funct. Anal.* **260**:8 (2011), 2424–2450. MR Zbl

[Alabau-Boussouira et al. 2017] F. Alabau-Boussouira, Y. Privat, and E. Trélat, "Nonlinear damped partial differential equations and their uniform discretizations", *J. Funct. Anal.* **273**:1 (2017), 352–403. MR Zbl

[Badra and Takahashi 2014] M. Badra and T. Takahashi, "On the Fattorini criterion for approximate controllability and stabilizability of parabolic systems", *ESAIM Control Optim. Calc. Var.* **20**:3 (2014), 924–956. MR Zbl

[Banks and Ito 1997] H. T. Banks and K. Ito, "Approximation in LQR problems for infinite-dimensional systems with unbounded input operators", *J. Math. Systems Estim. Control* **7**:1 (1997), art. id. 62459. MR Zbl

[Bardos et al. 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

[Borichev and Tomilov 2010] A. Borichev and Y. Tomilov, "Optimal polynomial decay of functions and operator semigroups", *Math. Ann.* **347**:2 (2010), 455–478. MR Zbl

[Boyer 2013] F. Boyer, "On the penalised HUM approach and its applications to the numerical approximation of null-controls for parabolic problems", pp. 15–58 in *CANUM 2012, 41e Congrès National d'Analyse Numérique*, edited by L. Chupin and A. Münch, ESAIM Proc. **41**, EDP Sci., Les Ulis, 2013. MR Zbl

[Boyer and Le Rousseau 2014] F. Boyer and J. Le Rousseau, "Carleman estimates for semi-discrete parabolic operators and application to the controllability of semi-linear semi-discrete parabolic equations", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **31**:5 (2014), 1035–1078. MR Zbl

[Boyer et al. 2019] F. Boyer, V. Hernández-Santamaría, and L. de Teresa, "Insensitizing controls for a semilinear parabolic equation: a numerical approach", *Math. Control Relat. Fields* **9**:1 (2019), 117–158. MR

[Coron 2007] J.-M. Coron, *Control and nonlinearity*, Mathematical Surveys and Monographs **136**, American Mathematical Society, Providence, RI, 2007. MR Zbl

[Coron and Lü 2014] J.-M. Coron and Q. Lü, "Local rapid stabilization for a Korteweg-de Vries equation with a Neumann boundary control on the right", *J. Math. Pures Appl.* (9) **102**:6 (2014), 1080–1120. MR Zbl

[Coron and Trélat 2004] J.-M. Coron and E. Trélat, "Global steady-state controllability of one-dimensional semilinear heat equations", *SIAM J. Control Optim.* **43**:2 (2004), 549–569. MR Zbl

[Curtain and Zwart 1995] R. F. Curtain and H. Zwart, *An introduction to infinite-dimensional linear systems theory*, Texts in Applied Mathematics **21**, Springer, 1995. MR Zbl

[Ekeland and Témam 1999] I. Ekeland and R. Témam, *Convex analysis and variational problems*, Classics in Applied Mathematics **28**, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999. MR Zbl

[Engel and Nagel 2000] K.-J. Engel and R. Nagel, *One-parameter semigroups for linear evolution equations*, Graduate Texts in Mathematics **194**, Springer, 2000. MR Zbl

[Ervedoza and Zuazua 2009] S. Ervedoza and E. Zuazua, "Uniformly exponentially stable approximations for a class of damped systems", *J. Math. Pures Appl.* (9) **91**:1 (2009), 20–48. MR Zbl

[Fenchel 1949] W. Fenchel, "On conjugate convex functions", *Canadian J. Math.* **1** (1949), 73–77. MR Zbl

[Flandoli et al. 1988] F. Flandoli, I. Lasiecka, and R. Triggiani, "Algebraic Riccati equations with nonsmoothing observation arising in hyperbolic and Euler–Bernoulli boundary control problems", *Ann. Mat. Pura Appl.* (4) **153** (1988), 307–382. MR Zbl

[Glowinski et al. 2008] R. Glowinski, J.-L. Lions, and J. He, *Exact and approximate controllability for distributed parameter systems: a numerical approach*, Encyclopedia of Mathematics and its Applications **117**, Cambridge University Press, 2008. MR Zbl

[Jacob and Schnaubelt 2007] B. Jacob and R. Schnaubelt, "Observability of polynomially stable systems", *Systems Control Lett.* **56**:4 (2007), 277–284. MR Zbl

[Jacob and Zwart 2009] B. Jacob and H. Zwart, "On the Hautus test for exponentially stable C_0-groups", *SIAM J. Control Optim.* **48**:3 (2009), 1275–1288. MR Zbl

[Komornik 1997] V. Komornik, "Rapid boundary stabilization of linear distributed systems", *SIAM J. Control Optim.* **35**:5 (1997), 1591–1613. MR Zbl

[Labbé and Trélat 2006] S. Labbé and E. Trélat, "Uniform controllability of semidiscrete approximations of parabolic control systems", *Systems Control Lett.* **55**:7 (2006), 597–609. MR Zbl

[Lasiecka and Triggiani 2000a] I. Lasiecka and R. Triggiani, *Control theory for partial differential equations: continuous and approximation theories, I: Abstract parabolic systems*, Encyclopedia of Mathematics and its Applications **74**, Cambridge University Press, 2000. MR Zbl

[Lasiecka and Triggiani 2000b] I. Lasiecka and R. Triggiani, *Control theory for partial differential equations: continuous and approximation theories, II: Abstract hyperbolic-like systems over a finite time horizon*, Encyclopedia of Mathematics and its Applications **75**, Cambridge University Press, 2000. MR Zbl

[Le Rousseau et al. 2017] J. Le Rousseau, G. Lebeau, P. Terpolilli, and E. Trélat, "Geometric control condition for the wave equation with a time-dependent observation domain", *Anal. PDE* **10**:4 (2017), 983–1015. MR Zbl

[Lions 1988] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués, I: Contrôlabilité exacte*, Recherches en Mathématiques Appliquées **8**, Masson, 1988. MR Zbl

[Lions 1992] J.-L. Lions, "Remarks on approximate controllability", *J. Anal. Math.* **59**:1 (1992), 103–116. MR Zbl

[Liu 1997] K. Liu, "Locally distributed control and damping for the conservative systems", *SIAM J. Control Optim.* **35**:5 (1997), 1574–1590. MR Zbl

[Liu and Zheng 1999] Z. Liu and S. Zheng, *Semigroups associated with dissipative systems*, Chapman & Hall/CRC Research Notes in Mathematics **398**, Chapman & Hall/CRC, Boca Raton, FL, 1999. MR Zbl

[Pazy 1983] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, Applied Mathematical Sciences **44**, Springer, 1983. MR Zbl

[Phung et al. 2017] K. D. Phung, G. Wang, and Y. Xu, "Impulse output rapid stabilization for heat equations", *J. Differential Equations* **263**:8 (2017), 5012–5041. MR Zbl

[Pritchard and Zabczyk 1981] A. J. Pritchard and J. Zabczyk, "Stability and stabilizability of infinite-dimensional systems", *SIAM Rev.* **23**:1 (1981), 25–52. MR Zbl

[Privat and Trélat 2015] Y. Privat and E. Trélat, "Optimal design of sensors for a damped wave equation", *Discrete Contin. Dyn. Syst.* **2015**:Dynamical systems, differential equations and applications. 10th AIMS Conference. Suppl. (2015), 936–944. MR Zbl

[Privat et al. 2015] Y. Privat, E. Trélat, and E. Zuazua, "Optimal shape and location of sensors for parabolic equations with random initial data", *Arch. Ration. Mech. Anal.* **216**:3 (2015), 921–981. MR Zbl

[Privat et al. 2016] Y. Privat, E. Trélat, and E. Zuazua, "Optimal observability of the multi-dimensional wave and Schrödinger equations in quantum ergodic domains", *J. Eur. Math. Soc.* (*JEMS*) **18**:5 (2016), 1043–1111. MR Zbl

[Rockafellar 1967] R. T. Rockafellar, "Duality and stability in extremum problems involving convex functions", *Pacific J. Math.* **21** (1967), 167–187. MR Zbl

[Russell 1978] D. L. Russell, "Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions", *SIAM Rev.* **20**:4 (1978), 639–739. MR Zbl

[Slemrod 1974] M. Slemrod, "A note on complete controllability and stabilizability for linear control systems in Hilbert space", *SIAM J. Control* **12** (1974), 500–508. MR Zbl

[Staffans 2005] O. Staffans, *Well-posed linear systems*, Encyclopedia of Mathematics and its Applications **103**, Cambridge University Press, 2005. MR Zbl

[Trélat 2018] E. Trélat, "Stabilization of semilinear PDEs, and uniform decay under discretization", pp. 31–76 in *Evolution equations: long time behavior and control*, edited by K. Ammari and S. Gerbi, London Math. Soc. Lecture Note Ser. **439**, Cambridge Univ. Press, 2018. MR Zbl

[Triggiani 1975] R. Triggiani, "On the stabilizability problem in Banach space", *J. Math. Anal. Appl.* **52**:3 (1975), 383–403. MR Zbl

[Tucsnak and Weiss 2009] M. Tucsnak and G. Weiss, *Observation and control for operator semigroups*, Birkhäuser, Basel, 2009. MR Zbl

[Urquiza 2005] J. M. Urquiza, "Rapid exponential feedback stabilization with unbounded control operators", *SIAM J. Control Optim.* **43**:6 (2005), 2233–2244. MR Zbl

[Weiss and Rebarber 2000] G. Weiss and R. Rebarber, "Optimizability and estimatability for infinite-dimensional linear systems", *SIAM J. Control Optim.* **39**:4 (2000), 1204–1232. MR Zbl

[Weiss and Zwart 1998] G. Weiss and H. Zwart, "An example in linear quadratic optimal control", *Systems Control Lett.* **33**:5 (1998), 339–349. MR Zbl

[Zabczyk 1995] J. Zabczyk, *Mathematical control theory: an introduction*, Birkhäuser, Boston, MA, 1995.

[Zuazua 2004] E. Zuazua, "Optimal and approximate control of finite-difference approximation schemes for the 1D wave equation", *Rend. Mat. Appl.* (7) **24**:2 (2004), 201–237. MR Zbl

[Zuazua 2005] E. Zuazua, "Propagation, observation, and control of waves approximated by finite difference methods", *SIAM Rev.* **47**:2 (2005), 197–243. MR Zbl

[Zwart 1996] H. J. Zwart, "Linear quadratic optimal control for abstract linear systems", pp. 175–182 in *Modelling and optimization of distributed parameter systems* (Warsaw, 1995), edited by K. Malanowski et al., Chapman & Hall, New York, 1996. MR Zbl

Received 30 Apr 2019. Revised 19 Jun 2019. Accepted 22 Jul 2019.

EMMANUEL TRÉLAT: emmanuel.trelat@sorbonne-universite.fr
Sorbonne Université, CNRS, Université de Paris, Inria, Laboratoire Jacques-Louis Lions, Paris, France

GENGSHENG WANG: wanggs@yeah.net
Center for Applied Mathematics, Tianjin University, Tianjin, China

YASHAN XU: yashanxu@fudan.edu.cn
School of Mathematical Sciences, Fudan University, KLMNS, Shanghai, China

PURE and APPLIED ANALYSIS

Vol. 2, No. 1, 2020
dx.doi.org/10.2140/paa.2020.2.123

msp

STATIONARY COMPRESSIBLE NAVIER–STOKES EQUATIONS WITH INFLOW CONDITION IN DOMAINS WITH PIECEWISE ANALYTICAL BOUNDARIES

PIOTR B. MUCHA AND TOMASZ PIASECKI

We show the existence of strong solutions in Sobolev–Slobodetskii spaces to the stationary compressible Navier–Stokes equations with inflow boundary condition. Our result holds provided a certain condition on the shape of the boundary around the points where characteristics of the continuity equation are tangent to the boundary, which holds in particular for piecewise analytical boundaries. The mentioned situation creates a singularity which limits regularity at such points. We show the existence and uniqueness of regular solutions in a vicinity of given laminar solutions under the assumption that the pressure is a linear function of the density. The proofs require the language of suitable fractional Sobolev spaces. In other words our result is an example where the application of fractional spaces is irreplaceable, although the subject is a classical system.

1. Introduction

We investigate the existence of regular solutions to stationary barotropic compressible Navier–Stokes equations in a two-dimensional bounded domain Ω with nonzero inflow/outflow through the boundary. The complete system reads

$$\rho v\cdot\nabla v-\mu\Delta v-(\mu+\nu)\nabla \operatorname{div} v+\nabla\pi(\rho)=0 \quad \text{in } \Omega, \tag{1a}$$

$$\operatorname{div}(\rho v)=0 \quad \text{in } \Omega, \tag{1b}$$

$$n\cdot 2\mu \boldsymbol{D}(v)\cdot\tau+fv\cdot\tau=b, \quad \text{on } \Gamma, \tag{1c}$$

$$n\cdot v=d \quad \text{on } \Gamma, \tag{1d}$$

$$\rho=\rho_{\mathrm{in}} \quad \text{on } \Gamma_{\mathrm{in}}, \tag{1e}$$

where the velocity field of the fluid v and the density ρ are the unknown functions describing the flow. We distinguish the parts of the boundary

$$\begin{aligned}\Gamma_{\mathrm{in}}&=\{x\in\Gamma : d<0\}, \quad \Gamma_{\mathrm{out}}=\{x\in\Gamma : d>0\},\\ \Gamma_0&=\{x\in\Gamma : d=0\}, \quad \Gamma_*=\overline{\Gamma}_0\cap\overline{\Gamma_{\mathrm{in}}\cup\Gamma_{\mathrm{out}}}.\end{aligned} \tag{2}$$

We show the existence of a solution in fractional Sobolev spaces $u\in W_p^{1+s}(\Omega)$, $\rho\in W_p^s(\Omega)$, where u is the velocity field of the fluid and ρ is the density. Our choice of functional spaces allows us to overcome the problem of singularity in the continuity equation and obtain boundedness of the density. Before we formulate the problem more precisely we give a brief overview of recent developments in this topic,

MSC2010: 35Q30, 76N10.

Keywords: compressible Navier–Stokes equations, slip boundary condition, inflow boundary condition, strong solutions.

focusing on the scope of interest of this paper, that is, on regular stationary solutions, mentioning also the most important results concerning global weak stationary solutions. For more complete overview of known results in the mathematical theory of compressible flows we refer to [Novotný and Straškraba 2004; Plotnikov and Sokołowski 2012].

The mathematical theory of stationary solutions to the Navier–Stokes equations describing compressible flows started to develop in early 80s with certain results on the existence of regular solutions, first in Hilbert spaces [Valli 1983] and later in L_p framework [Beirão da Veiga 1987]. However, all of these results required certain smallness assumptions on the data and concerned mostly homogeneous boundary conditions with vanishing normal component of the velocity.

The famous result of [Lions 1998] on the existence of weak solutions for homogeneous Dirichlet boundary conditions triggered the development of the global existence theory of weak solutions. The result was improved by Novo and Novotný [2002], who adopted the nonsteady approach of [Feireisl et al. 2001], and then it was extended to the case of slip boundary conditions in the barotropic case in [Mucha and Pokorný 2006; Pokorný and Mucha 2008] and for the full system, including thermal effects in [Mucha and Pokorný 2009]; see also the result for a system involving radiation effects in [Kreml et al. 2013]. Further improvements in the theory of regular solutions were made in [Novotný and Padula 1994; Novotný and Pileckas 1998] but mostly for homogeneous boundary data.

It should be emphasized that all above-mentioned global results concern the case of the normal component of the velocity vanishing on the boundary. If the normal component of the velocity does not vanish, substantial mathematical difficulties arise in the analysis of the continuity equation, which can be reduced to a stationary transport equation. Namely, the hyperbolicity of the continuity equation makes it necessary to prescribe the density on the part of the boundary where the fluid enters the domain, called briefly the inflow part. Solvability of either a time-dependent or stationary transport equation is of utmost importance in the mathematical analysis of the Navier–Stokes equations for compressible flow. For a recent application of the theory of the transport equation in the context of the existence of weak solutions to the compressible Navier–Stokes equation we can refer to [Bresch and Jabin 2018]. Important developments in the theory of the transport equation, strengthening the classical results of [DiPerna and Lions 1989], were made in [Crippa and De Lellis 2008; Mucha 2010].

The mathematical investigation of inflow/outflow problems began with Valli and Zajączkowski [1986], who investigated the time-dependent problem obtaining also an existence result in the stationary case. Then the development of existence theory for inhomogeneous boundary data was hindered by both mathematical difficulties and a shift in interest toward global existence of weak solutions, until [Kweon and Kellogg 1997]. More recently, the existence theory has been developed, motivated by applications in shape optimization, by Plotnikov, Ruban and Sokolowski [Plotnikov et al. 2008; 2009]; see also [Plotnikov and Sokołowski 2012]. All above results require certain smallness assumptions. Concerning large data problems, there are only few particular results on the global existence of weak solutions for nonstationary problems; see [Girinon 2011]. In the stationary case, due to nontrivial boundary terms it has been for a long time impossible to get basic a priori estimates and further problems are encountered with the issue of existence and uniqueness for the continuity equation. The first global existence result

was obtained very recently by Feireisl and Novotný [2018], under the assumption that the pressure is a nondecreasing C^1 function of the density satisfying $\lim_{\rho\to\bar\rho}\ p(\rho)=+\infty$ for some positive constant $\bar\rho$. The proof is based on appropriate regularization of both continuity and momentum equations. The key estimates for the approximate systems are obtained using a suitable extension of the boundary velocity which is constructed in such a way that it satisfies a certain smallness condition even though the data can be arbitrarily large.

At first glance a natural functional space for regular solutions is W^1_p for the density and W^2_p for the velocity. A regular solution is then understood as a function with weak derivatives satisfying the equations almost everywhere. However, except for some special classes of domains, we are not able to obtain the solutions in the above class for arbitrarily large p; see [Kweon and Kellogg 1997]. The reason is a singularity arising in the solution of the steady transport equation around the points where characteristics of this hyperbolic equation become tangent to the boundary; we refer to these points as singularity points.

On the other hand, the range $p>n$ is important since it gives boundedness of the density due to the imbedding theorem. The results from [Kweon and Kellogg 1997] cover a part of this range, namely $2<p<3$. However, a further increase of p is impossible even under relaxation of the boundary singularity. Further investigation of this singularity is therefore an interesting question in view of the development of the theory of regular solutions.

One possible way to obtain existence for $n<p<\infty$ is to investigate some special domains, such as a cylindrical domain in [Mucha and Piasecki 2014; Piasecki 2009; 2010; Guo et al. 2015] for the barotropic case and [Piasecki and Pokorný 2014] for a system with thermal effects or an unbounded domain contained between two parallel planes in [Kweon and Kellogg 1998]. A possible way to overcome the singularity problem described above in a general domain is an appropriate choice of functional spaces. In [Plotnikov et al. 2008; 2009] the existence and uniqueness of solutions in fractional Sobolev spaces (velocity in W^{1+s}_p and density in W^s_p) is shown under certain assumption relating the inflow velocity and shape of the boundary around the singularity points. However, this result requires an additional assumption that the gradient of the density and the second gradient of the velocity are in L_2.

In this paper we show existence of solutions in fractional Sobolev spaces as above. However we do not require the existence of $\nabla\rho$ and $\nabla^2 u$. Our analysis shows that we are able to show existence of the solutions for $sp>n$, which gives boundedness of the density. We need to impose a certain limitation on the boundary around the singularity points; however this assumption is weaker than in [Kweon and Kellogg 1997; Plotnikov et al. 2008; 2009] and turns out to be quite natural; in particular it is satisfied by analytical boundaries.

The only result giving uniqueness of solutions to the compressible Navier–Stokes equations for large data without information on $\nabla\rho$ is [Hoff 2006] for a time-dependent problem. The key idea there is to show uniqueness in quite low regularity, namely L_2 for the velocity and H^{-1} for the density. Then we have to estimate the H^{-1}-norm of the pressure with H^{-1}-norm of the density and for this purpose it is required that the pressure is a linear function of the density (or satisfies slightly more general constraint — see (1.16) in [Hoff 2006]).

Our result, which is to our knowledge the first one giving uniqueness of solutions without any information on the gradient of the density in the stationary case, combines the ideas from [Hoff 2006] in the context of uniqueness with the approach used to obtain existence of regular solutions in the series of papers [Piasecki 2010; Mucha and Piasecki 2014; Piasecki and Pokorný 2014]. It shows that the choice of fractional Sobolev spaces is in a sense natural for the considered problem and therefore is not only of purely mathematical interest. In particular it may indicate a possible direction for the development of the theory of global existence.

1.1. ***Functional spaces.*** In order to formulate more precisely the problem and main result we shall first recall the definitions of the functional spaces in which we work. We use standard Sobolev spaces W_p^k with natural k, which consist of functions with weak derivatives up to order k in $L_p(\Omega)$; for the definition we refer for example to [Adams and Fournier 2003]. However, most important for our result are Sobolev–Slobodetskii spaces W_p^s with fractional s. For the sake of completeness we recall the definition here; see [Triebel 1978, Section 4.4]. By $W_p^s(\Omega)$ we denote the space of functions for which the norm

$$\|f\|_{W_p^s(\Omega)} = \left(\int_\Omega |f|^p\,dx\right)^{1/p} + \left(\iint_{\Omega^2} \frac{|f(x)-f(y)|^p}{|x-y|^{2+sp}}\,dx\,dy\right)^{1/p} \tag{3}$$

is finite and $s \in (0,1)$. Furthermore, $W_p^{1+s}(\Omega)$ is a space of functions with first weak derivatives in $W_p^s(\Omega)$ with the norm

$$\|f\|_{W_p^{1+s}(\Omega)} = \|f\|_{W_p^s(\Omega)} + \|\nabla f\|_{W_p^s(\Omega)}. \tag{4}$$

Let us recall two important features of Sobolev–Slobodetskii spaces. We formulate them in a simplified way convenient for our applications.

Fact. Assume $\Omega \subset \mathbb{R}^2$ be bounded with $\partial\Omega \in C^2$, Let $u \in W_p^s$ with $sp > n$. Then for all $q \le \infty,\ \delta > 0$

$$\|u\|_{L_q} \le C(q,\Omega)\|u\|_{W_p^s}, \tag{5}$$

$$\|u\|_{L_q} \le \delta\|u\|_{W_p^s} + C(\delta)\|u\|_{L_2}. \tag{6}$$

The proofs of more general versions of the above facts can be found in [Triebel 1978]. Furthermore, for given $\epsilon > 0$ let us define

$$\Omega_\epsilon = \{x \in \Omega :\ \operatorname{dist}(x, \partial\Omega) > \epsilon\}.$$

Then by $B_{p,\infty}^r(\Omega)$ we denote a space of functions for which (see [Triebel 1978, Section 4.4] or [DeVore and Sharpley 1993, Section 2])

$$\|u\|_{B_{p,\infty}^r(\Omega)} = \sup_{0<|h|\le h_0} |h|^{-r}\|u(\cdot+h) - u(\cdot)\|_{L_p(\Omega_{|h|})} < \infty, \tag{7}$$

where $h_0 = \sup\{|h| : |\Omega_h| > 0\}$. Then for Ω bounded with $\partial\Omega \in C^2$ we have (see [Triebel 1978, Section 4.6])

$$W_p^r(\Omega) \subset B_{p,\infty}^r(\Omega). \tag{8}$$

Finally, let us define

$$V = \{v \in W_2^1(\Omega) : v \cdot n|_\Gamma = 0\}. \tag{9}$$

1.2. ***Problem formulation.*** Let us move to a precise statement of the problem under consideration. We investigate stationary flow of a barotropic fluid in a two-dimensional, bounded domain described by the system (1). The system is supplied with inhomogeneous slip boundary conditions on the velocity. In particular, the normal component of the velocity does not vanish and, as explained above, we have to prescribe the density on the part of the boundary where the flow enters the domain.

Let us have a closer look at the definition of different parts of the boundary (2). We see that Γ_* consists of points where the characteristics of the continuity equation (1b) become tangent to the boundary; we will call it the set of singularity points. Moreover, let Γ_{in}^s be a certain boundary neighborhood of the set of singularity points. Formally we define it as

$$\Gamma_{\text{in}}^s = \{x \in \Gamma_{\text{in}} : \operatorname{dist}(x, \Gamma_*) < 2\eta\}, \tag{10}$$

where $\eta > 0$ is a small number to be made precise later.

Our goal is to show the existence of a solution $(u, \rho) \in W_p^{1+s} \times W_p^s$ to the system (1), where $s > n/p$, which is close to the constant flow

$$(\bar{v}, \bar{\rho}) \equiv ([v^*, 0], 1), \tag{11}$$

where v^* is a positive constant. Our method works for a wider class of solutions in which x_1 is in a sense a dominating direction. Our motivation for the choice of this fractional-order space for the density has been explained above; we want to solve the problem of singularity in the solution of the continuity equation around the singularity points. On the other hand, by the imbedding theorem we have $W_p^s(\Omega) \in L_\infty(\Omega)$. Then the choice of the space W_p^{1+s} for the velocity follows naturally from the structure of (1). Obviously such a solution no longer satisfies the equations almost everywhere. Therefore in order to define the solutions we need a weak formulation of the problem (1). A natural way is to multiply (1a) by $\psi \in V$ and (1b) by a smooth function ϕ such that $\phi|_{\Gamma \setminus \Gamma_{\text{in}}=0}$, integrate by parts and apply boundary conditions (1c), (1d). However, because of inhomogeneous condition (1d) we would obtain boundary terms with derivatives of v. Therefore we first remove the inhomogeneity and introduce a weak formulation of the perturbed problem (30). We obtain the following definition:

Definition 1. A strong solution to the system (1) is a pair $(v, \rho) \in W_p^{1+s} \times W_p^s$ such that $v = \bar{v} + u + u_0$ and $\rho = \bar{\rho} + w$, where (u, w) is a solution to the system (30) in the sense of Definition 7 and u_0 is defined in (29).

In order to formulate our main result let us introduce the following quantity to measure the distance of the data of the problem (1) from $(\bar{v}, \bar{\rho})$:

$$D_0 = \|b - f\tau_1\|_{W_p^{s-1/p}(\Gamma)} + \|d - n_1\|_{W_p^{1+s-1/p}(\Gamma)} + \|\rho_{\text{in}} - 1\|_{W_p^s(\Gamma_{\text{in}})}. \tag{12}$$

As the formulation our main result involves certain properties of the boundary, it is useful to describe first the domain, introducing the necessary assumptions.

1.3. ***The domain*: *representative case.*** Conducting the proof for a general domain with multiple singularity points would lead to unnecessary complications which would likely hide the main ideas. For clarity

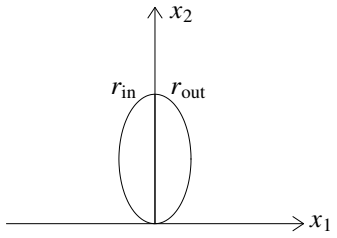

Figure 1. The domain: simple representative case with two singularity points.

of the proof we consider a simple domain with inflow and outflow parts of the boundary given by

$$\begin{aligned} \Gamma_{\text{in}} &= \{(\underline{x}_1(x_2), x_2) : x_2 \in (0, b)\}, \\ \Gamma_{\text{out}} &= \{(\bar{x}_1(x_2), x_2) : x_2 \in (0, b)\}, \end{aligned} \tag{13}$$

with

$$\lim_{x_2\to 0^+} \underline{x}_1(x_2) = \lim_{x_2\to b^-} \underline{x}_1(x_2) = \lim_{x_2\to 0^+} \bar{x}_1(x_2) = \lim_{x_2\to b^-} \bar{x}_1(x_2) = 0 \tag{14}$$

and

$$\begin{aligned} \lim_{x_2\to 0^+} \underline{x}_1'(x_2) &= \lim_{x_2\to b^-} \bar{x}_1'(x_2) = -\infty, \\ \lim_{x_2\to 0^+} \bar{x}_1'(x_2) &= \lim_{x_2\to b^-} \underline{x}_1'(x_2) = +\infty. \end{aligned} \tag{15}$$

Then we have two singularity points:

$$\Gamma_* = \Gamma_0 = \{(0, 0), (0, b)\}.$$

We assume further that these are the only singularity points; that is, the are no singularity points "inside" Γ_{in} and Γ_{out}. An example of such domain is shown in Figure 1. It is well known and was already mentioned in the introduction that existence of regular solutions to the inflow problem (1) requires certain assumptions on the shape of the boundary around the singularity points. We also need an assumption of this kind; in order to formulate it notice that around each singularity point the boundary is given as a function $x_2(x_1)$. We assume it satisfies the condition

$$\begin{aligned} &\text{there exists } N \in \mathbb{N} \text{ such that } |x_2(x_1) - x_2(y_1)| \geq C|x_1 - y_1|^N \\ &\quad \text{for all } x_1, y_1 \text{ such that } (x_1, x_2(x_1)), (y_1, x_2(y_1)) \in \Gamma_{\text{in}}^s. \end{aligned} \tag{16}$$

We emphasize that the above condition is required only on the inflow part, therefore it can be rewritten as

$$\text{there exists } \delta > 0 \text{ such that } |\underline{x}_1(x_2) - \underline{x}_1(y_2)| \leq C|x_2 - y_2|^\delta, \quad \text{with } \delta = \frac{1}{N}. \tag{17}$$

Remark 2. Notice that if (16) is satisfied by some N^* then it is also satisfied for $N > N^*$ for small $|x_2 - y_2|$. This condition means that the boundary around a singularity point must be less flat than some polynomial. Moreover we shall emphasize that (16) is a necessary but not sufficient condition; we also require sufficient global regularity of the boundary in order to solve an auxiliary elliptic problem. In order to avoid additional technicalities we assume that Ω is a C^3 domain. Such regularity is obviously not assured by (16); in particular a Lipschitz boundary may satisfy (16) for any $N \geq 1$.

Remark 3. A similar domain is considered in [Kweon and Kellogg 1997] and it is worth comparing to the condition (17) with a similar constraint there with $\delta \geq \frac{1}{2}$. Therefore our condition is clearly weaker and means that the boundary around the singularity points is less flat than some polynomial.

Another interesting observation concerning the condition (16) is that a polynomial behavior of the boundary near the singularity points turns out to be important in a completely different context of singularity of boundary layers in the stationary Munk equation; see [Dalibard and Saint-Raymond 2018].

Although condition (17) seems quite technical in the above formulation, it is in fact satisfied by a wide class of functions, in particular by piecewise analytical boundaries, which is shown in the following lemma:

Lemma 4. *Assume that x_2 is an analytic function of x_1 around the singularity points. Then* (16) *holds.*

Proof. By (15) we have $x_2'(x_1) = 0$ at the singularity points. Therefore it is enough to show that if $f : \mathbb{R} \to \mathbb{R}$ is analytic in some $[-r, r]$, $f(0) = f'(0) = 0$, and $f \neq 0$ then

$$|f(x)| \geq C|x|^N \quad \text{for } x \in (-l, l) \tag{18}$$

for some $C > 0$, $N \geq 2$, and $l < r$ sufficiently small. Since $f \neq 0$ and f is analytic, we must have $f^{(n)}(0) \neq 0$ for some $n \geq 2$. Let $f^{(k)}(0)$ be the first derivative not vanishing in 0. Then we have

$$f(x) = \frac{f^{(k)}(0)}{k!}x^k + R^{k+1}(x),$$

where $|R^{k+1}(x)| \leq M|x|^{k+1}$ for $x \in (-r, r)$. Hence

$$|f(x)| \geq \left|\frac{f^k(0)}{k!}\right||x|^k - M|x|^{k+1} = \left(\frac{f^k(0)}{k!} - M|x|\right)|x|^k. \qquad \square$$

1.4. ***The domain*: *general case.*** Our result holds for a wide class of domains where the inflow and outflow parts are defined in a natural way. A general setting is to consider inflow and outflow described as

$$\begin{aligned} \Gamma_{\text{in}} &= \{(\underline{x}_1(x_2), x_2) : x_2 \in (a, b)\}, \\ \Gamma_{\text{out}} &= \{(\bar{x}_1(x_2), x_2) : x_2 \in (a, b)\}, \end{aligned} \tag{19}$$

with singularity points given by $\{(\underline{x}_1(x_2), x_2) : x_2 = k_{\text{in}}^i\}$ and $\{(\bar{x}_1(x_2), x_2) : x_2 = k_{\text{out}}^j\}$, where

$$\begin{aligned} \lim_{x_2 \to k_{\text{in}}^{i-}} |\underline{x}_1'(x_2)| &= \lim_{x_2 \to k_{\text{in}}^{i+}} |\underline{x}_1'(x_2)| = \infty, \\ \lim_{x_2 \to k_{\text{out}}^{j-}} |\bar{x}_1'(x_2)| &= \lim_{x_2 \to k_{\text{out}}^{j+}} |\bar{x}_1'(x_2)| = \infty. \end{aligned}$$

Furthermore we assume

$$\begin{aligned} \lim_{x_2 \to a^+} \underline{x}_1(x_2) &= c_1, \quad \lim_{x_2 \to b^-} \underline{x}_1(x_2) = c_2, \\ \lim_{x_2 \to a^+} \bar{x}_1(x_2) &= d_1, \quad \lim_{x_2 \to b^-} \bar{x}_1(x_2) = d_2, \end{aligned} \tag{20}$$

with $c_i \leq d_i$. Then we have

$$\Gamma_0 = (c_1, d_1) \times \{a\} \cup (c_2, d_2) \times \{b\}. \tag{21}$$

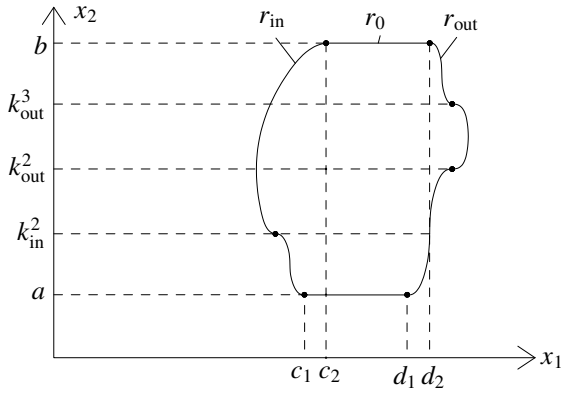

Figure 2. The domain: general case.

An example of an above-described general domain is shown in Figure 2. Again, around singularity points x_2 is given as a function $x_2^i(x_1)$ and we assume these functions satisfy (16).

1.5. ***Main result.*** We are now in a position to formulate our main result.

Theorem 5. *Assume that* Ω *satisfies the conditions introduced above*, *in particular* (17). *Assume the following regularity of the boundary data*:

$$b \in W_p^{s-1/p}(\Gamma), \quad d \in W_p^{1+s-1/p}(\Gamma), \tag{22}$$

$$\rho_{\text{in}} \in W_p^s(\Gamma_{\text{in}}) \cap W_p^r(\Gamma_{\text{in}}^s), \tag{23}$$

where

$$s < \delta, \quad r > \frac{s}{\delta}, \quad sp > n. \tag{24}$$

Let the pressure be in the form

$$\pi(\rho) = K\rho \tag{25}$$

for some positive constant K. *Let the viscosity* μ *be sufficiently large compared to* $|\Omega|$, $\|\bar{v}\|_{L_\infty}$ *and* K. *Assume further that the boundary data satisfy the following additional assumption*:

$$|(d - n_1)|\sqrt{1 + |\underline{x}_1'(x_2)|^2} \le \kappa < 1. \tag{A1}$$

Assume that D_0 *defined in* (12) *is small enough and let* f *be large enough on* Γ_{in}. *Then there exists a solution* $(v, \rho) \in W_p^{1+s} \times W_p^s$ *to the system* (1) *such that*

$$\|v - \bar{v}\|_{W_p^{1+s}} + \|\rho - \bar{\rho}\|_{W_p^s} \le E(D_0), \tag{26}$$

where $E(\cdot)$ *is a Lipschitz function*, $E(0) = 0$. *Moreover*, *this solution is unique in the class of solutions satisfying* (26).

Remark 6. Condition (23) is required to obtain the estimate for the density near the singularity points; the details are shown in the proofs of Lemmas 15 and 17. Therefore, for δ given by the geometry of the

domain in (17) we can choose an arbitrarily small r provided we choose a sufficiently small s, but we have to compensate for it with sufficiently large p. Note that in particular we can assume

$$\rho_{\text{in}} \in W_p^s(\Gamma_{\text{in}}) \cap W_p^1(\Gamma_{\text{in}}^s). \tag{27}$$

The rest of the paper is organized as follows. In the remainder of the present section we reformulate the problem (1) introducing perturbations as new unknowns, obtaining system (30). Then we recall basic properties of the functional spaces we use. In Section 2 we introduce the linearization of (30) and show a priori estimates. First we show the energy estimate. Then in Section 2.2 we move to the estimate in W_p^s for the steady transport equation, which is the main difficulty in the proof. This result makes it possible to conclude the estimate in $W_p^{1+s} \times W_p^s$ in Sections 2.3 and 2.4. In Section 3 we prove Theorem 5 with an iterative scheme using our estimates for the linear system to show the convergence of the sequence of approximations. Finally we finish with a short concluding section. Without loss of generality we assume in the proofs $v^* = 1$ in (11), except in the proof of Proposition 23 where we need to track the dependence of the viscosity on this constant.

