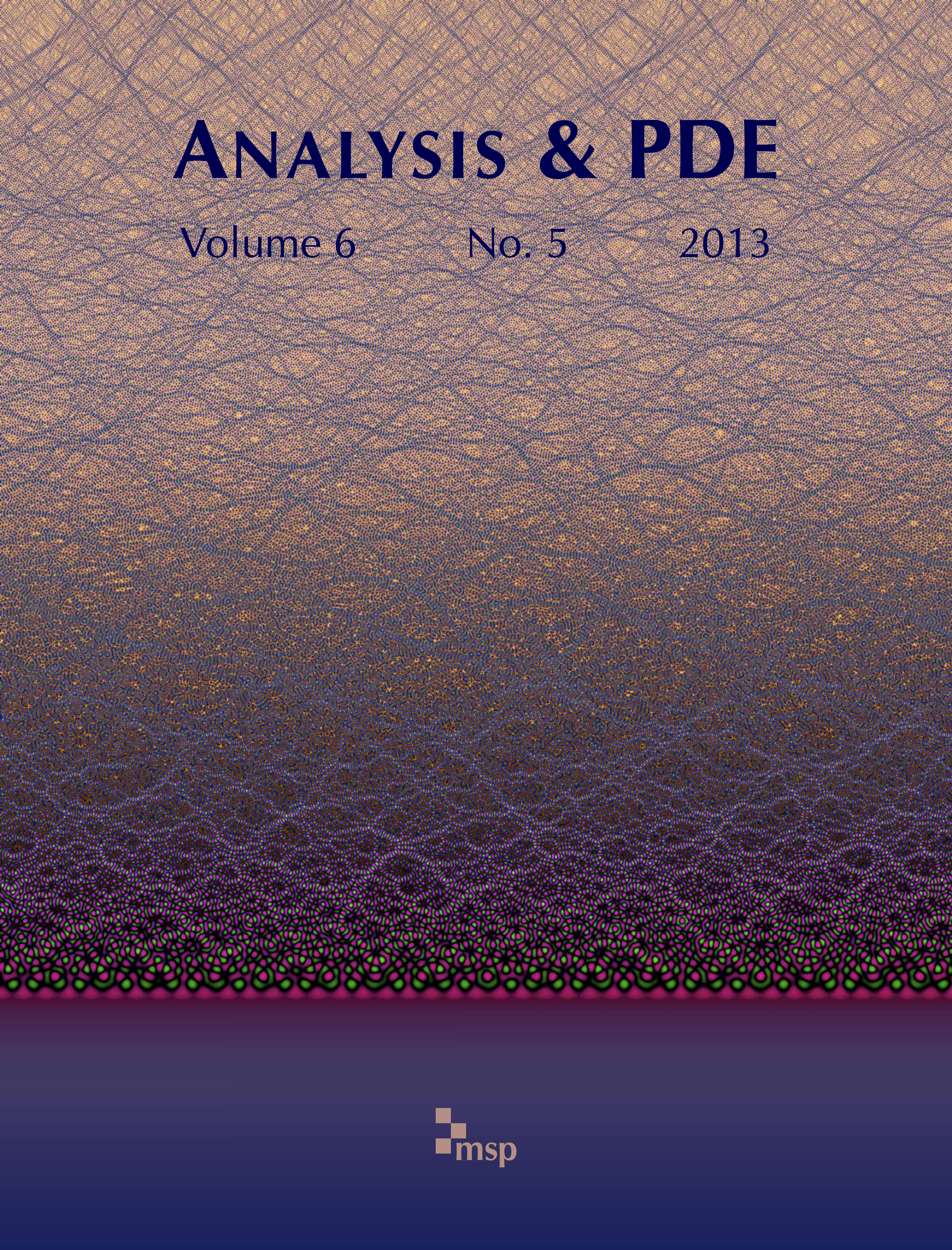

ANALYSIS & PDE

Volume 6 No. 5 2013

msp

Analysis & PDE

msp.org/apde

EDITORS

EDITOR-IN-CHIEF

Maciej Zworski
zworski@math.berkeley.edu
University of California
Berkeley, USA

BOARD OF EDITORS

Nicolas Burq	Université Paris-Sud 11, France nicolas.burq@math.u-psud.fr
Sun-Yung Alice Chang	Princeton University, USA chang@math.princeton.edu
Michael Christ	University of California, Berkeley, USA mchrist@math.berkeley.edu
Charles Fefferman	Princeton University, USA cf@math.princeton.edu
Ursula Hamenstaedt	Universität Bonn, Germany ursula@math.uni-bonn.de
Vaughan Jones	U.C. Berkeley & Vanderbilt University vaughan.f.jones@vanderbilt.edu
Herbert Koch	Universität Bonn, Germany koch@math.uni-bonn.de
Izabella Laba	University of British Columbia, Canada ilaba@math.ubc.ca
Gilles Lebeau	Université de Nice Sophia Antipolis, France lebeau@unice.fr
László Lempert	Purdue University, USA lempert@math.purdue.edu
Richard B. Melrose	Massachussets Institute of Technology, USA rbm@math.mit.edu
Frank Merle	Université de Cergy-Pontoise, France Frank.Merle@u-cergy.fr
William Minicozzi II	Johns Hopkins University, USA minicozz@math.jhu.edu
Werner Müller	Universität Bonn, Germany mueller@math.uni-bonn.de
Yuval Peres	University of California, Berkeley, USA peres@stat.berkeley.edu
Gilles Pisier	Texas A&M University, and Paris 6 pisier@math.tamu.edu
Tristan Rivière	ETH, Switzerland riviere@math.ethz.ch
Igor Rodnianski	Princeton University, USA irod@math.princeton.edu
Wilhelm Schlag	University of Chicago, USA schlag@math.uchicago.edu
Sylvia Serfaty	New York University, USA serfaty@cims.nyu.edu
Yum-Tong Siu	Harvard University, USA siu@math.harvard.edu
Terence Tao	University of California, Los Angeles, USA tao@math.ucla.edu
Michael E. Taylor	Univ. of North Carolina, Chapel Hill, USA met@math.unc.edu
Gunther Uhlmann	University of Washington, USA gunther@math.washington.edu
András Vasy	Stanford University, USA andras@math.stanford.edu
Dan Virgil Voiculescu	University of California, Berkeley, USA dvv@math.berkeley.edu
Steven Zelditch	Northwestern University, USA zelditch@math.northwestern.edu

PRODUCTION
production@msp.org
Silvio Levy, Scientific Editor

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

The subscription price for 2013 is US $160/year for the electronic version, and $310/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

Analysis & PDE (ISSN 1948-206X electronic, 2157-5045 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.

APDE peer review and production are managed by EditFLOW® from Mathematical Sciences Publishers.

PUBLISHED BY
mathematical sciences publishers
nonprofit scientific publishing
http://msp.org/

© 2013 Mathematical Sciences Publishers

ANALYSIS AND PDE
Vol. 6, No. 5, 2013
dx.doi.org/10.2140/apde.2013.6.1001

A LICHNEROWICZ ESTIMATE FOR THE FIRST EIGENVALUE OF CONVEX DOMAINS IN KÄHLER MANIFOLDS

VINCENT GUEDJ, BORIS KOLEV AND NADER YEGANEFAR

In this article, we prove a Lichnerowicz estimate for a compact *convex* domain of a Kähler manifold whose Ricci curvature satisfies $\mathrm{Ric} \geq k$ for some constant $k > 0$. When equality is achieved, the boundary of the domain is totally geodesic and there exists a nontrivial holomorphic vector field.

We show that a ball of sufficiently large radius in complex projective space provides an example of a *strongly pseudoconvex* domain which is *not convex*, and for which the *Lichnerowicz estimate* fails.

1. Introduction

Let (M^n, g) be a compact n-dimensional Riemannian manifold. Assume first that M has no boundary. A theorem of Lichnerowicz [1958] asserts that if the Ricci curvature Ric of M satisfies $\mathrm{Ric} \geq k$ for some constant $k > 0$, the first nonzero eigenvalue λ of the Laplace operator satisfies

$$\lambda \geq \frac{n}{n-1}k. \tag{1-1}$$

Here, $nk/(n-1)$ should be viewed as the first nonzero eigenvalue of the round n-dimensional sphere $S^n(k/(n-1))$ of constant curvature $k/(n-1)$. Moreover, by a result of Obata [1962], the equality case in (1-1) is obtained if and only if M is isometric to this sphere. Reilly [1977] considered a similar problem, but for compact manifolds with boundary. Namely, he proved that if M is as in the Lichnerowicz theorem, except that it has a boundary such that its mean curvature with respect to the outward normal vector field is nonnegative, then the first eigenvalue λ of the Laplace operator with the Dirichlet boundary condition still satisfies (1-1). He also proved that the equality case characterizes a hemisphere in $S^n(k/(n-1))$.

In another direction, Lichnerowicz showed that for Kähler manifolds, his estimate (1-1) can be improved, by showing that, in this case, we have

$$\lambda \geq 2k.$$

Moreover, if equality is achieved, there is a nontrivial holomorphic vector field on M.

The purpose of this note is to consider the case of compact Kähler manifolds with boundary. As in Reilly's result, we will have to impose some convexity property on the boundary.

MSC2010: 35P15, 58C40.
Keywords: Lichnerowicz estimate, first eigenvalue, convex domains in Kähler manifolds.

Theorem 1.1. *Let M be a compact convex domain in a Kähler manifold. Assume that the Ricci curvature satisfies* $\mathrm{Ric} \geq k$ *for some constant* $k > 0$. *Then the first eigenvalue* λ *of the Laplacian with the Dirichlet boundary condition satisfies*

$$\lambda \geq 2k.$$

Moreover, if equality is achieved, the boundary ∂M *is totally geodesic and there is a nontrivial holomorphic vector field on* M.

Remark 1.2. As we will see in the proof, the convexity hypothesis may be relaxed into another condition of mean curvature type. More precisely, let $\boldsymbol{n}$ denote the outward unit normal vector field on the boundary ∂M, and let II and H be respectively the second fundamental form and the mean curvature. Denote also by J the complex structure of M. If we assume that on the boundary we have

$$(n-1)H + \mathrm{II}(J\boldsymbol{n}, J\boldsymbol{n}) \geq 0, \tag{1-2}$$

the Lichnerowicz estimate $\lambda \geq 2k$ holds (see inequality (4-3) and the remark just before Section 3.2). Now, convexity means that II is a nonnegative bilinear symmetric form, so that it obviously implies condition (1-2).

Remark 1.3. Jean-François Grosjean [2002, Theorem 1.1] proves that there is a Lichnerowicz type estimate on compact (real) manifolds with convex boundary and positive Ricci curvature, if there exists a nontrivial parallel p-form with $2 \leq p \leq n/2$. In the Kähler case, we can of course consider the Kähler form which is a nontrivial parallel 2-form, so that the result of Grosjean gives a Lichnerowicz estimate. But this estimate is weaker than ours. Note however that our result was known to Grosjean and is stated without proof in [2002, page 504].

Remark 1.4. It is natural to ask whether our result remains true if one assumes pseudoconvexity of the boundary instead of its convexity. It turns out that a ball of sufficiently large radius in complex projective space provides an example of a strongly pseudoconvex domain which is not convex, and for which the Lichnerowicz estimate fails (see Proposition 5.1 for more details on this).

Remark 1.5. In the real setting, one can consider the Laplacian with the Neumann boundary condition, and again with the convexity condition, one can show that the Lichnerowicz estimate (1-1) still holds for the first nonzero eigenvalue [Pak et al. 1986]. In the Kähler setting, by using the method of proof of Theorem 1.1, it should also be possible to prove that the conclusion of this theorem is true for the first nonzero eigenvalue of the Laplacian with the Neumann boundary condition. It should also be possible to get a similar result for the first nonzero eigenvalue of the $\bar{\partial}$-Laplacian with the absolute $\bar{\partial}$-condition on the boundary.

An immediate consequence of our theorem is the following.

Corollary 1.6. *Assume that M is a strongly convex domain in a complex manifold which can be endowed with a Kähler metric whose Ricci curvature satisfies* $\mathrm{Ric} \geq k$ *for some constant* $k > 0$. *Then the first eigenvalue* λ *of the Laplacian with the Dirichlet boundary condition satisfies*

$$\lambda > 2k.$$

Our proof follows the same strategy as the original proofs of Lichnerowicz and Reilly. We will actually give two slightly different proofs. The first proof is more adapted to the complex setting (see Section 3). We use an appropriate Bochner formula for the $\bar{\partial}$-Laplacian $\Box$ acting on (0, 1)-forms and apply it to $\bar{\partial} f$, where the function f is an eigenfunction of $\Box$ for the first eigenvalue. After integrating the result on M and integrating by parts, we get a Reilly-type formula for the $\bar{\partial}$-Laplacian which may be of independent interest. The desired eigenvalue estimate follows if we can prove that some boundary term is nonpositive, which is the case under the convexity hypothesis.The second proof rests on the well-known Reilly formula for real manifolds; see [Reilly 1977]. This is done in Section 4.

2. Background material

In this section, we recall some well-known facts that will be used in the proof of our main result.

2.1. ***Decomposition of the Hessian.*** Let f be a real valued smooth function on a Kähler manifold (M, J, g). Its Riemannian Hessian ∇df can be decomposed as the sum of a J-symmetric bilinear form and a J-skew-symmetric bilinear form. More specifically, we have

$$\nabla df = \mathrm{H}^1 f + \mathrm{H}^2 f$$

where for tangent vectors A and B,

$$\mathrm{H}^1 f(A, B) = \tfrac{1}{2}\{\nabla df(A, B) + \nabla df(JA, JB)\}$$

and

$$\mathrm{H}^2 f(A, B) = \tfrac{1}{2}\{\nabla df(A, B) - \nabla df(JA, JB)\}.$$

The two following facts may be easily checked.

(1) The (1, 1)-form associated to $\mathrm{H}^1 f$ by the complex structure J is $i\partial\bar{\partial} f$:

$$\mathrm{H}^1 f(JA, B) = i\partial\bar{\partial} f(A, B).$$

(2) In local coordinates, $\mathrm{H}^2 f$ has components

$$(\mathrm{H}^2 f)_{pq} = \overline{(\mathrm{H}^2 f)_{\bar{p}\bar{q}}} = \frac{\partial^2 f}{\partial z_p \partial z_q} - \Gamma^r_{pq} \frac{\partial f}{\partial z_r},$$

and the other components vanish. $\mathrm{H}^2 f$ is called the *complex Hessian*.

Since $J^* = J^{-1}$, we have $\|\nabla df\| = \|(\nabla df)^J\|$, where

$$(\nabla df)^J(A, B) := \nabla df(JA, JB).$$

Therefore

$$2\|\mathrm{H}^1 f\|^2 = \|\nabla df\|^2 + \langle \nabla df, (\nabla df)^J \rangle \tag{2-1}$$

and

$$2\|\mathrm{H}^2 f\|^2 = \|\nabla df\|^2 - \langle \nabla df, (\nabla df)^J \rangle. \tag{2-2}$$

2.2. ***Reilly formula for the (real) Laplacian.*** Let (M, g) be a Riemannian manifold. Let f be a smooth function on M and ∇df, Δf, and $\operatorname{grad} f$ be its Riemannian Hessian, its Laplacian (Laplace Beltrami), and its gradient on M, respectively. Let $\boldsymbol{n}$ denotes the outward unit normal vector field on ∂M and let II and H be the second fundamental form and the mean curvature, respectively. We choose the convention $\mathrm{II}(X, Y) = \langle \nabla_X \boldsymbol{n}, Y\rangle$ for any $X, Y \in T\partial M$. The Laplacian and the gradient on the boundary ∂M with the induced metric are denoted by $\bar\Delta$ and $\overline{\operatorname{grad}}$, respectively. The Reilly formula [Reilly 1977] is given by

$$\begin{aligned}&\int_M \|\nabla df\|^2\\ &= \int_M (\Delta f)^2 - \int_M \operatorname{Ric}(\operatorname{grad} f, \operatorname{grad} f) + 2\int_{\partial M} \bar\Delta f \frac{\partial f}{\partial \boldsymbol{n}}\sigma - (n-1)\int_{\partial M} H\Big(\frac{\partial f}{\partial \boldsymbol{n}}\Big)^2\sigma - \int_{\partial M} \mathrm{II}(\overline{\operatorname{grad}} f, \overline{\operatorname{grad}} f)\sigma.\end{aligned}$$

Moreover if we assume that f is vanishing on ∂M, then $\bar\Delta f = 0$, $\overline{\operatorname{grad}} f = 0$ and

$$\int_M \|\nabla df\|^2 = \int_M (\Delta f)^2 - \int_M \operatorname{Ric}(\operatorname{grad} f, \operatorname{grad} f) - (n-1)\int_{\partial M} H\Big(\frac{\partial f}{\partial \boldsymbol{n}}\Big)^2\sigma. \tag{2-3}$$

2.3. ***Bochner formula for the (complex) Laplacian.*** Let (M, g) be a Kähler manifold, and denote by ∇ its Levi-Civita connection. If α is a $(0, 1)$-form, we denote by $D''\alpha$ the $(0, 2)$-part of $\nabla\alpha$. More precisely, $\nabla\alpha$ is a section of the bundle $T^*M \otimes (T^*)^{0,1}M$; this bundle decomposes as a direct sum

$$((T^*)^{1,0}M \otimes (T^*)^{0,1}M) \oplus ((T^*)^{0,1}M \otimes (T^*)^{0,1}M),$$

and $D''\alpha$ is the projection of $\nabla\alpha$ on the second factor of this decomposition. In local complex coordinates, we have

$$(D''\alpha)_{\bar p\bar q} = \frac{\partial \alpha_{\bar q}}{\partial \bar z_p} - \Gamma^{\bar r}_{\bar p\bar q}\alpha_{\bar r}.$$

Now let $(D'')^*$ be the formal adjoint of D''. For a section β of $(T^*)^{0,1}M \otimes (T^*)^{0,1}M$ one can see that locally

$$((D'')^*\beta)_{\bar p} = -g^{q\bar r}\frac{\partial \beta_{\bar r\bar p}}{\partial z_q}.$$

Then we have the following Bochner formula for the $\bar\partial$-Laplacian $\square$ acting on $(0, 1)$-forms:

$$\square = (D'')^* D'' + \operatorname{Ric}. \tag{2-4}$$

For future reference, we also give the integration by parts formula for D'' in the presence of a boundary; see, for example, [Taylor 2011, Proposition 9.1]. Here, we assume that M is compact, and we let $\boldsymbol{n}$ denote the outward unit normal vector field on ∂M. The $(0, 1)$ part of the dual 1-form ν corresponding to $\boldsymbol{n}$ by the metric will be denoted by $\nu^{0,1}$. Finally, we let σ denote the measure induced on the boundary by the metric. For smooth α and β, we then have

$$\langle D''\alpha, \beta\rangle_{L^2(M)} = \langle \alpha, (D'')^*\beta\rangle_{L^2(M)} + \int_{\partial M} \langle \nu^{0,1} \otimes \alpha, \beta\rangle\sigma. \tag{2-5}$$

3. Bochner formula and the first eigenvalue

In this section, we will give the first proof of Theorem 1.1. Let $\Box$ denote the $\bar\partial$-Laplacian on M, which is given on forms by

$$\Box = \bar\partial\bar\partial^* + \bar\partial^*\bar\partial.$$

Recall that on a Kähler manifold, we have $\Box = \frac{1}{2}\Delta$. We will denote by μ the first eigenvalue of $\Box$ with the Dirichlet boundary condition, so that

$$\mu = \tfrac{1}{2}\lambda.$$

Now let f be a real valued eigenfunction of $\Box$ corresponding to the first eigenvalue μ. Thus $f : \overline{M} \to \mathbb{R}$ is smooth, vanishes on the boundary ∂M, and satisfies $\Box f = \mu f$. (Note that it is possible to choose f to be real valued, because $\Box$ is equal to half the Laplace Beltrami operator Δ.) We write the Bochner formula (2-4) for the (0, 1)-form $\bar\partial f$ and take the L^2-inner product of the resulting equality with $\bar\partial f$ itself:

$$\langle \Box\bar\partial f, \bar\partial f\rangle_{L^2(M)} = \langle (D'')^* D''\bar\partial f, \bar\partial f\rangle_{L^2(M)} + \int_M \operatorname{Ric}(\bar\partial f, \bar\partial f). \tag{3-1}$$

Using the fact that $\Box\bar\partial = \bar\partial\Box$ and $f|_{\partial M} = 0$, we can integrate by parts the left hand side of (3-1) to get

$$\begin{aligned}\langle \Box\bar\partial f, \bar\partial f\rangle_{L^2(M)} &= \langle \bar\partial\Box f, \bar\partial f\rangle_{L^2(M)}\\ &= \langle \bar\partial(\mu f), \bar\partial f\rangle_{L^2(M)}\\ &= \mu\langle \Box f, f\rangle_{L^2(M)}\\ &= \mu^2\|f\|^2_{L^2(M)}.\end{aligned}$$

We can deal with the Ricci term in the right hand side of (3-1) in a similar way:

$$\begin{aligned}\int_M \operatorname{Ric}(\bar\partial f, \bar\partial f) &\ge k\langle \bar\partial f, \bar\partial f\rangle_{L^2(M)}\\ &= k\langle \Box f, f\rangle_{L^2(M)}\\ &= k\mu\|f\|^2_{L^2(M)}.\end{aligned}$$

Finally, we can integrate by parts the first term in the right hand side of (3-1) (see formula (2-5)) to get

$$\langle (D'')^* D''\bar\partial f, \bar\partial f\rangle_{L^2(M)} = \|D''\bar\partial f\|^2_{L^2(M)} - \int_{\partial M} \langle D''\bar\partial f, \nu^{0,1}\otimes\bar\partial f\rangle\sigma, \tag{3-2}$$

and, combining this with our previous estimates, we obtain

$$\mu(\mu-k)\|f\|^2_{L^2(M)} \ge \|D''\bar\partial f\|^2_{L^2(M)} - \int_{\partial M} \langle D''\bar\partial f, \nu^{0,1}\otimes\bar\partial f\rangle\sigma. \tag{3-3}$$

As a consequence, if we set

$$I = -\int_{\partial M} \langle D''\bar\partial f, \nu^{0,1}\otimes\bar\partial f\rangle\sigma,$$

we get $\mu \ge k$, provided we can prove that $I \ge 0$. In the next subsection, we see that this is indeed the case under suitable assumptions on the boundary.

3.1. ***Boundary term.*** To estimate the boundary term I, we first notice that as f is real valued, we have

$$(D''\bar{\partial} f)_{\bar{p}\bar{q}} = (\mathrm{H}^2 f)_{\bar{p}\bar{q}}$$

so that

$$I = -\int_{\partial M} \langle \mathrm{H}^2 f, \nu^{0,1} \otimes \bar{\partial} f \rangle \sigma = -\int_{\partial M} \mathrm{H}^2 f(\boldsymbol{n}^{0,1}, (\partial f)^\sharp)\sigma.$$

We then choose a boundary defining function ρ for ∂M. This means that ρ is a smooth real valued function such that $M = \{\rho \le 0\}$, $\partial M = \{\rho = 0\}$, and $d\rho$ does not vanish on ∂M. By multiplying ρ by a suitable smooth positive function if necessary, we may assume that

$$\boldsymbol{n} = \operatorname{grad} \rho.$$

Moreover, near a fixed (but arbitrary) point of the boundary ∂M, we fix a local orthonormal frame adapted to the complex structure J which has the form

$$v_1, Jv_1, \dots, v_m, Jv_m = \boldsymbol{n} = \operatorname{grad} \rho.$$

We also set

$$e_p = \frac{1}{\sqrt{2}}(v_p - iJv_p), \quad p = 1, \dots, m.$$

Note that as f vanishes on ∂M, its derivatives along tangent vectors to ∂M also vanish and, consequently,

$$(\partial f)^\sharp = \frac{-i}{\sqrt{2}}(\boldsymbol{n} \cdot f)\bar{e}_m, \qquad \boldsymbol{n}^{0,1} = \frac{-i}{\sqrt{2}}\bar{e}_m,$$

where $\boldsymbol{n} \cdot f$ means $df(\boldsymbol{n})$. Therefore,

$$I = \frac{1}{2}\int_{\partial M} (\boldsymbol{n} \cdot f)\nabla df(\bar{e}_m, \bar{e}_m)\sigma,$$

which can be decomposed as $I = I_1 + iI_2$ with

$$I_1 = \frac{1}{4}\int_{\partial M} (\boldsymbol{n} \cdot f)[\nabla df(J\boldsymbol{n}, J\boldsymbol{n}) - \nabla df(\boldsymbol{n}, \boldsymbol{n})]\sigma$$

and

$$I_2 = -\frac{1}{2}\int_{\partial M} (\boldsymbol{n} \cdot f)\nabla df(J\boldsymbol{n}, \boldsymbol{n})\sigma.$$

Actually I_2 vanishes because I is a real number. (This follows from the fact that in Equation (3-1), the left hand side and the Ricci term are real numbers, so that the term involving D'' is also a real number. This implies, by Equation (3-2), that the boundary term I is a real number as well. There is also a more conceptual reason for the vanishing of I_2; see Section 3.2.) We now turn our attention to I_1. As $\triangle f = \mu f = 0$ on ∂M, the trace of ∇df is also zero on ∂M:

$$\nabla df(J\boldsymbol{n}, J\boldsymbol{n}) - \nabla df(\boldsymbol{n}, \boldsymbol{n}) = \sum_{k=1}^{m-1}[\nabla df(v_k, v_k) + \nabla df(Jv_k, Jv_k)] + 2\nabla df(J\boldsymbol{n}, J\boldsymbol{n}).$$

We notice that all vectors appearing in the right hand side are tangent to the boundary. For such a vector u, we have on ∂M

$$\begin{aligned}\nabla df(u,u) &= -\langle \nabla_u u, \boldsymbol{n}\rangle(\boldsymbol{n}\cdot f)\\ &= \langle \nabla_u \boldsymbol{n}, u\rangle(\boldsymbol{n}\cdot f)\\ &= (\boldsymbol{n}\cdot f)\nabla d\rho(u,u).\end{aligned}$$

This implies

$$I_1 = \frac{1}{4}\int_{\partial M}(\boldsymbol{n}\cdot f)^2\left(\sum_{k=1}^{m-1}[\nabla d\rho(v_k,v_k)+\nabla d\rho(Jv_k,Jv_k)] + 2\nabla d\rho(J\boldsymbol{n},J\boldsymbol{n})\right)\sigma. \tag{3-4}$$

If we assume that ∂M is convex, all terms in the integrand of the right hand side are nonnegative, so that $I = I_1 \geq 0$ as desired. This proves that $\mu \geq k$ in the convex case.

It remains to deal with the equality case. If we assume that $\mu = k$, then, by (3-3), we must have $D''\bar\partial f = 0$ and $I = 0$. On the one hand, $D''\bar\partial f = 0$ means that the $(1,0)$-vector field associated to $\bar\partial f$ by the metric is a (nonzero) holomorphic vector field. On the other hand, from $I = 0$, we infer that the integrand in Equation (3-4) has to vanish identically on the boundary:

$$(\boldsymbol{n}\cdot f)^2\left(\sum_{k=1}^{m-1}\{\nabla d\rho(v_k,v_k)+\nabla d\rho(Jv_k,Jv_k)\} + 2\nabla d\rho(J\boldsymbol{n},J\boldsymbol{n})\right) = 0.$$

Assume by contradiction that ∂M is not totally geodesic (but is still convex of course). Then the term between the brackets is positive at some point and we will get the vanishing of $\boldsymbol{n}.f$ on an open subset of ∂M. But f is in the kernel of the elliptic operator $\Box - \mu$ and vanishes on ∂M. By the unique continuation principle for elliptic operators (see, for example, [Booß-Bavnbek and Wojciechowski 1993]), f has to vanish on M as well, which is absurd. Therefore, ∂M is totally geodesic. This completes the proof of Theorem 1.1.

Remark. With our conventions, $\nabla d\rho$ is nothing but the second fundamental form of ∂M. Thus, we recover condition (1-2) of Remark 1.2.

3.2. *A direct proof that the boundary term is real.*

The fact that

$$I_2 = -\frac{1}{2}\int_{\partial M}(\boldsymbol{n}\cdot f)\nabla df(J\boldsymbol{n},\boldsymbol{n})\sigma$$

vanishes is also a consequence of the fact that the expression

$$(\boldsymbol{n}\cdot f)\nabla df(J\boldsymbol{n},\boldsymbol{n})\sigma = (\boldsymbol{n}\cdot f)(J\boldsymbol{n}\cdot\boldsymbol{n}\cdot f)\sigma$$

is an exact differential form on the closed manifold ∂M. Indeed, the vector field $J\boldsymbol{n} = J\operatorname{grad}\rho$ is the Hamiltonian vector field associated to ρ. This means that if ω is the Kähler form,

$$i_{J\boldsymbol{n}}\omega = -d\rho.$$

Hence

$$di_{J\boldsymbol{n}}i_{\boldsymbol{n}}\omega^m = -md(\boldsymbol{n}\cdot\rho)\wedge\omega^{m-1} - m(m-1)d\rho\wedge di_{\boldsymbol{n}}\omega\wedge\omega^{m-2}.$$

Let $j : \partial M \to M$ be the inclusion map. Since the functions $\boldsymbol{n} \cdot \rho$ and ρ are constant on ∂M, we have

$$j^*(di_{J\boldsymbol{n}} i_{\boldsymbol{n}} \omega^m) = 0.$$

Now, $J\boldsymbol{n}$ is a vector field defined on a neighborhood of ∂M whose restriction to ∂M is tangent to ∂M, so that

$$j^*(i_{J\boldsymbol{n}} \beta) = i_{J\boldsymbol{n}} j^*(\beta)$$

for any differential form β. As a consequence, we get

$$di_{J\boldsymbol{n}} j^*(i_{\boldsymbol{n}} \omega^m) = 0.$$

Finally, we have

$$j^*(i_{\boldsymbol{n}} \omega^m) = \sigma$$

and

$$di_{J\boldsymbol{n}} \sigma = 0.$$

Defining a vector field X by

$$X = \tfrac{1}{2}(\boldsymbol{n} \cdot f)^2 J\boldsymbol{n},$$

it follows that, on ∂M, we have

$$di_X \sigma = (\boldsymbol{n} \cdot f)(J\boldsymbol{n} \cdot \boldsymbol{n} \cdot f)\sigma.$$

4. Reilly formula and the first eigenvalue

In this section, we present an alternative proof of our main result which was indicated by the referee. It is based on Reilly's formula, a well-known result in *real* Riemannian geometry, which is probably the tool used in [Grosjean 2002, page 504].

This complements nicely the arguments given in Section 3, which have a complex geometry flavor. The complex proof is a bit longer, as we first need to establish a Reilly-type formula for the $\bar{\partial}$-Laplacian. Given the importance of the $\bar{\partial}$-Laplacian in complex geometry, it is likely that this (*complex*) Reilly formula will have other applications.

Let M be a compact smooth domain in a Kähler manifold of complex dimension m and real dimension $n = 2m$, with metric g and Ricci curvature bounded from below by some positive constant k. The outward unit normal vector field on the boundary ∂M is denoted by $\boldsymbol{n}$. Our aim is to prove a Lichnerowicz estimate for the first eigenvalue by using the Reilly formula. We begin with some general facts.

Let G be a symmetric, covariant 2-tensor field and X a vector field. We have

$$\operatorname{div}(G(X, \cdot)) = (\operatorname{div} G)(X) + \langle G, DX^\flat \rangle,$$

where $DX^\flat$ is the symmetric part of the covariant 2-tensor field $\nabla X^\flat$. Specializing this formula for $G = (\nabla df)^J$ and $X = \operatorname{grad} f$, for some smooth real function f, we get

$$\operatorname{div} \alpha = \operatorname{Tr}[\nabla^2 df(\cdot, J\cdot, J \operatorname{grad} f)] + \langle (\nabla df)^J, \nabla df \rangle,$$

where

$$\alpha(X) := \nabla df(JX, J\operatorname{grad} f).$$

Given an orthonormal basis $(e_i)_{1\leq i\leq n}$ at a point x in M, we have

$$\begin{aligned}\operatorname{Tr}[\nabla^2 df(\cdot, J\cdot, J\operatorname{grad} f)] &= \tfrac{1}{2}\{\nabla^2 df(e_i, Je_i, J\operatorname{grad} f) - \nabla^2 df(Je_i, e_i, J\operatorname{grad} f)\}\\ &= -\tfrac{1}{2}[R(e_i, Je_i)\,df](J\operatorname{grad} f)\\ &= \tfrac{1}{2}R(e_i, Je_i, J\operatorname{grad} f, \operatorname{grad} f)\\ &= -\operatorname{Ric}(\operatorname{grad} f, \operatorname{grad} f).\end{aligned}$$

Hence we get

$$\operatorname{div}\alpha = -\operatorname{Ric}(\operatorname{grad} f, \operatorname{grad} f) + \langle(\nabla df)^J, \nabla df\rangle.$$

Integrating by parts we find

$$\int_M \langle(\nabla df)^J, \nabla df\rangle = \int_M \operatorname{Ric}(\operatorname{grad} f, \operatorname{grad} f) + \int_{\partial M}\alpha(\boldsymbol{n})\sigma,$$

but, for a point $m \in \partial M$, we have

$$\begin{aligned}\alpha(\boldsymbol{n})_m &= (\nabla df)_m(J\boldsymbol{n}, J\operatorname{grad} f)\\ &= (\nabla df)_m\Big(J\boldsymbol{n}, J\Big(\overline{\operatorname{grad}} f + \frac{\partial f}{\partial \boldsymbol{n}}\boldsymbol{n}\Big)\Big)\\ &= (\nabla df)_m(J\boldsymbol{n}, J\overline{\operatorname{grad}} f) + \frac{\partial f}{\partial \boldsymbol{n}}(\nabla df)_m(J\boldsymbol{n}, J\boldsymbol{n}).\end{aligned}$$

Now, recall that the second fundamental form II of ∂M is defined as follows (see [Gallot et al. 2004, Chapter 5] for details). Let U, V be local vector fields in M which extend some vector fields u, v on ∂M, in a neighborhood of $m \in \partial M$. We have

$$(\nabla_U V)_m = (\overline{\nabla}_u v)_m - \mathrm{II}_m(u, v)\boldsymbol{n},$$

from which we deduce that

$$(\nabla df)_m(u, v) = (\overline{\nabla} df)_m(u, v) + \frac{\partial f}{\partial \boldsymbol{n}}\mathrm{II}_m(u, v).$$

Therefore

$$\alpha(\boldsymbol{n})_m = (\nabla df)_m(J\boldsymbol{n}, J\overline{\operatorname{grad}} f) + \frac{\partial f}{\partial \boldsymbol{n}}\overline{\nabla} df(J\boldsymbol{n}, J\boldsymbol{n}) + \Big(\frac{\partial f}{\partial \boldsymbol{n}}\Big)^2 \mathrm{II}(J\boldsymbol{n}, J\boldsymbol{n}).$$

If we assume furthermore that f vanishes on the boundary, the first two terms of the right hand side of the equation above vanish as well, so we finally obtain

$$\int_M \langle(\nabla df)^J, \nabla df\rangle = \int_M \operatorname{Ric}(\operatorname{grad} f, \operatorname{grad} f) + \int_{\partial M}\Big(\frac{\partial f}{\partial \boldsymbol{n}}\Big)^2 \mathrm{II}(J\boldsymbol{n}, J\boldsymbol{n})\sigma. \tag{4-1}$$

On the left side of the Reilly formula (2-3), we can first use (2-2) to replace $\|\nabla df\|^2$ by

$$\|\nabla df\|^2 = 2\|\mathrm{H}^2 f\|^2 + \langle\nabla df, (\nabla df)^J\rangle,$$

and then use (4-1) to get

$$2\int_M \|\mathrm{H}^2 f\|^2 = \int_M (\Delta f)^2 - 2\int_M \mathrm{Ric}(\operatorname{grad} f, \operatorname{grad} f) - \int_{\partial M} [(n-1)H + \mathrm{II}(J\boldsymbol{n}, J\boldsymbol{n})]\left(\frac{\partial f}{\partial \boldsymbol{n}}\right)^2 \sigma. \tag{4-2}$$

Suppose now that f is a real valued eigenfunction of Δ corresponding to the first eigenvalue λ of Δ, so that $f : \overline{M} \to \mathbb{R}$ is smooth, vanishes on the boundary ∂M, and satisfies $\Delta f = \lambda f$. The hypothesis on the Ricci curvature implies that

$$\int_M \mathrm{Ric}(\operatorname{grad} f, \operatorname{grad} f) \geq k\|df\|_{L^2}^2 = k\langle \Delta f, f\rangle_{L^2} = k\lambda \|f\|_{L^2}^2.$$

From (4-2), we then infer

$$\lambda(\lambda - 2k)\|f\|_{L^2}^2 \geq \int_{\partial M} [(n-1)H + \mathrm{II}(J\boldsymbol{n}, J\boldsymbol{n})]\left(\frac{\partial f}{\partial \boldsymbol{n}}\right)^2 \sigma. \tag{4-3}$$

Finally, if we assume that the boundary is convex, II is by definition a symmetric bilinear form which is nonnegative, so that its trace H is also nonnegative. Therefore, the left hand side of the previous equation is nonnegative, and we get $\lambda \geq 2k$, as desired. For the equality case, we can argue as in the end of Section 3.1.

5. Counterexample in the pseudoconvex case

We use the notation introduced in Section 3. It is clear from the proof of Theorem 1.1 that in order to get the estimate $\mu \geq k$, it is enough to assume that on the boundary we have

$$\sum_{k=1}^{m-1} \{\nabla d\rho(v_k, v_k) + \nabla d\rho(Jv_k, Jv_k)\} + 2\nabla d\rho(J\boldsymbol{n}, J\boldsymbol{n}) \geq 0, \tag{5-1}$$

and not necessarily the convexity of ∂M. We may rewrite this condition as

$$\sum_{k=1}^{m-1} \mathrm{H}^1 \rho(v_k, v_k) + \nabla d\rho(J\boldsymbol{n}, J\boldsymbol{n}) \geq 0.$$

Here, $\sum_{k=1}^{m-1} \mathrm{H}^1 \rho(v_k, v_k)$ is the trace of the Levi form of the boundary, which would be nonnegative if ∂M were assumed to be only pseudoconvex. The extra term $\nabla d\rho(J\boldsymbol{n}, J\boldsymbol{n})$, however, can usually not be controlled in the pseudoconvex case. This suggests that the conclusion of Theorem 1.1 does not generally hold in this case, as we now explain.

We consider here the complex m-dimensional projective space $\mathbb{P}^m(\mathbb{C})$ equipped with the Fubini–Study metric normalized so that the holomorphic sectional curvature is 4 (the Einstein constant is thus $2(m+1)$ and the diameter is $\pi/2$).

Proposition 5.1. *Fix some point $x \in \mathbb{P}^m(\mathbb{C})$, some $r_0 \in]0, \pi/2[$, and let M be the geodesic ball centered at x, of radius r_0.*

(i) *If $r_0 \in]\pi/4, \pi/2[$, M is strongly pseudoconvex, not convex.*

(ii) *The first eigenvalue of M with Dirichlet boundary conditions goes to 0 as r_0 approaches $\pi/2$.*

Proof. The first point is a well-known result. For completeness, we outline the proof here. Denote by r the distance function from x, and set $\rho = r^2 - r_0^2$, so that ρ is a smooth defining function for M. We want to compute the eigenvalues of the Hessian of ρ. As

$$\nabla d\rho = 2r\nabla dr + 2dr \otimes dr,$$

we only have to compute the eigenvalues of ∇dr. To do this, we proceed as in the proof of [Greene and Wu 1979, Theorem A, page 19]. Recall that for a tangent vector u, the curvature $R(u, .)u$ of $\mathbb{P}^m(\mathbb{C})$ is given by [Berger et al. 1971, Proposition F.34]

$$R(u, .)u = \begin{cases} 0 & \text{on } \mathbb{R}u, \\ 4\mathrm{Id} & \text{on } \mathbb{R}Ju, \\ \mathrm{Id} & \text{on the orthogonal complement of } (u, Ju). \end{cases}$$

Let γ be a normal geodesic starting from x. We can choose a parallel frame along γ which has the form $v_1, Jv_1, \dots, v_m, Jv_m = \operatorname{grad} r$. Using the explicit expression of R, it is then easy to check that the space of Jacobi fields V along γ satisfying $V(0) = 0$ and $V \perp \dot\gamma$ has as a basis $V_i = \sin(r)v_i$, JV_i, $i = 1, \dots, m-1$ and $V_m = \sin(2r)v_m$. Using the second variation formula, we see that ∇dr is diagonalized in the basis $v_1, Jv_1, \dots, v_m, Jv_m$ with eigenvalues $\cot(r)$ (of order $2m-2$), $2\cot(2r)$, and 0. If $r = r_0 \in]\pi/4, \pi/2[$, we infer that the Levi form of ρ is positive definite, being equal to $2r_0\cot(r_0)\mathrm{Id}$ on the Levi distribution. In other words, M is strongly pseudoconvex. However, M is not convex because the principal curvature $2\cot(2r_0)$ is negative.

As for the second point of our proposition, it is, for example, a consequence of [Chavel and Feldman 1978, Theorem 1], which states the following: Let X be a compact Riemannian manifold and let $X' \subset X$ be a submanifold. For small $\varepsilon > 0$, let X'_ε be the ε-neighborhood of X' in X and denote by Ω_ε the set $X \setminus X'_\varepsilon$. Let (λ_j) be the spectrum of X and let $(\lambda_j(\varepsilon))$ be the spectrum of Ω_ε with Dirichlet boundary conditions. If the codimension of X' in X is at least 2, then, for all j, $\lambda_j(\varepsilon) \to \lambda_{j-1}$ as $\varepsilon \to 0$. In our case, we can take $X = \mathbb{P}^m(\mathbb{C})$ and $X' = \mathbb{P}^{m-1}(\mathbb{C})$, which we view as the cut locus of our fixed point x. If $\varepsilon = \pi/2 - r_0$, Ω_ε actually coincides with M and we get (ii). ☐

Acknowledgements

We thank Saïd Ilias for bringing [Grosjean 2002] to our attention and Jean-François Grosjean for explaining his work. We also thank the referee for several useful remarks, and for showing us the second proof of Theorem 1.1.

References

[Berger et al. 1971] M. Berger, P. Gauduchon, and E. Mazet, *Le spectre d'une variété riemannienne*, Lecture Notes in Mathematics **194**, Springer, Berlin, 1971. MR 43 #8025 Zbl 0223.53034

[Booß-Bavnbek and Wojciechowski 1993] B. Booß-Bavnbek and K. P. Wojciechowski, *Elliptic boundary problems for Dirac operators*, Birkhäuser, Boston, MA, 1993. MR 94h:58168 Zbl 0797.58004

[Chavel and Feldman 1978] I. Chavel and E. A. Feldman, "Spectra of domains in compact manifolds", *J. Funct. Anal.* **30**:2 (1978), 198–222. MR 80c:58027 Zbl 0392.58016

[Gallot et al. 2004] S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian geometry*, 3rd ed., Springer, Berlin, 2004. MR 2005e:53001 Zbl 1068.53001

[Greene and Wu 1979] R. E. Greene and H. Wu, *Function theory on manifolds which possess a pole*, Lecture Notes in Mathematics **699**, Springer, Berlin, 1979. MR 81a:53002 Zbl 0414.53043

[Grosjean 2002] J.-F. Grosjean, "A new Lichnerowicz-Obata estimate in the presence of a parallel p-form", *Manuscripta Math.* **107**:4 (2002), 503–520. MR 2003f:58060 Zbl 1017.58004

[Lichnerowicz 1958] A. Lichnerowicz, *Géométrie des groupes de transformations*, Travaux et Recherches Mathématiques, III. Dunod, Paris, 1958. MR 23 #A1329 Zbl 0096.16001

[Obata 1962] M. Obata, "Certain conditions for a Riemannian manifold to be isometric with a sphere", *J. Math. Soc. Japan* **14** (1962), 333–340. MR 25 #5479 Zbl 0115.39302

[Pak et al. 1986] E. Pak, H. Minn, O. K. Yoon, and D. P. Chi, "On the first eigenvalue estimate of the Dirichlet and Neumann problem", *Bull. Korean Math. Soc.* **23**:1 (1986), 21–25. MR 87k:58279 Zbl 0604.53018

[Reilly 1977] R. C. Reilly, "Applications of the Hessian operator in a Riemannian manifold", *Indiana Univ. Math. J.* **26**:3 (1977), 459–472. MR 57 #13799 Zbl 0391.53019

[Taylor 2011] M. E. Taylor, *Partial differential equations I*, Applied Mathematical Scences **115**, Springer, New York, 2011. Zbl 1206.35002

Received 25 Jan 2012. Revised 10 Jun 2012. Accepted 27 Sep 2012.

VINCENT GUEDJ: `vincent.guedj@math.univ-toulouse.fr`
Institut Universitaire de France and Institut Mathematiques de Toulouse, Universite Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse Cedex 9, France

BORIS KOLEV: `kolev@cmi.univ-mrs.fr`
Laboratoire d'Analyse, Topologie, Probabilités, CNRS & Aix-Marseille University, CMI, 39 rue F. Joliot-Curie, 13453 Marseille Cedex 13, France

NADER YEGANEFAR: `nader.yeganefar@cmi.univ-mrs.fr`
Laboratoire d'Analyse, Topologie, Probabilités, Aix-Marseille University, CMI, 39 rue F. Joliot-Curie, 13453 Marseille Cedex 13, France

ANALYSIS AND PDE
Vol. 6, No. 5, 2013
dx.doi.org/10.2140/apde.2013.6.1013

SHARP MODULUS OF CONTINUITY FOR PARABOLIC EQUATIONS ON MANIFOLDS AND LOWER BOUNDS FOR THE FIRST EIGENVALUE

BEN ANDREWS AND JULIE CLUTTERBUCK

We derive sharp estimates on the modulus of continuity for solutions of the heat equation on a compact Riemannian manifold with a Ricci curvature bound, in terms of initial oscillation and elapsed time. As an application, we give an easy proof of the optimal lower bound on the first eigenvalue of the Laplacian on such a manifold as a function of diameter.

1. Introductory comments

In our previous papers [Andrews and Clutterbuck 2009a; 2009b] we proved sharp bounds on the modulus of continuity of solutions of various parabolic boundary value problems on domains in Euclidean space. In this paper, our aim is to extend these estimates to parabolic equations on manifolds. Precisely, let (M, g) be a compact Riemannian manifold with induced distance function d, diameter

$$\sup\{d(x, y) : x, y \in M\} = D,$$

and lower Ricci curvature bound $\mathrm{Ric}(v, v) \geq (n-1)\kappa g(v, v)$. Let $a : T^*M \to \mathrm{Sym}_2(T^*M)$ be a parallel equivariant map (so that $a(S^*\omega)(S^*\mu, S^*\nu) = a(\omega)(\mu, \nu)$ for any ω, μ, ν in T_x^*M and $S \in O(T_xM)$, while $\nabla\,(a(\omega)(\mu, \nu)) = 0$ whenever $\nabla\omega = \nabla\mu = \nabla\nu = 0$). Then we consider solutions to the parabolic equation

$$\frac{\partial u}{\partial t} = a^{ij}(Du)\nabla_i\nabla_j u. \tag{1}$$

Our assumptions imply that the coefficients a^{ij} have the form

$$a(Du)(\xi, \xi) = \alpha(|Du|)\frac{(Du \cdot \xi)^2}{|Du|^2} + \beta(|Du|)\left(|\xi|^2 - \frac{(Du \cdot \xi)^2}{|Du|^2}\right) \tag{2}$$

for some smooth positive functions α and β. Of particular interest are the cases of the heat equation (with $\alpha = \beta = 1$) and the p-laplacian heat flows (with $\alpha = (p-1)|Du|^{p-2}$ and $\beta = |Du|^{p-2}$). Here we are principally concerned with the case of manifolds without boundary, but can also allow M to have a nontrivial convex boundary (in which case we impose Neumann boundary conditions $D_\nu u = 0$). Our main aim is to provide the following estimates on the modulus of continuity of solutions in terms of the initial oscillation, elapsed time, κ, and D.

Andrews was supported by Discovery Projects grants DP0985802 and DP120102462 of the Australian Research Council.
MSC2010: 35K05, 35K55, 35P15.
Keywords: eigenvalue lower bound, heat equation, modulus of continuity.

Theorem 1 (modulus of continuity estimate). *Let* (M, g) *be a compact Riemannian manifold* (*possibly with smooth*, *uniformly locally convex boundary*) *with diameter* D *and Ricci curvature bound* $\mathrm{Ric} \geq (n-1)\kappa g$ *for some constant* $\kappa \in \mathbb{R}$. *Let* $u : M \times [0, T) \to \mathbb{R}$ *be a smooth solution to* (1) *with Neumann boundary conditions if* $\partial M \neq \varnothing$. *Suppose that*

- $u(\,\cdot\,, 0)$ *has a smooth modulus of continuity* $\varphi_0 : [0, D/2] \to \mathbb{R}$ *with* $\varphi_0(0) = 0$ *and* $\varphi_0' \geq 0$;
- $\varphi : [0, D/2] \times \mathbb{R}_+ \to \mathbb{R}$ *satisfies*
  (i) $\varphi(z, 0) = \varphi_0(z)$ *for each* $z \in [0, D/2]$;
  (ii) $\partial\varphi/\partial t \geq \alpha(\varphi')\varphi'' - (n-1)\boldsymbol{T}_\kappa \beta(\varphi')\varphi'$;
  (iii) $\varphi' \geq 0$ *on* $[0, D/2] \times \mathbb{R}_+$.

Then $\varphi(\,\cdot\,, t)$ *is a modulus of continuity for* $u(\,\cdot\,, t)$ *for each* $t \in [0, T)$:

$$|u(x,t) - u(y,t)| \leq 2\varphi\left(\frac{d(x,y)}{2}, t\right).$$

Here we use the notation

$$\boldsymbol{C}_\kappa(\tau) = \begin{cases} \cos\sqrt{\kappa}\tau, & \kappa > 0, \\ 1, & \kappa = 0, \\ \cosh\sqrt{-\kappa}\tau, & \kappa < 0 \end{cases} \quad \text{and} \quad \boldsymbol{S}_\kappa(\tau) = \begin{cases} (1/\sqrt{\kappa})\sin\sqrt{\kappa}\tau, & \kappa > 0, \\ \tau, & \kappa = 0, \\ (1/\sqrt{-\kappa})\sinh\sqrt{-\kappa}\tau, & \kappa < 0, \end{cases} \tag{3}$$

and

$$\boldsymbol{T}_\kappa(s) := \kappa\frac{\boldsymbol{S}_\kappa(s)}{\boldsymbol{C}_\kappa(s)} = \begin{cases} \sqrt{\kappa}\tan\left(\sqrt{\kappa}s\right), & \kappa > 0, \\ 0, & \kappa = 0, \\ -\sqrt{-\kappa}\tanh\left(\sqrt{-\kappa}s\right), & \kappa < 0. \end{cases}$$

These estimates are sharp, holding exactly for certain symmetric solutions on particular warped product spaces. The modulus of continuity estimates also imply sharp gradient bounds which hold in the same situation. The central ingredient in our argument is a comparison result for the second derivatives of the distance function (Theorem 3) which is a close relative of the well-known Laplacian comparison theorem. We remark that the assumption of smoothness can be weakened: for example, in the case of the p-laplacian heat flow, we do not expect solutions to be smooth near spatial critical points, but nevertheless solutions are smooth at other points, and this is sufficient for our argument.

As an immediate application of the modulus of continuity estimates, we provide a new proof of the optimal lower bound on the smallest positive eigenvalue of the Laplacian in terms of D and κ. Precisely, if we define

$$\lambda_1(M, g) = \inf\left\{\int_M |Du|_g^2\, d\mathrm{Vol}(g) : \int_M u^2\, d\mathrm{Vol}(g) = 1,\ \int_M u\, d\mathrm{Vol}(g) = 0\right\}$$

and

$$\lambda_1(D, \kappa, n) = \inf\{\lambda_1(M, g) : \dim(M) = n,\ \mathrm{diam}(M) \leq D,\ \mathrm{Ric} \geq (n-1)\kappa g\},$$

then we characterize $\lambda_1(D, \kappa)$ precisely as the first eigenvalue of a certain one-dimensional Sturm–Liouville problem.

Theorem 2 (lower bound on the first eigenvalue). *Let μ be the first eigenvalue of the Sturm–Liouville problem*

$$\begin{gathered}\frac{1}{\boldsymbol{C}_\kappa^{n-1}}\left(\Phi'\boldsymbol{C}_\kappa^{n-1}\right)' + \mu\Phi = 0 \quad \textit{on } [-D/2, D/2],\\ \Phi'(\pm D/2) = 0.\end{gathered} \tag{4}$$

Then $\lambda_1(D, \kappa, n) = \mu$.

Previous results in this direction include those derived from gradient estimates in [Li 1979; Li and Yau 1980], with the sharp result for nonnegative Ricci curvature first proved in [Zhong and Yang 1984]. The complete result as stated above is implicit in [Kröger 1992, Theorem 2]. Chen and Wang [1994] used stochastic methods to prove an apparently equivalent result. The result appears to have been first explicitly stated in the form above in by Bakry and Qian [2000, Theorem 14], who also used gradient estimate methods. Our contribution is the rather simple proof using the long-time behavior of the heat equation (a method which was also central in our work on the fundamental gap conjecture [Andrews and Clutterbuck 2011], and which has also been employed successfully in [Ni 2013]), which seems considerably easier than the previously available arguments. In particular, the complications arising in previous works from possible asymmetry of the first eigenfunction are avoided in our argument. A similar argument proving the sharp lower bound for λ_1 on a Bakry–Emery manifold may be found in [Andrews and Ni 2012].

The estimate in Theorem 2 is sharp (that is, we obtain an equality and not just an inequality), since, for a given diameter D and Ricci curvature bound κ, we can construct a sequence of manifolds satisfying these bounds on which the first eigenvalue approaches μ_1; see the remarks after Corollary 1 in [Kröger 1992]. We include a discussion of these examples in Section 5, since the examples required for our purposes are a simpler subset of those constructed in [Kröger 1992]. We also include in Section 6 a discussion of the implications for a conjectured inequality of Li.

2. A comparison theorem for the second derivatives of distance

Theorem 3. *Let (M, g) be a complete connected Riemannian manifold with a lower Ricci curvature bound $\mathrm{Ric} \ge (n-1)\kappa g$, and let φ be a smooth function with $\varphi' \ge 0$. Then on $(M \times M) \setminus \{(x, x) : x \in M\}$ the function $v(x, y) = 2\varphi(d(x, y)/2)$ is a viscosity supersolution of*

$$\mathscr{L}[\nabla^2 v, \nabla v] = 2[\alpha(\varphi')\varphi'' - (n-1)\boldsymbol{T}_\kappa \beta(\varphi')\varphi']|_{d/2},$$

where

$$\mathscr{L}[B, \omega] = \inf\left\{\mathrm{tr}(AB) : \begin{array}{l} A \in \mathrm{Sym}_2(T^*_{x,y}(M \times M)),\\ A \ge 0,\\ A|_{T^*_x M} = a(\omega|_{T_x M}),\\ A|_{T^*_y M} = a(\omega|_{T_y M}) \end{array}\right\}$$

for any $B \in \mathrm{Sym}_2(T_{x,y}(M \times M)$ and $\omega \in T^*_{(x,y)}(M \times M)$.

Proof. By approximation it suffices to consider the case where φ' is strictly positive. Let x and y be fixed, with $y \neq x$ and $d = d(x, y)$, and let $\gamma : [-d/2, d/2] \to M$ be a minimizing geodesic from x to y (that is, with $\gamma(-d/2) = x$ and $\gamma(d/2) = y$) parametrized by arc length. Choose an orthonormal basis $\{E_i\}_{1\le i\le n}$ for T_xM with $E_n = \gamma'(-d/2)$. Use parallel transport along γ to produce an orthonormal basis $\{E_i(s)\}_{1\le i\le n}$ for $T_{\gamma(s)}M$ with $E_n(s) = \gamma'(s)$ for each $s \in [-d/2, d/2]$. Let $\{E_*^i\}_{1\le i\le n}$ be the dual basis for $T^*_{\gamma(s)}M$.

To prove the theorem, consider any smooth function ψ defined on a neighborhood of (x, y) in $M \times M$ such that $\psi \le v$ and $\psi(x, y) = v(x, y)$. We must prove that

$$\mathscr{L}[\nabla^2\psi, \nabla\psi]|_{(x,y)} \le 2[\alpha(\varphi')\varphi'' - (n-1)\beta(\varphi')\varphi'\boldsymbol{T}_\kappa]|_{d(x,y)/2}.$$

By definition of $\mathscr{L}$, it suffices to find a nonnegative $A \in \mathrm{Sym}_2(T^*_{x,y}(M\times M))$ such that $A|_{T_xM} = a(\nabla\psi|_{T_xM})$ and $A|_{T_yM} = a(\nabla\psi|_{T_yM})$, with $\mathrm{tr}(AD^2\psi) \le 2[\alpha(\varphi')\varphi'' - (n-1)\beta(\varphi')\varphi'\boldsymbol{T}_\kappa]|_{d/2}$.

Before choosing this, we observe that $\nabla\psi$ is determined by d and φ: We have $\psi \le 2\varphi \circ d/2$ with equality at (x, y). In particular, we have (since φ is nondecreasing)

$$\psi(\gamma(s), \gamma(t)) \le 2\varphi(d(\gamma(s), \gamma(t))/2) \le 2\varphi(L[\gamma|_{[s,t]}]/2) \le 2\varphi(|t-s|/2)$$

for all $s \neq t$, with equality when $t = d/2$ and $s = -d/2$. This gives $\nabla\psi(E_n, 0) = -\varphi'(d/2)$ and $\nabla\psi(0, E_n) = \varphi'(d/2)$. To identify the remaining components of $\nabla\psi$, we define

$$\gamma_i^y(r, s) = \exp_{\gamma(s)}(r(1/2 + s/d)E_i(s))$$

for $1 \le i \le n-1$. Then we have

$$\psi(x, \exp_y(rE_i)) \le 2\varphi(L[\gamma_i^y(r, \cdot)]/2)$$

with equality at $r = 0$. The right-hand side is a smooth function of r with derivative zero, from which it follows that $\nabla\psi(0, E_i) = 0$. Similarly, we have $\nabla\psi(E_i, 0) = 0$ for $i = 1, \dots, n-1$. Therefore we have

$$\nabla\psi|_{(x,y)} = \varphi'(d(x, y)/2)(-E_*^n, E_*^n).$$

In particular, by (2), we have

$$a(\nabla\psi|_{T_xM}) = \alpha(\varphi')E_n \otimes E_n + \beta(\varphi')\sum_{i=1}^{n-1} E_i \otimes E_i,$$

and similarly for y.

Now we choose A as follows:

$$A = \alpha(\varphi')(E_n, -E_n) \otimes (E_n, -E_n) + \beta(\varphi')\sum_{i=1}^{n-1}(E_i, E_i) \otimes (E_i, E_i). \tag{5}$$

This is manifestly nonnegative, and agrees with a on T_xM and T_yM as required. This choice gives

$$\mathrm{tr}(A\nabla^2\psi) = \alpha(\varphi)\nabla^2\psi((E_n, -E_n), (E_n, -E_n)) + \beta(\varphi')\sum_{i=1}^{n-1}\nabla^2\psi((E_i, E_i), (E_i, E_i)). \tag{6}$$

For each $i \in \{1, \dots, n-1\}$ let $\gamma_i : (-\varepsilon, \varepsilon) \times [-d/2, d/2] \to M$ be any smooth one-parameter family of curves with $\gamma_i(r, \pm d/2) = \exp_{\gamma(\pm d/2)}(rE_i(\pm d/2))$ for $i = 1, \dots, n-1$, and $\gamma_i(0, s) = \gamma(s)$. Then $d(\exp_x(rE_i), \exp_y(rE_i)) \le L[\gamma_i(r, \cdot)]$, and hence

$$\begin{aligned}\psi(\exp_x(rE_i), \exp_y(rE_i)) &\le v(\exp_x(rE_i), \exp_y(rE_i))\\ &= 2\varphi\left(\frac{d(\exp_x(rE_i), \exp_y(rE_i))}{2}\right)\\ &\le 2\varphi\left(\frac{L[\gamma_i(r, \cdot)]}{2}\right),\end{aligned}$$

since φ is nondecreasing. Since the functions on the left and the right are both smooth functions of r and equality holds for $r = 0$, it follows that

$$\nabla^2\psi((E_i, E_i), (E_i, E_i)) \le 2\frac{d^2}{dr^2}\left(\varphi\left(\frac{L[\gamma_i(r, \cdot)]}{2}\right)\right)\Big|_{r=0}. \tag{7}$$

Similarly, since $d - 2r = L[\gamma|_{[-d/2+r, d/2-r]}] \ge d(\gamma(-d/2+r), \gamma(d/2-r))$, we have

$$\nabla^2\psi(E_n, -E_n), (E_n, -E_n)) \le 2\frac{d^2}{dr^2}\left(\varphi\left(\frac{d}{2} - r\right)\right)\Big|_{r=0} = 2\varphi''\left(\frac{d}{2}\right). \tag{8}$$

Now we make a careful choice of the curves $\gamma_i(r, \cdot)$, motivated by the situation in the model space, in order to get a useful result on the right-hand side in inequality (7): To begin with, if $K > 0$, we assume that $d < \pi/\sqrt{K}$ (we will return to deal with the equality case later). We choose

$$\gamma_i(r, s) = \exp_{\gamma(s)}\left(\frac{r\boldsymbol{C}_\kappa(s)E_i}{\boldsymbol{C}_\kappa(d/2)}\right),$$

where $\boldsymbol{C}_\kappa$ is given by (3). Now we proceed to compute the right-hand side of (7): Denoting s derivatives of γ_i by γ' and r derivatives by $\dot{\gamma}$, we find

$$\begin{aligned}\frac{d}{dr}\Big(L[\gamma_i(r, \cdot)]\Big) &= \frac{d}{dr}\left(\int_{-d/2}^{d/2} \|\gamma'(r, s)\|\, ds\right)\\ &= \int_{-d/2}^{d/2} \frac{\langle\gamma', \nabla_r\gamma'\rangle}{\|\gamma'\|}\, ds.\end{aligned}$$

In particular this gives zero when $r = 0$. Differentiating again, we obtain (using $\|\gamma'(0, s)\| = 1$ and the expression $\dot{\gamma}(0, s) = (\boldsymbol{C}_\kappa(s)/\boldsymbol{C}_\kappa(d/2))E_i$)

$$\frac{d^2}{dr^2}(L[\gamma_i(r, \cdot)])|_{r=0} = \int_{-d/2}^{d/2} \|\nabla_r\gamma'\|^2 - \langle\gamma', \nabla_r\gamma'\rangle^2 + \langle\gamma', \nabla_r\nabla_r\gamma'\rangle\, ds.$$

Now we observe that

$$\nabla_r\gamma' = \nabla_s\dot{\gamma} = \nabla_s\left(\frac{\boldsymbol{C}_\kappa(s)}{\boldsymbol{C}_\kappa(d/2)}E_i\right) = \frac{\boldsymbol{C}_\kappa'(s)}{\boldsymbol{C}_\kappa(d/2)}E_i,$$

while

$$\nabla_r\nabla_r\gamma' = \nabla_r\nabla_s\dot\gamma = \nabla_s\nabla_r\dot\gamma - R(\dot\gamma, \gamma')\dot\gamma = -\frac{\boldsymbol{C}_\kappa(s)^2}{\boldsymbol{C}_\kappa(d/2)^2}R(E_i, E_n)E_i,$$

since by the definition of $\gamma_i(r, s)$ we have $\nabla_r\dot\gamma = 0$. This gives

$$\frac{d^2}{dr^2}(L[\gamma_i(r,\cdot)])|_{r=0} = \frac{1}{\boldsymbol{C}_\kappa(d/2)^2}\int_{-d/2}^{d/2}\{\boldsymbol{C}'_\kappa(s)^2 - \boldsymbol{C}_\kappa(s)^2 R(E_i, E_n, E_i, E_n)\}\,ds.$$

Summing over i from 1 to $n-1$ gives

$$\begin{aligned}\sum_{i=1}^{n-1}\frac{d^2}{dr^2}(L[\gamma_i(r,\cdot)])|_{r=0} &= \frac{1}{\boldsymbol{C}_\kappa(d/2)^2}\int_{-d/2}^{d/2}\left\{(n-1)\boldsymbol{C}'_\kappa(s)^2 - \boldsymbol{C}_\kappa(s)^2\sum_{i=1}^{n-1}R(E_i, E_n, E_i, E_n)\right\}ds\\ &= \frac{1}{\boldsymbol{C}_\kappa(d/2)^2}\int_{-d/2}^{d/2}\{(n-1)\boldsymbol{C}'_\kappa(s)^2 - \boldsymbol{C}_\kappa(s)^2\operatorname{Ric}(E_n, E_n)\}\,ds\\ &\le \frac{n-1}{\boldsymbol{C}_\kappa(d/2)^2}\int_{-d/2}^{d/2}\{\boldsymbol{C}'_\kappa(s)^2 - \kappa\boldsymbol{C}_\kappa(s)^2\}\,ds.\end{aligned}$$

In the case $\kappa = 0$, the integral is zero; in the case $\kappa < 0$, or the case $\kappa > 0$ with $d < \pi/\sqrt{\kappa}$, we have

$$\begin{aligned}\frac{1}{\boldsymbol{C}_\kappa(d/2)^2}\int_{-d/2}^{d/2}\{\boldsymbol{C}'_\kappa(s)^2 - \kappa\boldsymbol{C}_\kappa(s)^2\}\,ds &= \frac{1}{\boldsymbol{C}_\kappa(d/2)^2}\int_{-d/2}^{d/2}(-\kappa\boldsymbol{S}_\kappa\boldsymbol{C}'_\kappa - \kappa\boldsymbol{S}'_\kappa\boldsymbol{C}_\kappa)\,ds\\ &= -\frac{\kappa}{\boldsymbol{C}_\kappa(d/2)^2}\int_{-d/2}^{d/2}(\boldsymbol{C}_\kappa\boldsymbol{S}_\kappa)'\,ds\\ &= -\frac{2\kappa\boldsymbol{C}_\kappa(d/2)\boldsymbol{S}_\kappa(d/2)}{\boldsymbol{C}_\kappa(d/2)^2}\\ &= -2\boldsymbol{T}_\kappa(d/2).\end{aligned}$$

Finally, we have

$$\frac{d}{dr}\left(\varphi\left(\frac{L[\gamma_i(r,\cdot)]}{2}\right)\right)\Big|_{r=0} = \varphi'\frac{d}{dr}\left(\frac{L[\gamma_i(r,\cdot)]}{2}\right)\Big|_{r=0} = 0,$$

and so

$$\begin{aligned}\sum_{i=1}^{n-1}\frac{d^2}{dr^2}\left(\varphi\left(\frac{L[\gamma_i(r,\cdot)]}{2}\right)\right)\Big|_{r=0} &= \sum_{i=1}^{n-1}\left(\varphi'\frac{d^2}{dr^2}\left(\frac{L[\gamma_i(r,\cdot)]}{2}\right)\Big|_{r=0} + \varphi''\left(\frac{d}{dr}\left(\frac{L[\gamma_i(r,\cdot)]}{2}\right)\Big|_{r=0}\right)^2\right)\\ &\le -(n-1)\varphi'\boldsymbol{T}_\kappa|_{d/2}.\end{aligned}$$

Now, using the inequalities (7) and (8), we have from (6) that

$$\mathscr{L}[\nabla^2\psi, \nabla\psi] \le \operatorname{trace}(A\nabla^2\psi) \le 2[\alpha(\varphi')\varphi'' - (n-1)\beta(\varphi')\varphi'\boldsymbol{T}_\kappa]|_{d/2}, \tag{9}$$

as required.

In the case $d = \pi/\sqrt{K}$, we instead choose $\gamma_i(r, s) = \exp_{\gamma(s)}(r\boldsymbol{C}_{\kappa'}(s)E_i/(\boldsymbol{C}_{\kappa'}(d/2)))$, for arbitrary $\kappa' < \kappa$. Then the computation above gives

$$\sum_{i=1}^{n-1} \nabla^2\psi((E_i, E_i), (E_i, E_i)) \leq -2(n-1)\varphi'\boldsymbol{T}_\kappa.$$

Since the right-hand side approaches $-\infty$ as κ' increases to κ, we have a contradiction to the assumption that ψ is smooth. Hence no such ψ exists and there is nothing to prove. □

3. Estimate on the modulus of continuity for solutions of heat equations

In this section we prove Theorem 1, which extends the oscillation estimate from domains in $\mathbb{R}^n$ to compact Riemannian manifolds. The estimate is analogous to [Andrews and Clutterbuck 2009b, Theorem 4.1], the modulus of continuity estimate for the Neumann problem on a convex Euclidean domain.

Proof of Theorem 1. Recall that (M, g) is a compact Riemannian manifold, possibly with boundary (in which case we assume that the boundary is locally convex). Define an evolving quantity, Z, on the product manifold $M \times M \times [0, \infty)$:

$$Z(x, y, t) := u(y, t) - u(x, t) - 2\varphi(d(x, y)/2, t) - \epsilon(1 + t)$$

for small $\epsilon > 0$.

We have assumed that φ is a modulus of continuity for u at $t = 0$, and so $Z(\,\cdot\,, \,\cdot\,, 0) \leq -\epsilon < 0$. Note also that Z is continuous on $M \times M \times [0, \infty)$, and $Z(x, x, t) = -\varepsilon(1 + t) < 0$ for each $x \in M$ and $t \in [0, T)$. It follows that if Z ever becomes positive, there exists a first time $t_0 > 0$ and points $x_0 \neq y_0$ in M such that $Z(x_0, y_0, t_0) = 0$. There are two possibilities: either both x_0 and y_0 are in the interior of M, or at least one of them (say x_0) lies in the boundary ∂M.

We deal with the first case first: Clearly $Z(x, y, t) \leq 0$ for all $x, y \in M$ and $t \in [0, t_0]$. In particular, if we let $v(x, y) = 2\varphi(d(x, y)/2, t_0)$ and $\psi(x, y) = u(y, t_0) - u(x, t_0) - \varepsilon(1 + t_0)$, then

$$\psi(x, y) \leq v(x, y)$$

for all $x, y \in M$, while $\psi(x_0, y_0) = v(x_0, y_0)$. Since ψ is smooth, by Theorem 3 we have

$$\mathscr{L}[\nabla^2\psi, \nabla\psi] \leq 2[\alpha(\varphi')\varphi'' - (n-1)\boldsymbol{T}_\kappa\beta(\varphi')\varphi']|_{d(x_0,y_0)/2}.$$

Now we observe that since the mixed partial derivatives of $\nabla^2\psi$ all vanish, we have for any admissible A in the definition of $\mathscr{L}$ that

$$\operatorname{tr}(A\nabla^2\psi) = (a(Du)^{ij}\nabla_i\nabla_j u)|_{(y_0,t_0)} - (a(Du)^{ij}\nabla_i\nabla_j u)|_{(x_0,t_0)},$$

and therefore

$$\mathscr{L}[\nabla^2\psi, \nabla\psi] = (a(Du)^{ij}\nabla_i\nabla_j u)|_{(y_0,t_0)} - (a(Du)^{ij}\nabla_i\nabla_j u)|_{(x_0,t_0)}.$$

It follows that

$$a(Du)^{ij}\nabla_i\nabla_j u|_{(y_0,t_0)} - a(Du)^{ij}\nabla_i\nabla_j u|_{(x_0,t_0)} \leq 2[\alpha(\varphi')\varphi'' - (n-1)\boldsymbol{T}_\kappa\beta(\varphi')\varphi']|_{d(x_0,y_0)/2}. \tag{10}$$

We also know that the time derivative of Z is nonnegative at (x_0, y_0, t_0), since $Z(x_0, y_0, t) \le 0$ for $t < t_0$:

$$\frac{\partial Z}{\partial t}|_{(x_0,y_0,t_0)} = a(Du)^{ij}\nabla_i\nabla_j u|_{(y_0,t_0)} - a(Du)^{ij}\nabla_i\nabla_j u|_{(x_0,t_0)} - 2\frac{\partial\varphi}{\partial t} - \varepsilon \ge 0. \tag{11}$$

Combining the inequalities (10) and (11), we obtain

$$\frac{\partial\varphi}{\partial t} < \alpha(\varphi')\varphi'' - (n-1)\boldsymbol{T}_\kappa\beta(\varphi')\varphi',$$

where all terms are evaluated at the point $d(x_0, y_0)/2$. This contradicts assumption (ii) in Theorem 1.

Now we consider the second case, where $x_0 \in \partial M$. Under the assumption that ∂M is convex, there exists [Bartolo et al. 2002] a length-minimizing geodesic $\gamma : [0, d] \to M$ from x_0 to y_0, such that $\gamma(s)$ is in the interior of M for $0 < s < d$ and $\gamma'(0)\cdot\nu(x_0) > 0$, where $\nu(x_0)$ is the inward-pointing unit normal to ∂M at x_0. We compute

$$\frac{d}{ds}Z(\exp_{x_0}(s\nu(x_0)), y_0, t_0) = -\nabla_{\nu(x_0)}u - \varphi'(d/2)\nabla d(\nu(x_0), 0) = \varphi'(d/2)\gamma'(0)\cdot\nu(x_0) \ge 0.$$

In particular, $Z(\exp_{x_0}(s\nu(x_0)), y_0, t_0) > 0$ for all small positive s, contradicting the fact that $Z(x, y, t_0) \le 0$ for all $x, y \in M$.

Therefore Z remains negative for all $(x, y) \in M$ and $t \in [0, T)$. Letting ε approach zero proves the theorem. □

4. The eigenvalue lower bound

Now we provide the proof of the sharp lower bound on the first eigenvalue (Theorem 2), which follows very easily from the modulus of continuity estimate from Theorem 1.

Proposition 4. *For M and u as in Theorem 1 applied to the heat equation ($\alpha \equiv \beta \equiv 1$ in (2)), we have the oscillation estimate*

$$|u(y,t) - u(x,t)| \le C\mathrm{e}^{-\mu t},$$

where C depends on the modulus of continuity of $u(\,\cdot\,, 0)$, and μ is the smallest positive eigenvalue of the Sturm–Liouville equation

$$\begin{aligned}\Phi'' - (n-1)\boldsymbol{T}_\kappa\Phi' + \mu\Phi = \frac{1}{\boldsymbol{C}_\kappa^{n-1}}(\Phi'\boldsymbol{C}_\kappa^{n-1})' + \mu\Phi = 0 \quad \text{on } [-D/2, D/2],\\ \Phi'(\pm D/2) = 0.\end{aligned} \tag{12}$$

Proof. The eigenfunction-eigenvalue pair (Φ, μ) is defined as follows: For any $\sigma \in \mathbb{R}$ we define $\Phi_\sigma(x)$ to be the solution of the initial value problem

$$\begin{aligned}\Phi_\sigma'' - (n-1)\boldsymbol{T}_\kappa\Phi_\sigma' + \sigma\Phi_\sigma = 0;\\ \Phi_\sigma(0) = 0;\\ \Phi_\sigma'(0) = 1.\end{aligned}$$

Then $\mu = \sup\{\sigma : x \in [-D/2, D/2] \Longrightarrow \Phi'_\sigma(x) > 0\}$. In particular, for $\sigma < \mu$ the function Φ_σ is strictly increasing on $[-D/2, D/2]$, and $\Phi_\sigma(x)$ is decreasing in σ and converges smoothly to $\Phi(x) = \Phi_\mu(x)$ as σ approaches μ for $x \in (0, D/2]$ and $0 < \sigma < \mu$.

Now we apply Theorem 1: Since Φ is smooth, has positive derivative at $x = 0$ and is positive for $x \in (0, D/2]$, there exists $C > 0$ such that $C\Phi$ is a modulus of continuity for $u(\,.\,, 0)$. Then, for each $\sigma \in (0, \mu)$, $\varphi_0 = C\Phi_\sigma$ is also a modulus of continuity for $u(\,.\,, 0)$, with $\varphi_0(0) = 0$ and $\varphi_0' > 0$. Defining $\varphi(x, t) = C\Phi_\sigma(x)\mathrm{e}^{-\sigma t}$, all the conditions of Theorem 1 are satisfied, and we deduce that $\varphi(\,.\,, t)$ is a modulus of continuity for $u(\,.\,, t)$, each $t \ge 0$. Letting σ approach μ, we deduce that $C\Phi\mathrm{e}^{-\mu t}$ is also a modulus of continuity. That is, for all x, y and $t \ge 0$,

$$|u(y, t) - u(x, t)| \le C\mathrm{e}^{-\mu t}\Phi\left(\frac{d(x, y)}{2}\right) \le C \sup \Phi\mathrm{e}^{-\mu t}. \qquad \square$$

Proof of Theorem 2. Observe that if (φ, λ) is the first eigenfunction-eigenvalue pair, then $u(x, t) = e^{-\lambda t}\varphi(x)$ satisfies the heat equation on M for all $t > 0$. From Proposition 4, we have $|u(y, t) - u(x, t)| \le Ce^{-\mu t}$, and so $|\varphi(y) - \varphi(x)| \le Ce^{-(\mu-\lambda)t}$ for all $x, y \in M$ and $t > 0$. Since φ is nonconstant, letting $t \to \infty$ implies that $\mu - \lambda \le 0$. $\square$

5. Sharpness of the estimates

In the previous section we proved that $\lambda_1(D, \kappa, n) \ge \mu$. To complete the proof of Theorem 2, we must prove that $\lambda_1(D, \kappa, n) \le \mu$. To do this, we construct examples of Riemannian manifolds with given diameter bounds and Ricci curvature lower bounds such that the first eigenvalue is as close as desired to μ. The construction is similar to that given in [Kröger 1992; Bakry and Qian 2000], but we include it here because the construction also produces examples proving that the modulus of continuity estimates of Theorem 1 are sharp.

Fix κ and D, and let $M = S^{n-1} \times [-D/2, D/2]$ with the metric

$$g = ds^2 + a\boldsymbol{C}_\kappa^2(s)\bar{g},$$

where $\bar{g}$ is the standard metric on S^{n-1}, and $a > 0$. The Ricci curvatures of this metric are given by

$$\begin{aligned}
&\mathrm{Ric}(\partial_s, \partial_s) = (n-1)\kappa; \\
&\mathrm{Ric}(\partial_s, v) = 0 && \text{for } v \in TS^{n-1}; \\
&\mathrm{Ric}(v, v) = \left((n-1)\kappa + (n-2)\frac{1/a - \kappa}{\boldsymbol{C}_\kappa^2}\right)|v|^2 && \text{for } v \in TS^{n-1}.
\end{aligned}$$

In particular, the lower Ricci curvature bound $\mathrm{Ric} \ge (n-1)\kappa$ is satisfied for any a if $\kappa \le 0$ and for $a \le 1/\kappa$ if $\kappa > 0$.

To demonstrate the sharpness of the modulus of continuity estimate in Theorem 1, we construct solutions of (1) on M which satisfy the conditions of Theorem 1 and satisfy the conclusion with equality for positive times: Let $\varphi_0 : [0, D/2]$ be as given in Theorem 1, and extend by odd reflection to $[-D/2, D/2]$ and define φ to be the solution of the initial-boundary value problem

$$\begin{aligned}\frac{\partial\varphi}{\partial t}&=\alpha(\varphi')\varphi''+(n-1)\boldsymbol{T}_\kappa\beta(\varphi')\varphi';\\ \varphi(x,0)&=\varphi_0(x);\\ \varphi'(\pm D/2,t)&=0.\end{aligned}$$

Now define $u(z,s,t)=\varphi(s,t)$ for $s\in[-D/2,D/2]$, $z\in S^{n-1}$, and $t\geq 0$. Then a direct calculation shows that u is a solution of (1) on M. If φ_0 is concave on $[0,D/2]$, we have $|\varphi_0(a)-\varphi_0(b)|\leq 2\varphi_0(|b-1|/2)$ for all a and b in $[-D/2,D/2]$. For our choice of φ, this also remains true for positive times. Note also that for any $w,z\in S^{n-1}$ and $a,b\in[-D/2,D/2]$ we have $d((w,a),(z,b))\geq|b-a|$. Therefore we have

$$|u(w,a,t)-u(z,b,t)|=|\varphi(a,t)-\varphi(b,t)|\leq 2\varphi\left(\frac{|b-a|}{2},t\right)\leq 2\varphi\left(\frac{d((w,a),(z,b))}{2},t\right),$$

so that $\varphi(\,.\,,t)$ is a modulus of continuity for $u(\,.\,,t)$ as claimed. Furthermore, this holds with equality whenever $w=z$ and $b=-a$, so there is no smaller modulus of continuity and the estimate is sharp.

Now we proceed to the sharpness of the eigenvalue estimate. On the manifold constructed above, we have an explicit eigenfunction of the Laplacian, given by $\varphi(z,s)=\Phi(s)$, where Φ is the first eigenfunction of the one-dimensional Sturm–Liouville problem given in Proposition 4. That is, we have $\lambda_1(M,g)\leq\mu$. In this example we have the required Ricci curvature lower bound, and the diameter approaches D as $a\to 0$. Since μ depends continuously on D, the result follows.

A slightly more involved construction shows that the bound is sharp even in the smaller class of manifolds without boundary. This is achieved by smoothly attaching spherical caps to the ends of the above examples; see the similar construction in [Andrews and Ni 2012, Section 2].

6. Implications for the "Li conjecture"

In this section we mention some implications of the sharp eigenvalue estimate and a conjecture attributed to Peter Li. The result of Lichnerowicz [1958] is that $\lambda_1\geq n\kappa$ whenever $\mathrm{Ric}\geq(n-1)\kappa g_{ij}$ (so that, by the Bonnet–Myers estimate, $D\leq\pi/\sqrt{\kappa}$). An estimate from [Zhong and Yang 1984] gives $\lambda_1\geq\pi^2/D^2$ for $\mathrm{Ric}\geq 0$. Both of these are sharp, and the latter estimate should also be sharp as $D\to 0$ for any lower Ricci curvature bound. Interpolating linearly (in κ) between these estimates, we obtain Li's conjecture:

$$\lambda_1\geq\frac{\pi^2}{D^2}+(n-1)\kappa.$$

By construction this holds precisely at the endpoints $\kappa\to 0$ and $\kappa\to\pi^2/D^2$.

Several previous attempts to prove such inequalities have been made, particularly towards proving inequalities of the form $\lambda_1\geq\pi^2/D^2+a\kappa$ for some constant a, which are linear in κ and have the correct limit as $\kappa\to 0$. These include works of DaGang Yang [1999], Jun Ling [2006] and Ling and Lu [2010], the latter showing that $\alpha=\frac{34}{100}$ holds. These are all superseded by the result of Shi and Zhang [2007] which proves $\lambda_1\geq\sup_{s\in(0,1)}\{4s(1-s)\pi^2/D^2+(n-1)s\kappa\}$, so in particular $\lambda_1\geq\pi^2/D^2+((n-1)/2)\kappa$ by taking $s=\frac{1}{2}$.

We remark here that the inequality with $a = (n-1)/2$ is the best possible of this kind, and in particular the Li conjecture is false. This can be seen by computing an asymptotic expansion for the sharp lower bound μ given by Theorem 2. For fixed $D = \pi$ we perturb about $\kappa = 0$ (as in [Andrews and Ni 2012, Section 4]), obtaining

$$\mu = 1 + \frac{(n-1)}{2}\kappa + O(\kappa^2).$$

By scaling, this amounts to the estimate

$$\mu = \frac{\pi^2}{D^2} + \frac{(n-1)}{2}\kappa + O(\kappa D^2).$$

Since the lower bound $\lambda_1 \geq \mu$ is sharp, this shows that the inequality $\lambda_1 \geq \pi^2/D^2 + a\kappa$ is false for any $a > (n-1)/2$, and in particular for $a = n-1$.

References

[Andrews and Clutterbuck 2009a] B. Andrews and J. Clutterbuck, "Lipschitz bounds for solutions of quasilinear parabolic equations in one space variable", *J. Differential Equations* **246**:11 (2009), 4268–4283. MR 2010j:35237 Zbl 1171.35060

[Andrews and Clutterbuck 2009b] B. Andrews and J. Clutterbuck, "Time-interior gradient estimates for quasilinear parabolic equations", *Indiana Univ. Math. J.* **58**:1 (2009), 351–380. MR 2010k:35234 Zbl 1173.35030

[Andrews and Clutterbuck 2011] B. Andrews and J. Clutterbuck, "Proof of the fundamental gap conjecture", *J. Amer. Math. Soc.* **24**:3 (2011), 899–916. MR 2012d:35051 Zbl 1222.35130

[Andrews and Ni 2012] B. Andrews and L. Ni, "Eigenvalue comparison on Bakry–Emery manifolds", *Comm. Partial Differential Equations* **37**:11 (2012), 2081–2092. MR 3005536 Zbl 1258.35153

[Bakry and Qian 2000] D. Bakry and Z. Qian, "Some new results on eigenvectors via dimension, diameter, and Ricci curvature", *Adv. Math.* **155**:1 (2000), 98–153. MR 2002g:58048 Zbl 0980.58020

[Bartolo et al. 2002] R. Bartolo, A. Germinario, and M. Sánchez, "Convexity of domains of Riemannian manifolds", *Ann. Global Anal. Geom.* **21**:1 (2002), 63–83. MR 2003a:58016 Zbl 0995.58009

[Chen and Wang 1994] M. F. Chen and F. Y. Wang, "Application of coupling method to the first eigenvalue on manifold", *Sci. China Ser. A* **37**:1 (1994), 1–14. MR 96d:58141 Zbl 0799.53044

[Kröger 1992] P. Kröger, "On the spectral gap for compact manifolds", *J. Differential Geom.* **36**:2 (1992), 315–330. MR 94g:58236 Zbl 0738.58048

[Li 1979] P. Li, "A lower bound for the first eigenvalue of the Laplacian on a compact manifold", *Indiana Univ. Math. J.* **28**:6 (1979), 1013–1019. MR 81j:58084 Zbl 0429.35054

[Li and Yau 1980] P. Li and S. T. Yau, "Estimates of eigenvalues of a compact Riemannian manifold", pp. 205–239 in *Geometry of the Laplace operator* (Honolulu, HI, 1979), edited by R. Osserman and A. Weinstein, Proc. Sympos. Pure Math. **36**, Amer. Math. Soc., Providence, R.I., 1980. MR 81i:58050 Zbl 0441.58014

[Lichnerowicz 1958] A. Lichnerowicz, *Géométrie des groupes de transformations*, Travaux et Recherches Mathématiques **3**, Dunod, Paris, 1958. MR 23 #A1329 Zbl 0096.16001

[Ling 2006] J. Ling, "The first eigenvalue of a closed manifold with positive Ricci curvature", *Proc. Amer. Math. Soc.* **134**:10 (2006), 3071–3079. MR 2007d:58057 Zbl 1094.53032

[Ling and Lu 2010] J. Ling and Z. Lu, "Bounds of eigenvalues on Riemannian manifolds", pp. 241–264 in *Trends in partial differential equations* (Hangzhou, 2008), edited by B. Bian et al., Adv. Lect. Math. **10**, International Press, Somerville, MA, 2010. MR 2011k:58048 Zbl 1200.58020

[Ni 2013] L. Ni, "Estimates on the modulus of expansion for vector fields solving nonlinear equations", *J. Math. Pures Appl.* (9) **99**:1 (2013), 1–16. MR 3003280 Zbl 06136919

[Shi and Zhang 2007] Y. M. Shi and H. C. Zhang, “Lower bounds for the first eigenvalue on compact manifolds”, *Chinese Ann. Math. Ser. A* **28**:6 (2007), 863–866. MR 2009b:58074 Zbl 1164.53369

[Yang 1999] D. Yang, “Lower bound estimates of the first eigenvalue for compact manifolds with positive Ricci curvature”, *Pacific J. Math.* **190**:2 (1999), 383–398. MR 2001b:53039 Zbl 1016.58016

[Zhong and Yang 1984] J. Q. Zhong and H. C. Yang, “On the estimate of the first eigenvalue of a compact Riemannian manifold”, *Sci. Sinica Ser. A* **27**:12 (1984), 1265–1273. MR 87a:58162 Zbl 0561.53046

Received 1 Apr 2012. Accepted 21 May 2013.

BEN ANDREWS: *Mathematical Sciences Institute, Australian National University, Building 27, Canberra ACT 0200, Australia*

and

Mathematical Sciences Center, Tsinghua University
`ben.andrews@anu.edu.au`

JULIE CLUTTERBUCK: *Mathematical Sciences Institute, Australian National University, Canberra ACT 0200, Australia*
`Julie.Clutterbuck@anu.edu.au`

ANALYSIS AND PDE
Vol. 6, No. 5, 2013
dx.doi.org/10.2140/apde.2013.6.1025

SOME MINIMIZATION PROBLEMS IN THE CLASS OF CONVEX FUNCTIONS WITH PRESCRIBED DETERMINANT

NAM Q. LE AND OVIDIU SAVIN

We consider minimizers of linear functionals of the type

$$L(u)=\int_{\partial\Omega}u\,d\sigma-\int_{\Omega}u\,dx$$

in the class of convex functions u with prescribed determinant $\det D^2u=f$.

We obtain compactness properties for such minimizers and discuss their regularity in two dimensions.

1. Introduction

In this paper, we consider minimizers of certain linear functionals in the class of convex functions with prescribed determinant. We are motivated by the study of convex minimizers u for convex energies E of the type

$$E(u)=\int_{\Omega}F(\det D^2u)\,dx+L(u),\quad \text{with } L \text{ a linear functional,}$$

which appear in the work of Donaldson [2002; 2009] in the context of existence of Kähler metrics of constant scalar curvature for toric varieties. The minimizer u solves a fourth-order elliptic equation with two nonstandard boundary conditions involving the second- and third-order derivatives of u (see (1-4) below). In this paper, we consider minimizers of L (or E) in the case when the determinant $\det D^2u$ is prescribed. This allows us to understand better the type of boundary conditions that appear in such problems and to obtain estimates also for unconstrained minimizers of E.

The simplest minimization problem with prescribed determinant which is interesting in its own right is

$$\text{minimize}\int_{\partial\Omega}u\,d\sigma,\quad \text{with } u\in\mathcal{A}_0,$$

where Ω is a bounded convex set, $d\sigma$ is the surface measure of $\partial\Omega$, and $\mathcal{A}_0$ is the class of nonnegative solutions to the Monge–Ampère equation $\det D^2u=1$:

$$\mathcal{A}_0:=\{u:\bar{\Omega}\to[0,\infty)\mid u \text{ convex},\ \det D^2u=1\}.$$

Question. *Is the minimizer u smooth up to the boundary* $\partial\Omega$ *if* Ω *is a smooth, say uniformly convex, domain?*

MSC2010: primary 35J96; secondary 35J66.
Keywords: boundary regularity, convex minimizer, fourth-order elliptic equation, prescribed determinant.

In the present paper, we answer this question affirmatively in dimensions $n = 2$. First, we remark that the minimizer must vanish at x_0, the center of mass of $\partial\Omega$:

$$x_0 = \fint_{\partial\Omega} x \, d\sigma.$$

This follows easily since

$$u(x) - u(x_0) - \nabla u(x_0)(x - x_0) \in \mathcal{A}_0$$

and

$$\int_{\partial\Omega} \big[u(x) - u(x_0) - \nabla u(x_0)(x - x_0)\big] \, d\sigma = \int_{\partial\Omega} [u - u(x_0)] \, d\sigma \le \int_{\partial\Omega} u \, d\sigma,$$

with strict inequality if $u(x_0) > 0$. Thus we can reformulate the problem above as minimizing

$$\int_{\partial\Omega} u \, d\sigma - \mathcal{H}^{n-1}(\partial\Omega) u(x_0)$$

in the set of all solutions to the Monge–Ampère equation $\det D^2 u = 1$ which are not necessarily nonnegative. This formulation is more convenient since we can now perturb functions in all directions.

More generally, we consider linear functionals of the type

$$L(u) = \int_{\partial\Omega} u \, d\sigma - \int_{\Omega} u \, dA,$$

with $d\sigma$, dA nonnegative Radon measures supported on $\partial\Omega$ and Ω respectively. In this paper, we study the existence, uniqueness and regularity properties for minimizers of L, that is,

$$\text{minimize } L(u) \text{ for all } u \in \mathcal{A} \tag{P}$$

in the class $\mathcal{A}$ of subsolutions (solutions) to a Monge–Ampère equation $\det D^2 u \ge f$:

$$\mathcal{A} := \big\{u : \overline{\Omega} \to \mathbb{R} \;\big|\; u \text{ convex}, \ \det D^2 u \ge f\big\}.$$

Notice that we are minimizing a linear functional L over a convex set $\mathcal{A}$ in the cone of convex functions.

Clearly, the minimizer of the problem (P) satisfies $\det D^2 u = f$ in Ω. Otherwise we can find $v \in \mathcal{A}$ such that $v = u$ in a neighborhood of $\partial\Omega$, and $v \ge u$ in Ω with strict inequality in some open subset, and thus $L(v) < L(u)$.

We assume throughout that the following 5 conditions are satisfied:

(1) Ω is a bounded, uniformly convex, $C^{1,1}$ domain.

(2) f is bounded away from 0 and ∞.

(3) $d\sigma = \sigma(x) \, d\mathcal{H}^{n-1} \lfloor \partial\Omega$, with the density $\sigma(x)$ bounded away from 0 and ∞.

(4) $dA = A(x) \, dx$ in a small neighborhood of $\partial\Omega$, with the density $A(x)$ bounded from above.

(5) $L(u) > 0$ for all u convex but not linear.

The last condition is known as the *stability* of L (see [Donaldson 2002]), and in two dimensions, is equivalent to saying that for all linear functions l, we have

$$L(l)=0 \quad \text{and} \quad L(l^+)>0 \quad \text{if } l^+\not\equiv 0 \text{ in } \Omega,$$

where $l^+=\max(l,0)$ (see Proposition 2.4).

Notice that the stability of L implies that $L(l)=0$ for any linear function l, and hence $d\sigma$ and dA must have the same mass and the same center of mass.

A minimizer u of the functional L is determined up to linear functions, since both L and $\mathcal{A}$ are invariant under addition with linear functions. We "normalize" u by subtracting its tangent plane at, say, the center of mass of Ω. In Section 2, we shall prove in Proposition 2.5 that there exists a unique normalized minimizer to the problem (P).

We also prove a compactness theorem for minimizers.

Theorem 1.1 (compactness). *Let u_k be the normalized minimizers of the functionals L_k with data $(f_k, d\sigma_k, dA_k, \Omega)$ that has uniform bounds in k. Precisely, the inequalities* (2-1) *and* (2-4) *below are satisfied uniformly in k and $\rho\le f_k\le \rho^{-1}$. If*

$$f_k \rightharpoonup f,\ d\sigma_k \rightharpoonup d\sigma,\ dA_k \rightharpoonup dA,$$

then $u_k\to u$ uniformly on compact sets of Ω, where u is the normalized minimizer of the functional L with data $(f, d\sigma, dA, \Omega)$.

If u is a minimizer, then the Euler–Lagrange equation reads (see Proposition 3.6)

$$\text{if } \varphi:\Omega\to\mathbb{R} \text{ solves } U^{ij}\varphi_{ij}=0,\ \text{ then } L(\varphi)=0,$$

where U^{ij} are the entries of the cofactor matrix U of the Hessian D^2u. Since the linearized Monge–Ampère equation is also an equation in divergence form, we can always express the Ω-integral of a function φ in terms of a boundary integral. For this, we consider the solution v to the Dirichlet problem

$$U^{ij}v_{ij}=-dA \text{ in } \Omega, \quad v=0 \text{ on } \partial\Omega.$$

Integrating by parts twice and using $\partial_i(U^{ij})=\partial_j(U^{ij})=0$, we can compute

$$\begin{aligned}\int_\Omega \varphi\, dA &= -\int_\Omega \varphi\, U^{ij}v_{ij} = \int_\Omega \varphi_i\, U^{ij}v_j - \int_{\partial\Omega}\varphi U^{ij}v_j\nu_i\\ &= -\int_\Omega (U^{ij}\varphi_{ij})v + \int_{\partial\Omega}\varphi_i U^{ij} v\nu_j - \int_{\partial\Omega}\varphi U^{ij}v_j\nu_i = -\int_{\partial\Omega}\varphi U^{ij}v_i\nu_j.\end{aligned} \tag{1-1}$$

From the Euler–Lagrange equation, we obtain

$$U^{ij}v_i\nu_j=-\sigma \quad \text{on } \partial\Omega.$$

Since $v=0$ on $\partial\Omega$, we have $v_i=v_\nu\nu_i$, and hence

$$U^{ij}v_i\nu_j=U^{ij}\nu_i\nu_jv_\nu=U^{\nu\nu}v_\nu=\big(\det D^2_{x'}u\big)v_\nu,$$

with $x' \perp \nu$ denoting the tangential directions along $\partial\Omega$. In conclusion, if u is a smooth minimizer, then there exists a function v such that (u, v) solves the system

$$\begin{cases} \det D^2 u = f & \text{in } \Omega, \\ U^{ij} v_{ij} = -dA & \text{in } \Omega, \\ v = 0 & \text{on } \partial\Omega, \\ U^{\nu\nu} v_\nu = -\sigma & \text{on } \partial\Omega. \end{cases} \tag{1-2}$$

This system is interesting since the function v above satisfies two boundary conditions, Dirichlet and Neumann, while u has no boundary conditions. Heuristically, the boundary values for u can be recovered from the term $U^{\nu\nu} = \det D^2_{x'} u$, which appears in the Neumann boundary condition for v.

Our main regularity results for the minimizers u are in two dimensions.

Theorem 1.2. *Assume that $n = 2$, and the conditions (1)–(5) hold. If $\sigma \in C^\alpha(\partial\Omega)$, $f \in C^\alpha(\overline{\Omega})$, and $\partial\Omega \in C^{2,\alpha}$, then the minimizer $u \in C^{2,\alpha}(\overline{\Omega})$ and the system* (1-2) *holds.*

We obtain Theorem 1.2 by showing that u separates quadratically on $\partial\Omega$ from its tangent planes, and then we apply the boundary Hölder gradient estimates for v which were obtained in [Le and Savin 2013].

As a consequence of Theorem 1.2, we obtain higher regularity if the data $(f, d\sigma, dA, \Omega)$ is more regular.

Theorem 1.3. *Assume that $n = 2$ and the conditions (1)–(5) hold. If $\sigma \in C^\infty(\partial\Omega)$, $f \in C^\infty(\overline{\Omega})$, $A \in C^\infty(\overline{\Omega})$, and $\partial\Omega \in C^\infty$, then $u \in C^\infty(\overline{\Omega})$.*

In Section 6, we provide an example of Pogorelov type for a minimizer in dimensions $n \geq 3$ that shows that Theorem 1.3 does not hold in this generality in higher dimensions.

We explain briefly how Theorem 1.3 follows from Theorem 1.2. If $u \in C^{2,\alpha}(\overline{\Omega})$, then $U^{ij} \in C^\alpha(\overline{\Omega})$, and Schauder estimates give $v \in C^{2,\alpha}(\overline{\Omega})$, and thus $v_\nu \in C^{1,\alpha}(\partial\Omega)$. From the last equation in (1-2) we obtain $U^{\nu\nu} = \det D^2_{x'} u \in C^{1,\alpha}(\partial\Omega)$. This implies $u \in C^{3,\alpha}(\partial\Omega)$, and from the first equation in (1-2), we find $u \in C^{3,\alpha}(\overline{\Omega})$. We can repeat the same argument and obtain that $u \in C^{k,\alpha}$ for any $k \geq 2$.

As we mentioned above, our constraint minimization problem is motivated by the minimization of the Mabuchi energy functional from complex geometry in the case of toric varieties

$$M(u) = \int_\Omega -\log\det D^2 u + \int_{\partial\Omega} u \, d\sigma - \int_\Omega u \, dA.$$

In this case, $d\sigma$ and dA are canonical measures on $\partial\Omega$ and Ω. Minimizers of M satisfy the following fourth-order equation, called Abreu's equation [1998]:

$$u^{ij}_{ij} := \sum_{i,j=1}^{n} \frac{\partial^2 u^{ij}}{\partial x_i \partial x_j} = -A,$$

where u^{ij} are the entries of the inverse matrix of $D^2 u$. This equation and the functional M have been studied extensively by Donaldson [2002; 2005; 2008; 2009]; see also [Zhou and Zhu 2008]. In Donaldson's papers, the domain Ω was taken to be a polytope $P \subset \mathbb{R}^n$ and A was taken to be a positive constant. The existence of smooth solutions with suitable boundary conditions has important implications in complex geometry. It says that we can find Kähler metrics of constant scalar curvature for toric varieties.

More generally, one can consider minimizers of the convex functional

$$E(u)=\int_\Omega F(\det D^2u)+\int_{\partial\Omega} u\,d\sigma-\int_\Omega u\,dA, \tag{1-3}$$

where $F(t^n)$ is a convex and decreasing function of $t\geq 0$. The Mabuchi energy functional corresponds to $F(t)=-\log t$, whereas in our minimization problem (P) (with $f\equiv 1$),

$$F(t)=\begin{cases}\infty & \text{if } t<1,\\ 0 & \text{if } t\geq 1.\end{cases}$$

Minimizers of E satisfy a system similar to (1-2):

$$\begin{cases}-F'(\det D^2u)=v & \text{in } \Omega,\\ U^{ij}v_{ij}=-dA & \text{in } \Omega,\\ v=0 & \text{on } \partial\Omega,\\ U^{\nu\nu}v_\nu=-\sigma & \text{on } \partial\Omega.\end{cases} \tag{1-4}$$

A similar system but with different boundary conditions was investigated by Trudinger and Wang [2008a]. If the function F is strictly decreasing, then we see from the first and third equations above that $\det D^2u=\infty$ on $\partial\Omega$, and therefore we cannot expect minimizers to be smooth up to the boundary (as is the case with the Mabuchi functional $M(u)$).

If F is constant for large values of t (as in the case we considered), then $\det D^2u$ becomes finite on the boundary and smoothness up to the boundary is expected. More precisely, assume that

$$F\in C^{1,1}((0,\infty)),\quad G(t):=F(t^n) \text{ is convex in } t,\quad \text{and}\quad G'(0^+)=-\infty,$$

and there exists $t_0>0$ such that

$$F(t)=0 \text{ on } [t_0,\infty),\qquad F''(t)>0 \text{ on } (0,t_0].$$

Theorem 1.4. *Assume $n=2$ and the conditions (1)–(5) and the above hypotheses on F are satisfied. If $\sigma\in C^\alpha(\partial\Omega)$, $A\in C^\alpha(\overline{\Omega})$, and $\partial\Omega\in C^{2,\alpha}$, then the normalized minimizer u of the functional E defined in* (1-3) *satisfies $u\in C^{2,\alpha}(\overline{\Omega})$, and the system* (1-4) *holds in the classical sense.*

The paper is organized as follows. In Section 2, we discuss the notion of stability for the functional L and prove existence, uniqueness and compactness of minimizers of the problem (P). In Section 3, we state a quantitative version of Theorem 1.2, Proposition 3.1, and we also obtain the Euler–Lagrange equation. Proposition 3.1 is proved in Sections 4 and 5, first under the assumption that the density A is bounded from below and then in the general case. In Section 6, we give an example of a singular minimizer in dimension $n\geq 3$. Finally, in Section 7, we prove Theorem 1.4.

2. Stability inequality and existence of minimizers

Let Ω be a bounded convex set and define

$$L(u)=\int_{\partial\Omega} u\,d\sigma-\int_{\Omega} u\,dA$$

for all convex functions $u:\overline{\Omega}\to\mathbb{R}$ with $u\in L^1(\partial\Omega, d\sigma)$. We assume that

$$\sigma\geq\rho \text{ on } \partial\Omega \text{ and } A(x)\leq\rho^{-1} \text{ in a neighborhood of } \partial\Omega \tag{2-1}$$

for some small $\rho>0$ and that L is stable, that is,

$$L(u)>0 \text{ for all } u \text{ convex but not linear.} \tag{2-2}$$

Assume for simplicity that 0 is the center of mass of Ω. We notice that (2-2) implies $L(l)=0$ for any l linear, since l can be approximated by both convex and concave functions. We “normalize” a convex function by subtracting its tangent plane at 0, and this does not change the value of L. First we prove some lower semicontinuity properties of L with respect to normalized solutions.

Lemma 2.1 (lower semicontinuity). *Assume that* (2-1) *holds and* (u_k) *is a normalized sequence that satisfies*

$$\int_{\partial\Omega} u_k\,d\sigma\leq C,\qquad u_k\to u \textit{ uniformly on compact sets of } \Omega, \tag{2-3}$$

for some function $u:\Omega\to\mathbb{R}$. *Let* $\bar{u}$ *be the minimal convex extension of* u *to* $\overline{\Omega}$, *that is,*

$$\bar{u}=u \textit{ in } \Omega,\qquad \bar{u}(x)=\lim_{t\to 1^-} u(tx) \textit{ if } x\in\partial\Omega.$$

Then

$$\int_{\Omega} u\,dA=\lim\int_{\Omega} u_k\,dA,\qquad \int_{\partial\Omega}\bar{u}\,d\sigma\leq\liminf\int_{\partial\Omega} u_k\,d\sigma,$$

and thus

$$L(\bar{u})\leq\liminf L(u_k).$$

Remark. The upper graph of the function $\bar{u}$ is the closure of the upper graph of u in $\mathbb{R}^{n+1}$.

Proof. Since u_k are normalized, they are increasing on each ray out of the origin. For each $\eta>0$ small, we consider the set $\Omega_\eta:=\{x\in\Omega:\operatorname{dist}(x,\partial\Omega)<\eta\}$, and from (2-1) we obtain

$$\int_{\Omega_\eta} u_k\,dA\leq C\rho^{-1}\eta\int_{\partial\Omega} u_k\,d\sigma\leq C\eta.$$

Since this inequality holds for all small $\eta\to 0$, we easily obtain

$$\int_{\Omega} u\,dA=\lim\int_{\Omega} u_k\,dA.$$

For each $z\in\partial\Omega$ and $t<1$ we have $u_k(tz)\leq u_k(z)$. We let $k\to\infty$ in the inequality

$$\int_{\partial\Omega} u_k(tz)\,d\sigma\leq\int_{\partial\Omega} u_k(z)\,d\sigma$$

and obtain

$$\int_{\partial\Omega} u(tz)\, d\sigma \leq \liminf \int_{\partial\Omega} u_k(z)\, d\sigma,$$

and then we let $t \to 1^-$:

$$\int_{\partial\Omega} \bar{u}\, d\sigma \leq \liminf \int_{\partial\Omega} u_k\, d\sigma. \qquad \square$$

Remark 2.2. From the proof we see that if we are given functionals L_k with measures σ_k, A_k that satisfy (2-1) uniformly in k and

$$\sigma_k \rightharpoonup \sigma, \quad A_k \rightharpoonup A,$$

and if (2-3) holds for a sequence u_k, then the statement still holds; that is,

$$L(\bar{u}) \leq \liminf L_k(u_k).$$

By compactness, one can obtain a quantitative version of (2-2) known as *stability inequality*. This was done by Donaldson [2002, Proposition 5.2.2]. For completeness, we sketch its proof here.

Proposition 2.3. *Assume that* (2-1) *and* (2-2) *hold. Then we can find* $\mu > 0$ *such that*

$$L(u) := \int_{\partial\Omega} u d\sigma - \int_{\Omega} u dA \geq \mu \int_{\partial\Omega} u d\sigma \tag{2-4}$$

for all convex functions u normalized at 0.

Proof. Assume the conclusion does not hold; then there is a sequence of normalized convex functions (u_k) with

$$\int_{\partial\Omega} u_k d\sigma = 1, \quad \lim L(u_k) = 0,$$

and thus

$$\lim \int_{\Omega} u_k dA = 1.$$

Using convexity, we may assume that u_k converges uniformly on compact subsets of Ω to a limiting function $u \geq 0$. Let $\bar{u}$ be the minimal convex extension of u to $\overline{\Omega}$. Then, from Lemma 2.1, we obtain

$$L(\bar{u}) = 0, \quad \int_{\Omega} \bar{u}\, dA = 1,$$

and thus $\bar{u} \geq 0$ is not linear and we contradict (2-2). $\square$

Donaldson [2002, Proposition 5.3.1] showed that when $n = 2$, the stability condition can be checked easily.

Proposition 2.4. *Assume that* $n = 2$, *that* (2-1) *holds, and that for all linear functions l we have*

$$L(l) = 0 \quad \text{and} \quad L(l^+) > 0 \quad \text{if } l^+ \not\equiv 0 \text{ in } \Omega, \tag{2-5}$$

where $l^+ = \max(l, 0)$. *Then L is stable*; *that is, condition* (2-2) *is satisfied.*

Proof. For completeness, we sketch the proof. Assume by contradiction that $L(u) \le 0$ for some convex function u which is not linear in Ω. Let u^* be the convex envelope generated by the boundary values of $\bar{u}$ — the minimal convex extension of u to $\overline{\Omega}$. Notice that $u^* = \bar{u}$ on $\partial\Omega$. Since $L(u^*) \le L(\bar{u}) \le L(u)$, we find $L(u^*) \le 0$. Notice that u^* is not linear, since otherwise $0 = L(u^*) < L(\bar{u}) \le 0$ (we used that $\bar{u}$ is not linear). After subtracting a linear function, we may assume that u^* is normalized and u^* is not identically 0.

We obtain a contradiction by showing that u^* satisfies the stability inequality. By our hypotheses, there exists $\mu > 0$ small such that

$$L(l^+) \ge \mu \int_{\partial\Omega} l^+ \, d\sigma$$

for any l^+. Indeed, by (2-1), this inequality is valid if the "crease" $\{l = 0\}$ is near $\partial\Omega$, and for all other l's, it follows by compactness from (2-5). We approximate from below u^* by u^*_k, which is defined as the maximum of the tangent planes of u^* at some points $y_i \in \Omega$, $i = 1, \dots, k$. Since u^* is a convex envelope in two dimensions, u^*_k is a discrete sum of l^+'s, and hence it satisfies the stability inequality. Now we let $k \to \infty$; since $u^*_k \le u^*$, using Lemma 2.1, we obtain that u^* also satisfies the stability inequality. □

Proposition 2.5. *Assume that* (2-1) *and* (2-2) *hold. Then there exists a unique* (*up to linear functions*) *minimizer u of L subject to the constraint*

$$u \in \mathcal{A} := \{ v : \overline{\Omega} \to \mathbb{R} \mid v \text{ convex}, \ \det D^2 v \ge f \},$$

where $\rho \le f \le \rho^{-1}$ for some $\rho > 0$. The minimizer satisfies $\det D^2 u = f$, *and if $n = 2$, it is unique* (*up to linear functions*).