To remove inhomogeneity from the boundary condition (30d), we construct $u_0 \in W_p^{1+s}$ such that

$$u_0 \cdot n|_\Gamma = d - n_1, \quad \|u_0\|_{W_p^{1+s}} \le C\|d - n_1\|_{W_p^{1+s-1/p}(\Gamma)}. \tag{28}$$

For our purpose we find u_0 as a solution to the problem

$$-\mu\Delta u_0 - (\mu+\nu)\nabla \operatorname{div} u_0 = 0, \quad u_0 \cdot n = d - n_1. \tag{29}$$

In order to construct u_0 we supply (29) with the condition $u_0 \cdot \tau = 0$ and define $u_0 = u_0^1 + u_0^2$, where $u_0^1 \in W_p^{1+s}$ is any extension of the boundary data $d - n_1$ and u_0^2 solves

$$-\mu\Delta u_0^2 - (\mu+\nu)\nabla \operatorname{div} u_0^2 = \mu\Delta u_0^1 + (\mu+\nu)\nabla \operatorname{div} u_0^1, \quad u_0^2|_\Gamma = 0.$$

Introducing the perturbations

$$u = v - \bar{v} - u_0 \quad \text{and} \quad w = \rho - \bar{\rho},$$

we obtain the system

$$\partial_{x_1} u - \mu\Delta u - (\nu+\mu)\nabla \operatorname{div} u + K\,\nabla w = F(u,w) \quad \text{in } \Omega, \tag{30a}$$

$$\operatorname{div} u + \partial_{x_1} w + (u+u_0)\cdot\nabla w = G(u,w) \quad \text{in } \Omega, \tag{30b}$$

$$n \cdot 2\mu \boldsymbol{D}(u)\cdot\tau + f\, u\cdot\tau = B \quad \text{on } \Gamma, \tag{30c}$$

$$n \cdot u = 0 \quad \text{on } \Gamma, \tag{30d}$$

$$w = w_{\text{in}} \quad \text{on } \Gamma_{\text{in}}, \tag{30e}$$

where

$$\begin{aligned} F(u,w) &= -w\,(u+\bar{v}+u_0)\cdot\nabla(u+u_0) - \partial_{x_1}u_0 - (u+u_0)\cdot\nabla(u+u_0), \\ G(u,w) &= -(w+1)\,\operatorname{div} u_0 - w\,\operatorname{div} u, \\ B &= b - f\tau^{(1)} - 2\mu\, n\cdot\boldsymbol{D}(u_0)\cdot\tau. \end{aligned} \tag{31}$$

Notice that in general $F(u, w)$ contains a term $\mu\Delta u_0 + (\nu+\mu)\nabla \operatorname{div} u_0$, which vanishes due to our definition of u_0. It can be seen easily that

$$\begin{aligned}
\|F(u,w)\|_{L_p} &\le C[\|w\|_{W_p^s}\|u\|_{W_p^{1+s}} + E(\|\nabla u, u, w\|_{L_p}+1)],\\
\|G(u,w)\|_{W_p^s} &\le C[\|w\|_{W_p^s}\|u\|_{W_p^{1+s}} + E(\|w\|_{W_p^s}+1)],\\
\|B\|_{W_p^{s-1/p}} &\le C[\|b - f\tau\|_{W_p^{s-1/p}} + \|d-n_1\|_{W_p^{1+s-1/p}(\Gamma)}],
\end{aligned} \tag{32}$$

where E is a small constant dependent on $\|u_0\|_{W_p^{1+s}}$ and V^* is a dual space to V defined in (9).

From now on we focus on the system (30). Our goal is to show the existence of a solution $(u, w) \in W_p^{1+s} \times W_p^s$ for given small function $u_0 \in W_p^{1+s}$. Recall in particular that the solution to the system (30) is used in the definition of the solution to the original problem (1). In order to define the solution to (30) we need its weak formulation. For this purpose we apply the identity

$$\begin{aligned}
&\int_\Omega (-\mu\Delta u - (\nu+\mu)\nabla \operatorname{div} u)\cdot v\,dx\\
&\quad = \int_\Omega 2\mu \boldsymbol{D}(u) : \nabla v + \nu \operatorname{div} u \operatorname{div} v\,dx - \int_\Gamma n\cdot[2\mu\boldsymbol{D}(u)]\cdot v\,d\sigma - \int_\Gamma n\cdot[\nu(\operatorname{div} u)\mathbf{Id}]\cdot v\,d\sigma.
\end{aligned} \tag{33}$$

Then a natural definition is the following:

Definition 7. A regular W_p^s solution to the problem (30) is a pair $(u, w) \in W_p^{1+s} \times W_p^s$ such that

$$\begin{aligned}
&\int_\Omega \psi\,\partial_{x_1} u\,dx + \int_\Omega (2\mu\boldsymbol{D}(u) : \nabla\psi + \nu \operatorname{div} u \operatorname{div}\psi)\,dx\\
&\qquad + \int_\Gamma [f(u\cdot\tau)(\psi\cdot\tau) - b(\psi\cdot\tau)]\,d\sigma - \int_\Omega w \operatorname{div}\psi\,dx = \int_\Omega F(u,w)\cdot\psi\,dx \quad \text{for all } \psi \in V
\end{aligned} \tag{34}$$

and

$$\begin{aligned}
&\int_\Omega \phi \operatorname{div} u\,dx - \int_\Omega w((\bar v + u + u_0)\cdot\nabla\phi + \phi \operatorname{div}(u+u_0))\,dx + \int_{\Gamma_{\text{in}}} w_{\text{in}}\phi\,d\sigma = \int_\Omega G(u,w)\phi\,dx\\
&\qquad\qquad \text{for all } \phi \in C^1(\Omega) \text{ such that } \phi|_{\Gamma\setminus\Gamma_{\text{in}}} = 0.
\end{aligned} \tag{35}$$

2. Linearization and a priori bounds

In this section we derive a priori estimates for the following linearization of system (30):

$$\partial_{x_1} u - \mu\Delta u - (\nu+\mu)\nabla \operatorname{div} u + K\,\nabla w = F \qquad \text{in } \Omega, \tag{36a}$$

$$\operatorname{div} u + \partial_{x_1} w + U\cdot\nabla w = G \qquad \text{in } \Omega, \tag{36b}$$

$$n\cdot 2\mu\boldsymbol{D}(u)\cdot\tau + f\,u\cdot\tau = B \qquad \text{on } \Gamma, \tag{36c}$$

$$n\cdot u = 0 \qquad \text{on } \Gamma, \tag{36d}$$

$$w = w_{\text{in}} \qquad \text{on } \Gamma_{\text{in}}, \tag{36e}$$

where $U \in W_p^{1+s}$ is small and satisfies $U\cdot n|_\Gamma = d - n_1$ and on the right-hand side we have

$$F \in L_p(\Omega), \quad G \in W_p^s(\Omega), \quad \text{and} \quad B \in W_p^{s-1/p}(\Gamma). \tag{37}$$

We start with the energy estimate, then we deal with the steady transport equation, which is the main difficulty of our proof, and finally we show the estimate in the solution space.

2.1. ***Energy estimate.*** In this section we show an energy estimate for the solutions of (36).

Lemma 8. *Let* (u, w) *be a sufficiently smooth solution to system* (36) *with given functions* $(F, G, B) \in (V^* \times L_2 \times L_2(\Gamma))$. *Then*

$$\|u\|_{W_2^1} + \|w\|_{L_2} \le C[\|F\|_{V^*} + \|G\|_{L_2} + \|B\|_{L^2(\partial\Omega)}], \tag{38}$$

where C *is independent from the boundary data.*

Proof. Multiplying (36a) by u and integrating over Ω we get, using the boundary condition (36c) and the identity (33),

$$\int_\Omega 2\mu \boldsymbol{D}^2(u) + \nu \operatorname{div}^2 u \, dx + \int_{\partial\Omega} \left(f + \frac{n_1}{2}\right) u^2 \, d\sigma + \int_\Omega K \nabla w u \, dx = -\int_\Omega u \, \partial_{x_1} u \, dx + \int_\Omega F u \, dx + \int_{\partial\Omega} B(u \cdot \tau) \, d\sigma. \tag{39}$$

The boundary term on the left-hand side will be positive for $f \ge 0$ on Γ_{out} and $f \ge n_1/2$ on Γ_{in}. Next we integrate by parts the last term of the left-hand side of (39). Using (36b) we obtain

$$K \int_\Omega u \nabla w \, dx = -K \int_\Omega w \operatorname{div} u \, dx = K \int_\Omega (w \partial_{x_1} w + U \cdot w \nabla w - G w) \, dx = K \left[\frac{1}{2} \int_\Gamma w^2 n_1 \, d\sigma - \frac{1}{2} \int_\Omega w^2 \operatorname{div} U - \int G w \, dx \right]. \tag{40}$$

We will also use the Korn inequality

$$\int_\Omega 2\mu \mathbf{D}^2(u) + \nu \operatorname{div}^2 u \, dx \ge C_K \|u\|^2_{W_2^1(\Omega)}, \tag{41}$$

where $C_K = C_K(\Omega)$. Using (39), (40) and (41) we get

$$\|u\|^2_{W_2^1(\Omega)} + \int_{\Gamma_{\text{out}}} w^2 n_1 \, d\sigma \le C \left[\int_\Omega w^2 \operatorname{div} U \, dx + \gamma \int_\Omega G w \, dx + \int_\Omega F u \, dx + \int_\Gamma B(u \cdot \tau) \, d\sigma + \int_{\Gamma_{\text{in}}} w_{\text{in}}^2 n_1 \, d\sigma \right].$$

Next, using Hölder and Young inequalities, the fact that $w^2 n_1 > 0$ on Γ_{out}, and the trace theorem on the boundary term, we get for any $\delta > 0$

$$\|u\|^2_{W_2^1(\Omega)} \le (\delta + \|U\|_{W_p^{1+s}(\Omega)}) \|w\|^2_{L_2(\Omega)} + C(\delta) \|G\|^2_{L_2(\Omega)} + C[(\|F\|_{V^*} + \|B\|_{L_2(\Gamma)}) \|u\|_{W_2^1(\Omega)}],$$

which yields

$$\|u\|_{W_2^1(\Omega)} \le (\delta + \|U\|_{W_p^{1+s}(\Omega)}) \|w\|_{L_2(\Omega)} + C(\delta) \|G\|_{L_2(\Omega)} + C(\|F\|_{V^*} + \|B\|_{L_2(\Gamma)}). \tag{42}$$

To estimate the first term of the right-hand side we find a bound on $\|w\|_{L_2}$. From (36b) we have

$$\partial_{x_1} w + U \cdot \nabla w = G - \operatorname{div} u.$$

In order to estimate $\|w\|_{L_2}$, for $x \in \Omega$ let us denote by γ_x a characteristic of the operator $\partial_{x_1} + U \cdot \nabla$ connecting x with Γ_{in}, i.e., a solution to

$$\begin{aligned} &\dot{\gamma_x}(t) = -(1 - U_1(\gamma_x(t)), U_2(\gamma_x(t))), \\ &\gamma_x(0) = x. \end{aligned} \tag{43}$$

Notice that we take the tangential vector with opposite sign since we consider a backwards characteristic starting at a given point inside Ω. Denote by $x^{\text{in}}(x)$ the intersection of γ_x with Γ_{in}. Due to regularity and the smallness of U, γ_x is close to a straight line $\{x_2 = c\}$. Now we can write

$$w(x) = w_{\text{in}}(x^{\text{in}}(x)) + \int_{\gamma_x} (G - \operatorname{div} u)\, dl_{\gamma_x}. \tag{44}$$

By Jensen's inequality we have

$$\left(\int_{\gamma} (G - \operatorname{div} u)\, dl_{\gamma_x} \right)^2 \le |\gamma_x| \int_{\gamma_x} |G|^2 + |\operatorname{div} u|^2\, dl_{\gamma_x}.$$

Hence applying (44) we get

$$\|w\|_{L_2(\Omega)} \le C(\Omega)(\|w_{\text{in}}\|_{L_2(\Gamma_{\text{in}})} + \|G\|_{L_2(\Omega)} + \|u\|_{W_2^1(\Omega)}). \tag{45}$$

Now we combine (42) and (45). By the smallness of U we can fix δ in (42) small enough to put the term $\delta + \|U\|_{W_p^{1+s}}$ on the left, obtaining (38), which completes the proof of the lemma. □

2.2. ***Steady transport equation.*** In this section we show the estimate in W_p^s for the steady transport equation with inflow condition, which is a crucial step in showing an a priori estimate for the linear problem (36).

Proposition 9. *Let Ω be a set defined at the end of Section 1, satisfying* (17)*. Let w solve*

$$\overline{K} w + w_{x_1} + U \cdot \nabla w = H, \quad w|_{\Gamma_{\text{in}}} = w_{\text{in}}, \tag{46}$$

where $\overline{K}$ is a positive constant, $U \in W_\infty^1(\Omega)$ is small and satisfies $U \cdot n|_\Gamma = d - n_1$, where d and n_1 satisfy (A1)*. Assume $H \in W_p^s(\Omega)$ and let w_{in}, s, and p satisfy the assumptions of Theorem 5. Then*

$$\|w\|_{W_p^s(\Omega)} \le C[\|H\|_{W_p^s(\Omega)} + \|w_{\text{in}}\|_{W_p^s(\Gamma_{\text{in}}) \cap W_p^r(\Gamma_s)}], \tag{47}$$

where $C = C(s, p, \Omega)$ and r is from (24)*.*

Remark 10. Notice that the assumption $U \in W_\infty^1(\Omega)$ is weaker than $U \in W_p^{1+s}(\Omega)$ for $sp > 2$ due to the imbedding theorem.

Remark 11. Proposition 9 generalizes the result from [Piasecki 2018], where it is assumed $U = [u^1, 0]$, which enables a much simpler proof than the one presented below and no additional regularity of the density around the singularity points is required.

For simplicity we set $\overline{K} = 1$, which is allowed as we assume anyway sufficient smallness of the data. The proof of Proposition 9 is quite technical since we have to treat carefully the boundary terms. For the reader's convenience we divide it into several lemmas.

Lemma 12. *Let the assumptions of Proposition 9 hold. Then*

$$\|w\|_{W_p^s(\Omega)} + \frac{1}{p}\iint_{\Omega^2} \frac{\partial_{x_1}|w(x)-w(y)|^p + \partial_{y_1}|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\,dx\,dy$$
$$+ \frac{1}{p}\iint_{\Omega^2} \frac{U(x)\cdot\nabla_x|w(x)-w(y)|^p + U(y)\cdot\nabla_y|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\,dx\,dy \le \|H\|_{W_p^s(\Omega)}\|w\|_{W_p^s}^{p-1}. \quad (48)$$

Proof. Recalling the definition of Sobolev–Slobodetskii norm we write (46) in x and y. Using identities of the type $\nabla_x w(y) = 0$ we can write

$$w(x) + \partial_{x_1}[w(x)-w(y)] + U(x)\cdot\nabla_x[w(x)-w(y)] = H(x),$$
$$w(y) + \partial_{y_1}[w(y)-w(x)] + U(y)\cdot\nabla_y[w(y)-w(x)] = H(y).$$

We multiply the first equation by

$$\frac{|w(x)-w(y)|^{p-2}(w(x)-w(y))}{\phi_\epsilon(x,y)}$$

and the second by

$$\frac{|w(x)-w(y)|^{p-2}(w(y)-w(x))}{\phi_\epsilon(x,y)},$$

where

$$\phi_\epsilon(x,y) = \epsilon + |x-y|^{2+sp}.$$

Then we add the equations and integrate twice over Ω, with respect to x and y. Since

$$w(x)(w(x)-w(y)) + w(y)(w(y)-w(x)) = [w(x)-w(y)]^2$$

and

$$H(x)(w(x)-w(y)) + H(y)(w(y)-w(x)) = (H(x)-H(y))(w(x)-w(y)),$$

we obtain on the left-hand side

$$\iint_{\Omega^2} \frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\,dx\,dy \to_{\epsilon\to 0} \|w\|_{W_p^s}^p. \quad (49)$$

On the right-hand side we have using the Hölder inequality

$$\iint_{\Omega^2} \frac{(H(x)-H(y))|w(x)-w(y)|^{p-2}(w(x)-w(y))}{\phi_\epsilon(x,y)}$$
$$\le \left(\iint_{\Omega^2} \frac{|H(x)-H(y)|^p}{\phi_\epsilon(x,y)}\right)^{1/p}\left(\iint_{\Omega^2} \frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\right)^{1-1/p} \to_{\epsilon\to 0} \|H\|_{W_p^s}\|w\|_{W_p^s}^{p-1}, \quad (50)$$

which completes the proof of (48). □

The rest of the proof of Proposition 9 consists in dealing with the integral terms in (48). This is where all the difficulties are hidden and our assumptions on the boundary and boundary data will come into play. We start with observing that under the assumptions of Proposition 9 we have

$$n_1 + U\cdot n \ge 0 \quad \text{on } \Gamma_{\text{out}}. \quad (51)$$

Indeed, we have

$$n(x)|_{\Gamma_{\text{out}}} = \frac{[1, -\bar{x}_1'(x_2)]}{\sqrt{1+[\bar{x}_1'(x_2)]^2}}, \quad n(x)|_{\Gamma_{\text{in}}} = \frac{[-1, \underline{x}_1'(x_2)]}{\sqrt{1+[\underline{x}_1'(x_2)]^2}}; \tag{52}$$

therefore by (A1)

$$\left|\frac{U\cdot n}{n_1}\right| = \sqrt{1+[\bar{x}_1'(x_2)]^2}|U\cdot n| \le \kappa,$$

and so (51) holds.

We now proceed with transforming (48). Since we want to have a W_p^s-norm, we have to get rid of the derivatives of w and for this purpose we integrate by parts. This step is presented in the following lemma:

Lemma 13. *Let the assumptions of Proposition 9 be satisfied. Then*

$$\|w\|^p_{W^s_p(\Omega)} + \frac{2}{p}\int_\Omega\int_{\Gamma_{\text{in}}} \frac{|w(x)-w(y)|^p(n_1+U\cdot n)(x)}{\phi_\epsilon}\,dS(x)\,dy \le \|H\|_{W^s_p(\Omega)}\|w\|^{p-1}_{W^s_p(\Omega)} + C\|\nabla U\|_{L_\infty(\Omega)}\|w\|^p_{W^s_p(\Omega)}. \tag{53}$$

Proof. We will obtain (53) from (48). Let us start with the first integral term on the left-hand side of (48). We have

$$\frac{\partial_{x_1}|w(x)-w(y)|^p}{\phi_\epsilon(x,y)} = \partial_{x_1}\left(\frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\right) - |w(x)-w(y)|^p\partial_{x_1}\left(\frac{1}{\phi_\epsilon(x,y)}\right). \tag{54}$$

By the definition of $\phi_\epsilon(x,y)$ we have

$$\nabla_x\phi_\epsilon(x,y) = -\nabla_y\phi_\epsilon(x,y); \tag{55}$$

therefore using (54) we get

$$\begin{aligned}\iint_{\Omega^2}&\left\{\frac{\partial_{x_1}|w(x)-w(y)|^p}{\phi_\epsilon(x,y)} + \frac{\partial_{y_1}|w(y)-w(x)|^p}{\phi_\epsilon(x,y)}\right\}dx\,dy\\ &= \iint_{\Omega^2}\left\{\partial_{x_1}\left(\frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\right) + \partial_{y_1}\left(\frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\right)\right\}dx\,dy\\ &\quad - \iint_{\Omega^2}\left\{|w(x)-w(y)|^p\partial_{x_1}\left(\frac{1}{\phi_\epsilon(x,y)}\right) + |w(x)-w(y)|^p\partial_{y_1}\left(\frac{1}{\phi_\epsilon(x,y)}\right)\right\}dx\,dy.\end{aligned} \tag{56}$$

Taking into account (55), the integrand in the second integral on the right-hand side of (56) vanishes identically. Combining (55) with the identities

$$\nabla_x|w(x)-w(y)|^p = p|w(x)-w(y)|^{p-2}(w(x)-w(y))\nabla_x w(x), \tag{57}$$

$$\nabla_y|w(x)-w(y)|^p = -p|w(x)-w(y)|^{p-2}(w(x)-w(y))\nabla_y w(y), \tag{58}$$

we see that the first integral on the right-hand side of (56) adds up to

$$\frac{2}{p}\iint_{\Omega^2}\partial_{x_1}\left(\frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\right)dx\,dy = \frac{2}{p}\int_\Omega\int_\Gamma \frac{|w(x)-w(y)|^p n_1}{\phi_\epsilon(x,y)}\,dS(x)\,dy. \tag{59}$$

Now consider the second integral on the left-hand side of (48). We have

$$\iint_{\Omega^2} U(x)\frac{\nabla_x|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\,dx\,dy$$
$$= \iint_{\Omega^2} U(x)\nabla_x\left(\frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\right)dx\,dy - \iint_{\Omega^2} U(x)|w(x)-w(y)|^p\nabla_x\left(\frac{1}{\phi_\epsilon(x,y)}\right)dx\,dy \quad (60)$$

and

$$\iint_{\Omega^2} U(y)\frac{\nabla_y|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\,dx\,dy$$
$$= \iint_{\Omega^2} U(y)\nabla_y\left(\frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\right)dx\,dy - \iint_{\Omega^2} U(y)|w(x)-w(y)|^p\nabla_y\left(\frac{1}{\phi_\epsilon(x,y)}\right)dx\,dy. \quad (61)$$

Recalling (55) we see that the terms with $\nabla(1/\phi)$ cancel. But this time the integrand does not vanish identically and we have to check whether the sum of these integrals makes sense when $\epsilon \to 0$. This sum equals

$$\left|\iint_{\Omega^2}[U(x)-U(y)]|w(x)-w(y)|^p\frac{(2+sp)|x-y|^{sp}(x-y)}{(\epsilon+|x-y|^{2+sp})^2}\,dx\,dy\right|$$
$$\longrightarrow_{\epsilon\to 0}(2+sp)\left|\iint_{\Omega^2}\frac{U(x)-U(y)}{x-y}\frac{|w(x)-w(y)|^p}{|x-y|^{2+sp}}\,dx\,dy\right| \le C\|\nabla U\|_{L_\infty}\|w\|^p_{W^s_p}.$$

Now consider the terms with $\nabla|w(x)-w(y)|^p/\phi_\epsilon$ on the right-hand side of (60) and (61). Combining (57) and (58) with (55) we see that these terms add up to

$$\frac{2}{p}\iint_{\Omega^2} U(x)\nabla_x\left(\frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\right)dx\,dy$$
$$= -\int_\Omega dy\int_\Omega \frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\,\mathrm{div}_x\,U(x)\,dx + \int_\Omega dy\int_\Gamma \frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}U(x)\cdot n\,dS(x). \quad (62)$$

The first integral is straightforward:

$$\left|\int_\Omega dy\int_\Omega \frac{|w(x)-w(y)|^p}{\phi_\epsilon(x,y)}\,\mathrm{div}_x\,U(x)\,dx\right| \le C(\Omega)\|\nabla U\|_{L_\infty(\Omega)}\|w\|^p_{W^s_p(\Omega)}. \quad (63)$$

Combining (48), (59), (62) and (63) we get (53). □

In order to complete the proof of Proposition 9 it remains to find an estimate on the integral term in (53). Notice that it is sufficient to consider the integral over Γ_{in} since the outflow part is nonnegative due to (51) and therefore we already skipped it in (53). However, the inflow term is negative because of n_1 and therefore must be estimated. The treatment of this term is quite technical and in fact constitutes the core of the proof of Proposition 9. Let us start with introducing some further notation. In the estimate which will follow, $y = (y_1, y_2)$ will denote a point inside Ω, while $\underline{x} = (\underline{x}_1, \underline{x}_2)$ is a point on Γ_{in}. In several places we will replace an integral with respect to the boundary measure on Γ_{in} by an integral with respect to x_2. Then we will write $\underline{x}(x_2) = (\underline{x}_1(x_2), x_2)$. At this point we should also fix η in the

definition (10). For this purpose observe that (A1) implies

$$|(d-n_1)|\sqrt{1+|\underline{x}_1'(x_2)|^2} \le \frac{\tilde{\kappa}}{\sqrt{1+\lambda^2}} \tag{64}$$

for some $\kappa < \tilde{\kappa} < 1$ and $\lambda > 0$. Now in (10) we choose η such that

$$|x_2'(x_1)| < \lambda \quad \text{on } \Gamma_{\text{in}}^s. \tag{65}$$

Notice that such η exists due to assumed regularity of the boundary. Next we define

$$\begin{aligned}
\Gamma_{\text{in}}^r &:= \{x \in \Gamma_{\text{in}} :\ \eta < x_2 < b-\eta\},\\
\Omega_{\text{in}}^r &:= \{x \in \Omega :\ \eta < x_2 < b-\eta,\ x_1 - \underline{x}_1(x_2) < \eta\},\\
\Omega_{\text{in}}^s &:= \{x \in \Omega :\ x_2 < 2\eta \ \vee\ b-2\eta < x < b\},\\
\Omega_{\text{in}} &= \Omega_{\text{in}}^r \cup \Omega_{\text{in}}^s.
\end{aligned} \tag{66}$$

Therefore, Γ_{in}^r and Ω_{in}^r are, respectively, a part of Γ_{in} and its neighborhood, which are at some fixed distance from singularity points. In particular, we have

$$|\underline{x}_1'(x_2)| \le C(\eta) \quad \text{for } (\underline{x}_1(x_2), x_2) \in \Gamma_{\text{in}}^r. \tag{67}$$

Notice also that $\Gamma_{\text{in}}^s = \Omega_{\text{in}}^s \cap \Gamma_{\text{in}}$. Next, for $y = (y_1, y_2)$ let us denote by γ_y a characteristic curve of the operator $(\partial_{x_1} + U \cdot \nabla)$ connecting y with Γ_{in}, i.e., a solution to (43) with $\gamma_y(0) = y$. Furthermore, we denote the intersection of γ_y with Γ_{in} by $y^{\text{in}}(y)$ (see Figure 3). We will need the following auxiliary result:

Lemma 14. *Assume that*

$$y \in \Omega_{\text{in}}^r, \quad \underline{x} \in \Gamma_{\text{in}}^r. \tag{68}$$

Then

$$|\underline{x} - y^{\text{in}}(y)| + |y^{\text{in}}(y) - y| \le C|\underline{x} - y|, \tag{69}$$

$$|\gamma_y| \le (1+E)|y_1 - \underline{x}_1(y_2)| \tag{70}$$

for some constant $C > 0$ and a small constant E.

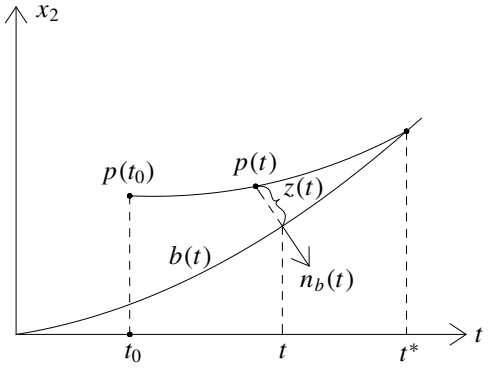

Figure 3. Explanation of notation used in Lemma 15.

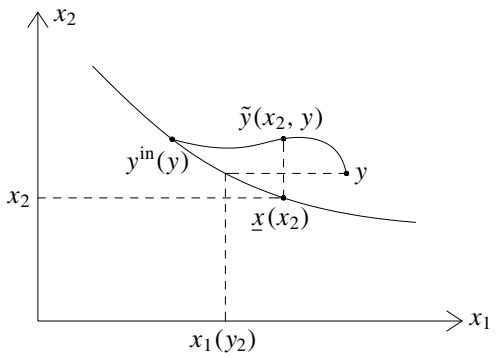

Figure 4. Illustration of notation. Case (b): intersection point $\tilde{y}$ well-defined.

Proof. Without loss of generality we can restrict to assuming that x_2 is a decreasing function of x_1 on Γ_{in} as the opposite case is analogous. Then it is convenient to distinguish three cases:

$$\begin{aligned} &\text{(a)}\ \ \underline{x}_1 \le y_1^{\mathrm{in}}(y) && \text{(see Figure 4).}\\ &\text{(b)}\ \ y_1^{\mathrm{in}}(y) < \underline{x}_1 < y_1 && \text{(see Figure 3).}\\ &\text{(c)}\ \ \underline{x}_1 \ge y_1. \end{aligned} \tag{71}$$

In the case (a) (69) is obvious with $C = 2$ as $|\underline{x} - y^{\mathrm{in}}(y)| \le |\underline{x} - y|$ and $|y^{\mathrm{in}}(y) - y| \le |\underline{x} - y|$. The case (c) is treated similarly to (b); therefore we don't show it on a separate figure and focus on the proof for (b). In that case it is convenient to define by $\tilde{y}(x_2, y)$ the intersection of a line $\{(\underline{x}_1(x_2), t) : t \in \mathbb{R}\}$ with a characteristic γ_y connecting Γ_{in} with y. Such an intersection is well-defined in the case (b); see Figure 3. We have for some $\vartheta \in [\underline{x}_2, y_2]$

$$\begin{aligned} |\underline{x} - y^{\mathrm{in}}(y)| &\le C(|\underline{x}_1 - y_1^{\mathrm{in}}(y)| + |\underline{x}_2 - y_2^{\mathrm{in}}(y)|)\\ &= (C\underline{x}_1'(\vartheta) + 1)|\underline{x}_2 - y_2^{\mathrm{in}}(y)|\\ &\le C|\underline{x}_2 - \tilde{y}_2(x_2, y)| \le C|\underline{x} - \tilde{y}(x_2, y)| \le C|\underline{x} - y|, \end{aligned} \tag{72}$$

where we have used (67). Therefore also

$$|y^{\mathrm{in}}(y) - y| \le |y^{\mathrm{in}}(y) - \underline{x}| + |\underline{x} - y| \le C|\underline{x} - y|,$$

which completes the proof of the first assertion. To show the second one we can assume without loss of generality that $y_2^{\mathrm{in}}(y) \ge y_2$. We have

$$|y_2^{\mathrm{in}}(y) - y_2| = \int_{y_1^{\mathrm{in}}(y)}^{y_1} U_2(t, \gamma(t))\, dt \le \|U\|_\infty |y_1^{\mathrm{in}}(y) - y_1|;$$

therefore using again (67) we get

$$\begin{aligned} |y_1 - y_1^{\mathrm{in}}(y)| &\le |y_1 - \underline{x}_1(y_2)| + |\underline{x}_1(y_2) - y_1^{\mathrm{in}}(y)| + |y_2^{\mathrm{in}}(y) - y_2|\\ &\le |y_1 - \underline{x}_1(y_2)| + (|\underline{x}_1'(\vartheta)| + 1)|y_2^{\mathrm{in}}(y) - y_2|\\ &\le |y_1 - \underline{x}_1(y_2)| + (|\underline{x}_1'(\vartheta)| + 1)\|U_2\|_\infty |y_1 - y_1^{\mathrm{in}}(y)|, \end{aligned}$$

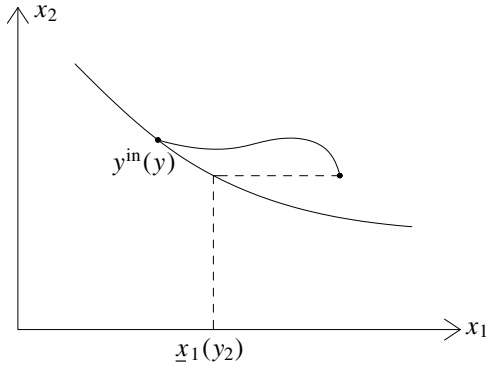

Figure 5. Case (a): no intersection point $\tilde{y}$.

so for sufficiently small $\|U\|_\infty$ we obtain

$$|y_1 - y_1^{\text{in}}(y)| \le (1+E)|y_1 - \underline{x}_1(y_2)|,$$

and to complete the proof it is enough to note that

$$|\gamma_y| = \int_{y_1^{\text{in}}(y)}^{y_1} \sqrt{[1+U_1(\gamma_y(s))]^2 + (U_2(\gamma_y(s)))^2}\, ds \le (1+E)|y_1 - y_1^{\text{in}}(y)|. \qquad \square$$

We shall emphasize that (69) holds except in a neighborhood of singularity points. The point is that $|\underline{x}_1'(x_2)|$ is unbounded as we approach the singularity. Therefore, as

$$|\underline{x}(x_2) - y^{\text{in}}(y_2)| \ge |\underline{x}_1(x_2) - y_1^{\text{in}}(y_2)| = |\underline{x}_1'(\vartheta)||x_2 - y_2| \tag{73}$$

for some $\vartheta \in [x_2, y_2]$, we no longer have (69). In order to control $|\underline{x} - y^{\text{in}}(y)|$ we need an appropriate estimate on a length of a characteristic connecting y with the boundary and here the assumptions (16) and (A1) come into play. The required result is:

Lemma 15. *Let $b(t) = (t, \underline{x}_2(-t))$ and let $p(t)$ satisfy*

$$\dot{p}(t) = (1 + U^1(p(t)), U^2(p(t))), \quad p(t_0) = x_2^0, \tag{74}$$

where x_2^0 is a given, small positive number, t_0 is sufficiently small so that

$$p(s) > b(s) \quad \text{for } s \le t_0,$$

and U satisfies the assumptions of Proposition 9. Let t_* denote the first coordinate of the first intersection of $p(t)$ and $b(t)$ (see Figure 5). Then

$$t_* - t_0 \le C(x_2^0 - b(t_0))^\delta,$$

where δ is from (17) *and $C = C(U, \Omega)$.*

Remark 16. We will apply the above lemma with $t = -x_1$ in some neighborhood of a singularity point $(0, 0)$ in the proof of Lemma 17 which will follow; therefore we consider a "backward" characteristic of (46).

Proof of Lemma 15. Assume first that $t_0 = 0$. Let us define

$$A(t) = \sqrt{1 + |\underline{x}_2'(t)|^2}.$$

Then the unit normal to $b(t)$ is $n_b(t) = A(t)^{-1}(-\underline{x}_2'(t), 1)$. Let $z(t)$ denote the distance between $b(t)$ and $p(t)$ measured along $n_b(t)$ (see Figure 5); that is,

$$p(t) = b(t) + n_b(t)z(t).$$

Differentiating this identity in t and multiplying the resulting equation by $n_b(t)$ we get

$$\dot{p}(t) \cdot n_b(t) = \dot{z}(t), \tag{75}$$

which by (74) can be rewritten as

$$\dot{z}(t) = U \cdot n_b(t, z) - \frac{\underline{x}_2'(t)}{A(t)}. \tag{76}$$

We have

$$|U \cdot n_b(t, z)| \le |U \cdot n_b(t, 0)| + \|\nabla U\|_{L_\infty} z;$$

therefore by (76)

$$\dot{z}(t) \le |U \cdot n_b(t, 0)| - \frac{\underline{x}_2'(t)}{A(t)} + \|\nabla U\|_{L_\infty} z(t). \tag{77}$$

Now, due to (64) and (65) we have

$$|U \cdot n_b(t, 0)| \le \frac{\tilde{\kappa}|\underline{x}_2'(t)|}{\sqrt{1+\lambda^2}} < \tilde{\kappa}\frac{|\underline{x}_2'(t)|}{A(t)}; \tag{78}$$

therefore

$$\dot{z}(t) \le -(1-\tilde{\kappa})\frac{\underline{x}_2'(t)}{A(t)} + E_U z(t) \le -\vartheta + E_U z(t), \tag{79}$$

with $\vartheta = (1-\tilde{\kappa})/\sqrt{1+\lambda^2}$ and $E_U = \|\nabla U\|_{L_\infty}$. Now t_* must satisfy

$$\int_0^{t^*} \dot{z}(t)\,dt = -x_2^0; \quad \text{therefore} \quad x_2^0 = \vartheta \underline{x}_2(t_*) - E_U \underbrace{\int_0^{t_*} z(t)\,dt}_{Z(t_*)}. \tag{80}$$

In order to simplify the above condition observe that

$$z(t) \le x_2^0 e^{E_U t} \le C x_2^0, \quad p(t) - b(t) \le C z(t),$$

and therefore

$$Z(t_*) \le t \sup_{s \le t}[p(t) - b(t)] \le C_2 t_* x_2^0. \tag{81}$$

Moreover, by (16) we have

$$\underline{x}_2(t_*) \ge C_1 t_*^N. \tag{82}$$

In view of (81) and (82), if we define $\tilde{t}_*$ by

$$P_{x_2^0}(\tilde{t}_*) := C_1 \vartheta \tilde{t}_*^N - C_2 E_U x_2^0 \tilde{t}_* - x_2^0 = 0,$$

then, because of (80), (81) and (82),

$$t_* \le \tilde{t}_*. \tag{83}$$

We claim that there exists $1 \le C_3 = C_3(C_1, C_2, \Omega)$ such that if $t_1 = C_3(x_2^0)^{1/N}$ then $P(t_1) \ge 0$. Indeed, we have

$$P(t_1) = x_2^0[C_1\vartheta C_3 - 1 - C_2E_U(C_3x_2^0)^{1/N}] \ge C_3(C_1 - C_2E_U(x_2^0)^{1/N}) - 1 > 0$$

for sufficiently large $C_3 = C_3(C_1, C_2, \Omega)$. Therefore,

$$\tilde{t}_* \le t_1 \le C_3(x_2^0)^{1/N},$$

which together with (83) completes the proof for $t_0 = 0$ as $\delta = 1/N$. For $t_0 > 0$ the reasoning is the same modulo minor adjustments, replacing x_2^0 with $-z(t_0)$ in (80). We are now in a position to prove the estimate on the boundary term on the left-hand side of (53).

Lemma 17. *Let the assumptions of Proposition 9 hold. Then*

$$\int_\Omega dy \int_{\Gamma_{\text{in}}} \frac{|w(x) - w(y)|^p (n_1 + U\cdot n)(x)}{\phi_\epsilon(x, y)}\, dS(x) \le C[\|H\|_{L_\infty(\Omega)} + \|w\|_{L_p(\Omega)} + \|w\|_{L_\infty(\Omega)} + \|w_{\text{in}}\|_{W_p^s(\Gamma_{\text{in}})} + \|w_{\text{in}}\|_{W_p^r(\Gamma_{\text{in}}^s)}]^p. \tag{84}$$

Proof. First of all, due to (52) and (A1) we have

$$\left|\frac{U\cdot n}{n_1}\right| \le \kappa,$$

which gives

$$|n_1 + U\cdot n| \le (1+\kappa)n_1. \tag{85}$$

Therefore it is enough to show (84) for $U\cdot n = 0$. As

$$\begin{aligned} dS(x_2)|_{\Gamma_{\text{out}}} &= \sqrt{1 + [\bar{x}_1'(x_2)]^2}dx_2, \\ dS(x_2)|_{\Gamma_{\text{in}}} &= \sqrt{1 + [\underline{x}_1'(x_2)]^2}dx_2, \end{aligned} \tag{86}$$

(52) yields

$$dS(x_2)|_{\Gamma_{\text{in}}} = -\frac{dx_2}{n_1}, \quad n_2 dS(x_2)|_{\Gamma_{\text{in}}} = \underline{x}_1'(x_2)dx_2. \tag{87}$$

By (87), assuming $U\cdot n = 0$ the left-hand side of (84) can be rewritten as

$$\int_\Omega \int_{\Gamma_{\text{in}}} \frac{|w(x) - w(y)|^p n_1(x)}{\phi_\epsilon(x, y)}\, dS(x)\, dy = \int_\Omega dy \underbrace{\int_0^b -\frac{|w(\underline{x}_1(x_2), x_2) - w(y)|^p}{\phi_\epsilon((\underline{x}_1(x_2), x_2), y)}\, dx_2}_{I_\epsilon(y)}, \tag{88}$$

and it remains to show the estimate (84) for (88). When we take $\epsilon \to 0$, we can expect some problems only when y is close to Γ_{in}. Therefore we write (88) as

$$\int_\Omega I_\epsilon(y)\, dy = \int_{\Omega_{\text{in}}^r} I_\epsilon(y)\, dy + \int_{\Omega_{\text{in}}^s} I_\epsilon(y)\, dy + \int_{\Omega\setminus\Omega_{\text{in}}} I_\epsilon(y)\, dy =: I_1^r + I_1^s + I_2,$$

where the sets $\Omega^r_{\rm in}$, $\Omega^s_{\rm in}$ and $\Omega_{\rm in}$ are defined in (66). The set $\Omega\setminus\Omega_{\rm in}$ consists of points at the distance from $\Gamma_{\rm in}$ bounded from below by η. Therefore the I_2 integral is straightforward as we have $|(\underline{x}_1(x_2),x_2)-y|>\eta$ for $y\in\Omega\setminus\Omega_{\rm in}$ and so

$$I_2\le C(\eta)[\|w_{\rm in}\|_{L_p(\Gamma_{\rm in})}+\|w\|_{L_p}]^p. \tag{89}$$

The estimates for I^r_2 and I^s_1 are more involved; we show them in detail.

Estimate of I^r_1: Let $\underline{x}(x_2)=(\underline{x}_1(x_2),x_2)\in\Gamma_{\rm in}$. We have

$$\frac{|w(\underline{x}(x_2))-w(y)|^p}{|\underline{x}(x_2)-y|^{2+sp}}\le C(p)\bigg[\frac{|w(\underline{x}(x_2))-w(y^{\rm in}(y))|^p}{|\underline{x}(x_2)-y|^{2+sp}}+\frac{|w(y^{\rm in}(y))-w(y)|^p}{|\underline{x}(x_2)-y|^{2+sp}}\bigg]. \tag{90}$$

For the first term on the right-hand side of (90) we have

$$\begin{aligned}&\int_{\Omega^r_{\rm in}}dy\int_0^b\frac{|w(\underline{x}(x_2))-w(y^{\rm in}(y))|^p}{|\underline{x}(x_2)-y|^{2+sp}}\,dx_2\\&\quad=\int_{2\eta}^{b-2\eta}dx_2\bigg[\int_{V_\eta(x_2)}\frac{|w(\underline{x}(x_2))-w(y^{\rm in}(y))|^p}{|\underline{x}(x_2)-y|^{2+sp}}\,dy+\int_{\Omega^r_{\rm in}\setminus V_\eta(x_2)}\frac{|w(\underline{x}(x_2))-w(y^{\rm in}(y))|^p}{|\underline{x}(x_2)-y|^{2+sp}}\,dy\bigg],\end{aligned} \tag{91}$$

where

$$V_\eta(x_2)=\{y\in\Omega^r_{\rm in}:\ x_2-\eta<y_2<x_2+\eta\}, \tag{92}$$

which implies that in the second integral $|\underline{x}(x_2)-y|$ is bounded from below by η; therefore

$$\begin{aligned}\int_0^b dx_2\int_{\Omega^r_{\rm in}\setminus V_\eta(x_2)}\frac{|w(\underline{x}(x_2))-w(y^{\rm in}(y))|^p}{|\underline{x}(x_2)-y|^{2+sp}}\,dy&\le C\bigg[\int_0^b|w(\underline{x}(x_2))|^p\,dx_2+\int_0^b|w(y^{\rm in}(y))|^p\,dy_2\bigg]\\&\le C\|w_{\rm in}\|_{L_p(\Gamma_{\rm in})},\end{aligned}$$

where in the last inequality we have used an obvious inequality

$$dy_2\le dS(y_2). \tag{93}$$

In order to treat the first integral in (91) we apply (69)–(70). Due to (70) we can assume without loss of generality for this estimate that γ_y is a straight line $y_2\equiv\text{const}$; therefore $y^{\rm in}(y)=\underline{y}(y)$. Then by the definitions of $V_\eta(x_2)$ and $\Omega^r_{\rm in}$ ((92) and (66), respectively)

$$\begin{aligned}&\int_{V_\eta(x_2)}\frac{|w(\underline{x}(x_2))-w(y^{\rm in}(y))|^p}{|\underline{x}(x_2)-y|^{2+sp}}\,dy\\&\quad=\int_{x_2-\eta}^{x_2+\eta}\int_{\underline{y}_1(y_2)}^{\underline{y}_1(y_2)+\eta}\frac{|w(\underline{x}(x_2))-w(y^{\rm in}(y_2))|^p}{|\underline{x}(x_2)-y|^{2+sp}}\,dy_1\,dy_2\\&\quad\le\int_{x_2-\eta}^{x_2+\eta}|w(\underline{x}(x_2))-w(y^{\rm in}(y_2))|^p\,dy_2\int_{\underline{y}_1(y_2)}^{\underline{y}_1(y_2)+\eta}(y_1-\underline{y}_1(y)+|\underline{x}(x_2)-y^{\rm in}(y_2)|)^{-2-sp}\,dy_1\\&\quad\le\int_{x_2-\eta}^{x_2+\eta}\frac{|w(\underline{x}(x_2))-w(y^{\rm in}(y_2))|^p}{|\underline{x}(x_2)-y^{\rm in}(y_2)|^{1+sp}}\,dy_2,\end{aligned}$$

where in the first inequality we have used (69). Therefore, as we have additionally (93),

$$\int_{2\eta}^{b-2\eta} dx_2 \int_{V_\eta(x_2)} \frac{|w(\underline{x}(x_2)) - w(y^{\mathrm{in}}(y))|^p}{|\underline{x}(x_2) - y|^{2+sp}}\, dy \le C\|w\|_{W^s_p(\Gamma_{\mathrm{in}})}. \tag{94}$$

Notice that in this part of the estimate the norm of boundary data appears naturally.

In the second term on the right-hand side of (90) we express $w(y) - w(y^{\mathrm{in}}(y))$ by an integral along γ_y using (46). Applying Jensen's inequality we get

$$\begin{aligned}
\int_{\Omega^r_{\mathrm{in}}} dy \int_0^b \frac{|w(y^{\mathrm{in}}(y)) - w(y)|^p}{|\underline{x}(x_2) - y|^{2+sp}}\, dx_2 &\le \int_{\Omega^r_{\mathrm{in}}} dy \int_0^b \frac{|\int_{\gamma_y} (H - w)\, dl_{\gamma_y}|^p}{|\underline{x}(x_2) - y|^{2+sp}}\, dx_2 \\
&\le (\|H\|_{L_\infty} + \|w\|_{L_\infty})^p \int_{\Omega^r_{\mathrm{in}}} dy \int_0^b \frac{|\gamma_y|^p}{|\underline{x}(x_2) - y|^{2+sp}}\, dx_2 \\
&\le C(\|H\|_{L_\infty} + \|w\|_{L_\infty})^p \int_{\Omega^r_{\mathrm{in}}} dy \int_0^b \frac{|y^{\mathrm{in}}(y) - y|^p}{|\underline{x}(x_2) - y|^{2+sp}}\, dx_2 \\
&\le C(\|H\|_{L_\infty} + \|w\|_{L_\infty})^p \int_0^b dx_2 \int_{\Omega^r_{\mathrm{in}}} \frac{|\underline{x}(x_2) - y|^p}{|\underline{x}(x_2) - y|^{2+sp}}\, dy,
\end{aligned} \tag{95}$$

where we have used (69) and (70). The last integral is finite for $s < 1$; we remember Ω^r_{in} is two-dimensional. Combining (90), (94) and (95) we get

$$I^r_1 \le C(\|H\|_{L_\infty(\Omega)} + \|w\|_{L_\infty(\Omega)} + \|w_{\mathrm{in}}\|_{W^s_p(\Gamma_{\mathrm{in}})})^p. \tag{96}$$

Estimate of I^s_1: Around the singularity points we have to proceed more carefully. If we consider again separately the cases (71), then in the case (a) (69) still holds with $C = 2$ so we can repeat directly the previous approach and it remains to consider the cases (b) and (c). The key difference is that, as we already explained, $|\underline{x}'_1(x_2)|$ is unbounded. Therefore we cannot repeat directly (94) and (95). We have to control $|\underline{x} - y^{\mathrm{in}}(y)|$ and here we use the assumption (17). Recall the notions of γ_y and $\tilde{y}$ introduced in (43) and after (71), respectively; see Figure 3. Let us denote by $\gamma_{\tilde{y}}$ the part of the γ_y connecting Γ_{in} with $\tilde{y}(x_2, y)$. We modify (90) adding the point $\tilde{y}(x_2, y)$:

$$\begin{aligned}
&\frac{|w(\underline{x}(x_2)) - w(y)|^p}{|\underline{x}(x_2) - y|^{2+sp}} \\
&\quad \le C(p)\left[\frac{|w(\underline{x}(x_2)) - w(y^{\mathrm{in}}(y))|^p}{|\underline{x}(x_2) - y|^{2+sp}} + \frac{|w(y^{\mathrm{in}}(y)) - w(\tilde{y}(x_2, y))|^p}{|\underline{x}(x_2) - y|^{2+sp}} + \frac{|w(\tilde{y}(x_2, y)) - w(y)|^p}{|\underline{x}(x_2) - y|^{2+sp}}\right].
\end{aligned} \tag{97}$$

We start with the second and third terms on the right-hand side of (97). Much as before, in both terms we replace the value of w with an integral along γ_y. The last term is analogous to I^r_1 and we get

$$\int_{\Omega^s_{\mathrm{in}}} dy \int_0^b \frac{|w(\tilde{y}(x_2, y)) - w(y)|^p}{|\underline{x}(x_2) - y|^{2+sp}}\, dx_2 \le C(\|H\|_{L_\infty(\Omega)} + \|w\|_{L_\infty(\Omega)})^p \quad \text{for } s < 1. \tag{98}$$

In the second term we have

$$\begin{aligned}|w(\tilde{y}(x_2,y))-w(y^{\mathrm{in}}(y))| &= \left|\int_{\gamma_{\tilde{y}}}(H-w)\,dl_{\gamma_{\tilde{y}}}\right| \\ &\le (\|H\|_{L^\infty}+\|w\|_{L^\infty})\int_{y_1^{\mathrm{in}}(y)}^{\underline{x}_1(x_2)}\sqrt{[1+U_1(\gamma_y(s))]^2+(U_2(\gamma_y(s)))^2}\,ds \\ &\le C(\|H\|_{L^\infty}+\|w\|_{L^\infty})|y_1^{\mathrm{in}}(y)-\underline{x}_1(x_2)|.\end{aligned}$$

Now we estimate the latter term. This is the most subtle part of the whole proof, and which has been carried out in Lemma 15. Applying this result with $t_0=-\underline{x}_1(x_2)$ we have

$$|y_1^{\mathrm{in}}(y)-\underline{x}_1(x_2)|\le C|x_2-\tilde{y}_2(x_2,y)|^{\delta}, \tag{99}$$

where δ is from (17). Next,

$$|x_2-\tilde{y}_2(x_2,y)|\le |x_2-y_2|+\|U_2\|_{L^\infty(\Omega)}|x_1-y_1|\le C|\underline{x}-y|;$$

therefore

$$\begin{aligned}\int_{\Omega_{\mathrm{in}}^s}dy\int_0^b\frac{|w(y^{\mathrm{in}}(y))-w(\tilde{y}(x_2,y))|^p}{|\underline{x}(x_2)-y|^{2+sp}} &\le \int_{\Omega_{\mathrm{in}}^s}dy\int_0^b\frac{|\int_{\gamma_{\tilde{y}}}(H-w)\,dl_{\gamma_{\tilde{y}}}|^p}{|\underline{x}(x_2)-y|^{2+sp}}\,dx_2 \\ &\le C(\|H\|_{L^\infty}+\|w\|_{L^\infty})^p\int_{\Omega_{\mathrm{in}}^s}dy\int_0^b\frac{|x_2-\tilde{y}_2(x_2,y)|^{\delta p}}{|\underline{x}(x_2)-y|^{2+sp}}\,dx_2 \\ &\le C(\|H\|_{L^\infty}+\|w\|_{L^\infty})^p\int_{\Omega_{\mathrm{in}}^s}dy\int_0^b|\underline{x}(x_2)-y|^{\delta p-2-sp}\,dx_2,\end{aligned}$$

where in the second inequality we used (99). The last integral is finite for $s<\delta$ and we conclude

$$\int_{\Omega_{\mathrm{in}}^s}dy\int_0^b\frac{|w(y^{\mathrm{in}}(y))-w(\tilde{y}(x_2,y))|^p}{|\underline{x}(x_2)-y|^{2+sp}}\,dx_2\le C(\|H\|_{L^\infty}+\|w\|_{L^\infty})^p\quad\text{for } s<\delta. \tag{100}$$

It remains to estimate the first term on the right-hand side of (97). For this purpose we assume $\underline{x}\ne y^{\mathrm{in}}(y)$ since otherwise this term vanishes. By Fubini's theorem we have $dy=dS(y_2)\,dy_1$; therefore we can write

$$\begin{aligned}\int_{\Omega_{\mathrm{in}}^s}dy\int_0^b\frac{|w(\underline{x}(x_2))-w(y^{\mathrm{in}}(y))|^p}{|\underline{x}(x_2)-y|^{2+sp}}\,dx_2 &\le C\int_{\Gamma_{\mathrm{in}}^s}dS(y_2)\int_0^b dx_2\int_{\underline{y}_1(y_2)}^{\bar{y}_1(y_2)}\frac{|w(\underline{x}(x_2))-w(\underline{y}(y_2))|^p}{[|\underline{x}_1(x_2)-y_1|+|x_2-y_2|]^{2+sp}}\,dy_1 \\ &\le C\int_{\Gamma_{\mathrm{in}}^s}dS(y_2)\int_0^b dx_2\frac{|w(\underline{x}(x_2))-w(\underline{y}(y_2))|^p}{|x_2-y_2|^{1+sp}}.\end{aligned} \tag{101}$$

Introducing $h=x_2-y_2$ and using the imbedding (8) we get

$$\begin{aligned}\int_{\Gamma_{\mathrm{in}}^s}dS(y_2)\int_0^b dx_2\frac{|w(\underline{x}(x_2))-w(\underline{y}(y_2))|^p}{|x_2-y_2|^{1+sp}} &\le C\int_0^{h_0}\frac{dh}{|h|^{1+sp}}\int_{\Gamma_{\mathrm{in}}^{s,h}}|w(\underline{x}(y_2+h))-w(\underline{y}(y_2))|^p\,dS(y_2) \\ &\le C\|w\|_{B^r_{p,\infty}(\Gamma_{\mathrm{in}}^s)}\int_0^{h_0}h^{-1-sp}\sup_{y_2\in(0,\eta)}|\underline{x}(y_2+h)-\underline{y}(y_2)|^{rp}\,dh \\ &\le C\|w\|_{W^r_p(\Gamma_{\mathrm{in}}^s)}\int_0^{h_0}h^{-1-sp}\sup_{y_2\in(0,\eta)}|\underline{x}(y_2+h)-\underline{y}(y_2)|^{rp}\,dh,\end{aligned} \tag{102}$$

where

$$h_0 = \sup\{x_2 : (\underline{x}_1(x_2), x_2) \in \Gamma^s_{\text{in}}\} \quad \text{and} \quad \Gamma^{s,h}_{\text{in}} = \{\underline{x}(x_2) : \underline{x}(x_2+h) \in \Gamma^s_{\text{in}}\}.$$

In the second inequality in (102) we have used (7) and in the last one (8). Now we apply Lemma 15 with

$$t_0 = -x_1, \quad p(t_0) = (y_1^{\text{in}}(y_2), y_2 + h), \quad p(t) = \gamma_{p(t_0)}(-t),$$

where $\gamma_y(\cdot)$ is a characteristic passing through y defined in (43). Then Lemma 15 implies

$$|\underline{x}(y_2+h) - \underline{y}(y_2)| \le C|\underline{x}(y_2+h) - p(t_0)| \le C|h|^{\delta}, \tag{103}$$

which together with (102) gives

$$\int_{\Gamma^s_{\text{in}}} dS(y_2) \int_0^b dx_2 \frac{|w(\underline{x}(x_2)) - w(\underline{y}(y_2))|^p}{|x_2 - y_2|^{1+sp}} \le C\|w\|_{W^r_p(\Gamma^s_{\text{in}})} \int_0^{h_0} h^{\delta rp - 1 - sp}. \tag{104}$$

The last integral is finite provided

$$\delta r > s,$$

which implies the second relation in (24). Combining (98), (100) and (104) we conclude

$$I_1^s \le C(\|H\|_{L_\infty} + \|w\|_{L_\infty} + \|w_{\text{in}}\|_{W^r_p(\Gamma_s)})^p. \tag{105}$$

Putting together (89), (96) and (105) we get (84) with $U \cdot n = 0$, which completes the proof due to (85). □

Proof of Proposition 9. Now we are ready to close the estimate (47). Combining (53) with (84) we obtain

$$\|w\|^p_{W^s_p} \le C(s, p, \Omega)\big[\|H\|_{W^s_p}\|w\|^{p-1}_{W^s_p} + \|w\|^p_{L_\infty} + \|w\|^p_{L_p} + \|w_{\text{in}}\|_{W^s_p(\Gamma_{\text{in}})} + \|w_{\text{in}}\|_{W^r_p(\Gamma_s)} + \|\nabla U\|_{L_\infty}\|w\|^p_{W^s_p}\big] \tag{106}$$

and applying the interpolation inequality (6) we conclude (47). □

Remark 18. Notice that from (106) it follows that the constant in (47) is of the form

$$C = \frac{C(s, p, \Omega)}{1 - \|\nabla U\|_{L_\infty}};$$

therefore under the assumption of smallness of $\|\nabla U\|_{L_\infty}$ we can assume that it does not depend on $\|\nabla U\|_{L_\infty}$.

Remark 19. The proof of Proposition 9 is quite technical; therefore it may be unclear to the reader how it can be extended to a more general domain. To clarify it, we shall emphasize that the most subtle part is the estimate of $|w(\underline{x}) - w(y)|^p / |\underline{x} - y|^{2+sp}$, where $\underline{x}$ is on the boundary and y is inside Ω, both close to singularity. The most delicate point is to estimate the distance between the point where a characteristic crossing y intersects the boundary and $\underline{x}$. This was carried out in Lemma 15. On the other hand, a larger distance between $\underline{x}$ and y does not create additional problems. Therefore the proof presented here can be extended to the general case of multiple singularity points. We have to consider again separately the "regular" part of the boundary where $\underline{x}'_2(x_1)$ is bounded by a given constant and therefore we get the estimates (89), (96), and neighborhoods of each singularity point where we repeat the estimate (105).

This concludes the proof of Lemma 15. □

2.3. ***Linear estimate in $W_p^{1+s} \times W_p^s$ and solution of the linear system.*** In this section we solve the linear problem (36) with right-hand side of regularity determined in (37). First we show the a priori estimate:

Proposition 20. *Let (u, w) solve the system* (36) *with the right-hand side of regularity determined in* (37), *with s, p, and the boundary data satisfying the assumptions of Theorem 5. Then we have*

$$\|u\|_{W_p^{1+s}(\Omega)} + \|w\|_{W_p^s(\Omega)} \le C\,[\|F\|_{L_p(\Omega)} + \|G\|_{W_p^s(\Omega)} + \|w_{\mathrm{in}}\|_{W_p^s(\Gamma_{\mathrm{in}})\cap W_p^r(\Gamma_s)} + \|B\|_{W_p^{1-1/p}(\Gamma)}]. \tag{107}$$

Remark 21. The main difficulty in the proof of (107) lies in the steady transport equation which has been dealt with in Proposition 9. The rest of the proof uses classical results from the theory of elliptic problems and can be divided into several steps.

Proof of Proposition 20. Step I: Let us take the rotation of $(36)_1$. Then we get the system

$$\begin{aligned} -\mu\Delta \operatorname{rot} u &= \operatorname{rot}(F - \partial_{x_1} u) \quad \text{in } \Omega, \\ \operatorname{rot} u &= \left(2\chi - \frac{f}{\nu}\right) u\cdot\tau + \frac{B}{\mu} \quad \text{at } \Gamma. \end{aligned} \tag{108}$$

In order to show the boundary relation for the vorticity we differentiate (36c) in the tangential direction and apply (36d). The details are shown in [Mucha and Rautmann 2006] in the case $u\cdot n = d,\ B = 0$. A minor modification for our case $u\cdot n = 0,\ B \neq 0$, yields (108). We find the following estimate for the solution to (108):

$$\|\operatorname{rot} u\|_{W_p^1(\Omega)} \le C(\|F\|_{L_p(\Omega)} + \|u, B\|_{W_p^{1-1/p}(\Gamma)} + \|\partial_{x_1} u\|_{L_p(\Omega)}). \tag{109}$$

The solvability of (108) belongs to the classical theory of elliptic linear problems; hence we omit the details of this construction.

Step II: Consider the Helmholtz decomposition of u:

$$u = \nabla\phi + \nabla^\perp A. \tag{110}$$

We have $0 = n\cdot\nabla^\perp A = \tau\cdot\nabla A$. Therefore A satisfies

$$\Delta A = \operatorname{rot} u, \quad A|_\Gamma = \text{const}. \tag{111}$$

As $\partial\Omega \in C^3$, from (111) we obtain

$$\|A\|_{W_p^3(\Omega)} \le C\|\operatorname{rot} u\|_{W_p^1(\Omega)}. \tag{112}$$

Substituting (110) into (36a) we get

$$\nabla(-(2\mu+\nu)\operatorname{div} u + Kw) = F - \partial_{x_1} u + \mu\Delta\nabla^\perp A + (\mu+\nu)\nabla\operatorname{div}\nabla^\perp A =: \bar{F}.$$

By (109) and (112) we have

$$\|\bar{F}\|_{L_p(\Omega)} \le C[\|F\|_{L_p(\Omega)} + \|\partial_{x_1} u\|_{L_p(\Omega)} + \|u, B\|_{W_p^{1-1/p}(\Gamma)}];$$

therefore setting

$$-(2\mu+\nu)\operatorname{div} u + Kw = H \tag{113}$$

we have for any $\delta > 0$

$$\begin{aligned}\|H\|_{W^s_p(\Omega)} &\le \delta\|\nabla H\|_{L_p(\Omega)} + C(\delta)\|H\|_{L_2(\Omega)}\\ &\le \delta[\|F\|_{L_p(\Omega)} + \|\partial_{x_1}u\|_{L_p(\Omega)} + \|u,\, B\|_{W_p^{1-1/p}(\Gamma)}] + C(\delta)\|\nabla u,\, w\|_{L_2(\Omega)}. \end{aligned} \tag{114}$$

Step III: Adding (36b) to (113) with suitable scaling we obtain

$$\partial_{x_1} w + U\cdot\nabla w + \overline{K} w = \overline{G}, \tag{115}$$

with $\overline{K} = K/(2\mu+\nu)$ and

$$\|\overline{G}\|_{W^s_p(\Omega)} \le \|G\|_{W^s_p(\Omega)} + [\text{RHS of (114)}].$$

Here we meet the main mathematical challenge of our result; we have to solve this transport-like problem in a domain which touches the characteristics at the ends of Γ_{in}. Its solvability is given by Proposition 9. In particular, the estimate (47) yields

$$\|w\|_{W^s_p(\Omega)} \le C(\|G\|_{W^s_p(\Omega)} + \|w_{\text{in}}\|_{W^s_p(\Gamma_{\text{in}})\cap W^r_p(\Gamma_s)}). \tag{116}$$

Remark 22. Notice that Remark 18 implies that we can assume that the constant in (116) does not depend on $\|U\|_{W^1_\infty}$.

Step IV: We collect the elements of our estimation. In order to get the information about the velocity we use estimates for $\operatorname{rot} u$ given by (109) and about $\operatorname{div} u$ coming from (113), (114) and (116). We obtain

$$\begin{aligned}\|u\|_{W_p^{1+s}(\Omega)} &\le C(\|\operatorname{rot} u\|_{W^s_p(\Omega)} + |\operatorname{div} u\|_{W^s_p(\Omega)})\\ &\le C(\|F\|_{L_p(\Omega)} + \|G\|_{W^s_p(\Omega)} + \|B\|_{W^{1-1/p}(\Gamma)} + \|\nabla u, w\|_{L_2(\Omega)} + \|\partial_{x_1}u\|_{L_p} + \|u\|_{W^{1-1/p}(\Gamma)})\\ &\qquad + \delta(\|u\|_{W_p^{1+s}(\Omega)} + \|w\|_{W^s_p(\Omega)}). \end{aligned} \tag{117}$$

The L_2 term on the right-hand side is treated with the energy estimate (38). To the boundary term $\|u\|_{W^{1-1/p}(\Gamma)}$ we apply first the trace theorem and then the interpolation inequality (6) to get

$$\|u\|_{W_p^{1-1/p}(\Gamma)} \le C\|u\|_{W^s_p(\Omega)} \le \delta\|u\|_{W_p^{1+s}(\Omega)} + C(\delta)\|\nabla u\|_{L_2(\Omega)}. \tag{118}$$

The term $\partial_{x_1}u$ is treated similarly with the interpolation inequality. We conclude (107) and complete the proof. □

Now it is a matter of standard theory to show the existence for the linear system (36).

Proposition 23. *Let the right-hand side of* (36) *be of regularity specified in* (37) *with* s, p, *and the boundary data satisfying the assumptions of Theorem 5. Then there exists a unique solution to* (36) *satisfying the estimate* (107).

Proof. As we have the a priori estimate, the proof of existence is standard. We start with showing the weak solutions using the Galerkin method and the energy estimate (38). Then, using our a priori estimate we show that our solution has the required regularity. □

2.4. ***Estimate for the nonlinear system.*** We are now ready to close the estimate in $W_p^{1+s} \times W_p^s$ for the nonlinear system. To this end we combine the linear estimate (107) with the bounds (32) on the right-hand side of the nonlinear problem obtaining

$$\|u\|_{W_p^{1+s}(\Omega)} + \|w\|_{W_p^s(\Omega)} \le E(\|u\|_{W_p^{1+s}(\Omega)} + \|w\|_{W_p^s(\Omega)} + 1) + C(\|u\|_{W_p^{1+s}(\Omega)} + \|w\|_{W_p^s(\Omega)})^2,$$

where E is a small constant dependent on the data and we can assume (see Remark 22) that C does not depend on $\|u\|_{W_\infty^1(\Omega)}$. For sufficiently small data we conclude

$$\|u\|_{W_p^{1+s}(\Omega)} + \|w\|_{W_p^s(\Omega)} \le C(\|u\|_{W_p^{1+s}(\Omega)} + \|w\|^2_{W_p^s(\Omega)} + E), \tag{119}$$

where C does not depend on $(\|w\|_{W_p^s(\Omega)}, \|u\|_{W_p^{1+s}(\Omega)})$. This estimate will be crucial in showing the existence and uniqueness of solutions in the next section.

Remark 24. Notice that the estimate (119) holds without any assumptions on relation between μ and K. However, this relation will come into play in the following section.

3. Proof of Theorem 5

In order to prove our main result we apply an iterative scheme. The idea is to combine the estimate (119) with the Cauchy condition in some weaker space. In [Piasecki 2010; Piasecki and Pokorný 2014] this space was $H^1 \times L_2$. However, we need some information on the gradient of the density which is not available in our framework. Here we overcome this obstacle showing the convergence in $L_2 \times H^{-1}$. However, then we have to estimate $\|\pi(\rho^n) - \pi(\rho^{n-1})\|_{H^{-1}}$ in terms of $\|\rho^n - \rho^{n-1}\|_{H^{-1}}$ and for this purpose we have to assume (25). A similar approach was applied in the context of uniqueness of weak solutions in the nonstationary case in [Hoff 2006], where the constraint (25) also appeared. Under this assumption, setting

$$v^{n+1} = u^{n+1} + u_0 + \bar{v}$$

we can define the sequence of approximations in the following way:

$$\operatorname{div}(w^{n+1}v^{n+1} \otimes v^{n+1}) - \mu\Delta u^{n+1} - (\mu+\nu)\nabla \operatorname{div} u^{n+1} + K\nabla w^{n+1} = 0 \qquad \text{in } \Omega, \tag{120a}$$

$$w_{x_1}^{n+1} + \operatorname{div}((u^n+u_0)w^{n+1}) = -\operatorname{div}(u^n+u_0) \qquad \text{in } \Omega, \tag{120b}$$

$$n \cdot 2\mu \boldsymbol{D}(u^{n+1}) \cdot \tau + f u^{n+1} \cdot \tau = B \qquad \text{on } \Gamma, \tag{120c}$$

$$n \cdot u^{n+1} = 0 \qquad \text{on } \Gamma, \tag{120d}$$

$$w^{n+1} = w_{\text{in}} \qquad \text{on } \Gamma_{\text{in}}. \tag{120e}$$

We start with the following auxiliary result:

Lemma 25. *Assume that ϕ solves*

$$(\partial_{x_1} + U \cdot \nabla)\phi = f, \quad \phi|_{\Gamma_{\text{out}}} = 0, \tag{121}$$

with $f \in H_0^1(\Omega)$ *and sufficiently small* $U \in W_\infty^1(\Omega)$ *such that*

$$U \cdot n|_{\Gamma_{\text{in}}} < 0, \quad U \cdot n|_{\Gamma_{\text{out}}} > 0. \tag{122}$$

Then

$$\|\nabla \phi\|_{L_2(\Omega)} \le C_\Omega \|f\|_{H_0^1(\Omega)}, \tag{123}$$

where $C_\Omega \sim \max\{|\Omega|, |\Omega|^2\}$.

Proof. The boundary condition (120b) implies

$$0 = \partial_\tau \phi|_{\Gamma_{\text{out}}} = \tau_1 \phi_{x_1} + \tau_2 \phi_{x_2}|_{\Gamma_{\text{out}}}. \tag{124}$$

On the other hand we have

$$(1 + u^1)\phi_{x_1} + u^2 \phi_{x_2}|_{\Gamma_{\text{out}}} = 0. \tag{125}$$

Subtracting (124) multiplied by u^2 from (125) multiplied by τ^2 we get

$$0 = [(1 + u^1)\tau^2 - u^2\tau^1]\phi_{x_1} = -[(1 + u^1)n^1 + u^2 n^2]\phi_{x_1} = 0;$$

therefore $\phi_{x_1}|_{\Gamma_{\text{out}}} = 0$ due to (122). Combining this identity with (124) we conclude

$$\nabla \phi|_{\Gamma_{\text{out}}} = 0. \tag{126}$$

Next we differentiate (120a) with respect to x_2 and multiply by ϕ_{x_2}. Setting $V = [1 + U^1, U^2]$ we obtain

$$\tfrac{1}{2} V \cdot \nabla \phi_{x_2}^2 = f_{x_2}\phi_{x_2} - V_{x_2}\phi_{x_2} \cdot \nabla \phi. \tag{127}$$

Now let us define

$$\Omega_a = \Omega \cap \{x_1 > a\}, \quad I_a = \Omega \cap \{x_1 = a\}.$$

Integrating (127) over Ω_a we get

$$\frac{1}{2} \int_{\partial \Omega_{x_1}} (V \cdot n)\phi_{x_2}^2 \, d\sigma = \int_{\Omega_{x_1}} \Big[\frac{1}{2} \phi_{x_2}^2 \operatorname{div} V + f_{x_2}\phi_{x_2} + V_{x_2}\phi_{x_2} \cdot \nabla \phi \Big] dx. \tag{128}$$

Notice that by (126) and the condition $V \cdot n|_{\Gamma_{\text{in}}} < 0$ we have

$$\int_{\partial \Omega \cap \partial \Omega_{x_1}} (V \cdot n)\phi_{x_2}^2 \, d\sigma \le 0.$$

Therefore the smallness of U^1 implies

$$\int_{\partial \Omega_{x_1}} (V \cdot n)\phi_{x_2}^2 \, d\sigma = -\int_{I_{x_1}} (1 + U^1)\phi_{x_2}^2 \, dx_2 + \int_{\partial \Omega \cap \partial \Omega_{x_1}} (V \cdot n)\phi_{x_2}^2 \, d\sigma \le -C_U \int_{I_{x_1}} \phi_{x_2}^2 \, d\sigma$$

for some $C_U > 0$. The latter inequality combined with (128) gives for any $\epsilon > 0$

$$\sup_{x_1} \int_{I_{x_1}} \phi_{x_2}^2 \, dx_2 \le \|\nabla U\|_{L_\infty(\Omega)} \|\nabla \phi\|_{L_\infty(\Omega)} + \epsilon \|\phi_{x_2}\|_{L_2(\Omega)}^2 + \frac{C}{\epsilon} \|\nabla f\|_{L_2(\Omega)}^2.$$

Using again the smallness of U we obtain

$$\|\phi_{x_2}\|^2_{L_2(\Omega)} \le |\Omega|[(\epsilon + \|\nabla U\|_{L_\infty(\Omega)})\|\nabla\phi\|^2_{L_2(\Omega)} + C(\epsilon)\|\nabla f\|^2_{L_2(\Omega)}]. \tag{129}$$

Finally from (120a) we have

$$\|\phi_{x_1}\|^2_{L_2(\Omega)} \le \left\| \frac{f}{1+U^1} \right\|^2_{L_2(\Omega)} + \left\| \frac{U^2}{1+U^1} \right\|_{L_\infty(\Omega)} \|\phi_{x_2}\|^2_{L_2(\Omega)}.$$

Combining this inequality with (129) we get

$$\|\nabla\phi\|^2_{L_2(\Omega)} \le [|\Omega|(\epsilon + \nabla U\|_{L_\infty(\Omega)}) + C\|U\|_{L_\infty(\Omega)}]\|\nabla\phi\|^2_{L_2(\Omega)} + \left[\frac{C|\Omega|}{\epsilon} + C_p\right]\|\nabla f\|^2_{L_2(\Omega)},$$

where C_p is the constant from the Poincaré inequality. Now, provided $\|U\|_{W^1_\infty}$ is sufficiently small, to close the estimate we need $\epsilon \sim 1/|\Omega|$. Taking into account that $C_p \sim |\Omega|$ we conclude (123). □

The following proposition implies convergence of the sequence (u^n, w^n) in $L_2 \times H^{-1}$.

Proposition 26. *Assume that μ is sufficiently large compared to $|\Omega|$, $\|\bar v\|_{L_\infty}$ and K. Then the sequence (u^n, w^n) defined by* (120) *satisfies*

$$\|u^{n+1} - u^n\|_{L_2(\Omega)} + \|w^{n+1} - w^n\|_{H^{-1}(\Omega)} \le M[\|u^n - u^{n-1}\|_{L_2(\Omega)} + \|w^n - w^{n-1}\|_{H^{-1}(\Omega)}], \tag{130}$$

with $M < 1$.

Remark 27. In order to track the dependence of μ on $\|\bar v\|_{L_\infty}$ we consider again (11) with a general constant v^*.

Proof. Subtracting (120b) for two consecutive steps we get

$$(w^{n+1} - w^n)_{x_1} + \operatorname{div}[(u^n + u_0)(w^{n+1} - w^n)] = -\operatorname{div}((w^n + 1)(u^n - u^{n-1})).$$

We test this equation with ϕ given by (121) with $U = u^n + u_0$ and f such that

$$\Delta f = w^{n+1} - w^n.$$

We obtain

$$\begin{aligned}\int_\Omega (u^n - u^{n-1})(w_n + 1)\cdot\nabla\phi\,dx &= -\int_\Omega (w^{n+1} - w^n)(\partial_{x_1} + (u^n + u_0)\cdot\nabla)\phi\,dx \\ &= -\int_\Omega f\,\Delta f\,dx = \|w^{n+1} - w^n\|^2_{H^{-1}}.\end{aligned}$$

Therefore by (123) we obtain

$$\|w^{n+1} - w^n\|^2_{H^{-1}} \le \|w^n + 1\|_{L_\infty}\|u^n - u^{n-1}\|_{L_2}\|\nabla\phi\|_{L_2} \le C_1 C_\Omega \|u^n - u^{n-1}\|_{L_2}\|\nabla f\|_{L_2},$$

where C_Ω is the constant from (123). Now, since $\|\nabla f\|_{L_2} \le C\|w^{n+1} - w^n\|_{H^{-1}}$, we conclude

$$\|w^{n+1} - w^n\|_{H^{-1}(\Omega)} \le C_2 C_\Omega \|u^n - u^{n-1}\|_{L_2(\Omega)}. \tag{131}$$

Now we have to estimate $\|u^n - u^{n-1}\|_{L_2(\Omega)}$ in terms of $\|w^n - w^{n-1}\|_{H^{-1}(\Omega)}$. Subtracting (120a) for two consecutive steps we obtain

$$-\mu\Delta(u^{n+1}-u^n)-(\mu+\nu)\nabla\operatorname{div}(u^{n+1}-u^n) = -K\nabla(w^{n+1}-w^n)-\operatorname{div}[(w^{n+1}-w^n)(v^{n+1}\otimes v^{n+1})] \\ -\operatorname{div}[w^n(v^{n+1}\otimes(v^{n+1}-v^n)+(v^{n+1}-v^n)\otimes v^n)]. \quad (132)$$

Notice that as $u^{n+1} - u^n$ satisfy homogeneous boundary conditions we have in the weak sense

$$\int_\Omega \psi\cdot[\mu\Delta u + (\mu+\nu)\nabla\operatorname{div} u]\,dx = \int_\Omega u\cdot[\mu\Delta\psi + (\mu+\nu)\nabla\operatorname{div}\psi]\,dx$$

for any $\psi \in H^2$ such that $\psi|_\Gamma = 0$. Therefore testing (132) with ψ being a solution to the problem

$$-\mu\Delta\psi - (\mu+\nu)\nabla\operatorname{div}\psi = \mu(u^{n+1}-u^n), \quad \psi|_\Gamma = 0,$$

we get

$$\begin{aligned}\mu\|u^{n+1}-u^n\|^2_{L_2(\Omega)} &= \int_\Omega (w^{n+1}-w^n)(K\operatorname{div}\psi + v^{n+1}\otimes v^{n+1} : \nabla\psi)\,dx \\ &\quad + \int_\Omega w^n[v^{n+1}\otimes(v^{n+1}-v^n)+(v^{n+1}-v^n)\otimes v^n] : \nabla\psi\,dx \\ &\le C_v[\|w^{n+1}-w^n\|_{H^{-1}(\Omega)}\|\nabla\psi\|_{L_2(\Omega)} + \|u^{n+1}-u^n\|_{L_2(\Omega)}\|\nabla\psi\|_{L_2(\Omega)}] \\ &\le C_{v,K}[\|w^{n+1}-w^n\|_{H^{-1}(\Omega)}\|u^{n+1}-u^n\|_{L_2(\Omega)} + \|u^{n+1}-u^n\|^2_{L_2(\Omega)}],\end{aligned}$$

where $C_{v,K} = C_{v,K}(\|v^{n+1}\|_{L_\infty(\Omega)}, K)$. In the above display we have used the facts that $v^{n+1} - v^n = u^{n+1} - u^n$ and $\|\nabla\psi\|_{L_2(\Omega)} \le C\|u^{n+1} - u^n\|_{L_2(\Omega)}$. Therefore, assuming $\mu > C_v$ we obtain

$$\|u^{n+1}-u^n\|_{L_2(\Omega)} \le \frac{C_{v,K}}{\mu - C_{v,K}}\|w^{n+1}-w^n\|_{H^{-1}(\Omega)}.$$

Combining this estimate with (131) we conclude (130) with $M = C_2C_\Omega C_{v,K}/(\mu - C_{v,K})$. Therefore $M < 1$ provided the viscosity is sufficiently large compared to $|\Omega|$, $\|\bar v\|_{L_\infty(\Omega)}$, and K. □

The solvability of the linear system established in Section 3 implies that the sequence (120) is well-defined. Moreover, setting

$$A_n = \|u^n\|_{W^{1+s}_p(\Omega)} + \|w^n\|_{W^s_p(\Omega)},$$

by (119) we have

$$A_{n+1} \le CA_n^2 + E$$

and we can assume $E \le 1/(2C)$. Then the sequence starting with $(u^0, w^0) = ([0, 0], 0)$ satisfies

$$A_n \le 2E, \quad (133)$$

where E is the constant from (119). Proposition 26 implies

$$(u^n, w^n) \to (u, w) \quad \text{in } L_2(\Omega)\times H^{-1}(\Omega).$$

On the other hand, (133) yields up to a subsequence $(u^n, w^n) \rightharpoonup (\bar u, \bar w)$ in $W^{1+s}_p(\Omega)\times W^s_p(\Omega)$. By the definition of (u^n, w^n), the limit (u, w) is a solution to (30)–(31) in the sense of Definition 7. The

uniqueness is shown in the same way as Proposition 26. Namely, taking two solutions (u^1, w^1) and (u^2, w^2) for the same data we show

$$\|u^1 - u^2\|_{L^2(\Omega)} + \|w^1 - w^2\|_{H^{-1}(\Omega)} = 0. \qquad \square$$

4. Concluding remarks

The solutions considered here are located somehow between weak and "traditional" regular solutions satisfying the equations almost everywhere. The result for the steady transport equation given by Proposition 9 is obtained for a general class of boundary singularities showing that the choice of the fractional Sobolev–Slobodetskii spaces is in a sense natural for the problem under consideration. The price we pay is that we have to assume linearity of the pressure. Getting rid of this constraint seems an interesting open problem. It is likely that the existence itself could be shown for more general pressure laws using for example approximation with more regular solutions which give some information about the gradient of the density. However, such a result would be highly technical and not really meaningful without uniqueness, which is more challenging and seems to require some novel approach to treat more general pressure laws. Finally we shall mention that the assumed C^3 regularity of the domain required to solve an elliptic problem appearing in the estimate for the velocity is not optimal and could be relaxed at the price of additional technicalities. However, as we are rather interested in a careful investigation of the boundary singularity in the stationary transport equation which is independent of global regularity of the boundary (for example a C^∞ boundary can fail to satisfy (16)), we keep this regularity assumption.