Proof. Let (u_k) be a sequence of normalized solutions such that $L(u_k) \to \inf_{\mathcal{A}} L$. By the stability inequality, we see that $\int_{\partial\Omega} u_k \, d\sigma$ are uniformly bounded, and after passing to a subsequence, we may assume that u_k converges uniformly on compact subsets of Ω to a function u. Then $u \in \mathcal{A}$, and from the lower semicontinuity we see that $L(u) = \inf_{\mathcal{A}} L$, that is, u is a minimizer. Notice that $\det D^2 u = f$. Indeed, if a quadratic polynomial P with $\det D^2 P > f$ touches u strictly from below at some point $x_0 \in \Omega$, in a neighborhood of x_0, then we can replace u in this neighborhood by $\max\{P + \epsilon, u\} \in \mathcal{A}$, and the energy decreases.

Next we assume w is another minimizer. We use the strict concavity of $M \mapsto \log(\det D^2 M)$ in the space of positive symmetric matrices M, and obtain that for almost every x where u, w are twice differentiable,

$$\log\det D^2\left(\frac{u+w}{2}\right)(x) \ge \tfrac{1}{2}\log\det D^2 u(x) + \tfrac{1}{2}\log\det D^2 w(x) \ge \log f(x).$$

This implies $(u + w)/2 \in \mathcal{A}$ is also a minimizer and $D^2 u = D^2 w$ almost everywhere in Ω. Since f is bounded above and below, we know that $u, w \in W^{2,1}_{\text{loc}}$ (see [De Philippis and Figalli 2013]) in the open set Ω' where both u, w are strictly convex. This gives that $u - w$ is linear on each connected component of Ω'. If $n = 2$, then $\Omega' = \Omega$, and hence $u - w$ is linear. □

Remark. Uniqueness is expected to hold in any dimension. For this one needs to show that the set of strict convexity of a solution to the Monge–Ampère equation is always connected.

Remark. The arguments above show that the stability condition is also necessary for the existence of a minimizer. Indeed, if u is a minimizer and $L(u_0)=0$ for some convex function u_0 that is not linear, then $u+u_0$ is also a minimizer and we contradict the uniqueness.

Proof of Theorem 1.1. We assume that the data $(f_k, d\sigma_k, dA_k, \Omega)$ satisfies (2-1), (2-4) uniformly in k and $\rho \le f_k \le \rho^{-1}$. For each k, let w_k be the convex solution to $\det D^2 w_k = f_k$ in Ω with $w_k = 0$ on $\partial\Omega$. Since f_k are bounded from above, we find $w_k \ge -C$, and so by the minimality of u_k,

$$L_k(u_k) \le L_k(w_k) \le C.$$

It follows from the stability inequality that

$$\int_{\partial\Omega} u_k \, d\sigma_k \le C,$$

and we may assume, after passing to a subsequence, that $u_k \to u$ uniformly on compact subsets of Ω.

We need to show that u is a minimizer for L with data $(f, d\sigma, dA, \Omega)$. For this it suffices to prove that for any continuous $v : \overline{\Omega} \to \mathbb{R}$ which solves $\det D^2 v = f$ in Ω, we have $L(u) \le L(v)$.

Let v_k be the solution to $\det D^2 v_k = f_k$ with boundary data $v_k = v$ on $\partial\Omega$. Using appropriate barriers, it is standard to check that $f_k \rightharpoonup f$, $f_k \le \rho^{-1}$ implies $v_k \to v$ uniformly in $\overline{\Omega}$. Then we let $k \to \infty$ in $L_k(u_k) \le L_k(v_k)$, use Remark 2.2, and obtain

$$L(u) \le \liminf L_k(u_k) \le \lim L_k(v_k) = L(v),$$

which finishes the proof. $\square$

3. Preliminaries and the Euler–Lagrange equation

We rewrite our main hypotheses in a quantitative way. We assume that for some small $\rho > 0$, we have:

(H1) The curvatures of $\partial\Omega$ are bounded from below by ρ and from above by ρ^{-1}.

(H2) $\rho \le f \le \rho^{-1}$.

(H3) $d\sigma = \sigma(x)\, d\mathcal{H}^{n-1} \lfloor \partial\Omega$, with $\rho \le \sigma(x) \le \rho^{-1}$.

(H4) $dA = A(x)\, dx$ in a small neighborhood

$$\Omega_\rho := \{x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) < \rho\}$$

of $\partial\Omega$ with $A(x) \le \rho^{-1}$.

(H5) For any convex function u normalized at the center of mass of Ω, we have

$$L(u) := \int_{\partial\Omega} u \, d\sigma - \int_{\Omega} u \, dA \ge \rho \int_{\partial\Omega} u \, d\sigma.$$

We denote by c, C positive constants depending on ρ, and their values may change from line to line whenever there is no possibility of confusion. We refer to such constants as *universal constants*.

Our main theorem, Theorem 1.2, follows from the next proposition, which deals with less regular data.

Proposition 3.1. *Assume that $n = 2$ and that conditions* (H1)–(H5) *hold.*

(i) *Then the minimizer u obtained in Proposition 2.5 satisfies $u \in C^{1,\beta}(\overline{\Omega}) \cap C^{1,1}(\partial\Omega)$ for some universal $\beta \in (0, 1)$ and u separates quadratically from its tangent planes on $\partial\Omega$, that is,*

$$C^{-1}|x-y|^2 \le u(y) - u(x) - \nabla u(x)(y-x) \le C|x-y|^2 \quad \textit{for all } x, y \in \partial\Omega,$$

for some $C > 0$ universal.

(ii) *If in addition $\sigma \in C^{\alpha}(\partial\Omega)$, then $u\,|_{\partial\Omega} \in C^{2,\gamma}(\partial\Omega)$ with $\gamma := \min\{\alpha, \beta\}$, and*

$$\|u\|_{C^{2,\gamma}(\partial\Omega)} \le C\|\sigma\|_{C^{\gamma}(\partial\Omega)}.$$

We remark that in part (ii), we obtain $u \in C^{2,\gamma}(\partial\Omega)$ even though f and A are assumed to be only L^{∞}.

Proof that Proposition 3.1 implies Theorem 1.2. Theorem 7.3 of [Savin 2013] states that a solution to the Monge–Ampère equation which separates quadratically from its tangent planes on the boundary satisfies the classical C^{α}-Schauder estimates. Thus, if the assumptions of Proposition 3.1(ii) are satisfied and $f \in C^{\alpha}(\overline{\Omega})$, then $u \in C^{2,\gamma}(\overline{\Omega})$ with its $C^{2,\gamma}$ norm bounded by a constant C depending on ρ, α, $\|\sigma\|_{C^{\alpha}(\partial\Omega)}$, $\|\partial\Omega\|_{C^{2,\alpha}}$, and $\|f\|_{C^{\alpha}(\overline{\Omega})}$. This implies that the system (1-2) holds. If $\alpha \le \beta$, then we are done. If $\alpha > \beta$, then we use $v_{\nu} \in C^{\alpha}(\partial\Omega)$ in the last equation of the system and obtain $u \in C^{2,\alpha}(\partial\Omega)$, which gives $u \in C^{2,\alpha}(\overline{\Omega})$. □

We prove Proposition 3.1 in the next two sections. Part (ii) follows from part (i) and the boundary Harnack inequality for the linearized Monge–Ampère equation, which was obtained in [Le and Savin 2013, Theorem 2.4]. This theorem states that if a solution to the Monge–Ampère equation with bounded right-hand side separates quadratically from its tangent planes on the boundary, then the classical boundary estimate of Krylov holds for solutions of the associated linearized equation.

In order to simplify the ideas, we prove the proposition in the case when the hypotheses (H1), (H2), (H4) are replaced by

(H1′) $\Omega = B_1$.

(H2′) $f \in C^{\infty}(\overline{\Omega})$, $\rho \le f \le \rho^{-1}$.

(H4′) $dA = A(x)\,dx$ with $\rho \le A(x) \le \rho^{-1}$ in Ω and $A \in C^{\infty}(\Omega)$.

We use (H1′) only for simplicity of notation. We will see from the proofs that the same arguments carry to the general case. We use (H2′) so that D^2u is continuous in Ω and the linearized Monge–Ampère equation is well defined. Our estimates do not depend on the smoothness of f, and thus the general case follows by approximation from Theorem 1.1. Later, in Section 5, we show that (H4′) can be replaced by (H4), that is, the bound for A from below is not needed.

First, we establish a result on uniform modulus of convexity for minimizers of L in two dimensions.

Proposition 3.2. *Let u be a minimizer of L that satisfies the hypotheses above. Then, for any $\delta < 1$, there exist $c(\delta) > 0$ depending on ρ, δ such that*

$$x \in B_{1-\delta} \quad \Longrightarrow \quad S_h(x) \Subset B_1 \ \textit{ if } h \le c(\delta),$$

where $S_h(x)$ *denotes the section of* u *centered at* x *at height* h:

$$S_h(x) = \{y \in \bar{B}_1 : u(y) < u(x) + \nabla u(x)(y-x) + h\}.$$

Although this result is well known (see [Trudinger and Wang 2008b, Remark 3.2] for example), we include its proof here for completeness.

Proof. Without loss of generality, assume u is normalized in B_1, that is, $u \geq 0$, $u(0) = 0$. From the stability inequality (2-4), we obtain

$$\int_{\partial B_1} u\,dx \leq C.$$

This integral bound and the convexity of u imply

$$|u|, |Du| \leq C(\delta) \text{ in } B_{1-\delta/2},$$

for any $\delta < 1$. We show that our statement follows from these bounds. Assume by contradiction that the conclusion is not true. Then we can find a sequence of convex functions u_k satisfying the bounds above such that

$$u_k(y_k) \leq u_k(x_k) + \nabla u_k(x_k)(y_k - x_k) + h_k \tag{3-1}$$

for sequences $x_k \in B_{1-\delta}$, $y_k \in \partial B_{1-\delta/2}$ and $h_k \to 0$. Because Du_k is uniformly bounded, after passing to a subsequence if necessary, we may assume

$$u_k \to u_* \text{ uniformly on } \overline{B}_{1-\delta/2}, \quad x_k \to x_*, \quad y_k \to y_*.$$

Moreover, u_* satisfies $\rho \leq \det D^2 u_* \leq \rho^{-1}$, and

$$u_*(y_*) = u_*(x_*) + \nabla u_*(x_*)(y_* - x_*),$$

that is, the graph of u_* contains a straight line in the interior. However, any subsolution v to $\det D^2 v \geq \rho$ in two dimensions does not have this property and we reach a contradiction. □

Since $f \in C^\alpha$, we obtain that $u \in C^{2,\alpha}(B_1)$, and thus the linearized Monge–Ampère equation is well defined in B_1. The next lemma deals with general linear elliptic equations in B_1 which may become degenerate as we approach ∂B_1.

Lemma 3.3. *Let* $\mathscr{L}v := a^{ij}(x)v_{ij}$ *be a linear elliptic operator with continuous coefficients* $a^{ij} \in C^\alpha(B_1)$ *that satisfy the ellipticity condition* $(a^{ij}(x))_{ij} > 0$ *in* B_1. *Given a continuous boundary data* φ, *there exists a unique solution* $v \in C(\overline{B}_1) \cap C^2(\Omega)$ *to the Dirichlet problem*

$$\mathscr{L}v = 0 \text{ in } B_1, \quad v = \varphi \text{ on } \partial B_1.$$

Proof. For each small δ, we consider the standard Dirichlet problem for uniformly elliptic equations $\mathcal{L}v_\delta = 0$ in $B_{1-\delta}$, $v_\delta = \varphi$ on $\partial B_{1-\delta}$. Since v_δ satisfies the comparison principle with linear functions, it follows that the modulus of continuity of v_δ at points on the boundary $\partial B_{1-\delta}$ depends only on the modulus of continuity of φ. Thus, from the maximum principle, we see that v_δ converges uniformly to a solution v of the Dirichlet problem above. The uniqueness of v follows from the standard comparison principle. □

Remark 3.4. The modulus of continuity of v at points on ∂B_1 depends only on the modulus of continuity of φ.

Remark 3.5. If $\mathcal{L}_m$ is a sequence of operators satisfying the hypotheses of Lemma 3.3 with $a_m^{ij} \to a^{ij}$ uniformly on compact subsets of B_1 and $\mathcal{L}_m v_m = 0$ in B_1, $v_m = \varphi$ on ∂B_1, then $v_m \to v$ uniformly in $\overline{B}_1$.

Indeed, since v_m have a uniform modulus of continuity on ∂B_1 and, for all large m, a uniform modulus of continuity in any ball $B_{1-\delta}$, we see that we can always extract a uniform convergent subsequence in $\overline{B}_1$. Now it is straightforward to check that the limiting function v satisfies $\mathcal{L}v = 0$ in the viscosity sense.

Next, we establish an integral form of the Euler–Lagrange equations for the minimizers of L.

Proposition 3.6. *Assume that u is the normalized minimizer of L in the class $\mathcal{A}$. If $\varphi \in C^2(\Omega) \cap C^0(\overline{\Omega})$ is a solution to the linearized Monge–Ampère equation*

$$U^{ij}\varphi_{ij} = 0 \text{ in } \Omega,$$

then

$$L(\varphi) := \int_{\partial\Omega} \varphi \, d\sigma - \int_\Omega \varphi \, dA = 0.$$

Proof. Consider the solution $u_\epsilon = u + \epsilon\varphi_\epsilon$ to

$$\begin{cases} \det D^2 u_\epsilon = f & \text{in } B_1, \\ u_\epsilon = u + \epsilon\varphi & \text{on } \partial B_1. \end{cases}$$

Since φ_ϵ satisfies the comparison principle and comparison with planes, its existence follows as in Lemma 3.3 by solving the Dirichlet problems in $B_{1-\delta}$ and then letting $\delta \to 0$.

In B_1, φ_ϵ satisfies

$$0 = \frac{1}{\epsilon}\bigl(\det D^2 u_\epsilon - \det D^2 u\bigr) = \frac{1}{\epsilon}\int_0^1 \frac{d}{dt} \det D^2(u + t\epsilon\varphi_\epsilon)\, dt = a_\epsilon^{ij}\partial_{ij}\varphi_\epsilon,$$

where (a_ϵ^{ij}) is the integral from 0 to 1 of the cofactor matrix of $D^2(u + t\epsilon\varphi_\epsilon)$, that is,

$$(a_\epsilon^{ij})_{ij} = \int_0^1 \det D^2(u + t\epsilon\varphi_\epsilon)\bigl(D^2(u + t\epsilon\varphi_\epsilon)\bigr)^{-1}\, dt.$$

Because u is strictly convex in two dimensions and $u_\epsilon \to u$ uniformly on $\overline{B}_1$, $D^2 u_\epsilon \to D^2 u$ uniformly on compact sets of B_1. Thus, as $\epsilon \to 0$, $a_\epsilon^{ij} \to U^{ij}$ uniformly on compact sets of B_1 and by Remark 3.5, we find $\varphi_\epsilon \to \varphi$ uniformly in $\overline{B}_1$. By the minimality of u, we find

$$0 \le \lim_{\epsilon \to 0^+} \frac{1}{\epsilon}\big(L(u_\epsilon) - L(u)\big) = \int_{\partial B_1} \varphi \, d\sigma - \int_{B_1} \varphi \, dA.$$

By replacing φ with $-\varphi$, we obtain the opposite inequality. □

4. Proof of Proposition 3.1

In this section, we prove Proposition 3.1 where (H1′), (H2′) and (H4′) are satisfied. Given a convex function $u \in C^\infty(B_1)$ (not necessarily a minimizer of L) with $\rho \le \det D^2 u \le \rho^{-1}$, we let v be the solution to the Dirichlet problem

$$U^{ij} v_{ij} = -A \text{ in } B_1 \quad v = 0 \text{ on } \partial B_1. \tag{4-1}$$

Notice that $\Psi := C(1 - |x|^2)$ is an upper barrier for v if C is large enough, since

$$U^{ij}\Psi_{ij} \le -C \operatorname{tr} U \le -C(\det D^2 U)^{1/n} = -C(\det D^2 u)^{(n-1)/n} \le -C\rho^{(n-1)/n} \le -A,$$

and hence

$$0 \le v(x) \le C(1 - |x|^2) \sim \operatorname{dist}(x, \partial B_1). \tag{4-2}$$

As in Lemma 3.3, the function v is the uniform limit of the corresponding v_δ that solve the Dirichlet problem in $B_{1-\delta}$. Indeed, since v_δ also satisfies (4-2), we see that

$$|v_{\delta_1} - v_{\delta_2}|_{L^\infty} \le C \max\{\delta_1, \delta_2\}.$$

Let φ be the solution of the homogeneous problem

$$U^{ij}\varphi_{ij} = 0 \text{ in } B_1, \quad \varphi = l^+ \text{ on } \partial B_1,$$

where $l^+ = \max\{0, l\}$ for some linear function $l = b + \nu \cdot x$ of slope $|\nu| = 1$. Denote by $\mathcal{S} := \overline{B}_1 \cap \{l = 0\}$ the segment of intersection of the crease of l with $\overline{B}_1$. Then:

Lemma 4.1.
$$\int_{B_1} \varphi \, dA = \int_{B_1} l^+ \, dA + \int_{\mathcal{S}} u_{\tau\tau} v \, d\mathcal{H}^1,$$

where τ is the unit vector in the direction of $\mathcal{S}$, and hence $\tau \perp \nu$.

Proof. It suffices to show the equality in the case when $u \in C^\infty(\overline{B}_1)$. The general case follows by writing the identity in $B_{1-\delta}$ with v_δ (which increases as δ decreases), and then letting $\delta \to 0$.

Let $\tilde{l}_\epsilon$ be a smooth approximation of l^+ with

$$D^2 l_\epsilon \rightharpoonup \nu \otimes \nu \, dH^1 \lfloor \mathcal{S} \text{ as } \epsilon \to 0,$$

and let φ_ϵ solve the corresponding Dirichlet problem with boundary $\tilde{l}_\epsilon$. Then we integrate by parts and use $\partial_i U^{ij} = 0$:

$$\begin{aligned}\int_{B_1} (\varphi_\epsilon - \tilde{l}_\epsilon)\, dA &= -\int_{B_1} (\varphi_\epsilon - \tilde{l}_\epsilon) U^{ij} v_{ij}\, dx = \int_{B_1} \partial_i (\varphi_\epsilon - \tilde{l}_\epsilon) U^{ij} v_j\, dx \\ &= -\int_{B_1} \partial_{ij} (\varphi_\epsilon - \tilde{l}_\epsilon) U^{ij} v\, dx = \int_{B_1} U^{ij} \partial_{ij} \tilde{l}_\epsilon v\, dx.\end{aligned}$$

We let $\epsilon \to 0$ and obtain

$$\int_{B_1} (\varphi - l^+)\, dA = \int_{\mathcal{S}} U^{\nu\nu} v\, d\mathcal{H}^1,$$

which is the desired conclusion, since $U^{\nu\nu} = u_{\tau\tau}$. □

From Lemma 4.1 and Proposition 3.6, we obtain:

Corollary 4.2. *If u is a minimizer of L in the class $\mathcal{A}$, then*

$$\int_{\mathcal{S}} u_{\tau\tau} v\, d\mathcal{H}^1 = \int_{\partial B_1} l^+\, d\sigma - \int_{B_1} l^+\, dA.$$

The hypotheses on σ and A imply that if the segment $\mathcal{S}$ has length $2h$ with $h \le h_0$ small, universal then

$$ch^3 \le \int_{\mathcal{S}} u_{\tau\tau} v\, d\mathcal{H}^1 \le Ch^3,$$

for some c, C universal.

Lemma 4.3. *Let X_1 and X_2 be the endpoints of the segment $\mathcal{S}$ defined as above. Then*

$$\int_{\mathcal{S}} u_{\tau\tau} (1 - |x|^2)\, d\mathcal{H}^1 = 4h \left(\frac{u(X_1) + u(X_2)}{2} - ⨍_{\mathcal{S}} u\, d\mathcal{H}^1 \right), \tag{4-3}$$

where $2h$ denotes the length of $\mathcal{S}$.

Proof. Again we may assume that $u \in C^2(\overline{B}_1)$, since the general case follows by approximating B_1 by $B_{1-\delta}$. Assume for simplicity that $\tau = e_1$. Then

$$\int_{\mathcal{S}} u_{\tau\tau} (1 - |x|^2)\, d\mathcal{H}^1 = \int_{-h}^{h} \partial_t^2 u(t, a)(h^2 - t^2)\, dt$$

for some fixed a, and integrating by parts twice, we obtain (4-3). □

We remark that the right-hand side in (4-3) represents twice the area between the segment with end points $(X_1, u(X_1))$, $(X_2, u(X_2))$ and the graph of u above $\mathcal{S}$.

Definition 4.4. We say that u admits a tangent plane at a point $z \in \partial B_1$ if there exists a linear function l_z such that

$$x_{n+1} = l_z(x)$$

is a supporting hyperplane for the graph of u at $(z, u(z))$ but for any $\epsilon > 0$,

$$x_{n+1} = l_z(x) - \epsilon z \cdot (x - z)$$

is not a supporting hyperplane. We call l_z a tangent plane for u at z.

Remark 4.5. Notice that if $\det D^2 u \le C$, then the set of points where u admits a tangent plane is dense in ∂B_1. Indeed, using standard barriers, it is not difficult to check that any point on ∂B_1 where the boundary data $u|_{\partial B_1}$ admits a quadratic polynomial from below satisfies the definition above. In the definition above, we assumed $u = \bar{u}$ on ∂B_1 with $\bar{u}$ defined as in the Lemma 2.1; therefore $u|_{\partial B_1}$ is lower semicontinuous.

Assume that u admits a tangent plane at z, and define

$$\tilde{u} = u - l_z.$$

Lemma 4.6. *There exists $\eta > 0$ small, universal such that the section*

$$\tilde{S}_z := \{x \in \overline{B}_1 \mid \tilde{u} < \eta(x-z)\cdot(-z)\}$$

satisfies

$$\tilde{S}_z \subset B_1 \setminus B_{1-\rho}, \quad |\tilde{S}_z| \ge c,$$

for some small c universal.

Proof. We notice that (4-3) is invariant under additions with linear functions. We apply it to $\tilde{u}$ with $X_1 = z$, $X_2 = x$ and use $\tilde{u} \ge 0$, $\tilde{u}(z) = 0$ together with (4-2) and Corollary 4.2 to obtain

$$\tilde{u}(x) \ge c|x-z|^2, \qquad x \in \partial B_1 \cap B_{h_0}(z).$$

From the uniform strict convexity of $\tilde{u}$, which was obtained in Proposition 3.2, we find that the inequality above holds for all $x \in \partial B_1$ for possibly a different value of c. Thus, by choosing η sufficiently small, we obtain

$$\tilde{S}_z \subset B_1, \qquad \tilde{S}_z \cap B_{1-\rho} = \varnothing,$$

where the second statement follows also from Proposition 3.2.

Next we show that $|\tilde{S}_z|$ cannot be arbitrarily small. Otherwise, by the uniform strict convexity of $\tilde{u}$, we obtain that $\tilde{S}_z \subset B_{\epsilon^4}(z)$ for some small $\epsilon > 0$. Assume for simplicity of notation that $z = -e_2$. Then the function

$$w := \eta(x_2+1) + \frac{\epsilon}{2}x_1^2 + \frac{1}{2\rho\epsilon}(x_2+1)^2 - 2\epsilon(x_2+1)$$

is a lower barrier for $\tilde{u}$ in $B_1 \cap B_{\epsilon^4}(z)$. Indeed, notice that if ϵ is sufficiently small, then

$$w \le \eta(x_2+1) \le \tilde{u} \text{ on } \partial(B_1 \cap B_{\epsilon^4}(z)), \qquad \det D^2 w = \rho^{-1} \ge \det D^2 \tilde{u}.$$

In conclusion, $\tilde{u} \ge w \ge (\eta/2)(x_2+1)$ and we contradict that $x_{n+1} = 0$ is a tangent plane for $\tilde{u}$ at z. □

Lemma 4.7. *Let u be the normalized minimizer of L. Then $\|u\|_{C^{0,1}(\overline{B}_1)} \le C$, and u admits tangent planes at all points of ∂B_1. Also, u separates at least quadratically from its tangent planes, that is,*

$$u(x) \ge l_z(x) + c|x-z|^2 \quad \textit{for all } x, z \in \partial B_1.$$

Proof. Let z be a point on ∂B_1 where u admits a tangent plane l_z. From the previous lemma, we know that u satisfies the quadratic separation inequality at z and also that $\tilde{u} = u - l_z$ is bounded from above and below in $\tilde{S}_z$, that is,

$$|u - l_z| \le C \text{ in } \tilde{S}_z.$$

We obtain

$$\int_{\tilde{S}_z} |l_z|\, dx - C \le \int_{\tilde{S}_z} u\, dx \le \int_{B_1} u\, dx \le C \int_{\partial B_1} u\, d\sigma \le C,$$

and since $\tilde{S}_z \subset B_1$ has measure bounded from below, we find

$$l_z(z),\ |\nabla l_z| \le C.$$

By Remark 4.5, this holds for almost every $z \in \partial B_1$ and, by approximation, we find that any point in ∂B_1 admits a tangent plane that satisfies the bounds above. This also shows that u is Lipschitz and the lemma is proved. ☐

Lemma 4.8. *The function v satisfies the lower bound*

$$v(x) \ge c \ \mathrm{dist}(x, \partial B_1),$$

for some small c universal.

Proof. Let $z \in \partial B_1$ and let l be a linear functional with

$$l(x) = l_z(x) - b\, z \cdot (x - z), \quad \text{for some } 0 \le b \le \eta,$$

where l_z denotes a tangent plane at z. We consider all sections

$$S = \{x \in \overline{B}_1 \mid u < l\}$$

which satisfy

$$\inf_S (u - l) \le -c_0,$$

for some appropriate c_0 small, universal. We denote the collection of such sections $\mathcal{M}_z$. From Lemma 4.6, we see that $\mathcal{M}_z \neq \varnothing$ since $\tilde{S}_z$ (or $b = \eta$) satisfies the property above. Notice also that $S \subset \tilde{S}_z \subset B_1$ and $z \in \partial S$. For any section $S \in \mathcal{M}_z$, we consider its center of mass z^S, and from the property above we see that $z^S \in B_{1-c}$ for some small $c > 0$ universal.

First, we show that the lower bound for v holds on the segment $[z, z^S]$. Indeed, since

$$U^{ij}[c(l-u)]_{ij} = -2c \det D^2 u \ge -2c\rho^{-1} \ge -A = U^{ij} v_{ij}$$

and $c(l - u) \le 0 = v$ on ∂B_1, we conclude that

$$c(l-u)^+ \le v \text{ in } B_1. \tag{4-4}$$

Now we use the convexity of u and the fact that the property of S implies $(u - l)(z^S) < -c$, and conclude that

$$v(x) \ge c(l-u)(x) \ge c|x - z| \ge c\, \mathrm{dist}(x, \partial B_1) \quad \text{for all } x \in [z, z^S].$$

Now it remains to prove that the collection of segments $[z, z^S]$, $z \in \partial B_1$, $S \in \mathcal{M}_z$ cover a fixed neighborhood of ∂B_1. To this aim, we show that the multivalued map

$$z \in \partial B_1 \mapsto F(z) := \{z^S \mid S \in \mathcal{M}_z\}$$

has the following properties:

(1) the map F is *closed* in the sense that

$$z_n \to z_* \text{ and } z_n^{S_n} \to y_* \Rightarrow y^* \in F(z_*);$$

(2) $F(z)$ is a connected set for any z.

The first property follows easily from the following facts: z^S varies continuously with the linear map l that defines $S = \{u < l\}$; and if $l_{z_n} \to l_*$, then $l_* \le l_{z_*}$ for some tangent plane l_{z_*}.

To prove the second property, we notice that if we increase continuously the value of the parameter b (which defines l) up to η, then all the corresponding sections also belong to $\mathcal{M}_z$. This means that in $F(z)$ we can continuously connect z^S with $z^{\tilde{S}_z}$ for some section $\tilde{S}_z$. On the other hand, the set of all possible $z^{\tilde{S}_z}$ is connected, since the set l_z of all tangent planes at z is connected in the space of linear functions.

Since $F(z) \subset B_{1-c}$, it follows that for all $\delta < c$, the intersection map

$$z \mapsto G_\delta(z) = \{[z, y] \cap \partial B_{1-\delta} \mid y \in F(z)\}$$

also has properties (1) and (2) above. Now it is easy to check that the image of G_δ covers the whole $\partial B_{1-\delta}$, and hence the collection of segments $[z, z^S]$ covers $B_1 \setminus B_{1-c}$ and the lemma is proved. □

Now we are ready to prove the first part of Proposition 3.1.

Proof of Proposition 3.1(i). In Lemma 4.7, we obtained the quadratic separation from below for $\tilde{u} = u - l_z$. Next we show that $\tilde{u}$ separates at most quadratically on ∂B_1 in a neighborhood of z.

Assume for simplicity of notation that $z = -e_2$. We apply (4-3) to $\tilde{u}$ with $X_1 = (-h, a)$, $X_2 = (h, a)$, and then use Corollary 4.2 and Lemma 4.8 to obtain

$$\frac{\tilde{u}(X_1) + \tilde{u}(X_2)}{2} - \fint_{\mathcal{S}} \tilde{u} \le Ch^2.$$

On the other hand, for small h, the segment $[z, z^{\tilde{S}_z}]$ intersects $[X_1, X_2]$ at a point $y = (t, a)$ with $|t| \le Ch^2 \le h/2$. Moreover, since $y \in \tilde{S}_z$, we have $\tilde{u}(y) \le \eta(a+1) \le Ch^2$. On the segment $[X_1, X_2]$, $\tilde{u}$ satisfies the conditions of Lemma 4.9 which we prove below, and hence

$$\tilde{u}(X_1), \tilde{u}(X_2) \le Ch^2.$$

In conclusion, u separates quadratically on ∂B_1 from its tangent planes and therefore satisfies the hypotheses of the Localization Theorem in [Savin 2013; Le and Savin 2013]. From [Le and Savin 2013, Theorem 2.4 and Proposition 2.6], we conclude that

$$\|u\|_{C^{1,\beta}(\bar{B}_1)},\ \|v\|_{C^\beta(\bar{B}_1)},\ \|v_\nu\|_{C^\beta(\partial B_1)} \le C, \tag{4-5}$$

for some $\beta < 1$, C universal. □

Lemma 4.9. *Let $f : [-h, h] \to \mathbb{R}^+$ be a nonnegative convex function such that*

$$\frac{f(-h)+f(h)}{2} - \frac{1}{2h}\int_{-h}^{h} f(x)\,dx \le Mh^2, \qquad f(t) \le Mh^2,$$

for some $t \in [-h/2, h/2]$. Then

$$f(\pm h) \le Ch^2$$

for some C depending on M.

Proof. The inequality above states that the area between the line segment with end points $(-h, f(-h))$, $(h, f(h))$ and the graph of f is bounded by $2Mh^3$. By convexity, this area is greater than the area of the triangle with vertices $(-h, f(-h))$, $(t, f(t))$, $(h, f(h))$. Now the inequality of the heights $f(\pm h)$ follows from elementary euclidean geometry. ☐

Finally, we are ready to prove the second part of Proposition 3.1.

Proof of Proposition 3.1 (*ii*). Let φ be such that

$$U^{ij}\varphi_{ij} = 0 \text{ in } B_1, \quad \varphi \in C^{1,1}(\partial B_1) \cap C^0(\overline{B}_1).$$

Since u satisfies the quadratic separation assumption and f is smooth up to the boundary, we obtain from [Le and Savin 2013, Theorem 2.5 and Proposition 2.6]

$$\|v\|_{C^{1,\beta}(\overline{B}_1)},\ \|\varphi\|_{C^{1,\beta}(\overline{B}_1)} \le K, \quad \text{and} \quad |U^{ij}| \le K|\log\delta|^2 \text{ on } B_{1-\delta},$$

for some constant K depending on ρ, $\|f\|_{C^\beta(\overline{B}_1)}$, and $\|\varphi\|_{C^{1,1}(\partial B_1)}$.

We will use the following identity in two dimensions:

$$U^{ij}v_j\nu_i = U^{\tau\nu}v_\tau + U^{\nu\nu}v_\nu.$$

Integrating by parts twice, we obtain, as in (1-1),

$$\int_{B_{1-\delta}} \varphi\,dA = -\int_{B_{1-\delta}} \varphi\,U^{ij}v_{ij}\,dx = \int_{\partial B_{1-\delta}} \varphi_i U^{ij} v\nu_j - \int_{\partial B_{1-\delta}} \varphi U^{ij} v_j\nu_i = -\int_{\partial B_{1-\delta}} \varphi U^{\nu\nu} v_\nu + o(1),$$

where in the last equality we used the estimates

$$|v| \le C\delta, \quad |v_\tau| \le K\delta^\beta, \quad |\varphi|, |\nabla\varphi| \le K, \quad U^{ij} \le K|\log\delta|^2 \quad \text{on } \partial B_{1-\delta}.$$

Since on ∂B_r

$$U^{\nu\nu} = u_{\tau\tau} = r^{-2}u_{\theta\theta} + r^{-1}u_\nu,$$

$u \in C^{1,\beta}(\overline{B}_1)$ and $u(re^{i\theta})$ converges uniformly as $r \to 1$, and $u_{\theta\theta}$ is uniformly bounded from below, we obtain

$$U^{\nu\nu}\,d\mathcal{H}^1\lfloor_{\partial B_r} \rightharpoonup (u_{\theta\theta} + u_\nu)\,d\mathcal{H}^1\lfloor_{\partial B_1} \quad \text{as } r \to 1.$$

We let $\delta \to 0$ in the equality above and find

$$\int_{B_1} \varphi\,dA = -\int_{\partial B_1} \varphi\,(u_{\theta\theta} + u_\nu)v_\nu\,d\mathcal{H}^1.$$

Now the Euler–Lagrange equation, Proposition 3.6, gives

$$(u_{\theta\theta}+u_\nu)v_\nu=-\sigma \quad \text{on } \partial B_1.$$

We use that $\|v_\nu\|_{C^\beta(\partial B_1)}\le C$ and, from Lemma 4.8, $v_\nu\le -c$ on ∂B_1 and obtain

$$\|u\|_{C^{2,\gamma}(\partial B_1)}\le C\|\sigma\|_{C^\gamma(\partial B_1)}. \qquad \square$$

5. The general case for A

In this section, we remove the assumptions that A is bounded from below by ρ in B_1 and we also assume that A is bounded from above only in a neighborhood of the boundary. Precisely, we assume that $A\ge 0$ in B_1 and $A\le\rho^{-1}$ in $B_1\backslash\overline{B}_{1-\rho}$. We may also assume A is smooth in B_1, since the general case follows by approximation. Notice that $\int_{B_1}A\,dx$ is bounded from above and below since it equals $\int_{\partial B_1}d\sigma$.

Let v be the solution of the Dirichlet problem

$$U^{ij}v_{ij}=-A, \qquad v=0 \text{ on } \partial B_1. \tag{5-1}$$

In Section 4, we used that A is bounded from above when we obtained $v\le C(1-|x|^2)$, and we used that A is bounded from below in Lemma 4.8 (see (4-4)). We need to show that these bounds for v also hold in a neighborhood of ∂B_1 under the weaker hypotheses above. First, we show:

Lemma 5.1. *$v\le C$ on $\partial B_{1-\rho/2}$ and $v\ge c(\delta)$ on $B_{1-\delta}$, with C universal and $c(\delta)>0$ depending also on δ.*

Proof. As before, we may assume that $u\in C^\infty(\overline{B}_1)$, since the general case follows by approximating B_1 by $B_{1-\epsilon}$.

We multiply the equation in (5-1) by $(1-|x|^2)$, integrate by parts twice, and obtain

$$\int_{B_1}2v\operatorname{tr}U\,dx=\int_{B_1}A(x)(1-|x|^2)\,dx\le C,$$

and since $\operatorname{tr}U\ge c$, we obtain

$$\int_{B_1}v\,dx\le C.$$

We know this:

(1) $v\ge 0$ solves a linearized Monge–Ampère equation with bounded right-hand side in $B_1\setminus B_{1-\rho}$.

(2) u has a uniform modulus of convexity on compact sets of B_1.

Now we use the Harnack inequality of Caffarelli and Gutierrez [1997] and conclude that

$$\sup_{\mathcal{V}}v\le C(\inf_{\mathcal{V}}v+1), \qquad \mathcal{V}:=B_{1-\rho/4}\setminus\overline{B}_{1-3\rho/4},$$

and the integral inequality above gives $\sup_{\mathcal{V}}v\le C$.

Next we prove the lower bound. We multiply the equation in (5-1) by $\varphi\in C_0^\infty(B_1)$ with

$$\varphi=0 \text{ if } |x|\ge 1-\delta/2, \quad \varphi=1 \text{ in } B_{1-\delta}, \quad \|D^2\varphi\|\le C/\delta^2,$$

integrate by parts twice, and obtain

$$C(\delta)\int_{\mathcal{U}} v \operatorname{tr} U \geq -\int_{B_1} v\, U^{ij}\varphi_{ij} = \int_{B_1} A\varphi \geq c, \qquad \mathcal{U} := B_{1-\delta/2}\setminus \overline{B}_{1-\delta},$$

where the last inequality holds provided that δ is sufficiently small. Since u is normalized, we obtain (see Proposition 3.2) $|\nabla u| \leq C(\delta)$ in $\mathcal{U}$, and thus

$$\int_{\mathcal{U}} \operatorname{tr}\, U = \int_{\mathcal{U}} \triangle u = \int_{\partial\mathcal{U}} u_\nu \leq C(\delta).$$

The last two inequalities imply $\sup_{\mathcal{U}} v \geq c(\delta)$, and hence there exists $x_0 \in \mathcal{U}$ such that $v(x_0) \geq c(\delta)$. We use (1), (2) above and the Harnack inequality and find $v \geq c(\delta)$ in $B_{\bar\delta}(x_0)$ for some small $\bar\delta$ depending on ρ and δ. Since v is a supersolution, that is, $U^{ij}v_{ij} \leq 0$, we can apply the weak Harnack inequality of Caffarelli and Gutierrez [1997, Theorem 4]. From property (2) above, we see that we can extend the lower bound of v from $B_{\bar\delta}(x_0)$ all the way to $\mathcal{U}$, and by the maximum principle, this bound holds also in $B_{1-\delta/2}$. □

The upper bound in Lemma 5.1 gives as in (4-2) the upper bound for v in a neighborhood of ∂B_1, that is,

$$v(x) \leq C(1-|x|^2) \quad \text{on} \quad B_1 \setminus B_{1-\rho/2}.$$

This implies, as in Section 4, that Lemma 4.7 holds, that is, u separates at least quadratically from its tangent planes on ∂B_1. It remains to show that also Lemma 4.8 holds. Since A is not strictly positive, $c(l-u)$ is no longer a subsolution for the equation (5-1) and we cannot bound v below as we did in (4-4). In the next lemma, we construct another barrier which allows us to bound v from below on the segment $[z, z^S]$.

Lemma 5.2. *Let $\tilde u : B_1 \to \mathbb{R}$ be a convex function with $\tilde u \in C(\overline{B}_1)\cap C^2(B_1)$, and*

$$\rho \leq \det D^2\tilde u \leq \rho^{-1}.$$

Assume that the section $S := \{\tilde u < 0\}$ is included in B_1 and is tangent to ∂B_1 at a point $z \in \partial B_1$, and also that

$$\inf_S \tilde u \leq -\mu,$$

for some $\mu > 0$. If

$$\tilde U^{ij} v_{ij} \leq 0 \text{ in } B_1, \quad v \geq 0 \text{ on } \partial B_1,$$

then

$$v(x) \geq c(\mu,\rho)|x-z| \inf_{S'} v \quad \text{for all } x \in [z, z^S], \qquad S' := \left\{\tilde u \leq \tfrac{1}{2}\inf_S \tilde u\right\},$$

where z^S denotes the center of mass of S and $c(\mu,\rho)$ is a positive constant depending on μ and ρ.

The functions $\tilde u = u - l$ and v in the proof of Lemma 4.8 satisfy the lemma above, if η in Lemma 4.6 is small, universal. Using also the lower bound on v from Lemma 5.1, we find

$$v \geq c|x-z| \text{ on } [z, z^S],$$

for some c universal, and the rest of the proof of Lemma 4.8 follows as before. This shows that Proposition 3.1 holds also with our assumptions on the measure A.

Proof of Lemma 5.2. We construct a lower barrier for v of the type

$$w := e^{k\bar{w}} - 1, \quad \bar{w} := -\tilde{u} + \frac{\epsilon}{2}(|x|^2 - 1),$$

for appropriate constants k large and $\epsilon \ll \mu$ small. Notice that $w \le 0$ on ∂B_1, since $\bar{w} \le 0$ on ∂B_1. Also

$$\bar{w} \ge c|x - z| \text{ on } [z, z^S],$$

since, by convexity, $-\tilde{u} \ge c|x - z|$ on $[z, z^S]$ for some c depending on μ and ρ. It suffices to check that

$$\tilde{U}^{ij} w_{ij} \ge 0 \text{ on } B_1 \setminus S',$$

since then we obtain $v \ge (\inf_{S'} v)cw$ in $B_1 \setminus S'$, which easily implies the conclusion. In $B_1 \setminus S'$ we have $|\nabla \bar{w}| \ge c(\mu) > 0$, provided that ϵ is sufficiently small, and thus

$$\tilde{U}^{ij} \bar{w}_i \bar{w}_j = (\det D^2\tilde{u})(\nabla \bar{w})^T (D^2\tilde{u})^{-1} \nabla \bar{w} \ge c\Lambda^{-1},$$

where Λ is the largest eigenvalue of $D^2\tilde{u}$. Then we use that $\operatorname{tr} \tilde{U} \ge c\lambda^{-1} \ge c\Lambda^{1/(n-1)}$, where λ is the smallest eigenvalue of $D^2\tilde{u}$, and obtain

$$\begin{aligned}\tilde{U}^{ij} w_{ij} &= ke^{k\bar{w}}\big(\tilde{U}^{ij}\bar{w}_{ij} + k\tilde{U}^{ij}\bar{w}_i\bar{w}_j\big) \ge ke^{k\bar{w}}\big(-n\rho^{-1} + \epsilon \operatorname{tr} \tilde{U} + kc\Lambda^{-1}\big)\\ &\ge ke^{k\bar{w}}\big(-n\rho^{-1} + c(\epsilon\Lambda^{1/(n-1)} + k\Lambda^{-1})\big) \ge 0,\end{aligned}$$

if k is chosen large depending on ϵ, ρ, μ and n. □

6. Singular minimizers in dimension $n \ge 3$.

Let

$$u(x) := |x'|^{2-2/n} h(x_n)$$

be the singular solution to $\det D^2u = 1$ constructed by Pogorelov, with h a smooth even function, defined in a neighborhood of 0 and $h(0) = 1$, satisfying an ODE

$$\left(\left(1 - \frac{2}{n}\right)hh'' - \left(2 - \frac{2}{n}\right)h'^2\right)h^{n-2} = c.$$

We let

$$v(x) := |x'|^{2-2/n} q(x_n)$$

be obtained as the infinitesimal difference between u and a rescaling of u,

$$v(x', x_n) := \lim_{\epsilon \to 0} \frac{1}{\epsilon}\big[u(x', x_n) - (1+\epsilon)^{-\gamma} u(x', (1+\epsilon)x_n)\big],$$

for some small $\gamma < \dfrac{2}{n}$. Notice that

$$q(t) = \gamma h(t) - h'(t)t$$

and $q > 0$ in a small interval $(-a, a)$ and q vanishes at its end points. Also,

$$U^{ij} v_{ij} = n\gamma - 2 < 0 \quad \text{in } \Omega := \mathbb{R}^{n-1} \times [-a, a],$$
$$v = 0, \quad U^{\nu\nu} v_\nu = U^{nn} v_n = -\sigma_0 \quad \text{on } \partial\Omega,$$

for some constant $\sigma_0 > 0$. The last equality follows since U^{nn} is homogeneous of degree $-(n-1)(2/n)$ in $|x'|$ and v_n is homogeneous of degree $2 - 2/n$ in $|x'|$.

Notice that u, v are solutions of the system (1-2) in the infinite cylinder Ω for uniform measures A and σ. In order to obtain a solution in a finite domain Ω_0, we modify v outside a neighborhood of the line $|x'| = 0$ by subtracting a smooth convex function ψ which vanishes in B_1 and increases rapidly outside B_1. Precisely, we let

$$\tilde{v} := v - \psi, \quad \Omega_0 := \{\tilde{v} > 0\},$$

and then we notice that u, $\tilde{v}$, solve the system (1-2) in the smooth bounded domain Ω_0 for smooth measures A and σ.

Since

$$|U^{ij}| \le C r^{(2/n)-2} \quad \text{if } |x'| \ge r,$$

we integrate by parts in the domain $\Omega_0 \setminus \{|x'| \le \epsilon\}$ and then let $\epsilon \to 0$ and find

$$\int_{\Omega_0} \varphi \, dA = -\int_{\Omega_0} U^{ij} \varphi_{ij} v + \int_{\partial\Omega_0} \varphi \, d\sigma, \quad \text{for all } \varphi \in C^2(\overline{\Omega}_0),$$

or

$$L(\varphi) = \int_{\Omega_0} U^{ij} \varphi_{ij} v.$$

This implies that L is stable, that is, $L(\varphi) > 0$ for any convex φ which is not linear. Also, if $w \in C^2(\overline{\Omega}_0)$ satisfies $\det D^2 w = 1$, then $U^{ij}(w - u)_{ij} \ge 0$, and we obtain

$$L(w) - L(u) = \int_{\Omega_0} U^{ij} (w - u)_{ij} v \ge 0,$$

that is, u is a minimizer of L.

We remark that the domain Ω_0 has flat boundary in a neighborhood of the line $\{|x'| = 0\}$, and therefore is not uniformly convex. However, this is not essential in our example. One can construct, for example, a function $\bar{v}$ in a uniformly convex domain by modifying v as

$$\bar{v} := |x'|^{2-2/n} q(x_n(1 + \delta |x'|^2)),$$

for some small $\delta > 0$.

7. Proof of Theorem 1.4

We assume for simplicity that $\Omega = B_1$. The existence of a minimizer u for the convex functional E follows as in Section 2. First, we show that

$$t_1 \le \det D^2 u \le t_0 \tag{7-1}$$

for some t_1 depending on F and ρ. The upper bound follows easily. If $\det D^2u > t_0$ in a set of positive measure, then the function w defined as

$$\det D^2 w = \min\{t_0, \det D^2 u\}, \quad w = u \text{ on } \partial B_1,$$

satisfies $E(w) < E(u)$, since $F(\det D^2 w) = F(\det D^2 u)$ and $L(w) < L(u)$.

In order to obtain the lower bound in (7-1), we need the following lemma.

Lemma 7.1. *Let w be a convex function in B_1 with*

$$(\det D^2 w)^{1/n} = g \in L^n(B_1).$$

Let $w + \varphi$ be another convex function in B_1 with the same boundary values as w such that

$$(\det D^2(w+\varphi))^{1/n} = g - h, \quad \text{for some } h \geq 0.$$

Then

$$\int_{B_1} \varphi g^{n-1} \leq C(n) \int_{B_1} h g^{n-1}.$$

Proof. By approximation, we may assume that w, φ are smooth in $\overline{B}_1$. Using the concavity of the map $M \mapsto (\det M)^{1/n}$ in the space of symmetric matrices $M \geq 0$, we obtain

$$(\det D^2(w+\varphi))^{1/n} \leq (\det D^2 w)^{1/n} + \frac{1}{n}(\det D^2 w)^{(1/n)-1} W^{ij}\varphi_{ij},$$

and hence

$$-nhg^{n-1} \leq W^{ij}\varphi_{ij}.$$

We multiply both sides by $\Phi := \frac{1}{2}(1-|x|^2)$ and integrate. Since both φ and Φ vanish on ∂B_1 we integrate by parts twice and obtain

$$-C(n)\int_{B_1} hg^{n-1} \leq \int_{B_1} W^{ij}\Phi_{ij}\varphi = -\int_{B_1} (\operatorname{tr} W)\varphi.$$

Using

$$\operatorname{tr} W \geq c(n)(\det W)^{1/n} = c(n)(\det D^2 w)^{(n-1)/n} = c(n) g^{n-1},$$

we obtain the desired conclusion. □

Now we prove the lower bound in (7-1). Define w such that $w = u$ on ∂B_1 and

$$\det D^2 w = \max\{t_1, \det D^2 u\}$$

for some small t_1. Since $G(t) = F(t^n)$ is convex and $\det D^2 w \geq t_1$, we have

$$G\big((\det D^2 w)^{1/n}\big) \leq G\big((\det D^2 u)^{1/n}\big) + G'(t_1^{1/n})\big((\det D^2 w)^{1/n} - (\det D^2 u)^{1/n}\big).$$

We write

$$u - w = \varphi, \quad (\det D^2 w)^{1/n} = g, \quad (\det D^2 u)^{1/n} = g - h,$$

and we rewrite the inequality above as

$$F(\det D^2w) \le F(\det D^2u) + G'(t_1^{1/n})h.$$

From Lemma 7.1, we obtain

$$\int_{B_1} h\, g^{n-1} \ge c(n) \int_{B_1} \varphi\, g^{n-1},$$

and since h is supported on the set where the value of $g = t_1^{1/n}$ is minimal, we find that

$$\int_{B_1} h \ge c(n) \int_{B_1} \varphi.$$

This gives

$$\int_{B_1} F(\det D^2w) - F(\det D^2u) \le c(n)G'(t_1^{1/n}) \int_{B_1} \varphi,$$

and thus, using the minimality of u and $G'(0^+) = -\infty$,

$$0 \le E(w) - E(u) \le \int_{B_1} \varphi dA + c(n)G'(t_1^{1/n}) \int_{B_1} \varphi \le 0,$$

if t_1 is small enough. In conclusion, $\varphi = 0$ and $u = w$ and (7-1) is proved.

We write

$$\det D^2u = f, \quad t_1 \le f \le t_0.$$

Any minimizer for L in the class of functions whose determinant equals f is a minimizer for E as well. In order to apply Theorem 1.2, we need f to be Holder continuous. However, we can approximate f by smooth functions f_n and find smooth minimizers u_n for approximate linear functionals L_n with the constraint

$$\det D^2u_n = f_n.$$

By Proposition 3.1 (see (4-5)),

$$\|u_n\|_{C^{1,\beta}(\overline{B}_1)},\ \|v_n\|_{C^{\beta}(\overline{B}_1)} \le C,$$

and hence we may assume (see Theorem 1.1) that, after passing to a subsequence, $u_n \to u$ and $v_n \to v$ uniformly for some function $v \in C^\beta(\overline{B}_1)$. We show that

$$v = -F'(f). \tag{7-2}$$

Then by the hypotheses on F, we obtain $\det D^2u = f \in C^\beta(\overline{B}_1)$, and from Theorem 1.2, we easily obtain

$$\|u\|_{C^{2,\alpha}(\overline{B}_1)},\ \|v\|_{C^{2,\alpha}(\overline{B}_1)} \le C$$

for some C depending on ρ, α, $\|\sigma\|_{C^\alpha(\overline{B}_1)}$, $\|A\|_{C^\alpha(\overline{B}_1)}$, and F.

In order to prove (7-2), we need a uniform integral bound (in two dimensions) between solutions to the Monge–Ampère equation and solutions of the corresponding linearized equation.

The proof of the following lemma will be given at the end of the section.

Lemma 7.2. *Assume $n = 2$ and let w be a smooth convex function in $\overline{B}_1$ with*

$$\lambda \le \det D^2 w := g \le \Lambda$$

for some positive constants λ, Λ. Let $w + \epsilon\varphi$ be a convex function with

$$\det D^2(w + \epsilon\varphi) = g + \epsilon h, \quad \varphi = 0 \text{ on } \partial B_1$$

for some smooth function h with $\|h\|_{L^\infty} \le 1$. If $\epsilon \le \epsilon_0$, then

$$\int_{B_1} |h - W^{ij}\varphi_{ij}| \le C\epsilon$$

for some C, ϵ_0 depending only on λ, Λ.