Acknowledgements

This work has been supported by Ideas Plus grant ID 2011 0006 61. The authors would like to thank the anonymous referee for a careful reading of the manuscript and numerous remarks which contributed to improving the clarity of the proofs.

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

[Bresch and Jabin 2018] D. Bresch and P.-E. Jabin, "Global existence of weak solutions for compressible Navier–Stokes equations: thermodynamically unstable pressure and anisotropic viscous stress tensor", *Ann. of Math.* (2) **188**:2 (2018), 577–684. MR Zbl

[Crippa and De Lellis 2008] G. Crippa and C. De Lellis, "Estimates and regularity results for the DiPerna–Lions flow", *J. Reine Angew. Math.* **616** (2008), 15–46. MR Zbl

[Dalibard and Saint-Raymond 2018] A.-L. Dalibard and L. Saint-Raymond, *Mathematical study of degenerate boundary layers: a large scale ocean circulation problem*, Mem. Amer. Math. Soc. **1206**, Amer. Math. Soc., Providence, RI, 2018. MR Zbl

[DeVore and Sharpley 1993] R. A. DeVore and R. C. Sharpley, "Besov spaces on domains in $\mathbb{R}^d$", *Trans. Amer. Math. Soc.* **335**:2 (1993), 843–864. MR Zbl

[DiPerna and Lions 1989] R. J. DiPerna and P.-L. Lions, "Ordinary differential equations, transport theory and Sobolev spaces", *Invent. Math.* **98**:3 (1989), 511–547. MR Zbl

[Feireisl and Novotný 2018] E. Feireisl and A. Novotný, "Stationary solutions to the compressible Navier–Stokes system with general boundary conditions", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **35**:6 (2018), 1457–1475. MR Zbl

[Feireisl et al. 2001] E. Feireisl, A. Novotný, and H. Petzeltová, "On the existence of globally defined weak solutions to the Navier–Stokes equations", *J. Math. Fluid Mech.* **3**:4 (2001), 358–392. MR Zbl

[Girinon 2011] V. Girinon, "Navier–Stokes equations with nonhomogeneous boundary conditions in a bounded three-dimensional domain", *J. Math. Fluid Mech.* **13**:3 (2011), 309–339. MR Zbl

[Guo et al. 2015] Y. Guo, S. Jiang, and C. Zhou, "Steady viscous compressible channel flows", *SIAM J. Math. Anal.* **47**:5 (2015), 3648–3670. MR Zbl

[Hoff 2006] D. Hoff, "Uniqueness of weak solutions of the Navier–Stokes equations of multidimensional, compressible flow", *SIAM J. Math. Anal.* **37**:6 (2006), 1742–1760. MR Zbl

[Kreml et al. 2013] O. Kreml, Š. Nečasová, and M. Pokorný, "On the steady equations for compressible radiative gas", *Z. Angew. Math. Phys.* **64**:3 (2013), 539–571. MR Zbl

[Kweon and Kellogg 1997] J. R. Kweon and R. B. Kellogg, "Compressible Navier–Stokes equations in a bounded domain with inflow boundary condition", *SIAM J. Math. Anal.* **28**:1 (1997), 94–108. MR Zbl

[Kweon and Kellogg 1998] J. R. Kweon and R. B. Kellogg, "Smooth solution of the compressible Navier–Stokes equations in an unbounded domain with inflow boundary condition", *J. Math. Anal. Appl.* **220**:2 (1998), 657–675. MR Zbl

[Lions 1998] P.-L. Lions, *Mathematical topics in fluid mechanics, II: Compressible models*, Oxford Lecture Series in Mathematics and its Applications **10**, The Clarendon Press, Oxford University Press, New York, 1998. MR Zbl

[Mucha 2010] P. B. Mucha, "Transport equation: extension of classical results for div $b \in$ BMO", *J. Differential Equations* **249**:8 (2010), 1871–1883. MR Zbl

[Mucha and Piasecki 2014] P. B. Mucha and T. Piasecki, "Compressible perturbation of Poiseuille type flow", *J. Math. Pures Appl.* (9) **102**:2 (2014), 338–363. MR Zbl

[Mucha and Pokorný 2006] P. B. Mucha and M. Pokorný, "On a new approach to the issue of existence and regularity for the steady compressible Navier–Stokes equations", *Nonlinearity* **19**:8 (2006), 1747–1768. MR Zbl

[Mucha and Pokorný 2009] P. B. Mucha and M. Pokorný, "On the steady compressible Navier–Stokes–Fourier system", *Comm. Math. Phys.* **288**:1 (2009), 349–377. MR Zbl

[Mucha and Rautmann 2006] P. B. Mucha and R. Rautmann, "Convergence of Rothe's scheme for the Navier–Stokes equations with slip conditions in 2D domains", *ZAMM Z. Angew. Math. Mech.* **86**:9 (2006), 691–701. MR Zbl

[Novo and Novotný 2002] S. Novo and A. Novotný, "On the existence of weak solutions to the steady compressible Navier–Stokes equations when the density is not square integrable", *J. Math. Kyoto Univ.* **42**:3 (2002), 531–550. MR Zbl

[Novotný and Pileckas 1998] A. Novotny and K. Pileckas, "Steady compressible Navier–Stokes equations with large potential forces via a method of decomposition", *Math. Methods Appl. Sci.* **21**:8 (1998), 665–684. MR Zbl

[Novotný and Padula 1994] A. Novotný and M. Padula, "L^p-approach to steady flows of viscous compressible fluids in exterior domains", *Arch. Rational Mech. Anal.* **126**:3 (1994), 243–297. MR Zbl

[Novotný and Straškraba 2004] A. Novotný and I. Straškraba, *Introduction to the mathematical theory of compressible flow*, Oxford Lecture Series in Mathematics and its Applications **27**, Oxford University Press, 2004. MR Zbl

[Piasecki 2009] T. Piasecki, "Steady compressible Navier–Stokes flow in a square", *J. Math. Anal. Appl.* **357**:2 (2009), 447–467. MR Zbl

[Piasecki 2010] T. Piasecki, "On an inhomogeneous slip-inflow boundary value problem for a steady flow of a viscous compressible fluid in a cylindrical domain", *J. Differential Equations* **248**:8 (2010), 2171–2198. MR Zbl

[Piasecki 2018] T. Piasecki, "Steady transport equation in Sobolev–Slobodetskii spaces", *Colloq. Math.* **154**:1 (2018), 65–76. MR Zbl

[Piasecki and Pokorný 2014] T. Piasecki and M. Pokorný, "Strong solutions to the Navier–Stokes–Fourier system with slip-inflow boundary conditions", *ZAMM Z. Angew. Math. Mech.* **94**:12 (2014), 1035–1057. MR Zbl

[Plotnikov and Sokołowski 2012] P. Plotnikov and J. Sokołowski, *Compressible Navier–Stokes equations: theory and shape optimization*, Instytut Matematyczny Polskiej Akademii Nauk. Monografie Matematyczne (New Series) **73**, Springer, 2012. MR Zbl

[Plotnikov et al. 2008] P. I. Plotnikov, E. V. Ruban, and J. Sokolowski, "Inhomogeneous boundary value problems for compressible Navier–Stokes equations: well-posedness and sensitivity analysis", *SIAM J. Math. Anal.* **40**:3 (2008), 1152–1200. MR Zbl

[Plotnikov et al. 2009] P. I. Plotnikov, E. V. Ruban, and J. Sokolowski, "Inhomogeneous boundary value problems for compressible Navier–Stokes and transport equations", *J. Math. Pures Appl.* (9) **92**:2 (2009), 113–162. MR Zbl

[Pokorný and Mucha 2008] M. Pokorný and P. B. Mucha, "3D steady compressible Navier–Stokes equations", *Discrete Contin. Dyn. Syst. Ser. S* **1**:1 (2008), 151–163. MR Zbl

[Triebel 1978] H. Triebel, *Interpolation theory, function spaces, differential operators*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1978. MR Zbl

[Valli 1983] A. Valli, "Periodic and stationary solutions for compressible Navier–Stokes equations via a stability method", *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **10**:4 (1983), 607–647. MR Zbl

[Valli and Zajączkowski 1986] A. Valli and W. M. Zajączkowski, "Navier–Stokes equations for compressible fluids: global existence and qualitative properties of the solutions in the general case", *Comm. Math. Phys.* **103**:2 (1986), 259–296. MR Zbl

[Beirão da Veiga 1987] H. Beirão da Veiga, "An L^p-theory for the n-dimensional, stationary, compressible Navier–Stokes equations, and the incompressible limit for compressible fluids: the equilibrium solutions", *Comm. Math. Phys.* **109**:2 (1987), 229–248. MR Zbl

Received 25 Jun 2019. Accepted 27 Sep 2019.

PIOTR B. MUCHA: p.mucha@mimuw.edu.pl
Institute of Applied Mathematics and Mechanics, University of Warsaw, Warsaw, Poland

TOMASZ PIASECKI: tpiasecki@mimuw.edu.pl
Institute of Applied Mathematics and Mechanics, University of Warsaw, Warsaw, Poland

PURE and APPLIED ANALYSIS

Vol. 2, No. 1, 2020
dx.doi.org/10.2140/paa.2020.2.157

msp

OPTIMAL CONSTANTS IN NONTRAPPING RESOLVENT ESTIMATES AND APPLICATIONS IN NUMERICAL ANALYSIS

JEFFREY GALKOWSKI, EUAN A. SPENCE AND JARED WUNSCH

We study the resolvent for nontrapping obstacles on manifolds with Euclidean ends. It is well known that for such manifolds the outgoing resolvent satisfies $\|\chi R(k)\chi\|_{L^2\to L^2} \le Ck^{-1}$ for $k > 1$, but the constant C has been little studied. We show that, for high frequencies, the constant is bounded above by $\frac{2}{\pi}$ times the length of the longest generalized bicharacteristic of $|\xi|_g^2 - 1$ remaining in the support of χ. We show that this estimate is optimal in the case of manifolds without boundary. We then explore the implications of this result for the numerical analysis of the Helmholtz equation.

1. Introduction

Let (M, g) be a manifold with Euclidean ends and $\Omega \Subset M$ an obstacle with smooth boundary. Assume that all Melrose–Sjöstrand generalized bicharacteristics (i.e., geodesics) escape to infinity. Let $\Delta_{\Omega,g}$ be the Dirichlet realization of the Laplacian on $M \setminus \Omega$. It is well known that for any $\chi \in \mathcal{C}_c^\infty(M)$, there exists a constant $C > 0$ and $k_0 > 0$ so that

$$\|\chi(-\Delta_{\Omega,g} - k^2 - i0)^{-1}\chi\|_{L^2(M\setminus\Omega)\to L^2(M\setminus\Omega)} \le C|k|^{-1}, \quad k > k_0.$$

In this paper we study how the constant $C > 0$ depends on the classical dynamics on $((M, g), \Omega)$.

Suppose that there are no geodesics tangent to $\partial\Omega$ to infinite order and let $\varphi_t : S^*M \to S^*M$ denote the Melrose–Sjöstrand generalized bicharacteristic flow [Hörmander 1994, Section 24.3]. Next, fix $R_1 > 0$ so that $\Omega \subset B(0, R_1)$ and (M, g) is Euclidean outside $B(0, R_1)$. Then define, for any $R \ge R_1$,

$$L(g, \Omega, R) := \inf\{t > 0 : \varphi_t(S^*_{B(0,R)}M) \cap S^*_{B(0,R)}M = \varnothing\}. \tag{1-1}$$

(We will omit the Ω from the notation when $\Omega = \varnothing$.)

In the statement of the following estimates, we will use a family of Sobolev spaces with appropriate semiclassical scaling, using the global definition, for $s \in \mathbb{R}$,

$$\|u\|^2_{H^s(M\setminus\Omega)} = \langle(-\Delta_{\Omega,g} + k^2)^s u, u\rangle. \tag{1-2}$$

Theorem 1. *Let (M, g) be a manifold with Euclidean ends with $g \in \mathcal{C}^{1,1}$. Suppose that $\Omega \Subset M$ has smooth boundary, g is $\mathcal{C}^\infty$ near Ω, and Ω is nowhere tangent to the geodesic flow to infinite order. Assume the*

MSC2010: 35J05, 35P25, 65N30.
Keywords: resolvent, Helmholtz equation, nontrapping, variable wave speed, finite element method.

generalized geodesic flow on $M\setminus\Omega$ is nontrapping. Then for every $R > R_1$, $\chi \in \mathcal{C}_c^\infty(B(0,R);[0,1])$ there exists $k_0 > 0$ so that, for $k > k_0$,

$$\|\chi(-\Delta_{\Omega,g} - k^2 - i0)^{-1}\chi\|_{L^2(M\setminus\Omega)\to L^2(M\setminus\Omega)} \le \frac{2L(g,\Omega,R)}{\pi k}. \tag{1-3}$$

More generally, for $0 \le s \le 2$,

$$\|\chi(-\Delta_{\Omega,g} - k^2 - i0)^{-1}\chi\|_{L^2(M\setminus\Omega)\to H^s(M\setminus\Omega)} \le \frac{2^{s/2+1}L(g,\Omega,R)}{\pi}k^{s-1}. \tag{1-4}$$

The constant k_0 may be chosen uniformly as g varies within a sufficiently small open neighborhood in $\mathcal{C}^{2,\alpha}$ ($\alpha > 0$) of a given nontrapping metric in $\mathcal{C}^{2,\alpha}$ (where all metrics are taken smooth near $\partial\Omega$).

Conversely, when $\Omega = \varnothing$ and $g \in \mathcal{C}^\infty$, for every $R' > R_1$ with $B(0,R') \subset \{\chi \equiv 1\}$, there is $k_0 > 0$ so that, for $k > k_0$,

$$\|\chi(-\Delta_g - k^2 - i0)^{-1}\chi\|_{L^2(M\setminus\Omega)\to L^2(M\setminus\Omega)} \ge \frac{2L(g,R')}{\pi k}. \tag{1-5}$$

Notice that in the interior of $M \setminus \Omega$, the Melrose–Sjöstrand flow is equal to the geodesic flow except that, on the unit cotangent bundle, the speed of the flow is 2 rather than 1. Thus, in the case where $\Omega = \varnothing$, the theorem states that the growth of the resolvent as a map on L^2 as $k \to \infty$ is controlled above and below by $\frac{1}{\pi}$ times the length of the longest geodesic contained entirely in $B(0,R_1)$.

1A. ***More general operators.*** In Section 6B below, we recall that several important physical applications of the Helmholtz equation involve an operator that differs slightly from $-\Delta_{\Omega,g}$, namely the divergence form operator $-\sum \partial_i g^{ij}\,\partial_j$, which is self-adjoint with respect to the Euclidean volume form on $\mathbb{R}^n$. In order to deal with this and similar operators, we will prove a slightly more general result.

Theorem 2. *Let (M,g) be a manifold with Euclidean ends, $g \in \mathcal{C}^{1,1}$. Suppose that $\Omega \subset M$ has smooth boundary, g is $\mathcal{C}^\infty$ in a neighborhood of Ω, and Ω is nowhere tangent to the geodesic flow to infinite order. Assume the generalized geodesic flow on $M\setminus\Omega$ is nontrapping. Suppose that $P(g) \in \mathrm{Diff}^2(M)$ so that*

$$P(g) + \Delta_g = \sum_j L_j\partial_{x_j} + L, \tag{1-6}$$

where $L_j, L \in \mathcal{C}_c^{0,\alpha}(B(0,R_1))$ for some $\alpha > 0$. Suppose further that ν is a density on M so that $P_\Omega(g)$ is self-adjoint with respect to $L^2(M\setminus\Omega;\nu)$, where $P_\Omega(g)$ is the Dirichlet realization of $P(g)$. Let

$$\|u\|^2_{H^s_\nu(M\setminus\Omega)} := \langle (P_\Omega(g)+k^2)^s u, u\rangle_{L^2(M\setminus\Omega;\nu)}. \tag{1-7}$$

Then for every $R > R_1$, $\chi \in \mathcal{C}_c^\infty(B(0,R);[0,1])$, there exists $k_0 > 0$ so that, for $k > k_0$ and $0 \le s \le 2$,

$$\|\chi(P_\Omega(g) - k^2 - i0)^{-1}\chi\|_{L^2_\nu(M\setminus\Omega)\to H^s_\nu(M\setminus\Omega)} \le \frac{2^{s/2+1}L(g,\Omega,R)}{\pi}k^{s-1}. \tag{1-8}$$

The constant k_0 may be chosen uniformly as g varies within sufficiently small open neighborhoods of in $\mathcal{C}^{2,\alpha}$ ($\alpha > 0$) of a given nontrapping metric in $\mathcal{C}^{2,\alpha}$ and L_j, L vary in small neighborhoods in $\mathcal{C}^{0,\alpha}$ (where all g, L_j, L are taken smooth near $\partial\Omega$ and the subset is assumed to be contained in a small open neighborhood in $\mathcal{C}^N$ for some sufficiently large N near $\partial\Omega$).

1B. ***Motivation from, and applications to, numerical analysis.*** In the last few years, there has been growing interest from the numerical-analysis community in proving bounds on solutions of the Helmholtz equation where the constants are explicit in the metric (i.e., the coefficients); see [Brown et al. 2017; Chaumont-Frelet 2016; Barucq et al. 2017; Ohlberger and Verfürth 2018; Sauter and Torres 2018; Moiola and Spence 2019; Graham et al. 2019; Graham and Sauter 2020]. Almost all of these previously obtained bounds used variants of the Morawetz commutator $x \cdot \nabla$, and thus are restricted to star-shaped domains and certain classes of coefficients (although this class includes discontinuous coefficients; see, e.g., [Graham et al. 2019]); the exception are the one-dimensional bounds in [Sauter and Torres 2018], which use the fact that the solution of the Helmholtz equation with piecewise-constant coefficients in one dimension can be expressed in terms of the solution of a linear system of algebraic equations.

This interest from the numerical-analysis community is because the analysis of *any* numerical method for solving the Helmholtz equation with variable coefficients requires a resolvent estimate, and if the constant in the resolvent estimate is not given explicitly in terms of the coefficients, then the numerical analysis will not be explicit in the coefficients. For example, having proved a bound explicit in the coefficients, [Chaumont-Frelet 2016; Graham and Sauter 2020] were then concerned with analyzing standard finite-element methods applied to Helmholtz problems with variable coefficients, and [Brown et al. 2017; Chaumont-Frelet 2016; Barucq et al. 2017; Ohlberger and Verfürth 2018] were concerned with designing and analyzing methods tailored to the coefficients.

The resolvent estimate (1-8) will therefore be a fundamental ingredient in the numerical analysis of variable-coefficient Helmholtz problems in nontrapping scenarios. In Section 6 we illustrate this fact by proving an error estimate for the finite-element method applied to the variable-coefficient Helmholtz equation posed in the exterior of a nontrapping Dirichlet obstacle, with this estimate explicit in k, the coefficients, and the parameters of the discretization; see Theorem 3 below. The key point is that Theorem 3 shows how the condition on the discretization for the error estimate to hold depends on the length of the longest ray, showing that this condition becomes more restrictive as the length of the longest ray grows.

In Section 6 we also briefly outline the implications of the estimate (1-8) for (i) preconditioning finite-element discretizations of the Helmholtz equation (see Remark 6.8), and (ii) "uncertainty quantification" of the Helmholtz equation (see Remark 6.9).

2. Manifolds with Euclidean ends

We now define the notion of a manifold with Euclidean ends. Note that the canonical example of a manifold with Euclidean ends is the space $\mathbb{R}^n$ with a metric g so that $I - g$ has compact support. In order to allow more general topologies, we define the general notion of a manifold with Euclidean ends.

Definition 2.1. (M, g) is an n-dimensional *manifold with Euclidean ends* if it is a noncompact, complete Riemannian manifold such that

- there exists a function $r \in \mathcal{C}^\infty(M; \mathbb{R})$ such that the sets $\{r \leq c\}$ are compact for all c, and

- there exists $R_1 > 0$ such that $\{r \geq R_1\}$ is the disjoint union of finitely many components, each of which is isometric to $\mathbb{R}^n \setminus B(0, R_1)$ with the Euclidean metric, and the pullback of r under the isometry is the Euclidean norm.

The connected components of $\{r \geq R_1\}$ are called the *infinite ends* of M. The notation $B(0, R)$ is defined for $R \geq R_1$ and has the following meaning:

$$B(0, R) := \{r < R\}.$$

2A. ***The outgoing resolvent on manifolds with Euclidean ends.*** We now review (following the treatment in [Dyatlov and Zworski 2019, Section 4.2]) some properties of the outgoing resolvent on a manifold with Euclidean ends. Let $E_i,\ i = 1, \ldots, m$, be the infinite ends of M, and let $R_0(k)$ denote the free resolvent on $\mathbb{R}^n$. That is, $R_0(k) : L^2_{\text{comp}}(\mathbb{R}^n) \to L^2_{\text{loc}}(\mathbb{R}^n)$ is the meromorphic continuation of $(-\Delta - k^2)^{-1}$ from the half plane $\operatorname{Im} k > 0$. Define $\widetilde{R}_0(k) : L^2_{\text{comp}}(M) \to L^2_{\text{loc}}(M)$ by

$$\widetilde{R}_0(k) f := \sum_{i=1}^{m} 1_{E_i} R_0(k) 1_{E_i} f. \tag{2-1}$$

Next, let $\chi_i \in \mathcal{C}^\infty_c(M; [0, 1]),\ i = 0, \ldots, 3$, so that

$$\chi_i \equiv 1 \text{ on } \operatorname{supp} \chi_{i-1}, \quad \operatorname{supp}(1 - \chi_i) \subset M \setminus \overline{B(0, R_1)}, \quad \operatorname{supp} \chi_i \subset B(0, R) \tag{2-2}$$

for some $R > R_1$. Then for $k_0 \in \mathbb{C}$ with $\operatorname{Im} k_0 \gg 1$, let

$$Q(k, k_0) := Q_0(k) + Q_1(k_0),$$
$$Q_0(k) := (1 - \chi_0)\widetilde{R}_0(k)(1 - \chi_1), \quad Q_1(k) := \chi_2(-\Delta_g - k_0^2)^{-1}\chi_1.$$

Following the proof in [Dyatlov and Zworski 2019, Section 4.2], we can write

$$R(k) := (P_\Omega(g) - k^2)^{-1} = Q(k, k_0)(I + K(k, k_0)\chi_3)^{-1}(I - K(k, k_0)(1 - \chi_3)), \tag{2-3}$$

where $K(k, k_0) : L^2(M) \to L^2_{\text{comp}}(M)$ and in particular, $(1 - \chi_3)K(k, k_0) = 0$. This implies $(1 - \chi_3)R(k)$ lies in the image of $\widetilde{R}_0(k)$.

3. The case of a manifold without boundary

To illustrate the methods we begin by proving (1-3) in the case of a manifold without boundary. The idea of the proof is identical when the boundary is nonempty, but the proofs of propagation statements are more involved. Thus we take $\partial M = \varnothing$ throughout this section.

3A. ***Defect measures.*** We will argue by contradiction. Let $h = k^{-1}$ and $P(h) := h^2 P_\Omega(g) - 1$. We also write L^2 for $L^2_\nu(M \setminus \Omega)$ and H^s for $H^s_\nu(M \setminus \Omega)$.

Remark 3.1. We will sometimes write $\|\cdot\|_{H^s_h}$ for $h^s\|\cdot\|_{H^s}$; since $\|\cdot\|_{H^s}$ has a semiclassical scaling built into it, this means that

$$\|u\|^2_{H^s_h} = \langle (h^2\Delta_{\Omega,g} + 1)^s u, u\rangle.$$

Note that if $\chi_0,\ \chi_1 \in \mathcal{C}_c^\infty(M;[0,1])$ and $\chi_1 \equiv 1$ on $\operatorname{supp}\chi_0$, then

$$\frac{\|\chi_0(P-i0)^{-1}\chi_0 f\|_{L^2}}{\|f\|_{L^2}} = \frac{\|\chi_0\chi_1(P-i0)^{-1}\chi_1\chi_0 f\|_{L^2}}{\|f\|_{L^2}} \le \frac{\|\chi_0\chi_1(P-i0)^{-1}\chi_1\chi_0 f\|_{L^2}}{\|\chi_0 f\|_{L^2}}$$
$$\le \frac{\|\chi_1(P-i0)^{-1}\chi_1\chi_0 f\|_{L^2}}{\|\chi_0 f\|_{L^2}} \le \|\chi_1(P-i0)^{-1}\chi_1\|_{L^2\to L^2}.$$

In particular,

$$\|\chi_0(P-i0)^{-1}\chi_0\|_{L^2\to L^2} \le \|\chi_1(P-i0)^{-1}\chi_1\|_{L^2\to L^2},$$

and, since $R > R_1$, *we may assume without loss of generality that* $\chi \equiv 1$ *on* $B(0,R_1)$.

If (1-3) fails, there exists a sequence of discrete values of $h = h_j \downarrow 0$ and a sequence $0 \neq f(h) \in L^2$ such that

$$\|\chi(P-i0)^{-1}\chi h f(h)\|_{L^2} = 1$$

and

$$\lim_{h\to 0} \frac{\|\chi(P-i0)^{-1}\chi h f(h)\|_{L^2}}{\|f(h)\|_{L^2}} = M \ge 2L(g,R)$$

(note that we allow $M=\infty$). There exists $R' < R$ such that we still have $\operatorname{supp}\chi \subset B(0,R')$; hence we have the strict inequality $M > L(g,R')$. Let

$$u(h) = (P-i0)^{-1}\chi h f \in L^2_{\mathrm{loc}}$$

so that $\|\chi u\|_{L^2} = 1$.

Lemma 3.2. *For all* $\tilde\chi \in \mathcal{C}_c^\infty(M)$, *there exists* $C > 0$ *so that* $\|\tilde\chi u\|_{L^2} \le C$.

Proof. Let $\chi_i \in \mathcal{C}_c^\infty(M;[0,1])$, $i = 0, 1$, so that $\chi \equiv 1$ on $\operatorname{supp}\chi_1$, and $\chi_1 \equiv 1$ on $\operatorname{supp}\chi_0$, and $\chi_0 \equiv 1$ on $B(0,R_1)$. Then

$$(1-\chi_0)P(1-\chi_1) = \sum_i (1-\chi_0)(-h^2\Delta_{\mathbb{R}^n} - 1)1_{E_i}(1-\chi_1),$$

where E_i denotes the i-th Euclidean end of M.

Then,

$$P(1-\chi_1)u = \sum_i (1-\chi_0)1_{E_i}(-h^2\Delta_{\mathbb{R}^n} - 1)1_{E_i}(1-\chi_1)u$$
$$= (1-\chi_1)\chi h f - \sum_i (1-\chi_0)[-h^2 1_{E_i}\Delta_{\mathbb{R}^n}1_{E_i}, \chi_1]u$$
$$= (1-\chi_1)\chi h f - \sum_i [-h^2 1_{E_i}\Delta_{\mathbb{R}^n}1_{E_i}, \chi_1]u.$$

Next extend $u1_{E_i}$ by 0 to a function v_i on $\mathbb{R}^n$ so that $v_i \equiv u1_{E_i}$ on $\mathbb{R}^n \setminus B(0,R_1)$. Then

$$(-h^2\Delta_{\mathbb{R}^n} - 1)(1-\chi_1)1_{E_i}v_i = 1_{E_i}(1-\chi_1)\chi h f - [h^2\Delta_{\mathbb{R}^n}, 1_{E_i}\chi_1]v_i.$$

Since by (2-3) v_i is h^{-1} outgoing,

$$(1-\chi_1)1_{E_i}v_i = h^{-2}R_0(h^{-1})(1_{E_i}(1-\chi_1)\chi h f - [h^2\Delta_{\mathbb{R}^n}, 1_{E_i}\chi_1]v_i).$$

In particular, since $\chi \equiv 1$ on $\operatorname{supp}\chi_1$,

$$\|(1-\chi_1)1_{E_i}v_i\|_{L^2} \le C(\|f\|_{L^2} + \|[\Delta_{\mathbb{R}^n}, 1_{E_i}\chi_1]v_i\|_{H^{-1}}) \le C(\|f\|_{L^2} + \|\chi u\|_{L^2}).$$

Now,

$$(1-\chi_1)u = (1-\chi_1)\sum_i 1_{E_i}v_i.$$

Therefore,

$$\|(1-\chi_1)u\|_{L^2} \le C(\|f\|_{L^2} + \|\chi u\|_{L^2}).$$

Let $\tilde\chi \in \mathcal{C}_c^\infty(M;[0,1])$. Then using again that $\chi \equiv 1$ on $\operatorname{supp}\chi_1$,

$$\begin{aligned}\limsup_{h\to 0}\|\tilde\chi u\|_{L^2} &\le \limsup_{h\to 0}(\|\tilde\chi(1-\chi)u\|_{L^2} + \|\tilde\chi\chi u\|_{L^2})\\ &\le \limsup_{h\to 0}(\|\tilde\chi(1-\chi_1)u\|_{L^2} + \|\chi u\|_{L^2})\\ &\le C\limsup_{h\to 0}(\|f\|_{L^2} + \|\chi u\|_{L^2}) \le C\left(\frac{1}{2L(g,R')}+1\right),\end{aligned}$$

completing the proof of the lemma. □

By Lemma 3.2, u is uniformly bounded in L^2_{loc} and, taking subsequences, we may assume that u has defect measure μ, χf has defect measure α, and u and χf have joint defect measure μ^j; in other words, for $h = h_j$ in this chosen subsequence, and for every $a \in \mathcal{C}_c^\infty(T^*M)$,

$$\lim_{h\downarrow 0}\langle a(x,hD)u,u\rangle = \int a\,d\mu, \quad \lim_{h\downarrow 0}\langle a(x,hD)\chi f,\chi f\rangle = \int a\,d\alpha, \quad \lim_{h\downarrow 0}\langle a(x,hD)\chi f,u\rangle = \int a\,d\mu^j. \tag{3-1}$$

Note that we use a quantization procedure that sends symbols with compact support in x to operators with compactly supported kernel.

The Cauchy–Schwarz inequality gives us a simple inequality satisfied by these three measures. We use the notation

$$\mu(a) \equiv \int a\,d\mu$$

for the pairing of a function and a measure.

Lemma 3.3. *For any $a \in \mathcal{C}_c^0(T^*M;\mathbb{R})$,*

$$|\mu^j(a)| \le \sqrt{\mu(|a|)}\sqrt{\alpha(|a|)}.$$

Proof. If $b = \sqrt{a}$ is in $\mathcal{C}_c^\infty(T^*M)$,

$$\begin{aligned}\langle a(x,hD)\chi f,u\rangle &= \langle b^2(x,hD)\chi f,u\rangle = \langle b(x,hD)\chi f, b(x,hD)u\rangle + O(h)\\ &\le \sqrt{\langle b(x,hD)\chi f, b(x,hD)\chi f\rangle}\sqrt{\langle b(x,hD)u, b(x,hD)u\rangle} + O(h)\\ &\le \left(\int b^2\,d\alpha\right)^{1/2}\left(\int b^2\,d\mu\right)^{1/2} + o(1),\end{aligned}$$

which proves the desired inequality by letting $h \to 0$.

Next, let $a \in \mathcal{C}_c^\infty(T^*M)$ with $a \geq 0$, $\psi_i \in \mathcal{C}_c^\infty(T^*M; [0,1])$, $i = 0, 1$, have $\psi_i \equiv 1$ in a neighborhood of $\operatorname{supp} a$ and $\psi_1 \equiv 1$ in a neighborhood of $\operatorname{supp} \psi_0$. Then we apply the previous result to $a_\epsilon = \psi_0^2(a + \epsilon\psi_1)$ and let $\epsilon \downarrow 0$ to see that $|\mu^j(a)| \leq \sqrt{\mu(a)}\sqrt{\alpha(a)}$.

Now, letting $0 \leq a \in \mathcal{C}_c^0(T^*M)$, we have $0 \leq a_n \in \mathcal{C}_c^\infty(T^*M)$ with $a_n \to a$ uniformly. Then we have

$$\mu^j(a_n) \to \mu^j(a), \quad \mu(a_n) \to \mu(a), \quad \alpha(a_n) \to \alpha(a).$$

In particular,

$$|\mu^j(a)| \leq \sqrt{\mu(a)}\sqrt{\alpha(a)}, \quad 0 \leq a \in \mathcal{C}_c^0(T^*M).$$

Next, for $a \in \mathcal{C}_c^0(T^*M; \mathbb{R})$, write $a = a_+ - a_-$ with $0 \leq a_\pm \in \mathcal{C}_c^0(T^*M)$. Then,

$$|\mu^j(a)| \leq |\mu^j(a_+)| + |\mu^j(a_-)| \leq \sqrt{\mu(a_+)}\sqrt{\alpha(a_+)} + \sqrt{\mu(a_-)}\sqrt{\alpha(a_-)} \leq \sqrt{\mu(|a|)}\sqrt{\alpha(|a|)}. \qquad \square$$

Since $Pu = h\chi f$, for $a \in \mathcal{C}_c^\infty(T^*M; \mathbb{R})$,

$$\begin{aligned} ih^{-1}\langle [P, a(x, hD)]u, u\rangle &= ih^{-1}(\langle a(x, hD)u, Pu\rangle - \langle a(x, hD)Pu, u\rangle \\ &= 2\operatorname{Im}\langle a(x, hD)\chi f, u\rangle. \end{aligned} \tag{3-2}$$

Sending $h \to 0$ yields

$$\mu(H_p a) = 2\operatorname{Im}\mu^j(a). \tag{3-3}$$

For the proof of the following standard result, we use the formula (2-3) relating

$$R(k) := (P_\Omega(g) - k^2 - i0)^{-1}$$

to the free outgoing resolvent on $\mathbb{R}^n$ and refer the reader to, e.g., [Burq 2002, Proposition 3.5].

Lemma 3.4. *Let*

$$\mathcal{I} := \left\{ \rho \in S^*M : \bigcup_{t \geq 0} \varphi_{-t}(\rho) \cap \operatorname{supp} \chi = \varnothing \right\}$$

be the directly incoming set. Then $\mu(\mathcal{I}) = 0$.

Next, we show that μ is supported on the characteristic variety and that u is oscillating at frequency roughly h^{-1}.

Lemma 3.5. *The measure* μ *is supported on* S^*M. *In addition, for* $b \in S^1$,

$$\lim_{h \to 0} \|b(x, hD)\chi u\|_{L^2}^2 = \mu(|b|^2\chi^2).$$

Proof. Suppose that $a \in \mathcal{C}_c^\infty(T^*M)$ with $a \equiv 0$ in a neighborhood of $|\xi|_g = 1$. Then there exists $E \in \Psi^{-2}$ compactly supported so that

$$a(x, hD) = EP + O_{L^2_{\text{loc}} \to L^2_{\text{comp}}}(h^\infty).$$

Therefore,

$$\langle a(x, hD)u, u\rangle = \langle Eh\chi f, u\rangle + O(h^\infty) \to 0.$$

Hence, $\mu(a) = 0$. In particular, $\operatorname{supp} \mu \subset S^*M$.

Fix $\psi \in \mathcal{C}_c^\infty(T^*M)$ with $\psi \equiv 1$ on $\operatorname{supp}\chi \cap \{|\xi|_g^2 \le 2\}$. Then, there exists $E \in \Psi^0$ compactly supported so that

$$\chi(x)b(x,hD)^*b(x,hD)\chi(x)(1-\psi(x,hD)) = EP + O_{L^2_{\mathrm{loc}}\to L^2_{\mathrm{comp}}}(h^\infty).$$

In particular,

$$\begin{aligned}
&\|b(x,hD)\chi u\|_{L^2}^2\\
&= \langle \chi(x)b(x,hD)^*b(x,hD)\chi(x)u, u\rangle\\
&= \langle \chi(x)b(x,hD)^*b(x,hD)\chi(x)\psi(x,hD)u, u\rangle + \langle \chi(x)b(x,hD)^*b(x,hD)\chi(x)(1-\psi(x,hD))u, u\rangle\\
&= \langle \chi(x)b(x,hD)^*b(x,hD)\chi(x)\psi(x,hD)u, u\rangle + \langle hE\chi f, u\rangle + O(h^\infty)\\
&\to \mu(|b|^2\chi^2\psi) = \mu(|b|^2\chi^2).
\end{aligned}$$

□

3B. ***Hölder continuous metrics.*** We now make the necessary adjustments to allow the metric g to be Hölder continuous. We refer the reader to [Taylor 2000, Chapter 3, Section 11] for an analogous account of propagation of singularities for the wave equation as well as a review of the history of propagation of singularities theorems for operators with rough coefficients. Stronger results (i.e., with weaker regularity hypotheses) are probably possible in line with [Burq and Zuily 2015, Remark 3.3], but low regularity is not our main focus here.

Lemma 3.6. *Let g_0, L_0, $L_{j,0}$ satisfy the hypotheses of Theorem 2. Suppose that $g(h) \in \mathcal{C}^{1,\alpha}$ and $L(h), L_j(h) \in \mathcal{C}^{0,\alpha}$ for some $\alpha > 0$ satisfy*

$$\begin{aligned}
\lim_{h\to 0} \|g(h) - g_0\|_{C^{1,\alpha}} &= 0,\\
\lim_{h\to 0} \|L(h) - L_0\|_{C^{0,\alpha}} &= 0,\\
\lim_{h\to 0} \|L_j(h) - L_{j,0}\|_{C^{0,\alpha}} &= 0, \quad j = 1,\dots,n.
\end{aligned}$$

Then the measure μ is supported in S^*M, for $b \in S^1$,

$$\mu(|b|^2\chi^2) = \lim_{h\to 0} \|\operatorname{Op}_h(b)\chi u\|_{L^2}^2 \tag{3-4}$$

and, for $a \in \mathcal{C}_c^\infty(T^*M)$,

$$\mu(H_p a) = 2\operatorname{Im}\mu^j(a), \tag{3-5}$$

where $p = |\xi|_{g_0}^2 - 1$.

This lemma (together with the results of Section 3C below) suffices to prove our estimate *provided the bicharacteristic flow is unique*. Note that the lemma only requires the hypothesis that $g \in \mathcal{C}^{1,\alpha}$ for $\alpha > 0$ and that this regularity suffices to prove existence of solutions to Hamilton's equations. However, it is not adequate for proving uniqueness. Hence Theorem 1 is only stated for $\alpha = 1$, which is sufficient to guarantee the existence of a well-defined single-valued bicharacteristic flow. (A quantitative result using the dynamics of the multivalued flow for $\alpha < 1$ would be an interesting direction for further research.)

To prove Lemma 3.6, we will need a more general set of symbol classes. Let $r \in \mathbb{N}$, $0 < \alpha < 1$, $0 \le \rho < 1$. We say that $p(x,\xi) \in \mathcal{C}^{r,\alpha} S^m_\rho$ if

$$\|D^\beta_\xi p(\,\cdot\,,\xi)\|_{C^{r,\alpha}} \le C_\beta h^{-r\rho}\langle\xi\rangle^{m-|\beta|}.$$

We say $p \in S^m_\rho$ if $p \in \bigcap_r \mathcal{C}^{r,\alpha} S^m_\rho$. We need the following boundedness property for operators with symbol in $\mathcal{C}^{r,\alpha} S^m$.

Lemma 3.7. *For $r \ge 0$ and $\alpha > 0$, the map $\mathrm{Op}_h : \mathcal{C}^{r,\alpha} S^m \to \mathcal{L}(H^m_h, L^2)$ with $p \mapsto p(x, hD)$ is continuous with norm bounded independently of h.*

Proof. It is enough to show that $\mathrm{Op}_h : \mathcal{C}^{r,\alpha} S^0 \to \mathcal{L}(L^2, L^2)$ is continuous with norm bounded independently of h. For this, we rescale to $h = 1$. That is, let $T_h : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ be given by $(T_h u)(x) = h^{n/4} u(h^{1/2}x)$. Then T_h is unitary and

$$\mathrm{Op}_h(a)u = T_h^* \,\mathrm{Op}_1(a_h) T_h u,$$

where

$$a_h(x,\xi) = a(h^{1/2}x, h^{1/2}\xi).$$

Now, for $a \in \mathcal{C}^{r,\alpha} S^m$,

$$\|D^\beta_\xi a_h(\,\cdot\,,\xi)\|_{\mathcal{C}^{r,\alpha}} \le C_\beta h^{|\beta|/2}\langle h^{1/2}\xi\rangle^{-|\beta|}.$$

In particular, a_h is uniformly bounded in $\mathcal{C}^{r,\alpha} S^m$. Therefore, by [Taylor 2011, Chapter 13, Theorem 9.1] $\mathrm{Op}_1(a_h) : L^2 \to L^2$ uniformly in h with norm bounded by a finite sum of $\mathcal{C}^{r,\alpha} S^0$ seminorms. Since T_h is unitary, this completes the proof. □

Lemma 3.8. *Let $a \in \mathcal{C}^\infty_c(T^*M)$, $m \in \mathbb{R}$, and $0 < \alpha < 1$. Then*

$$\mathcal{C}^{1,\alpha} S^m \to \mathcal{L}(L^2, L^2), \quad p \mapsto h^{-1}[p(x,hD), a(x,hD)],$$

is continuous with norm bounded independently of h.

Proof. Let $p \in \mathcal{C}^{1,\alpha} S^m$ and choose $\rho \in (2/(2+\alpha), 1)$ so that $(1+\alpha/2)\rho > 1$. Let $\psi \in \mathcal{C}^\infty_c(\mathbb{R})$ with $\psi \equiv 1$ on $[-2, 2]$. Define $p_h(x,\xi) = (\psi(h^\rho |D_x|)p)(x,\xi)$. Then, $p_h \in S^m_\rho$. Moreover,

$$\begin{aligned} &|D_x D^\beta_\xi p_h(x,\xi)| \le C_\beta \langle\xi\rangle^{m-|\beta|}, \\ &|D^\gamma_x D^\beta_\xi p_h(x,\xi)| \le C_\beta h^{-\rho(|\gamma|-1)}\langle\xi\rangle^{m-|\beta|}, \quad |\gamma| > 1. \end{aligned}$$

Note that the symbol classes S^m_ρ have a symbol calculus and, in particular,

$$h^{-1}[\mathrm{Op}_h(p_h), a(x,hD)] = \frac{1}{i}\,\mathrm{Op}_h(\{p_h, a\}) + O(h^{1-\rho})_{L^2\to L^2}. \tag{3-6}$$

Next, note that by the characterization of Hölder spaces by Littlewood–Paley decomposition [Taylor 2011, Chapter 13, Theorems 8.1, 8.2],

$$\|(p - p_h)(\,\cdot\,,\xi)\|_{\mathcal{C}^{0,\alpha/2}} \le h^{\rho(1+\alpha/2)}\|p(\,\cdot\,,\xi)\|_{\mathcal{C}^{1,\alpha}} = o(h)\|p(\,\cdot\,,\xi)\|_{C^{1,\alpha}}.$$

In particular, $p - p_h \in o(h)\mathcal{C}^{0,\alpha/2}S^2$ and hence

$$h^{-1}[\mathrm{Op}_h(p-p_h), a(x,hD)] = o(1)_{L^2\to L^2}.$$

Combining this with (3-6) completes the proof. □

Proof of Lemma 3.6. We start with the proof of (3-4). Note that $P = \mathrm{Op}_h(p_0) + h\,\mathrm{Op}_h(p_1)$, where $p_0(h) = |\xi|^2_{g(h)} - 1 \in \mathcal{C}^{1,\alpha}S^2$ and $p_1(h) \in \mathcal{C}^{0,\alpha}S^1$ uniformly as $h \to 0$.

Fix $\psi \in \mathcal{C}^\infty_c(\mathbb{R})$ with $\psi \equiv 1$ on $[-2,2]$ and let $\epsilon > 0$. Let $p_{i,\epsilon}(x,\xi) = (\psi(\epsilon|D_x|)p_i)(x,\xi)$. Then, $p_{i,\epsilon} \in S^{2-i}$ and again by Littlewood–Paley,

$$\|D^\beta_\xi(p_0 - p_{0,\epsilon})(\,\cdot\,,\xi)\|_{\mathcal{C}^{0,\alpha}} \le C_\beta \epsilon \langle\xi\rangle^{2-|\beta|}.$$

In particular, for $\epsilon > 0$ small and $h < h_0$ small, p_ϵ is elliptic on $||\xi|_g - 1| \ge C\epsilon$. Therefore, for $a \in S^0(T^*M)$ supported away from $p = 0$, there is $\epsilon > 0$, h_0 small enough so that $p_{0,\epsilon}$ is elliptic on $\operatorname{supp} a$. Hence there is $E_\epsilon \in \Psi^{-\infty}$ (uniformly for $\epsilon > 0$ small) so that

$$a(x,hD) = E_\epsilon\,\mathrm{Op}_h(p_{0,\epsilon}) + O(h^\infty)_{\Psi^{-\infty}}.$$

In particular,

$$a(x,hD) = E_\epsilon(P - \mathrm{Op}_h(p_0 - p_{0,\epsilon}) - h\,\mathrm{Op}_h(p_1)) + O(h^\infty)_{\Psi^{-\infty}}.$$

Therefore,

$$\begin{aligned}\mu(a) &= \lim_{h\to 0}\langle E_\epsilon P u, u\rangle - \langle E_\epsilon[\mathrm{Op}_h(p_0 - p_{0,\epsilon}) + h\,\mathrm{Op}_h(p_1)]u, u\rangle\\ &= \lim_{h\to 0}\langle E_\epsilon h\chi f, u\rangle + O(\epsilon) = O(\epsilon).\end{aligned}$$

Since the left-hand side does not depend on ϵ, sending $\epsilon \to 0$, we have $\mu(a) = 0$ and hence $\operatorname{supp}\mu \subset S^*M$.

Next, fix $K > 0$ large enough so that, for $h < h_0$, $\{|\xi|_{g(h)} \le 2\} \subset \{|\xi| \le K\}$. We apply the previous argument with $a(x,hD) = \chi(x)b(x,hD)^*b(x,hD)\chi(x)(1 - \psi(K^{-1}|hD|))$. In particular, there exists $E_\epsilon \in \Psi^0$ compactly supported so that

$$\chi(x)b(x,hD)^*b(x,hD)\chi(x)(1-\psi(|hD|_g)) = E_\epsilon P + O_{L^2_{\mathrm{loc}}\to L^2_{\mathrm{comp}}}(\epsilon).$$

Therefore,

$$\begin{aligned}\lim_{h\to 0}\|b(x,hD)\chi u\|^2_{L^2} &= \lim_{h\to 0}\langle\chi(x)b(x,hD)^*b(x,hD)\chi(x)u, u\rangle\\ &= \lim_{h\to 0}\langle\chi(x)b(x,hD)^*b(x,hD)\chi(x)\psi(K^{-1}|hD|)u, u\rangle\\ &\qquad + \lim_{h\to 0}\langle\chi(x)b(x,hD)^*b(x,hD)\chi(x)(1-\psi(K^{-1}|hD|))u, u\rangle\\ &= \lim_{h\to 0}\langle\chi(x)b(x,hD)^*b(x,hD)\chi(x)\psi(K^{-1}|hD|)u, u\rangle + \lim_{h\to 0}\langle hE_\epsilon f, u\rangle + O(\epsilon)\\ &= \mu(|b|^2\chi^2\psi(|\xi|_g)) + O(\epsilon) = \mu(|b|^2\chi^2) + O(\epsilon).\end{aligned}$$

Again, since the left hand side is independent of ϵ, this proves (3-4).

We now prove (3-5). Let $a \in \mathcal{C}^\infty_c(T^*M;\mathbb{R})$. Then,

$$\frac{i}{h}\langle[P, a(x,hD)]u, u\rangle = i[(a(x,hD)u, h\chi f) - (ha(x,hD)\chi f, u)] \to 2\,\mathrm{Im}\,\mu^j(a).$$

Therefore, it remains to show that

$$\frac{i}{h}\langle [P, A]u, u\rangle \to \mu(H_p a). \tag{3-7}$$

First, observe that by Lemma 3.7 and the fact that

$$\|D_\xi^\beta (p_1 - p_{1,\epsilon})(\,\cdot\,, \xi)\|_{\mathcal{C}^{0,\alpha/2}} \le C_\beta \epsilon^{\alpha/2}\langle \xi\rangle^{1-|\beta|},$$

we obtain

$$\|ha(x, hD)\operatorname{Op}_h(p_1 - p_{1,\epsilon})\| + \|h\operatorname{Op}_h(p_1 - p_{1,\epsilon})a(x, hD)u\|_{L^2} \le Ch\epsilon^{\alpha/2}.$$

Therefore,

$$\|[h\operatorname{Op}_h(p_1), a(x, hD)u]\|_{L^2} \le Ch\epsilon^{\alpha/2}\|u\|_{L^2} + O_\epsilon(h^2). \tag{3-8}$$

Next, observe that by Lemma 3.8,

$$\|[\operatorname{Op}_h(p_0 - p_{0,\epsilon}), a(x, hD)]u\|_{L^2} \le C\epsilon h\|u\|_{L^2}. \tag{3-9}$$

Finally, with $q_\epsilon := \lim_{h\to 0} p_{0,\epsilon}$,

$$\lim_{h\to 0}\left\langle \frac{i}{h}[\operatorname{Op}_h(p_{0,\epsilon}), a(x, hD)]u, u\right\rangle = \mu(H_{q_\epsilon} a). \tag{3-10}$$

Combining (3-8), (3-9), and (3-10) gives

$$\left|\lim_{h\to 0}\frac{i}{h}\langle [P, A]u, u\rangle - \mu(H_{q_\epsilon} a)\right| \le C\epsilon^{\alpha/2}.$$

Since $p_0 \in \mathcal{C}^{1,\alpha}S^2$ uniformly in h, we have $H_{q_\epsilon} \to H_p$. In particular, sending $\epsilon \to 0$ and applying the dominated convergence theorem gives (3-7). □

3C. ***Appearance of the Volterra operator.*** To complete the proof of (1-3) in the boundaryless case, we prove the following measure-theoretic proposition. This lemma will be modified slightly in the case of a manifold with boundary to account for the fact that the generalized bicharacteristic flow is not generated by a vector field.

Proposition 3.9. *Suppose that μ, μ^j, and α are Radon measures on T^*M with α finite and μ finite on compact subsets of T^*M. Let X be a continuous vector field on T^*M and $\varphi_t : T^*M \to T^*M$ be given by $\varphi_t(q) = \exp(tX)(q)$. Let $\Sigma \subset T^*M$ be a hypersurface transverse to X so that the map $F : \mathbb{R}\times\Sigma \to T^*M$, $(t, q) \mapsto \varphi_t(q)$, is a homeomorphism onto its image. Suppose that $\mu(F((-\infty, 0)\times\Sigma)) = 0$ and, for $a \in \mathcal{C}_c^0(T^*M; \mathbb{R})$,*

$$|\mu^j(a)| \le \sqrt{\mu(|a|)\alpha(|a|)}. \tag{3-11}$$

Furthermore, for $\psi \in \mathcal{C}_c^1(\mathbb{R})$, $a \in \mathcal{C}_c^0(\Sigma)$, let $f_{\psi,a}(F(t, q)) = \psi(t)a(q)$ and assume that

$$\mu(f_{\partial_t\psi,a}) = 2\operatorname{Im}\mu^j(f_{\psi,a}). \tag{3-12}$$

Then,

$$\mu(F([0, L]\times\Sigma)) \le \frac{4L^2}{\pi^2}\alpha(F([0, L]\times\Sigma))).$$

Proof. For $0 \le a \in \mathcal{C}^0_c(T^*M)$, define the Radon measures $\mu^\sharp_a$, $\operatorname{Im}\mu^{j,\sharp}$, and $\alpha^\sharp_a$ on $\mathbb{R}$ by

$$\mu^\sharp_a(\psi) = \mu(f_{\psi,a}), \quad \mu^{j,\sharp}_a(\psi) = \mu^j(f_{\psi,a}), \quad \alpha^\sharp_a(\psi) = \mu(f_{\psi,a}).$$

Lemma 3.10. *The measure $\mu^\sharp$ is absolutely continuous with respect to Lebesgue measure and*

$$\mu^\sharp_a = \dot\mu^\sharp_a\, dt,$$

with $|\dot\mu^\sharp_a| \le C \sup|a|$.

Proof. First, observe that for $\psi \in \mathcal{C}^1_c(\mathbb{R})$, (3-12) implies

$$\mu^\sharp_a(\partial_t\psi) = 2\operatorname{Im}\mu^j(f_{\psi,a}).$$

Let $[b,c] \subset \mathbb{R}$, $0 \le \chi_\epsilon \in \mathcal{C}^\infty_c(\mathbb{R};[0,1])$, with $\chi \equiv 1$ on $[b,c]$ and $\operatorname{supp}\chi_\epsilon \subset (b-\epsilon, c+\epsilon)$. Define

$$\psi_\epsilon(t) = -\int \chi_\epsilon(s)\, ds + \int_{-\infty}^t \chi_\epsilon(s)\, ds.$$

Then, since $\mu((-\infty,0)\times\Sigma) = 0$, we have $\mu^j(F((-\infty,0)\times\Sigma)) = 0$ and hence

$$\begin{aligned}
|\mu^\sharp_a(\chi_\epsilon)| &= 2|\operatorname{Im}\mu^j(1_{F([0,\infty)\times\Sigma)} f_{\psi_\epsilon,a})| \\
&\le C\sup|a|\sup|\psi_\epsilon| \\
&\le C\sup|a|\left|\int \chi_\epsilon\, ds\right| \le C\sup|a|(c-b+2\epsilon).
\end{aligned}$$

Since $\chi_\epsilon \to 1_{[b,c]}$, we have by the dominated convergence theorem,

$$|\mu^\sharp_a([a,b])| \le C\sup|a|(c-b).$$

In particular, $\mu^\sharp_a$ is absolutely continuous with respect to Lebesgue measure since the intervals generate the Borel sets on $\mathbb{R}$. Moreover, its Radon–Nikodym derivative is linear in a and bounded by $C\sup|a|$. □

By (3-11), since $\mu^\sharp_a$ is absolutely continuous with respect to Lebesgue measure, $\mu^{j,\sharp}_a$ is also. Let $\dot\mu^\sharp_a$, $\dot\mu^{j,\sharp}_a$, and $\dot\alpha^\sharp_a$ be the Radon–Nikodym derivatives of $\mu^\sharp_a$, $\mu^{j,\sharp}_a$, and $\alpha^\sharp_a$ respectively with respect to Lebesgue measure. In particular,

$$\mu^\sharp_a = \dot\mu^\sharp_a\, dt, \quad \operatorname{Im}\mu^{j,\sharp}_a = \operatorname{Im}\dot\mu^{j,\sharp}_a\, dt, \quad \alpha^\sharp_a = \dot\alpha^\sharp_a\, dt + \lambda,$$

where $\lambda \perp dt$. Now, by (3-11),

$$\frac{1}{2r}|\mu^{j,\sharp}_a([t-r,t+r])| \le \frac{1}{2r}\sqrt{\mu^\sharp_a([t-r,t+r])}\sqrt{\alpha^\sharp_a([t-r,t+r])}.$$

So, by the Lebesgue differentiation theorem [Folland 1999, Theorem 3.22], for Lebesgue a.e. t,

$$|\dot\mu^{j,\sharp}_a(t)| \le \sqrt{\dot\mu^\sharp_a(t)}\sqrt{\dot\alpha^\sharp_a(t)}. \tag{3-13}$$

Lemma 3.11. *With $\dot\mu^\sharp_a(t)$ as above, $\dot\mu^\sharp_a(t)$ is continuous and*

$$\dot\mu^\sharp_a(t) \le 2\int_0^{\max(t,0)} |\operatorname{Im}\dot\mu^{j,\sharp}_a(s)|\, ds \le 2\int_0^t \sqrt{\dot\mu^\sharp_a(s)}\sqrt{\dot\alpha^\sharp_a(s)}\, ds. \tag{3-14}$$

Proof. By (3-12), for $\psi \in \mathcal{C}_c^1(\mathbb{R})$,

$$\int \psi'(t)\dot{\mu}_a^\sharp(t)\,dt = 2\int \psi(t)\,\mathrm{Im}\,\dot{\mu}_a^{j,\sharp}(t)\,dt. \tag{3-15}$$

Let $\chi \in \mathcal{C}_c^\infty(\mathbb{R};[0,1])$ with $\int \chi(s)\,ds \equiv 1$ and define $\chi_{\epsilon,t'}(s) = \epsilon^{-1}\chi(\epsilon^{-1}(s-t'))$ and

$$\psi_{\epsilon,t'}(s) = -1 + \int_{-\infty}^{t} \chi_{\epsilon,t'}(s)\,ds.$$

Then by (3-15), together with the fact that $\mu_a^{j,\sharp}(t) = 0$ for $t < 0$,

$$\int \chi_{\epsilon,t'}(t)\dot{\mu}_a^\sharp(t)\,dt = 2\int_0^\infty \psi_{\epsilon,t'}(t)\,\mathrm{Im}\,\dot{\mu}_a^{j,\sharp}(t)\,dt.$$

Applying the Lebesgue differentiation theorem on the left and dominated convergence on the right, we find for Lebesgue a.e. t'

$$\dot{\mu}_a^\sharp(t') = -2\int_0^{\max(t',0)} \mathrm{Im}\,\dot{\mu}_a^{j,\sharp}(t)\,dt.$$

Note that this implies $\dot{\mu}_a^\sharp(t')$ is continuous. Applying (3-13) gives (3-14). □

Lemma 3.12. *Let $0 \le b(t), c(t) \in L^2$ with $b(t)$ continuous and*

$$b^2(t) \le 2\int_0^t b(s)c(s)\,ds.$$

Then

$$b(t) \le \int_0^t c(s)\,ds. \tag{3-16}$$

Proof. Supposing (3-16) to be false, there exists $\epsilon > 0$ and $t > 0$ such that

$$b(t) \ge \int_0^t c(s)\,ds + \epsilon;$$

hence we may define

$$t_0 := \inf\left\{t > 0 : \int_0^t c(s)\,ds + \epsilon \le b(t)\right\}.$$

Then,

$$\begin{aligned}
\left[\int_0^{t_0} c(s)\,ds + \epsilon\right]^2 &= b^2(t_0) \le 2\int_0^{t_0} b(s)c(s)\,ds \\
&\le 2\int_0^{t_0}\int_0^s c(r)c(s)\,dr\,ds + 2\epsilon\int_0^{t_0} c(s)\,ds \\
&= \left[\int_0^{t_0}\int_0^s c(r)c(s)\,dr\,ds + \int_0^{t_0}\int_s^{t_0} c(r)c(s)\,dr\,ds\right] + 2\epsilon\int_0^{t_0} c(s)\,ds \\
&= \left[\int_0^{t_0} c(s)\,ds\right]^2 + 2\epsilon\int_0^{t_0} c(s)\,ds \\
&= \left[\int_0^{t_0} c(s)\,ds + \epsilon\right]^2 - \epsilon^2,
\end{aligned}$$

which is a contradiction. □

Using the continuity of $\dot{\mu}_a^\sharp$ together with (3-14) to apply Lemma 3.12 with $b(t)=\sqrt{\dot{\mu}_a^\sharp(t)}$, $c(t)=\sqrt{\dot{\alpha}_a^\sharp(s)}$, we have

$$\sqrt{\dot{\mu}_a^\sharp(t)} \le \int_0^t \sqrt{\dot{\alpha}_a^\sharp(s)}\, ds.$$

Now,

$$\begin{aligned}\mu(F([0,L]\times\Sigma) &= \int_0^L \dot{\mu}_1^\sharp(t)\, dt \le \int_0^L \left| \int_0^t \sqrt{\dot{\alpha}_1^\sharp(s)}\, ds \right|^2 dt \\ &= \left\| V\sqrt{\dot{\alpha}_1^\sharp} \right\|^2_{L^2([0,L])} \le \|V\|^2_{L^2([0,L])\to L^2([0,L])} \int_0^L \dot{\alpha}_1^\sharp(t)\, dt,\end{aligned}$$

where $V : L^2([0, L]) \to L^2([0, L])$ is the Volterra operator. Since the Volterra operator acting on $L^2([0, L])$ has norm $(2L)/\pi$ [Halmos 1967, Problem 188] and

$$\alpha(F([0,L]\times\Sigma)) = \int_0^L \dot{\alpha}_1^\sharp(t)\, dt,$$

we have

$$\mu(F([0,L]\times\Sigma)) \le \frac{4L^2}{\pi^2}\alpha(F([0,L]\times\Sigma)). \qquad \square$$

3D. ***Completion of the proof of* (1-3) *in the boundaryless case.*** Let

$$\Sigma = \{\rho \in T^*_{\partial B(0,R)}M : (H_p r)(\rho) < 0\},$$

where r is the radial variable defined in the Euclidean end(s) as in Definition 2.1. Thus, these are the inward-pointing covectors over the boundary of the ball of radius R. By convexity of Euclidean balls,

$$\bigcup_{t>0} \exp(tH_p)(\Sigma) \supset \mathcal{I}^c, \quad \bigcup_{t\le 0} \exp(tH_p)(\Sigma) \subset \mathcal{I},$$

and F as in Proposition 3.9 is a diffeomorphism onto its image.

Then, by (3-3) and Lemmas 3.3, 3.4, and 3.6, $(\mu, \mu^j, \alpha, \Sigma)$ satisfy the hypotheses of Proposition 3.9 with $\varphi_t = \exp(tH_p)$. Therefore,

$$\mu\left(\bigcup_{t=0}^L \varphi_t(\Sigma \cap S^*M)\right) \le \frac{4L^2}{\pi^2}\alpha(T^*M),$$

where we may take $L = L(g, R')$. So, since $0 \le \chi^2 \le 1$ and

$$\operatorname{supp}\chi \cap S^*M \subset \bigcup_{t=0}^{L(g,R')} \varphi_t(\Sigma \cap S^*M),$$

we conclude

$$\limsup_{h\to 0} \|\chi u\|_{L^2} = \sqrt{\mu(\chi^2)} \le \frac{2L(g,R')}{\pi}\sqrt{\alpha(T^*M)} \le \liminf \frac{2L(g,R')}{\pi}\|f(h)\|_{L^2}.$$

This completes the proof of (1-3) in the case of a manifold without boundary.

3E. ***Uniformity of k_0 and Sobolev estimates.*** In this section we prove the uniformity of k_0 statement from Theorem 1, as well as the more general Sobolev mapping property (1-4). We will omit the terms L, L_j from this discussion for brevity, but will on the other hand prove a slightly more general statement about metric perturbations as follows.

We begin with uniformity in the case $s=0$, i.e., in the basic L^2 estimate. Fix $R_1>0$ and let $\mathcal{C}$ be a subset of $\mathcal{C}^{1,1}$ metrics so that, for $g\in\mathcal{C}$, g is nontrapping and $\operatorname{supp}(I-g)\subset B(0,R_1)$. Furthermore, suppose that for any $\{g_k\}_{k=1}^\infty\subset\mathcal{C}$, there exists a subsequence $\{g_{k_m}\}_{m=1}^\infty$, $\gamma>0$, and $g\in\mathcal{C}$ so that $\|g_{k_m}-g\|_{\mathcal{C}^{1,\gamma}}\to 0$ and $\lim_{m\to\infty} L(g_{k_m},R_1)\geq L(g,R_1)$. Let $\chi\in\mathcal{C}_c^\infty(B(0,R);[0,1])$ and $R_1<R'<R''<R$ so that $\operatorname{supp}\chi\subset B(0,R')$.

We start by showing that there exists $k_0>0$ so that, for all $g\in\mathcal{C}$ and $k>k_0$,

$$\|\chi(P(g)-k^2-i0)^{-1}\chi\|_{L^2\to L^2}\leq\frac{2L(g,R'')}{\pi k}. \tag{3-17}$$

Suppose not. Then since $L(g,R')<L(g,R'')$, there exists $\{g_k\}_{k=1}^\infty\subset\mathcal{C}$, $h_k\to 0$, $f_k\in L^2(M)$, $u_k\in H^s_{\mathrm{loc}}(M)$, with $\|\chi u_k\|_{L^2}=1$, and $\delta>0$ so that

$$(h_k^2P(g_k)-1)u_k=h_k\chi f_k,\quad 1=\|\chi u_k\|_{L^2}\geq\frac{2L(g_k,R')\|f_k\|_{L^2(M)}}{\pi}+\delta. \tag{3-18}$$

Extracting subsequences, we may assume that $g_k\to g$ in $\mathcal{C}^{1,\gamma}$ for some $\gamma>0$ and $\lim L(g_k,R')\geq L(g,R')$. Note also that by Lemma 3.2, u_k is uniformly bounded in L^2_{loc} and hence, by extracting further subsequences, we may assume that u_k has defect measure μ, χf_k has defect measure α, and u_k, χf_k have joint defect measure μ^j. Let Σ denote the set of directly incoming points on $T^*_{\partial B(0,R')}M$. By Lemmas 3.3, 3.4, and 3.6, $(\mu,\mu^j,\alpha,\Sigma)$ satisfy the hypotheses of Proposition 3.9 with $\varphi_t=\exp(tH_p)$ and $p=|\xi|_g^2-1$. Therefore,

$$\mu\left(\bigcup_{t=0}^L\varphi_t(\Sigma\cap S^*M)\right)\leq\frac{4L^2}{\pi^2}\alpha(T^*M),$$

and we arrive at a contradiction just as above.

Finally, to obtain the bound (1-4), we begin by considering the case $s=2$. We compute

$$(P(g)+k^2)\chi(P(g)-k^2-i0)^{-1}\chi f=\chi^2 f+([-\Delta_{\mathbb{R}^n},\chi]+2k^2\chi)(P(g)-k^2-i0)^{-1}\chi f.$$

We next consider $[-\Delta_{\mathbb{R}^n},\chi]$ and show that there exists $k_1>0$ so that, for $k>k_1$ and $g\in\mathcal{C}$,

$$\|[-\Delta_{\mathbb{R}^n},\chi](P-k^2-i0)^{-1}\chi\|_{L^2\to L^2}\leq\sup_{S^*M}|H_{|\xi|^2}\chi|\frac{2L(g_k,R'')}{\pi}. \tag{3-19}$$

Suppose that (3-19) does not hold. Then there is $\delta>0$ and a sequence $h_k\to 0$, $f_k\in L^2(M)$, $u_k\in H^2(M)$ with $\|f_k\|_{L^2}=1$, u_k so that

$$(h_k^2P(g_k)-1)u_k=\chi h_k f_k,\quad u_k|_{\partial M}=0,$$
$$\|h_k^{-1}[-h_k^2\Delta_{\mathbb{R}^n},\chi]u_k\|_{L^2}\geq\sup_{S^*M}|H_{|\xi|^2}\chi|\frac{2(L(g_k,R')+\delta)}{\pi}.$$

Let $\tilde\chi \in \mathcal{C}_c^\infty(B(0,R');[0,1])$ with $\tilde\chi \equiv 1$ on $\operatorname{supp}\chi$. As before, we may assume that u_k has a defect measure and then, by Lemma 4.2 applied with χ replaced by $\tilde\chi \in C_c^\infty(B(0,R'))$,

$$\begin{aligned}\sup_{S^*M}|H_{|\xi|^2}\chi^2|\frac{4(L(g_k,R')+\delta)^2}{\pi^2} &\le \|h^{-1}[-h^2\Delta_{\mathbb{R}^n},\chi]\tilde\chi u_k\|_{L^2}^2 \to \mu(|H_{|\xi|^2}\chi|^2\tilde\chi^2)\\ &\le \sup_{S^*M}|H_{|\xi|^2}\chi|^2\mu(\tilde\chi^2) \le \sup_{S^*M}|H_{|\xi|^2}\chi|^2\frac{4(L(g,R'))^2}{\pi^2},\end{aligned}$$

a contradiction.

Thus, by (3-17), for $k>k_0$, and any $g\in\mathcal{C}$,

$$\begin{aligned}&\|(P(g)+k^2)\chi(P(g)-k^2-i0)^{-1}\chi f\|_{L^2}\\ &\quad\le 2k^2\|\chi(P(g)-k^2-i0)^{-1}\chi f\|_{L^2}+\|\chi^2 f\|+\|[-\Delta_{\mathbb{R}^n},\chi](P(g)-k^2-i0)^{-1}\chi f\|\\ &\quad\le \frac{4L(g,R'')k}{\pi}\|\chi f\|_{L^2}+\|[-\Delta_{\mathbb{R}^n},\chi](P(g)-k^2-i0)^{-1}\chi f\|_{L^2}+\|\chi^2 f\|_{L^2}.\end{aligned}$$

Using (3-19), we have for $k>k_1$, and any $g\in\mathcal{C}$,

$$\|[-\Delta_{\mathbb{R}^n},\chi](P(g)-k^2-i0)^{-1}\chi f\|_{L^2}\le \sup_{S^*M}|H_{|\xi|^2}\chi|\frac{2L(g_k,R'')}{\pi}\|f\|_{L^2}.$$

In particular, letting

$$k>\max\left(k_0,k_1,\frac{\sup_{S^*M}|H_{|\xi|^2}\chi|2L(g,R'')+\pi}{4(L(g,R)-L(g,R''))}\right),$$

we have for all $g\in\mathcal{C}$,

$$\begin{aligned}\|\chi(P(g)-k^2-i0)^{-1}\chi f\|_{H^2}&=\|(P(g)+k^2)\chi(P(g)-k^2-i0)^{-1}\chi f\|_{L^2}\\ &\le \frac{4L(g,R)k}{\pi}\|f\|_{L^2}.\end{aligned}\tag{3-20}$$

Since $L(g,R'')\le L(g,R)$, the general case of $0\le s\le 2$ then follows by interpolation between (3-17) and (3-20).

If we choose $\mathcal{C}$ to be a subset of an open neighborhood in the $\mathcal{C}^{2,\alpha}$ topology of a given $\mathcal{C}^{2,\alpha}$ nontrapping metric, then provided this neighborhood is chosen sufficiently small, all the metrics in $\mathcal{C}$ are nontrapping, and the subsequence condition is guaranteed by the compactness of the embedding $\mathcal{C}^{2,\alpha}\hookrightarrow\mathcal{C}^{1,\gamma}$; indeed we have subsequential convergence in $\mathcal{C}^{1,1}$, which ensures that $\lim L(g_{k_m},R_1)=L(g,R_1)$. This choice of $\mathcal{C}$ then gives the uniformity assertion from Theorem 1. This completes the proof of (1-3) in the case of a manifold without boundary.

4. Manifolds with boundary

We now turn to the case of manifolds with boundary. Our treatment of the propagation of defect measures in this setting is motivated by [Miller 2000; Burq and Lebeau 2001]; see also [Burq 1997].

Let $\widetilde{M}$ be an open manifold extending M and extend P to an operator on $\widetilde{M}$. We then let $\underline{u}$ be the extension of u by 0. As in the boundaryless case, we will assume that we have a sequence $h=h_j\downarrow 0$ with

$$Pu(h)=h\chi f(h)\quad\text{in } M,\qquad u|_{x_1=0}=0,$$

and we take

$$\|f(h)\| = 1, \quad \lim_{h\to 0} \|\chi u\|_{L^2} = M > L(\Omega, g, R).$$

Since propagation of singularities is a local consideration, we may employ Riemannian normal coordinates (x_1, x') in which ∂M is given by $x_1 = 0$, $M = \{x_1 > 0\}$, and the operator P is given by

$$P = (hD_{x_1})^2 + R(x_1, x', hD_{x'}) + hE \tag{4-1}$$

for some self-adjoint $E \in \mathrm{Diff}_h^1$ and for $R \in \mathrm{Diff}_h^2$ with symbol $r(x_1, x', \xi')$.

We proceed as before, letting μ be a defect measure of $\underline{u}$, α be a defect measure of $\underline{\chi f}$, and μ^j a joint defect measure of $\underline{u}$ and $\underline{\chi f}$. Finally, let $\dot{\nu}$ be the defect measures of $hD_{x_1}u|_{x_1=0}$ on ∂M. Thus,

$$\begin{aligned} &\lim_{h\downarrow 0} \langle a(x, hD)\underline{u}, \underline{u}\rangle = \int a\, d\mu, && \lim_{h\downarrow 0} \langle a(x, hD)\underline{\chi f}, \underline{\chi f}\rangle = \int a\, d\alpha, \\ &\lim_{h\downarrow 0} \langle a(x, hD)\underline{\chi f}, \underline{u}\rangle = \int a\, d\mu^j, && \lim_{h\downarrow 0} \langle b(x', hD')hD_x u, hD_x u\rangle_{\partial M} = \int_{\partial M} b\, d\dot{\nu}. \end{aligned} \tag{4-2}$$

4A. ***Local study of the measure μ.***

Lemma 4.1. *For $a = a_0(x, \xi') + a_1(x, \xi')\xi_1$ with $a_i \in \mathcal{C}_c^\infty([0, \epsilon) \times T^*\partial M)$, we have*

$$\mu(H_p a) = 2\,\mathrm{Im}\,\mu^j(a) - \dot{\nu}(a_1|_{x_1=0}).$$

Proof. Here we work in a small neighborhood of the boundary and write

$$A = a_0(x, hD') + a_1(x, hD')hD_{x_1}.$$

Equation (3-2) must now be modified owing to boundary terms arising in integration by parts. In particular,

$$\langle (hD_{x_1})^2 Au, u\rangle_M = ih(\langle hD_{x_1}Au, u\rangle_{L^2(\partial M)} + \langle Au, hD_{x_1}u\rangle_{L^2(\partial M)}) + \langle Au, (hD_{x_1})^2 u\rangle_{L^2(M)}.$$

There is also a term arising when we integrate by parts to change $\langle f, Au\rangle$ to $\langle Au, f\rangle$ but this term is $O(h)$. Now,

$$\begin{aligned} \langle hD_{x_1}Au, u\rangle_{L^2(\partial M)} &= \langle a_0 hD_{x_1}u + a_1(hD_{x_1})^2 u, u\rangle_{L^2(\partial M)} + o(1) \\ &= \langle a_0 hD_{x_1}u + a_1(1 + h^2\Delta_{\partial M})u, u\rangle_{L^2(\partial M)} + \langle a_1 h\chi f, u\rangle_{\partial M} + o(1). \end{aligned}$$

With Dirichlet boundary conditions, this is actually just 0. Moreover,

$$\langle Au, hD_{x_1}u\rangle_{L^2(\partial M)} = \langle [a_0 + a_1 hD_{x_1}]u, hD_{x_1}u\rangle = \langle a_1 hD_{x_1}u, hD_{x_1}u\rangle.$$

So (3-2) now reads

$$ih^{-1}\langle [P, A]u, u\rangle_M = \mathrm{Im}\langle A\chi f, u\rangle_M - \langle a_1 hD_{x_1}u, hD_{x_1}u\rangle_{\partial M} + o(1). \qquad \square$$

We want to upgrade this to a statement for all functions in $\mathcal{C}_c^\infty(T^*M)$. First, we need to show that μ is supported on S^*M.