Now let h be a smooth function, $\|h\|_{L^\infty} \le 1$, and we solve the equations

$$\det D^2(u_n + \epsilon\varphi_n) = f_n + \epsilon h, \quad \varphi_n = 0 \text{ on } \partial B_1,$$

with u_n, f_n as above. From (1-1) we see that

$$L_n(\varphi_n) = \int_{B_1} (U_n^{ij}\partial_{ij}\varphi_n)v_n,$$

and hence, by the lemma above,

$$\left| L_n(\varphi_n) - \int_{B_1} h v_n \right| \le C\epsilon$$

with C universal. We let $n \to \infty$ and obtain

$$\left| L(\varphi) - \int_{B_1} h v \right| \le C\epsilon,$$

with φ the solution of

$$\det D^2(u + \epsilon\varphi) = f + \epsilon h, \quad \varphi = 0 \text{ on } \partial B_1.$$

The inequality $E(u + \epsilon\varphi) \ge E(u)$ implies

$$\int_{B_1} \bigl(F(f + \epsilon h) - F(f) + \epsilon h v\bigr) \ge -C\epsilon^2,$$

and hence, as $\epsilon \to 0$,

$$\int_{B_1} (F'(f) + v)h \ge 0 \quad \text{for any smooth } h,$$

which gives (7-2). □

Proof of Lemma 7.2. Using the concavity of $(\det D^2 w)^{1/n}$, we obtain

$$(g + \epsilon h)^{1/n} \le g^{1/n} + \frac{\epsilon}{n} g^{1/n-1} W^{ij}\varphi_{ij},$$

and thus, for $\epsilon \le \epsilon_0$,

$$h - C\epsilon \le W^{ij}\varphi_{ij}. \tag{7-3}$$

Since $n = 2$, we have

$$\det D^2(w + \epsilon\varphi) = \det D^2 w + \epsilon W^{ij}\varphi_{ij} + \epsilon^2 \det D^2\varphi,$$

and hence

$$h - W^{ij}\varphi_{ij} = \epsilon \det D^2\varphi.$$

From the pointwise inequality (7-3), we see that in order to prove the lemma, it suffices to show that

$$\int_{B_1} \det D^2\varphi \geq -C.$$

Let $\Phi = (\Phi^{ij})$ be the cofactor matrix of $D^2\phi$. Integrating by parts and using $\varphi = 0$ on ∂B_1, we find

$$\int_{B_1} 2\det D^2\varphi = \int_{B_1} \Phi^{ij}\varphi_{ij} = \int_{\partial B_1} \Phi^{ij}\varphi_i \nu_j = \int_{\partial B_1} \Phi^{\nu\nu}\varphi_\nu = \int_{\partial B_1} \varphi_\nu^2 \geq 0,$$

where we used $\Phi^{\nu\nu} = \varphi_{\tau\tau} = \varphi_\nu$. ☐

References

[Abreu 1998] M. Abreu, "Kähler geometry of toric varieties and extremal metrics", *Internat. J. Math.* **9**:6 (1998), 641–651. MR 99j:58047 Zbl 0932.53043

[Caffarelli and Gutiérrez 1997] L. A. Caffarelli and C. E. Gutiérrez, "Properties of the solutions of the linearized Monge–Ampère equation", *Amer. J. Math.* **119**:2 (1997), 423–465. MR 98e:35060 Zbl 0878.35039

[De Philippis and Figalli 2013] G. De Philippis and A. Figalli, "$W^{2,1}$ regularity for solutions of the Monge–Ampère equation", *Invent. Math.* **192**:1 (2013), 55–69. MR 3032325 Zbl 06160861

[Donaldson 2002] S. K. Donaldson, "Scalar curvature and stability of toric varieties", *J. Differential Geom.* **62**:2 (2002), 289–349. MR 2005c:32028 Zbl 1074.53059

[Donaldson 2005] S. K. Donaldson, "Interior estimates for solutions of Abreu's equation", *Collect. Math.* **56**:2 (2005), 103–142. MR 2006d:35035 Zbl 1085.53063

[Donaldson 2008] S. K. Donaldson, "Extremal metrics on toric surfaces: a continuity method", *J. Differential Geom.* **79**:3 (2008), 389–432. MR 2009j:58018 Zbl 1151.53030

[Donaldson 2009] S. K. Donaldson, "Constant scalar curvature metrics on toric surfaces", *Geom. Funct. Anal.* **19**:1 (2009), 83–136. MR 2010j:32041 Zbl 1177.53067

[Le and Savin 2013] N. Q. Le and O. Savin, "Boundary regularity for solutions to the linearized Monge–Ampère equations", *Arch. Ration. Mech. Anal.* **210**:3 (2013), 813–836. MR 3116005 Zbl 06168117

[Savin 2013] O. Savin, "Pointwise $C^{2,\alpha}$ estimates at the boundary for the Monge–Ampère equation", *J. Amer. Math. Soc.* **26**:1 (2013), 63–99. MR 2983006 Zbl 06168117

[Trudinger and Wang 2008a] N. S. Trudinger and X.-J. Wang, "Boundary regularity for the Monge–Ampère and affine maximal surface equations", *Ann. of Math.* (2) **167**:3 (2008), 993–1028. MR 2010h:35168 Zbl 1176.35046

[Trudinger and Wang 2008b] N. S. Trudinger and X.-J. Wang, "The Monge–Ampère equation and its geometric applications", pp. 467–524 in *Handbook of geometric analysis, I*, edited by L. Ji et al., Adv. Lect. Math. (ALM) **7**, International Press, Somverville, MA, 2008. MR 2010g:53065 Zbl 1156.35033

[Zhou and Zhu 2008] B. Zhou and X. Zhu, "Minimizing weak solutions for Calabi's extremal metrics on toric manifolds", *Calc. Var. Partial Differential Equations* **32**:2 (2008), 191–217. MR 2009a:53081 Zbl 1141.53061

Received 5 Apr 2012. Revised 12 Dec 2012. Accepted 28 Feb 2013.

NAM Q. LE: `namle@math.columbia.edu`
Department of Mathematics, Columbia University, New York, NY 10027, United States

OVIDIU SAVIN: `savin@math.columbia.edu`
Department of Mathematics, Columbia University, New York, NY 10027, United States

mathematical sciences publishers

msp

ANALYSIS AND PDE
Vol. 6, No. 5, 2013
dx.doi.org/10.2140/apde.2013.6.1051

ON THE SPECTRUM OF DEFORMATIONS OF COMPACT DOUBLE-SIDED FLAT HYPERSURFACES

DENIS BORISOV AND PEDRO FREITAS

We study the asymptotic behavior of the eigenvalues of the Laplace–Beltrami operator on a compact hypersurface in $\mathbb{R}^{n+1}$ as it is *flattened* into a singular double-sided flat hypersurface. We show that the limit spectral problem corresponds to the Dirichlet and Neumann problems on one side of this flat (Euclidean) limit, and derive an explicit three-term asymptotic expansion for the eigenvalues where the remaining two terms are of orders $\varepsilon^2 \log \varepsilon$ and ε^2.

1. Introduction

In recent years there have been several papers studying the effect that flattening a domain has on the eigenvalues of the Laplace operator [Borisov and Cardone 2011; Borisov and Freitas 2009; 2010; Friedlander and Solomyak 2009]; see also [Nazarov 2001; Panasenko 2005] and the references therein for similar problems with boundary conditions other than Dirichlet. In these papers the main objective has been the derivation of the asymptotics of these eigenvalues in terms of a scalar parameter measuring how thin the domain becomes in one direction, as this parameter approaches zero. As far as we are aware, almost if not all such existing examples in the literature are concerned with domains in Euclidean space where the limiting problem degenerates to a domain of zero measure and therefore eigenvalues approach infinity.

A slightly different set of problems which has been considered consists of domains which are perturbations of singular sets such as thin tubular neighborhoods of graphs, i.e., domains which locally are like thin tubes — see [Exner and Post 2005; 2009], for instance, and also [Grieser 2008] for a review. As in the papers cited above, again the limiting domains have zero measure and the spectrum behaves in quite a different way from the model considered here.

In this paper we study a situation which, although different from that described in the first paragraph, has in common with it the process by which the limiting domain is approached. More precisely, consider the case of a given domain Ω in $\mathbb{R}^{n+1}$ satisfying certain restrictions which for the purpose here may be stated roughly as being bounded from above and below by the graphs of two functions — see Section 2 for a precise formulation. The domain Ω is then flattened towards a domain ω in $\mathbb{R}^n$ via a (continuous) one-parameter family of domains Ω_ε. These domains are obtained as the functions mentioned above are

---

Both authors were partially supported by FCT's projects PTDC/MAT/101007/2008 and PEst-OE/MAT/UI0208/2011. Borisov was partially supported by RFBR, by the Federal Task Program, and by a fellowship of the Dynasty Foundation for young mathematicians.
MSC2000: primary 35P15; secondary 35J05.
Keywords: Laplace–Beltrami operator, eigenvalue, flat manifolds.

multiplied by the parameter ε. The problem that shall concern us here is the study of the evolution of the eigenvalues of the Laplace–Beltrami operator on the one-parameter family of compact hypersurfaces $\mathcal{S}_\varepsilon$ which are the boundaries of the domains Ω_ε described above, as ε approaches zero. One of the differences in this instance is that while the domain Ω_0 has zero $(n+1)$-measure as stated above, $\mathcal{S}_0$ retains positive n-measure, developing instead a singularity on the boundary of the domain ω (when considered as a domain in $\mathbb{R}^n$). We thus expect these eigenvalues to remain finite as the parameter ε approaches zero, and to converge to a limiting spectral problem on the double-sided flat hypersurface. This is indeed the case, and the relevant spectral problems turn out to be the Dirichlet and Neumann problems on the domain ω, with the two next asymptotic terms after that being of orders $\varepsilon^2 \log\varepsilon$ and ε^2. These results have been announced in [Borisov and Freitas 2012].

In order to understand the origin of the $\varepsilon^2 \log\varepsilon$ term in the expansion, it turns out that it is sufficient to consider the case where n equals one, that is when the boundary is basically S^1. Because of this, it is not necessary to take into consideration the geometric intricacies of the problem which appear in higher dimensions and it is possible to obtain the full description of eigenvalues in terms of elliptic integrals.

More precisely, for an ellipse of radii 1 and ε we have that the eigenvalues are given by

$$\lambda_k(\varepsilon) = \frac{k^2\pi^2}{4E^2(1-\varepsilon^2)} \quad \text{for } k \in \mathbb{Z}, \quad \text{where } E(m) = \int_0^{\pi/2} \sqrt{1-m\sin^2(\theta)}\, d\theta$$

is the complete elliptic integral of the second type yielding one quarter of the perimeter of the ellipse for $m = 1-\varepsilon^2$.

Combining the above with the asymptotic expansion for E yields

$$\lambda_k(\varepsilon) = \frac{k^2\pi^2}{4} + \frac{k^2\pi^2}{4}\varepsilon^2 \log\varepsilon + \frac{k^2\pi^2}{2}\left(\frac{1}{4} - \log 2\right)\varepsilon^2 + \mathcal{O}(\varepsilon^{2+\rho}), \quad \rho \in (0,1).$$

In some sense, the purpose of the analysis that we shall carry out in what follows is to show that the above result may actually be extended to higher dimensions. It should be noted here that this expansion depends on the relation between the different variables at the endpoints of the segment, which in this case is of the form $x_1^2 + \varepsilon^2 x_2^2 = 1$. Clearly different relations between the leading powers will lead to different expansions.

More generally, the issue is that the points of the boundary of Ω where there is a tangent in the direction along which the domain is being flattened will play a special role. Throughout the paper we assume this set of points to be contained in a hyperplane orthogonal to the scaling direction, and that this tangency is simple. In the vicinity of these points we take the cross-section of our surface as indicated in Figure 1 which, with the assumptions made, will be similar to the one-dimensional ellipse described above. Our results then state that in the higher-dimensional case the asymptotics for the eigenvalues still behave in a similar fashion and thus the logarithmic terms appearing above persist in this more general setting.

Apart from the intrinsic interest of the behavior of the spectrum close to double-sided flat domains, we point out that such manifolds have appeared in the literature in connection with eigenvalues as maximizers of the invariant eigenvalues among all surfaces isometric to surfaces of revolution in $\mathbb{R}^3$ [Abreu and Freitas 2002] and for hypersurfaces of revolution diffeomorphic to a sphere and isometrically embedded

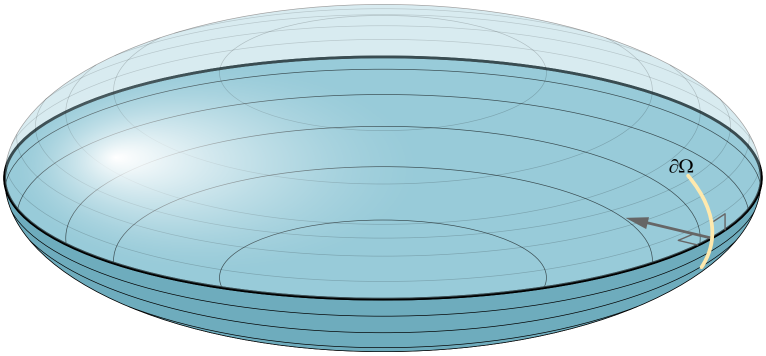

Figure 1. Surface $\mathcal{S}_\varepsilon$ with a cross-section at the edge.

in $\mathbb{R}^{n+1}$ [Colbois et al. 2008]. In fact, it is shown in those papers that these optimal singular *double flat disks* maximize the whole invariant spectrum and not just a specific eigenvalue. Another source of interest for such asymptotic expansions lies with the fact that, in some cases, they turn out to be fairly good approximations for low eigenvalues also for values of the parameter ε away from zero — see [Borisov and Freitas 2009; 2010; Freitas 2007].

We remark in passing that another problem for which it is conjectured that the optimal shape is given by a double-sided flat disk is Alexandrov's conjecture relating the area and diameter of surfaces of nonnegative curvature.

The structure of the paper is as follows. In the next section we give a precise formulation of the problem under consideration and state our main results, namely, the nature of the limiting problem and the relation of the limit and approximating operators. This includes the form of the asymptotic expansion and the expressions for the first three coefficients and an application to the case of the surface of an ellipsoid. Section 3 is then devoted to several preliminaries and auxiliary material used in Sections 4 and 5, where the proofs of the main results are presented.

2. Problem formulation and main results

Let $x'=(x_1,\dots,x_n)$, $x=(x',x_{n+1})$ be Cartesian coordinates in $\mathbb{R}^n$ and $\mathbb{R}^{n+1}$, respectively, $n\geqslant 2$, and let ω be a bounded domain in $\mathbb{R}^n$ with infinitely smooth boundary. Let also $h_\pm=h_\pm(x')\in C^\infty(\omega)\cap C(\overline{\omega})$ denote two arbitrary functions and define the manifold

$$\mathcal{S}_\varepsilon := \{x : x'\in\overline{\omega}, x_{n+1}=\varepsilon h_+(x')\}\cup\{x : x'\in\overline{\omega}, x_{n+1}=-\varepsilon h_-(x')\}, \tag{2-1}$$

where ε is a small positive parameter. We assume $\mathcal{S}_\varepsilon$ to be infinitely differentiable and to have no self-intersections. To ensure this, we make the following assumptions on $h_\pm$, the first of which ensures the absence of self-intersections:

(A1) The following relations hold true:

$$h_+(x')+h_-(x')>0,\quad x'\in\omega,\qquad h_+(x')=h_-(x')=0,\quad x'\in\partial\omega.$$

To state the second assumption we need to introduce some additional notation. Let $\nu = \nu(P)$, $P \in \partial\omega$, be the inward normal to $\partial\omega$, and denote by τ the distance to a point measured in the direction of ν. Consider the equations

$$t = h_+(P + \tau\nu(P)), \quad t > 0, \qquad t = -h_-(P + \tau\nu(P)), \quad t < 0. \tag{2-2}$$

Our second assumption concerns the solvability of these equations with respect to τ and implies the smoothness of $\mathcal{S}_\varepsilon$ in a neighborhood of $\partial\omega$:

(A2) There exists $t_0 > 0$ such that for all $t \in [-t_0, t_0]$, $P \in \partial\omega$, the equations (2-2) have a unique solution given by

$$\tau = a(t, P) \in C^\infty([-t_0, t_0] \times \partial\omega),$$

such that

$$\frac{\partial^2 a}{\partial t^2} > 0 \quad \text{for all } P \in \partial\omega. \tag{2-3}$$

We observe that assumptions (A1) and (A2) imply that

$$h_+(x') \geqslant 0, \quad h_-(x') \leqslant 0 \quad \text{in a small neighborhood of } \partial\omega.$$

The main object of our study is the Laplace–Beltrami operator $\mathcal{H}_\varepsilon$ on $\mathcal{S}_\varepsilon$. We introduce it rigorously as the self-adjoint operator associated with a symmetric lower-semibounded sesquilinear form

$$\mathfrak{h}_\varepsilon[u, v] := (\nabla u, \nabla v)_{L_2(\mathcal{S}_\varepsilon)} \quad \text{on } W_2^1(\mathcal{S}_\varepsilon).$$

We recall that on an arbitrary manifold with metric tensor g this may be written in local coordinates $y = (y_1, \dots, y_n)$ as

$$-\det{}^{-\frac{1}{2}} g \sum_{i,j=1}^{n} \frac{\partial}{\partial y_i} g^{ij} \det{}^{\frac{1}{2}} g \frac{\partial}{\partial y_j},$$

where g^{ij} are the entries of the inverse to the metric tensor. If in our case we take x' as local coordinates on $\mathcal{S}_\varepsilon$, then on each side $\mathcal{S}_\varepsilon^\pm$ the operator $\mathcal{H}_\varepsilon$ may be written in the form

$$\mathcal{H}_\varepsilon = -(1 + \varepsilon^2|\nabla_{x'} h_\pm|^2)^{-\frac{1}{2}} \operatorname{div}_{x'} (1 + \varepsilon^2|\nabla_{x'} h_\pm|^2)^{\frac{1}{2}} (\mathrm{E} + \varepsilon^2 \mathrm{Q}_\pm)^{-1} \nabla_{x'}, \tag{2-4}$$

where E is the $n \times n$ identity matrix and $\mathrm{Q}_\pm$ is the matrix with entries $\frac{\partial h_\pm}{\partial x_i} \frac{\partial h_\pm}{\partial x_j}$. On the boundary $\partial\omega$ the coefficients of such operator have singularities, and this is why in a neighborhood of $\partial\omega$ it is more convenient to employ the coordinates (τ, s), where s are some local coordinates on $\partial\omega$. We do not give here the expression of the operator $\mathcal{H}_\varepsilon$ in such coordinates, as it requires the introduction of additional (cumbersome) notation. These two parametrizations are discussed in detail in Section 3.

The purpose of the present paper is to describe the asymptotic behavior of the resolvent and the spectrum of $\mathcal{H}_\varepsilon$ as $\varepsilon \to +0$. In this limit, the hypersurface $\mathcal{S}_\varepsilon$ collapses to a flat two-sided domain $\boldsymbol{\omega} = (\omega_+, \omega_-)$, where $\omega_\pm$ are two copies of ω understood as the *upper* and *lower* sides of $\boldsymbol{\omega}$. Because of this, it is natural to expect that the limiting operator for $\mathcal{H}_\varepsilon$ as $\varepsilon \to +0$ is the Laplacian on $\boldsymbol{\omega}$, i.e., that on $\omega_\pm$ subject to certain boundary conditions. Indeed, this is true, and it is our first main result. Namely, we introduce the space $L_2(\boldsymbol{\omega})$ as consisting of the vectors $\boldsymbol{u} = (u_+, u_-)$, where the functions $u_\pm$ are defined on $\omega_\pm$ and $u_\pm \in L_2(\omega_\pm)$. We can naturally identify $L_2(\boldsymbol{\omega})$ with $L_2(\omega) \oplus L_2(\omega)$. In the same way we introduce the Sobolev spaces $W_2^j(\boldsymbol{\omega})$ assuming that for each $\boldsymbol{u} \in W_2^j(\boldsymbol{\omega})$ the functions $u_\pm \in W_2^j(\omega_\pm)$ satisfy the boundary conditions

$$\left.\frac{\partial^i u_+}{\partial \tau^i}\right|_{\partial\omega} = (-1)^i \left.\frac{\partial^i u_-}{\partial \tau^i}\right|_{\partial\omega}, \quad i = 0, 1, \dots, j-1. \tag{2-5}$$

The meaning of these boundary conditions is that the functions $u_\pm$ should be "glued smoothly" while moving from ω_+ to ω_- via $\partial\omega = \partial\omega_\pm$. We observe that $W_2^j(\boldsymbol{\omega})$ is embedded into $W_2^j(\omega) \oplus W_2^j(\omega)$, but does not coincide. It is also clear that for any $u \in W_2^1(\omega)$ the function $\boldsymbol{u} := (u, u)$ belongs to $W_2^1(\boldsymbol{\omega})$. Similarly, if $u \in W_2^2(\omega)$, $u|_{\partial\omega} = 0$, or, respectively, $u \in W_2^2(\omega)$, $\frac{\partial u}{\partial \tau}|_{\partial\omega} = 0$, then $\boldsymbol{u} = (u, -u) \in W_2^2(\boldsymbol{\omega})$, or, respectively, $\boldsymbol{u} = (u, u) \in W_2^2(\boldsymbol{\omega})$.

Let $\mathcal{H}_0$ be the self-adjoint operator in $L_2(\boldsymbol{\omega})$ associated with the closed symmetric lower-semibounded sesquilinear form

$$\mathfrak{h}_0[\boldsymbol{u}, \boldsymbol{v}] := (\nabla \boldsymbol{u}, \nabla \boldsymbol{v})_{L_2(\boldsymbol{\omega})} \quad \text{on } W_2^1(\boldsymbol{\omega}).$$

By $\mathfrak{D}(\cdot)$ we denote the domain of an operator, and the symbol $\|\cdot\|_{X\to Y}$ indicates the norm of an operator acting from the Hilbert space X to a Hilbert space Y.

Given any vector $\boldsymbol{u} = (u_+, u_-)$ defined on $\boldsymbol{\omega}$, by $\mathcal{I}_\varepsilon \boldsymbol{u}$ we denote the function on $\mathcal{S}_\varepsilon$ being $u_+(x')$ on $\{x : x' \in \overline{\omega}, x_{n+1} = \varepsilon h_+(x')\}$ and $u_-(x')$ on $\{x : x' \in \overline{\omega}, x_{n+1} = -\varepsilon h_-(x')\}$. And vice versa, given any function u defined on $\mathcal{S}_\varepsilon$, by $\mathcal{I}_\varepsilon^{-1} u$ we denote the vector $\boldsymbol{u} = (u_+, u_-)$, where $u_\pm = u_\pm(x') := u(x')$, $x' \in \omega$, $x_{n+1} = \varepsilon h_\pm(x')$.

Theorem 2.1. *For each $z \in \mathbb{C} \setminus \mathbb{R}$ there exists $C(z) > 0$ such that the following estimate holds true*:

$$\left\|(\mathcal{H}_\varepsilon - z)^{-1} - \mathcal{I}_\varepsilon(\mathcal{H}_0 - z)^{-1}\mathcal{I}_\varepsilon^{-1}\right\|_{L_2(\mathcal{S}_\varepsilon)\to W_2^1(\mathcal{S}_\varepsilon)} \leqslant C(z)\varepsilon^{2/3}. \tag{2-6}$$

Remark 2.2. The statement of this theorem includes the fact that the operator $\mathcal{I}_\varepsilon(\mathcal{H}_0 - z)^{-1}\mathcal{I}_\varepsilon^{-1}$ is well-defined as a bounded one from $L_2(\mathcal{S}_\varepsilon)$ into $W_2^1(\mathcal{S}_\varepsilon)$.

In view of the embedding of $W_2^1(\boldsymbol{\omega})$ into $W_2^1(\omega) \oplus W_2^1(\omega)$, and the compact embedding of the latter into $L_2(\omega) \oplus L_2(\omega) = L_2(\boldsymbol{\omega})$, the operator $\mathcal{H}_\varepsilon$ has a compact resolvent. Hence, it has a pure discrete spectrum accumulating only at infinity. The same is true for the Dirichlet and Neumann Laplacians $-\Delta_\omega^{(D)}$ and $-\Delta_\omega^{(N)}$ on ω. Recall that $-\Delta_\omega^{(D)}$ is the Friedrichs extension in $L_2(\omega)$ of $-\Delta$ from $C_0^\infty(\Omega)$, and $-\Delta_\omega^{(N)}$ is the self-adjoint operator in $L_2(\omega)$ associated with the sesquilinear form $(\nabla u, \nabla v)_{L_2(\Omega)}$ on $W_2^1(\omega)$. In what follows $\sigma_d(\cdot)$ denotes the discrete spectrum of an operator.

Our next result follows from Theorem 2.1 and [Reed and Simon 1980, Theorems VIII.23, VIII.24].

Theorem 2.3. *The eigenvalues of $\mathcal{H}_\varepsilon$ converge to those of $\mathcal{H}_0$ as ε goes to zero. In particular*, *if $\lambda \notin \sigma_d(\mathcal{H}_0)$, then $\lambda \notin \sigma_d(\mathcal{H}_\varepsilon)$ for ε small enough. For each m-multiple eigenvalue $\lambda \in \sigma_d(\mathcal{H}_0)$ there exist exactly m eigenvalues* (*counting multiplicities*) *of $\mathcal{H}_\varepsilon$ converging to λ as $\varepsilon \to +0$. Let $\mathcal{P}_0$ be the projector on the eigenspace associated with λ, $\mathcal{P}_\varepsilon$ be the total projector associated with the eigenvalues* *of $\mathcal{H}_\varepsilon$ converging to λ. Then the following convergence holds true*:

$$\|\mathcal{P}_\varepsilon - \mathcal{I}_\varepsilon \mathcal{P}_0 \mathcal{I}_\varepsilon^{-1}\|_{L_2(\mathcal{G}_\varepsilon)\to W_2^1(\mathcal{G}_\varepsilon)} \to 0, \quad \varepsilon \to +0.$$

Let now λ be an eigenvalue of $\mathcal{H}_0$ with multiplicity m and $\boldsymbol{\psi}_i = (\psi_+^{(i)}, \psi_-^{(i)})$ be associated eigenfunctions orthonormalized in $L_2(\boldsymbol{\omega})$. It will be shown in the next section in Lemma 4.2 that the asymptotics

$$\psi_\pm^{(i)}(x') = \Psi_i^{(0)}(P) \pm \Psi_i^{(1)}(P)\tau + \mathcal{O}(\tau^2), \quad P \in \partial\omega, \quad \tau \to +0, \tag{2-7}$$

hold true, where

$$\Psi_i^{(0)} = \psi_+^{(i)}\big|_{\partial\omega} = \psi_-^{(i)}\big|_{\partial\omega} \in C^\infty(\partial\omega), \quad \Psi_i^{(1)} = \frac{\partial\psi_+^{(i)}}{\partial\tau}\bigg|_{\partial\omega} = -\frac{\partial\psi_-^{(i)}}{\partial\tau}\bigg|_{\partial\omega} \in C^\infty(\partial\omega).$$

By $-\Delta_{\partial\omega}$ we denote the Laplace–Beltrami operator on $\partial\omega$, where the metric $\mathrm{G}_{\partial\omega}$ on $\partial\omega$ is induced by the Euclidean one in $\mathbb{R}^n$. For any smooth functions u, v on $\partial\omega$, we shall denote the pointwise scalar product of their gradients by $\nabla u \cdot \nabla v$.

Let

$$\omega^\delta := \omega \setminus \{x' : 0 < \tau < \delta\}. \tag{2-8}$$

Employing the coefficients of the asymptotics (2-7), we introduce two real symmetric matrices $\Lambda^{(0)}$, $\Lambda^{(1)}$ with entries

$$\Lambda_{ij}^{(0)} := \int_{\partial\omega} \frac{1}{a_2}\big(\lambda\Psi_i^{(0)}\Psi_j^{(0)} - \nabla\Psi_i^{(0)}\cdot\nabla\Psi_j^{(0)} + \Psi_i^{(1)}\Psi_j^{(1)}\big)\,d\omega, \tag{2-9}$$

$$\begin{aligned}\Lambda_{ij}^{(1)} := -\lim_{\delta\to+0}\bigg[&\frac{1}{2}\int_{\omega^\delta} |\nabla_{x'}h_+|^2\big(\lambda\psi_+^{(i)}\psi_+^{(j)} - (\nabla_{x'}\psi_+^{(i)}, \nabla_{x'}\psi_+^{(j)})_{\mathbb{R}^d}\big)\,dx' \\ &+ \frac{1}{2}\int_{\omega^\delta} |\nabla_{x'}h_-|^2\big(\lambda\psi_-^{(i)}\psi_-^{(j)} - (\nabla_{x'}\psi_-^{(i)}, \nabla_{x'}\psi_-^{(j)})_{\mathbb{R}^d}\big)\,dx' \\ &+ \int_{\omega^\delta} (\nabla_{x'}h_+, \nabla_{x'}\psi_+^{(i)})_{\mathbb{R}^d}(\nabla_{x'}h_+, \nabla_{x'}\psi_+^{(j)})_{\mathbb{R}^d}\,dx' \\ &+ \int_{\omega^\delta} (\nabla_{x'}h_-, \nabla_{x'}\psi_-^{(i)})_{\mathbb{R}^d}(\nabla_{x'}h_-, \nabla_{x'}\psi_-^{(j)})_{\mathbb{R}^d}\,dx' \\ &+ \ln\delta\int_{\partial\omega}\frac{1}{4a_2}\big(\Psi_i^{(1)}\Psi_j^{(1)} + \lambda\Psi_i^{(0)}\Psi_j^{(0)} - \nabla\Psi_i^{(0)}\cdot\nabla\Psi_j^{(0)}\big)\,ds\bigg] \\ -\int_{\partial\omega}&\frac{1+4\ln 2+\ln a_2}{4a_2}\big(\Psi_i^{(1)}\Psi_j^{(1)} + \lambda\Psi_i^{(0)}\Psi_j^{(0)} - \nabla\Psi_i^{(0)}\cdot\nabla\Psi_j^{(0)}\big)\,ds,\end{aligned} \tag{2-10}$$

where

$$a_2(P) := \frac{1}{2}\frac{\partial^2 a}{\partial t^2}(0, P).$$

It will be shown in Section 4 that the matrix $\Lambda^{(1)}$ is well-defined. By the theorem on simultaneous diagonalization of two quadratic forms, in what follows the eigenfunctions $\boldsymbol{\psi}_i$ are supposed to be orthonormalized in $L_2(\boldsymbol{\omega})$ and the matrix $\Lambda^{(0)} + \frac{1}{\ln\varepsilon}\Lambda^{(1)}$ to be diagonal. The eigenfunctions $\boldsymbol{\psi}_i$ chosen in this way depend on ε, but it is clear that the norms $\|\psi_{\pm}^{(i)}\|_{C^k(\overline{\omega})}$ are bounded uniformly in ε for all $k \geqslant 0$, $i = 1, \dots, m$.

Theorem 2.4. *Let λ be an m-multiple eigenvalue of $\mathcal{H}_0$ and $\boldsymbol{\psi}_i$, $i = 1, \dots, m$, be the associated eigenfunctions of $\mathcal{H}_0$ chosen as described above. Then there exist exactly m eigenvalues $\lambda_k(\varepsilon)$, $k = 1, \dots, m$ (counting multiplicity) of $\mathcal{H}_\varepsilon$ converging to λ. These eigenvalues satisfy the asymptotic expansions*

$$\lambda_k(\varepsilon) = \lambda + \varepsilon^2 \ln \varepsilon \, \mu_k\left(\frac{1}{\ln \varepsilon}\right) + \mathcal{O}(\varepsilon^{2+\rho}), \tag{2-11}$$

where μ_k are the eigenvalues of the matrix $\Lambda^{(0)} + \frac{1}{\ln \varepsilon}\Lambda^{(1)}$, and ρ is any constant in $(0, 1/2)$. The eigenvalues $\mu_k\left(\frac{1}{\ln \varepsilon}\right)$ are holomorphic in $\frac{1}{\ln \varepsilon}$ and converge to the eigenvalues of $\Lambda^{(0)}$ as $\varepsilon \to 0$.

In addition to the asymptotic expansions for the eigenvalues $\lambda_i(\varepsilon)$ given in this theorem, we also obtain the asymptotics for the total projector associated with these eigenvalues. However, to formulate this result we have to introduce additional notation and it is thus more convenient to postpone its statement which will then be made at the end of Section 5 — see Theorem 5.3.

Let us describe briefly the main ideas employed in the proofs of the main results. The proof of the uniform resolvent convergence in Theorem 2.1 is based on the analysis of the quadratic forms associated with the perturbed and the limiting operators and on the accurate estimates of the functions in certain weighted Sobolev spaces. The proof of the first theorem uses essentially the method of matching asymptotic expansions [Il′in 1992] for formal construction of the asymptotics for the eigenfunctions associated with $\lambda_k(\varepsilon)$. These asymptotics are constructed as a combination of outer and inner expansions. The former depends on x' and its coefficients have singularities at $\partial\omega$. In the vicinity of $\partial\omega$ we introduce a special rescaled variable $\xi := a^{1/2}(x_{n+1}\varepsilon^{-1}, P)\varepsilon^{-1}$ as $x_{n+1} > 0$ and $\xi := -a^{1/2}(x_{n+1}\varepsilon^{-1}, P)\varepsilon^{-1}$ as $x_{n+1} < 0$. This variable then describes the slope of $\mathcal{S}_\varepsilon$ in the vicinity of ε — see also the equations (3-11) giving the parametrization of $\mathcal{S}_\varepsilon$ in the vicinity of $\partial\omega$. After rewriting the eigenvalue equation in the variables (ξ, s), where s are local coordinates on $\partial\omega$, its leading term is in fact the Laplace–Beltrami operator on the ellipse giving rise to the logarithmic terms in the asymptotics for both the eigenvalues and the eigenfunctions.

Despite the fact that we are only presenting the leading terms of the asymptotics for $\lambda_k(\varepsilon)$ and for the associated total projector in Theorems 2.4 and 5.3, respectively, our approach also allows us to construct the complete asymptotic expansions if required. Although this would need to be checked in a way similar to what was done here for the first few terms, the ansatzes (5-1) and (5-39) suggest that the complete asymptotic expansion for the eigenvalues should be

$$\lambda_k(\varepsilon) = \lambda + \varepsilon^2 \ln \varepsilon \mu_k(\varepsilon) + \sum_{i=2}^{\infty} \varepsilon^{2i} \ln^i \varepsilon \mu_k^{(i)}\left(\frac{1}{\ln \varepsilon}\right),$$

where $\mu_k^{(i)}$ are functions holomorphic in $\frac{1}{\ln \varepsilon}$. These higher-order terms would then still reflect the behavior observed in the ellipse example given in the Introduction.

Although the above formulas for $\Lambda_{ij}^{(0)}$ and (especially) $\Lambda_{ij}^{(1)}$ may look quite cumbersome at a first glance, they will actually simplify when computed for particular cases as some of the terms involved will vanish depending on whether we are considering Dirichlet or Neumann boundary conditions on $\partial\omega$. We note that a similar effect was already present when computing the coefficients in the expansions obtained in [Borisov and Freitas 2009; 2010]. This is particularly clear in the second of these papers dealing with dimensions higher than two, where the general expression is quite complicated and needs to be computed specifically in each case. When this is done for general ellipsoids in any dimension, for instance, it yields a much simpler one-line expression.

We shall illustrate this by considering a thin ellipsoidal surface. To this end take ω to be the unit disk centered at the origin with

$$h_{\pm}(x') := \sqrt{1-r^2}, \quad r = |x'|, \quad \tau = 1-r, \quad a_2 = \tfrac{1}{2}. \tag{2-12}$$

Under such definition this surface converges to the unit disk ω regarded as a double-sided surface. In this instance the limiting eigenvalues may be found via separation of variables and they will be of the form κ^2, where κ are the zeroes of the Bessel function J_κ and its derivative J'_κ, corresponding to eigenfunctions satisfying Dirichlet and Neumann boundary conditions on $\partial\omega$, respectively. The following examples illustrating both cases are taken from [Borisov and Freitas 2012], where the details may be found.

We consider the case of Dirichlet boundary conditions first; i.e.,

$$J_0(\kappa) = 0, \quad \lambda = \kappa^2, \quad \psi(x) = -\frac{J_0(\kappa r)}{\sqrt{2\pi}\, J_1(\kappa)}, \quad \boldsymbol{\psi} = (\psi, -\psi), \quad \Psi^{(0)} = 0, \quad \Psi^{(1)} = -\frac{\kappa}{\sqrt{2\pi}}.$$

Substituting these formulas and (2-12) into (2-9) and (2-10), we then obtain

$$\Lambda_{11}^{(0)} = 2\lambda \quad \text{and} \quad \Lambda_{11}^{(1)} = -\frac{\lambda}{J_1^2(\kappa)} \int_0^1 \frac{r^3}{1-r^2} \big(J_0^2(\kappa r) + J_1^2(\kappa r) - J_1^2(\kappa) \big)\, dr - \lambda \ln 2.$$

The asymptotics (2-11) thus become

$$\lambda_\kappa(\varepsilon) = \lambda + \varepsilon^2 (2\lambda \ln \varepsilon + \Lambda_{11}^{(1)}) + \mathcal{O}(\varepsilon^{2+\rho})$$

and, for a particular eigenvalue, the remaining integral may be computed numerically. We illustrate this by considering the case corresponding to the first Dirichlet eigenvalue on the disk which yields

$$\lambda_1(\varepsilon) = j_{0,1}^2 + \varepsilon^2 (2 j_{0,1}^2 \ln \varepsilon + \Lambda_{11}^{(1)}) + \mathcal{O}(\varepsilon^{2+\rho}) \approx 5.7831 + 11.5664\, \varepsilon^2 \ln \varepsilon - 6.0871\, \varepsilon^2 + \mathcal{O}(\varepsilon^{2+\rho}).$$

As an example of a limiting multiple eigenvalue we consider the first nontrivial Neumann eigenvalue of the disk. In two dimensions this is a double eigenvalue with associated (normalized) eigenfunctions

$$\psi_1(x) = \frac{J_1(\kappa' r) \cos\theta}{J_0(\kappa') \sqrt{\pi(\kappa'^2 - 1)}}, \quad \psi_2(x) = \frac{J_1(\kappa' r) \sin\theta}{J_0(\kappa') \sqrt{\pi(\kappa'^2 - 1)}},$$

where θ is the polar angle corresponding to x and κ' is the first nontrivial zero of J'_1.

The eigenfunctions in $L_2(\boldsymbol{\omega})$ are then given by $\boldsymbol{\psi}_i = (\psi_i, \psi_i)$, $i = 1, 2$, from which we have

$$\Psi_1^{(0)} = \frac{J_1(\kappa') \cos\theta}{J_0(\kappa') \sqrt{\pi(\kappa'^2 - 1)}}, \quad \Psi_2^{(0)} = \frac{J_1(\kappa') \sin\theta}{J_0(\kappa') \sqrt{\pi(\kappa'^2 - 1)}}$$

and $\Psi_i^{(1)} = 0$, $i = 1, 2$. Proceeding as before, we have

$$\Lambda_{11}^{0} = \Lambda_{22}^{0} = \frac{2J_1^2(\kappa')}{J_0^2(\kappa')} = 2\kappa'^2 = 2\lambda \quad \text{and} \quad \Lambda_{ij}^{0} = 0 \quad (i \neq j).$$

For the next term we now obtain

$$\Lambda_{ii}^{(1)} = -\frac{\kappa'^2}{J_0^2(\kappa')(\kappa'^2-1)} \int_0^1 \frac{r^3}{1-r^2}\left[J_1^2(\kappa' r) - J_1^2(\kappa') + J_0^2(\kappa' r) + J_0^2(\kappa') - \frac{2}{\kappa' r} J_0(\kappa' r) J_1(\kappa' r)\right] dr - \lambda \ln 2$$

for $i = 1, 2$ and $\Lambda_{ij} = 0$ for $i \neq j$.

From this, and again computing the relevant integrals numerically, we obtain

$$\lambda_i(\varepsilon) = (j_{1,1}')^2 + \varepsilon^2(2\lambda \ln \varepsilon + \Lambda_{ii}^{(1)}) + \mathcal{O}(\varepsilon^{2+\rho}) \approx 3.3900 + 6.7799\,\varepsilon^2 \ln \varepsilon - 1.8555\,\varepsilon^2 + \mathcal{O}(\varepsilon^{2+\rho}), \quad i = 1, 2.$$

Due to the radial symmetry of ω, it is clear that these two eigenvalues should coincide, and the associated eigenfunctions converge to $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$.

3. Preliminaries

In this section we discuss two parametrizations of the surface $\mathcal{S}_\varepsilon$ and prove three auxiliary lemmas which will be used in the next sections for proving Theorems 2.1, 2.4.

First parametrization of $\mathcal{S}_\varepsilon$. The first parametrization is that used in the definition of $\mathcal{S}_\varepsilon$ in (2-1); i.e., each point on $\mathcal{S}_\varepsilon$ is described as $x_{n+1} = \pm\varepsilon h_\pm(x')$, $x' \in \overline{\omega}$, where the sign corresponds to the upper or lower part of $\mathcal{S}_\varepsilon$. Let us first calculate the metrics on $\mathcal{S}_\varepsilon$ in terms of the variables x'.

The tangential vectors to $\mathcal{S}_\varepsilon$ at the point $x' \in \omega$, $x_{n+1} = \varepsilon h_\pm(x')$ are

$$\left(0, \dots, 0, 1, 0, \dots, 0, \varepsilon\frac{\partial h_\pm}{\partial x_i}\right), \quad i = 1, \dots, n,$$

where "1" stands on i-th position. Thus, the metric tensor has the form

$$\mathrm{G}_\pm(x', \varepsilon) := \begin{pmatrix} 1 + \varepsilon^2\big(\frac{\partial h_\pm}{\partial x_1}\big)^2 & \varepsilon^2 \frac{\partial h_\pm}{\partial x_1}\frac{\partial h_\pm}{\partial x_2} & \varepsilon^2 \frac{\partial h_\pm}{\partial x_1}\frac{\partial h_\pm}{\partial x_3} & \dots & \varepsilon^2 \frac{\partial h_\pm}{\partial x_1}\frac{\partial h_\pm}{\partial x_n} \\ \varepsilon^2 \frac{\partial h_\pm}{\partial x_2}\frac{\partial h_\pm}{\partial x_1} & 1 + \varepsilon^2\big(\frac{\partial h_\pm}{\partial x_2}\big)^2 & \varepsilon^2 \frac{\partial h_\pm}{\partial x_2}\frac{\partial h_\pm}{\partial x_3} & \dots & \varepsilon^2 \frac{\partial h_\pm}{\partial x_2}\frac{\partial h_\pm}{\partial x_n} \\ \varepsilon^2 \frac{\partial h_\pm}{\partial x_3}\frac{\partial h_\pm}{\partial x_1} & \varepsilon^2 \frac{\partial h_\pm}{\partial x_3}\frac{\partial h_\pm}{\partial x_2} & 1 + \varepsilon^2\big(\frac{\partial h_\pm}{\partial x_3}\big)^2 & \dots & \varepsilon^2 \frac{\partial h_\pm}{\partial x_3}\frac{\partial h_\pm}{\partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varepsilon^2 \frac{\partial h_\pm}{\partial x_n}\frac{\partial h_\pm}{\partial x_1} & \varepsilon^2 \frac{\partial h_\pm}{\partial x_n}\frac{\partial h_\pm}{\partial x_2} & \varepsilon^2 \frac{\partial h_\pm}{\partial x_{n-1}}\frac{\partial h_\pm}{\partial x_3} & \dots & 1 + \varepsilon^2\big(\frac{\partial h_\pm}{\partial x_n}\big)^2 \end{pmatrix}.$$

It easy to see that

$$\mathrm{G}_\pm(x', \varepsilon) = \mathrm{E} + \varepsilon^2 \mathrm{Q}_\pm, \quad \mathrm{Q}_\pm := (\nabla_{x'} h_\pm)(\nabla_{x'} h_\pm)^*, \tag{3-1}$$

where $\nabla_{x'} h_\pm$ is treated as a column vector, and "$*$" denotes transposition.

Lemma 3.1. *The matrix* $G_\pm$ *has two eigenvalues, the* $(n-1)$*-multiple eigenvalue* 1, *and the simple eigenvalue* $(1+\varepsilon^2|\nabla_{x'}h_\pm|^2)$. *The following identity holds true*:

$$d\mathcal{S}_\varepsilon = J_\varepsilon^\pm\, dx', \quad J_\varepsilon^\pm := \sqrt{1+\varepsilon^2|\nabla_{x'}h_\pm|^2}, \quad dx' = dx_1\, dx_2 \cdots dx_n. \tag{3-2}$$

Proof. From (3-1) we may write the eigenvalue problem for the matrix $G_\pm$ as

$$(\mathrm{E}+\varepsilon^2 vv^*)u = zu \quad \text{and} \quad (z-1)u = \varepsilon^2 vv^*u,$$

where $v=\nabla_{x'}h_\pm$. We thus see that any vector orthogonal to v is an eigenvector for the above equation with eigenvalue z equal to one. This yields an eigenvalue of multiplicity $n-1$ if v is not zero, and n in case v vanishes. In the former case, we easily see that v is also an eigenvector, now with eigenvalue $1+\varepsilon^2|v|^2$, which will have multiplicity one. The determinant of $G_\pm$ is thus $g^\pm = 1+\varepsilon^2|v|^2$, yielding the volume element to be $\sqrt{1+\varepsilon^2|v|^2}$ as desired. □

In what follows we shall make use of the differential expression for the operator $\mathcal{H}_\varepsilon$, namely, its expansion with respect to ε. The expression itself is given by (2-4), while using (3-1) allows us to expand some of the terms in this expression in powers of ε:

$$(\mathrm{E}+\varepsilon^2\mathrm{Q}_\pm)^{-1} = \mathrm{E}-\varepsilon^2\mathrm{Q}_\pm+\mathcal{O}(\varepsilon^4), \quad (1+\varepsilon^2|\nabla_{x'}h_\pm|^2)^{\pm\frac{1}{2}} = 1\pm\varepsilon^2\frac{|\nabla_{x'}h_\pm|^2}{2}+\mathcal{O}(\varepsilon^4),$$

where the plus and minus signs correspond to the upper and lower parts of S_ε, respectively. We substitute these formulas into (2-4) and get

$$\mathcal{H}_\varepsilon = -\Delta_{x'} - \varepsilon^2\left(\frac{|\nabla_{x'}h_\pm|^2}{2}\Delta_{x'} + \mathrm{div}_{x'}\left(\frac{|\nabla_{x'}h_\pm|^2}{2}-\mathrm{Q}_\pm\right)\nabla_{x'}\right)+\mathcal{O}(\varepsilon^4). \tag{3-3}$$

The disadvantage of the parametrization by the variables x' is that the functions $h_\pm$ are not smooth in a vicinity of $\partial\omega$ and their derivatives blow up at the boundary $\partial\omega$. We shall show this below while introducing the second parametrization. The main idea of the second parametrization is to use special coordinates in a vicinity of $\partial\omega$ so that they involve smooth functions only; this parametrization is purely local and will be used only in a vicinity of $\partial\omega$. It is natural to expect the existence of such coordinates since the surface $\mathcal{S}_\varepsilon$ is infinitely differentiable.

Second parametrization of $\boldsymbol{\mathcal{S}_\varepsilon}$**.** In a neighborhood of $\partial\omega$ we introduce new coordinates (τ, s), where $s=(s_1,\dots,s_{n-1})$ are local coordinates on $\partial\omega$ corresponding to a C^∞-atlas, and τ, we remind, is the distance to a point measured in the direction of the inward normal $\nu=\nu(s)$ to $\partial\omega$. Let $r=r(s)$ be the vector-function describing $\partial\omega$. We have

$$x' = r(s)+\tau\nu(s), \quad \nabla_{(\tau,s)} = \mathrm{M}(\tau,s)\nabla_{x'}, \quad \mathrm{M}=\mathrm{M}(\tau,s) = \begin{pmatrix} \nu \\ \dfrac{\partial r}{\partial s_1}+\tau\dfrac{\partial \nu}{\partial s_1} \\ \vdots \\ \dfrac{\partial r}{\partial s_{n-1}}+\tau\dfrac{\partial \nu}{\partial s_{n-1}} \end{pmatrix}, \tag{3-4}$$

where $\nu(s)$ and the other vectors in the definition of M are treated as rows. The vectors $\frac{\partial r}{\partial s_i}$ are tangential to M and linearly independent, while $\nu(s)$ is orthogonal to $\partial\omega$. Thus, the matrix M is invertible for all sufficiently small τ and all $s \in \partial\omega$. The inequalities

$$C_1 \leqslant \mathrm{M}(\tau, s) \leqslant C_2, \quad C_2^{-1} \leqslant \mathrm{M}^{-1}(\tau, s) \leqslant C_1^{-1}, \quad s \in \partial\omega, \quad \tau \in [-\tau_0, \tau_0] \tag{3-5}$$

are valid, where C_1, C_2 are positive constants independent of (τ, s). It follows from these estimates and (3-4) that the matrix $M^{-1}(\tau, s)$ is infinitely differentiable in the neighborhood $\{x : |\tau| < \tau_0\}$ of $\partial\omega$.

Consider now the equations (2-2). By assumption (A2) they have the smooth solution $\tau = a(x_{n+1}, P)$ and, for small x_{n+1}, the function a behaves as

$$a(x_{n+1}, P) = a_2(P)x_{n+1}^2 + \mathcal{O}(x_{n+1}^3).$$

Hence,

$$\begin{gathered} h_\pm(P + \tau\nu(P)) = x_{n+1} = \pm a_2^{-\frac{1}{2}}(P)\tau^{\frac{1}{2}} + \mathcal{O}(\tau), \quad \tau \to +0, \qquad \nabla_{x'}h_\pm = \mathrm{M}^{-1}\nabla_{(\tau,s)}h_\pm, \\ C_3\tau^{-1} \leqslant |\nabla_{x'}h_\pm|^2 \leqslant C_4\tau^{-1}, \quad \tau \in (0, \tau_0], \end{gathered} \tag{3-6}$$

where C_3, C_4 are positive constants independent of (τ, s). As we see from the last estimates, the functions $h_\pm$ are not smooth at the point $\tau = 0$, i.e., at $\partial\omega$.

We employ once again assumption (A2) and pass from the equations $x_{n+1} = \pm\varepsilon h_\pm(x')$ to

$$\tau = a(t, P), \quad x_{n+1} = \varepsilon t, \quad x' = r(s) + \tau\nu(s). \tag{3-7}$$

It follows from (2-3) that the function $a(t, P)$ can be represented as $t^2\widetilde{a}(t, P)$, where $\widetilde{a} \in C^\infty([-t_0, t_0] \times \partial\omega)$ and $\widetilde{a} > 0$ for sufficiently small t_0.

We introduce a new variable $\zeta = t\widetilde{a}^{\frac{1}{2}}(t, P)$. From assumption (A2) we conclude that

$$t = b(\zeta, P) \in C^\infty([-\zeta_0, \zeta_0] \times \partial\omega) \tag{3-8}$$

for a fixed small constant ζ_0, and the Taylor series for a and b read

$$a(t, P) = \sum_{i=2}^{\infty} a_i(P)t^i, \quad t \to +0, \tag{3-9}$$

$$b(\zeta, P) = \sum_{i=1}^{\infty} b_i(P)\zeta^i, \quad \zeta \to 0, \qquad b_1 := a_2^{-\frac{1}{2}}, \tag{3-10}$$

where a_i, $b_i \in C^\infty(\partial\omega)$. We define a rescaled variable $\xi := \zeta\varepsilon^{-1}$. The final form of the second parametrization for $\mathcal{S}_\varepsilon$ is

$$x' = r(s) + \varepsilon^2\xi^2\nu(s), \quad x_{n+1} = \varepsilon^2 b_\varepsilon(\xi, r(s)), \quad \xi \in [-\zeta_0\varepsilon^{-1}, \zeta_0\varepsilon^{-1}], \tag{3-11}$$

where $b_\varepsilon(\xi, P) := \varepsilon^{-1}b(\varepsilon\xi, P)$ and ζ_0 is a fixed sufficiently small number. We observe that, by the definition of ζ,

$$\tau = a(t, P) = \zeta^2 = \varepsilon^2\xi^2. \tag{3-12}$$

As in (3-3), we shall also employ the expansion in ε of the differential expression for $\mathcal{H}_\varepsilon$ corresponding to the second parametrization. We find first the tangential vectors to S_ε corresponding to the parametrization (3-11):

$$T_{s_i}=\left(\frac{\partial r}{\partial s_i}+\varepsilon^2\xi^2\frac{\partial \nu}{\partial s_i},\varepsilon^2\frac{\partial b_\varepsilon}{\partial s_i}\right),\quad T_\xi=\varepsilon^2\left(2\xi\nu,\frac{\partial b_\varepsilon}{\partial \xi}\right). \tag{3-13}$$

It is clear that the vectors $\frac{\partial r}{\partial s_i}$, $\frac{\partial \nu}{\partial s_i}$ belong to the tangential plane and are orthogonal to ν. Employing this fact and (3-13), we calculate the metric tensor:

$$\begin{aligned}(T_\xi,T_\xi)_{\mathbb{R}^{n+1}}&=\varepsilon^4\left(4\xi^2+\left(\frac{\partial b_\varepsilon}{\partial \xi}\right)^2\right),\quad (T_\xi,T_{s_i})_{\mathbb{R}^{n+1}}=\varepsilon^4\frac{\partial b_\varepsilon}{\partial \xi}\frac{\partial b_\varepsilon}{\partial s_i},\\ (T_{s_i},T_{s_j})_{\mathbb{R}^{n+1}}&=\left(\frac{\partial r}{\partial s_i}+\varepsilon^2\xi^2\frac{\partial \nu}{\partial s_i},\frac{\partial r}{\partial s_j}+\varepsilon^2\xi^2\frac{\partial \nu}{\partial s_j}\right)_{\mathbb{R}^{n+1}}+\varepsilon^4\frac{\partial b_\varepsilon}{\partial s_i}\frac{\partial b_\varepsilon}{\partial s_j}.\end{aligned}$$

By the Weingarten equations we see that

$$((T_{s_i},T_{s_j})_{\mathbb{R}^{n+1}})_{i,j=\overline{1,n}}=\mathrm{A},$$

where

$$\begin{aligned}\mathrm{A}:=&\,\mathrm{G}_{\partial\omega}-2\varepsilon^2\xi^2\mathrm{B}+\varepsilon^4\xi^4\mathrm{B}\mathrm{G}_{\partial\omega}^{-1}\mathrm{B}+\varepsilon^4(\nabla_s b_\varepsilon)(\nabla_s b_\varepsilon)^*\\ =&\,\mathrm{G}_{\partial\omega}(\mathrm{E}-\varepsilon^2\xi^2\mathrm{G}_{\partial\omega}^{-1}\mathrm{B})^2+\varepsilon^4(\nabla_s b_\varepsilon)(\nabla_s b_\varepsilon)^*,\end{aligned} \tag{3-14}$$

$\mathrm{G}_{\partial\omega}$ is the metric tensor of $\partial\omega$ associated with the coordinates s, B is the second fundamental form of $\partial\omega$ corresponding to the orientation defined by ν. Hence, the metric tensor G_ε of S_ε associated with the parametrization (3-11) reads

$$\mathrm{G}_\varepsilon=\begin{pmatrix}\varepsilon^4(4\xi^2+(\frac{\partial b_\varepsilon}{\partial \xi})^2) & \varepsilon^4\mathrm{p}^*\\ \varepsilon^4\mathrm{p} & \mathrm{A}\end{pmatrix},\quad \mathrm{p}:=\frac{\partial b_\varepsilon}{\partial \xi}\nabla_s b_\varepsilon.$$

By direct calculations we check that

$$\mathrm{G}_\varepsilon^{-1}=\begin{pmatrix}\varepsilon^{-4}\beta-\beta\mathrm{p}^*\mathrm{A}^{-1}\\ -\beta\mathrm{A}^{-1}\mathrm{p}\mathrm{A}^{-1}+\varepsilon^4\beta\mathrm{A}^{-1}\mathrm{p}\mathrm{p}^*\mathrm{A}^{-1}\end{pmatrix},\quad \beta:=\left(4\xi^2+\left(\frac{\partial b_\varepsilon}{\partial \xi}\right)^2-\varepsilon^4\mathrm{p}^*\mathrm{A}^{-1}\mathrm{p}\right)^{-1}. \tag{3-15}$$

The quantities in (3-15) are well-defined provided ζ_0 is sufficiently small. Indeed, by (3-9),

$$\mathrm{A}=\mathrm{G}_{\partial\omega}+\mathcal{O}(\zeta^2),\quad \mathrm{p}=\mathcal{O}(1),\quad \frac{\partial b}{\partial \zeta}(\zeta,P)=\mathcal{O}(1),\quad \zeta\to 0,$$

which implies the existence of A^{-1} and β. In what follows we assume that ζ_0 is chosen in such a way.