Lemma 4.2. *Let μ be a defect measure associated to u as above. Then* $\operatorname{supp}\mu \subset S^*M$ *and* $\mu(\chi^2) = 1$. *Moreover, for $b \in S^1(T^*M)$, supported away from ∂M,*

$$\lim_{h\to 0} \|\operatorname{Op}_h(b)\chi u\|_{L^2}^2 = \mu(|b|^2\chi^2).$$

Proof. Note that

$$P\underline{u} = 1_{x_1\geq 0}Pu + h(-\delta(x_1)\otimes h\partial_{x_1}u|_{x_1=0} + h\delta'(x_1)\otimes u|_{x_1=0}).$$

Now, suppose $a \in \mathcal{C}_c^\infty(T^*\widetilde{M})$ has $\operatorname{supp} a \cap \{p = 0\} = \varnothing$. Then, there exist $E \in \Psi^{-2}$ so that $a(x, hD) = EP + O_{\mathcal{D}'\to\mathcal{C}^\infty}(h^\infty)$. Therefore,

$$a(x, hD)\underline{u} = a(x, hD)1_{x_1\geq 0}Pu + hE(-\delta(x_1)\otimes h\partial_{x_1}u|_{x_1=0}).$$

Since $E : H_h^s \to H_h^{s+2}$ is bounded and $\|h\partial_{x_1}u|_{x_1=0}\| \leq C$, we have

$$a(x, hD)\underline{u} = a(x, hD)1_{x_1\geq 0}Pu + O_{L^2}(h^{1/2}).$$

In particular, this implies $\mu(a) = 0$ completing the proof.

To see that $\mu(\chi^2) = 1$, let $\psi \in \mathcal{C}_c^\infty(\mathbb{R})$ with $\psi \equiv 1$ on $|s| \leq 2$. Then, as above there exists $E \in S^{-2}$ so that

$$\chi(1-\psi)(|hD|)\chi = EP + O_{\mathcal{D}'\to\mathcal{S}}(h^\infty)$$

and hence again

$$\chi(1-\psi)(|hD|))\chi\underline{u} = O_{L^2}(h^{1/2}).$$

In particular,

$$\begin{aligned} 1 = \langle\chi^2\underline{u}, \underline{u}\rangle &= \langle\chi\psi(|hD|)\chi\underline{u}, \underline{u}\rangle + O(h^{1/2}) \\ &= \int \chi^2(x)\psi(|\xi|)\,d\mu = \int \chi^2\,d\mu, \end{aligned}$$

where in the last line we use the fact that μ is supported on S^*M.

The last claim follows the same arguments as in the boundaryless case. □

We now introduce notation for the even and odd parts of smooth functions on T^*M.

Definition 4.3. For $a \in \mathcal{C}^\infty(T^*M)$, let

$$a_o(x, \xi_1, \xi') = \frac{a(x, \xi_1, \xi') - a(x, -\xi_1, \xi')}{2\xi_1}, \quad a_e(x, \xi_1, \xi') = \frac{a(x, \xi_1, \xi') + a(x, -\xi_1, \xi')}{2}.$$

Thus, $a = a_e + \xi_1 a_o$ and a_e, a_o are both even functions of ξ_1.

Next we upgrade the test functions for which we can compute $\mu(H_p a)$.

Lemma 4.4. *For $a \in \mathcal{C}_c^\infty(T^*\widetilde{M})$,*

$$\mu(H_p a) = 2\operatorname{Im}\mu^j(a) - \dot{v}(a_o|_{x_1=0}).$$

In particular, for $a = a(x, x_1\xi_1, \xi') \in \mathcal{C}_c^\infty({}^bT^*\overline{\mathbb{R}^n_+})$,

$$\mu(H_p a) = 2\operatorname{Im}\mu^j(a).$$

Remark 4.5. Here, ${}^bT^*\overline{\mathbb{R}^n_+}$ denotes the b-cotangent bundle of this local model for a manifold with boundary, constructed so that smooth functions on this space are simply functions of $(x, x_1\xi_1, \xi')$. See Section 4B for a more precise description of this space and for additional references. We are regarding a here as extended to a smooth function on $T^*\widetilde{M}$; the choice of extension is immaterial, as μ is supported on M.

Proof. Since a_o and a_e are both even functions, we may write

$$a_{e/o}(x, \xi_1, \xi') = \tilde{a}_{e/o}(x, \xi_1^2, \xi')$$

for smooth functions $\tilde{a}_{e/o}$; thus,

$$a = \tilde{a}_e(x, \xi_1^2, \xi') + \tilde{a}_o(x, \xi_1^2, \xi')\xi_1.$$

Since H_p is tangent to $S^*\widetilde{M}$, we have $H_p 1_{S^*\widetilde{M}} a = 1_{S^*\widetilde{M}} H_p a$. In particular, since $\operatorname{supp}\mu \subset S^*\widetilde{M}$,

$$\mu(H_p a) = \mu(H_p 1_{S^*\widetilde{M}} a).$$

Now,

$$S^*\widetilde{M} := \{\xi_1^2 - r(x, \xi') = 0\}.$$

Therefore,

$$1_{S^*\widetilde{M}} a = 1_{S^*\widetilde{M}}[\tilde{a}_e(x, r(x, \xi'), \xi') + \tilde{a}_o(x, r(x, \xi'), \xi')\xi_1].$$

Let $b_{e/o} = a_{e/o}(x, r(x, \xi'), \xi')$. Then we have

$$\begin{aligned}\mu(H_p a) &= \mu(H_p 1_{S^*\widetilde{M}}[b_e + b_o\xi_1]) \\ &= \mu(H_p(b_e + b_o\xi_1)) \\ &= 2\operatorname{Im}\mu^j(b_e + b_o\xi_1) - \dot{v}(b_o|_{x_1=0}).\end{aligned}$$

Now, since $\operatorname{supp}\mu \subset S^*M$, we have $\operatorname{supp}\mu^j \subset S^*M$ (by Lemma 3.3) and hence

$$\mu^j(b_e + b_o\xi_1) = \mu^j(a).$$

Therefore, we have

$$\mu(H_p a) = 2\operatorname{Im}\mu^j(a) - \dot{v}(b_o|_{x_1=0}).$$ □

Before proceeding further, we recall the decomposition of $T^*_{\partial M}M$ into elliptic, hyperbolic, and glancing regions. In the notation of (4-1), with r the symbol of $R(x_1, x', hD_{x'})$, we write

$$\begin{aligned}(y, \xi_1, \xi') &\in \mathcal{E} \quad \text{if } r(0, x', \xi') > 0, \\ (y, \xi_1, \xi') &\in \mathcal{H} \quad \text{if } r(0, x', \xi') < 0, \\ (y, \xi_1, \xi') &\in \mathcal{G} \quad \text{if } r(0, x', \xi') = 0\end{aligned}$$

to denote the elliptic, hyperbolic, and glancing sets respectively.

We further split the glancing set into subsets as follows: we let

$$\begin{aligned}\mathcal{G}_d &:= \mathcal{G} \cap \{H_p^2 x_1 > 0\}, \\ \mathcal{G}_g &:= \mathcal{G} \cap \{H_p^2 x_1 < 0\};\end{aligned}$$

these are the *diffractive* and *gliding* sets respectively. We will also employ the filtration of the remainder of the glancing set given by the order of contact with the boundary as follows:

$$\mathcal{G}^k := \{\rho \in \mathcal{G} : (H_p^j x_1)(\rho) = 0 \text{ for } 0 \leq j < k, \text{ and } (H_p^k x_1)(\rho) \neq 0\}.$$

(We remark that this definition differs from the G^k defined in [Hörmander 1994, Section 24.3] in that our $\mathcal{G}^k$ does not include higher-order glancing as well.)

With an open set $U \subset {}^bT^*M$ fixed, we will also use the notation

$$\widetilde{\mathcal{G}}^k := \{\rho \in U : \rho \text{ is connected to } \mathcal{G}^k \text{ by a generalized bicharacteristic}\}.$$

We further let

$$\widetilde{\mathcal{G}}^{\geq k} := \bigcup_{k' \geq k} \widetilde{\mathcal{G}}^{k'}.$$

Near glancing: We now commence the study of the measure μ near the glancing set $\mathcal{G}$. Here we follow an inductive strategy employed in [Burq and Lebeau 2001].

Lemma 4.6. *There exists a positive measure μ^∂ on the glancing set $\mathcal{G}$ so that*

$$\mu = \mu 1_{x_1 > 0} + \delta(x_1) \otimes \delta(H_p x_1) \otimes \mu^\partial.$$

In particular, $\mu(\mathcal{H}) = 0$.

Proof. Let $a_\epsilon = \varphi(x, \xi')\epsilon\chi(\epsilon^{-1}x_1)$, with $\chi \in \mathcal{C}_c^\infty(\mathbb{R})$, $\chi(0) = \chi'(0) = 1$.

By Lemma 4.4,

$$\mu(H_p a_\epsilon) = 2 \operatorname{Im} \mu^j(a_\epsilon).$$

By the dominated convergence theorem,

$$2 \operatorname{Im} \mu^j(a_\epsilon) \to 0.$$

On the other hand,

$$H_p a_\epsilon = \chi'(\epsilon^{-1}x_1)\varphi(x, \xi')H_p x_1 + O(\epsilon).$$

So, by dominated convergence,

$$\mu(H_p a_\epsilon) \to \mu(1_{x_1=0} H_p x_1 \varphi).$$

In particular,

$$\mu 1_{x_1=0}(H_p x_1 \varphi) = 0$$

for any φ and hence $\mu 1_{x_1=0}$ is supported on $\mathcal{G}$. Since it is a positive measure, $\mu 1_{x_1=0}$ takes the form specified. □

Lemma 4.7. *We have*

$$-(H_p^2 x_1)\mu^\partial = 4\dot{\nu} 1_{\mathcal{G}}.$$

In particular, *since μ^∂ and $\dot{\nu}$ are positive measures*,

$$\operatorname{supp} \mu^\partial \cup \operatorname{supp} \dot{\nu} \subset \{H_p^2 x_1 \leq 0\};$$

hence

$$\mu(\mathcal{G}_d) = 0,$$

and, for $E \subset \mathcal{G}$,

$$\dot{\nu}(E) = \dot{\nu}(E \cap \{H_p^2 x_1 < 0\}).$$

Proof. Since $H_p x_1 = 2\xi_1$,

$$H_p(2a(x,\xi)\xi_1) = aH_p^2 x_1 + 2\xi_1 H_p a.$$

Let $a_\epsilon = \chi(\epsilon^{-1}x_1)\chi(\epsilon^{-1}r(x,\xi'))2a(x,\xi')\xi_1$, where $\chi \in \mathcal{C}_c^\infty(\mathbb{R})$ has $\chi(0) = 1,\ \chi'(0) = 0$. Then

$$\begin{aligned} H_p a_\epsilon &= \epsilon^{-1}\chi'(\epsilon^{-1}x_1)\chi(\epsilon^{-1}r(x,\xi'))2a\xi_1 H_p x_1 \\ &\quad + \epsilon^{-1}\chi(\epsilon^{-1}x_1)\chi'(\epsilon^{-1}r(x,\xi'))2a\xi_1 H_p r \\ &\quad + 2\xi_1\chi(\epsilon^{-1}x_1)\chi(\epsilon^{-1}r(x,\xi'))H_p a + a\chi(\epsilon^{-1}x_1)\chi(\epsilon^{-1}r(x,\xi'))H_p^2 x_1 \\ &= a\chi(\epsilon^{-1}x_1)\chi(\epsilon^{-1}r(x,\xi'))H_p^2 x_1 + O(1)(|\chi'(\epsilon^{-1}x_1)| + |\chi'(\epsilon^{-1}r(x,\xi'))| + \epsilon^{1/2}). \end{aligned}$$

Here we have used that on S^*M, we have $H_p r = -H_p\xi_1^2 = O(\xi_1)$. Then by dominated convergence,

$$\mu(H_p a_\epsilon) \to \tfrac{1}{2}\mu^\partial([H_p^2 x_1]a).$$

On the other hand, since $a_\epsilon = O(\sqrt{\epsilon})$ on S^*M,

$$\mu(H_p a_\epsilon) = 2\,\mathrm{Im}\,\mu^j(a_\epsilon) - \dot{\nu}(2a(0,x',\xi')\chi(\epsilon^{-1}r)) \to -\dot{\nu}(21_{\mathcal{G}}a).$$

So, $-H_p^2 x_1 \mu^\partial = 4\dot{\nu}1_{\mathcal{G}}$ as claimed. □

4B. ***Invariance of the measure*** μ**.**

The b-cotangent bundle, ${}^bT^*M$: In our discussion of the invariance of μ, it will be convenient to have some facts about the b-cotangent bundle, ${}^bT^*M$. We refer the reader to [Hörmander 1994, Section 18.3] (where the notation $\widetilde{T}^*M$ is used) and [Vasy 2008] for more details on the ${}^bT^*M$. Recall that ${}^bT^*M$ is the dual to the vector bundle of vector fields tangent to ∂M. In local coordinates with $\partial M = \{x_1 = 0\}$, the bundle of vector fields tangent to M is generated over $\mathcal{C}^\infty(M)$ by the vector fields

$$\{x_1\partial_{x_1}, \partial_{x_2}, \ldots, \partial_{x_n}\}.$$

Over $x_1 > 0$, ${}^bT^*M$ is canonically diffeomorphic to T^*M with

$$(x,\zeta,\xi') \mapsto (x,\zeta/x_1,\xi').$$

In fact, we can identify T^*M as a subset of ${}^bT^*M$ using the canonical projection map $\pi : T^*M \to {}^bT^*M$ given in local coordinates by

$$\pi : (x,\xi_1,\xi') \mapsto (x,x_1\xi_1,\xi').$$

Using this map, we can push forward the measure μ to a measure on ${}^bT^*M$. It is actually this pushforward $\pi_*\mu$ for which we will show invariance under the generalized bicharacteristic flow.

Define

$$H_p^G = H_p + \frac{H_p^2 x_1}{H_{x_1}^2 p} H_{x_1}.$$

That is, H_p^G is the gliding vector field. Then H_p^G is tangent to $\mathcal{G}$; see [Hörmander 1994, Section 24.3]. Let also $\varphi_t : {}^bT^*M \to {}^bT^*M$ denote the generalized bicharacteristic flow on ${}^bT^*M$.

Invariance: The aim of the rest of this section is to show that

$$\pi_*\mu(q \circ \varphi_t) - \pi_*\mu(q) = \int_0^t 2 \operatorname{Im} \pi_*\mu^j (q \circ \varphi_s)\, ds, \quad q \in \mathcal{C}_c^\infty({}^bT^*M). \tag{4-3}$$

The main lemma is as follows.

Lemma 4.8. *Let μ and μ_1 be measures on T^*M supported by S^*M with $\mu_1 \ll \mu$. The following are equivalent*:

(1) $$\pi_*\mu(q \circ \varphi_t) - \pi_*\mu(q) = \int_0^t \pi_*\mu_1(q \circ \varphi_s)\, ds, \quad q \in \mathcal{C}_c^\infty({}^bT^*M). \tag{4-4}$$

(2) *For $a = a(x_1, x', x_1\xi_1, \xi')$*

$$\mu(H_p a) = \mu_1(a) \tag{4-5}$$

and $\mu(\mathcal{G}_d \cup \mathcal{H}) = 0$.

Once we prove this, Lemmas 4.4 and 4.7 imply that μ satisfies condition (ii) with $\mu_1 = 2 \operatorname{Im} \mu^j$ and hence satisfies (4-3).

Proof of Lemma 4.8. Here we follow [Burq and Lebeau 2001, Section 3.3].

(i) implies (ii): Let $a \in \mathcal{C}_c^\infty(T^*M)$ with $a = a(x, x_1\xi_1, \xi')$. Then there exists $q \in \mathcal{C}_c^\infty({}^bT^*M)$ such that $a = \pi^* q$. Thus, $q \circ \varphi_s(\rho)$ is differentiable from the left and right for all s and

$$\partial_s^{\mp}(q \circ \varphi_s)(\rho) = H_p a(\pi_\pm^{-1}(\varphi_s(\rho))), \tag{4-6}$$

where $\pi_\pm^{-1}$ denotes the outward- and inward-pointing inverses of π. In particular, for s such that $\varphi_s(\rho) \notin \mathcal{H}$,

$$\partial_s(q \circ \varphi_s)(\rho) = H_p a(\varphi_s(\rho)).$$

Now, $\pi_*\mu(\mathcal{H}) = 0$ since $\pi_*\mu$ satisfies (4-4) and $\mathcal{H}$ is transverse to the flow φ_s. Therefore, since $\pi_*\mu$ satisfies (4-4), differentiating yields

$$\pi_*\mu(\partial_t q \circ \varphi_t|_{t=s}) = \pi_*\mu_1(q \circ \varphi_s)$$

and setting $s = 0$ gives

$$\mu(H_p a) = \mu_1(a).$$

Finally, since $\mathcal{G}_d$ is transverse to the flow, and $\pi_*\mu$ satisfies (i),

$$\pi_*\mu(1_{\mathcal{G}_d}) = \mu(1_{\mathcal{G}_d}) = 0.$$

This completes the proof of (ii).

(ii) implies (i):

Lemma 4.9. *For $b \in \mathcal{C}_c^\infty(T^*M)$,*

$$\mu(H_p b) = \mu_1(b) + \mu(1_{x_1=0}\partial_{x_1} r b_o|_{\xi_1^2 = r(x,\xi')}) + u^0(b_o|_{x_1=0,\xi_1^2=r(0,x',\xi')}),$$

where u^0 is a measure supported $\mathcal{H}$, $1_{x_1=0}\mu$ is supported by $\mathcal{G}\setminus\mathcal{G}_d$, and

$$b_o = \frac{b(x,\xi_1,\xi') - b(x,-\xi_1,\xi')}{2\xi_1}$$

is the ξ_1 odd part of b.

Proof. Write

$$b = b_e + \xi_1 b_o,$$

where

$$b_e = \frac{b(x,\xi_1,\xi') + b(x,-\xi_1,\xi')}{2}.$$

Then, $b_e = \tilde{b}_e(x,\xi_1^2,\xi')$, $b_0 = \tilde{b}_o(x,\xi_1^2,\xi')$, and so on S^*M,

$$b_e = \tilde{b}_e(x, r(x,\xi'),\xi') =: a_e(x,\xi'), \quad b_o = \tilde{b}_o(x,r(x,\xi'),\xi') =: a_o(x,\xi').$$

Since H_p is tangent to S^*M,

$$\mu(H_p b) = \mu(H_p(a_e + \xi_1 a_o)) = \mu(H_p \xi_1 a_o) + \mu_1(a_e).$$

Now, let

$$\chi \in \mathcal{C}_c^\infty(\mathbb{R}), \quad \chi \equiv 1 \text{ near } 0, \quad \chi' \le 0 \text{ on } x_1 \ge 0, \quad \chi_\epsilon(x_1) = \chi(\epsilon^{-1}x_1). \tag{4-7}$$

Then,

$$\mu(H_p(1-\chi_\epsilon)\xi_1 a_o)) = \mu_1((1-\chi_\epsilon)\xi_1 a_o) \to \mu_1(1_{x_1>0}\xi_1 a_o).$$

Now,

$$\mu(H_p(\chi_\epsilon \xi_1 a_o)) = \mu(\chi_\epsilon \xi_1 H_p a_o + \chi_\epsilon \partial_{x_1} r a_o + 2\epsilon^{-1}\chi_\epsilon' \xi_1^2 a_o) =: I_\epsilon + II_\epsilon + III_\epsilon.$$

By the dominated convergence theorem,

$$I_\epsilon \to \mu(1_{x_1=0}\xi_1 H_p a_o), \quad II_\epsilon \to \mu(1_{x_1=0}\partial_{x_1} r a_o).$$

Note that since $\mu(\mathcal{H}) = 0$ and $\xi_1 = 0$ on $\{x_1 = 0\}\setminus\mathcal{H}$, we have

$$I_\epsilon \to 0, \quad II_\epsilon \to \mu(1_{x_1=0}\partial_{x_1} r a_o). \tag{4-8}$$

Now, observe that

$$\mu_1((1-\chi_\epsilon)\xi_1 a_o) = \mu(H_p(1-\chi_\epsilon)\xi_1 a_o) = \mu((1-\chi_\epsilon)H_p\xi_1 a_o) - III_\epsilon.$$

So,

$$III_\epsilon = \mu((1-\chi_\epsilon)H_p\xi_1 a_o) - \mu_1((1-\chi_\epsilon)\xi_1 a_o)$$

and in particular

$$\lim_{\epsilon\to 0} III_\epsilon(a_o) = \mu(1_{x_1>0}H_p\xi_1 a_o) - \mu_1(1_{x_1>0}\xi_1 a_o). \tag{4-9}$$

On the other hand for $a_o \geq 0$,

$$III_\epsilon(a_o) \leq 0.$$

Therefore, III_ϵ is a family of measures with a weak limit. In particular, $III_\epsilon \to u^0$ for some measure u_0 supported in $x_1 = 0$. In fact, this also shows that

$$a_o \mapsto \mu(1_{x_1>0} H_p(\xi_1 a_o)) - \mu_1(\xi_1 a_o) =: u^0(a_o) \tag{4-10}$$

is a measure. We now check that u^0 is supported in $\mathcal{H}$. Note that once we show this the proof will be complete, since by (4-8) and (4-9) together with $\mu_1(\mathcal{H}) = 0$, we then have

$$\begin{aligned}\mu(H_p b) &= \mu_1(a_e) + \mu_1(\xi_1 a_o) + \mu(1_{x_1=0}\partial_{x_1} r a_o) + u^0(a_o)\\ &= \mu_1(b) + \mu(1_{x_1=0}\partial_{x_1} r a_o) + u^0(a_o)\end{aligned}$$

as desired.

Let χ, χ_ϵ be as in (4-7). Note that for each $\epsilon > 0$, $a_o \chi_\epsilon(x_1)\chi_{\sqrt{\epsilon}}(\xi_1)$ can be written as a smooth function of $(x_1, x', x_1\xi_1, \xi')$. Using (4-10), we now have the decomposition

$$u^0(a_o\chi_\epsilon(x_1)\chi_{\sqrt{\epsilon}}(\xi_1)) = u_1^0(\epsilon) + u_2^0(\epsilon) + u_3^0(\epsilon) + u_4^0(\epsilon),$$

where

$$\begin{aligned}u_1^0(\epsilon) &= \mu(1_{x_1>0}\chi_\epsilon(x_1)\chi_{\sqrt{\epsilon}}(\xi_1)\xi_1 H_p a_o),\\ u_2^0(\epsilon) &= \mu(1_{x_1>0}\epsilon^{-1}\chi_\epsilon'(x_1)\chi_{\sqrt{\epsilon}}(\xi_1) 2\xi_1^2 a_o),\\ u_3^0(\epsilon) &= \mu(1_{x_1>0}\chi_\epsilon(x_1)(\xi_1\epsilon^{-1/2}\chi'_{\sqrt{\epsilon}}(\xi_1) + \chi_{\sqrt{\epsilon}}(\xi_1))\partial_{x_1} r a_o),\\ u_4^0(\epsilon) &= -\mu_1(\xi_1 a_o \chi_\epsilon(x_1)\chi_{\sqrt{\epsilon}}(\xi_1)).\end{aligned}$$

Clearly $u_1^0(\epsilon) \to 0$ by dominated convergence. The same is true of $u_2^0(\epsilon)$: the function $\chi_\epsilon'(x_1)\chi_{\sqrt{\epsilon}}(\xi_1)2\xi_1^2 a_o$ is uniformly bounded as $\epsilon \downarrow 0$ since $\xi_1^2 \leq C\epsilon$ on $\operatorname{supp}\chi_{\sqrt{\epsilon}}(\xi_1)$. Next,

$$u_3^0(\epsilon) \to 0$$

as $\xi_1\epsilon^{-1/2}\chi'_{\sqrt{\epsilon}}(\xi_1) + \chi_{\sqrt{\epsilon}}(\xi_1)$ likewise remains uniformly bounded. Finally,

$$u_4^0(\epsilon) \to -\mu_1(1_{x_1=0}1_{\xi_1=0}\xi_1 a_o) = 0.$$

In particular, this implies by the dominated convergence theorem that

$$u^0(a_o 1_{x_1=\xi_1=0}) = 0$$

and so u^0 is supported by $\mathcal{H}$. □

Invariance away from glancing: We now prove the integral invariance statement (i) away from glancing.

Lemma 4.10. *Suppose $\tilde{a} = \tilde{a}(x, \sigma, \xi') \in \mathcal{C}_c^\infty$ and let $a = \tilde{a}(x, x_1\xi_1, \xi')$. Furthermore, assume that for $\rho \in \operatorname{supp} a$, on $[-T, 0]$ $\varphi_{-t}(\rho)$ intersects $x_1 = 0$ at most once and in a hyperbolic point. Then,*

$$\mu(a) - \mu(a \circ \varphi_{-T}) = \int_{-T}^0 \mu_1(a \circ \varphi_{-s})\, ds.$$

Proof. Let $-T(\rho)$ be the time that $\varphi_{-T(\rho)}(\rho) \in \mathcal{H}$. Then

$$\begin{aligned}\int_{-T}^{0} H_p(a\circ\varphi_s)(\rho)\,ds &= \int_{-T}^{-T(\rho)-0} H_p(a\circ\varphi_s)\,ds + \int_{-T(\rho)+0}^{0} H_p(a\circ\varphi_s)\,ds\\ &= a\circ\varphi_{-T(\rho)-0}(\rho) - a\circ\varphi_{-T}(\rho) + a - a\circ\varphi_{-T(\rho)+0}(\rho).\end{aligned}$$

Now, since at $x_1 = 0,\ a(0,x',\xi_1,\xi') = a(0,x',-\xi_1,\xi'),\ a\circ\varphi_{-T(\rho)-0} = a\circ\varphi_{-T(\rho)+0}$, we have

$$\int_{-T}^{0} H_p(a\circ\varphi_s)(\rho)\,ds = a(\rho) - a\circ\varphi_{-T}(\rho).$$

Now, by Lemma 4.9

$$\iint_{-T}^{0} H_p(a\circ\varphi_s)(\rho)\,ds\,d\mu(\rho) = (H_p^*\mu)\left(\int_{-T}^{0} a\circ\varphi_s\,ds\right) = \mu_1\left(\int_{-T}^{0} a\circ\varphi_s\,ds\right) + u^0\left[\left(\int_{-T}^{0} a\circ\varphi_s\,ds|_{\mathcal{H}}\right)_o\right].$$

Note that

$$\rho \mapsto \int_{-T}^{0} a\circ\varphi_s\,ds|_{\mathcal{H}}(\rho)$$

is even under $\xi_1 \mapsto -\xi_1$, since points $(x=0,\,y,\,\pm\xi,\,\eta)$ over the hyperbolic set are both mapped to the same point in T^*M° under the short-time flow; thus the u^0 term vanishes. Therefore,

$$\int_{-T}^{0}\int a\circ\varphi_{-s}(\rho)\,d\mu_1\,ds = \mu_1\left(\int_{-T}^{0} a\circ\varphi_s\,ds\right) = \int a - a\circ\varphi_{-T}\,d\mu. \qquad \square$$

Invariance on the glancing set: We now turn to analysis at the glancing set. Our strategy will be to prove (i) for all $t < t_0$ fixed and small. To do this, we may break up the support of q by a partition of unity and work locally, assuming that q is replaced by a function a supported in a neighborhood U of a point $\rho \in \mathcal{G}$; we will show that for each such U, if t_0 is taken to be small, (i) holds with q replaced by a.

We will do this using a "layer stripping" argument. First, we "strip away" the hyperbolic layer using the invariance established in Lemma 4.10. In particular, recall that $\widetilde{\mathcal{G}}$ denotes the set of generalized bicharacteristics in U which encounter $\mathcal{G}$. Then we have already seen that (i) holds on $\widetilde{\mathcal{G}}^c$. Since $\widetilde{\mathcal{G}}^c$ is open, $\pi_* 1_{\widetilde{\mathcal{G}}^c}\mu$ satisfies

$$\pi_* 1_{\widetilde{\mathcal{G}}^c}\mu(q\circ\varphi_t) - \pi_* 1_{\widetilde{\mathcal{G}}^c}\mu(q) = \int_0^t \pi_* 2\,\mathrm{Im}\, 1_{\widetilde{\mathcal{G}}^c}\mu_1(q\circ\varphi_s)\,ds, \quad q \in \mathcal{C}_c^\infty({}^bT^*M). \tag{4-11}$$

In particular, since (i) implies (ii), $1_{\widetilde{\mathcal{G}}^c}\mu$ satisfies

$$1_{\widetilde{\mathcal{G}}^c}\mu(H_p a) = 1_{\widetilde{\mathcal{G}}^c}\mu_1(a). \tag{4-12}$$

Subtracting this from (4-5) for μ, $\mu 1_{\widetilde{\mathcal{G}}}$ satisfies

$$\mu 1_{\widetilde{\mathcal{G}}}(H_p a) = 1_{\widetilde{\mathcal{G}}}\mu_1(a). \tag{4-13}$$

Thus we may turn our attention to studying the new measures $\mu 1_{\widetilde{\mathcal{G}}}$, $1_{\widetilde{\mathcal{G}}}\mu_1$. By a slight abuse of notation, *we assume henceforth that μ and μ_1 are supported on $\widetilde{\mathcal{G}}$.*

As we proceed through the proof, we will "strip away" the higher-order tangent layers. We start by showing that at points where the bicharacteristics of H_p are tangent to exactly order 2, the measure is invariant in the sense of (i). Once we have shown this, we can argue as we did to remove the hyperbolic set and replace the measure μ by $\mu 1_{\widetilde{\mathcal{G}}^{\geq 3}}$. We then prove invariance on $\widetilde{\mathcal{G}}^k$ by induction. As before, the fact that (i) implies (ii) allows us to assume that μ is supported on $\widetilde{\mathcal{G}}^{\geq k}$. Then, since the flow is transverse to $\mathcal{G}^k$ we may work with a measure supported on $\widetilde{\mathcal{G}}^k$ when showing invariance at order k. We will obtain this invariance by using the transversality of the flow to $\mathcal{G}^k$ to show that there is no jump in the measure across these points. In particular, this argument will show that μ accumulates no mass across $\mathcal{G}^k$ and thus that the invariance continues to hold.

The case of $\mathcal{G}_d$: We choose the neighborhood of U of $\rho \in \widetilde{\mathcal{G}}_d$ so small that $U \cap (\mathcal{H} \cup \mathcal{G} \setminus \mathcal{G}_d) = \varnothing$. Next, suppose that $a \in \mathcal{C}_c^\infty(U)$. Recall that

$$\mu(H_p a) = \mu_1(a) + \mu(1_{x_1=0} \partial_{x_1} r a_o|_{\xi_1^2 = r(x,\xi')}) + u^0(a_o),$$

with u^0 supported in $\mathcal{H} \cap \operatorname{supp} \mu$ and $\mu(\mathcal{G}_d) = 0$. Since $\operatorname{supp} \mu \subset \widetilde{\mathcal{G}}$ and $\operatorname{supp} a \subset U$, we have

$$\mu(H_p a) = \mu_1(a),$$

and hence (4-4) holds since the generalized bicharacteristics through $\mathcal{G}_d$ are bicharacteristics of H_p.

The case of $\mathcal{G}_g$: Let $\widetilde{\mathcal{G}}_g$ be the set of generalized bicharacteristics encountering $\mathcal{G}_g$ in U. Recall that we may assume μ and μ_1 are supported on $\widetilde{\mathcal{G}}_g$ and may further choose U small enough so that $U \cap \widetilde{\mathcal{G}}_g \subset \mathcal{G}_g$; hence we may in fact assume at this stage that μ and μ_1 are supported by $\mathcal{G}_g$.

Let $a \in \mathcal{C}_c^\infty(U)$. For $q = q(x_1, x', x_1\xi_1, \xi')$, we have $1_{\mathcal{G}} H_p^G q = 1_{\mathcal{G}} H_p q$. Moreover, since $H_p^G x_1 = H_p^G \xi_1 = 0$, letting $a_0 = a|_{\xi_1=0}$,

$$1_{\mathcal{G}} H_p^G a = 1_{\mathcal{G}} H_p^G a_0 = 1_{\mathcal{G}} H_p a_0. \tag{4-14}$$

Now, since on U, $\operatorname{supp} \mu \subset \mathcal{G}_g$,

$$\mu(H_p^G a) = \mu(1_{\mathcal{G}_g} H_p^G a) = \mu(1_{\mathcal{G}_g} H_p^G a_0) = \mu(H_p a_0) = \mu_1(a_0) = \mu_1(1_{\mathcal{G}_g} a_0) = \mu_1(a). \tag{4-15}$$

Since H_p^G generates the bicharacteristic flow in $\mathcal{G}_g$, this implies (4-4).

Remark 4.11. While it may seem at first that the argument used for $\mathcal{G}_g$ applies directly to higher-order tangencies, we observe that it does not. In fact, the equality $\mu(H_p q) = \mu_1(q)$ for $q = q(x, x_1\xi_1, \xi')$ a priori only implies that μ mass can either "stick" to the boundary as in $\mathcal{G}_g$ or instantly detach from the boundary as in $\mathcal{G}_d$. It does not rule out either case. The facts that, in $\mathcal{G}_g$, trajectories which detach from the boundary instantly leave $\overline{M}$, while in $\mathcal{G}_d$, $\mu(\mathcal{G}_d) = 0$, rule out detaching from and sticking to the boundary at $\mathcal{G}_g$ and $\mathcal{G}_d$ respectively. In order to determine whether mass at a point of higher-order tangency sticks or detaches, we will use the transversality of the flow to such points to show that mass can neither accumulate nor dissipate.

Higher-order tangencies: Recall that

$$\mathcal{G}^k := \{\rho \in \mathcal{G} : (H_p^j x_1)(\rho) = 0 \text{ for } 0 \leq j < k \text{ and } (H_p^k x_1)(\rho) \neq 0\}.$$

We have shown that (4-3) holds for $a \in \mathcal{C}_c^\infty(T^*M)$ with $\operatorname{supp} a \cap \mathcal{G}^k = \varnothing$ for $k \geq 3$. We now proceed by induction on k.

Let $r_0(x', \xi') = r(0, x', \xi')$ and $r_1(x,', \xi') = \partial_{x_1} r(0, x', \xi')$. Then note that by [Hörmander 1994, Lemma 24.3.1],

$$\mathcal{G}^k = \{x_1 = 0 = 1 - r(0, x', \xi'),\ H_{r_0}^j r_1 = 0,\ 0 \leq j < k-2,\ H_{r_0}^{k-2} r_1 \neq 0\}.$$

For each k, let $\widetilde{\mathcal{G}}^k$ denote the set of generalized bicharacteristics encountering $\mathcal{G}^k$ in a small fixed neighborhood of $\mathcal{G}^k$.

Suppose (4-4) holds for $q \in \mathcal{C}_c^\infty(T^*M)$ with $\operatorname{supp} q \cap \widetilde{\mathcal{G}}^k = \varnothing$ for $k \geq M-1$. We show that (4-4) holds for $a \in \mathcal{C}_c^\infty(T^*M)$ with $a \cap \widetilde{\mathcal{G}}^k = \varnothing$ for $k \geq M$.

Let $\rho \in \mathcal{G}^M$ and U a neighborhood of ρ so that $U \cap \mathcal{G}^k = \varnothing$ for $k > M$. As before, we may assume μ and μ_1 are supported on $\widetilde{\mathcal{G}}^{\geq M}$; thus by our choice of U, we may in fact assume without loss of generality that μ, μ_1 are supported on $\widetilde{\mathcal{G}}^M$. Let (x_2, ρ) be coordinates on $T^*(\partial M)$ so that $\partial_{x_2} = H_{r_0}$. Shrinking U if necessary, by the implicit function theorem there exists $\Theta(\rho)$ such that on U

$$\widetilde{\mathcal{G}}^M \cap \mathcal{G}^M = \{(x, \xi) \in \widetilde{\mathcal{G}}^M : x_2 = \Theta(\rho)\}.$$

Let

$$\widetilde{\mathcal{G}}_0 := \widetilde{\mathcal{G}}^M \cap \mathcal{G}^M,$$
$$\widetilde{\mathcal{G}}^+ := \{(x, \xi) \in \widetilde{\mathcal{G}}^M : x_2 > \Theta(\rho)\}, \quad \widetilde{\mathcal{G}}^- : \{(x, \xi) \in \widetilde{\mathcal{G}}^M : x_2 < \Theta(\rho)\}.$$

We write

$$\mu 1_{\widetilde{\mathcal{G}}^M} = \mu_- + \mu_+ + \Upsilon, \quad \mu_- = 1_{\widetilde{\mathcal{G}}^-}\mu, \quad \mu_+ = 1_{\widetilde{\mathcal{G}}^+}\mu, \quad \Upsilon = 1_{\widetilde{\mathcal{G}}_0}\mu.$$

Then, for $q = q(x_1, x', x_1\xi_1, \xi')$ supported in U,

$$\Upsilon(H_p q) = -\mu_-(H_p q) - \mu_+(H_p q) + \mu_1(q). \tag{4-16}$$

Let

$$\widetilde{H}(\rho) = \begin{cases} H_p(\rho), & \rho \notin \mathcal{G}_g, \\ H_p^G(\rho), & \rho \in \mathcal{G}_g; \end{cases}$$

this vector field generates the bicharacteristic flow on $\widetilde{\mathcal{G}}^M$, so by our inductive hypothesis,

$$\widetilde{H}^* \mu = \mu_1 \tag{4-17}$$

on $\widetilde{\mathcal{G}}^M \backslash \mathcal{G}^M$. Moreover, since on $\widetilde{\mathcal{G}}^0$ we have $H_p^M x_1 \neq 0$, the flow within U (perhaps after further shrinking that set) for forward and backward time is either in $\{x_1 > 0\}$ or lies entirely in $\mathcal{G}_g$ (with separate cases for forward/backward flow depending on the parity of M and the sign of $H_p^M x_1$). In either case, the flow stays away from $\mathcal{H}$, and hence is generated by $\widetilde{H}$ on $\widetilde{\mathcal{G}}^M \cap U$.

Now as in Lemma 3.10 and the surrounding discussion, given any continuous a defined on $\mathcal{G}_0$ and $\psi \in \mathcal{C}_c^\infty(\mathbb{R})$, we extend a to be constant on the flow, and set

$$\mu_a^\sharp(\psi) = \mu(\psi(s) a(\rho));$$

here we are using $\rho \in \mathcal{G}_0$ together with $s \in \mathbb{R}$ as local (nonsmooth) coordinates on $\widetilde{\mathcal{G}}^M$ defined by the flow, mapping $(-\delta, \delta) \times \widetilde{\mathcal{G}}_0$ homeomorphically to $\widetilde{\mathcal{G}}^M$ via $(s, \rho) \mapsto \exp(t\widetilde{H})(\rho)$. Just as in Lemma 3.10, the flow invariance established on $\widetilde{\mathcal{G}}^M \backslash \mathcal{G}_0$ implies that $\mu_a^\sharp$ is AC with respect to Lebesgue measure on $\mathbb{R} \setminus 0$; moreover the Radon–Nikodym derivative is itself bounded by a multiple of $\sup|a|$, and hence is a function of s with values in measures on $\mathcal{G}_0$, denoted by $G(s)$. Finally, let $\operatorname{tr} \dot{\mu}_\pm = G(0^\pm)$ denote the restrictions of these functions to $\mathcal{G}_0$ from left and right; these are well-defined since the relation (4-17) together with the arguments in Lemma 3.11 show that G is continuous on $s \le 0$ and on $s \ge 0$ separately but that there may be a jump at $s = 0$.

Moreover, denoting the function $\psi(s)a(\rho)$ by $\psi \otimes a$, for $0 \notin \operatorname{supp} \psi$, we have $\mu(\psi' \otimes a) = \mu_1(\psi \otimes a)$. So in particular,

$$\int [G(a)](s)\psi'(s)\, ds = \mu_1(\psi \otimes a).$$

Now, let $\chi \in \mathcal{C}_c^\infty(\mathbb{R})$ with $\chi \equiv 1$ on $[-1, 1]$ and $\operatorname{supp} \chi \subset (-2, 2)$ and $\chi_\epsilon(s) = \chi(\epsilon^{-1}s)$. Then,

$$\begin{aligned}\mu((1-\chi_\epsilon(s))\widetilde{H}(\psi \otimes a)) &= \mu(\widetilde{H}[(1-\chi_\epsilon)\psi] \otimes a) + \mu([\psi \otimes a]\widetilde{H}\chi_\epsilon)\\ &= \mu(\widetilde{H}[(1-\chi_\epsilon)\psi \otimes a]) + \mu(\epsilon^{-1}\psi\chi'(\epsilon^{-1}\cdot) \otimes a)\\ &= \mu_1((1-\chi_\epsilon)\psi \otimes a) + \mu(\epsilon^{-1}\psi\chi'(\epsilon^{-1}\cdot) \otimes a).\end{aligned}$$

Thus,

$$\begin{aligned}\mu(\epsilon^{-1}\psi\chi'(\epsilon^{-1}\cdot) \otimes a) &= \epsilon^{-1}\int_{\mathbb{R}\setminus 0} [G(a)](s)\chi'(\epsilon^{-1}s)\psi(s)\, ds\\ &= \psi(0)(G_a(0^-) - G_a(0^+)) - \int_{\mathbb{R}\setminus 0} G_a(s)\chi_\epsilon(s)\psi'(s)\, ds + \mu_1(\chi_\epsilon\psi \otimes a)\\ &\to \psi(0)(G_a(0^-) - G_a(0^+)) + \mu_1((\psi \otimes a)1_{\widetilde{\mathcal{G}}^0}).\end{aligned}$$

In particular,

$$\begin{aligned}\mu_+(\widetilde{H}\psi \otimes a) + \mu_-(\widetilde{H}\psi \otimes a) &= \psi(0)([G(a)](0^-) - [G(a)](0^+)) + \mu_1(\psi \otimes a)\\ &= \operatorname{tr}\dot{\mu}_+(\psi \otimes a) - \operatorname{tr}\dot{\mu}_-(\psi \otimes a) + \mu_1(\psi \otimes a). \qquad (4\text{-}18)\end{aligned}$$

Now we claim that we may also apply (4-16) to the function $q = \psi \otimes a$. We begin by approximating a with a smooth function on $\mathcal{G}_0$. Since $\widetilde{H}$ locally generates the flow and is at least Lipschitz across $\mathcal{G}_0$, we find that the flow on $\widetilde{\mathcal{G}}^M$ is at least $\mathcal{C}^1$. We note further that $\widetilde{\mathcal{G}}^M$ is without loss of generality disjoint from $\mathcal{H}$ (after appropriately shrinking U) since depending on the parity and sign of $H_p^M x_1$ along $\mathcal{G}_0$ the flow is either gliding or enters $x_1 > 0$ in each of $\widetilde{\mathcal{G}}^\pm$. In particular, $\pi(\widetilde{\mathcal{G}}^M) \subset {}^bT^*M$ is $\mathcal{C}^1$. Moreover,

$$(-\delta, \delta) \times \mathcal{G}_0 \to \pi(\widetilde{\mathcal{G}}^M), \quad (s, q) \mapsto \varphi_s(q),$$

are $\mathcal{C}^1$ coordinates on $\widetilde{\mathcal{G}}^M$. Hence $b = \pi^*\psi \otimes a$ is $\mathcal{C}^1$ on $\pi(\widetilde{\mathcal{G}}^M)$ and we may extend b to $\tilde{b}$, a $\mathcal{C}^1$ function on ${}^bT^*M$. Equivalently, setting $\tilde{q} = \pi^*\tilde{b}$, we have $\tilde{q} = \tilde{b}(x, x_1\xi_1, \xi')$ with $\tilde{b}$ a $\mathcal{C}^1$ function of its arguments and $\tilde{q}|_{\widetilde{\mathcal{G}}^M} = \psi \otimes a$.

Since $\tilde{b}$ is $\mathcal{C}^1$, we may approximate it in $\mathcal{C}^1$ by smooth functions $\tilde{b}_\epsilon$ and hence we obtain using the dominated convergence theorem that (4-16) holds for $\tilde{q}$. Finally, since $\tilde{q}|_{\widetilde{\mathcal{G}}^M} = \psi \otimes a$ and $\operatorname{supp} \mu \subset \widetilde{\mathcal{G}}^M$,

we obtain (4-16) for $q = \psi \otimes a$. Moreover, since $\mu_\pm$ are each supported in the interior or $\mathcal{G}_g$, we can replace $\widetilde{H}$ by H_p in (4-18).

Now (4-16) and (4-18) yield

$$\Upsilon(H_p\psi \otimes a) = -\operatorname{tr}(\dot{\mu}_+)(\psi \otimes a) + \operatorname{tr}(\dot{\mu}_-)(\psi \otimes a), \tag{4-19}$$

and hence

$$|\Upsilon(H_p\psi \otimes a)| \le C|\sup \psi \otimes a|.$$

Since Υ is supported at $\widetilde{\mathcal{G}}^0$, and the collection of functions $\psi \otimes a$ is dense, this implies that, for all q,

$$|\Upsilon(H_p q)| \le C|\sup q|.$$

Then, using that Υ is a measure supported at $\widetilde{\mathcal{G}}_0$ and H_p is transverse to $\widetilde{\mathcal{G}}_0$, this implies $\Upsilon = 0$. Finally, inserting $\Upsilon = 0$ into (4-19) yields

$$\operatorname{tr}(\dot{\mu}_-) = \operatorname{tr}(\dot{\mu}_+).$$

Moreover, since μ_1 is absolutely continuous with respect to μ, we have $\mu_1 1_{\widetilde{\mathcal{G}}^0} = 0$.

Let $\psi_{t_0}(s) = \psi(s + t_0)$. Then $(\psi \otimes a) \circ \varphi_t = \psi_t \otimes a$ and

$$\begin{aligned}
\mu(\psi_{t_0} \otimes a) - \mu(\psi \otimes a) &= \int_0^{t_0} \partial_t \int [G(a)](s)\psi(s+t)\,ds\,dt \\
&= \int_0^{t_0}\int_{-\infty}^{0} [G(a)](s)\psi'(s+t)\,ds + \int_0^{\infty} [G(a)](s)\psi'(s+t)\,dt \\
&= \int_0^{t_0}\int_{-\infty}^{\infty} \mu_1(\psi_t \otimes a)\,ds + \operatorname{tr}\dot{\mu}_-(a)\psi(t) - \operatorname{tr}\dot{\mu}_+(a)\psi(t)dt \\
&= \int_0^{t_0} \mu_1(\psi_t \otimes a)\,dt.
\end{aligned}$$

In particular, we have

$$\mu((\psi \otimes a) \circ \varphi_{t_0}) - \mu(\psi \otimes a) = \int_0^{t_0} \mu_1([\psi \otimes a] \circ \varphi_t)\,dt.$$

Since the collection of functions of the form $\psi \otimes a$ is dense in $\mathcal{C}^0({}^bT^*M)$, this completes the proof that μ is invariant on $\widetilde{\mathcal{G}}^M$.

As there are no infinite-order tangencies, $\mathcal{G} = \bigcup_M \mathcal{G}^M$ and this completes the proof of (4-4). □

Finally, to complete the proof of (1-3) we use Proposition 3.9 and the following lemma.

Lemma 4.12. *Suppose that μ satisfies* (4-3). *Let $\psi \in \mathcal{C}_c^1(\mathbb{R})$, Σ be a hypersurface transverse to the generalized bicharacteristic flow so that $\varphi_t : \mathbb{R} \times \Sigma \to {}^bT^*M$ is a homeomorphism onto its image, and $a \in \mathcal{C}_c^0(\Sigma)$. Define $f_{\psi,a}(\varphi_t(\rho)) = \psi(t)a(\rho)$. Then*

$$\pi_*\mu(f_{\partial_t\psi,a}) = 2\operatorname{Im}\mu^j(f_{\psi,a}).$$

Proof. We have, for $q \in \mathcal{C}_c^\infty({}^bT^*M)$,

$$\partial_t\pi_*\mu(q \circ \varphi_t)|_{t=0} = 2\operatorname{Im}\mu^j(q).$$

Therefore, it is enough to show that we can move the derivative inside $\pi_*\mu$. This follows from the dominated convergence theorem since $\pi_*\mu(\mathcal{H}) = 0$. Now approximation of $f_{\psi,a}$ by smooth functions gives the result. □

In order to obtain the required uniformity, we adjust the set $\mathcal{C}$ by fixing neighborhood U of Ω in which we assume that for $\{g_k\}_{k=1}^{\infty} \subset \mathcal{C}$ we not only have $\|g - g_{k_m}\|_{\mathcal{C}^{1,\gamma}} \to 0$ for some $\gamma > 0$ but also $\|g - g_{k_m}\|_{\mathcal{C}^{\infty}(U)} \to 0$. Arguing as in Section 3E then completes the proof of Theorem 2 when $\Omega \neq \varnothing$.

5. Lower bounds on manifolds without boundary

The idea behind our proof of the lower bounds in Theorem 1 is that near any segment of a geodesic γ, which is not trapped, P is *globally* microlocally equivalent to hD_{x_1} and hence it is enough to construct examples which saturate the $L^2_{\text{comp}} \to L^2_{\text{loc}}$ bounds for the solution operator for hD_{x_1} where we impose the condition that the solution be supported in $x_1 \geq 0$.

Let U be an open set. Assume there exists a null bicharacteristic curve γ such that

$$\{\pi(\gamma(s)) : s \in (0, L)\} \subset U, \quad \pi(\gamma(s)) \notin U \text{ for } s \notin (0, L). \tag{5-1}$$

For any interval I (open or closed), let

$$\Gamma_I \equiv \{\gamma(s) : s \in I\} \subset T^*M, \quad \Gamma_I^0 = \{x_1 \in I : \xi_1 = 0, x' = \xi' = 0\} \subset T^*\mathbb{R}^n.$$

Lemma 5.1. *There exist neighborhoods $V \subset T^*M$ of $\Gamma_{[0,L]}$, $U \subset T^*\mathbb{R}^n$ of $\Gamma^0_{[0,L]}$ and a symplectomorphism $\kappa : U \to V$ with*

$$\kappa : \Gamma^0_{[0,L]} \to \Gamma_{[0,L]}, \quad \kappa^* p = \xi_1.$$

Proof. Let $\rho_0 = \gamma(0)$. Then by Darboux's theorem, there exist neighborhoods $V_1 \subset T^*M$ of ρ_0, $U_1 \subset T^*\mathbb{R}^n$ of 0 and a symplectomorphism $\kappa_1 : U_1 \to V_1$ so that

$$\kappa_1^* p = \xi_1, \quad \kappa_1 : \Gamma^0_{[0,L]} \cap U_1 \to \Gamma_{[0,L]} \cap V_1.$$

We will extend κ_1 so that its image covers a neighborhood of $\Gamma_{[0,L]}$. For this, let $\Sigma := \kappa_1(\{x_1 = 0\} \cap U_1)$. Shrinking U_1 if necessary, we assume that there is $\epsilon > 0$ so that

$$(-\epsilon, L + \epsilon) \times \Sigma \to V \subset T^*M, \quad (t, q) \mapsto \varphi_t(q),$$

is a diffeomorphism onto its image, V. Then, let

$$x_i(\varphi_t(q)) = x_i(\kappa_1(q)), \quad 1 < i \leq n, \qquad \xi_i(\varphi_t(q)) = \xi_i(\kappa_i(q)), \quad 1 \leq i \leq n, \qquad x_1(\varphi_t(q)) = t.$$

In particular, we have

$$H_p x_i = 0, \quad 1 < i \leq n, \qquad H_p \xi_i = 0, \quad 1 \leq i \leq n, \qquad H_p x_1 = 1.$$

By Jacobi's identity, for $f, g \in \mathcal{C}^\infty(T^*M)$,

$$H_p\{g, f\} = \{g, H_p f\} - \{f, H_p g\}.$$

Therefore, since the corresponding identities hold at Σ,

$$\{x_i, x_j\} = 0, \quad \{\xi_i, \xi_j\} = 0, \quad \{\xi_i, x_j\} = \delta_{ij}.$$

In particular, this implies that $\kappa^{-1} : V \to T^*\mathbb{R}^n$

$$\kappa^{-1}(q) := (x(q), \xi(q))$$

is a symplectomorphism onto its image, U. Furthermore, since p is invariant under H_p, we have $\kappa^* p = \xi_1$, and hence also $\kappa(\Gamma^0_{[0,L]}) = \Gamma_{[0,L]}$. □

Proposition 5.2. *Let γ be a geodesic and $U \subset M$ satisfy* (5-1). *Then for any $\epsilon > 0$ there exists $\chi \in \mathcal{C}^\infty_c(U)$ and $f \in L^2$ with* supp $f \subset \{\chi = 1\}$ *and $u \in L^2_{\mathrm{loc}}$ with*

$$Pu = f$$

and

$$\mathrm{WF}_h\, u \subset \bigcup_{s \geq 0} [\exp(s H_p)\, \mathrm{WF}_h\, f],$$

and where

$$\|\chi u\|_{L^2} \geq \left(\frac{2L - \epsilon}{\pi h}\right) \|f\|_{L^2}.$$

Note that u is purely microlocally outgoing by the wavefront set statement, and hence is given by the outgoing resolvent applied to f (modulo $O(h^\infty)$).

Proof. We first let κ be a symplectomorphism from $T^*\mathbb{R}^n$ into T^*M with

$$\kappa : \Gamma^0_{[0,L]} \to \Gamma_{[0,L]}, \quad \kappa^* p = \xi_1,$$

where

$$\Gamma^0_I = \{x_1 \in I : \xi_1 = 0, x' = \xi' = 0\};$$

this exists by Lemma 5.1.

Now quantize κ to a microlocally unitary FIO $T : \mathcal{C}^\infty_c(\mathbb{R}^n) \to \mathcal{C}^\infty_c(M)$ with parametrix S such that $ST - I$, $TS - I$ have no microsupport near $\Gamma^0_{[0,L]}$ and $\Gamma^0_{[0,L]}$ respectively and such that

$$ThD_1 = PT + O(h^\infty) \text{ microlocally near } \Gamma^0_{[0,L]}.$$

(There are no obstructions to such a quantization as long as we work on a contractible neighborhood of Γ^0.) Assume without loss of generality that T is defined on a δ_0-neighborhood of $\Gamma^0_{[0,L]}$.

Now for any $\delta < \delta_0$ we define on $\mathbb{R}^n$

$$f_0 = \begin{cases} \varphi(x') \cos \pi(x_1 - \delta)/2(L - 2\delta), & x_1 \in [\delta, L - \delta], \\ 0, & \text{otherwise}, \end{cases}$$

where φ is smooth and compactly supported in $B(0, \delta)$, with

$$\int |\varphi(x')|^2 \, dx' = 1.$$

Let

$$v_0 = \begin{cases} 0, & x_1 < \delta, \\ ih^{-1}(2(L-2\delta)/\pi)\varphi(x')\sin\pi(x_1-\delta)/2(L-2\delta), & x_1 \in [\delta, L-\delta], \\ ih^{-1}(2(L-2\delta)/\pi)\varphi(x'), & x_1 > L-\delta. \end{cases}$$

Thus of course $hD_1 v_0 = f_0$ as distributions. Let $\psi(x_1)$ be a smooth cutoff function supported in $(0, L-\delta/4)$ and equal to 1 on $[\delta, L-\delta/2]$. Set $u_0 = \psi(x_1)v_0$, $w = Tu_0$, and $f = Tf_0$ Then modulo $O(h^\infty)$,

$$Pw \equiv PTu_0 \equiv ThD_{x_1}u_0 \equiv Tf_0 + Tr_0 \equiv f + r,$$

where r_0 is the error term coming from the $hD_1(\psi)$ term. Note that

$$\mathrm{WF}\, r_0 \subset \{x_1 \in [L-\delta/2, L-\delta/4] : |x'| < \delta,\ \xi = 0\},$$

which is thus in an arbitrarily small neighborhood of the endpoint of $\Gamma^0_{[0,L]}$. Thus by the construction of T, $r \equiv Tr_0$ has wavefront set in an arbitrarily small neighborhood of $\gamma(L)$ (as $\delta \downarrow 0$).

Note also that we may construct a cutoff function $\chi \in \mathcal{C}^\infty_c(U)$ such that

$$\chi = \begin{cases} 1 & \text{on } \mathrm{WF}_h\, f, \\ 0 & \text{on } \pi\{\exp(sH_p)(\mathrm{WF}_h\, r) : s \geq 0\}. \end{cases}$$

This follows from our hypotheses since on $\mathbb{R}^n$ f is supported on $x_1 \in [\delta, L-\delta]$ in a small neighborhood of Γ^0, while

$$\mathrm{WF}_h\, r_0 \subset x_1 \in [L-\delta/2, L-\delta/4],$$

and the forward flowout is thus contained (in these local coordinates) in $\{x_1 \geq L-\delta/2\}$. Since the Hamilton flow is nonradial and x_1 maps to the flow parameter under κ, we know that $\mathrm{WF}_h\, r_0$ is separated from a neighborhood of $\Gamma_{[0,L-\delta]}$ in *base* variables, not just in the cotangent bundle, and shrinking δ if necessary we obtain the separation in the base variables. Note for use below that since $\chi = 1$ on $\mathrm{WF}_h\, f$, pulling back by T yields a fortiori $\kappa^*(\chi) = 1$ on $\mathrm{WF}_h\, f_0$ and hence

$$\kappa^*(\chi) = 1 \quad \text{on } \{x_1 \in [\delta, L-\delta] : |x'| < \delta,\ \xi = 0\}. \tag{5-2}$$

Finally we set

$$u = w - (P - i0)^{-1} r$$

so that

$$Pu = f.$$

Owing to the propagation of singularities for the outgoing resolvent, $\mathrm{WF}_h(u-w)$ lies in the *forward flowout* of $\mathrm{WF}_h\, r$; by construction this $\mathrm{WF}_h\, r$ is disjoint from the support of χ and by our geometric hypotheses, its forward flowout *remains* disjoint from U: here we use the fact that $\pi\gamma(L+s) \notin U$ for $s > 0$ and indeed this point escapes to infinity; hence $\pi\exp(sH_p)(\gamma(L)) \notin \mathrm{WF}_h\, f$ for all $s > 0$. By continuity, the same is true with $\gamma(L)$ replaced by any point in $\mathrm{WF}_h\, r$ for δ sufficiently small. Thus,

$$\chi u = \chi w + O(h^\infty).$$

By microlocal unitarity of T we compute by the Egorov theorem

$$\begin{aligned}\|\chi u\|^2 &= \|\chi w\|^2 + O(h^\infty) = \|S\chi T S w\|^2 + O(h^\infty) = \|\mathrm{Op}(\kappa^*\chi)u_0\|^2 + O(h)\\ &\geq \int_\delta^{1-\delta}\int_\delta^{L-\delta} |v_0|^2\, dx'\, dx_1 + O(h)\\ &\geq \int_\delta^{1-\delta}\int_\delta^{L-\delta} h^{-2}\frac{4(L-2\delta)^2}{\pi^2}|\varphi(x')|^2 \sin^2\frac{\pi(x_1-\delta)}{2(L-2\delta)}\, dx'\, dx_1 + O(h)\\ &= \frac{\pi}{4}\frac{1}{h^2\mu^3} + O(h),\end{aligned}$$

where

$$\mu = \frac{\pi}{2(L-2\delta)};$$

here we have used the fact that $\kappa^*\chi$ equals 1 on the zero section over $[\delta, L-\delta]\times B(0,\delta)$, while the semiclassical wavefront set of u_0 lies within the zero section; hence the last inequality follows by existence of a microlocal parametrix for $\mathrm{Op}(\kappa^*\chi)$ on this set.

Meanwhile,

$$\|f\|^2 = \int_\delta^{1-\delta}\int_\delta^{L-\delta} |\varphi(x')|^2 \cos^2\frac{\pi(x_1-\delta)}{2(L-2\delta)}\, dx'\, dx_1 = \frac{\pi}{4}\frac{1}{\mu}.$$

Thus

$$\frac{\|\chi u\|}{\|f\|} \geq \frac{1}{h\mu} = \frac{2(L-2\delta)}{\pi h}.$$ □

Finally, we apply Proposition 5.2 in the case of a nontrapping manifold with Euclidean ends and $P = -h^2\Delta_g - 1$. Let $R_1 < R' < R''$ so that $B(0,R'')\subset\{\chi\equiv 1\}$. By the definition of $L(g,R'')$, there is a null-bicharacteristic (a speed-2 geodesic lifted to S^*M) γ with $\pi\gamma(0)\in\partial B(0,R'')$, $\pi\gamma(s)\in B(0,R'')$ for $s\in(0,L(g,R''))$ and $\gamma(s)\notin B(0,R'')$ for $s > L(g,R'')$. In particular, since $L(g,R') < L(g,R'')$, Proposition 5.2 applies to γ with $U = B(0,R'')$ and $\epsilon < 2L(g,R'') - 2L(g,R')$. In particular, there exist $f\in L^2(M)$ with $\mathrm{supp}\, f\subset B(0,R'')$ so that

$$Pu = hf, \quad \mathrm{WF}_h\, u \subset \bigcup_{s\geq 0}[\exp(sH_p)\,\mathrm{WF}_h\, f] \tag{5-3}$$

and

$$\|u\|_{L^2(B(0,R''))} \geq \frac{2L(g,R'') - 2L(g,R'') + 2L(g,R')}{\pi}\|f\|_{L^2} = \frac{2L(g,R')}{\pi}\|f\|_{L^2}.$$

Next, observe that (5-3) implies

$$u = (-h^2\Delta_g - 1 - i0)^{-1}hf + O(h^\infty\|f\|_{L^2})_{L^2_{\mathrm{loc}}}.$$

In particular, letting $k = h^{-1}$,

$$\|\chi(-\Delta_g - k^2 - i0)^{-1}f\|_{L^2} \geq \left(\frac{2L(g,R')}{\pi k} - C_N k^{-N}\right)\|f\|_{L^2}$$

completing the proof of the lower bound in Theorem 1.

6. Application to numerical analysis of the finite-element method

6A. ***Summary.*** In this section, we focus on the implications the bound (1-8) has on the numerical analysis of solving the Helmholtz exterior Dirichlet problem by the finite-element method (FEM). We mention two other numerical-analysis applications of (1-8) in Remarks 6.8 and 6.9 at the end.

We consider the *h-version* of the FEM; i.e., the solution is approximated in spaces of piecewise polynomials of fixed degree on a mesh with mesh width h_{FEM} (we use this notation to avoid a notational clash with the semiclassical parameter $h := k^{-1}$ in the rest of the paper). The question of how fast h_{FEM} must decrease with k to maintain accuracy as $k \to \infty$ was thoroughly investigated in the case of the *constant-coefficient* Helmholtz equation (i.e., (6-2) below with $A = I$ and $\nu = 1$) by [Ihlenburg and Babuška 1995; 1997] when $d = 1$ and by [Melenk 1995; Melenk and Sauter 2010; 2011; Esterhazy and Melenk 2012] when $d = 2, 3$. For example, in the case of piecewise-linear polynomials, and when the solution of the boundary value problem is nontrapping, the FEM satisfies a *quasioptimal* error estimate (of the form (6-22) and (6-29) below) when

$$h_{\mathrm{FEM}} k^2 \le c \tag{6-1}$$

for a sufficiently small constant c (the case when the boundary value problem is trapping is more complicated; see [Chandler-Wilde et al. 2017, Section 1.4] for some initial results).

In this section, we use the bound (1-8) to prove the analogue of the result above in the case of the *variable-coefficient* Helmholtz equation ((6-2) below) posed in the exterior of a nontrapping Dirichlet obstacle (see Theorem 3 below). In particular, we show how the constant c in (6-1) depends on the constant in (1-8). The key point is that our bound on h_{FEM} shows explicitly how the constant c in (6-1) *decreases* (i.e., the requirement on h_{FEM} for quasioptimality becomes more stringent) as the length of the longest ray *increases*.

6B. ***The Helmholtz exterior Dirichlet problem and FEM set-up.***

Motivation from applications. Several important physical applications involve the PDE

$$\nabla \cdot (A \nabla u) + k^2 \nu u = -f \tag{6-2}$$

posed in $\mathbb{R}^n \setminus \Omega$, where A is a symmetric positive-definite matrix-valued function of position and ν is a strictly positive scalar-valued function of position. One example is via reductions of the time-harmonic Maxwell equations

$$\operatorname{curl} H + ik\epsilon E = (ik)^{-1} J, \quad \operatorname{curl} E - ik\mu H = 0 \quad \text{in } \mathbb{R}^3, \tag{6-3}$$

when all the fields involved depend only on two Cartesian space variables, say x and y. In the transverse-magnetic (TM) mode, J and E are given by $J = (0, 0, J_z(x, y))$ and $E = (0, 0, E_z(x, y))$ so, when additionally the permittivity ϵ is a scalar and the permeability μ satisfies

$$\mu = \begin{pmatrix} \tilde{\mu} & 0 \\ 0 & 1 \end{pmatrix} \tag{6-4}$$

for $\tilde{\mu}$ a 2×2 symmetric positive-definite matrix, E_z satisfies (6-2) with $\nu = \epsilon$,

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}^T (\tilde{\mu})^{-1} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{6-5}$$

and $f = J_z$. Similarly, in the case of the transverse-electric (TE) mode, J and H are given by $J = (J_x(x, y), J_y(x, y), 0)$ and $H = (0, 0, H_z(x, y))$, so that when μ is a scalar and ϵ satisfies an equation analogous to (6-4), H_z satisfies (6-2) with A given by (6-5) with $\tilde{\mu}$ replaced by $\tilde{\epsilon}$, $\nu = \mu$, and

$$f = -\frac{1}{ik}\nabla \cdot \left[\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} (\tilde{\epsilon})^{-1} \begin{pmatrix} J_x \\ J_y \end{pmatrix} \right].$$

Additionally, the so-called acoustic-approximation of the elastodynamic wave equation is (6-2) with A the inverse of the (scalar) density and ν the inverse of the (scalar) bulk modulus; see, e.g., the derivation in [Chaumont-Frelet 2015, Section 1.2.6].

Placing the PDE (6-2) *in the framework of Section 1A.* We let $M := \mathbb{R}^n \setminus \Omega$, where Ω is a compact set such that its complement is connected,

$$g^{ij} := \frac{A_{ij}}{\nu}, \quad L_j := \sum_{i=1}^{n} \frac{g^{ij}}{\sqrt{\det g}} \partial_i(\sqrt{\det g}) + \sum_{i=1}^{n} \partial_i(g^{ij}) - \sum_{i=1}^{n} \frac{1}{\nu} \partial_i(A_{ij}), \quad L := 0,$$

and

$$P(g)\varphi := -\frac{1}{\nu} \sum_{i,j=1}^{n} \partial_i(A_{ij}\partial_j\varphi),$$

and then observe that (1-6) is satisfied, since

$$\Delta_g \varphi := \sum_{i,j=1}^{n} \frac{1}{\sqrt{\det g}} \partial_i(\sqrt{\det g}\, g^{ij}\, \partial_j\varphi).$$

With these definitions, the PDE (6-2) is $(P(g) - k^2)u = f/\nu$. The operator P_Ω is then self-adjoint with respect to $L^2(\mathbb{R}^n \setminus \Omega; \nu)$ and the H^1 norm defined by (1-7) becomes

$$\|\varphi\|^2_{H^1_\nu(\mathbb{R}^n\setminus\Omega)} := \int_{\mathbb{R}^n\setminus\Omega} ((A\nabla\varphi) \cdot \overline{\nabla\varphi} + k^2\nu|\varphi|^2)\, dx. \tag{6-6}$$

In this section (*Section 6*) *only*, to make contact with the standard numerical-analysis literature, *we use nonsemiclassically scaled Sobolev spaces*; i.e., the norms on the spaces H^s do *not* include powers of k unless otherwise indicated. Indeed, in this section we write (6-6) as the weighted H^1 norm

$$\|\varphi\|^2_{H^1_{k,A,\nu}(\mathbb{R}^n\setminus\Omega)} := \|A^{1/2}\nabla\varphi\|^2_{L^2(\mathbb{R}^n\setminus\Omega)} + k^2\|\nu^{1/2}\varphi\|^2_{L^2(\mathbb{R}^n\setminus\Omega)}, \tag{6-7}$$

and we use analogous notation for norms over subsets of $\mathbb{R}^n \setminus \Omega$. We also define the norms

$$\|\varphi\|^2_{H^m(\mathbb{R}^n\setminus\Omega)} := \sum_{|\alpha|\le m} \int_{\mathbb{R}^n\setminus\Omega} |\partial^\alpha\varphi|^2\, dx;$$

i.e., if there are no A, ν in the subscript, the H^m norm is the "standard" H^m norm on $\mathbb{R}^n$.

By (1-1), $L(\nu A^{-1}, \Omega, R)$ is the length of the longest generalized bicharacteristic (informally, ray) in $B(0, R) \setminus \Omega$. Recall that, in the case $\Omega = \varnothing$, the (generalized) bicharacteristics are the projections in x of

the solutions $(x(s), \xi(s)) \in \mathbb{R}^n \times \mathbb{R}^n$ of the Hamiltonian system

$$\frac{dx_i}{ds}(s) = \frac{\partial}{\partial \xi_i} H(x(s), \xi(s)), \quad \frac{d\xi_i}{ds}(s) = -\frac{\partial}{\partial x_i} H(x(s), \xi(s)), \tag{6-8}$$

where the Hamiltonian $H(x, \xi)$ is given by

$$H(x, \xi) := \frac{1}{\nu(x)} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}(x) \xi_i \xi_j - 1. \tag{6-9}$$

Exterior Dirichlet problem and its variational formulation. With Ω a compact set such that its complement is connected, we define $\Omega_+ := \mathbb{R}^n \setminus \Omega$. Since the vast majority of numerical-analysis applications of the Helmholtz equation are in two and three dimensions, we restrict attention to $d = 2, 3$. Let $\gamma : H^1_{\mathrm{loc}}(\Omega) \to H^{1/2}(\partial\Omega)$ be the trace operator. Let A be a symmetric positive-definite matrix-valued function of position such that $\mathrm{supp}(I - A)$ is a compact subset of $\mathbb{R}^n$. Let ν be a strictly positive scalar-valued function of position such that $\mathrm{supp}(1 - \nu)$ is a compact subset of $\mathbb{R}^n$. Let $A_{\min}$ and $A_{\max}$ be such that

$$A_{\min} \le A(x) \le A_{\max} \quad \text{for all } x \in \Omega_+, \text{ in the sense of quadratic forms,} \tag{6-10}$$

and $\nu_{\min}$ and $\nu_{\max}$ be such that

$$\nu_{\min} \le \nu(x) \le \nu_{\max} \quad \text{for all } x \in \Omega_+. \tag{6-11}$$

We consider solving *both* the exterior Dirichlet problem

$$\nabla \cdot (A \nabla u) + k^2 \nu u = -f \quad \text{in } \Omega_+, \tag{6-12a}$$

$$\gamma u = 0 \quad \text{on } \partial\Omega, \tag{6-12b}$$

$$\frac{\partial u}{\partial r}(x) - iku(x) = o\left(\frac{1}{r^{(d-1)/2}}\right) \quad \text{as } r := |x| \to \infty, \text{ uniformly in } \hat{x} := x/r, \tag{6-12c}$$

and the sound-soft scattering problem

$$\nabla \cdot (A \nabla u) + k^2 \nu u = 0 \quad \text{in } \Omega_+, \tag{6-13a}$$

$$\gamma u = 0 \quad \text{on } \partial\Omega, \tag{6-13b}$$

$$(u - u^I) \text{ satisfies the radiation condition (6-12c),} \tag{6-13c}$$

where u^I is solution of $\Delta u^I + k^2 u^I = 0$ (such as a plane wave or point source) that is smooth in a neighborhood of $\mathrm{supp}(I - A) \cup \mathrm{supp}(1 - \nu) \cup \Omega$.

The standard variational formulations of these problems are posed on $\Omega_R := \Omega_+ \cap B(0, R) = B(0, R) \setminus \Omega$, where R is chosen large enough such that $\mathrm{supp}(I - A)$, $\mathrm{supp}(1 - \nu)$, and $\mathrm{supp}\ f$ are all compactly contained in $B(0, R)$, and also such that u^I is smooth in $B(0, R)$. Let

$$H^1_{0,D}(\Omega_R) := \{v \in H^1(\Omega_R) : \gamma v = 0 \text{ on } \partial\Omega\}, \tag{6-14}$$

let $\Gamma_R := \partial B(0, R)$, and let $T_R : H^{1/2}(\Gamma_R) \to H^{-1/2}(\Gamma_R)$ be the Dirichlet-to-Neumann (DtN) map for the equation $\Delta u + k^2 u = 0$ posed in the exterior of $B(0, R)$ with the Sommerfeld radiation condition

(6-12c); the definition of T_R in terms of Hankel functions and polar coordinates (when $d = 2$) and spherical polar coordinates (when $d = 3$) is given in, e.g., [Chandler-Wilde and Monk 2008, Equations 3.5 and 3.6; Nédélec 2001, Section 2.6.3; Melenk and Sauter 2010, Equations 3.7 and 3.10]. The variational formulation of (6-12) is

$$\text{find } u \in H^1_{0,D}(\Omega_R) \text{ such that } a(u, v) = F(v) \text{ for all } v \in H^1_{0,D}(\Omega_R), \tag{6-15}$$

where

$$a(u, v) := \int_{\Omega_R} ((A\nabla u) \cdot \overline{\nabla v} - k^2 \nu u \bar{v})\, dx - \langle T_R(\gamma u), \gamma v\rangle_{\Gamma_R} \quad \text{and} \quad F(v) := \int_{\Omega_R} f\, \bar{v}\, dx, \tag{6-16}$$

and where $\langle \cdot, \cdot \rangle_{\Gamma_R}$ denotes the duality pairing on Γ_R that is linear in the first argument and antilinear in the second. The variational formulation of (6-13) is (6-15) but with $F(v)$ instead given by

$$F(v) := \int_{\Gamma_R} \left(\frac{\partial u^I}{\partial r} - T_R u^I \right) \gamma \bar{v}\, ds.$$

Finite-element method. Let $\mathcal{T}_{h_{\mathrm{FEM}}}$ be a family of triangulations of Ω_R (in the sense of, e.g., [Ciarlet 1991, page 67]) that is shape regular (see, e.g., [Brenner and Scott 2008, Definition 4.4.13; Ciarlet 1991, page 128]). Let

$$\mathcal{H}_{h_{\mathrm{FEM}}} := \{v \in \mathcal{C}(\overline{\Omega}_R) : v|_K \text{ is a polynomial of degree 1 for each } K \in \mathcal{T}_{h_{\mathrm{FEM}}}\},$$

and observe that the dimension of $\mathcal{H}_{h_{\mathrm{FEM}}}$ is proportional to ${h_{\mathrm{FEM}}}^{-n}$. Our main result, Theorem 3 below, is valid when Ω is $\mathcal{C}^\infty$ (or, more precisely, $\mathcal{C}^m$ for some large m not given explicitly). For such Ω it is not possible to fit $\partial\Omega$ exactly with simplicial elements (i.e., when each element of $\mathcal{T}_{h_{\mathrm{FEM}}}$ is a simplex), and fitting $\partial\Omega$ with isoparametric elements (see, e.g, [Ciarlet 1991, Chapter VI]) or curved elements (see, e.g., [Bernardi 1989]) is impractical, and therefore some analysis of nonconforming error is necessary; since this is very standard (see, e.g., [Brenner and Scott 2008, Chapter 10]), we ignore this issue here.