By $K_i=K_i(s)$, $i=1,\dots,n-1$, we denote the principal curvatures of $\partial\omega$, and $K:=\sum_{i=1}^{n-1}K_i$. We note that $(n-1)^{-1}K$ is the mean curvature of $\partial\omega$ and let

$$\mathrm{a}:=\det\bigl((\mathrm{E}-\varepsilon^2\xi^2\mathrm{G}_{\partial\omega}^{-1}\mathrm{B})^2+\varepsilon^4\mathrm{G}_{\partial\omega}^{-1}(\nabla_s b_\varepsilon)(\nabla_s b_\varepsilon)^*\bigr).$$

Lemma 3.2. *The following identities hold true*:

$$b_\varepsilon=\sum_{i=1}^{\infty}b_i(P)\varepsilon^{i-1}\xi^i,\quad \mathrm{A}^{-1}=\mathrm{G}_{\partial\omega}^{-1}+\mathcal{O}(\varepsilon^2\xi^2),\quad \mathrm{p}=\xi b_1\nabla_s b_1+\mathcal{O}(\varepsilon\xi^2),\tag{3-16}$$

$$\det\mathrm{G}_\varepsilon=\varepsilon^4\beta^{-1}\det\mathrm{A},\tag{3-17}$$

$$\det\mathrm{A}=\mathrm{a}\det\mathrm{G}_{\partial\omega},\quad \mathrm{a}=\sum_{i=0}^{2}\varepsilon^{2i}\alpha_{2i}+\mathcal{O}(\varepsilon^4\xi^4),\tag{3-18}$$

$$\alpha_0:=1,\quad \alpha_2:=-2\xi^2K.\tag{3-19}$$

Proof. The identities (3-16) follow directly from the definitions of b_ε, A, and p.

We make linear transformations in (3-15) to calculate the determinant of G_ε:

$$(\det\mathrm{G}_\varepsilon)^{-1}=\det{}^{-1}\mathrm{G}_\varepsilon=\begin{vmatrix}\varepsilon^{-4}\beta & -\beta\mathrm{p}^*\mathrm{A}^{-1}\\ 0 & \mathrm{A}^{-1}\end{vmatrix}=\varepsilon^{-4}\beta\det{}^{-1}\mathrm{A},$$

which proves (3-17).

It is easy to see that

$$\det\mathrm{A}=\mathrm{a}\det\mathrm{G}_{\partial\omega}.\tag{3-20}$$

In view of (3-14) we get

$$\begin{aligned}\mathrm{a}&=\det\bigl(\mathrm{E}+\varepsilon^4(\mathrm{E}-\varepsilon^2\xi^2\mathrm{G}_{\partial\omega}^{-1}\mathrm{B})^{-2}\mathrm{G}_{\partial\omega}^{-1}(\nabla_s b_\varepsilon)(\nabla_s b_\varepsilon)^*\bigr)\det(\mathrm{E}-\varepsilon^2\xi^2\mathrm{G}_{\partial\omega}^{-1}\mathrm{B})^2\\ &=\bigl(1+\varepsilon^4\operatorname{Tr}(\mathrm{E}-\varepsilon^2\xi^2\mathrm{G}_{\partial\omega}^{-1}\mathrm{B})^{-2}\mathrm{G}_{\partial\omega}^{-1}(\nabla_s b_\varepsilon)(\nabla_s b_\varepsilon)^*+\mathcal{O}(\varepsilon^8\xi^2)\bigr)\prod_{i=1}^{n-1}(1-\varepsilon^2\xi^2K_i)^2\\ &=\bigl(1+\varepsilon^4\operatorname{Tr}\mathrm{G}_{\partial\omega}^{-1}(\nabla_s b_\varepsilon)(\nabla_s b_\varepsilon)^*+\mathcal{O}(\varepsilon^6\xi^4)\bigr)\bigl(1-2\varepsilon^2\xi^2K+\mathcal{O}(\varepsilon^4\xi^4)\bigr)\\ &=\bigl(1+\varepsilon^4|\nabla b_\varepsilon|^2+\mathcal{O}(\varepsilon^6\xi^4)\bigr)\bigl(1-2\varepsilon^2\xi^2K+\mathcal{O}(\varepsilon^4\xi^4)\bigr).\end{aligned}$$

We substitute the obtained formula and (3-10) into (3-20) and arrive at (3-18). □

Employing (3-14), (3-16), by direct calculations we check

$$\mathrm{p}^*\mathrm{A}^{-1}\mathrm{p}=\left(\frac{\partial b_\varepsilon}{\partial\xi}\right)^2(\nabla_s b_\varepsilon)^*\mathrm{G}_{\partial\omega}^{-1}(\nabla_s b_\varepsilon)+\mathcal{O}(\varepsilon^2\xi^2)=\left(\frac{\partial b_\varepsilon}{\partial\xi}\right)^2|\nabla b_\varepsilon|^2+\mathcal{O}(\varepsilon^2\xi^2)=b_1^2\xi^2|\nabla b_1|^2+\mathcal{O}(\varepsilon\xi^2).$$

Hence, by (3-17), (3-18) and the definition of β,

$$\varepsilon^{-2}\det{}^{\frac12}\mathrm{G}_\varepsilon=\beta^{-\frac12}\det{}^{\frac12}\mathrm{A}=\beta^{-1}\beta_{\mathrm{A}}\det{}^{\frac12}\mathrm{G}_{\partial\omega},\quad \beta_{\mathrm{A}}:=\beta^{\frac12}\mathrm{a}^{\frac12}=\sum_{i=0}^{4}\varepsilon^i\beta_{i-4}+\mathcal{O}\bigl(\varepsilon^5(|\xi|^2+\xi^4)\bigr),$$

where $\beta_i=\beta_i(\xi,P)\in C^\infty(\mathbb{R}\times\partial\omega)$ are some functions. In particular,

$$\beta_{-4}:=\frac{1}{(4\xi^2+b_1^2)^{\frac12}},\quad \beta_{-3}:=-\frac{2b_1b_2\xi}{(4\xi^2+b_1^2)^{\frac32}},$$

$$\beta_{-2}:=-\frac{3b_1b_3\xi^2}{(4\xi^2+b_1^2)^{\frac32}}-\frac{4\xi^2(2\xi^2-b_1^2)b_2^2}{(4\xi^2+b_1^2)^{\frac52}}-\frac{\xi^2K}{(4\xi^2+b_1^2)^{\frac12}},\tag{3-21}$$

while the functions β_{-1}, β_0 satisfy the uniform in ξ and P estimates

$$|\beta_{-1}| \leqslant \frac{C|\xi|^3}{1+|\xi|^3}, \quad |\beta_0| \leqslant C\xi^2(1+|\xi|).$$

The obtained formulas, Lemma 3.2, and (3-15) allow us to write the expansion for $\mathrm{G}_\varepsilon^{-1}$:

$$\varepsilon^{-2}(\det{}^{\frac{1}{2}} \mathrm{G}_\varepsilon)\mathrm{G}_\varepsilon^{-1} = \det{}^{\frac{1}{2}} \mathrm{G}_{\partial\omega} \sum_{i=-4}^{0} \varepsilon^i \mathrm{G}_i + \mathcal{O}(\varepsilon), \tag{3-22}$$

$$\mathrm{G}_i := \begin{pmatrix} \beta_i & 0 \\ 0 & 0 \end{pmatrix}, \quad i = -4, \dots, -1, \quad \mathrm{G}_0 := \begin{pmatrix} \beta_0 & -b_1\xi\beta_{-4}(\nabla_s b_1)^* \mathrm{G}_{\partial\omega}^{-1} \\ -b_1\xi\beta_{-4}\mathrm{G}_{\partial\omega}^{-1}\nabla_s b_1 & \beta_{-4}^{-1}\mathrm{G}_{\partial\omega}^{-1} \end{pmatrix}. \tag{3-23}$$

Taking into account (3-17), (3-18), we write the operator $\mathcal{H}_\varepsilon$ in terms of the variables (s_0, s), where $s_0 := \xi$:

$$\mathcal{H}_\varepsilon = -\frac{1}{\det^{\frac{1}{2}} \mathrm{G}_\varepsilon} \sum_{i,j=0}^{n-1} \frac{\partial}{\partial s_i} G_\varepsilon^{ij} \det{}^{\frac{1}{2}} \mathrm{G}_\varepsilon \frac{\partial}{\partial s_j} = -\frac{\varepsilon^{-2}\beta_{\mathrm{A}}}{\mathrm{a} \det^{\frac{1}{2}} \mathrm{G}_{\partial\omega}} \sum_{i,j=0}^{n-1} \frac{\partial}{\partial s_i} G_\varepsilon^{ij} \det{}^{\frac{1}{2}} \mathrm{G}_\varepsilon \frac{\partial}{\partial s_j}, \tag{3-24}$$

and G_ε^{ij} are the entries of the inverse matrix in (3-15). It follows from the last formula and (3-15) that

$$\mathcal{H}_\varepsilon = \varepsilon^{-4}\mathrm{a}^{-1}\beta_{\mathrm{A}} \frac{\partial}{\partial\xi} \beta_{\mathrm{A}} \frac{\partial}{\partial\xi} + \mathcal{O}(1).$$

We employ the obtained equation, (3-24), (3-22) and (3-23), and expand the coefficients of $\mathcal{H}_\varepsilon$ in powers of ε leading us to the identities

$$\mathcal{H}_\varepsilon = \sum_{i=-4}^{0} \varepsilon^i \mathcal{L}_i + \mathcal{O}(\varepsilon), \tag{3-25}$$

$$\begin{gathered} \mathcal{L}_{-4} := \mathcal{L}^{(-4)}, \quad \mathcal{L}_{-3} := \mathcal{L}^{(-3)}, \quad \mathcal{L}_{-2} := \mathcal{L}^{(-2)} + \alpha^{(2)}\mathcal{L}^{(-4)}, \mathcal{L}_{-1} := \mathcal{L}^{(-1)} + \alpha^{(2)}\mathcal{L}^{(-3)}, \\ \mathcal{L}_0 := \mathcal{L}^{(0)} + \alpha^{(2)}\mathcal{L}^{(-2)} + \alpha^{(4)}\mathcal{L}^{(-4)}, \quad \alpha^{(2)} := 2\xi^2 K, \quad \alpha^{(4)} = \alpha^{(4)}(\xi, s), \end{gathered} \tag{3-26}$$

$$\mathcal{L}^{(i)} := -\sum_{j=0}^{i+4} \beta_{j-4} \frac{\partial}{\partial\xi} \beta_{i-j} \frac{\partial}{\partial\xi}, \quad i = -4, \dots, -1, \tag{3-27}$$

$$\begin{aligned} \mathcal{L}^{(0)} := &-\sum_{l=0}^{4} \beta_{l-4} \frac{\partial}{\partial\xi} \beta_{-l} \frac{\partial}{\partial\xi} + b_1\beta_{-4} \frac{\partial}{\partial\xi} \xi\beta_{-4}(\nabla_s b_1)^* \mathrm{G}_{\partial\omega}^{-1} \nabla_s \\ &+ \beta_{-4} \det{}^{-\frac{1}{2}} \mathrm{G}_{\partial\omega} \operatorname{div}_s b_1\beta_{-4}\xi \det{}^{\frac{1}{2}} \mathrm{G}_{\partial\omega} (\nabla_s b_1)^* \mathrm{G}_{\partial\omega}^{-1} \frac{\partial}{\partial\xi} \\ &- \beta_{-4} \det{}^{-\frac{1}{2}} \mathrm{G}_{\partial\omega} \operatorname{div}_s \beta_{-4}^{-1} (\det{}^{\frac{1}{2}} \mathrm{G}_{\partial\omega}) \mathrm{G}_{\partial\omega}^{-1} \nabla_s. \end{aligned} \tag{3-28}$$

Auxiliary lemmas. We proceed to the auxiliary lemmas which will be used for proving Theorem 2.4.

Lemma 3.3. *In a vicinity of $\partial\omega$ the identities*

$$\det \mathrm{M} = (\det^{\frac{1}{2}} \mathrm{G}_{\partial\omega}) \prod_{i=1}^{n-1} (1-\tau K_i), \quad -\Delta_{x'} = -\frac{1}{\det \mathrm{M}} \operatorname{div}_{(\tau,s)} (\det \mathrm{M}) \widehat{\mathrm{M}} \nabla_{(\tau,s)} \tag{3-29}$$

hold true, *where*

$$\widehat{\mathrm{M}} := (\mathrm{M}^{-1})^* \mathrm{M}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & (\mathrm{E} - \tau \mathrm{G}_{\partial\omega}^{-1} \mathrm{B})^{-2} \mathrm{G}_{\partial\omega}^{-1} \end{pmatrix}. \tag{3-30}$$

Proof. It follows from (3-4) and the Weingarten formulas that

$$\mathrm{M} = \begin{pmatrix} \nu \\ \frac{\partial r}{\partial s_i} - \tau \sum_{k=1}^{n-1} B_i^k \frac{\partial r}{\partial s_k} \end{pmatrix},$$

where B_i^k are the entries of the matrix $\mathrm{G}_{\partial\omega}^{-1}\mathrm{B}$, and all vectors are treated as rows.

A straightforward direct calculation allows us to check that the inverse matrix M^{-1} reads

$$\mathrm{M}^{-1} = \begin{pmatrix} \nu \\ \sum_{k=1}^{n-1} c_i^k \frac{\partial r}{\partial s_k} \end{pmatrix}^*, \tag{3-31}$$

where $*$ indicates matrix transposition, and c_i^k are the entries of the matrix $\mathrm{C} = (\mathrm{E} - \tau \mathrm{G}_{\partial\omega}^{-1} \mathrm{B})^{-1} \mathrm{G}_{\partial\omega}^{-1}$.

Let $u_1, u_2 \in C_0^\infty(\omega)$ be any two functions with the corresponding supports located in a neighborhood of $\partial\omega$, where the coordinates (τ, s) are well-defined. We integrate by parts:

$$\begin{aligned} (-\Delta_{x'} u, v)_{L_2(\omega)} &= (\nabla_{x'} u, \nabla_{x'} v)_{L_2(\omega)} = (\mathrm{M}^{-1} \nabla_{(\tau,s)} u, (\det \mathrm{M}) \mathrm{M}^{-1} \nabla_{(\tau,s)} v)_{L_2((0,\tau_0)\times\partial\omega)} \\ &= \big(-\operatorname{div}_{(\tau,s)} (\det \mathrm{M}) (\mathrm{M}^{-1})^* (\mathrm{M}^{-1}) \nabla_{(\tau,s)} u, v\big)_{L_2((0,\tau_0)\times\partial\omega)} \\ &= \big(-(\det^{-1} \mathrm{M}) \operatorname{div}_{(\tau,s)} (\det \mathrm{M}) (\mathrm{M}^{-1})^* \mathrm{M}^{-1} \nabla_{(\tau,s)} u, v\big)_{L_2(\omega)}. \end{aligned}$$

Hence,

$$-\Delta_{x'} = -(\det^{-1} \mathrm{M}) \operatorname{div}_{(\tau,s)} (\det \mathrm{M}) (\mathrm{M}^{-1})^* \mathrm{M}^{-1} \nabla_{(\tau,s)}. \tag{3-32}$$

In view of (3-31) we have

$$(\mathrm{M}^{-1})^* \mathrm{M}^{-1} = \begin{pmatrix} \nu \\ \sum_{k=1}^{n-1} c_i^k \frac{\partial r}{\partial s_k} \end{pmatrix} \begin{pmatrix} \nu \\ \sum_{k=1}^{n-1} c_i^k \frac{\partial r}{\partial s_k} \end{pmatrix}^* = \begin{pmatrix} 1 & 0 \\ 0 & \mathrm{C} \mathrm{G}_{\partial\omega} \mathrm{C} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & (\mathrm{E} - \tau \mathrm{G}_{\partial\omega}^{-1} \mathrm{B})^{-2} \mathrm{G}_{\partial\omega}^{-1} \end{pmatrix},$$

$$\det^{-2} \mathrm{M} = \det (\mathrm{M}^{-1})^* \mathrm{M}^{-1} = \det (\mathrm{E} - \tau \mathrm{G}_{\partial\omega}^{-1} \mathrm{B})^{-2} \det \mathrm{G}_{\partial\omega}^{-1},$$

$$\det \mathrm{M} = \det^{\frac{1}{2}} \mathrm{G}_{\partial\omega} \det (\mathrm{E} - \tau \mathrm{G}_{\partial\omega}^{-1} \mathrm{B}) = \det^{\frac{1}{2}} \mathrm{G}_{\partial\omega} \prod_{i=1}^{n-1} (1 - \tau K_i).$$

The obtained formulas and (3-32) imply the statement of the lemma. ☐

We recall that the set ω^δ was introduced in (2-8).

Lemma 3.4. *Let the functions $f_\pm \in C^\infty(\omega_\pm)$ satisfy the differentiable asymptotics*

$$f_\pm(x') = \sum_{j=-4}^{\infty} f^{\pm}_{j/2}(P)\tau^{\frac{j}{2}}, \quad \tau \to +0 \tag{3-33}$$

uniformly in $P \in \partial\omega_\pm$, where $f^\pm_{j/2} \in C^\infty(\partial\omega_\pm)$, and $V^{(0)}$, $V^{(1)} \in C^\infty(\partial\omega)$ are some functions. Suppose the condition

$$\begin{aligned}\lim_{\delta\to+0}\Big[&(f_+, \psi_+^{(i)})_{L_2(\omega^\delta)} + (f_-, \psi_-^{(i)})_{L_2(\omega^\delta)} - \delta^{-1}\int_{\partial\omega}(f^+_{-2} + f^-_{-2})\Psi_i^{(0)}\,ds\\ &- 2\delta^{-1/2}\int_{\partial\omega}(f^+_{-3/2} + f^-_{-3/2})\Psi_i^{(0)}\,ds\\ &- \ln\delta\int_{\partial\omega}\big((K(f^+_{-2}+f^-_{-2}) - f^+_{-1} - f^-_{-1})\Psi_i^{(0)} - (f^+_{-2} - f^-_{-2})\Psi_i^{(1)}\big)\,ds\Big]\\ &- \int_{\partial\omega}(f^+_{-2} - f^-_{-2})\Psi_i^{(1)}\,ds + \int_{\partial\omega}(f^+_{-2}+f^-_{-2})\Psi_i^{(0)}K\,ds\\ &+ 2\int_{\partial\omega}\big(V^{(0)}\Psi_i^{(1)} - V^{(1)}\Psi_i^{(0)}\big)\,ds = 0, \quad i = 1,\dots,m,\end{aligned} \tag{3-34}$$

holds true. Then there exist the unique solutions $u_\pm \in C^\infty(\omega_\pm)$ to the equations

$$(-\Delta_{x'} - \lambda)u_\pm = f_\pm, \quad x \in \omega_\pm, \tag{3-35}$$

these solutions satisfy differentiable asymptotics

$$\begin{aligned}u_\pm(x') = f^\pm_{-2}(P)\ln\tau + U^{(0)}(P) \pm V^{(0)}(P) + 4f^\pm_{-3/2}(P)\tau^{1/2} + \tau(V^{(1)}(P) \pm U^{(1)}(P))&\\ + \tau(1-\ln\tau)\big(f^\pm_{-1}(P) - K(P)f^\pm_{-2}(P)\big) + \mathcal{O}(\tau^{3/2}), \quad \tau \to 0,&\end{aligned} \tag{3-36}$$

uniformly in $P \in \partial\omega_\pm$, where $U^{(0)}$, $U^{(1)} \in C^\infty(\partial\omega_\pm)$ are some functions, and the condition

$$(U_0, \Psi_i^{(0)})_{L_2(\partial\omega)} + (U_1, \Psi_i^{(1)})_{L_2(\partial\omega)} = 0, \quad i = 1,\dots,m, \tag{3-37}$$

holds true.

Proof. Let $\chi(\tau)$ be the cut-off function introduced in the proof of Lemma 4.4. We introduce the functions

$$\begin{aligned}\hat{u}_\pm(x') := \big(f^\pm_{-2}(P)\ln\tau \pm V^{(0)}(P) + 4f^\pm_{-3/2}(P)\tau^{1/2} + \tau(1-\ln\tau)\big(f^\pm_{-1}(P) - K(P)f^\pm_{-2}(P)\big)&\\ + \tau V^{(1)}(P) - \tfrac{4}{3}\tau^{3/2}\big(f^\pm_{-1/2}(P) - 2K(P)f^\pm_{-3/2}(P)\big)\big)\chi(\tau).&\end{aligned}$$

Employing Lemma 3.3, one can check that

$$(-\Delta_{x'} - \lambda)\hat{u}_\pm(x') = \chi(\tau)\sum_{j=-4}^{-1} f^\pm_{j/2}(P)\tau^j + \widehat{f}_\pm(x'), \tag{3-38}$$

where $\widehat{f}_\pm \in C^\infty(\omega_\pm) \cap L_2(\omega_\pm)$.

We construct the solutions to (3-35) as

$$u_\pm = \widehat{u}_\pm + \widetilde{u}_\pm.$$

Substituting this identity and (3-38) into (3-35), we obtain the equations for $\widetilde{u}_\pm$:

$$(-\Delta_{x'} - \lambda)\widetilde{u}_\pm = \widetilde{f}_\pm, \quad \widetilde{f}_\pm := f_\pm - \chi \sum_{j=-4}^{-1} f^\pm_{j/2}\tau^j - \widehat{f}_\pm, \tag{3-39}$$

and by (3-33) we have $\widetilde{f}_\pm \in L_2(\omega_\pm)$. Hence, we can rewrite these equations as

$$(\mathcal{H}_0 - \lambda)\widetilde{\boldsymbol{u}} = \widetilde{\boldsymbol{f}}, \quad \widetilde{\boldsymbol{u}} := (\widetilde{u}_+, \widetilde{u}_-), \quad \widetilde{\boldsymbol{f}} := (\widetilde{f}_+, \widetilde{f}_-). \tag{3-40}$$

Since λ is a discrete eigenvalue of $\mathcal{H}_0$, the solvability condition of the last equation is

$$(\widetilde{\boldsymbol{f}}, \boldsymbol{\psi}_i)_{L_2(\boldsymbol{\omega})} = 0, \quad k = 1, \dots, m,$$

which can be rewritten as

$$(\widetilde{f}_+, \psi_+^{(i)})_{L_2(\omega)} + (\widetilde{f}_+, \psi_+^{(i)})_{L_2(\omega)} = 0, \quad k = 1, \dots, m,$$

or, equivalently,

$$\lim_{\delta\to 0}\big((\widetilde{f}_+, \psi_+^{(i)})_{L_2(\omega^\delta)} + (\widetilde{f}_-, \psi_-^{(i)})_{L_2(\omega^\delta)}\big) = 0, \quad k = 1, \dots, m. \tag{3-41}$$

Integrating by parts and taking into account (3-38), (3-39), we get

$$(\widetilde{f}_\pm, \psi_\pm^{(i)})_{L_2(\omega^\delta)} = \big(f_\pm + (\Delta_{x'} + \lambda)\widetilde{u}_\pm, \psi_\pm^{(i)}\big)_{L_2(\omega^\delta)} = (f_\pm, \psi_\pm^{(i)})_{L_2(\omega^\delta)} - \int_{\partial\omega^\delta}\left(\psi_\pm^{(i)}\frac{\partial\widetilde{u}_\pm}{\partial\tau} - \widetilde{u}_\pm\frac{\partial\psi_\pm^{(i)}}{\partial\tau}\right)ds.$$

Here we have used that the normal derivative on $\partial\omega^\delta$ is that with respect to τ up to the sign. We parametrize the points of $\partial\omega^\delta$ by those on $\partial\omega$ via the relation $x' = r(s) + \delta\nu(s)$. In view of (3-4) and (3-29) we have

$$\int_{\partial\omega^\delta} \cdot \, ds = \int_{\partial\omega} \cdot \prod_{j=1}^{n-1}(1 - \tau K_j)\, ds. \tag{3-42}$$

Taking this formula into account, we continue the calculations:

$$\begin{aligned}
(\widetilde{f}_\pm, \psi_\pm^{(i)})_{L_2(\omega^\delta)} &= (f_\pm, \psi_\pm^{(i)})_{L_2(\omega^\delta)} - \int_{\partial\omega}\left(\psi_\pm^{(i)}\frac{\partial\widetilde{u}_\pm}{\partial\tau} - \widetilde{u}_\pm\frac{\partial\psi_\pm^{(i)}}{\partial\tau}\right)\bigg|_{x'=r(s)+\delta\nu(s)} \prod_{j=1}^{n-1}(1 - \tau K_j)\, ds \\
&= (f_\pm, \psi_\pm^{(i)})_{L_2(\omega^\delta)} - \delta^{-1}\int_{\partial\omega} f^\pm_{-2}\Psi_k^{(0)}\, ds - 2\delta^{-1/2}\int_{\partial\omega} f^\pm_{-3/2}\Psi_k^{(0)}\, ds \\
&\qquad - \ln\delta \int_{\partial\omega}\big((K f^\pm_{-2} - f^\pm_{-1})\Psi_i^{(0)} \mp f^\pm_{-2}\Psi_i^{(1)}\big)\, ds \\
&\qquad + \int_{\partial\omega} f^\pm_{-2}\big(\Psi_i^{(0)}K \mp \Psi_i^{(1)}\big)\, ds + \int_{\partial\omega}\big(V^{(0)}\Psi_i^{(1)} - V^{(1)}\Psi_i^{(0)}\big)\, ds + \mathcal{O}(\delta^{1/2}).
\end{aligned}$$

We substitute the last identities into (3-41) and arrive at (3-34). Thus, the condition (3-34) implies the existence of solutions to (3-35).

The functions $\widetilde{u}_\pm \in W_2^2(\omega_\pm)$ satisfy (2-5) in the sense of traces. Define

$$U^{(0)} := \widetilde{u}_\pm|_{\partial\omega}, \quad U^{(1)} := \frac{\partial \widetilde{u}_\pm}{\partial \tau}\bigg|_{\partial\omega}, \quad U^{(0)}, U^{(1)} \in L_2(\partial\omega).$$

The solution to (3-40) is defined up to a linear combination of the eigenfunctions. In view of the belongings $U^{(0)}$, $U^{(1)} \in L_2(\partial\omega)$ we can choose the mentioned linear combination of the eigenfunctions so that the condition (3-37) is satisfied. Then the solution to (3-40) is unique and the same is obviously true for (3-35). To prove the asymptotics (3-36) it is sufficient to study the smoothness of $\widetilde{u}_\pm$ at $\partial\omega$.

By standard smoothness improving theorems we conclude that $\widetilde{u}_\pm \in C^\infty(\omega)$. Moreover, given any $N > 0$, it is easy to construct the function $\widehat{u}_\pm^{(N)}$ similar to $\widehat{u}_\pm$ such that

$$\widehat{u}_\pm^{(N)}(x') = \widehat{u}_\pm(x') + \mathcal{O}(\tau^2), \quad \tau \to 0, \quad (-\Delta_{x'} - \lambda)\widehat{u}_\pm^{(N)}(x') = \chi(\tau) \sum_{j=-4}^{N} f_{j/2}^\pm(P)\tau^j + \widehat{f}_\pm^{(N)}(x'),$$

where $\widehat{f}_\pm^{(N)} \in C^\infty(\omega_\pm) \cap C^{N_1}(\overline{\omega}_\pm)$, and $N_1 = N_1(N) \to +\infty$, $N \to +\infty$. Then, proceeding as above, we can construct the solutions to (3-35) as $u_\pm = \widetilde{u}_\pm + \widehat{u}_\pm$, where $\widetilde{\boldsymbol{u}}^{(N)} := (\widetilde{u}_+^{(N)}, \widetilde{u}_-^{(N)})$ solves the equation

$$(\mathcal{H}_0 - \lambda)\widetilde{\boldsymbol{u}}^{(N)} = \widetilde{\boldsymbol{f}}^{(N)}, \quad \widetilde{\boldsymbol{f}}^{(N)} := (\widetilde{f}_+^{(N)}, \widetilde{f}_-^{(N)}), \quad \widetilde{f}_\pm^{(N)}(x') := f_\pm(x') - \chi(\tau) \sum_{j=-4}^{N} f_{j/2}^\pm(P)\tau^j - \widehat{f}_\pm^{(N)}.$$

It is clear that $\widetilde{f}_\pm^{(N)}$ belongs to $C^{N_2}(\overline{\omega}_\pm)$, where $N_2 = N_2(N) \to +\infty$ as $N \to +\infty$. Hence, by the smoothness improving theorems, $\widetilde{u}_\pm^{(N)} \in C^{N_3}(\overline{\omega}_\pm)$, $N_3 = N_3(N) \to +\infty$, $N \to +\infty$. Choosing N large enough, we arrive at the asymptotics (3-36). □

Lemma 3.5. *For all u, $v \in C^\infty(\overline{\omega})$ in a small vicinity of $\partial\omega$ the identities*

$$\operatorname{div}_{x'} \mathrm{Q}_\pm \nabla_{x'} u = \frac{1}{\det \mathrm{M}} \operatorname{div}_{(\tau,s)} (\det \mathrm{M}) \widehat{\mathrm{M}} \nabla_{(\tau,s)} h_\pm (\nabla_{(\tau,s)} h_\pm)^* \widehat{\mathrm{M}} \nabla_{(\tau,s)} u, \tag{3-43}$$

$$(\nabla_{x'} u, \nabla_{x'} v)_{\mathbb{R}^d} = \frac{\partial u}{\partial \tau}\frac{\partial v}{\partial \tau} + \nabla u \cdot (\mathrm{E} - \tau \mathrm{B} \mathrm{G}_{\partial\omega}^{-1})^{-2} \nabla v \tag{3-44}$$

hold true.

Proof. Let u, $v \in C^\infty(\overline{\omega})$ be two arbitrary functions with supports in a small vicinity $\{x' : 0 \leqslant \tau < \tau_0\}$, where τ_0 is a small fixed number. We choose τ_0 so that in this vicinity the coordinates (τ, s) are well-defined.

Taking (3-1) and (3-4) into account, we pass to the variables (τ, s) and integrate by parts to obtain

$$\begin{aligned}\int_\omega v\,\mathrm{div}_{x'}\,\mathrm{Q}_\pm \nabla_{x'} u\,dx' &= -\int_\omega (\nabla_{x'} v, \nabla_{x'} h_\pm (\nabla_{x'} h_\pm)^* \nabla_{x'} u)_{\mathbb{R}^n}\,dx' \\ &= -\int_{[0,\tau_0)\times\partial\omega} \big(\mathrm{M}^{-1}\nabla_{(\tau,s)} v, \mathrm{M}^{-1}\nabla_{(\tau,s)} h_\pm (\nabla_{(\tau,s)} h_\pm)^* \widehat{\mathrm{M}} \nabla_{(\tau,s)} u\big)_{\mathbb{R}^n} (\det \mathrm{M})\,d\tau\,ds \\ &= \int_{[0,\tau_0)\times\partial\omega} v\,\mathrm{div}_{(\tau,s)} (\det \mathrm{M}) \widehat{\mathrm{M}} \nabla_{(\tau,s)} h_\pm (\nabla_{(\tau,s)} h_\pm)^* \widehat{\mathrm{M}} \nabla_{(\tau,s)} u\,d\tau\,ds \\ &= \int_\omega v (\det{}^{-1} \mathrm{M})\,\mathrm{div}_{(\tau,s)} (\det \mathrm{M}) \widehat{\mathrm{M}} \nabla_{(\tau,s)} h_\pm (\nabla_{(\tau,s)} h_\pm)^* \widehat{\mathrm{M}} \nabla_{(\tau,s)} u\,dx',\end{aligned}$$

which proves (3-43).

The identity (3-44) follows from (3-4) and (3-30):

$$\begin{aligned}(\nabla_{x'} u, \nabla_{x'} v)_{\mathbb{R}^n} &= (\mathrm{M}^{-1}\nabla_{x'} u, \mathrm{M}^{-1}\nabla_{x'} v)_{\mathbb{R}^n} = (\nabla_{x'} u, \widehat{\mathrm{M}} \nabla_{x'} v)_{\mathbb{R}^n} \\ &= \frac{\partial u}{\partial \tau}\frac{\partial v}{\partial \tau} + \big(\nabla_s u, (\mathrm{E} - \tau \mathrm{G}_{\partial\omega}^{-1}\mathrm{B})^{-2} \mathrm{G}_{\partial\omega}^{-1} \nabla_s u\big)_{\mathbb{R}^n} = \frac{\partial u}{\partial \tau}\frac{\partial v}{\partial \tau} + \nabla u \cdot (\mathrm{E} - \tau \mathrm{B}\mathrm{G}_{\partial\omega}^{-1})^{-2} \nabla v. \quad \square\end{aligned}$$

4. Uniform resolvent convergence

In this section we prove Theorem 2.1. We begin with two auxiliary lemmas.

Lemma 4.1. *The identity* $\mathfrak{D}(\mathcal{H}_0) = W_2^2(\boldsymbol{\omega})$ *holds true and for each* $\boldsymbol{u} \in \mathfrak{D}(\mathcal{H}_0)$ *the operator* $\mathcal{H}_0$ *acts as* $\mathcal{H}_0(\boldsymbol{u}) = (-\Delta_{x'} u_+, -\Delta_{x'} u_-)$. *For each* $z \in \mathbb{C} \setminus \mathbb{R}$ *the estimate*

$$\|(\mathcal{H}_0 - z)^{-1}\|_{L_2(\boldsymbol{\omega}) \to W_2^2(\boldsymbol{\omega})} \leqslant \frac{C}{|\mathrm{Im}(z)|} \tag{4-1}$$

holds for some constant C, *where* $\mathrm{Im}(z)$ *denotes the imaginary part of* z.

Proof. The first part follows from the definitions and the considerations above for the space $W_2^2(\boldsymbol{\omega})$. The second part of the statement follows from the fact that the operator $\mathcal{H}_0$ is self-adjoint with compact resolvent. □

The description of the spectrum of $\mathcal{H}_0$ as being made up of the union of the Dirichlet and Neumann spectra is given in the following lemma, together with some properties which will be useful in the sequel.

Lemma 4.2. *The spectrum of* $\mathcal{H}_0$ *coincides with the union of spectra of* $-\Delta_\omega^{(D)}$ *and* $-\Delta_\omega^{(N)}$ *counting multiplicities. Namely, if* λ *is an* $m^{(D)}$*-multiple eigenvalue of* $-\Delta_\omega^{(D)}$ *with the associated eigenfunctions* $\psi_i^{(D)}$, $i = 1, \dots, m^{(D)}$, *and is an* $m^{(N)}$*-multiple eigenvalue of* $-\Delta_\omega^{(N)}$ *with the associated eigenfunctions* $\psi_i^{(N)}$, $i = 1, \dots, m^{(N)}$, *then* λ *is an* $(m^{(D)} + m^{(N)})$*-multiple eigenvalue of* $\mathcal{H}_0$ *with the associated eigenfunctions* $\boldsymbol{\psi}_i = (\psi_i^{(D)}, -\psi_i^{(D)})$ *and* $\boldsymbol{\psi}_i = (\psi_i^{(N)}, \psi_i^{(N)})$. *For any eigenfunction* $\boldsymbol{\psi} = (\psi_+, \psi_-)$ *of* $\mathcal{H}_0$ *we have* $\psi_\pm \in C^\infty(\overline{\omega})$ *and the asymptotics*

$$\psi_\pm(x') = \Psi^{(0)}(P) \pm \tau \Psi^{(1)}(P) + \mathcal{O}(\tau^2), \quad P \in \partial\omega,$$

where

$$\Psi^{(0)} = \psi_+|_{\partial\omega} = \psi_-|_{\partial\omega} \in C^\infty(\partial\omega), \quad \Psi^{(1)} = \left.\frac{\partial\psi_+}{\partial\tau}\right|_{\partial\omega} = -\left.\frac{\partial\psi_-}{\partial\tau}\right|_{\partial\omega} \in C^\infty(\partial\omega)$$

and

$$x' = P + \tau\nu(P) \quad \textit{for small positive } \tau.$$

Proof. Clearly if λ is an eigenvalue of $-\Delta_\omega^{(D)}$ with eigenfunction u, then λ is an eigenvalue of $\mathscr{H}_0$ with eigenfunction $(u, -u)$. Similarly, an eigenvalue of $-\Delta_\omega^{(N)}$ with eigenfunction v will also be an eigenvalue of $\mathscr{H}_0$ with eigenfunction (v, v).

Assume now that (u, v) is an eigenfunction of $\mathscr{H}_0$ and consider the functions $w_1 = u - v$ and $w_2 = u + v$. Then, provided they do not vanish identically, w_1 and w_2 will be eigenfunctions of $-\Delta_\omega^{(D)}$ and $-\Delta_\omega^{(N)}$, respectively. In case w_1 vanishes identically, then $u = v$ and u will be an eigenfunction of $-\Delta_\omega^{(N)}$, while if w_2 vanishes $u = -v$ and this will be an eigenfunction of $-\Delta_\omega^{(D)}$.

The remaining part of the lemma follows from standard arguments. □

By $L_2(\boldsymbol{\omega}, J_\varepsilon\, dx')$ we indicate the subspace of $L_2(\boldsymbol{\omega})$ consisting of the functions $\boldsymbol{u}$ with the finite norm

$$\|\boldsymbol{u}\|^2_{L_2(\boldsymbol{\omega}, J_\varepsilon dx')} = \|u_+\|^2_{L_2(\omega_+, J_\varepsilon^+ dx')} + \|u_-\|^2_{L_2(\omega_-, J_\varepsilon^- dx')}, \quad \|u_\pm\|^2_{L_2(\omega, J_\varepsilon^\pm dx')} = \int_{\omega_\pm} |u_\pm(x')|^2 J_\varepsilon^\pm(x')\, dx'.$$

In the same way we introduce the space $W_2^1(\boldsymbol{\omega}, J_\varepsilon\, dx')$ as consisting of $\boldsymbol{u} \in W_2^1(\boldsymbol{\omega})$ with the finite norm

$$\|\boldsymbol{u}\|^2_{W_2^1(\boldsymbol{\omega}, J_\varepsilon\, dx')} = \|\nabla_{x'}\boldsymbol{u}\|^2_{L_2(\boldsymbol{\omega}, J_\varepsilon\, dx')} + \|\boldsymbol{u}\|^2_{L_2(\boldsymbol{\omega}, J_\varepsilon\, dx')},$$

where $\nabla_{x'}\boldsymbol{u} = (\nabla_{x'}u_+, \nabla_{x'}u_-)$.

Lemma 4.3. *The spaces $L_2(\mathscr{S}_\varepsilon)$ and $L_2(\boldsymbol{\omega}, J_\varepsilon\, dx')$ are isomorphic and the isomorphism is the operator $\mathscr{I}_\varepsilon : L_2(\boldsymbol{\omega}, J_\varepsilon\, dx') \to L_2(\mathscr{S}_\varepsilon)$. If $\boldsymbol{u} \in W_2^1(\boldsymbol{\omega}, J_\varepsilon\, dx')$, then $\mathscr{I}_\varepsilon\boldsymbol{u} \in W_2^1(\mathscr{S}_\varepsilon)$, and if $u \in W_2^1(\mathscr{S}_\varepsilon)$, then $\mathscr{I}_\varepsilon^{-1}u \in W_2^1(\boldsymbol{\omega}, J_\varepsilon\, dx')$. The inequality*

$$\|J_\varepsilon^{-\frac{1}{2}}\nabla_{x'}\boldsymbol{u}\|_{L_2(\boldsymbol{\omega})} \leqslant \|\nabla\mathscr{I}_\varepsilon\boldsymbol{u}\|_{L_2(\mathscr{S}_\varepsilon)} \leqslant \|\nabla_{x'}\boldsymbol{u}\|_{L_2(\boldsymbol{\omega}, J_\varepsilon\, dx')} \tag{4-2}$$

holds true, where $J_\varepsilon^{-\frac{1}{2}}\nabla_{x'}\boldsymbol{u} := ((J_\varepsilon^+)^{-\frac{1}{2}}\nabla_{x'}u_+, (J_\varepsilon^-)^{-\frac{1}{2}}\nabla_{x'}u_-)$, $\boldsymbol{u} = (u_+, u_-)$.

Proof. The fact that $\mathscr{I}_\varepsilon$ is a bijection between the two spaces follows directly from its definition.

Regarding the inequalities we have

$$\begin{aligned}
\|J_\varepsilon^{-\frac{1}{2}}\nabla_{x'}\boldsymbol{u}\|^2_{L_2(\boldsymbol{\omega})} &= \int_{\omega_+} (J_\varepsilon^+)^{-1}|\nabla_{x'}u_+|^2\, dx' + \int_{\omega_-} (J_\varepsilon^-)^{-1}|\nabla_{x'}u_-|^2\, dx' \\
&= \int_{\omega_+} J_\varepsilon^+(J_\varepsilon^+)^{-2}|\nabla_{x'}u_+|^2\, dx' + \int_{\omega_-} J_\varepsilon^-(J_\varepsilon^-)^{-2}|\nabla_{x'}u_-|^2\, dx' \\
&\leq \int_{\omega_+} J_\varepsilon^+(\nabla_{x'}u_+)^* G_+^{-1}\nabla_{x'}u_+\, dx' + \int_{\omega_-} J_\varepsilon^-(\nabla_{x'}u_-)^* G_-^{-1}\nabla_{x'}u_-\, dx' \\
&= \|\nabla\mathscr{I}_\varepsilon\boldsymbol{u}\|_{L_2(\mathscr{S}_\varepsilon)} \leq \int_{\omega_+} J_\varepsilon^+|\nabla_{x'}u_+|^2\, dx' + \int_{\omega_-} J_\varepsilon^-|\nabla_{x'}u_-|^2\, dx' = \|\nabla_{x'}\boldsymbol{u}\|_{L_2(\boldsymbol{\omega}, J_\varepsilon\, dx')},
\end{aligned}$$

where we have used the knowledge of the eigenvalues of $G_\pm$ and the fact that $1 \le J_\varepsilon^\pm$. □

Define $\omega_\delta := \omega \cap \{x' : 0 < \tau < \delta\}$. We recall that the set ω^δ was introduced in (2-8), and in what follows $\boldsymbol{\omega}^\delta$ is ω^δ considered as a two-sided domain.

Lemma 4.4. *If* $\boldsymbol{u} \in W_2^1(\boldsymbol{\omega})$, *or, respectively,* $\boldsymbol{u} \in W_2^2(\boldsymbol{\omega})$, *then* $\boldsymbol{u} \in L_2(\boldsymbol{\omega}, J_\varepsilon\, dx')$, *or, respectively,* $\boldsymbol{u} \in W_2^1(\boldsymbol{\omega}, J_\varepsilon\, dx')$. *The inequalities*

$$\|\boldsymbol{u}\|_{L_2(\boldsymbol{\omega}, J_\varepsilon\, dx')} \leqslant C\|\boldsymbol{u}\|_{W_2^1(\boldsymbol{\omega})}, \tag{4-3}$$

$$\|\boldsymbol{u}\|_{L_2(\boldsymbol{\omega}_{\varepsilon^{4/3}}, J_\varepsilon\, dx')} \leqslant C\varepsilon^{2/3}\|\boldsymbol{u}\|_{W_2^1(\boldsymbol{\omega})}, \tag{4-4}$$

$$\|\boldsymbol{u}\|_{L_2(\boldsymbol{\omega}_{\varepsilon^{4/3}})} \leqslant C\varepsilon^{2/3}\|\mathscr{I}_\varepsilon \boldsymbol{u}\|_{W_2^1(\mathscr{S}_\varepsilon)}, \tag{4-5}$$

$$\|\boldsymbol{u}\|_{W_2^1(\boldsymbol{\omega}, J_\varepsilon\, dx')} \leqslant C\|\boldsymbol{u}\|_{W_2^2(\boldsymbol{\omega})},$$

$$\|\boldsymbol{u}\|_{W_2^1(\boldsymbol{\omega}_{\varepsilon^{4/3}}, J_\varepsilon\, dx')} \leqslant C\varepsilon^{2/3}\|\boldsymbol{u}\|_{W_2^2(\boldsymbol{\omega})} \tag{4-6}$$

hold true, where C *denotes positive constants independent of* ε *and* $\boldsymbol{u}$.

Proof. Let $\boldsymbol{u} \in W_2^1(\boldsymbol{\omega})$; then $u_\pm \in W_2^1(\omega)$, and for almost all $P \in \partial\omega$ the function $u_\pm(P + \cdot\,\nu(P))$ belongs to $W_2^1(0, \tau_0)$. Let $\chi = \chi(\tau)$ be an infinitely differentiable cut-off function vanishing as $\tau \geqslant \tau_0$ and being one as $\tau \leqslant \tau_0/2$. Then $u_\pm = u_\pm\chi$ for $\tau \in [0, \tau_0/2]$, and

$$u_\pm = \int_{\tau_0}^{\tau} \frac{\partial(u_\pm\chi)}{\partial\tau}\, d\tau, \quad |u_\pm(P + \tau\nu(P))|^2 \leqslant C\|u_\pm(P + \cdot\,\nu(P))\|^2_{W_2^1(0,\tau_0)}, \quad \tau \in [0, \tau_0/2],$$

where C is a positive constant independent of P and $u_\pm$. We multiply the last inequality by $J_\varepsilon^\pm$, integrate over $\partial\omega$, and take into account (3-5) to obtain

$$\int_{\partial\omega} \big|u_\pm(P + \tau\nu(P))\big|^2 |\det{}^{-1} \mathrm{M}|\, d\omega \leqslant C\|u_\pm\|^2_{W_2^1(\omega^{\tau_0})},$$

where C is a positive constant independent of $P \in \partial\omega$ and $u_\pm$. The above estimate, inequality (3-6), the definition (3-2) of $J_\varepsilon^\pm$ and the smoothness of $h_\pm$ imply

$$\int_\omega |u_\pm|^2 J_\varepsilon^\pm\, dx' = \int_{\omega_\delta} |u_\pm|^2 J_\varepsilon^\pm\, dx' + \int_{\omega^\delta} |u_\pm|^2 J_\varepsilon^\pm\, dx', \quad \delta \in (0, \tau_0/2],$$

$$\int_{\omega^\delta} |u_\pm|^2 J_\varepsilon^\pm\, dx' \leqslant C(\delta)\|u_\pm\|^2_{L_2(\omega^\delta)},$$

$$\int_{\omega_\delta} |u_\pm|^2 J_\varepsilon^\pm\, dx' = \int_0^\delta d\tau \int_{\partial\omega} |u_\pm|^2 J_\varepsilon^\pm |\det{}^{-1} \mathrm{M}|\, d\omega \leqslant C\|u_\pm\|^2_{W_2^1(\omega)} \int_0^\delta \sqrt{1 + C_4\varepsilon^2\tau^{-1}}\, d\tau, \tag{4-7}$$

where the constants C and $C(\delta)$ are independent of ε and $u_\pm$, and C is independent of δ. Taking $\delta = \tau_0/2$, we see that $\boldsymbol{u} \in L_2(\boldsymbol{\omega}, J_\varepsilon\, dx')$ and thus the estimate (4-3) holds. If we now take $\delta = \varepsilon^{4/3}$ in (4-7) instead and use the identity

$$\int_0^\delta \sqrt{1 + \varepsilon^2 C_4\tau^{-1}}\, d\tau = \mathring{J}_\varepsilon^\pm(\delta) := \sqrt{\delta^2 + C_4\varepsilon^2\delta} + \frac{C_4}{2}\varepsilon^2 \ln \frac{C_4\varepsilon^2 + 2\delta + 2\sqrt{\delta^2 + C_4\varepsilon^2\delta}}{C_4\varepsilon^2},$$

we obtain (4-4).

Let us prove (4-5). We integrate by parts as follows:

$$\begin{aligned}
&\int_{\omega_{\varepsilon^{4/3}}} |u_\pm|^2 J_\varepsilon^\pm \, dx' \leqslant C \int_{\partial\omega} d\omega \int_0^{\varepsilon^{4/3}} |u_\pm|^2 J_\varepsilon^\pm \, d\tau, \\
&\int_0^{\varepsilon^{4/3}} |u_\pm|^2 J_\varepsilon^\pm \, d\tau = |u_\pm|^2 \mathring{J}_\varepsilon^\pm \Big|_{\tau=0}^{\tau=\varepsilon^{4/3}} - 2\int_0^{\varepsilon^{4/3}} \mathring{J}_\varepsilon^\pm(\tau) \operatorname{Re} u_\pm \frac{\partial \overline{u}_\pm}{\partial\tau} \, d\tau \\
&\qquad\qquad \leqslant \mathring{J}_\varepsilon^\pm(\varepsilon^{4/3}) \biggl(|u_\pm|^2\Big|_{\tau=\varepsilon^{4/3}} + \int_0^{\varepsilon^{4/3}} |u_\pm|^2 J_\varepsilon^\pm \, d\tau + \int_0^{\varepsilon^{4/3}} \frac{1}{J_\varepsilon^\pm} \left| \frac{\partial u_\pm}{\partial\tau} \right|^2 d\tau \biggr), \\
&\int_{\omega_{\varepsilon^{4/3}}} |u_\pm|^2 J_\varepsilon^\pm \, dx' \leqslant C\varepsilon^{4/3} \biggl(\int_{\partial\omega} |u_\pm|^2\Big|_{\tau=\varepsilon^{4/3}} d\omega + \int_{\omega_{\varepsilon^{4/3}}} \Bigl(\frac{1}{J_\varepsilon^\pm} |\nabla_{x'} u_\pm|^2 + J_\varepsilon^\pm |u_\pm|^2 \Bigr) dx' \biggr).
\end{aligned}$$

By the embedding of $W_2^1(\omega^{\varepsilon^{4/3}})$ into $L_2(\{x : \tau = \varepsilon^{4/3}\})$ we have the estimate

$$\int_{\partial\omega} |u_\pm|^2\Big|_{\tau=\varepsilon^{4/3}} d\omega \leqslant C \|u_\pm\|^2_{W_2^1(\omega^{\varepsilon^{4/3}})} \leqslant C \|\mathcal{I}_\varepsilon \boldsymbol{u}\|^2_{W_2^1(\mathcal{S}_\varepsilon)},$$

where the constants C are independent of ε and $\boldsymbol{u}$. These two last estimates together with (4-2) yield (4-5).