The finite-element method for the variational problem (6-15) is the Galerkin method applied to the variational problem (6-15), i.e.,

$$\text{find } u_{h_{\mathrm{FEM}}} \in \mathcal{H}_{h_{\mathrm{FEM}}} \text{ such that } a(u_{h_{\mathrm{FEM}}}, v_{h_{\mathrm{FEM}}}) = F(v_{h_{\mathrm{FEM}}}) \text{ for all } v_{h_{\mathrm{FEM}}} \in \mathcal{H}_{h_{\mathrm{FEM}}}. \tag{6-17}$$

6C. ***Main result.*** Before stating the main result (Theorem 3 below) we define the following constants — all independent of k but dependent on one or more of A, ν, Ω, R, and k_0 — upon which the main result depends.

C_{int}: Recall that the *nodal interpolant* $I_{h_{\mathrm{FEM}}} : \mathcal{C}(\overline{\Omega}_R) \to \mathcal{H}_{h_{\mathrm{FEM}}}$ is well-defined for functions in $H^2(\Omega_R)$ (for $d = 2, 3$) and satisfies

$$\|\nu^{1/2}(v - I_{h_{\mathrm{FEM}}} v)\|_{L^2(\Omega_R)} + h_{\mathrm{FEM}} \|A^{1/2} \nabla (v - I_{h_{\mathrm{FEM}}} v)\|_{L^2(\Omega_R)} \le C_{\mathrm{int}} (h_{\mathrm{FEM}})^2 \|v\|_{H^2(\Omega_R)} \tag{6-18}$$

for all $v \in H^2(\Omega_R)$, for some constant $C_{\mathrm{int}} = C_{\mathrm{int}}(A, \nu)$. The only reason C_{int} depends on A and ν is because of the weights A and ν in the norms on the left-hand side of (6-18). Indeed, by, e.g., [Brenner

and Scott 2008, Equation 4.4.28],

$$\|v - I_{h_{\mathrm{FEM}}} v\|_{L^2(\Omega_R)} + h_{\mathrm{FEM}} \|\nabla(v - I_{h_{\mathrm{FEM}}} v)\|_{L^2(\Omega_R)} \le \widetilde{C}_{\mathrm{int}} (h_{\mathrm{FEM}})^2 \|v\|_{H^2(\Omega_R)}$$

for all $v \in H^2(\Omega_R)$, for some $\widetilde{C}_{\mathrm{int}}$ that depends only on the shape-regularity constant of the mesh, and thus (6-18) holds with

$$C_{\mathrm{int}}(A, \nu) := \widetilde{C}_{\mathrm{int}} \max\{(A_{\max})^{1/2}, (\nu_{\max})^{1/2}\}.$$

C_{DtN}: There exists $C_{\mathrm{DtN}} = C_{\mathrm{DtN}}(A, \nu, R, k_0)$ such that

$$|\langle T_R(\gamma u), \gamma v\rangle_{\Gamma_R}\rangle| \le C_{\mathrm{DtN}} \|u\|_{H^1_{k,A,\nu}(\Omega_R)} \|v\|_{H^1_{k,A,\nu}(\Omega_R)} \quad \text{for all } u, v \in H^1(\Omega_R) \text{ and for all } k \ge k_0; \tag{6-19}$$

see [Melenk and Sauter 2010, Lemma 3.3]. As above, the only reason C_{DtN} depends on A and ν is because of the weights A and ν in (6-7). Indeed, [Melenk and Sauter 2010, Lemma 3.3] bounds the left-hand side of (6-19) by $H^{1/2}(\Gamma_R)$ and $L^2(\Gamma_R)$ norms of γu and γv, and then uses trace theorems to prove that

$$|\langle T_R(\gamma u), \gamma v\rangle_{\Gamma_R}\rangle| \le \widetilde{C}_{\mathrm{DtN}}(R, k_0) \|u\|_{H^1_{k,I,1}(\Omega_R)} \|v\|_{H^1_{k,I,1}(\Omega_R)},$$

for some $\widetilde{C}_{\mathrm{DtN}}(R, k_0)$ given explicitly in the proof of [Melenk and Sauter 2010, Lemma 3.3]. Therefore, (6-19) holds with

$$C_{\mathrm{DtN}}(A, \nu, R, k_0) := \widetilde{C}_{\mathrm{DtN}}(R, k_0) \max\left\{ \frac{1}{(A_{\min})^{1/2}}, \frac{1}{(\nu_{\min})^{1/2}} \right\}.$$

C_{H^2}: There exists $C_{H^2} = C_{H^2}(A, \Omega, R)$ such that, if $\tilde{f} \in L^2(\Omega_{R+1})$ and $v \in H^1(\Omega_{R+1})$ satisfy

$$\nabla \cdot (A \nabla v) = \tilde{f} \quad \text{in } \Omega_{R+1} \text{ and } \gamma v = 0 \text{ on } \partial\Omega,$$

then

$$\|v\|_{H^2(\Omega_R)} \le C_{H^2}(\|A^{1/2} \nabla v\|_{L^2(\Omega_{R+1})} + \|v\|_{L^2(\Omega_{R+1})} + \|\tilde{f}\|_{L^2(\Omega_{R+1})}); \tag{6-20}$$

since $A \in \mathcal{C}^{0,1}$ and $\Omega \in \mathcal{C}^{1,1}$, such a C_{H^2} exists by, e.g., [McLean 2000, Theorem 4.18], and C_{H^2} depends on $A_{\min}$ in (6-10), the $\mathcal{C}^{0,1}$ norm of A, and the $\mathcal{C}^{1,1}$ norm of the parametrization of $\partial\Omega$. Note that the $R + 1$ in the norms on the right-hand side of (6-20) can be replaced by $R + b$ for any $b > 0$, but the constant C_{H^2} blows up as $b \to 0$.

Theorem 3 (quasioptimality of the Galerkin method). *Let Ω, A, and ν be as in Section 6B; i.e., Ω is a compact set such that its complement Ω_+ is connected, A is a symmetric positive-definite matrix-valued function of position such that* $\mathrm{supp}(I - A)$ *is a compact subset of $B(0, R)$ and ν a strictly positive scalar-valued function of position such that* $\mathrm{supp}(1 - \nu)$ *is a compact subset of $B(0, R)$. Furthermore, let Ω be $\mathcal{C}^\infty$, and let $A, \nu \in \mathcal{C}^{1,1}(\Omega_R)$ be such that A and ν are $\mathcal{C}^\infty$ in a neighborhood of $\partial\Omega$. Finally assume that Ω is nowhere tangent to infinite order to the geodesic flow generated by A and ν via the Hamiltonian* (6-9), *and that this flow on Ω_+ is nontrapping.*

There exists a $k_0 > 0$ such that if

$$1 \geq h_{\mathrm{FEM}} k^2 \sqrt{1+(h_{\mathrm{FEM}}k)^2}\, L(\nu A^{-1}, \Omega, R+2)\, C_{\mathrm{int}} C_{H^2}(1+C_{\mathrm{DtN}}) \left(\frac{\nu_{\max}}{\nu_{\min}}\right)^{1/2} \frac{4\sqrt{2}}{\pi} \times \left((\nu_{\max})^{1/2} + \frac{1+(\nu_{\min})^{1/2}}{k_0} + \frac{1}{k_0^2(\nu_{\min})^{1/2}}\right), \quad (6\text{-}21)$$

then the Galerkin equations (6-17) *have a unique solution which satisfies*

$$\|u - u_{h_{\mathrm{FEM}}}\|_{H^1_{k,A,\nu}(\Omega_R)} \leq 2(1+C_{\mathrm{DtN}}) \Big(\min_{v_{h_{\mathrm{FEM}}} \in \mathcal{H}_{h_{\mathrm{FEM}}}} \|u - v_{h_{\mathrm{FEM}}}\|_{H^1_{k,A,\nu}(\Omega_R)} \Big). \quad (6\text{-}22)$$

(*Note that k_0 depends on A, ν, and Ω but — as in Theorem 2 — is uniform on small $\mathcal{C}^{2,\alpha}$ neighborhoods of A and ν; see Section 3E.*)

Remark 6.1 (the mesh threshold (6-21)). If one assumes that $h_{\mathrm{FEM}}k \leq C$, then the mesh threshold (6-21) can be written in the form (6-1), i.e., $h_{\mathrm{FEM}}k^2 \leq c$. The key point is that if A, ν, and Ω are as in the statement of the theorem with the $\mathcal{C}^{0,1}$ norms of A and ν and the $\mathcal{C}^{1,1}$ norm of Ω all bounded by $\widetilde{C}$, say, then all the constants in c apart from $L(\nu A^{-1}, \Omega, R+2)$ (namely, $A_{\max}$, C_{int}, C_{H^2}, C_{DtN}) are bounded in terms of $\widetilde{C}$, $A_{\min}$ and $\nu_{\min}$, but $L(\nu A^{-1}, \Omega, R+2)$ can be arbitrarily large. Furthermore, c *decreases* as $L(\nu A^{-1}, \Omega, R+2)$ *increases*; i.e., the condition on the mesh threshold for quasioptimality becomes more restrictive as the length of the longest ray grows.

Remark 6.2 (quasioptimality for higher-order finite-element spaces). We have only considered the lowest-order conforming finite-element space of H^1, namely continuous piecewise-linear polynomials. In the case of the constant-coefficient Helmholtz equation, the mesh threshold for quasioptimality (analogous to (6-1)) has been determined by [Melenk and Sauter 2010; 2011; Esterhazy and Melenk 2012] for spaces of arbitrary order p (possibly depending on k), with the threshold then explicit in k, h, and p. These results are obtained by a careful splitting of the solution that does not immediately generalize to the case $A \neq I$. However, in the case of p fixed, the mesh thresholds for quasioptimality from [Melenk and Sauter 2010; 2011; Esterhazy and Melenk 2012] have recently been obtained in [Chaumont-Frelet and Nicaise 2019] using a simpler method (relying on well-known elliptic regularity results such as (6-20)), although the constants are not given explicitly in p. In the case when the DtN map is approximated by an impedance boundary condition, the results of [Chaumont-Frelet and Nicaise 2019] immediately apply to the case $A \neq I$, and they can in principle be extended to the full scattering problem considered here.

Remark 6.3 (comparison to existing results in the literature). The only existing results in the literature on quasioptimality (explicit in all parameters and coefficients) for the Galerkin method applied to the variable-coefficient Helmholtz equation are in [Chaumont-Frelet 2016; Graham and Sauter 2020]. The main similarity between these two works and the present paper is that they all use the "Schatz argument" described in Lemma 6.6 below (and pioneered for Helmholtz problems in [Sauter 2006; Banjai and Sauter 2007]). The main differences between [Chaumont-Frelet 2016; Graham and Sauter 2020] and the present paper are that (i) both [Chaumont-Frelet 2016] and [Graham and Sauter 2020] consider the Helmholtz equation in a bounded domain ([Chaumont-Frelet 2016] in one dimension, [Graham and Sauter

2020] in one, two, or three dimensions) with an impedance boundary condition (recall that this is the simplest-possible approximation to the DtN map operator T_R), and (ii) to get an a priori bound explicit in k and the coefficients, both [Chaumont-Frelet 2016] and [Graham and Sauter 2020] impose conditions on the coefficients and the domain that are *stronger* than the analogue of nontrapping for the interior impedance problem (i.e., the assumption that every ray reaches the boundary in a uniform time).

6D. ***Proof of Theorem 3.*** The heart of the proof is Lemma 6.6 below. This gives a condition for quasioptimality to hold in terms of how well the solution of the adjoint problem is approximated by the finite-element space, and relies on the fact that $a(\cdot,\cdot)$ satisfies a Gårding inequality. This argument essentially goes back to [Schatz 1974] (using the Aubin–Nitsche technique; see, e.g., the references in [Spence 2015, Remark 26]) and was extensively used in the analysis of the FEM for Helmholtz problems by [Sauter 2006; Melenk and Sauter 2010; 2011; Esterhazy and Melenk 2012].

Before stating Lemma 6.6 we need to introduce some notation. Let $C_{\mathrm{cont}} = C_{\mathrm{cont}}(A, \nu, R, k_0)$ be the *continuity constant* of the sesquilinear form $a(\cdot,\cdot)$ (defined in (6-16)) in the norm $\|\cdot\|_{H^1_{k,A,\nu}(\Omega_R)}$; i.e.,

$$a(u, v) \le C_{\mathrm{cont}} \|u\|_{H^1_{k,A,\nu}(\Omega_R)} \|v\|_{H^1_{k,A,\nu}(\Omega_R)} \quad \text{for all } u, v \in H^1_{0,D}(\Omega_R).$$

By the Cauchy–Schwarz inequality and (6-19) we have

$$C_{\mathrm{cont}} \le 1 + C_{\mathrm{DtN}}. \tag{6-23}$$

Definition 6.4 (the adjoint sesquilinear form $a^*(\cdot,\cdot)$). The adjoint sesquilinear form, $a^*(u, v)$, to the sesquilinear form $a(\cdot,\cdot)$ defined in (6-16) is given by

$$a^*(u, v) := \overline{a(v, u)} = \int_{\Omega_R} ((A\nabla u)\cdot\overline{\nabla v} - k^2 \nu u \bar{v}) - \langle \gamma u, T_R(\gamma v)\rangle_{\Gamma_R}. \tag{6-24}$$

Lemma 6.5. *Given $F \in (H^1_{0,D}(\Omega_R))'$, if u is the solution to the variational problem*

$$a^*(u, v) = F(v) \quad \textit{for all } v \in H^1_{0,D}(\Omega_R), \tag{6-25}$$

then $\bar{u}$ satisfies

$$a(\bar{u}, v) = \overline{F(\bar{v})} \quad \textit{for all } v \in H^1_{0,D}(\Omega_R). \tag{6-26}$$

Proof of Lemma 6.5. By (6-25),

$$\overline{a^*(u, \bar{v})} = \overline{F(\bar{v})} \quad \text{for all } v \in H^1_{0,D}(\Omega_R).$$

The result then follows from the definition of $a^*(\cdot,\cdot)$ and the following property of the DtN map T_R:

$$\langle T_R\psi, \bar{\varphi}\rangle_\Gamma = \langle T_R\varphi, \bar{\psi}\rangle_\Gamma \quad \text{for all } \varphi, \psi \in H^{1/2}(\Gamma_R).$$

This property follows from the fact that, if u and v are solutions of the homogeneous Helmholtz equation $\Delta u + k^2 u = 0$ in $\mathbb{R}^n \setminus \overline{B(0, R)}$, both satisfying the Sommerfeld radiation condition (6-12c), then

$$\int_{\Gamma_R} (\gamma u)\, \frac{\partial v}{\partial n} = \int_{\Gamma_R} (\gamma v)\, \frac{\partial u}{\partial n}$$

by Green's second identity; see, e.g., [Spence 2015, Lemma 6.13]. □

Lemma 6.6 (conditions for quasioptimality). *Given $f \in L^2(\Omega_R)$, let $S^* f$ be the solution of the adjoint problem* (6-25) *with $F(v)$ defined in* (6-16)*. Let*

$$\eta(\mathcal{H}_{h_{\mathrm{FEM}}}) := \sup_{0\neq f\in L^2(\Omega_R)} \min_{v_{h_{\mathrm{FEM}}}\in\mathcal{H}_{h_{\mathrm{FEM}}}} \frac{\|S^* f - v_{h_{\mathrm{FEM}}}\|_{H^1_{k,A,\nu}(\Omega_R)}}{\|f\|_{L^2(\Omega_R)}}. \tag{6-27}$$

If

$$\eta(\mathcal{H}_{h_{\mathrm{FEM}}}) \le \frac{1}{2C_{\mathrm{cont}}(\nu_{\max})^{1/2}k} \tag{6-28}$$

then the Galerkin equations (6-17) *have a unique solution which satisfies*

$$\|u - u_{h_{\mathrm{FEM}}}\|_{H^1_{k,A,\nu}(\Omega_R)} \le 2C_{\mathrm{cont}}\Big(\min_{v_{h_{\mathrm{FEM}}}\in\mathcal{H}_{h_{\mathrm{FEM}}}} \|u - v_{h_{\mathrm{FEM}}}\|_{H^1_{k,A,\nu}(\Omega_R)}\Big). \tag{6-29}$$

Proof. Since

$$\mathrm{Re}(-\langle T_R\varphi, \varphi\rangle_{\Gamma_R}) \ge 0 \quad \text{for all } \varphi \in H^{1/2}(\Gamma_R) \text{ and for all } k \ge k_0 \tag{6-30}$$

(see [Chandler-Wilde and Monk 2008, Corollary 3.1] or [Nédélec 2001, Theorem 2.6.4]), $a(\cdot,\cdot)$ satisfies the Gårding inequality

$$\mathrm{Re}(a(v, v)) \ge \|v\|^2_{H^1_{k,A,\nu}(\Omega_R)} - 2k^2\nu_{\max}\|v\|^2_{L^2(\Omega_R)}$$

and the result follows from the account of the Schatz argument in, e.g., [Spence 2015, Theorem 6.32]. □

The condition (6-28) on $\eta(\mathcal{H}_{h_{\mathrm{FEM}}})$ is implicitly a condition on the mesh width h_{FEM}. To make this condition explicit, we observe that the polynomial-approximation bound (6-18) implies that a bound on $\eta(\mathcal{H}_{h_{\mathrm{FEM}}})$ (defined by (6-27)) can be obtained from a bound on the H^2 norm of the (adjoint of the) exterior Dirichlet problem.

Lemma 6.7 (bound on H^2 norm). *If Ω, A, and ν are as in Theorem 3, then there exists a $k_0 > 0$ such that the solution of the exterior Dirichlet problem with f, $I - A$, and $1 - \nu$ supported in Ω_R satisfies*

$$\|u\|_{H^2(\Omega_R)}$$
$$\le k\, C_{H^2}\frac{2\sqrt{2}}{\pi(\nu_{\min})^{1/2}}L(\nu A^{-1}, \Omega, R+2)\left((\nu_{\max})^{1/2} + \frac{1+(\nu_{\min})^{1/2}}{k_0} + \frac{1}{k_0^2(\nu_{\min})^{1/2}}\right)\|f\|_{L^2(\Omega_R)}. \tag{6-31}$$

Proof. Choosing χ such that $\mathrm{supp}\,\chi \subset B_{R+2}$ but $\chi = 1$ on B_{R+1}, Theorem 2 with $s = 1$ implies that there exists $k_0 > 0$ such that

$$\|A^{1/2}\nabla u\|^2_{L^2(\Omega_{R+1})} + k^2\|\nu^{1/2}v\|^2_{L^2(\Omega_{R+1})} \le \left(\frac{2\sqrt{2}}{\pi}L(\nu A^{-1}, \Omega, R+2)\right)^2\frac{1}{\nu_{\min}}\|f\|^2_{L^2(\Omega_+)} \tag{6-32}$$

for all $k \ge k_0$. Since the χ in our argument can be chosen independent of A and ν, the constant k_0 can be chosen uniformly as A and ν vary within a sufficiently small open neighborhood in $\mathcal{C}^{2,\alpha}$ ($\alpha > 0$), just as in Theorem 2.

We then apply (6-20) with $v = u$, $\tilde{f} = -k^2\nu u - f$, and recall that $\mathrm{supp}\, f \subset \Omega_R$, to obtain

$$\|u\|_{H^2(\Omega_R)} \le C_{H^2}\left(\frac{2\sqrt{2}}{\pi(\nu_{\min})^{1/2}}L(\nu A^{-1}, \Omega, R+2)\left(k(\nu_{\max})^{1/2} + 1 + \frac{1}{k(\nu_{\min})^{1/2}}\right) + 1\right)\|f\|_{L^2(\Omega_R)}.$$

The result (6-31) then follows by noting that $L(\nu A^{-1}, \Omega, R+2) \geq 2$; this last inequality holds since the geodesic needs to at least travel from the boundary of the ball of radius $R+2$ into the ball of radius R and out again (a distance of at least 4), and in this annular region the metric is Euclidean and the flow has speed 2. □

We can now prove Theorem 3.

Proof of Theorem 3. Using the polynomial approximation result (6-18) and the definitions of $\eta(\mathcal{H}_{h_{\rm FEM}})$ (6-27) and $\|\cdot\|_{H^1_{k,A,\nu}(\Omega_R)}$ (6-7), we have

$$\eta(\mathcal{H}_{h_{\rm FEM}}) \leq C_{\rm int} h_{\rm FEM}\sqrt{1+(h_{\rm FEM}k)^2} \sup_{0\neq f\in L^2(\Omega_R)} \frac{\|S^* f\|_{H^2(\Omega_R)}}{\|f\|_{L^2(\Omega_R)}}.$$

By Lemma 6.5, the solution of the adjoint exterior Dirichlet problem with data f is the complex conjugate of the solution of the exterior Dirichlet problem with data $\bar{f}$. Therefore, the bound (6-31) for the solution of the exterior Dirichlet problem also holds for the solution of the adjoint problem, and implies

$$\eta(\mathcal{H}_{h_{\rm FEM}}) \leq C_{\rm int} h_{\rm FEM}\sqrt{1+(h_{\rm FEM}k)^2}\; k\, C_{H^2} \frac{2\sqrt{2}}{\pi(\nu_{\rm min})^{1/2}}\, L(\nu A^{-1}, \Omega, R+2) \times \left((\nu_{\rm max})^{1/2} + \frac{1+(\nu_{\rm min})^{1/2}}{k_0} + \frac{1}{k_0^2(\nu_{\rm min})^{1/2}}\right);$$

the result then follows from using this bound on $\eta(\mathcal{H}_{h_{\rm FEM}})$ in Lemma 6.6. □

Remark 6.8 (how the bound (1-8) can be used in the analysis of preconditioning strategies). The Galerkin method (6-17) is equivalent to a linear system of equations; denote the matrix of this linear system by A. Linear systems involving A are difficult to solve because (a) the dimension of A is proportional to ${h_{\rm FEM}}^{-n}$ and (by Theorem 3) $h_{\rm FEM}$ must decrease like k^{-2} for the Galerkin solution to be quasioptimal, therefore A is large, and (b) since A is large and sparse (when using standard piecewise-polynomial bases of $\mathcal{H}_{h_{\rm FEM}}$), iterative methods are usually used to solve the linear system, but A is sign-indefinite when k is sufficiently large, and the efficient iterative solution of linear systems involving such matrices is difficult. One therefore seeks to *precondition* A; i.e., to find a B such that (i) B^{-1} approximates A^{-1} and (ii) the action of B^{-1} is cheap to compute, and one then applies the iterative solver to $\mathsf{B}^{-1}\mathsf{A}$.

A very popular and successful preconditioner for A is based on choosing B^{-1} to be a cheap approximation of $(\mathsf{A}_\alpha)^{-1}$, where A_α is the Galerkin matrix arising from the exterior Dirichlet problem with $k^2 \mapsto k^2 + i\alpha$, i.e., with artificial absorption α added. The rationale behind this method (introduced in [Erlangga et al. 2006]) is that the larger α is, the less oscillatory the problem is, and hence the easier it is to find a cheap approximation to $(\mathsf{A}_\alpha)^{-1}$. However, the larger α is, the further $(\mathsf{A}_\alpha)^{-1}$ is from A^{-1}; hence the question of what is the optimal α is nontrivial.

Although one of the advantages of preconditioning with absorption, compared to other Helmholtz-preconditioning strategies, is its easy applicability to variable-coefficient problems (see, e.g., [Oosterlee et al. 2010]), the only existing analysis is for constant-coefficient Helmholtz problems (i.e., (6-2) with $A = I$ and $\nu = 1$). As a component of determining the optimal α, [Gander et al. 2015, Theorem 1.4] proved that, when $A = I,\ \nu = 1,\ \Omega$ is star-shaped, and a quasioptimal error estimate (similar to (6-22))

holds for Galerkin discretizations of the problem with absorption, there exists $C > 0$ (independent of k and h_{FEM}, but dependent on Ω and $\mathcal{T}_{h_{\mathrm{FEM}}}$) such that

$$\|\mathsf{I} - (\mathsf{A}_\alpha)^{-1}\mathsf{A}\| \le C\alpha/k; \tag{6-33}$$

i.e., choosing α to be a sufficiently small multiple of k guarantees that $(\mathsf{A}_\alpha)^{-1}$ is a good approximation to A^{-1}, uniformly as $k \to \infty$. Using the bound (1-8) in the arguments of [Gander et al. 2015] one can show that the bound (6-33) holds when A, ν, and Ω satisfy the conditions of Theorem 3 with C a constant multiple of $L(\nu A^{-1}, \Omega, R+1)$; i.e., as the length of the longest ray increases, less absorption is allowed for $(\mathsf{A}_\alpha)^{-1}$ to be a good approximation to A^{-1}. This result is then consistent with the numerical experiments in [Gander et al. 2015, Tables 8 and 9]: the geometry corresponding to Table 9 supports longer rays than the geometry corresponding to Table 8, and the numerical results in the tables (the number of iterations required to solve $(\mathsf{A}_\alpha)^{-1}\mathsf{A}$ with GMRES) show that a lower amount of absorption is allowed in the former case than in the latter for $(\mathsf{A}_\alpha)^{-1}$ to be a good preconditioner for A.

Remark 6.9 (how the bound (1-8) can be used in "uncertainty quantification"). In the last 10 years there has been a surge of interest in "uncertainty quantification (UQ)" of PDEs, understood as theory and algorithms for computing statistics of quantities of interest involving PDEs *either* posed on a random domain *or* having random coefficients.

There is a large literature on UQ for the Poisson equation

$$-\nabla \cdot (A(\omega)\nabla u(\omega)) = f(\omega) \tag{6-34}$$

(where ω is an element of the underlying probability space), due, in part, to its large number of applications (e.g., in modeling groundwater flow). The fact that a priori bounds on the solution of (6-34) that are explicit in the coefficient A can easily be obtained is the starting point for the rigorous analysis of UQ algorithms; see, e.g., [Babuška et al. 2004; 2007; Gittelson 2010; Mugler and Starkloff 2011; Charrier 2012; Charrier et al. 2013]. For example, for (6-34) posed in a bounded Lipschitz domain D with homogeneous Dirichlet boundary conditions and $A \in L^\infty(D)$ with $A_{\min} > 0$ (in the sense of quadratic forms as in (6-10)), the Lax–Milgram theorem implies

$$\|u\|_{H^1(D)} \le \frac{C_D}{A_{\min}} \|f\|_{L^2(D)}, \tag{6-35}$$

where C_D is the constant appearing in the Poincaré inequality $\|v\|_{H^1(D)} \le C_D^{1/2} \|\nabla v\|_{L^2(D)}$ for all $v \in H_0^1(D)$. In contrast, there has been essentially no rigorous theory of UQ for the Helmholtz equation with large k because a priori bounds on the solution that are explicit in both k and the coefficients have not been available.

The recent paper [Pembery and Spence 2018] presented general measure-theory arguments that convert a bound on the (deterministic) Helmholtz equation that is explicit in both k and the coefficients into a bound, and associated well-posedness result, on the Helmholtz equation with random coefficients (and random data). The paper [Pembery and Spence 2018] used as input to these general arguments the deterministic bounds from [Graham et al. 2019] for star-shaped Lipschitz Ω and coefficients satisfying

radial monotonicity-like conditions which are stronger than nontrapping. We highlight that the bound (1-8) can be used with the arguments of [Pembery and Spence 2018] to obtain well-posedness results and a priori bounds on the Helmholtz equation (6-2) where the coefficients and domain are such that the problem is almost surely nontrapping.

Acknowledgements

Galkowski was supported by NSF Postdoctoral Research Fellowship DMS-1502661 and thanks Maciej Zworski for helpful conversations. Spence was supported by EPSRC grant EP/R005591/1. Wunsch was partially supported by NSF grant DMS–1600023. The authors are grateful to an anonymous referee for helpful comments on the manuscript.

References

[Babuška et al. 2004] I. Babuška, R. Tempone, and G. E. Zouraris, "Galerkin finite element approximations of stochastic elliptic partial differential equations", *SIAM J. Numer. Anal.* **42**:2 (2004), 800–825. MR Zbl

[Babuška et al. 2007] I. Babuška, F. Nobile, and R. Tempone, "A stochastic collocation method for elliptic partial differential equations with random input data", *SIAM J. Numer. Anal.* **45**:3 (2007), 1005–1034. MR Zbl

[Banjai and Sauter 2007] L. Banjai and S. Sauter, "A refined Galerkin error and stability analysis for highly indefinite variational problems", *SIAM J. Numer. Anal.* **45**:1 (2007), 37–53. MR Zbl

[Barucq et al. 2017] H. Barucq, T. Chaumont-Frelet, and C. Gout, "Stability analysis of heterogeneous Helmholtz problems and finite element solution based on propagation media approximation", *Math. Comp.* **86**:307 (2017), 2129–2157. MR Zbl

[Bernardi 1989] C. Bernardi, "Optimal finite-element interpolation on curved domains", *SIAM J. Numer. Anal.* **26**:5 (1989), 1212–1240. MR Zbl

[Brenner and Scott 2008] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, 3rd ed., Texts in Applied Mathematics **15**, Springer, 2008. MR Zbl

[Brown et al. 2017] D. L. Brown, D. Gallistl, and D. Peterseim, "Multiscale Petrov–Galerkin method for high-frequency heterogeneous Helmholtz equations", pp. 85–115 in *Meshfree methods for partial differential equations, VIII*, edited by M. Griebel and M. A. Schweitzer, Lect. Notes Comput. Sci. Eng. **115**, Springer, 2017. MR Zbl

[Burq 1997] N. Burq, "Mesures semi-classiques et mesures de défaut", exposé 826 , 167–195 in *Séminaire Bourbaki, Vol. 1996/97*, Astérisque **245**, 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 and Lebeau 2001] N. Burq and G. Lebeau, "Mesures de défaut de compacité, application au système de Lamé", *Ann. Sci. École Norm. Sup.* (4) **34**:6 (2001), 817–870. 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

[Chandler-Wilde and Monk 2008] S. N. Chandler-Wilde and P. Monk, "Wave-number-explicit bounds in time-harmonic scattering", *SIAM J. Math. Anal.* **39**:5 (2008), 1428–1455. MR Zbl

[Chandler-Wilde et al. 2017] S. N. Chandler-Wilde, E. A. Spence, A. Gibbs, and V. P. Smyshlyaev, "High-frequency bounds for the Helmholtz equation under parabolic trapping and applications in numerical analysis", preprint, 2017. arXiv

[Charrier 2012] J. Charrier, "Strong and weak error estimates for elliptic partial differential equations with random coefficients", *SIAM J. Numer. Anal.* **50**:1 (2012), 216–246. MR Zbl

[Charrier et al. 2013] J. Charrier, R. Scheichl, and A. L. Teckentrup, "Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods", *SIAM J. Numer. Anal.* **51**:1 (2013), 322–352. MR Zbl

[Chaumont-Frelet 2015] T. Chaumont-Frelet, *Approximation par éléments finis de problèmes d'Helmholtz pour la propagation d'ondes sismiques*, Ph.D. thesis, INSA Rouen, 2015, available at https://tel.archives-ouvertes.fr/tel-01246244/file/CHAUMONT.pdf.

[Chaumont-Frelet 2016] T. Chaumont-Frelet, "On high order methods for the heterogeneous Helmholtz equation", *Comput. Math. Appl.* **72**:9 (2016), 2203–2225. MR Zbl

[Chaumont-Frelet and Nicaise 2019] T. Chaumont-Frelet and S. Nicaise, "Wavenumber explicit convergence analysis for finite element discretizations of general wave propagation problems", *IMA J. Num. Anal.* (online publication May 2019).

[Ciarlet 1991] P. G. Ciarlet, "Basic error estimates for elliptic problems", pp. 17–351 in *Finite element methods, Part 1*, edited by P. G. Ciarlet and J.-L. Lions, Handbook of Numerical Analysis **2**, North-Holland, Amsterdam, 1991. MR Zbl

[Dyatlov and Zworski 2019] S. Dyatlov and M. Zworski, *Mathematical theory of scattering resonances*, Graduate Studies in Mathematics **200**, American Mathematical Society, Providence, RI, 2019. Zbl

[Erlangga et al. 2006] Y. A. Erlangga, C. W. Oosterlee, and C. Vuik, "A novel multigrid based preconditioner for heterogeneous Helmholtz problems", *SIAM J. Sci. Comput.* **27**:4 (2006), 1471–1492. MR Zbl

[Esterhazy and Melenk 2012] S. Esterhazy and J. M. Melenk, "On stability of discretizations of the Helmholtz equation", pp. 285–324 in *Numerical analysis of multiscale problems*, edited by I. G. Graham et al., Lect. Notes Comput. Sci. Eng. **83**, Springer, 2012. MR Zbl

[Folland 1999] G. B. Folland, *Real analysis: modern techniques and their applications*, 2nd ed., John Wiley & Sons, New York, 1999. MR Zbl

[Gander et al. 2015] M. J. Gander, I. G. Graham, and E. A. Spence, "Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed?", *Numer. Math.* **131**:3 (2015), 567–614. MR Zbl

[Gittelson 2010] C. J. Gittelson, "Stochastic Galerkin discretization of the log-normal isotropic diffusion problem", *Math. Models Methods Appl. Sci.* **20**:2 (2010), 237–263. MR Zbl

[Graham and Sauter 2020] I. G. Graham and S. A. Sauter, "Stability and finite element error analysis for the Helmholtz equation with variable coefficients", *Math. Comp.* **89**:321 (2020), 105–138. MR Zbl

[Graham et al. 2019] I. G. Graham, O. R. Pembery, and E. A. Spence, "The Helmholtz equation in heterogeneous media: a priori bounds, well-posedness, and resonances", *J. Differential Equations* **266**:6 (2019), 2869–2923. MR Zbl

[Halmos 1967] P. R. Halmos, *A Hilbert space problem book*, D. Van Nostrand Co., Princeton, N.J., 1967. MR Zbl

[Hörmander 1994] L. Hörmander, *The analysis of linear partial differential operators, III: Pseudo-differential operators*, Grundlehren der Mathematischen Wissenschaften **274**, Springer, 1994. MR

[Ihlenburg and Babuška 1995] F. Ihlenburg and I. Babuška, "Finite element solution of the Helmholtz equation with high wave number, I: The h-version of the FEM", *Comput. Math. Appl.* **30**:9 (1995), 9–37. MR Zbl

[Ihlenburg and Babuška 1997] F. Ihlenburg and I. Babuška, "Finite element solution of the Helmholtz equation with high wave number, II: The h-p version of the FEM", *SIAM J. Numer. Anal.* **34**:1 (1997), 315–358. MR Zbl

[McLean 2000] W. McLean, *Strongly elliptic systems and boundary integral equations*, Cambridge University Press, 2000. MR Zbl

[Melenk 1995] J. M. Melenk, *On generalized finite-element methods*, Ph.D. thesis, University of Maryland, 1995, available at https://search.proquest.com/docview/304214298. MR

[Melenk and Sauter 2010] J. M. Melenk and S. Sauter, "Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions", *Math. Comp.* **79**:272 (2010), 1871–1914. MR Zbl

[Melenk and Sauter 2011] J. M. Melenk and S. Sauter, "Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation", *SIAM J. Numer. Anal.* **49**:3 (2011), 1210–1243. 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

[Moiola and Spence 2019] A. Moiola and E. A. Spence, "Acoustic transmission problems: wavenumber-explicit bounds and resonance-free regions", *Math. Models Methods Appl. Sci.* **29**:2 (2019), 317–354. MR Zbl

[Mugler and Starkloff 2011] A. Mugler and H.-J. Starkloff, "On elliptic partial differential equations with random coefficients", *Stud. Univ. Babeş-Bolyai Math.* **56**:2 (2011), 473–487. MR

[Nédélec 2001] J. C. Nédélec, *Acoustic and electromagnetic equations: integral representations for harmonic problems*, Applied Mathematical Sciences **144**, Springer, 2001.

[Ohlberger and Verfürth 2018] M. Ohlberger and B. Verfürth, "A new heterogeneous multiscale method for the Helmholtz equation with high contrast", *Multiscale Model. Simul.* **16**:1 (2018), 385–411. MR

[Oosterlee et al. 2010] C. W. Oosterlee, C. Vuik, W. A. Mulder, and R.-E. Plessix, "Shifted-Laplacian preconditioners for heterogeneous Helmholtz problems", pp. 21–46 in *Advanced computational methods in science and engineering*, edited by B. Koren and C. Vuik, Lecture Notes in Computational Science and Engineering **71**, Springer, 2010. Zbl

[Pembery and Spence 2018] O. R. Pembery and E. A. Spence, "The Helmholtz equation in random media: well-posedness and a priori bounds", preprint, 2018. arXiv

[Sauter 2006] S. A. Sauter, "A refined finite element convergence theory for highly indefinite Helmholtz problems", *Computing* **78**:2 (2006), 101–115. MR Zbl

[Sauter and Torres 2018] S. Sauter and C. Torres, "Stability estimate for the Helmholtz equation with rapidly jumping coefficients", *Z. Angew. Math. Phys.* **69**:6 (2018), art. id. 139. MR Zbl

[Schatz 1974] A. H. Schatz, "An observation concerning Ritz–Galerkin methods with indefinite bilinear forms", *Math. Comp.* **28** (1974), 959–962. MR Zbl

[Spence 2015] E. A. Spence, "Overview of variational formulations for linear elliptic PDEs", pp. 93–159 in *Unified transform for boundary value problems*, edited by A. S. Fokas and B. Pelloni, SIAM, Philadelphia, PA, 2015. MR Zbl

[Taylor 2000] M. E. Taylor, *Tools for PDE: pseudodifferential operators, paradifferential operators, and layer potentials*, Mathematical Surveys and Monographs **81**, American Mathematical Society, Providence, RI, 2000. MR Zbl

[Taylor 2011] M. E. Taylor, *Partial differential equations, III: Nonlinear equations*, 2nd ed., Applied Mathematical Sciences **117**, Springer, 2011. MR Zbl

[Vasy 2008] A. Vasy, "Propagation of singularities for the wave equation on manifolds with corners", *Ann. of Math.* (2) **168**:3 (2008), 749–812. MR Zbl

Received 10 Jan 2019. Revised 10 Aug 2019. Accepted 26 Oct 2019.

JEFFREY GALKOWSKI: jeffrey.galkowski@northeastern.edu
Department of Mathematics, Northeastern University, Boston, MA, United States

EUAN A. SPENCE: e.a.spence@bath.ac.uk
Department of Mathematical Sciences, University of Bath, Bath, United Kingdom

JARED WUNSCH: jwunsch@math.northwestern.edu
Department of Mathematics, Northwestern University, Evanston, IL, United States

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. 1 2020

2578-5893 (2020)2:1;1-B