To prove the second part of the lemma related to the case $\boldsymbol{u} \in W_2^2(\omega)$ it is sufficient to note that since $u_\pm, \nabla_{x'} u_\pm \in W_2^1(\omega)$, by the first part of the lemma these functions belong to $L_2(\omega, J_\varepsilon^\pm \, dx')$, and the estimates (4-3), (4-4) are valid for $\boldsymbol{u}$ replaced by $\nabla_{x'}\boldsymbol{u}$. This completes the proof. □

Proof of Theorem 2.1. Let $f \in L_2(\mathcal{S}_\varepsilon)$; then $\boldsymbol{f} := \mathcal{I}_\varepsilon f \in L_2(\boldsymbol{\omega}, J_\varepsilon \, dx') \subset L_2(\boldsymbol{\omega})$. Let $u^{(\varepsilon)} := (\mathcal{H}_\varepsilon - z)^{-1} f$, $\boldsymbol{u}^{(0)} := (\mathcal{H}_0 - z)^{-1} \mathcal{I}_\varepsilon^{-1} f$. By the definitions of $\mathcal{H}_\varepsilon$ and $\mathcal{H}_0$ we have

$$\mathfrak{h}_\varepsilon[u^{(\varepsilon)}, \varphi] - z(u^{(\varepsilon)}, \varphi)_{L_2(\mathcal{S}_\varepsilon)} = (f, \varphi)_{L_2(\mathcal{S}_\varepsilon)} \quad \text{for each } \varphi \in W_2^1(\mathcal{S}_\varepsilon), \tag{4-8}$$

$$\mathfrak{h}_0[\boldsymbol{u}^{(0)}, \boldsymbol{\varphi}] - z(\boldsymbol{u}^{(0)}, \boldsymbol{\varphi})_{L_2(\boldsymbol{\omega})} = (\boldsymbol{f}, \boldsymbol{\varphi})_{L_2(\boldsymbol{\omega})} \quad \text{for each } \varphi \in W_2^1(\boldsymbol{\omega}). \tag{4-9}$$

Since $\boldsymbol{u}^{(0)} \in W_2^2(\boldsymbol{\omega})$, by Lemmas 3.1 and 4.4, $u^{(0)} := \mathcal{I}_\varepsilon \boldsymbol{u}^{(0)} \in W_2^1(\mathcal{S}_\varepsilon)$. Hence, $v^{(\varepsilon)} := u^{(\varepsilon)} - u^{(0)} \in W_2^1(S_\varepsilon)$ and this can be used as a test function in (4-8):

$$\mathfrak{h}_\varepsilon[u^{(\varepsilon)}, v^{(\varepsilon)}] - z(u^{(\varepsilon)}, v^{(\varepsilon)})_{L_2(S_\varepsilon)} = (f, v^{(\varepsilon)})_{L_2(S_\varepsilon)}.$$

The identity $u^{(\varepsilon)} = v^{(\varepsilon)} + u^{(0)}$ yields

$$\|\nabla v^{(\varepsilon)}\|^2_{L_2(S_\varepsilon)} - z\|v^{(\varepsilon)}\|^2_{L_2(S_\varepsilon)} = (f, v_\varepsilon)_{L_2(S_\varepsilon)} - (\nabla u^{(0)}, \nabla v^{(\varepsilon)})_{L_2(S_\varepsilon)} + z(u^{(0)}, v^{(\varepsilon)})_{L_2(S_\varepsilon)}. \tag{4-10}$$

We parametrize S_ε as $x' = x'$, $x_{n+1} = \pm\varepsilon h_\pm(x')$, and use the definition of the scalar product of $\nabla u^{(0)}$ and $\nabla v^{(\varepsilon)}$ in $L_2(S_\varepsilon)$. It implies

$$\begin{aligned}
&(f, v^{(\varepsilon)})_{L_2(S_\varepsilon)} - (\nabla u^{(0)}, \nabla v^{(\varepsilon)})_{L_2(S_\varepsilon)} + z(u^{(0)}, v^{(\varepsilon)})_{L_2(S_\varepsilon)} \\
&\qquad = (f_+, J_\varepsilon^+ v_+^{(\varepsilon)})_{L_2(\omega_+)} + (f_-, J_\varepsilon^- v_-^{(\varepsilon)})_{L_2(\omega_-)} - \bigl((J_\varepsilon^+ G_+^{-1} \nabla_{x'} u_+^{(0)}, \nabla_{x'} v_+^{(\varepsilon)})_{L_2(\omega_+)} \\
&\qquad\qquad + (J_\varepsilon^- G_-^{-1} \nabla_{x'} u_-^{(0)}, \nabla_{x'} v_-^{(\varepsilon)})_{L_2(\omega_-)} \bigr) + z(u_+^{(0)}, J_\varepsilon^+ v_+^{(\varepsilon)})_{L_2(\omega_+)} + z(u_-^{(0)}, J_\varepsilon^- v_-^{(\varepsilon)})_{L_2(\omega_-)},
\end{aligned}$$

where $\boldsymbol{v}^{(\varepsilon)} = (v_+^{(\varepsilon)}, v_-^{(\varepsilon)}) = \mathcal{I}_\varepsilon^{-1} v^{(\varepsilon)}$ and $G_\pm^{ij}$ are the entries of the inverse matrix $G_\pm^{-1}$. We substitute the last formula into (4-10) and then sum it with (4-9), where we take $\boldsymbol{\varphi} = \boldsymbol{v}^{(\varepsilon)} \in W_2^1(\boldsymbol{\omega}, J_\varepsilon\, dx') \subset W_2^1(\boldsymbol{\omega})$:

$$\|\nabla v^{(\varepsilon)}\|^2_{L_2(S_\varepsilon)} - z\|v^{(\varepsilon)}\|^2_{L_2(S_\varepsilon)} = R^+ + R^-, \tag{4-11}$$

$$\begin{aligned} R^\pm := (f_\pm, (J_\varepsilon^\pm - 1)v_\pm^{(\varepsilon)})_{L_2(\omega)} - (J_\varepsilon^\pm G_\pm^{-1} \nabla_{x'} u_\pm^{(0)}, \nabla_{x'} v_\pm^{(\varepsilon)})_{L_2(\omega_\pm)} \\ - (\nabla_{x'} u_\pm^{(0)}, \nabla_{x'} v_\pm^{(\varepsilon)})_{L_2(\omega)} + z(u_\pm^{(0)}, (J_\varepsilon^\pm - 1)v_\pm^{(\varepsilon)})_{L_2(\omega)}. \end{aligned}$$

Let us estimate $R^\pm$ which we shall write as

$$R^\pm = R_1^\pm + R_2^\pm, \tag{4-12}$$

where

$$\begin{aligned} R_1^\pm &:= (f_\pm, (J_\varepsilon^\pm - 1)v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)} - (J_\varepsilon^\pm G_\pm^{-1} \nabla_{x'} u_\pm^{(0)}, \nabla_{x'} v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)} \\ &\qquad - (\nabla_{x'} u_\pm^{(0)}, \nabla_{x'} v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)} + z(u_\pm^{(0)}, (J_\varepsilon^\pm - 1)v^{(\varepsilon)})_{L_2(\omega^\delta)}, \\ R_2^\pm &:= (f_\pm, (J_\varepsilon^\pm - 1)v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)} - (J_\varepsilon^\pm G_\pm^{-1} \nabla_{x'} u_\pm^{(0)}, \nabla_{x'} v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)} \\ &\qquad - (\nabla_{x'} u_\pm^{(0)}, \nabla_{x'} v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)} + z(u_\pm^{(0)}, (J_\varepsilon^\pm - 1)v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)}, \end{aligned}$$

and $\delta := \varepsilon^{4/3}$. As $x' \in \omega_\delta$, by (3-6) we have

$$\varepsilon^2|\nabla_{x'} h_\pm|^2 \leqslant C\varepsilon^{2/3}, \quad \|\mathrm{G}_\pm^{-1} - \mathrm{E}\| \leqslant C\varepsilon^{2/3}, \quad |J_\varepsilon^\pm - 1| \leqslant C\varepsilon^{2/3}, \quad |(J_\varepsilon^\pm)^{-1} - 1| \leqslant C\varepsilon^{2/3}.$$

Hereinafter by C we indicate nonessential positive constants independent of ε, $u^{(\varepsilon)}$, $\boldsymbol{u}^{(0)}$, and f. Hence, by Lemmas 3.1, 4.4 and Schwarz's inequality,

$$\begin{gathered} \big|(f_\pm, (J_\varepsilon^\pm - 1)v_\pm^{(\varepsilon)})_{L_2(\omega_\delta)}\big| \leqslant C\varepsilon^{2/3}\|f_\pm\|_{L_2(\omega, J_\varepsilon^\pm dx')}\|v_\pm^{(\varepsilon)}\|_{L_2(\omega, J_\varepsilon^\pm dx')} \leqslant C\varepsilon^{2/3}\|f\|_{L_2(S_\varepsilon)}\|v^{(\varepsilon)}\|_{L_2(S_\varepsilon)}, \\ \big|z(u_\pm^{(0)}, (J_\varepsilon^\pm - 1)v_\pm^{(\varepsilon)})_{L_2(\omega_\delta)}\big| \leqslant C\varepsilon^{2/3}\|u^{(0)}\|_{L_2(\omega)}\|v^{(\varepsilon)}\|_{L_2(S_\varepsilon)}, \end{gathered}$$

$$\begin{aligned} &\big|(\nabla_{x'} u_\pm^{(0)}, \nabla_{x'} v_\pm^{(\varepsilon)})_{L_2(\omega_\delta)} - (J_\varepsilon^\pm G_\pm^{-1} \nabla_{x'} u_\pm^{(0)}, \nabla_{x'} v_\pm^{(\varepsilon)})_{L_2(\omega_\delta)}\big| \\ &\qquad \leqslant C\varepsilon^{2/3}\|\boldsymbol{u}^{(0)}\|_{W_2^1(\boldsymbol{\omega})}\|\nabla_{x'} v_\pm^{(\varepsilon)}\|_{L_2(\omega_\delta)} \leqslant C\varepsilon^{2/3}\|\boldsymbol{u}^{(0)}\|_{W_2^1(\boldsymbol{\omega})}\|J_\varepsilon^{-\frac{1}{2}} \nabla_{x'} \boldsymbol{v}^{(\varepsilon)}\|_{L_2(\boldsymbol{\omega}_\delta)} \\ &\qquad \leqslant C\varepsilon^{2/3}\|\boldsymbol{u}^{(0)}\|_{W_2^1(\boldsymbol{\omega})}\|J_\varepsilon^{-1} \nabla_{x'} \boldsymbol{v}^{(\varepsilon)}\|_{L_2(\boldsymbol{\omega}_\delta, J_\varepsilon\, dx')} \leqslant C\varepsilon^{2/3}\|\boldsymbol{u}^{(0)}\|_{W_2^1(\boldsymbol{\omega})}\|\nabla v^{(\varepsilon)}\|_{L_2(S_\varepsilon)}, \end{aligned}$$

and therefore

$$|R_1^+ + R_1^-| \leqslant C\varepsilon^{2/3}\|\boldsymbol{u}^{(0)}\|_{W_2^1(\boldsymbol{\omega})}\|v^{(\varepsilon)}\|_{W_2^1(S_\varepsilon)}. \tag{4-13}$$

To estimate $R_2^\pm$ we employ (4-3), (4-4), (4-5). We begin with the first term in $R_2^\pm$ applying again Schwarz's inequality and (4-5) to obtain

$$\begin{aligned} |(f_\pm, (J_\varepsilon^\pm - 1)v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)}| &\leqslant \|f_\pm\|_{L_2(\omega^\delta, J_\varepsilon^\pm\, dx')}\|\big(1 - (J_\varepsilon^\pm)^{-1}\big)v_\pm^{(\varepsilon)}\|_{L_2(\omega^\delta, J_\varepsilon^\pm\, dx')} \\ &\leqslant \|f\|_{L_2(S_\varepsilon)}\|v_\pm^{(\varepsilon)}\|_{L_2(\omega^\delta, J_\varepsilon^\pm\, dx')} \leqslant C\varepsilon^{2/3}\|f\|_{L_2(S_\varepsilon)}\|v^{(\varepsilon)}\|_{W_2^1(S_\varepsilon)}. \end{aligned} \tag{4-14}$$

Employing (4-2), (4-3) and (4-5) in the same way we get two more estimates:

$$\begin{aligned}\bigl|z\bigl(u_\pm^{(0)},(J_\varepsilon^\pm-1)v^{(\varepsilon)}\bigr)_{L_2(\omega^\delta)}\bigr| &\leqslant C\|u_\pm^{(0)}\|_{L_2(\omega^\delta,J_\varepsilon^\pm\,dx')}\|v_\pm^{(\varepsilon)}\|_{L_2(\omega^\delta,J_\varepsilon^\pm\,dx')}\\ &\leqslant C\varepsilon^{2/3}\|\boldsymbol{u}^{(0)}\|_{W_2^1(\boldsymbol{\omega})}\|v^{(\varepsilon)}\|_{W_2^1(S_\varepsilon)},\\ \bigl|(\nabla_{x'}u_\pm^{(0)},\nabla_{x'}v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)}\bigr| &\leqslant \|(J_\varepsilon^\pm)^{\frac{1}{2}}\nabla_{x'}u_\pm^{(0)}\|_{L_2(\omega^\delta)}\|(J_\varepsilon^\pm)^{-\frac{1}{2}}\nabla_{x'}v_\pm^{(\varepsilon)}\|_{L_2(\omega^\delta)}\\ &\leqslant C\varepsilon^{2/3}\|\boldsymbol{u}^{(0)}\|_{W_2^2(\boldsymbol{\omega})}\|\nabla v^{(\varepsilon)}\|_{L_2(S_\varepsilon)}. \qquad (4\text{-}15)\end{aligned}$$

Since

$$(G_\pm^{-1}\nabla_{x'}u_\pm^{(0)},\nabla_{x'}v_\pm^{(\varepsilon)})_{\mathbb{R}^n}=\nabla\mathcal{I}_\varepsilon\boldsymbol{u}^{(0)}\cdot\nabla v^{(\varepsilon)},$$

by Schwarz's inequality we have

$$\begin{aligned}\bigl|(G_\pm^{-1}\nabla_{x'}u_\pm^{(0)},\nabla_{x'}v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)}\bigr| &\leqslant \|\nabla v^{(\varepsilon)}\|_{L_2(S_\varepsilon)}(G_\pm^{-1}\nabla_{x'}u_\pm^{(0)},\nabla_{x'}u_\pm^{(0)})^{\frac{1}{2}}_{L_2(\omega^\delta)}\\ &\leqslant \|\nabla v^{(\varepsilon)}\|_{L_2(S_\varepsilon)}\|(J_\varepsilon^\pm)^{\frac{1}{2}}\nabla_{x'}u_\pm^{(0)}\|_{L_2(\omega^\delta)}.\end{aligned}$$

Here we have used the inequality

$$\sum_{i,j=1}^{n}G_\pm^{ij}\xi_i\xi_j\leqslant\sum_{i=1}^{n}|\xi_i|^2,$$

which follows from Lemma 3.1. Using (4-6) we get

$$\bigl|(G_\pm^{-1}\nabla_{x'}u_\pm^{(0)},\nabla_{x'}v_\pm^{(\varepsilon)})_{L_2(\omega^\delta)}\bigr|\leqslant\|\nabla v^{(\varepsilon)}\|_{L_2(S_\varepsilon)}\|\boldsymbol{u}^{(0)}\|_{W_2^1(\boldsymbol{\omega}^\delta)}\leqslant C\varepsilon^{2/3}\|\nabla v^{(\varepsilon)}\|_{L_2(S_\varepsilon)}\|\boldsymbol{u}^{(0)}\|_{W_2^2(\boldsymbol{\omega})},$$

which with (4-14) and (4-15) yields

$$|R_2^+ + R_2^-|\leqslant C\varepsilon^{2/3}\|\boldsymbol{u}^{(0)}\|_{W_2^2(\boldsymbol{\omega})}\|v^{(\varepsilon)}\|_{W_2^1(S_\varepsilon)}.$$

Together with (4-1), (4-11), (4-12), (4-13) it follows that

$$\bigl|\|\nabla v^{(\varepsilon)}\|^2_{L_2(S_\varepsilon)}-z\|v^{(\varepsilon)}\|^2_{L_2(S_\varepsilon)}\bigr|\leqslant C\varepsilon^{2/3}\|\boldsymbol{u}^{(0)}\|_{W_2^2(\boldsymbol{\omega})}\|v^{(\varepsilon)}\|_{W_2^1(S_\varepsilon)}\leqslant C\varepsilon^{2/3}\|\boldsymbol{f}\|_{L_2(\boldsymbol{\omega})}\|v^{(\varepsilon)}\|_{W_2^1(S_\varepsilon)}.$$

Since

$$\bigl|\|\nabla v^{(\varepsilon)}\|^2_{L_2(S_\varepsilon)}-z\|v^{(\varepsilon)}\|^2_{L_2(S_\varepsilon)}\bigr|\geqslant C\|v^{(\varepsilon)}\|^2_{W_2^1(S_\varepsilon)},$$

we arrive at (2-6), completing the proof. □

Remark 4.5. The proof above uses the estimates from Lemma 4.4 which include a measure of the boundary behavior by means of the weight function J_ε. A different approach which may also be used to prove convergence of the resolvent in similar situations is based on inequalities of Hardy type instead, possibly allowing for a better control of the behavior near the boundary — see [Krejčiřík and Zuazua 2010] for an illustration of this principle.

In the proof of Theorem 2.4 in the next section we shall use the following auxiliary lemma which is convenient to prove in this section.

Lemma 4.6. *Let* λ *be a* m*-multiple eigenvalue of* $\mathcal{H}_0$, *and* $\lambda_i(\varepsilon)$, $i = 1, \dots, m$, *be the eigenvalues of* $\mathcal{H}_\varepsilon$ *taken counting multiplicity and converging to* λ, *and* $\psi_\varepsilon^{(i)}$ *be the associated eigenfunctions orthonormalized in* $L_2(S_\varepsilon)$. *For* z *close to* λ *the representation*

$$(\mathcal{H}_\varepsilon - z)^{-1} = \sum_{i=1}^{m} \frac{\psi_\varepsilon^{(i)}}{\lambda_i(\varepsilon) - z} (\,\cdot\,, \psi_\varepsilon^{(i)})_{L_2(S_\varepsilon)} + \mathcal{R}_\varepsilon(z)$$

holds true, *where the operator* $\mathcal{R}_\varepsilon(z) : L_2(S_\varepsilon) \to W_2^1(S_\varepsilon)$ *is bounded uniformly in* ε *and* z. *The range of* $\mathcal{R}_\varepsilon(z)$ *is orthogonal to all* $\psi_\varepsilon^{(i)}$, $i = 1, \dots, m$.

Proof. We choose a fixed δ so that the disk $B_\delta(\lambda) := \{z : |z - \lambda| < \delta\}$ contains no eigenvalues of $\mathcal{H}_0$ except λ and

$$\operatorname{dist}\{\partial B_\delta(\lambda), \sigma_d(\mathcal{H}_0)\} \geqslant \delta.$$

Then, by Theorem 2.3, for sufficiently small ε this disk contains the eigenvalues $\lambda_i(\varepsilon)$, $i = 1, \dots, m$, and no other eigenvalues of $\mathcal{H}_\varepsilon$, and

$$\operatorname{dist}\bigl\{B_\delta(\lambda), \sigma_d(\mathcal{H}_\varepsilon) \setminus \{\lambda_i(\varepsilon), i = 1, \dots, m\}\bigr\} \geqslant \frac{\delta}{2}. \tag{4-16}$$

Denote by V_ε the orthogonal complement to $\psi_\varepsilon^{(i)}$, $i = 1, \dots, m$, in $L_2(S_\varepsilon)$. By [Kato 1966, Chapter V, Section 3.5, Equations (3.21)] the representation (3-29) holds true, where $\mathcal{R}_\varepsilon(z)$ is the part of the resolvent $(\mathcal{H}_\varepsilon - z)^{-1}$ acting in V_ε and

$$\|\mathcal{R}_\varepsilon(z)\|_{V_\varepsilon \to V_\varepsilon} \leqslant \frac{1}{\operatorname{dist}\bigl\{B_\delta(\lambda), \sigma_d(\mathcal{H}_\varepsilon) \setminus \{\lambda_i(\varepsilon), i = 1, \dots, m\}\bigr\}} \leqslant \frac{2}{\delta} \tag{4-17}$$

for $z \in B_\delta(\lambda)$, where we have used (4-16). Hence, the range of $\mathcal{R}_\varepsilon(z)$ is orthogonal to $\psi_\varepsilon^{(i)}$, $i = 1, \dots, m$. It is easy to check that the function $u_\varepsilon := \mathcal{R}_\varepsilon(z) f$, $f \in L_2(S_\varepsilon)$, solves the equation

$$(\mathcal{H}_\varepsilon - z) u_\varepsilon = f_\varepsilon, \quad f_\varepsilon := f - \sum_{i=1}^{m} \psi_\varepsilon^{(i)} (f, \psi_\varepsilon^{(i)})_{L_2(S_\varepsilon)}, \quad \|f_\varepsilon\|_{L_2(S_\varepsilon)} \leqslant \|f\|_{L_2(S_\varepsilon)}.$$

Hence, by the definition of $\mathcal{H}_\varepsilon$ and (4-17),

$$\|\nabla u_\varepsilon\|^2_{L_2(S_\varepsilon)} = z\|u_\varepsilon\|^2_{L_2(S_\varepsilon)} + (f_\varepsilon, u_\varepsilon)_{L_2(S_\varepsilon)} \leqslant |z|\,\|u_\varepsilon\|^2_{L_2(S_\varepsilon)} + \|f_\varepsilon\|_{L_2(S_\varepsilon)} \|u_\varepsilon\|_{L_2(S_\varepsilon)} \leqslant C(\delta)\|f\|^2_{L_2(S_\varepsilon)},$$

where the constant $C(\delta)$ is independent of ε and f. The last estimate and (4-17) complete the proof. ☐

5. Asymptotic expansions

In this section we give the proof of Theorem 2.4 which will be divided into two parts. We first build the asymptotic expansions formally, where the core of the formal construction is the method of matching asymptotic expansions [Il′in 1992]. The second part is devoted to the justification of the asymptotics, i.e., obtaining estimates for the error terms.

The formal construction consists of determining the outer and inner expansions on the base of the perturbed eigenvalue problem and the matching of these expansions. The outer expansion is used to approximate the perturbed eigenfunctions outside a small neighborhood of $\partial\omega$. It is constructed in terms of the variables x' using the first parametrization of $\mathcal{S}_\varepsilon$ given in the previous sections. In a vicinity of $\partial\omega$ the perturbed eigenfunctions are approximated by the inner expansion which is based on the second parametrization of $\mathcal{S}_\varepsilon$ and is constructed in terms of the variables (ξ, s).

Outer expansion: First term. By Theorem 2.3 there exist exactly m eigenvalues of $\mathcal{H}_\varepsilon$ converging to λ counting multiplicities. We denote these eigenvalues by $\lambda_k(\varepsilon)$, $k = 1, \dots, m$, while the symbols $\psi_\varepsilon^{(k)}$ will denote the associated eigenfunctions. We construct the asymptotics for $\lambda_k(\varepsilon)$ as

$$\lambda_k(\varepsilon) = \lambda + \varepsilon^2 \ln\varepsilon\, \mu_k\left(\frac{1}{\ln\varepsilon}\right) + \cdots . \tag{5-1}$$

Hereinafter terms like $\ln\varepsilon A$ are understood as $(\ln\varepsilon)A$. In accordance with the method of matching asymptotic expansions we form the asymptotics for $\psi_\varepsilon^{(k)}$ as the sum of outer and inner expansions. The outer expansion is built as

$$\psi_{\varepsilon,\mathrm{ex}}^{(k)} = \mathcal{I}_\varepsilon(\boldsymbol{\psi}_k + \varepsilon^2 \ln\varepsilon\, \boldsymbol{\phi}_k + \cdots), \tag{5-2}$$

where $\boldsymbol{\phi}_k = (\phi_+^{(k)}, \phi_-^{(k)})$, $\phi_\pm^{(k)} = \phi_\pm^{(k)}(x', \varepsilon)$, and the eigenfunctions $\boldsymbol{\psi}_k$ are chosen as described before the statement of Theorem 2.4. We also recall that these functions depend on ε in the case where λ is a multiple eigenvalue.

We substitute the identities (5-1), (5-2), and (3-3) into the eigenvalue equation

$$\mathcal{H}_\varepsilon \psi_\varepsilon^{(k)} = \lambda_k(\varepsilon)\psi_\varepsilon^{(k)}, \tag{5-3}$$

and take into account the eigenvalue equations for $\boldsymbol{\psi}_i$. It implies the equations for $\boldsymbol{\phi}_k$, namely,

$$\begin{gathered}(-\Delta_{x'} - \lambda)\phi_\pm^{(k)} = \frac{1}{\ln\varepsilon} f_{2,\pm}^{(k)} + \mu_k \psi_\pm^{(k)}, \quad x' \in \omega_\pm, \qquad f_{2,\pm}^{(k)} := \mathcal{H}_\pm^{(2)} \psi_\pm^{(k)}, \\ \mathcal{H}_\pm^{(2)} := -\operatorname{div}_{x'} \mathrm{Q}_\pm \nabla_{x'} - \frac{|\nabla_{x'} h_\pm|^2}{2} \Delta_{x'} + \tfrac{1}{2} \operatorname{div}_{x'} |\nabla_{x'} h_\pm|^2 \nabla_{x'}.\end{gathered} \tag{5-4}$$

The functions $\psi_\pm^{(i)}$ are infinitely differentiable in $\overline{\omega}_\pm$, and thus

$$\psi_\pm^{(k)}(x', \varepsilon) = \Psi_k^{(0)}(P, \varepsilon) \pm \Psi_k^{(1)}(P, \varepsilon)\tau + \Psi_k^{(2,\pm)}(P, \varepsilon)\tau^2 + \mathcal{O}(\tau^3), \quad P \in \partial\omega, \tag{5-5}$$

as $\tau \to +0$, where, by the definition of the domain of $\mathcal{H}_0$,

$$\begin{gathered}\Psi_k^{(0)} := \psi_+^{(k)}\big|_{\partial\omega} = \psi_-^{(k)}\big|_{\partial\omega}, \quad \Psi_k^{(1)} := \frac{\partial\psi_+^{(k)}}{\partial\tau}\bigg|_{\partial\omega} = -\frac{\partial\psi_-^{(k)}}{\partial\tau}\bigg|_{\partial\omega}, \quad \Psi_k^{(2,\pm)} := \frac{1}{2}\frac{\partial^2\psi_\pm^{(k)}}{\partial\tau^2}\bigg|_{\partial\omega}, \\ \Psi_k^{(j)}, \Psi_k^{(2,\pm)} \in C^\infty(\partial\omega).\end{gathered}$$

The functions $\Psi_k^{(i)}$ depend on ε only if λ is a multiple eigenvalue, since the same is true for the functions $\boldsymbol{\psi}_k$.

In view of the identity (3-12) we rewrite (5-5) as

$$\begin{aligned}
&\psi_{\pm}^{(k)}(x',\varepsilon)=\Psi_k^{(0)}(P,\varepsilon)\pm\Psi_k^{(1)}(P,\varepsilon)\zeta^2+\Psi_k^{(2,\pm)}(P,\varepsilon)\zeta^4+\mathcal{O}(\zeta^6), && \zeta\to+0.\\
&\psi_{\pm}^{(k)}(x',\varepsilon)=\Psi_k^{(0)}(P,\varepsilon)\pm\varepsilon^2\Psi_k^{(1)}(P,\varepsilon)\xi^2+\varepsilon^4\Psi_k^{(2,\pm)}(P,\varepsilon)\xi^4+\mathcal{O}(\varepsilon^6\xi^6), && \varepsilon\xi\to 0.
\end{aligned}\tag{5-6}$$

Inner expansion. In accordance with the method of matching asymptotic expansions the identities (5-2), (5-6) yield that the inner expansion for the eigenfunctions $\psi_\varepsilon^{(k)}$ should read

$$\psi_{\varepsilon,\,\mathrm{in}}^{(k)}(\xi,P,\varepsilon)=\sum_{i=0}^{4}\varepsilon^i v_i^{(k)}(\xi,P,\varepsilon)+\cdots,\tag{5-7}$$

where the coefficients must satisfy the following asymptotics as $\xi\to\pm\infty$:

$$v_0^{(k)}(\xi,P,\varepsilon)=\Psi_k^{(0)}(P,\varepsilon)+o(1),\tag{5-8}$$

$$v_1^{(k)}(\xi,P,\varepsilon)=o(|\xi|),\tag{5-9}$$

$$v_2^{(k)}(\xi,P,\varepsilon)=\pm\Psi_k^{(1)}(P,\varepsilon)\xi^2+o(|\xi|^2),\tag{5-10}$$

$$v_3^{(k)}(\xi,P,\varepsilon)=o(|\xi|^3),$$

$$v_4^{(k)}(\xi,P,\varepsilon)=\Psi_k^{(2,\pm)}(P,\varepsilon)\xi^4+o(|\xi|^4).$$

These asymptotics mean that the first term of the outer expansion is matched with the inner expansion.

We substitute (5-1), (5-7), (3-25), (3-21) into the eigenvalue equation (5-3) and equate the coefficients of ε^{-4}. This implies the equation for $v_0^{(k)}$:

$$\mathcal{L}_{-4}v_0^{(k)}\equiv-\frac{1}{\sqrt{4\xi^2+b_1^2}}\frac{\partial}{\partial\xi}\frac{1}{\sqrt{4\xi^2+b_1^2}}\frac{\partial v_0^{(k)}}{\partial\xi}=0\quad\text{on }\mathbb{R}\times\partial\omega.$$

The solution to the last equation satisfying (5-8) is obviously

$$v_0^{(k)}(\xi,P,\varepsilon)\equiv\Psi_k^{(0)}(P,\varepsilon).\tag{5-11}$$

We then substitute this identity and (5-1), (5-7), (3-25), (3-26), (3-27), (3-25) into (5-3) and equate the coefficients at ε^i, $i=-3,\dots,0$, leading us to the equations for $v_i^{(k)}$, $i=1,\dots,4$:

$$\mathcal{L}_{-4}v_1^{(k)}=0\qquad\text{on }\mathbb{R}\times\partial\omega,\tag{5-12}$$

$$\mathcal{L}_{-4}v_2^{(k)}=0\qquad\text{on }\mathbb{R}\times\partial\omega,\tag{5-13}$$

$$\mathcal{L}_{-4}v_3^{(k)}+\mathcal{L}_{-3}v_2^{(k)}+\mathcal{L}_{-2}v_1^{(k)}=0\qquad\text{on }\mathbb{R}\times\partial\omega,\tag{5-14}$$

$$\mathcal{L}_{-4}v_4^{(k)}+\mathcal{L}_{-3}v_3^{(k)}+\mathcal{L}_{-2}v_2^{(k)}+\mathcal{L}_{-1}v_1^{(k)}+\mathcal{L}_0v_0^{(k)}=\lambda v_0^{(k)}\qquad\text{on }\mathbb{R}\times\partial\omega,\tag{5-15}$$

where we have used that

$$\mathcal{L}_i v_0^{(k)}\equiv 0,\quad i=-3,\dots,-1,$$

due to (3-26), (3-27), (5-11). The only solution to (5-12) satisfying (5-9) is independent of ξ:

$$v_1^{(k)}(\xi, P, \varepsilon) \equiv C_1^{(k,0)}(P, \varepsilon), \tag{5-16}$$

where $C_1^{(k,0)}$ is an unknown function to be determined.

Equation (5-13) can be solved, and the solution satisfying (5-10) is

$$v_2^{(k)}(\xi, P, \varepsilon) = \Psi_k^{(1)}(P, \varepsilon) X_1(\xi, b_1(P)) + C_2^{(k,0)}(P, \varepsilon), \tag{5-17}$$

$$X_1(\xi, b) := \tfrac{1}{2}\xi(4\xi^2 + b^2)^{\frac{1}{2}} + \frac{b^2}{4}\ln\bigl(2\xi + (4\xi^2 + b^2)^{\frac{1}{2}}\bigr) - \frac{b^2}{4}\ln b, \tag{5-18}$$

where $C_2^{(k,0)}$ is an unknown function to be determined.

In view of (5-16), (5-17), (3-26), (3-27) and (5-13), Equation (5-14) may be written as

$$\beta_{-4}\frac{\partial}{\partial \xi}\beta_{-4}\frac{\partial v_3^{(k)}}{\partial \xi} = -\beta_{-4}\frac{\partial}{\partial \xi}\beta_{-3}\frac{\partial v_2^{(k)}}{\partial \xi} \quad \text{on } \mathbb{R} \times \partial\omega.$$

Employing the formulas (3-21), (5-17) and (5-18), we solve the last equation:

$$\begin{aligned} v_3^{(k)}(\xi, P, \varepsilon) &= \frac{\Psi_0^{(k,1)}(P, \varepsilon) b_1(P) b_2(P)}{2\beta_{-4}(\xi, P)} + C_3^{(k,1)}(P, \varepsilon) X_1(\xi) + C_3^{(k,0)}(P, \varepsilon) \\ &= \tfrac{1}{2}\Psi_k^{(1)}(P, \varepsilon) b_1(P) b_2(P)(4\xi^2 + b_1^2(P))^{\frac{1}{2}} + C_3^{(k,1)}(P, \varepsilon) X_1(\xi) + C_3^{(k,0)}(P, \varepsilon), \end{aligned} \tag{5-19}$$

where $C_3^{(k,1)}$ and $C_3^{(k,0)}$ are unknown functions to be determined.

We substitute (5-16), (5-17), (5-18), (5-19), (3-26), (3-27), (3-28), (3-19) and (3-21) into (5-15) and then solve it to obtain

$$\begin{aligned} v_4^{(k)} = {} & \tfrac{1}{16}\Psi_0^{(k,1)}\xi\left(K(4\xi^2 + b_1^2)^{\frac{3}{2}} + 12 b_1 b_3 (4\xi^2 + b_1^2)^{\frac{1}{2}} + \frac{8b_2^2(8\xi^2 + 3b_1^2)}{(4\xi^2 + b_1^2)^{\frac{1}{2}}}\right) \\ & + \tfrac{1}{2}C_3^{(k,1)} b_1 b_2 (4\xi^2 + b_1^2)^{\frac{1}{2}} - \tfrac{1}{2}X_1^2(\Delta_{\partial\omega} + \lambda)\Psi_k^{(0)} + \tfrac{1}{2}X_2 b_1 \nabla b_1 \cdot \nabla\Psi_k^{(0)} + C_4^{(k,1)} X_1 + C_4^{(k,0)}, \end{aligned}$$

where $X_1 = X_1(\xi, b_1(P))$,

$$X_2 = X_2(\xi, b) := \xi^2 - b^2 X_3\left(\frac{2\xi + \sqrt{4\xi^2 + b^2}}{b}\right), \quad X_3(z) := \tfrac{1}{8}\ln^2 z + \frac{1}{16}\left(z^2 - \frac{1}{z^2}\right)\ln z - \frac{1}{32}\left(z^2 + \frac{1}{z^2}\right),$$

and $C_4^{(k,0)} = C_4^{(k,0)}(P, \varepsilon)$ and $C_4^{(k,1)} = C_4^{(k,1)}(P, \varepsilon)$ are unknown functions to be determined.

To determine the coefficient $\boldsymbol{\phi}^{(k)}$ in the outer expansion and the functions $C_i^{k,j}$ in the inner one, we should match the constructed functions $v_i^{(k)}$ with the outer expansion. In order to do it, we must find the asymptotics for the functions $v_i^{(k)}$ as $\xi \to \pm\infty$. We observe that the functions $X_1, X_2 \in C^\infty(\mathbb{R} \times (0, +\infty))$ satisfy the identities

$$\begin{aligned} X_1(\xi, b) &= \pm\xi^2 \pm \frac{b^2}{8}(2\ln|\xi| + 1 + 4\ln 2 - 2\ln b) + \mathcal{O}(\xi^{-2}), && \xi \to \pm\infty, \\ X_2(\xi, b) &= \xi^2\bigl(\tfrac{3}{2} - 2\ln 2 + \ln b - \ln|\xi|\bigr) + \mathcal{O}(\ln^2|\xi|), && \xi \to \pm\infty, \end{aligned}$$

uniformly in $b \geqslant b_0 > 0$, with b_0 any fixed constant. Taking these asymptotics into account, we write the asymptotics for $v_i^{(k)}$ as $\xi \to \pm\infty$ and then pass to the variables (τ, P):

$$\begin{aligned}\sum_{i=0}^{4}\varepsilon^i v_i^{(k)}(\xi,P,\varepsilon)={}&\Psi_k^{(0)}(P,\varepsilon)\pm\Psi_k^{(1)}(P,\varepsilon)\tau+\tfrac{1}{2}\big(\pm\Psi_k^{(1)}(P,\varepsilon)K(P)-\Delta_{\partial\omega}\Psi_k^{(0)}(P,\varepsilon)-\lambda\Psi_k^{(0)}(P,\varepsilon)\big)\tau^2\\&+\varepsilon(\pm C_3^{(k,1)}(P,\varepsilon)\tau+C_1^{(k,0)})+\varepsilon^2\big(\ln\varepsilon\, W_{2,1,\pm}^{(k)}(x',\varepsilon)+W_{2,0,\pm}^{(k)}(x',\varepsilon)\big)+\mathcal{O}(\varepsilon^3+\varepsilon^4\tau^{-1}),\end{aligned}$$

where

$$W_{2,1,\pm}^{(k)} := \frac{1}{4}b_1^2\left(\mp\Psi_k^{(1)}+\tau\left(\Delta_{\partial\omega}+\frac{2}{b_1}\nabla b_1\cdot\nabla+\lambda\right)\Psi_k^{(0)}\right), \tag{5-20}$$

$$\begin{aligned}W_{2,0,\pm}^{(k)} :={}& \pm\tfrac{1}{8}b_1^2\Psi_k^{(1)}\ln\tau\pm\frac{b_1^2}{8}(1+4\ln 2-2\ln b_1)\Psi_k^{(1)}+C_2^{(k,0)}+\Psi_k^{(1)}b_1b_2\tau^{1/2}\\&-\tfrac{1}{8}b_1^2\tau\ln\tau\left(\Delta_{\partial\omega}+\frac{2}{b_1}\nabla b_1\cdot\nabla+\lambda\right)\Psi_k^{(0)}+\tau\big(-\tfrac{1}{8}b_1^2(1+4\ln 2-2\ln b_1)(\Delta_{\partial\omega}+\lambda)\Psi_k^{(0)}\\&-\tfrac{1}{2}\big(2\ln 2-\ln b_1-\tfrac{3}{2}\big)b_1\nabla b_1\cdot\nabla\Psi_k^{(0)}\pm\tfrac{1}{16}(3Kb_1^2+32b_2^2+24b_1b_3)\Psi_k^{(1)}\pm C_4^{(k,1)}\big).\end{aligned} \tag{5-21}$$

Taking into account the obtained formulas and (5-2), in accordance with the method of matching asymptotic expansions we conclude that

$$C_3^{(k,1)}(P,\varepsilon) = C_1^{(k,0)}(P,\varepsilon) \equiv 0, \tag{5-22}$$

while the solutions to (5-4) should satisfy the asymptotics

$$\phi_\pm^{(k)}(x',\varepsilon) = W_{2,1,\pm}^{(k)}(x',\varepsilon) + \frac{1}{\ln\varepsilon}W_{2,0,\pm}^{(k)}(x',\varepsilon) + o(\tau), \quad \tau \to 0. \tag{5-23}$$

Moreover, the identity

$$\tfrac{1}{2}\big(\pm\Psi_k^{(1)}K - \Delta_{\partial\omega}\Psi_k^{(0)} - \lambda\Psi_k^{(0)}\big) = \Psi_k^{(2,\pm)} \tag{5-24}$$

should hold.

Outer expansion: Second term. We substitute (3-29) and (5-5) into the eigenvalue equation for $\psi_\pm^{(k)}$ and equate the coefficient of τ^0. This leads us to identity (5-24).

We proceed to the problem (5-4), (5-23). To study its solvability we shall make use of one more auxiliary lemma. Recall that the matrices M and $\widehat{\mathrm{M}}$ are defined in (3-4) and (3-30), respectively.

Lemma 5.1. *The functions* $f_{2,\pm}^{(k)}$ *introduced in* (5-4) *satisfy the hypothesis of Lemma 3.4. In particular, the asymptotics* (3-33) *holds true with*

$$f_{-2}^{\pm}=\pm\frac{b_1^2}{8\ln\varepsilon}\Psi_k^{(1)}, \quad f_{-3/2}^{\pm}=\frac{b_1b_2}{4\ln\varepsilon}\Psi_k^{(1)}, \quad f_{-1}^{\pm}=-\frac{b_1^2}{4\ln\varepsilon}\left(\Psi_k^{(2,\pm)}-\frac{1}{b_1}\nabla b_1\cdot\nabla\Psi_k^{(0)}\mp K\Psi_k^{(1)}\right). \tag{5-25}$$

Proof. We begin with an obvious identity:

$$f_{2,\pm}^{(k)} = \frac{1}{\ln\varepsilon}\big(-\operatorname{div}_{x'}\mathrm{Q}_\pm\nabla_{x'}\psi_\pm^{(k)} + \tfrac{1}{2}\big(\nabla_{x'}|\nabla_{x'}h_\pm|^2, \nabla_{x'}\psi_\pm^{(k)}\big)_{\mathbb{R}^n}\big), \tag{5-26}$$

which follows from the definition of $f^{(k)}_{2,\pm}$ in (5-4). To prove the lemma, we shall pass to the variables (τ, s) in the obtained identity. It follows from (3-7), (3-12) and the definition of S_ε that

$$h_\pm(x') = t, \quad \pm t > 0.$$

Hence, by (3-8), (3-10),

$$h_\pm(x') = b(\pm\sqrt{\tau}, P) = \sum_{i=1}^{\infty} b_i(P)(\pm\sqrt{\tau})^i, \quad \tau \to +0. \tag{5-27}$$

Thus, employing (3-4) and (5-26), we conclude that the functions $f^{(k)}_{2,0,\pm}$ satisfy the hypothesis of Lemma 3.4 and in particular the asymptotics (3-33) holds true. It remains to prove the identities (5-25).

It follows from (3-44) that

$$|\nabla_{x'} h_\pm|^2 = \left|\frac{\partial h_\pm}{\partial \tau}\right|^2 + \nabla h_\pm \cdot (\mathrm{E} - \tau \mathrm{B}\mathrm{G}^{-1}_{\partial\omega})^{-2} \nabla h_\pm. \tag{5-28}$$

We substitute (5-27) into the obtained identity and arrive at the asymptotics for $|\nabla_{x'} h_\pm|^2$:

$$|\nabla_{x'} h_\pm|^2 = \sum_{j=-2}^{\infty} h^\pm_{j/2}(P)\tau^{j/2}, \quad h^\pm_{-1} = \tfrac{1}{4} b_1^2, \quad h^\pm_{-1/2} = \pm b_1 b_2, \quad \tau \to +0. \tag{5-29}$$

Employing these formulas and (3-4), (3-30), (5-5) and (3-44) we rewrite the second term in the right-hand side of (5-26) as

$$\begin{aligned}\tfrac{1}{2}\big(\nabla_{x'}|\nabla_{x'} h_\pm|^2, \nabla_{x'}\psi^{(k)}_\pm\big)_{\mathbb{R}^n} &= \frac{1}{2}\frac{\partial |\nabla_{x'} h_\pm|^2}{\partial\tau}\frac{\partial \psi^{(k)}_\pm}{\partial \tau} + \tfrac{1}{2}\nabla|\nabla_{x'} h_\pm|^2 \cdot (\mathrm{E} - \tau\mathrm{B}\mathrm{G}^{-1}_{\partial\omega})^{-2}\nabla\psi^{(k)}_\pm \\ &= \sum_{j=-4}^{\infty} f^{\pm,2}_{j/2}\tau^{j/2},\end{aligned} \tag{5-30}$$

where $f^{\pm,2}_{j/2} \in C^\infty(\partial\omega)$ are some functions, and, in particular,

$$f^{\pm,2}_{-2} = \mp\frac{1}{8\ln\varepsilon} b_1^2 \Psi^{(1)}_k, \quad f^{\pm,2}_{-3/2} = -\frac{1}{4\ln\varepsilon} b_1 b_2 \Psi^{(1)}_k, \quad f^{\pm,2}_{-1} = -\frac{b_1^2}{4\ln\varepsilon}\left(\Psi^{(2,\pm)}_k + \frac{1}{b_1}\nabla b_1 \cdot \nabla\Psi^{(0)}_k\right). \tag{5-31}$$

To obtain the same asymptotics for the first term in the right-hand side of (5-26), we employ first (3-43):

$$-\operatorname{div}_{x'} \mathrm{Q}_\pm \nabla_{x'}\psi^{(k)}_\pm = -\frac{1}{\det \mathrm{M}} \operatorname{div}_{(\tau,s)} (\det \mathrm{M}) \nabla_{(\tau,s)} h_\pm (\nabla_{(\tau,s)} h_\pm)^* \widehat{\mathrm{M}} \nabla_{(\tau,s)} \psi^{(k)}_\pm. \tag{5-32}$$

It follows from the equations (3-29), (3-30), (5-27) that

$$(\nabla_{(\tau,s)} h_\pm)^* \widehat{\mathrm{M}} \nabla_{(\tau,s)} \psi^{(k)}_\pm = \frac{\partial h_\pm}{\partial\tau}\frac{\partial\psi^{(k)}_\pm}{\partial\tau} + \nabla h_\pm \cdot (\mathrm{E} - \tau\mathrm{B}\mathrm{G}^{-1}_{\partial\omega})^{-2}\nabla\psi^{(k)}_\pm = \sum_{j=-1}^{\infty} c^\pm_{j/2}\tau^{j/2}, \quad \tau \to +0,$$

$$(\det \mathrm{M})\widehat{\mathrm{M}}\nabla_{(\tau,s)} h_\pm = \sum_{j=-1}^{\infty} \mathbf{c}^\pm_{j/2}\tau^{j/2}, \quad \tau \to +0,$$

where $c^{\pm}_{j/2} = c^{\pm}_{j/2}(P) \in C^{\infty}(\partial\omega)$ are some functions, $\mathbf{c}^{\pm}_{j/2} = \mathbf{c}^{\pm}_{j/2}(P) \in C^{\infty}(\partial\omega)$ are some n-dimensional vector-functions, and

$$c^{\pm}_{-1/2} = \tfrac{1}{2} b_1, \quad c^{\pm}_{0} = \pm b_2 \Psi^{(1)}_k, \quad \mathbf{c}^{\pm}_{-1/2} = \pm \tfrac{1}{2} b_1 \mathbf{e}_1, \quad \mathbf{c}^{\pm}_{0} = b_2 \mathbf{e}_1,$$

and $\mathbf{e}_1 = (1, 0, \dots, 0)^*$. We substitute the last identities into (5-32), which yields

$$-\operatorname{div}_{x'} \mathrm{Q}_{\pm} \nabla_{x'} \psi^{(k)}_{\pm} = \sum_{j=-4}^{\infty} f^{\pm,1}_{j/2} \tau^{j/2}, \quad \tau \to +0,$$

$$f^{\pm,1}_{-2} = \pm \frac{1}{4 \ln \varepsilon} b_1^2 \Psi^{(1)}_k, \quad f^{\pm,1}_{-3/2} = \frac{1}{2 \ln \varepsilon} b_1 b_2 \Psi^{(1)}_k, \quad f^{\pm,1}_{-1} = \pm \frac{1}{4 \ln \varepsilon} b_1^2 K \Psi^{(1)}_k.$$

The last identity, (5-30), (5-31), (5-26) imply the formulas (5-25). □

Taking into account (5-5), we apply Lemma 5.1 to problem (5-4). It implies that the right-hand side of (5-4) satisfies the hypothesis of Lemma 3.4 with the first four coefficients given by (5-25).

Given some functions $V^{(0)}_k$, $V^{(1)}_k \in C^{\infty}(\partial\omega)$, suppose the solvability condition (3-34) holds true. Then by (3-36), (5-24), (5-25) there exists the unique solution to (5-4) with the asymptotics

$$\begin{aligned}
\phi^{(k)}_{\pm} = {} & \frac{1}{\ln\varepsilon}\big(\pm\tfrac{1}{8} b_1^2 \Psi^{(1)}_k \ln\tau + b_1 b_2 \Psi^{(1)}_k \tau^{1/2} + \tau(1-\ln\tau)\big(-\tfrac{1}{4} b_1^2 \Psi^{(2,\pm)}_k + \tfrac{1}{4} b_1 \nabla b_1 \cdot \nabla \Psi^{(0)}_k \pm \tfrac{1}{8} K b_1^2 \Psi^{(1)}_k\big)\big) \\
& + U^{(0)}_k \pm V^{(0)}_k + \tau(V^{(1)}_k \pm U^{(1)}_k) + \mathcal{O}(\tau^{3/2}) \\
= {} & \frac{1}{\ln\varepsilon}\left(\pm\tfrac{1}{8} b_1^2 \Psi^{(1)}_k \ln\tau + b_1 b_2 \Psi^{(1)}_k \tau^{1/2} + \tau(1-\ln\tau)\left(\Delta_{\partial\omega} + \frac{2}{b_1}\nabla b_1 \cdot \nabla + \lambda\right)\Psi^{(0)}_k\right) \\
& + U^{(0)}_k \pm V^{(0)}_k + \tau(V^{(1)}_k \pm U^{(1)}_k), \quad \tau \to +0,
\end{aligned} \tag{5-33}$$

where $U^{(0)}_k$, $U^{(1)}_k \in C^{\infty}(\partial\omega)$ are some functions satisfying (3-37). We compare the last asymptotics with (5-20), (5-21), (5-23), take into consideration the identity (5-24) and arrive at the formulas for $V^{(0)}_k$, $V^{(1)}_k$, $C^{(k,0}_2$ and $C^{(k,1)}_4$:

$$\begin{aligned}
V^{(0)}_k &= -\frac{b_1^2}{4}\Psi^{(1)}_k + \frac{b_1^2}{8 \ln \varepsilon}(1 + 4\ln 2 - 2\ln b_1)\Psi^{(1)}_k, \quad C^{(k,0)}_2 = \ln\varepsilon\, U^{(0)}_k, \\
V^{(1)}_k &= \frac{b_1^2}{4}\left(\Delta_{\partial\omega} + \frac{2}{b_1}\nabla b_1 \cdot \nabla + \lambda\right)\Psi^{(0)}_k \\
&\qquad - \frac{b_1^2}{4 \ln \varepsilon}\left((2\ln 2 - \ln b_1 + 1)(\Delta_{\partial\omega} + \lambda)\Psi^{(0)}_k + \frac{4\ln 2 - 2\ln b_1 - 2}{b_1}\nabla b_1 \cdot \nabla \Psi^{(0)}_k\right), \\
C^{(k,1)}_4 &= \ln\varepsilon\, U^{(1)}_k - \tfrac{1}{16}(3Kb_1^2 + 32 b_2^2 + 24 b_1 b_3)\Psi^{(1)}_k.
\end{aligned}$$

In what follows the functions $V^{(0)}_k$, $V^{(1)}_k$, $C^{(k,0}_2$ and $C^{(k,1)}_4$ are supposed to be chosen in accordance with the above given formulas. Bearing these formulas, (5-24) and (5-25) in mind, we write the solvability conditions (3-34) for (5-4):

$$\begin{aligned}\frac{1}{\ln\varepsilon}\lim_{\delta\to+0}\Big[&(f^{(k)}_{2,+},\psi^{(i)}_{+})_{L_2(\omega^\delta)}+(f^{(k)}_{2,-},\psi^{(i)}_{-})_{L_2(\omega^\delta)}-\delta^{-1/2}\int_{\partial\omega}b_1b_2\Psi^{(1)}_k\Psi^{(0)}_i\,ds\\&+\ln\delta\int_{\partial\omega}\frac{b_1^2}{4}\Big(\Psi^{(1)}_i\Psi^{(1)}_k+\Psi^{(0)}_i\Big(\Delta_{\partial\omega}+\frac{2}{b_1}\nabla b_1\cdot\nabla+\lambda\Big)\Psi^{(0)}_k\Big)\,ds\Big]\\&+\int_{\partial\omega}\frac{b_1^2}{2\ln\varepsilon}(2\ln 2-\ln b_1+1)\Psi^{(0)}_i(\Delta_{\partial\omega}+\lambda)\Psi^{(0)}_k\,ds\\&+\int_{\partial\omega}\frac{b_1}{\ln\varepsilon}(2\ln 2-\ln b_1-1)\Psi^{(0)}_i\nabla b_1\cdot\nabla\Psi^{(0)}_k\,ds+\int_{\partial\omega}\frac{b_1^2}{2\ln\varepsilon}(2\ln 2-\ln b_1)\Psi^{(1)}_k\Psi^{(1)}_i\,ds\\&-\int_{\partial\omega}\frac{b_1^2}{2}\Big(\Psi^{(1)}_k\Psi^{(1)}_i+\Psi^{(0)}_i\Big(\Delta_{\partial\omega}+\frac{2}{b_1}\nabla b_1\cdot\nabla+\lambda\Big)\Psi^{(0)}_k\Big)\,ds+\mu_k\delta_{ik}=0,\quad i,k=1,\dots,m.\end{aligned}\tag{5-34}$$

Let us simplify the obtained identity. We first rewrite the formulas (5-4) of $f^{(k)}_{2,\pm}$ in a more convenient form employing the eigenvalue equation for $\psi^{(k)}_{\pm}$ and the definition of the matrix $\mathrm{Q}_{\pm}$:

$$f^{(k)}_{2,\pm}=-\operatorname{div}_{x'}\Phi^{(k)}_{\pm}\nabla_{x'}h_{\pm}+\frac{\lambda}{2}|\nabla_{x'}h_{\pm}|^2\psi^{(k)}_{\pm}+\tfrac{1}{2}\operatorname{div}_{x'}|\nabla_{x'}h_{\pm}|^2\nabla_{x'}\psi^{(k)}_{\pm},\quad \Phi^{(k)}_{\pm}:=(\nabla_{x'}h_{\pm},\nabla_{x'}\psi^{(k)}_{\pm})_{\mathbb{R}^n}.$$

Employing this representation, we integrate by parts to obtain

$$\begin{aligned}(f^{(k)}_{2,\pm},\psi^{(i)}_{\pm})_{L_2(\omega^\delta)}=&\int_{\partial\omega^\delta}\Big(\Phi^{(k)}_{\pm}\frac{\partial h_{\pm}}{\partial\tau}-\frac{1}{2}|\nabla_{x'}h_{\pm}|^2\frac{\partial\psi^{(i)}_{\pm}}{\partial\tau}\Big)\psi^{(i)}_{\pm}\,ds+\int_{\omega^\delta}\Phi^{(i)}_{\pm}\Phi^{(k)}_{\pm}\,dx'\\&+\frac{\lambda}{2}\int_{\omega^\delta}|\nabla_{x'}h_{\pm}|^2\psi^{(i)}_{\pm}\psi^{(k)}_{\pm}\,dx'-\frac{1}{2}\int_{\omega^d}|\nabla_{x'}h_{\pm}|^2(\nabla_{x'}\psi^{(i)}_{\pm},\nabla_{x'}\psi^{(k)}_{\pm})_{\mathbb{R}^d}\,dx'.\end{aligned}\tag{5-35}$$

Applying (3-44), we have

$$\Phi^{(k)}_{\pm}=\frac{\partial h_{\pm}}{\partial\tau}\frac{\partial\psi^{(k)}_{\pm}}{\partial\tau}+\nabla h_{\pm}\cdot(\mathrm{E}-\tau\mathrm{B}\mathrm{G}^{-1}_{\partial\omega})^{-2}\nabla\psi^{(k)}_{\pm}$$

in a vicinity of $\partial\omega$. Hence, by (5-5), (5-27) and (5-28),

$$\Phi^{(k)}_{\pm}=\frac{b_1}{2\sqrt{\tau}}\Psi^{(1)}_k+\mathcal{O}(1),\quad \tau\to+0,\tag{5-36}$$

$$\begin{aligned}\Big(\Phi^{(k)}_{\pm}\frac{\partial h_{\pm}}{\partial\tau}&-\frac{1}{2}|\nabla_{x'}h_{\pm}|^2\frac{\partial\psi^{(i)}_{\pm}}{\partial\tau}\Big)\psi^{(i)}_{\pm}\prod_{j=1}^{n-1}(1-\tau K_j)\,ds\\&=\pm\frac{1}{8\tau}b_1^2\Psi^{(1)}_i\Psi^{(1)}_k+\frac{1}{2\sqrt{\tau}}b_1b_2\Psi^{(0)}_i\Psi^{(1)}_k+\tfrac{1}{8}b_1^2\Psi^{(1)}_i\Psi^{(1)}_k\mp\tfrac{1}{8}b_1^2K\Psi^{(0)}_i\Psi^{(1)}_k\\&\quad+\tfrac{1}{4}(b_1^2\Psi^{(2,\pm)}_k\pm 3b_1b_3\Psi^{(1)}_k\pm 2b_2^2\Psi^{(1)}_k+2b_1\nabla b_1\cdot\Psi^{(0)}_k)\Psi^{(0)}_i+\mathcal{O}(\sqrt{\tau}),\quad \tau\to+0.\end{aligned}$$

Substituting the last identity into (5-35) and using (3-42) and (5-24), we get

$$\begin{aligned}
&(f^{(k)}_{2,+}, \psi^{(i)}_+)_{L_2(\omega^\delta)} + (f^{(k)}_{2,-}, \psi^{(i)}_-)_{L_2(\omega^\delta)} \\
&\quad = \int_{\omega^\delta} \frac{|\nabla_{x'} h_+|^2}{2} \big(\lambda \psi^{(i)}_+ \psi^{(k)}_+ - (\nabla_{x'} \psi^{(i)}_+, \nabla_{x'} \psi^{(k)}_+)_{\mathbb{R}^d}\big)\, dx' \\
&\qquad + \int_{\omega^\delta} \frac{|\nabla_{x'} h_-|^2}{2} \big(\lambda \psi^{(i)}_- \psi^{(k)}_- - (\nabla_{x'} \psi^{(i)}_-, \nabla_{x'} \psi^{(k)}_-)_{\mathbb{R}^d}\big)\, dx' \\
&\qquad + \int_{\omega^\delta} (\Phi^{(i)}_+ \Phi^{(k)}_+ + \Phi^{(i)}_- \Phi^{(k)}_-)\, dx' + \delta^{-1/2} \int_{\partial\omega} b_1 b_2 \Psi^{(0)}_i \Psi^{(0)}_k \, ds \\
&\qquad + \int_{\partial\omega} \frac{b_1^2}{4} \Psi^{(1)}_i \Psi^{(1)}_k \, ds - \int_{\partial\omega} \frac{b_1^2}{4} \Psi^{(0)}_i (\Delta_{\partial\omega} + \lambda) \Psi^{(0)}_k \, ds \\
&\qquad + \int_{\partial\omega} b_1 \Psi^{(0)}_i \nabla b_1 \cdot \nabla \Psi^{(0)}_k \, ds + \mathcal{O}(\delta^{1/2}), \quad \delta \to +0.
\end{aligned}$$

We integrate by parts once again, this time over $\partial\omega$, and we have

$$\int_{\partial\omega} b_1^2 \Psi^{(0)}_i \left(\Delta_{\partial\omega} + \frac{2}{b_1} \nabla b_1 \cdot \nabla + \lambda \right) \Psi^{(0)}_k \, ds = \int_{\partial\omega} b_1^2 \big(\lambda \Psi^{(0)}_i \Psi^{(0)}_k - \nabla \Psi^{(0)}_i \cdot \nabla \Psi^{(0)}_k \big)\, ds. \tag{5-37}$$

Substituting the last two identities into (5-34) yields

$$\begin{aligned}
&\frac{1}{\ln\varepsilon} \lim_{\delta\to+0} \bigg[\int_{\omega^\delta} \frac{|\nabla_{x'} h_+|^2}{2} \big(\lambda \psi^{(i)}_+ \psi^{(k)}_+ - (\nabla_{x'} \psi^{(i)}_+, \nabla_{x'} \psi^{(k)}_+)_{\mathbb{R}^d}\big)\, dx' \\
&\qquad + \int_{\omega^\delta} \frac{|\nabla_{x'} h_-|^2}{2} \big(\lambda \psi^{(i)}_- \psi^{(k)}_- - (\nabla_{x'} \psi^{(i)}_-, \nabla_{x'} \psi^{(k)}_-)_{\mathbb{R}^d}\big)\, dx' + \int_{\omega^\delta} (\Phi^{(i)}_+ \Phi^{(k)}_+ + \Phi^{(i)}_- \Phi^{(k)}_-)\, dx' \\
&\qquad + \ln\delta \int_{\partial\omega} \frac{b_1^2}{4} \big(\Psi^{(1)}_i \Psi^{(1)}_k + \lambda \Psi^{(0)}_i \Psi^{(0)}_k - \nabla \Psi^{(0)}_i \cdot \nabla \Psi^{(0)}_k \big)\, ds \bigg] \\
&\quad + \int_{\partial\omega} \frac{b_1^2}{4\ln\varepsilon} (1 + 4\ln 2 - 2\ln b_1) \big(\Psi^{(1)}_i \Psi^{(1)}_k + \Psi^{(0)}_i (\Delta_{\partial\omega} + \lambda) \Psi^{(0)}_k \big)\, ds \\
&\quad + \int_{\partial\omega} \frac{b_1}{\ln\varepsilon} (2\ln 2 - \ln b_1) \Psi^{(0)}_i \nabla b_1 \cdot \nabla \Psi^{(0)}_k \, ds \\
&\quad - \int_{\partial\omega} \frac{b_1^2}{2} \left(\Psi^{(1)}_k \Psi^{(1)}_i + \Psi^{(0)}_i \left(\Delta_{\partial\omega} + \frac{2}{b_1} \nabla b_1 \cdot \nabla + \lambda \right) \Psi^{(0)}_k \right) ds + \mu_k \delta_{ik} = 0,
\end{aligned} \tag{5-38}$$

as $i, k = 1, \dots, m$. It follows from (5-36), (5-29) and (5-5) that

$$\begin{aligned}
&|\nabla_{x'} h_+|^2 \big(\lambda \psi^{(i)}_+ \psi^{(k)}_+ - (\nabla_{x'} \psi^{(i)}_+, \nabla_{x'} \psi^{(k)}_+)_{\mathbb{R}^d}\big) + |\nabla_{x'} h_-|^2 \big(\lambda \psi^{(i)}_- \psi^{(k)}_- - (\nabla_{x'} \psi^{(i)}_-, \nabla_{x'} \psi^{(k)}_-)_{\mathbb{R}^d}\big) \\
&\qquad = \frac{b_1^2}{2\tau} (\lambda \Psi^{(0)}_i \Psi^{(0)}_k - \nabla \Psi^{(0)}_i \cdot \nabla \Psi^{(0)}_k) + \mathcal{O}(\tau^{-1/2}), \quad \tau \to +0,
\end{aligned}$$

$$\Phi^{(i)}_\pm \Phi^{(k)}_\pm = \frac{b_1^2}{4\tau} \Psi^{(1)}_i \Psi^{(1)}_k + \mathcal{O}(\tau^{-1/2}), \quad \tau \to +0.$$

Hence, the limit in (5-38) is finite. To calculate the boundary integrals in (5-38) we integrate by parts:

$$\begin{aligned}\int_{\partial\omega}\frac{b_1^2}{4}(1+4\ln 2-2\ln b_1)\big(\Psi_i^{(1)}\Psi_k^{(1)}+\Psi_i^{(0)}(\Delta_{\partial\omega}+\lambda)\Psi_k^{(0)}\big)\,ds+\int_{\partial\omega} b_1(2\ln 2-\ln b_1)\Psi_i^{(0)}\nabla b_1\cdot\nabla\Psi_k^{(0)}\,ds\\ =\int_{\partial\omega}\frac{b_1^2}{4}(1+4\ln 2-2\ln b_1)\big(\Psi_i^{(1)}\Psi_k^{(1)}+\lambda\Psi_i^{(0)}\Psi_k^{(0)}-\nabla\Psi_i^{(0)}\cdot\nabla\Psi_k^{(0)}\big)\,ds.\end{aligned}$$

Due to this identity, (5-37), the definition of b_1 in (3-10) and the definitions (2-9) and (2-10) of the matrices $\Lambda^{(0)}$ and $\Lambda^{(1)}$, respectively, we can rewrite (5-38) in the final form

$$\mu_k\delta_{ik}=\Lambda_{ik}^{(0)}+\frac{1}{\ln\varepsilon}\Lambda_{ik}^{(1)}.$$

Since the matrix on the right-hand side of the last identity is diagonal, we conclude that the solvability condition for the problem (5-4), (5-23) is satisfied provided μ_k are the eigenvalues of the matrix $\Lambda^{(0)}+\frac{1}{\ln\varepsilon}\Lambda^{(1)}$. It follows from [Kato 1966, Chapter II, Section 6.1, Theorem 6.1] that the eigenvalues of this matrix are holomorphic in $\frac{1}{\ln\varepsilon}$ and converge to those of $\Lambda^{(0)}$ as $\varepsilon\to 0$.

In view of the choice of μ_i the problems (5-4), (5-33) are solvable. We observe that each of the functions $\phi_\pm^{(k)}$ is defined up to a linear combination of the eigenfunctions $\psi_\pm^{(i)}$. The exact values of the coefficients of these linear combinations can be determined while constructing the next terms in the asymptotic expansions for $\lambda_k(\varepsilon)$ and $\psi_\varepsilon^{(k)}$. The formal constructing of the asymptotic expansions is complete.

Justification of the asymptotics. In order to justify the obtained asymptotics, one has to construct additional terms. This is a general and standard situation for singularly perturbed problems. In our case one should construct the terms of order up to $\mathcal{O}(\varepsilon^4)$ in the outer expansion for the eigenfunctions and for the eigenvalues, and the terms of order up to $\mathcal{O}(\varepsilon^6)$ in the inner expansion for the eigenfunctions. The asymptotics with the additional terms read

$$\begin{aligned}&\lambda_k(\varepsilon)=\lambda+\varepsilon^2\ln\varepsilon\,\mu_k\left(\frac{1}{\ln\varepsilon}\right)+\varepsilon^4\ln^2\varepsilon\,\eta_k(\varepsilon)+\cdots,\\ &\psi_{\varepsilon,\mathrm{ex}}^{(k)}=\mathcal{I}_\varepsilon(\boldsymbol{\psi}_k+\varepsilon^2\ln\varepsilon\,\boldsymbol{\phi}_k+\varepsilon^4\ln^2\varepsilon\,\boldsymbol{\theta}_k+\cdots),\quad \psi_{\varepsilon,\mathrm{in}}^{(k)}=v_0^{(k)}+\sum_{i=2}^{6}\varepsilon^i v_i^{(k)}+\cdots,\end{aligned}\tag{5-39}$$

where $\boldsymbol{\theta}_k=(\theta_+^{(k)},\theta_-^{(k)})$, $\theta_\pm^{(k)}=\theta_\pm^{(k)}(x',\varepsilon)$, $v_i^{(k)}=v_i^{(k)}(\xi,P,\varepsilon)$, and we used that $v_1^{(k)}=0$ by (5-16), (5-22). The equations for $\theta_\pm^{(k)}$ are

$$\begin{aligned}(-\Delta_{x'}-\lambda)\theta_\pm^{(k)}&=\frac{1}{\ln\varepsilon}\mathcal{H}_\pm^{(2)}\phi_\pm^{(k)}+\frac{1}{\ln^2\varepsilon}\mathcal{H}_\pm^{(4)}\psi_\pm^{(k)}+\mu_k\phi_\pm^{(k)}+\eta_k\psi_\pm^{(k)},\quad x'\in\omega_\pm,\\ \mathcal{H}_\pm^{(4)}&:=\tfrac{3}{8}|\nabla_{x'}h_\pm|^4\Delta_{x'}-\tfrac{1}{2}|\nabla_{x'}h_\pm|^2\operatorname{div}_{x'}\big(\tfrac{1}{2}|\nabla_{x'}h_\pm|^2\mathrm{E}-\mathrm{Q}_\pm\big)\nabla_{x'}\\ &\qquad-\operatorname{div}_{x'}\big(\tfrac{1}{8}|\nabla_{x'}h_\pm|^4\mathrm{E}+\tfrac{1}{2}\mathrm{Q}_\pm|\nabla_{x'}h_\pm|^2+\mathrm{Q}_\pm^2\big)\nabla_{x'}.\end{aligned}$$

The functions $\theta_{\pm}^{(k)}$ should satisfy the asymptotics

$$\begin{aligned}
\theta_{\pm}^{(k)}(x',\varepsilon) &= W_{4,2,\pm}^{(k)}(x',\varepsilon) + \frac{1}{\ln\varepsilon} W_{4,1,\pm}^{(k)}(x',\varepsilon) + \frac{1}{\ln^2\varepsilon} W_{4,0,\pm}^{(k)}(x',\varepsilon) + o(1), \quad \tau \to +0,\\
W_{4,2,\pm}^{(k)} &= -\tfrac{1}{32} b_1^3 \big(b_1(\Delta_{\partial\omega}+\lambda)\Psi_k^{(0)} + 2\nabla b_1 \cdot \nabla\Psi_k^{(0)}\big),\\
W_{4,1,\pm}^{(k)} &= \tfrac{1}{32} b_1^3 (\ln\tau + 1 + 4\ln 2 - 2\ln b_1)\big(b_1(\Delta_{\partial\omega}+\lambda)\Psi_k^{(0)} + 2\nabla b_1 \cdot \nabla\Psi_k^{(0)}\big),\\
W_{4,0,\pm}^{(k)} &= \pm\frac{1}{128}\frac{\Psi_k^{(1)} b_1^4}{\tau} + \frac{1}{8}\frac{\Psi_k^{(1)} b_1^3 b_2}{\sqrt{\tau}}\\
&\quad - \tfrac{1}{128} b_1^3 \big(b_1(\Delta_{\partial\omega}+\lambda)\Psi_k^{(0)} + 2\nabla b_1 \cdot \nabla\Psi_k^{(0)}\big)(\ln\tau + 4\ln 2 - 2\ln b_1 + 1)^2\\
&\quad - \tfrac{1}{128} b_1^3 \big(b_1(\Delta_{\partial\omega}+\lambda)\Psi_k^{(0)} - 2\nabla b_1 \cdot \nabla\Psi_k^{(0)}\big) \pm \tfrac{1}{256}\Psi_k^{(1)}(3Kb_1^4 + 48 b_1^3 b_3 + 128 b_1^2 b_2^2).
\end{aligned}$$

The equations for the functions $v_5^{(k)}$, $v_6^{(k)}$ are obtained in the same way as those for $v_i^{(k)}$, $i = 0, \dots, 4$, from

$$\mathscr{L}_{-4} v_5^{(k)} + \sum_{i=-3}^{-1} \mathscr{L}_i v_{1-i}^{(k)} \mathscr{L}_1 v_0^{(k)} = 0 \qquad \text{on } \mathbb{R}\times\partial\omega,$$

$$\mathscr{L}_{-4} v_6^{(k)} + \sum_{i=-3}^{0} \mathscr{L}_i v_{2-i}^{(k)} + \mathscr{L}_2 v_0^{(k)} = \lambda v_2^{(k)} + \ln\varepsilon\,\eta_k v_0^{(k)} \quad \text{on } \mathbb{R}\times\partial\omega,$$

where the operators $\mathscr{L}_1$, $\mathscr{L}_2$ are the next terms in the expansion (3-25). It can be shown that the problem for $\theta_{\pm}^{(k)}$ is solvable for some $\eta_k(\varepsilon)$. The equations for $v_5^{(k)}$ and $v_6^{(k)}$ can be solved explicitly. The arbitrary coefficients $C_{5,1}^{(k)}$, $C_{5,0}^{(k)}$, $C_{6,1}^{(k)}$, $C_{6,0}^{(k)}$ appearing in $v_5^{(k)}$, $v_6^{(k)}$ can be determined while matching the inner and outer expansions.

We now introduce the partial sums

$$\widehat{\lambda}_{\varepsilon}^{(k)} = \lambda + \varepsilon^2 \ln\varepsilon\,\mu_k\left(\frac{1}{\ln\varepsilon}\right) + \varepsilon^4 \ln^2\varepsilon\,\eta_k(\varepsilon),$$

$$\widehat{\psi}_{\varepsilon,\mathrm{ex}}^{(k)} = \mathscr{I}_{\varepsilon}(\boldsymbol{\psi}_k + \varepsilon^2\ln\varepsilon\,\boldsymbol{\phi}_k + \varepsilon^4\ln^2\varepsilon\,\boldsymbol{\theta}_k), \quad \widehat{\psi}_{\varepsilon,\mathrm{in}}^{(k)} = v_0^{(k)} + \sum_{i=2}^{6}\varepsilon^i v_i^{(k)}$$

and define the final approximation for the eigenfunctions as

$$\widehat{\psi}_{\varepsilon}^{(k)}(x) = \widehat{\psi}_{\varepsilon,\mathrm{ex}}^{(k)}(x)\chi\left(\frac{\tau}{\varepsilon^{\alpha}}\right) + \widehat{\psi}_{\varepsilon,\mathrm{in}}^{(k)}(\xi, P)\left(1 - \chi\left(\frac{\tau}{\varepsilon^{\alpha}}\right)\right),$$

where $\alpha \in (0,1)$ is a fixed constant, and χ is the cut-off function introduced in the proof of Lemma 4.4.

Lemma 5.2. *The function* $\widehat{\psi}_{\varepsilon}^{(k)} \in C^{\infty}(S_{\varepsilon})$ *satisfies the convergence*

$$\|\widehat{\psi}_{\varepsilon}^{(k)} - \mathscr{I}_{\varepsilon}\boldsymbol{\psi}_k\|_{L_2(S_{\varepsilon})} \to 0, \quad \varepsilon \to +0, \tag{5-40}$$

and the equation

$$(\mathscr{H}_{\varepsilon} - \widehat{\lambda}_{\varepsilon}^{(k)})\widehat{\psi}_{\varepsilon}^{(k)} = F_{\varepsilon}^{(k)}, \tag{5-41}$$

where for the right-hand side the uniform in ε *estimate*

$$\|F_\varepsilon^{(k)}\|_{L_2(S_\varepsilon)} \leqslant C\varepsilon^{5\alpha/2} \tag{5-42}$$

holds true. The relations

$$(\mathcal{I}_\varepsilon \boldsymbol{\psi}_i, \mathcal{I}_\varepsilon \boldsymbol{\psi}_j)_{L_2(S_\varepsilon)} \to \delta_{ij}, \quad \varepsilon \to +0, \tag{5-43}$$

are valid.

The proof of this lemma is not very difficult and is based on lengthy and rather technical, but straightforward, calculations. Because of this, and in order not to overload the text with long technical formulas, we shall skip these here.

It follows from Lemma 4.6 and (5-41) that

$$\widehat{\psi}_\varepsilon^{(k)} = \sum_{i=1}^m \frac{\psi_\varepsilon^{(i)}}{\lambda_i(\varepsilon)-\lambda_k(\varepsilon)} (F_\varepsilon^{(k)}, \psi_\varepsilon^{(i)})_{L_2(S_\varepsilon)} + \mathcal{R}_\varepsilon(\lambda_k(\varepsilon))F_\varepsilon^{(k)}, \tag{5-44}$$

and, by (5-42),

$$\|\mathcal{R}_\varepsilon(\lambda_k(\varepsilon))F_\varepsilon^{(k)}\|_{W_2^1(S_\varepsilon)} \leqslant C\varepsilon^{5\alpha/2}, \quad k=1,\dots,m, \tag{5-45}$$

where the constant C is independent of ε. We calculate the scalar products of the functions $\widehat{\psi}_\varepsilon^{(k)}$ in $L_2(S_\varepsilon)$ taking into consideration (5-44) and the properties of the operator $\mathcal{R}_\varepsilon$ described in Lemma 4.6:

$$(\widehat{\psi}_\varepsilon^{(k)}, \widehat{\psi}_\varepsilon^{(p)})_{L_2(S_\varepsilon)} = \sum_{i=1}^m \gamma_i^{(k)}(\varepsilon)\gamma_i^{(p)}(\varepsilon) + \big(\mathcal{R}_\varepsilon(\lambda_k(\varepsilon))F_\varepsilon^{(k)}, \mathcal{R}_\varepsilon(\lambda_\varepsilon^{(p)})F_\varepsilon^{(p)}\big)_{L_2(S_\varepsilon)},$$

$$\gamma_\varepsilon^{(k)}(\varepsilon) := \frac{1}{\lambda_i(\varepsilon)-\widehat{\lambda}_\varepsilon^{(k)}} (F_\varepsilon^{(k)}, \psi_\varepsilon^{(i)})_{L_2(S_\varepsilon)}.$$

The identities obtained and (5-45), (5-40), (5-43) yield

$$\sum_{i=1}^m \gamma_i^{(k)}(\varepsilon)\gamma_i^{(p)}(\varepsilon) \to \delta_{kp}, \quad \varepsilon \to +0. \tag{5-46}$$

In particular, as $p=k$ it implies

$$|\gamma_i^{(k)}(\varepsilon)| \leqslant \tfrac{3}{2} \tag{5-47}$$

for sufficiently small ε. We introduce the matrix $\mathrm{R}_\varepsilon := (\gamma_i^{(k)}(\varepsilon))$ and rewrite (5-46) as $\mathrm{R}_\varepsilon \mathrm{R}_\varepsilon^* \to \mathrm{E}$, $\varepsilon \to +0$, where * denotes matrix transposition. Thus, $|\det \mathrm{R}_\varepsilon| \to 1$ as $\varepsilon \to +0$. Therefore, for each sufficiently small ε there exists a permutation $(i_1(\varepsilon), i_2(\varepsilon), \dots, i_m(\varepsilon))$ such that

$$\left|\prod_{i=1}^m \gamma_{i_k(\varepsilon)}^{(k)}(\varepsilon)\right| \geqslant \frac{1}{2m!}. \tag{5-48}$$

For a given ε we rearrange the eigenvalues $\lambda_i(\varepsilon)$ so that $i_k(\varepsilon)=k$, which by (5-47), (5-48) yields

$$|\gamma_i^{(i)}(\varepsilon)| \geqslant \frac{2^{m-2}}{3^{m-1}m!}, \quad i=1,\dots,m.$$

In view of the definition of $\gamma_k^{(k)}(\varepsilon)$, (5-42), and the normalization of $\psi_\varepsilon^{(i)}$ it follows that

$$|\lambda_i(\varepsilon) - \widehat{\lambda}_i(\varepsilon)| \leqslant \frac{3^{m-1} m!}{2^{m-2}} \big|(F_\varepsilon^{(i)}, \psi_\varepsilon^{(i)})_{L_2(S_\varepsilon)}\big| \leqslant C\varepsilon^{5\alpha/2}.$$

Choosing $\alpha > 4/5$, we arrive at the asymptotics (2-11).

Define now

$$\widetilde{\psi}_\varepsilon^{(k)} = \mathcal{I}_\varepsilon(\boldsymbol{\psi}_k + \varepsilon^2 \ln\varepsilon\, \boldsymbol{\phi}_k)\chi\left(\frac{\tau}{\varepsilon^\alpha}\right) + \left(v_0^{(k)} + \sum_{i=2}^{4} \varepsilon^i v_i^{(k)}\right)\left(1-\chi\left(\frac{\tau}{\varepsilon^\alpha}\right)\right).$$

By direct calculations one can check that

$$\|\widehat{\psi}_\varepsilon^{(k)} - \widetilde{\psi}_\varepsilon^{(k)}\|_{W_2^1(S_\varepsilon)} = \mathcal{O}(\varepsilon^{\frac{5\alpha}{2}}).$$

This identity and (5-45) imply

$$\sum_{i=1}^{m} \gamma_i^{(k)}(\varepsilon)\psi_\varepsilon^{(i)} = \psi_\varepsilon^{(k)} + \mathcal{O}(\varepsilon^{\frac{5\alpha}{2}}), \quad k = 1, \dots, m.$$

Since the right-hand sides of these identities are linearly independent, the functions $\sum_{i=1}^{m} \gamma_i^{(k)}(\varepsilon)\psi_\varepsilon^{(i)}$ form a basis spanned over the eigenfunctions $\psi_\varepsilon^{(i)}$, $i = 1, \dots, m$. Hence, we arrive at:

Theorem 5.3. *Let $\mathcal{P}_\varepsilon$ be the total projector associated with the eigenvalues $\lambda_i(\varepsilon)$, $i = 1, \dots, m$, and $\widetilde{\mathcal{P}}_\varepsilon$ be the projector on the space spanned over $\widetilde{\psi}_\varepsilon^{(i)}$, $i = 1, \dots, m$. Then*

$$\mathcal{P}_\varepsilon = \widetilde{\mathcal{P}}_\varepsilon + \mathcal{O}(\varepsilon^{2+\rho}),$$

where ρ is any constant in $(0, 1/2)$.

References

[Abreu and Freitas 2002] M. Abreu and P. Freitas, "On the invariant spectrum of S^1-invariant metrics on S^2", *Proc. London Math. Soc.* (3) **84**:1 (2002), 213–230. MR 2003a:58041 Zbl 1015.58012

[Borisov and Cardone 2011] D. Borisov and G. Cardone, "Complete asymptotic expansions for eigenvalues of Dirichlet Laplacian in thin three-dimensional rods", *ESAIM Control Optim. Calc. Var.* **17**:3 (2011), 887–908. MR 2012h:35034 Zbl 1223.35248

[Borisov and Freitas 2009] D. Borisov and P. Freitas, "Singular asymptotic expansions for Dirichlet eigenvalues and eigenfunctions of the Laplacian on thin planar domains", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **26**:2 (2009), 547–560. MR 2010b:35059 Zbl 1168.35401

[Borisov and Freitas 2010] D. Borisov and P. Freitas, "Asymptotics of Dirichlet eigenvalues and eigenfunctions of the Laplacian on thin domains in $\mathbb{R}^d$", *J. Funct. Anal.* **258**:3 (2010), 893–912. MR 2011a:35362 Zbl 1180.35397

[Borisov and Freitas 2012] D. I. Borisov and P. Freitas, "Eigenvalue asymptotics for almost flat compact hypersurfaces", *Dokl. Akad. Nauk* **442**:2 (2012), 151–155. In Russian; translated in *Dokl. Math.* **85**:1 (2012), 18–22. MR 2962172 Zbl 1247.58009

[Colbois et al. 2008] B. Colbois, E. B. Dryden, and A. El Soufi, "Extremal G-invariant eigenvalues of the Laplacian of G-invariant metrics", *Math. Z.* **258**:1 (2008), 29–41. MR 2008m:58066 Zbl 1127.58026

[Exner and Post 2005] P. Exner and O. Post, "Convergence of spectra of graph-like thin manifolds", *J. Geom. Phys.* **54**:1 (2005), 77–115. MR 2006a:58038 Zbl 1095.58007

[Exner and Post 2009] P. Exner and O. Post, "Approximation of quantum graph vertex couplings by scaled Schrödinger operators on thin branched manifolds", *J. Phys. A* **42**:41 (2009), 415305. MR 2011a:81098 Zbl 1179.81080

[Freitas 2007] P. Freitas, "Precise bounds and asymptotics for the first Dirichlet eigenvalue of triangles and rhombi", *J. Funct. Anal.* **251**:1 (2007), 376–398. MR 2008k:35339 Zbl 1137.35049

[Friedlander and Solomyak 2009] L. Friedlander and M. Solomyak, "On the spectrum of the Dirichlet Laplacian in a narrow strip", *Israel J. Math.* **170** (2009), 337–354. MR 2010f:35271 Zbl 1173.35090

[Grieser 2008] D. Grieser, "Thin tubes in mathematical physics, global analysis and spectral geometry", pp. 565–593 in *Analysis on graphs and its applications*, edited by P. Exner et al., Proc. Sympos. Pure Math. **77**, Amer. Math. Soc., Providence, RI, 2008. MR 2010g:58025 Zbl 1158.58001

[Il′in 1992] A. M. Il′in, *Matching of asymptotic expansions of solutions of boundary value problems*, Translations of Mathematical Monographs **102**, American Mathematical Society, Providence, RI, 1992. MR 93g:35016 Zbl 0754.34002

[Kato 1966] T. Kato, *Perturbation theory for linear operators*, Die Grundlehren der mathematischen Wissenschaften **132**, Springer, New York, 1966. MR 34 #3324 Zbl 0148.12601

[Krejčiřík and Zuazua 2010] D. Krejčiřík and E. Zuazua, "The Hardy inequality and the heat equation in twisted tubes", *J. Math. Pures Appl.* (9) **94**:3 (2010), 277–303. MR 2011e:35142 Zbl 1217.35029

[Nazarov 2001] S. A. Nazarov, *Asymptotic theory of thin plates and rods, 1: Dimension reduction and integral estimates*, Nauchnaya Kniga, Novosibirsk, 2001. In Russian.

[Panasenko 2005] G. Panasenko, *Multi-scale modelling for structures and composites*, Springer, Berlin, 2005. MR 2006g:74071 Zbl 1078.74002

[Reed and Simon 1980] M. Reed and B. Simon, *Methods of modern mathematical physics, I: Functional analysis*, 2nd ed., Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1980. MR 85e:46002 Zbl 0459.46001

Received 2 May 2012. Revised 4 Oct 2012. Accepted 14 Feb 2013.

DENIS BORISOV: `borisovdi@yandex.ru`
Department of Physics and Mathematics, Bashkir State Pedagogical University, Ufa, 3a Oktyabrskoy Revolutsii st., 450000, Russian Federation

and

Institute of Mathematics USC RAS, Ufa, 112 Chernyshevsky str., 450000, Russian Federation

PEDRO FREITAS: `freitas@cii.fc.ul.pt`
Department of Mathematics, Faculty of Human Kinetics, and *Group of Mathematical Physics, University of Lisbon, Complexo Interdisciplinar, Av. Prof. Gama Pinto 2, P-1649-003 Lisboa, Portugal*

ANALYSIS AND PDE
Vol. 6, No. 5, 2013
dx.doi.org/10.2140/apde.2013.6.1089

STABILIZATION FOR THE SEMILINEAR WAVE EQUATION WITH GEOMETRIC CONTROL CONDITION

ROMAIN JOLY AND CAMILLE LAURENT

In this article, we prove the exponential stabilization of the semilinear wave equation with a damping effective in a zone satisfying the geometric control condition only. The nonlinearity is assumed to be subcritical, defocusing and analytic. The main novelty compared to previous results is the proof of a unique continuation result in large time for some undamped equation. The idea is to use an asymptotic smoothing effect proved by Hale and Raugel in the context of dynamical systems. Then, once the analyticity in time is proved, we apply a unique continuation result with partial analyticity due to Robbiano, Zuily, Tataru and Hörmander. Some other consequences are also given for the controllability and the existence of a compact attractor.

Dans cet article, on prouve la décroissance exponentielle de l'équation des ondes semilinéaires avec un amortissement actif dans une zone satisfaisant seulement la condition de contrôle géométrique. La nonlinéarité est supposée sous-critique, défocalisante et analytique. La principale nouveauté par rapport aux résultats précédents est la preuve d'un résultat de prolongement unique en grand temps pour une solution non amortie. L'idée est d'utiliser un effet régularisant asymptotique prouvé par Hale et Raugel dans le contexte des systèmes dynamiques. Ensuite, une fois l'analyticité en temps prouvée, on applique un théorème de prolongement unique avec analyticité partielle dû à Robbiano, Zuily, Tataru et Hörmander. Des applications à la contrôlabilité et à l'existence d'attracteur global compact pour l'équation des ondes sont aussi données.

1. Introduction

In this article, we consider the semilinear damped wave equation

$$\begin{cases} \Box u+\gamma(x)\partial_t u+\beta u+f(u)=0 & (t,x)\in\mathbb{R}_+\times\Omega, \\ u(t,x)=0 & (t,x)\in\mathbb{R}_+\times\partial\Omega, \\ (u,\partial_t u)=(u_0,u_1)\in H_0^1(\Omega)\times L^2(\Omega), \end{cases} \tag{1-1}$$

where $\Box=\partial_{tt}^2-\Delta$, with Δ being the Laplace–Beltrami operator with Dirichlet boundary conditions. The domain Ω is a connected $\mathscr{C}^\infty$ three-dimensional Riemannian manifold with boundaries, which is either:

(i) Compact.

(ii) A compact perturbation of $\mathbb{R}^3$, that is $\mathbb{R}^3\setminus D$, where D is a bounded smooth domain, endowed with a smooth metric equal to the euclidean one outside of a ball.

Lauren was financed by the ERC grant GeCoMethods and by the project EMAQS ANR-2011-BS01-017-01 during part of the writing of this article.
MSC2010: primary 35B40, 35B60, 35B65, 93D20, 35L71; secondary 35B41.
Keywords: damped wave equation, stabilization, analyticity, unique continuation property, compact attractor.

(iii) A manifold with periodic geometry (cylinder, $\mathbb{R}^3$ with a periodic metric, …).

The nonlinearity $f \in \mathscr{C}^1(\mathbb{R}, \mathbb{R})$ is assumed to be defocusing, energy subcritical and such that 0 is an equilibrium point. More precisely, we assume that there exists $C > 0$ such that

$$f(0) = 0, \quad sf(s) \geq 0, \quad |f(s)| \leq C(1 + |s|)^p, \quad |f'(s)| \leq C(1 + |s|)^{p-1}, \tag{1-2}$$

with $1 \leq p < 5$.

We assume $\beta \geq 0$ to be such that $\Delta - \beta$ is a negative-definite operator, that is that we have a Poincaré inequality $\int_\Omega |\nabla u|^2 + \beta |u|^2 \geq C \int_\Omega |u|^2$ with $C > 0$. In particular, this may require $\beta > 0$ if $\partial\Omega = \varnothing$ or if Ω is unbounded.

The damping $\gamma \in L^\infty(\Omega)$ is a nonnegative function. We assume that there exist an open set $\omega \subset \Omega$, $\alpha \in \mathbb{R}$, $x_0 \in \Omega$ and $R \geq 0$ such that

$$\Omega \setminus B(x_0, R) \subset \omega \qquad \text{and} \qquad \gamma(x) \geq \alpha > 0 \quad \text{for all } x \in \omega. \tag{1-3}$$

Moreover, we assume that ω satisfies the geometric control condition introduced in [Rauch and Taylor 1974; Bardos et al. 1992]:

(GCC) There exists $L > 0$ such that any generalized geodesic of Ω of length L meets the set ω where the damping is effective.

The associated energy $E \in \mathscr{C}^0(X, \mathbb{R}_+)$ is given by

$$E(u) := E(u, \partial_t u) = \frac{1}{2} \int_\Omega (|\partial_t u|^2 + |\nabla u|^2 + \beta |u|^2) + \int_\Omega V(u), \tag{1-4}$$

where $V(u) = \int_0^u f(s)ds$. Due to assumption (1-2) and the Sobolev embedding $H^1(\Omega) \hookrightarrow L^6(\Omega)$, this energy is well defined and, if u solves (1-1), we have, at least formally,

$$\partial_t E(u(t)) = -\int_\Omega \gamma(x) |\partial_t u(x, t)|^2 \, dx \leq 0. \tag{1-5}$$

The system is therefore dissipative. We are interested in the exponential decay of the energy of the nonlinear damped wave equation (1-1), that is, the property:

(ED) For any $E_0 \geq 0$, there exist $K > 0$ and $\lambda > 0$ such that, for all solutions u of (1-1) with $E(u(0)) \leq E_0$,

$$E(u(t)) \leq K e^{-\lambda t} E(u(0)) \quad \text{for all } t \geq 0.$$

Property (ED) means that the damping term $\gamma \partial_t u$ stabilizes any solution of (1-1) to zero, which is an important property from the dynamical and control points of view.

Our main theorem is as follows.

Theorem 1.1. *Assume that the damping γ satisfies* (1-3) *and the geometric control condition* (GCC). *If f is real analytic and satisfies* (1-2), *then the exponential decay property* (ED) *holds.*

Theorem 1.1 applies for nonlinearities f that are globally analytic. Of course, the nonlinearities $f(u) = |u|^{p-1}u$ are not analytic if $p \notin \{1, 3\}$, but we can replace these usual nonlinearities by similar ones as $f(u) = (u/\operatorname{th}(u))^{p-1}u$, which are analytic for all $p \in [1, 5)$. Note that the estimates (1-2) are only required for $s \in \mathbb{R}$, so that they do not imply that f is a polynomial. Moreover, we can show that (ED) holds in fact for almost all the nonlinearities f satisfying (1-2), including nonanalytic ones.

More precisely, we set

$$\mathfrak{C}^1(\mathbb{R}) = \{f \in \mathscr{C}^1(\mathbb{R}) \mid \text{there exist } C > 0 \text{ and } p \in [1,5) \text{ such that (1-2) holds}\} \tag{1-6}$$

and endow this set with the Whitney topology (or any other reasonable topology). We recall that the Whitney topology is the topology generated by the neighborhoods

$$\mathcal{N}_{f,\delta} = \big\{g \in \mathfrak{C}^1(\mathbb{R}) \mid \max\big(|f(u) - g(u)|, |f'(u) - g'(u)|\big) < \delta(u) \text{ for all } u \in \mathbb{R}\big\}, \tag{1-7}$$

where f is any function in $\mathfrak{C}^1(\mathbb{R})$ and δ is any positive continuous function. The set $\mathfrak{C}^1(\mathbb{R})$ is a Baire space, which means that any generic set, that is, any set containing a countable intersection of open and dense sets, is dense in $\mathfrak{C}^1(\mathbb{R})$ (see Proposition 7.1). The Baire property ensures that the genericity of a set in $\mathfrak{C}^1(\mathbb{R})$ is a good notion for "the set contains almost all nonlinearities f".

Theorem 1.2. *Assume that the damping γ satisfies* (1-3) *and the geometric control condition* (GCC). *There exists a generic set $\mathfrak{G} \subset \mathfrak{C}^1(\mathbb{R})$ such that the exponential decay property* (ED) *holds for all $f \in \mathfrak{G}$.*

The statements of both theorems lead to some remarks.

- Of course, our results and their proofs should easily extend to any space dimension $d \geq 3$ if the exponent p of the nonlinearity satisfies $p < (d+2)/(d-2)$.

- Actually, it may be possible to get $\lambda > 0$ in (ED) uniform with respect to the size of the data. We can take for instance $\lambda = \tilde{\lambda} - \varepsilon$, where $\tilde{\lambda}$ is the decay rate of the linear equation. The idea is that once we know the existence of a decay rate, we know that the solution is close to zero for a large time. Then, for small solutions, the nonlinear term can be neglected to get almost the same decay rate as the linear equation. We refer for instance to [Laurent et al. 2010] in the context of KdV equation. Notice that the possibility to get the same result with a constant K independent of E_0 is an open problem.

- The assumption on β is important to ensure some coercivity of the energy and to preclude the spatially constant functions to be undamped solutions for the linear equation. It has been proved in [Dehman and Gérard 2002] for $\mathbb{R}^3$ and in [Laurent 2011] for a compact manifold that exponential decay can fail without this term β.

- The geometric control condition is known to be not only sufficient but also necessary for the exponential decay of the linear damped equation. The proof of the optimality uses some sequences of solutions which are asymptotically concentrated outside of the damping region. We can use the same idea in our nonlinear stabilization context. First, the observability for a certain time eventually large is known to be equivalent to the exponential decay of the energy. This was for instance noticed in [Dehman and Gérard 2002, Proposition 2] in a similar context; see also Proposition 2.5 of this paper. Then we take as initial data the same sequence that would give a counterexample for the linear observability. The linearizability property (see [Gérard 1996]) allows to obtain that the nonlinear solution is asymptotically close to the linear one. This contradicts the observability property for the nonlinear solution as it does for the linear case. Hence the geometric control condition is also necessary for the exponential decay of the nonlinear equation.

• Our geometrical hypotheses on Ω may look strange, however they are only assumed for sake of simplicity. In fact, our results should apply more generally for any smooth manifold with bounded geometry, that is, such Ω that can be covered by a set of $\mathscr{C}^\infty$ charts $\alpha_i : U_i \mapsto \alpha_i(U_i) \subset \mathbb{R}^3$ such that $\alpha_i(U_i)$ is equal either to $B(0,1)$ or to $B_+(0,1) = \{x \in B(0,1), x_1 > 0\}$ (in the case with boundaries) and such that, for any $r \geq 0$ and $s \in [1,\infty]$, the $W^{r,s}$ norm of a function u in $W^{r,s}(\Omega,\mathbb{R})$ is equivalent to the norm $\big(\sum_{i\in\mathbb{N}} \|u \circ \alpha_i^{-1}\|^s_{W^{r,s}(\alpha_i(U_i))}\big)^{1/s}$.

The stabilization property (ED) for Equation (1-1) has been studied in [Haraux 1985a; Zuazua 1990; 1991; Dehman 2001] for $p < 3$. For $p \in [3,5)$, our main reference is the work of Dehman, Lebeau and Zuazua [Dehman et al. 2003]. This work is mainly concerned with the stabilization problem previously described on the Euclidean space $\mathbb{R}^3$ with flat metric and stabilization active outside of a ball. The main purpose of this paper is to extend their result to a nonflat geometry where multiplier methods cannot be used or do not give the optimal result with respect to the geometry. Other stabilization results for the nonlinear wave equation can be found in [Aloui et al. 2011] and the references therein. Some works have been done in the difficult critical case $p = 5$; we refer to [Dehman and Gérard 2002; Laurent 2011].

The proofs in these articles use three main ingredients:

(i) The exponential decay of the linear equation, which is equivalent to the geometric control condition (GCC).

(ii) A more or less involved compactness argument.

(iii) A unique continuation result implying that $u \equiv 0$ is the unique solution of

$$\begin{cases} \Box u + \beta u + f(u) = 0, & \\ \partial_t u = 0 & \text{on } [-T,T]\times\omega. \end{cases} \tag{1-8}$$

The results are mainly of the type “geometric control condition” plus “unique continuation” implies “exponential decay”. This type of implication is even stated explicitly in some related works for the nonlinear Schrödinger equation [Dehman et al. 2006; Laurent 2010].

In the subcritical case $p < 5$, the less understood point is the unique continuation property (iii). In the previous works as [Dehman et al. 2003], the authors use unique continuation results based on Carleman estimates. The resulting geometric assumptions are not very natural and are stronger than (GCC). Indeed, the unique continuation was often proved with some Carleman estimates that required some strong geometric conditions. For instance for a flat metric, the usual geometric assumption that appear are often of “multiplier type” that is ω is a neighborhood of $\{x \in \partial\Omega \mid (x - x_0)\cdot n(x) > 0\}$ which are known to be stronger than the geometric control condition (see [Miller 2002] for a discussion about the links between these assumptions). Moreover, on curved spaces, this type of condition often needs to be checked by hand in each situation, which is mostly impossible.

Our main improvement in this paper is the proof of unique continuation in infinite time under the geometric control condition only. We show that, if the nonlinearity f is analytic (or generic), then one can use the result of Robbiano and Zuily [1998] to obtain a unique continuation property (iii) for infinite time $T = +\infty$ with the geometric control condition (GCC) only.

The central argument of the proof of our main result, Theorem 1.1, is the unique continuation property of [Robbiano and Zuily 1998] (see Section 3). This result applies for solutions u of (1-8) being smooth in space and analytic in time. If f is analytic, then the solutions of (1-1) are of course not necessarily analytic in time since the damped wave equations are not smoothing in finite time. However, the damped wave equations admit an asymptotic smoothing effect, i.e., are smoothing in infinite time. Hale and Raugel [2003] have shown that, for compact trajectories, this asymptotic smoothing effect also concerns the analyticity (see Section 5). In other words, combining [Robbiano and Zuily 1998] and [Hale and Raugel 2003] shows that the unique solution of (1-8) is $u \equiv 0$ if f is analytic and if $T = +\infty$. This combination has already been used by dynamicists for $p < 3$ (Hale and Raugel, private communication; [Joly 2007]).

One of the main interests of this paper is the use of arguments coming from both the dynamical study and the control theory of the damped wave equations. The reader familiar with the control theory could find interesting the use of the asymptotic smoothing effect to get unique continuation property with smooth solutions. The one familiar with the dynamical study of PDEs could be interested in the use of Strichartz estimates to deal with the case $p \in [3, 5)$. The main part of the proof of Theorem 1.1 is written with arguments coming from the dynamical study of PDEs. They are simpler than the corresponding ones of control theory, but far less accurate since they do not give any estimation for the time of observability. Anyway, such accuracy is not important here since we use the unique continuation property for (1-8) with $T = +\infty$. We briefly recall in Section 8 how these propagation of compactness and regularity properties could have been proved with some arguments more usual in the control theory.

Moreover, we give two applications of our results in both contexts of control theory and dynamical systems. First, as it is usual in control theory, some results of stabilization can be coupled with local control theorems to provide global controllability in large time.

Theorem 1.3. *Assume that f satisfies the conditions of Theorem 1.1 or belongs to the generic set $\mathfrak{G}$ defined by Theorem 1.2. Let $R_0 > 0$ and ω satisfying the geometric control condition. Then there exists $T > 0$ such that for any (u_0, u_1) and $(\tilde{u}_0, \tilde{u}_1)$ in $H_0^1(\Omega) \times L^2(\Omega)$ with*

$$\|(u_0, u_1)\|_{H^1 \times L^2} \leq R_0 \quad \textit{and} \quad \|(\tilde{u}_0, \tilde{u}_1)\|_{H^1 \times L^2} \leq R_0$$

there exists $g \in L^\infty([0, T], L^2(\Omega))$ supported in $[0, T] \times \omega$ such that the unique strong solution of

$$\begin{cases} \Box u + \beta u + f(u) = g & \text{on } [0, T] \times \Omega, \\ (u(0), \partial_t u(0)) = (u_0, u_1), \end{cases}$$

satisfies $(u(T), \partial_t u(T)) = (\tilde{u}_0, \tilde{u}_1)$.

The second application of our results concerns the existence of a compact global attractor. A compact global attractor is a compact set, which is invariant by the flow of the PDE and which attracts the bounded sets. The existence of such an attractor is an important dynamical property because it roughly says that the dynamics of the PDE may be reduced to dynamics on a compact set, which is often finite-dimensional. See [Hale 1988; Raugel 2002] for reviews of this concept. Theorems 1.1 and 1.2 show that $\{0\}$ is a global attractor for the damped wave equation (1-1). Of course, it is possible to obtain a more complex attractor by considering an equation of the type

$$\begin{cases} \partial_{tt}^2 u + \gamma(x)\partial_t u = \Delta u - \beta u - f(x,u) & (x,t) \in \Omega \times \mathbb{R}_+, \\ u(x,t) = 0 & (x,t) \in \partial\Omega \times \mathbb{R}_+, \\ (u, \partial_t u) = (u_0, u_1) \in H_0^1 \times L^2, \end{cases} \tag{1-9}$$

where $f \in C^\infty(\Omega \times \mathbb{R}, \mathbb{R})$ is real analytic with respect to u and satisfies the following properties. There exist $C > 0$, $p \in [1, 5)$ and $R > 0$ such that for all $(x, u) \in \Omega \times \mathbb{R}$,

$$|f(x,u)| \le C(1+|u|)^p, \quad |f_x'(x,u)| \le C(1+|u|)^p, \quad |f_u'(x,u)| \le C(1+|u|)^{p-1}, \tag{1-10}$$

$$x \in \partial\Omega \quad \Longrightarrow \quad f(x,0) = 0, \tag{1-11}$$

$$(x \notin B(x_0, R) \quad \text{or} \quad |u| \ge R) \quad \Longrightarrow \quad f(x,u)u \ge 0, \tag{1-12}$$

where x_0 denotes a fixed point of the manifold.

Theorem 1.4. *Assume f is as above. Then the dynamical system generated by* (1-9) *in $H_0^1(\Omega) \times L^2(\Omega)$ is gradient and admits a compact global attractor $\mathcal{A}$.*

Of course, we would get the same result for f in a generic set similar to the one of Theorem 1.2.

We begin this paper by setting our main notations and recalling the basic properties of Equation (1-1) in Section 2. We recall the unique continuation property of Robbiano and Zuily in Section 3, whereas Sections 4 and 5 are concerned by the asymptotic compactness and the asymptotic smoothing effect of the damped wave equation. The proofs of our main results, Theorem 1.1 and 1.2, are given in Sections 6 and 7, respectively. An alternative proof, using more usual arguments from control theory, is sketched in Section 8. Finally, Theorems 1.3 and 1.4 are discussed in Section 9.

2. Notations and basic properties of the damped wave equation

In this paper, we use the following notations:

$$U = (u, u_t), \quad F = (0, f), \quad A = \begin{pmatrix} 0 & \mathrm{Id} \\ \Delta - \beta & -\gamma \end{pmatrix}.$$

In this setting, (1-1) becomes

$$\partial_t U(t) = AU(t) + F(U).$$

We set $X = H_0^1(\Omega) \times L^2(\Omega)$ and for $s \in [0, 1]$, we denote by X^s the space

$$X^s = D\big((-\Delta + \beta)^{(s+1)/2}\big) \times D\big((-\Delta+\beta)^{s/2}\big) = (H^{1+s}(\Omega) \cap H_0^1(\Omega)) \times H_0^s(\Omega).$$

Notice that $X^0 = X$ and $X^1 = D(A)$ (even if γ is only in L^∞).

We recall that E denotes the energy defined by (1-4). We also emphasize that (1-2) and the invertibility of $\Delta - \beta$ implies that a set is bounded in X if and only if its energy E is bounded. Moreover, for all $E_0 \ge 0$, there exists $C > 0$ such that

$$E(u,v) \le E_0 \quad \text{for all } (u,v) \in X \qquad \Longrightarrow \qquad \frac{1}{C}\,\|(u,v)\|_X^2 \le E(u,v) \le C\,\|(u,v)\|_X^2. \tag{2-1}$$

To simplify some statements in the proofs, we assume without loss of generality that $3 < p < 5$. This will avoid some meaningless statements with negative Lebesgue exponents since $p = 3$ is the exponent where Strichartz estimates are not necessary and can be replaced by Sobolev embeddings.

We recall that Ω is endowed with a metric g. We denote by d the distance on Ω defined by

$$d(x, y) = \inf\{l(c) \mid c \in \mathscr{C}^\infty([0, 1], \Omega) \text{ with } c(0) = x \text{ and } c(1) = y\},$$

where $l(c)$ is the length of the path c according to the metric g. A ball $B(x, R)$ in Ω is naturally defined by

$$B(x, R) = \{y \in \Omega, d(x, y) < R\}.$$

For instance, if $\Omega = \mathbb{R}^3 \setminus B_{\mathbb{R}^3}(0, 1)$, the distance between $(0, 0, 1)$ and $(0, 0, -1)$ is π (and not 2) and the ball $B((0, 0, 1), \pi)$ has nothing to do with the classical ball $B_{\mathbb{R}^3}((0, 0, 1), \pi)$ of $\mathbb{R}^3$.

Cauchy problem. The global existence and uniqueness of solutions of the subcritical wave equation (1-1) with $\gamma \equiv 0$ has been studied in [Ginibre and Velo 1985; 1989]. Their method also applies for $\gamma \not\equiv 0$ since this term is linear and well defined in the energy space X. Moreover, their argument to prove uniqueness also yields the continuity of the solutions with respect to the initial data.

The central argument is the use of Strichartz estimates.

Theorem 2.1 (Strichartz estimates). *Let $T > 0$ and (q, r) satisfy*

$$\frac{1}{q} + \frac{3}{r} = \frac{1}{2}, \quad q \in [7/2, +\infty]. \tag{2-2}$$

There exists $C = C(T, q) > 0$ such that for every $G \in L^1([0, T], L^2(\Omega))$ and every $(u_0, u_1) \in X$, the solution u of

$$\begin{cases} \Box u + \gamma(x)\partial_t u = G(t), \\ \quad (u, \partial_t u)(0) = (u_0, u_1), \end{cases}$$

satisfies the estimate

$$\|u\|_{L^q([0,T],L^r(\Omega))} \leq C \left(\|u_0\|_{H^1(\Omega)} + \|u_1\|_{L^2(\Omega)} + \|G\|_{L^1([0,T],L^2(\Omega))}\right).$$

The result was stated in the Euclidean space $\mathbb{R}^3$ by Strichartz [1977] and Ginibre and Velo with $q \in (2, +\infty]$. Kapitanskiĭ [1990] extended the result to variable coefficients. On a bounded domain, the first estimates were proved by Burq, Lebeau and Planchon [Burq et al. 2008] for $q \in [5, +\infty]$ and extended to a larger range by Blair, Smith and Sogge in [Blair et al. 2009]. Note that, thanks to the counterexamples of Ivanovici [2012], we know that we cannot expect some Strichartz estimates in the full range of exponents in the presence of boundaries.

From these results, we deduce the estimates for the damped wave equation by absorption for T small enough. We can iterate the operation in a uniform number of steps. Actually, for the purpose of the semilinear wave equation, it is sufficient to consider the Strichartz estimate $L^{2p/(p-3)}([0, T], L^{2p}(\Omega))$, which gives $u^p \in L^{2/(p-3)}([0, T], L^2(\Omega)) \subset L^1([0, T], L^2(\Omega))$ because $1 < 2/(p-3) < +\infty$.

Theorem 2.2 (Cauchy problem). *Let f satisfy* (1-2). *Then for any $(u_0, u_1) \in X = H_0^1(\Omega) \times L^2(\Omega)$ there exists a unique solution $u(t)$ of the subcritical damped wave equation* (1-1). *Moreover, this solution is defined for all $t \in \mathbb{R}$ and its energy $E(u(t))$ is nonincreasing in time.*

For any $E_0 \geq 0$, $T \geq 0$ and (q, r) satisfying (2-2), *there exists a constant C such that if u is a solution of* (1-1) *with $E(u(0)) \leq E_0$, then*

$$\|u\|_{L^q([0,T],L^r(\Omega))} \leq C \left(\|u_0\|_{H^1(\Omega)} + \|u_1\|_{L^2(\Omega)} \right).$$

In addition, for any $E_0 \geq 0$ and $T \geq 0$, there exists a constant C such that if u and $\tilde{u}$ are two solutions of (1-1) *with $E(u(0)) \leq E_0$ and $E(\tilde{u}(0)) \leq E_0$, then*

$$\sup_{t\in[-T,T]} \|(u, \partial_t u)(t) - (\tilde{u}, \partial_t \tilde{u})(t)\|_X \leq C\|(u, \partial_t u)(0) - (\tilde{u}, \partial_t \tilde{u})(0)\|_X.$$

Proof. The existence and uniqueness for small times is a consequence of the Strichartz estimates and of the subcriticality of the nonlinearity; see [Ginibre and Velo 1989]. The solution can be globalized backward and forward in time thanks to the energy estimates (1-5) for smooth solutions. Indeed,

$$E(t) \leq E(s) + C \int_t^s E(\tau)\, d\tau,$$

and thus Gronwall inequality for $t \leq s$ and the decay of energy for $t \geq s$ show that the energy does not blow up in finite time. This allows us to extend the solution for all times since the energy controls the norm of the space X by (2-1).

For the uniform continuity estimate, we notice that $w = u - \tilde{u}$ is solution of

$$\begin{cases} \Box w + \beta w + \gamma(x)\partial_t w = -w g(u, \tilde{u}), \\ (w, \partial_t w)(0) = (u, \partial_t u)(0) - (\tilde{u}, \partial_t \tilde{u})(0), \end{cases}$$

where $g(s, \tilde{s}) = \int_0^1 f'(s + \tau(\tilde{s} - s))\, d\tau$ fulfills $|g(s, \tilde{s})| \leq C(1 + |s|^{p-1} + |\tilde{s}|^{p-1})$. Let $q = 2p/(p-3)$, then the Strichartz and Hölder estimates give

$$\begin{aligned} \|(w, \partial_t w)(t)\|_{L^\infty([0,T],X)\cap L^q([0,T],L^{2p})} &\leq C\|(w, \partial_t w)(0)\|_X + C\|w g(u, \tilde{u})\|_{L^1([0,T],L^2)} \\ &\leq C\|(w, \partial_t w)(0)\|_X + CT\|w\|_{L^\infty([0,T],L^2)} \\ &\quad + T^\theta \|w\|_{L^q([0,T],L^{2p})} \Big(\|u\|^{p-1}_{L^q([0,T],L^{2p})} + \|\tilde{u}\|^{p-1}_{L^q([0,T],L^{2p})} \Big) \end{aligned}$$

with $\theta = (5 - p)/2 > 0$. We get the expected result for T small enough by absorption since we already know a uniform bound (depending on E_0) for the Strichartz norms of u and $\tilde{u}$. Then we iterate the operation to get the result for large T. ☐

Exponential decay of the linear semigroup. In this paper, we will strongly use the exponential decay for the linear semigroup in the case where γ may vanish but satisfies the geometric assumptions of this paper. In this case, (1-3) enables us to control the decay of energy outside a large ball and the geometric control condition (GCC) enables to control the energy trapped in this ball.

Proposition 2.3. *Assume that* $\gamma \in L^\infty(\Omega)$ *satisfies* (1-3) *and* (GCC). *There exist two positive constants* C *and* λ *such that*

$$|||e^{At}|||_{\mathscr{L}(X^s)} \le Ce^{-\lambda t} \quad \text{for all } s \in [0,1] \text{ and all } t \ge 0.$$

The exponential decay of the damped wave equation under the geometric control condition is well known since the works of Rauch and Taylor [1974] on a compact manifold and Bardos, Lebeau and Rauch [Bardos et al. 1988; 1992] on a bounded domain. Yet we did not find any reference for unbounded domains ([Aloui and Khenissi 2002; Khenissi 2003] concern unbounded domains but local energy only). It is noteworthy that the decay of the linear semigroup in unbounded domains seems not to have been extensively studied for the moment.

We give a proof of Proposition 2.3 using a microlocal defect measure as done in [Lebeau 1996; Burq 1997a] (see also [Burq and Gérard 1997] for the proof of the necessity). The only difference with respect to these results is that the manifold that we consider may be unbounded. Since a microlocal defect measure only reflects the local propagation, we thus have to use the property of equipartition of the energy to deal with the energy at infinity and to show a propagation of compactness (see [Dehman et al. 2003] for the flat case).

Lemma 2.4. *Let* $T > L$, *where* L *is given by* (GCC). *Assume that* $(U_{n,0}) \subset X$ *is a bounded sequence, which weakly converges to* 0 *and assume that* $U_n(t) = (u_n(t), \partial_t u_n(t)) = e^{At}U_{n,0}$ *satisfies*

$$\int_0^T \int_\Omega \gamma(x)|\partial_t u_n|^2 \to 0. \tag{2-3}$$

Then $(U_{n,0})$ *converges to* 0 *strongly in* X.

Proof. Let μ be a microlocal defect measure associated to (u_n) (see [Gérard 1991; Tartar 1990; Burq 1997b] for the definition). Note that (2-3) implies that μ can also be associated to the solution of the wave equation without damping, so the weak regularity of γ is not problematic for the propagation and we get that μ is concentrated on $\{\tau^2 - |\xi|_x^2 = 0\}$, where (τ, ξ) are the dual variables of (t, x). Moreover, (2-3) implies that $\gamma\tau^2\mu = 0$ and so $\mu \equiv 0$ on $S^*(]0,T[\times\omega)$. Then, by using the propagation of the measure along the generalized bicharacteristic flow of Melrose–Sjöstrand and the geometric control condition satisfied by ω, we obtain $\mu \equiv 0$ everywhere. We do not give more details about propagation of microlocal defect measures and refer to the Appendix of [Lebeau 1996] or Section 3 of [Burq 1997b] (see also [Gérard and Leichtnam 1993] for some close propagation results in a different context). Since $\mu \equiv 0$, we know that

$$U_n \to 0 \quad \text{on } H^1 \times L^2(]0,T[\times B(x_0,R))$$

for every $R > 0$.

To finish the proof, we need the classical equipartition of the energy to get the convergence to 0 in the whole manifold Ω. Since γ is uniformly positive outside a ball $B(x_0,R)$, (2-3) and the previous arguments imply that

$$\partial_t u_n \to 0 \quad \text{in } L^2([0,T] \times \Omega).$$

Let $\varphi \in C_0^\infty(]0,T[)$ with $\varphi \geq 0$ and $\varphi(t)=1$ for $t \in [\varepsilon, T-\varepsilon]$. We multiply the equation by $\varphi(t)u_n$ and we obtain

$$0 = -\iint_{[0,T]\times\Omega} \varphi(t)|\partial_t u_n|^2 - \iint_{[0,T]\times\Omega} \varphi'(t)\partial_t u_n u_n + \iint_{[0,T]\times\Omega} \varphi(t)|\nabla u_n|^2 + \iint_{[0,T]\times\Omega} \varphi(t)\beta|u_n|^2 + \iint_{[0,T]\times\Omega} \varphi(t)\gamma(x)\partial_t u_n u_n.$$

The L^2 norm of $u_n(t)$ is bounded, while $\partial_t u_n \to 0$ in $L^2([0,T]\times\Omega)$, so the first, second and fifth terms converge to zero. Then the above equation yields

$$\iint_{[0,T]\times\Omega} \varphi(t)\left(\beta|u_n|^2 + |\nabla u_n|^2\right) \to 0.$$

Finally, notice that the energy identity $\|U_{n,0}\|_X^2 = \|U_n(t)\|_X^2 + \int_0^T \int_\Omega \gamma(x)|\partial_t u_n|^2$ shows that

$$\iint_{[0,T]\times\Omega} \varphi(t)\left(\beta|u_n|^2 + |\nabla u_n|^2\right) \sim \|U_{n,0}\|_X^2 \int_0^T \varphi(t),$$

and thus that $\|U_{n,0}\|_X$ goes to zero. □

Proof of Proposition 2.3. Once Lemma 2.4 is established, the proof follows the arguments of the classical case, where Ω is bounded. We briefly recall them.

We first treat the case $s=0$. As in Proposition 2.5, the exponential decay of the energy is equivalent to the observability estimate, that is, the existence of $C>0$ and $T>0$ such that, for any trajectory $U(t)=e^{At}U_0$ in X,

$$\int_0^T \int_\Omega \gamma(x)|\partial_t u|^2 \geq C\|U(0)\|_X^2. \tag{2-4}$$

We argue by contradiction: Assume that (2-4) does not hold for any positive T and C. Then there exists a sequence of initial data $U_n(0)$ with $\|U_n(0)\|_X = 1$ and such that

$$\int_0^n \int_\Omega \gamma(x)|\partial_t u_n(t,x)|^2 dt dx \to 0 \quad \text{as } n \to +\infty,$$

where $(u_n, \partial_t u_n)(t) = U_n(t) = e^{At}U_n(0)$. Let $\widetilde{U}_n = U_n(n/2+\cdot)$. We have

$$\int_{-n/2}^{n/2} \int_\Omega \gamma(x)|\partial_t \tilde{u}_n(t,x)|^2 dt dx \to 0 \quad \text{as } n \to +\infty,$$

and, for any $t \in [-n/2, n/2]$,

$$\|\widetilde{U}_n(t)\|_X^2 = \|\widetilde{U}_n(-n/2)\|_X^2 - \int_{-n/2}^t \int_\Omega \gamma(x)|\partial_t \tilde{u}_n(s,x)|^2 ds dx \to 1 \quad \text{as } n \to +\infty.$$

We can thus assume that $U_n(0)$ converges to $U_\infty(0) \in X$, weakly in X. For any $T>0$, $U_n(t)$ and $\partial_t U_n(t)$ are bounded in $L^\infty([-T,T],X)$ and $L^\infty([-T,T],L^2(\Omega)\times H^{-1}(\Omega))$, respectively. Thus, by using Ascoli's theorem, we may also assume that $U_n(t)$ strongly converges to $U_\infty(t)$ in $L^\infty([-T,T], L^2(K)\times H^{-1}(K))$, where K is any compact of Ω. Hence $(u_\infty, \partial_t u_\infty)(t) = U_\infty(t) = e^{At}U_\infty(0)$ is a solution of

$$\begin{cases} \Box u_\infty + \beta u_\infty = 0 & \text{on } \mathbb{R}\times\Omega, \\ \partial_t u_\infty = 0 & \text{on } \mathbb{R}\times\omega. \end{cases} \tag{2-5}$$

in $L^2\times H^{-1}$. Since $U_\infty(0)\in X$ belongs to X, we deduce that, in fact, $U_\infty(t)$ solves (2-5) in X.

To finish the proof of Proposition 2.3, we have to show that $U_\infty \equiv 0$. Indeed, applying Lemma 2.4, we would get that U_n converges strongly to 0, which contradicts the hypothesis $\|U_n(0)\|_X = 1$. Note that $U_\infty \equiv 0$ is a direct consequence of a unique continuation property as in Corollary 3.2. However, Corollary 3.2 requires Ω to be smooth, whereas Proposition 2.3 could be more general. Therefore, we recall another classical argument to show that $U_\infty \equiv 0$.

Denote by N the set of functions $U_\infty(0)\in X$ satisfying (2-5), which is obviously a linear subspace of X. We will prove that $N=\{0\}$. Since $\gamma(x)|\partial_t u_\infty|^2 \equiv 0$ for functions u_∞ in N and since N is a closed subspace, Lemma 2.4 shows that any weakly convergent subsequence of N is in fact strongly convergent. By the Riesz theorem, N is therefore finite-dimensional. For any $t\in\mathbb{R}$, e^{tA} applies N into itself and thus $A_{|N}$ is a bounded linear operator. Assume that $N\neq\{0\}$, then $A_{|N}$ admits an eigenvalue λ with eigenvector $Y=(y_0,y_1)\in N$. This means that $y_1=\lambda y_0$ and that $(\Delta-\beta)y_0=\lambda^2 y_0$. Moreover, we know that $y_1=0$ on ω and so, if $\lambda\neq 0$, that $y_0=0$ on ω. This implies $y_0\equiv 0$ by the unique continuation property of elliptic operators. Finally, if $\lambda=0$, we have $(\Delta-\beta)y_0=0$ and $y_0=0$, because, by assumption, $\Delta-\beta$ is a negative definite operator.

So we have proved $N=\{0\}$ and therefore $U_\infty=0$, that is, $\widetilde{U}_n(0)$ converges to 0 weakly in X. We can then apply Lemma 2.4 on any interval $[-n/2, -n/2+T]$, where L is the time in the geometric control condition (GCC) and obtain a contradiction to $\|U_n(0)\|_X = 1$.

Let us now consider the cases $s\in(0,1]$. The basic semigroup properties (see [Pazy 1983]) show that, if $U\in X^1=D(A)$, then $e^{At}U$ belongs to $D(A)$ and

$$\begin{aligned} \|e^{At}U\|_{X^1} = \|Ae^{At}U\|_X + \|e^{At}U\|_X &= \|e^{At}AU\|_X + \|e^{At}U\|_X \\ &\le Ce^{-\lambda t}(\|AU\|_X + \|U\|_X) = Ce^{-\lambda t}\|U\|_{D(A)}. \end{aligned}$$

This shows Proposition 2.3 for $s=1$. Notice that we do not have to require any regularity for γ to obtain this result. Then Proposition 2.3 for $s\in(0,1)$ follows by interpolating between the cases $s=0$ and $s=1$ (see [Tartar 2007]). □

First nonlinear exponential decay properties. Theorem 2.2 shows that the energy E is nonincreasing along the solutions of (1-1). The purpose of this paper is to obtain the exponential decay of this energy in the sense of property (ED) stated above. We first recall the well-known criterion for exponential decay.

Proposition 2.5. *The exponential decay property* (ED) *holds if and only if there exist T and C such that*

$$E(u(0)) \le C(E(u(0)) - E(u(T))) = C\iint_{[0,T]\times\Omega} \gamma(x)|\partial_t u(x,t)|^2\,dtdx \tag{2-6}$$

for all solutions u of (1-1) *with $E(u(0))\le E_0$.*

Proof. If (ED) holds then (2-6) holds for T large enough since $E(u(0))-E(u(T))\geq(1-Ke^{-\lambda T})E(u(0))$. Conversely, if (2-6) holds, using $E(u(T))\leq E(u(0))$, we get $E(u(T))\leq C/(C+1)E(u(0))$ and thus $E(u(kT))\leq(C/(C+1))^kE(u(0))$. Using again the decay of the energy to fill the gaps $t\in(kT,(k+1)T)$, this shows that (ED) holds. □

First, we prove exponential decay in the case of positive damping, which will be helpful to study what happens outside a large ball since (1-3) is assumed in the whole paper. Note that the fact that $-\Delta+\beta$ is positive is necessary to avoid for instance the constant undamped solutions.

Proposition 2.6. *Assume that $\omega=\Omega$, that is that $\gamma(x)\geq\alpha>0$ everywhere. Then* (ED) *holds.*

Proof. We recall here the classical proof. We introduce a modified energy

$$\widetilde{E}(u)=\int_\Omega \tfrac{1}{2}(|\partial_t u|^2+|\nabla u|^2+\beta|u|^2)+V(u)+\varepsilon u\partial_t u$$

with $\varepsilon>0$. Since $\int_\Omega|\nabla u|^2+\beta|u|^2$ controls $\|u\|_{L^2}^2$, $\widetilde{E}$ is equivalent to E for ε small enough and it is sufficient to obtain the exponential decay of the auxiliary energy $\widetilde{E}$. Using $\gamma\geq\alpha>0$ and $uf(u)\geq 0$, a direct computation shows for ε small enough that

$$\begin{aligned}\widetilde{E}(u(T))-\widetilde{E}(u(0))&=\int_0^T\int_\Omega -\gamma(x)|\partial_t u|^2+\varepsilon|\partial_t u|^2+\varepsilon\gamma(x)u\partial_t u-\varepsilon(|\nabla u|^2+\beta|u|^2)-\varepsilon uf(u)\\&\leq -C\int_0^T\|(u,\partial_t u)\|_{H^1\times L^2}^2\leq -C\int_0^T\widetilde{E}(t)\,dt\leq -CT\widetilde{E}(T),\end{aligned}$$

where $C>0$ is a constant that may change from line to line. Thus, $\widetilde{E}(u(0))-\widetilde{E}(u(T))\geq CT\widetilde{E}(u(T))$ with $CT>0$ and therefore $\widetilde{E}(u(0))\geq\mu\widetilde{E}(u(T))$ with $\mu>1$. As in the proof of Proposition 2.3, this last property implies the exponential decay of $\widetilde{E}$ and thus the one of E. □

3. A unique continuation result for equations with partially holomorphic coefficients

Comparatively to previous articles on the stabilization of the damped wave equations as [Dehman et al. 2003], one of the main novelties of this paper is the use of a unique continuation theorem requiring partially analyticity of the coefficients, but very weak geometrical assumptions as shown in Corollary 3.2. We use here the following result of Robbiano and Zuily [1998]. This result has also been proved independently by Hörmander [1997] and has been generalized by Tataru [1999]. Note that the idea of using partial analyticity for unique continuation was introduced by Tataru [1995] but it requires some global analyticity assumptions that are not fulfilled in our case. All these results use very accurate microlocal analysis and hold in a much more general framework than the one of the wave equation. However, for sake of simplicity, we restrict the statement to this case.

Theorem 3.1. *Let $d\geq 1$, $(x_0,t_0)\in\mathbb{R}^d\times\mathbb{R}$ and let $\mathcal{U}$ be a neighborhood of (x_0,t_0). Let $(A_{i,j}(x,t))_{i,j=1,\dots,d}$, $b(x,t)$, $(c_i(x,t))_{i=1,\dots,d}$ and $d(x,t)$ be bounded coefficients in $\mathscr{C}^\infty(\mathcal{U},\mathbb{R})$. Let v be a strong solution of*

$$\partial_{tt}^2 v=\operatorname{div}(A(x,t)\nabla v)+b(x,t)\partial_t v+c(x,t)\nabla v+d(x,t)v,\qquad (x,t)\in\mathcal{U}\subset\mathbb{R}^d\times\mathbb{R}.\tag{3-1}$$

Let $\varphi\in\mathscr{C}^2(\mathcal{U},\mathbb{R})$ such that $\varphi(x_0,t_0)=0$ and $(\nabla\varphi,\partial_t\varphi)(x,t)\neq 0$ for all $(x,t)\in\mathcal{U}$. Assume that:

(i) *The coefficients* A, b, c *and* d *are analytic in time.*

(ii) $A(x_0, t_0)$ *is a symmetric positive definite matrix.*

(iii) *The hypersurface* $\{(x, t) \in \mathcal{U}, \varphi(x, t) = 0\}$ *is not characteristic at* (x_0, t_0), *that is, that we have* $|\partial_t \varphi(x_0, t_0)|^2 \neq \langle \nabla\varphi(x_0, t_0) \mid A(x_0, t_0)\nabla\varphi(x_0, t_0)\rangle$.

(iv) $v \equiv 0$ *in* $\{(x, t) \in \mathcal{U}, \varphi(x, t) \leq 0\}$.

Then $v \equiv 0$ *in a neighborhood of* (x_0, t_0).

Proof. We only have to show that Theorem 3.1 is a direct translation of Theorem A of [Robbiano and Zuily 1998] in the framework of the wave equation. To use the notations of [ibid.], we let x_a be the time variable and x_b the space variable and we set $(x_0, t_0) = x^0 = (x_b^0, x_a^0)$. Equation (3-1) corresponds to the differential operator

$$P = \xi_a^2 - {}^t\xi_b A(x_b, x_a)\xi_b - b(x_b, x_a)\xi_a - c(x_b, x_a)\xi_b - d(x_b, x_a)$$

with principal symbol $p_2 = \xi_a^2 - {}^t\xi_b A(x_b, x_a)\xi_b$.

All the statements of Theorem 3.1 are obvious translations of Theorem A of [ibid.], except maybe for the fact that hypothesis (iii) implies the hypothesis of pseudoconvexity of [ibid.]. We compute $\{p_2, \varphi\} = 2\xi_a\varphi'_a - 2{}^t\xi_b A(x_a, x_b)\varphi'_b$. Let us set $\zeta = (x_a^0, x_b^0, i\varphi'_a(x^0), \xi_b + i\varphi'_b(x^0))$, then $\{p_2, \varphi\}(\zeta) = 0$ if and only if

$$i(\varphi'_a(x^0))^2 - i{}^t\varphi'_b(x^0))A(x^0)\varphi'_b(x^0) -{}^t \xi_b A(x^0)\varphi'_b(x^0) = 0.$$

This is possible only if $(\varphi'_a(x^0))^2 = {}^t\varphi'_b(x^0)A(x^0)\varphi'_b(x^0)$, that is if the hypersurface $\varphi = 0$ is characteristic at (x_0, t_0). Thus, if this hypersurface is not characteristic, then the pseudoconvexity hypothesis of Theorem A of [ibid.] holds. □

The previous theorem allows us to prove some unique continuation result with some optimal time and geometric assumptions. This allows us to prove unique continuation where the geometric condition is only, roughly speaking, that we do not contradict the finite speed of propagation.

Corollary 3.2. *Let* $T > 0$ (*or* $T = +\infty$) *and let* b, $(c_i)_{i=1,2,3}$ *and* d *be coefficients in* $\mathcal{C}^\infty(\Omega \times [0, T], \mathbb{R})$. *Assume moreover that* b, c *and* d *are analytic in time and that* v *is a strong solution of*

$$\partial_{tt}^2 v = \Delta v + b(x, t)\partial_t v + c(x, t)\nabla v + d(x, t)v, \quad (x, t) \in \Omega \times (-T, T). \tag{3-2}$$

Let $\mathcal{O}$ *be a nonempty open subset of* Ω *and assume that* $v(x, t) = 0$ *in* $\mathcal{O} \times (-T, T)$. *Then* $v(x, 0) \equiv 0$ *in* $\mathcal{O}_T = \{x_0 \in \Omega, d(x_0, \mathcal{O}) < T\}$.

As consequences:

(a) *If* $T = +\infty$, *then* $v \equiv 0$ *everywhere.*

(b) *If* $v \equiv 0$ *in* $\mathcal{O} \times (-T, T)$ *and* $\overline{\mathcal{O}}_T = \Omega$, *then* $v \equiv 0$ *everywhere.*

Proof. Since Ω is assumed to be connected, both consequences are obvious from the first statement.

Let x_0 be given such that $d(x_0, \mathcal{O}) < T$. There is a point $x_* \in \mathcal{O}$ linked to x_0 by a smooth curve of length $l < T$ that stays away from the boundary. We introduce a sequence of balls $B(x_0, r), \ldots, B(x_K, r)$ with $r \in (0, T/K)$, $x_{k-1} \in B(x_k, r)$ and $x_K = x_*$, such that $B(x_k, r)$ stays away from the boundary and is

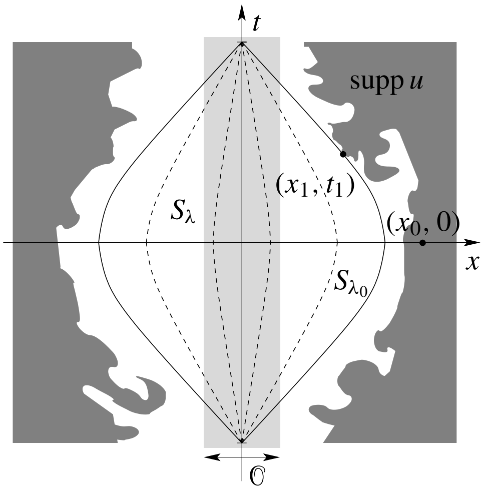

Figure 1. The proof of Corollary 3.2.

small enough such that it is diffeomorphic to an open set of $\mathbb{R}^3$ via the exponential map. Note that such a sequence of balls exists because the smooth curve linking x_0 to x_K is compact and of length smaller than T. We also notice that it is sufficient to prove Corollary 3.2 in each ball $B(x_k, r)$. Indeed, this would enable us to apply Corollary 3.2 in $B(x_K, r) \times (-T, T)$ to obtain that v vanishes in a neighborhood of x_{K-1} for $t \in (-T + r, T - r)$ and then to apply it recursively in $B(x_{K-1}, r) \times (-T + r, T - r)$, …, $B(x_1, r) \times (-T + (K - 1)r, T - (K - 1)r)$ to obtain that $v(x_0, 0) = 0$.

From now on, we assume that $x_0 \in B(x_*, r)$ and that v vanishes in a neighborhood $\mathbb{O}$ of x_* for $t \in (-r, r)$. Since $d(x_0, x_*) < r$, we can introduce a nonnegative function $h \in \mathscr{C}^\infty([-r, r], \mathbb{R})$ such that $h(0) > d(x_0, x_*)$, $h(\pm r) = 0$ and $|h'(t)| < 1$ for all $t \in [-r, r]$. We set $\mathcal{U} = B(x_*, r) \times (-r, r)$ and for any $\lambda \in [0, 1]$, we define

$$\varphi_\lambda(x, t) = d(x, x_*)^2 - \lambda h(t)^2.$$

Since r is assumed to be smaller than the radius of injectivity of the exponential map, φ_λ is a smooth well-defined function. We prove Corollary 3.2 by contradiction. Assume that $v(x_0, 0) \neq 0$. We denote by V_λ the volume $\{(x, t) \in \mathcal{U}, \varphi_\lambda(x, t) \leq 0\}$. We notice that $V_{\lambda_1} \subset V_{\lambda_2}$ if $\lambda_1 < \lambda_2$, that for small λ, V_λ is included in $\mathbb{O} \times (-r, r)$ where v vanishes, and that V_1 contains $(x_0, 0)$ where v does not vanish. Thus

$$\lambda_0 = \sup\{\lambda \in [0, 1] : v(x, t) = 0 \text{ for all} (x, t) \in V_\lambda\}$$

is well defined and belongs to $(0, 1)$. For t close to $-r$ or r, $h(t)$ is small and the section $\{x, (x, t) \in V_{\lambda_0}\}$ of V_{λ_0} is contained in $\mathbb{O}$ where v vanishes. Therefore, by compactness, the hypersurface $S_{\lambda_0} = \partial V_{\lambda_0}$ must touch the support of v at some point $(x_1, t_1) \in \mathcal{U}$ (see Figure 1).

In local coordinates, Δ can be written as $\operatorname{div}(A(x)\nabla.) + c(x) \cdot \nabla$. Moreover,

$$\langle \nabla\varphi_\lambda | A \nabla\varphi_\lambda \rangle = |\nabla_g d(\cdot, x_*)|_g^2 = 1,$$

where the index g means that the gradient and norm are taken according to the metric. Therefore, the hypersurface S_{λ_0} is noncharacteristic at (x_1, t_1) in the sense of hypothesis (iii) of Theorem 3.1 since $|\partial_t \varphi_\lambda(x, t)| = |\lambda h'(t_1)| < 1$. Thus, we can apply Theorem 3.1 with $\varphi = \varphi_{\lambda_0}$ at the point (x_1, t_1), mapping everything in the three-dimensional Euclidean frame via the exponential chart. We get that v must vanish in a neighborhood of (x_1, t_1). This is obviously a contradiction since (x_1, t_1) has been taken in the support of v. □

4. Asymptotic compactness

As soon as t is positive, a solution $u(t)$ of a parabolic PDE becomes smooth and stays in a compact set. The smoothing effect in finite time of course fails for the damped wave equations. However, these PDEs admit in some sense a smoothing effect in infinite time. This effect is called *asymptotic compactness* if one is interested in extracting asymptotic subsequences as in Proposition 4.3, or *asymptotic smoothness* if one uses the regularity of globally bounded solutions as in Proposition 4.4. For the reader interested in these notions, we refer to [Hale 1988]. The proof of this asymptotic smoothing effect is based on the variation of constant formula $U(t) = e^{At}U_0 + \int_0^t e^{A(t-s)}F(U(s))ds$ and two properties:

- The exponential decay of the linear group (Proposition 2.3), which implies that the linear part $e^{At}U_0$ asymptotically disappears.
- The regularity of the nonlinearity F implying the compactness of the nonlinear term $\int_0^t e^{A(t-s)}F(U(s))ds$ (Corollary 4.2 below). Note that the subcriticality of f is the key point of this property and that our arguments cannot be extended as they stand to the critical case $p = 5$.

The purpose of this section is to prove some compactness and regularity results about undamped solutions as (1-8). Note that these results could also have been obtained with a more "control theoretic" proof (see Section 8 for a sketch of the alternative proof) based on propagation results or observability estimates. Here, we have chosen to give a different one using asymptotic regularization, which is more common in dynamical systems. The spirit of the proof remains quite similar: we prove that the nonlinearity is more regular than it seems *a priori* and use some properties of the damped linear equation.

Regularity of the nonlinearity. Since f is subcritical, it is shown in [Dehman et al. 2003] that the nonlinear term of (1-1) yields a gain of smoothness.

Theorem 4.1 (Dehman, Lebeau and Zuazua [2003]). *Let $\chi \in \mathscr{C}_0^\infty(\mathbb{R}^3, \mathbb{R})$, $R > 0$ and $T > 0$. Let $s \in [0, 1)$ and let $\varepsilon = \min(1-s, (5-p)/2, (17-3p)/14) > 0$ with p and f as in* (1-2). *There exist (q, r) satisfying* (2-2) *and $C > 0$ such that the following property holds: If $v \in L^\infty([0, T], H^{1+s}(\mathbb{R}^3))$ is a function with finite Strichartz norms $\|v\|_{L^q([0,T],L^r(\mathbb{R}^3))} \le R$, then $\chi(x)f(v) \in L^1([0, T], H^{s+\varepsilon}(\mathbb{R}^3))$ and moreover*

$$\|\chi(x)f(v)\|_{L^1([0,T],H^{s+\varepsilon}(\mathbb{R}^3))} \le C\|v\|_{L^\infty([0,T],H^{1+s}(\mathbb{R}^3))}.$$

The constant C depends only on χ, s, T, (q, r), R and the constant in estimate (1-2).

Theorem 4.1 is a copy of Theorem 8 of [Dehman et al. 2003], except for two points.

First, we would like to apply the result to a solution v of the damped wave equation on a manifold possibly with boundaries, where not all Strichartz exponents are available. This leads to the constraint $q \geq \frac{7}{2}$ for the Strichartz exponents (q, r) of (2-2) (see Theorem 2.2). In the proof of Theorem 8 of [ibid.], the useful Strichartz estimate corresponds to $r = 3(p-1)/(1-\varepsilon)$ and $q = 2(p-1)/(p-3+2\varepsilon)$ and it is required that $q \geq p-1$, which yields $\varepsilon \leq (5-p)/2$. In this paper, we require also that $q \geq 7/2$, which yields in addition $\varepsilon \leq (17-3p)/14$. Notice that $p < 5$ and thus both bounds are positive.

The second difference is that, in [ibid.] the function f is assumed to be of class $\mathscr{C}^3$ and to satisfy

$$|f''(u)| \leq C(1+|u|)^{p-2} \quad \text{and} \quad |f^{(3)}(u)| \leq C(1+|u|)^{p-3} \tag{4-1}$$

in addition of (1-2). Since Theorem 4.1 concerns the $L^1(H^{s'})$ norm of $\chi(x)f(v)$ for $s' = s+\varepsilon \leq 1$, we can omit assumption (4-1). Actually, we make the assumption $\varepsilon \leq 1-s$ which is not present in [ibid.] and a careful study of their proof shows that (1-2) is not necessary under that assumption.

Indeed, let $\tilde{f}(u) = \text{th}^3(u)|u|^p$. The function $\tilde{f}$ is of class $\mathscr{C}^3$ and satisfies (1-2) and (4-1). Hence Theorem 8 of [ibid.] can be applied to $\tilde{f}$ and we can bound the $L^1(H^{s'})$ norm of $\tilde{f}$ as in Theorem 4.1. On the other hand, we notice that

$$|\tilde{f}(u)| \underset{\pm\infty}{\sim} |u|^p, \quad \tilde{f}'(u) \underset{\pm\infty}{\sim} p|u|^{p-1}, \quad \tilde{f}'(u) \geq 0.$$

Then, since f satisfies (1-2), there exists $C > 0$ such that $|f(u)| \leq C(1+|\tilde{f}(u)|)$ and $|f'(u)| \leq C(1+\tilde{f}'(u))$. Thus, if we assume that $v > u$ to fix the notations,

$$\begin{aligned} |f(v)-f(u)| &\leq (v-u)\int_0^1 |f'(u+\tau(v-u))|\,d\tau \\ &\leq C(v-u) + C(v-u)\int_0^1 \tilde{f}'(u+\tau(v-u))\,d\tau \\ &\leq C(v-u) + C(\tilde{f}(v)-\tilde{f}(u)) \leq C|v-u| + C|\tilde{f}(v)-\tilde{f}(u)|. \end{aligned}$$

For $0 < s < 1$, using the above inequalities and the definition of the $H^{s'}$ norm as

$$\|\chi f(u)\|_{H^{s'}}^2 = \|\chi f(u)\|_{L^2}^2 + \iint_{\mathbb{R}^6} \frac{|\chi(x)f(u(x)) - \chi(y)f(u(y))|^2}{|x-y|^{2s'}}dxdy,$$

we obtain

$$\|\chi f(u)\|_{L^1(H^{s'})} \leq C\|u\|_{L^\infty(H^1)} + C\|\tilde{\chi}\tilde{f}(u)\|_{L^1(H^{s'})},$$

where $\tilde{\chi}$ is another cut-off function with larger support. Hence for $0 < s < 1$, the conclusion of Theorem 4.1 holds not only for $\tilde{f}$ but also for f. If $s' = 1$, we just apply the chain rule and the proof is easier.

Note that the above arguments show that the constant C depends on f through estimate (1-2) only. Notice in addition that since f is only $\mathscr{C}^1$, we cannot expect $\chi f(v)$ to be more regular than H^1 and that is why we also assume $\varepsilon \leq 1-s$.

In this paper, we use a generalization of Theorem 8 of [Dehman et al. 2003] for noncompact manifolds with boundaries.

Corollary 4.2. *Let* $R > 0$ *and* $T > 0$. *Let* $s \in [0, 1)$ *and let* $\varepsilon = \min(1-s, (5-p)/2, (17-3p)/14) > 0$ *with* p *as in* (1-2). *There exist* (q, r) *satisfying* (2-2) *and* $C > 0$ *such that the following property holds*: *If* $v \in L^\infty([0, T], H^{1+s}(\Omega) \cap H^1_0(\Omega))$ *is a function with finite Strichartz norms* $\|v\|_{L^q([0,T],L^r(\Omega))} \leq R$, *then* $f(v) \in L^1([0, T], H_0^{s+\varepsilon}(\Omega))$ *and moreover*

$$\|f(v)\|_{L^1([0,T],H_0^{s+\varepsilon}(\Omega))} \leq C\|v\|_{L^\infty([0,T],H^{s+1}(\Omega)\cap H_0^1(\Omega))}.$$

The constant C *depends only on* Ω, (q, r), R *and the constant in estimate* (1-2).

Proof. Since we assumed that Ω has a bounded geometry in the sense that Ω is either compact or a compact perturbation of a manifold with periodic metric, Ω can be covered by a set of $\mathscr{C}^\infty$ charts $\alpha_i : U_i \to \alpha_i(U_i) \subset \mathbb{R}^3$ such that $\alpha_i(U_i)$ is equal either to $B(0, 1)$ or to $B_+(0, 1) = \{x \in B(0, 1), x_1 > 0\}$ and such that, for any $s \geq 0$ the norm of a function $u \in H^s(\Omega)$ is equivalent to the norm

$$\left(\sum_{i\in\mathbb{N}} \|u \circ \alpha_i^{-1}\|^2_{H^s(\alpha_i(U_i))}\right)^{1/2}.$$

Moreover, the Strichartz norm $L^q([0, T], L^r(\alpha_i(U_i))$ of $v \circ \alpha_i^{-1}$ is uniformly controlled from above by the Strichartz norm $L^q([0, T], L^r(U_i))$ of v, which is bounded by R.

Therefore, it is sufficient to prove that Corollary 4.2 holds for Ω being either $B(0, 1)$ or $B_+(0, 1)$. Say that $\Omega = B_+(0, 1)$, the case $\Omega = B(0, 1)$ being simpler. To apply Theorem 4.1, we extend v in a neighborhood of $B_+(0, 1)$ as follows. For $x \in B_+(0, 2)$, we use the radial coordinates $x = (r, \sigma)$ and we set

$$\tilde{v}(x) = \tilde{v}(r, \sigma) = 5v(1-r, \sigma) - 20v(1-r/2, \sigma) + 16v(1-r/4, \sigma).$$

Then, for $x = (x_1, x_2, x_3) \in B_-(0, 2)$, we set

$$\tilde{v}(x) = 5v(-x_1, x_2, x_3) - 20v(-x_1/2, x_2, x_3) + 16v(-x_1/4, x_2, x_3).$$

Notice that $\tilde{v}$ is an extension of v in $B(0, 2)$, which preserves the $\mathscr{C}^2$ regularity, and that the H^s norm for $s \leq 2$ as well as the Strichartz norms of $\tilde{v}$ are controlled by the corresponding norms of v. Let $\chi \in \mathscr{C}_0^\infty(\mathbb{R}^3)$ be a cut-off function such that $\chi \equiv 1$ in $B_+(0, 1)$ and $\chi \equiv 0$ outside $B(0, 2)$. Applying Theorem 4.1 to $\chi(x) f(\chi(x)\tilde{v})$ yields a control of $\|f(v)\|_{L^1([0,T],H^{s+\varepsilon}(B_+(0,1)))}$ by $\|v\|_{L^\infty([0,T],H^{s+1}(\Omega))}$. Finally, notice that $f(0) = 0$ and thus the Dirichlet boundary condition on v naturally implies the one on $f(v)$. □

Asymptotic compactness and regularization effect. As explained in the beginning of this section, using the Duhamel formula $U(t) = e^{At}U_0 + \int_0^t e^{A(t-s)}F(U(s))ds$ and Corollary 4.2, we obtain two propositions related to the asymptotic smoothing effect of the damped wave equations.

Proposition 4.3. *Let* $f \in \mathscr{C}^1(\mathbb{R})$ *satisfy* (1-2), *let* (u_0^n, u_1^n) *be a sequence of initial data which is bounded in* $X = H_0^1(\Omega) \times L^2(\Omega)$ *and let* (u_n) *be the corresponding solutions of the damped wave equation* (1-1). *Let* $(t_n) \in \mathbb{R}$ *be a sequence of times such that* $t_n \to +\infty$ *when* n *goes to* $+\infty$.

Then there exist subsequences $(u_{\psi(n)})$ *and* $(t_{\psi(n)})$ *and a global solution* u_∞ *of* (1-1) *such that*

$$(u_{\psi(n)}, \partial_t u_{\psi(n)})(t_{\psi(n)} + .) \to (u_\infty, \partial_t u_\infty)(\cdot) \quad \text{in } \mathscr{C}^0([-T, T], X) \quad \text{for all } T > 0.$$

Proof. We use the notations of Section 2. Due to the equivalence between the norm of X and the energy given by (2-1) and the fact that the energy is decreasing in time, we know that $U_n(t)$ is uniformly bounded in X with respect to n and $t \geq 0$. So, up to taking a subsequence, it weakly converges to a limit $U_\infty(0)$ which gives a global solution U_∞. We notice that, due to the continuity of the Cauchy problem with respect to the initial data stated in Theorem 2.2, it is sufficient to show that $U_{\psi(n)}(t_{\psi(n)}) \to U_\infty(0)$ for some subsequence $\psi(n)$. We have

$$\begin{aligned} U_n(t_n) &= e^{At_n}U_n(0) + \int_0^{t_n} e^{sA}F(U_n(t_n-s))\,ds \\ &= e^{At_n}U_n(0) + \sum_{k=0}^{\lfloor t_n\rfloor -1} e^{kA}\int_0^1 e^{sA}F(U_n(t_n-k-s))\,ds + \int_{\lfloor t_n\rfloor}^{t_n} e^{sA}F(U_n(t_n-s))\,ds \\ &= e^{At_n}U_n(0) + \sum_{k=0}^{\lfloor t_n\rfloor -1} e^{kA}I_{k,n} + I_n. \end{aligned} \tag{4-2}$$

Theorem 2.2 shows that the Strichartz norms $\|u_n(t_n-k-\,.\,)\|_{L^q([0,1],L^r(\Omega))}$ are uniformly bounded since the energy of U_n is uniformly bounded. Therefore Corollary 4.2 and Proposition 2.3 show that the terms $I_{n,k} = \int_0^1 e^{sA}F(U_n(t_n-k-s))ds$, as well as I_n, are bounded by some constant M in $H^{1+\varepsilon}(\Omega)\times H^\varepsilon(\Omega)$ uniformly in n and k. Using Proposition 2.3 again and summing up, we get that the last terms of (4-2) are bounded in $H^{1+\varepsilon}(\Omega)\times H^\varepsilon(\Omega)$ uniformly in n by

$$\left\| \sum_{k=0}^{\lfloor t_n\rfloor -1} e^{kA}I_{k,n} + I_n \right\|_{X^\varepsilon} \leq \sum_{k=0}^{\lfloor t_n\rfloor -1} Ce^{-\lambda k}M + M \leq M\left(1+\frac{C}{1-e^{-\lambda}}\right).$$

Moreover, Proposition 2.3 shows that $e^{At_n}U_n(0)$ goes to zero in X when n goes to $+\infty$. Therefore, by a diagonal extraction argument and the Rellich theorem, we can extract a subsequence $U_{\psi(n)}(t_{\psi(n)})$ that converges to $U_\infty(0)$ in $H_0^1(B)\times L^2(B)$ for all bounded set B of Ω.

To finish the proof of Proposition 4.3, we have to show that this convergence holds in fact in X and not only locally. Let $\eta > 0$ be given. Let $T > 0$ and let $\widetilde{U}_n$ be the solution of (1-1) with $\widetilde{U}_n(0) = U_n(t_n - T)$ and with γ being replaced by $\tilde{\gamma}$, where $\tilde{\gamma}(x) \equiv \gamma(x)$ for large x and $\tilde{\gamma} \geq \alpha > 0$ everywhere. By Proposition 2.6, $\|\widetilde{U}_n(T)\|_X \leq \eta$ if T is chosen sufficiently large and if n is large enough so that $t_n - T > 0$. Since the information propagates at finite speed in the wave equation, $U_n(t_n) \equiv \widetilde{U}_n(T)$ outside a large enough bounded set and thus $U_{\psi(n)}(t_{\psi(n)})$ has a X norm smaller than η outside this bounded set. On the other hand, we can assume that the norm of $U_\infty(0)$ is also smaller than η outside the bounded set. Then, choosing n large enough, $\|U_{\psi(n)}(t_{\psi(n)}) - U_\infty(0)\|_X$ becomes smaller than 3η. $\square$

The trajectories U^∞ appearing in Proposition 4.3 are trajectories which are bounded in X for all times $t \in \mathbb{R}$. The following result shows that these special trajectories are more regular than the usual trajectories of the damped wave equation.

Proposition 4.4. *Let $f \in \mathscr{C}^1(\mathbb{R})$ satisfying* (1-2) *and let $E_0 \geq 0$. There exists a constant M such that if u is a solution of* (1-1) *that exists for all times $t \in \mathbb{R}$ and satisfies $\sup_{t\in\mathbb{R}} E(u(t)) \leq E_0$, then*

$t \mapsto U(t) = (u(t), \partial_t u(t))$ is continuous from $\mathbb{R}$ into $D(A)$ and

$$\sup_{t\in\mathbb{R}} \|(u(t), \partial_t u(t))\|_{D(A)} \le M.$$

In addition, M depends only on E_0 and the constants in (1-2).

Proof. We use a bootstrap argument. For any $t \in \mathbb{R}$ and $n \in \mathbb{N}$,

$$U(t) = e^{nA}U(t-n) + \sum_{k=0}^{n-1} e^{kA} \int_0^1 e^{sA} F(U(t-k-s))\,ds.$$

Using Proposition 2.3, when n goes to $+\infty$, we get

$$U(t) = \sum_{k=0}^{+\infty} e^{kA} \int_0^1 e^{sA} F(U(t-k-s))\,ds. \tag{4-3}$$

Moreover, arguing exactly as in the proof of Proposition 4.3, we show that Proposition 2.3 and Corollary 4.2 imply that (4-3) also holds in X^{ε}. Hence, $U(t)$ is uniformly bounded in X^{ε}. Then, using again Proposition 2.3 and Corollary 4.2, (4-3) also holds in $X^{2\varepsilon}$, and so on. Repeating the arguments and noting that, until the last step, ε only depends on p, we obtain that $U(t)$ is uniformly bounded in $X^1 = D(A)$.

Since the constant C of Corollary 4.2 only depends on f through estimate (1-2), the same holds for the bound M here. ☐

Proposition 4.5. *The Sobolev embedding $H^2(\Omega) \hookrightarrow \mathscr{C}^0(\Omega)$ holds and there exists a constant $\mathscr{K}$ such that*

$$\sup_{x\in\Omega} |u(x)| \le \mathscr{K}\|u\|_{H^2} \quad \textit{for all } u \in H^2(\Omega).$$

In particular, the solution u in the statement of Proposition 4.4 belongs to $\mathscr{C}^0(\overline{\Omega}\times\mathbb{R}, \mathbb{R})$ and

$$\sup_{(x,t)\in\overline{\Omega}\times\mathbb{R}} |u(x,t)| \le \mathscr{K}M.$$

Proof. Proposition 4.5 follows directly from the fact that Ω has a bounded geometry and from the classical Sobolev embedding $H^2 \hookrightarrow \mathscr{C}^0$ in the ball $B(0, 1)$ of $\mathbb{R}^3$. ☐

5. Smoothness and uniqueness of nondissipative complete solutions

In this section, we consider only a nondissipative complete solution, that is, a solution u^* existing for all times $t \in \mathbb{R}$ for which the energy E is constant. In other words, $u^*(t)$ solves

$$\begin{cases} \partial_{tt}^2 u^* = \Delta u^* - \beta u^* - f(u^*) & (x,t) \in \Omega\times\mathbb{R}, \\ u^*(x,t) = 0 & (x,t) \in \partial\Omega\times\mathbb{R}, \\ \partial_t u^*(x,t) = 0 & (x,t) \in \operatorname{supp}\gamma\times\mathbb{R}. \end{cases} \tag{5-1}$$

Since the energy E is not dissipated by $u^*(t)$, we can write $E(u^*)$ instead of $E(u^*(t))$. Yet, an interesting fact that will be used several times in the sequel is that such u^* is, at the same time, solution of both damped and undamped equations.

The purpose of this section is:

- First show that u^* is analytic in time and smooth in space. The central argument is to use a theorem of J. K. Hale and G. Raugel [2003].
- Then use the unique continuation result of L. Robbiano and C. Zuily stated in Corollary 3.2 to show that u^* is necessarily an equilibrium point of (1-1).
- Finally show that the assumption $sf(s) \geq 0$ implies that $u^* \equiv 0$.

We point out that the first two steps are valid and very helpful in a more general framework than the one of our paper.

Smoothness and partial analyticity of u*. First we recall here the result of Section 2.2 of [Hale and Raugel 2003], adapting the statement to suit our notations:

Theorem 5.1. *Let Y be a Banach space. Let $P_n \in \mathscr{L}(Y)$ be a sequence of continuous linear maps and let $Q_n = \mathrm{Id} - P_n$. Let $A : D(A) \to Y$ be the generator of a continuous semigroup e^{tA} and let $G \in \mathscr{C}^1(Y)$. We assume that V is a complete mild solution in Y of*

$$\partial_t V(t) = AV(t) + G(V(t)) \quad \textit{for all } t \in \mathbb{R}.$$

We further assume that:

(i) *$\{V(t), t \in \mathbb{R}\}$ is contained in a compact set K of Y.*

(ii) *For any $y \in Y$, $P_n y$ converges to y when n goes $+\infty$ and (P_n) and (Q_n) are sequences of $\mathscr{L}(Y)$ bounded by K_0.*

(iii) *The operator A splits as $A = A_1 + B_1$, where B_1 is bounded and A_1 commutes with P_n.*

(iv) *There exist M and $\lambda > 0$ such that $\|e^{At}\|_{\mathscr{L}(Y)} \leq Me^{-\lambda t}$ for all $t \geq 0$.*

(v) *G is analytic in the ball $B_Y(0, r)$, where r is such that $r \geq 4K_0 \sup_{t\in\mathbb{R}} \|V(t)\|_Y$. More precisely, there exists $\rho > 0$ such that G can be extended to an holomorphic function of $B_Y(0, r) + iB_Y(0, \rho)$.*

(vi) *$\{DG(V(t))V_2 \,|t \in \mathbb{R}, \|V_2\|_Y \leq 1\}$ is a relatively compact set of Y.*

Then the solution $V(t)$ is analytic from $t \in \mathbb{R}$ into Y.

More precisely, Theorem 5.1 is Theorem 2.20 (which relates to Theorem 2.12) of [Hale and Raugel 2003] applied with hypotheses (H3mod) and (H5).

Proposition 4.4 shows that u^* is continuous in both space and time variables. We apply Theorem 5.1 to show that because f is analytic, u^* is also analytic with respect to the time.

Proposition 5.2. *Let $f \in \mathscr{C}^1(\mathbb{R})$ satisfying* (1-2) *and let $E_0 \geq 0$. Let $\mathscr{K}$ and M be the constants given by Propositions 4.4 and 4.5. Assume that f is analytic in $[-4\mathscr{K}M, 4\mathscr{K}M]$. Then for any nondissipative complete solution $u^*(t)$ solving* (5-1) *and satisfying $E(u^*) \leq E_0$, $t \mapsto u^*(\cdot, t)$ is analytic from $\mathbb{R}$ into X^α with $\alpha \in (1/2, 1)$. In particular, for all $x \in \Omega$, $u^*(x, t)$ is analytic with respect to the time.*

Proof. Theorem 5.1 uses strongly some compactness properties. Therefore, we need to truncate our solution to apply the theorem on a bounded domain (of course, this is not necessary and easier if Ω is already bounded).

Let $\chi \in \mathscr{C}_0^\infty(\overline{\Omega})$ be such that $\partial\chi/\partial\nu = 0$ on $\partial\Omega$, $\chi \equiv 1$ in $\{x \in \Omega, \gamma(x) = 0\}$ and $\operatorname{supp}\chi$ is included in a smooth bounded subdomain $\mathcal{O}$ of Ω. Since Proposition 4.4 shows that $u^* \in \mathscr{C}^0(\mathbb{R}, D(A))$ and since u^* is constant with respect to the time in $\operatorname{supp}\gamma$, $(1-\chi)u^*$ is obviously analytic from $\mathbb{R}$ into $D(A)$. It remains to obtain the analyticity of χu^*.

In this proof, the damping γ needs to be more regular than just $L^\infty(\Omega)$. We replace γ by a damping $\tilde{\gamma} \in C^\infty(\Omega)$, which has the same geometrical properties (GCC) and (1-3) and which vanishes where γ does. Notice that $\gamma\partial_t u^* \equiv 0 \equiv \tilde{\gamma}\partial_t u^*$, therefore replacing γ by $\tilde{\gamma}$ has no consequences here.

Let $v = \chi u^*$, we have

$$\begin{cases} \partial_{tt}^2 v + \tilde{\gamma}(x)\partial_t v = \Delta v - \beta v + g(x, v) & (x, t) \in \mathcal{O} \times \mathbb{R}_+, \\ v(x, t) = 0 & (x, t) \in \partial\mathcal{O} \times \mathbb{R}_+, \end{cases} \tag{5-2}$$

with $g(x, v) = -\chi(x) f(v + (1-\chi)u^*(x)) - 2(\nabla\chi\, \nabla u^*)(x) - (u^*\Delta\chi)(x)$. We apply Theorem 5.1 with the following setting: Let $Y = X^\alpha = H^{1+\alpha}(\mathcal{O}) \cap H_0^1(\mathcal{O}) \times H_0^\alpha(\mathcal{O})$ with $\alpha \in (1/2, 1)$. Let $V = (v, \partial_t v)$ and let $G(v) = (0, g(., v))$. We set

$$A = A_1 + B_1 = \begin{pmatrix} 0 & \mathrm{Id} \\ \Delta - \beta & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & -\tilde{\gamma} \end{pmatrix}.$$

Let $(\lambda_k)_{k\geq 1}$ be the negative eigenvalues of the Laplacian operator on $\mathcal{O}$ with Dirichlet boundary conditions and let (φ_k) be corresponding eigenfunctions. We set P_n to be the canonical projections of X on the subspace generated by $((\varphi_k, 0))_{k=1,\dots,n}$ and $((0, \varphi_k))_{k=1,\dots,n}$.

To finish the proof of Proposition 5.2, we only have to check that the hypotheses of Theorem 5.1 hold.

The trajectory V is compact since we know by Proposition 4.4 that it is bounded in X^1, which gives (i).

Hypothesis (ii) and (iii) hold with $K_0 = 1$ by construction of P_n and because B_1 is bounded in Y since $\tilde{\gamma}$ belongs to $C^\infty(\Omega)$. Hypothesis (iv) follows from Proposition 2.3.

We recall that $u^*(x, \cdot)$ is constant outside $\chi^{-1}(1)$ and belongs locally to $H^{1+\alpha}$ since $u^* \in D(A)$. Therefore, the terms $(1-\chi)u^*(x)$, $\nabla\chi\, \nabla u^*$ and $u^*\Delta\chi$ appearing in the definition of g are in H^1. Moreover, they satisfy Dirichlet boundary condition on $\partial\Omega$ since $u^* \equiv 0$ and $\partial_\nu\chi \equiv 0$ there. Of course, they also satisfy Dirichlet boundary condition on the other parts of $\partial\mathcal{O}$ since $\chi \equiv 0$ outside $\mathcal{O}$. Notice that $\alpha > 1/2$ and thus $H^{1+\alpha}(\mathcal{O}) \cap H_0^1(\mathcal{O})$ is an algebra included in $\mathscr{C}^0$. Therefore (1-2) shows that G is of class $\mathscr{C}^1$ in the bounded sets of Y. Since $u \in [-4\mathcal{H}M, 4\mathcal{H}M] \mapsto f(u) \in \mathbb{R}$ is analytic, it can be extended to a holomorphic function in $[-4\mathcal{H}M, 4\mathcal{H}M] + i[-\rho, \rho]$ for small $\rho > 0$. Using again the embedding $H^{1+\alpha}(\mathcal{O}) \hookrightarrow \mathscr{C}^0(\mathcal{O})$ and the definitions of $\mathcal{H}$ and M, we deduce that (v) holds.

Finally, for $V_2 = (v_2, \partial_t v_2)$ with $\|V_2\|_Y \leq 1$, $DG(V(t))V_2 = (0, -\chi(x) f'(v(t) + (1-\chi)u^*(x))v_2)$ is relatively compact in Y since $v(t)$ is bounded in $H^2 \cap H_0^1$ due to Proposition 4.4 and therefore $v_2 \in H^{1+\alpha} \mapsto \chi(x) f'(v(t) + (1-\chi)u^*(x))v_2 \in H^\alpha$ is a compact map. This yields (vi). □

Once the time-regularity of u^* is proved, the space-regularity follows directly.

Proposition 5.3. *Let f and u^* be as in Proposition 5.2. Then $u^* \in \mathscr{C}^\infty(\Omega \times \mathbb{R})$.*

Proof. Proposition 5.2 shows that u^* and all its time-derivatives belong to X^α with $\alpha \in (1/2, 1)$. Due to the Sobolev embeddings, this implies that any time-derivative of u^* is Hölder continuous. Writing

$$\Delta u^* = \partial_{tt}^2 u^* + \beta u^* + f(u^*) \tag{5-3}$$

and using the local elliptic regularity properties (see [Miranda 1970] and the references therein), we get that u^* is locally of class $\mathscr{C}^{2,\lambda}$ in space for some $\lambda \in (0,1)$. Thus, u^* is of class $\mathscr{C}^{2,\lambda}$ in both time and space. Then we can use a bootstrap argument in (5-3) to show that u^* is of class $\mathscr{C}^{2k,\lambda}$ for all $k \in \mathbb{N}$. ☐

Identification of u*. The smoothness and the partial analyticity of u^* shown in Propositions 5.2 and 5.3 enable us to use the unique continuation result of [Robbiano and Zuily 1998].

Proposition 5.4. *Let f and u^* be as in Proposition 5.2. Then u^* is constant in time, i.e., u^* is an equilibrium point of the damped wave equation* (1-1).

Proof. Setting $v = \partial_t u^*$, we get $\partial_{tt}^2 v = \Delta v - \beta v - f'(u^*)v$. Propositions 5.2 and 5.3 show that u^* is smooth and analytic with respect to the time and moreover $v \equiv 0$ in $\operatorname{supp} \gamma$. Thus, the unique continuation result stated in Corollary 3.2 yields $v \equiv 0$ everywhere. ☐

The sign assumption on f directly implies that 0 is the only possible equilibrium point of (1-1).

Corollary 5.5. *Let $f \in \mathscr{C}^1(\mathbb{R})$ satisfying* (1-2) *and let $E_0 \geq 0$. Let $\mathscr{K}$ and M be the constants given by Propositions 4.4 and 4.5 and assume that f is analytic in $[-4\mathscr{K}M, 4\mathscr{K}M]$. Then the unique solution u^* of* (5-1) *with $E(u^*) \leq E_0$ is $u^* \equiv 0$.*

Proof. Due to Proposition 5.4, u^* is solution of $\Delta u^* - \beta u^* = f(u^*)$. By multiplying by u^* and integrating by parts, we obtain $\int_\Omega |\nabla u^*|^2 + \beta |u^*|^2 \, dx = -\int_\Omega u^* f(u^*) \, dx$, which is nonpositive due to assumption (1-2). Since $\beta \geq 0$ is such that $\Delta - \beta$ is negative definite, this shows that $u^* \equiv 0$. ☐

6. Proof of Theorem 1.1

Due to Proposition 2.5, Theorem 1.1 directly follows from the following result.

Proposition 6.1. *Let $f \in \mathscr{C}^1(\mathbb{R})$ satisfy* (1-2) *and let $E_0 \geq 0$. Let $\mathscr{K}$ and M be the constants given by Propositions 4.4 and 4.5. Assume that f is analytic in $[-4\mathscr{K}M, 4\mathscr{K}M]$ and that γ is as in Theorem 1.1. Then there exist $T > 0$ and $C > 0$ such that any u solution of* (1-1) *with $E(u)(0) \leq E_0$ satisfies*

$$E(u)(0) \leq C \iint_{[0,T]\times\Omega} \gamma(x) \, |\partial_t u|^2 \, dt dx.$$

Proof. We argue by contradiction: we assume that there exists a sequence (u_n) of solutions of (1-1) and a sequence of times (T_n) converging to $+\infty$ such that

$$\iint_{[0,T_n]\times\Omega} \gamma(x) \, |\partial_t u_n|^2 \, dt dx \leq \frac{1}{n} E(u_n)(0) \leq \frac{1}{n} E_0. \tag{6-1}$$

Set $\alpha_n = (E(u_n)(0))^{1/2}$. Since $\alpha \in [0, \sqrt{E_0}]$, we can assume that α_n converges to a limit α when n goes to $+\infty$. We distinguish two cases: $\alpha > 0$ and $\alpha = 0$.

First case. $\alpha_n \to \alpha > 0$. Notice that, due to (2-1), $\|(u_n, \partial_t u_n)(0)\|_X$ is uniformly bounded from above and from below by positive numbers. We set $u_n^* = u_n(T_n/2 + .)$. Due to the asymptotic compactness property stated in Proposition 4.3, we can assume that u_n^* converges to a solution u^* of (1-1) in $\mathscr{C}^0([-T, T], X)$ for all time $T > 0$. We notice that

$$E(u_n(0)) \geq E(u_n^*(0)) = E(u_n(0)) - \iint_{[0,T_n/2]\times\Omega} \gamma(x)\,|\partial_t u_n|^2 \geq (1 - 1/n)E(u_n(0))$$

and thus $E(u^*(0)) = \alpha^2 > 0$. Moreover, (6-1) shows that $\gamma\partial_t u_n^*$ converges to zero in $L^2([-T, T], L^2(\Omega))$ for any $T > 0$ and thus $\partial_t u^* \equiv 0$ in $\operatorname{supp}\gamma$. In other words, u^* is a nondissipative solution of (1-1), i.e., a solution of (5-1) with $E(u^*) = \alpha^2 \leq E_0$. Corollary 5.5 shows that $u^* \equiv 0$, which contradicts the positivity of $E(u^*(0))$.

Second case. $\alpha_n \to 0$. The assumptions on f allow to write $f(s) = f'(0)s + R(s)$ with

$$|R(s)| \leq C(|s|^2 + |s|^p) \quad \text{and} \quad |R'(s)| \leq C(|s| + |s|^{p-1}). \tag{6-2}$$

Let us make the change of unknown $w_n = u_n/\alpha_n$. Then w_n solves

$$\Box w_n + \gamma(x)\partial_t w_n + (\beta + f'(0))w_n + \frac{1}{\alpha_n}R(\alpha_n w_n) = 0 \tag{6-3}$$

and

$$\iint_{[0,T_n]\times\Omega} \gamma(x)\,|\partial_t w_n|^2\,dtdx \leq \frac{1}{n}. \tag{6-4}$$

Set $W_n = (w_n, \partial_t w_n)$. Due to the equivalence between norm and energy given by (2-1), the scaling $w_n = u_n/\alpha_n$ implies that $\|(w_n(0), \partial_t w_n(0))\|_X$ is uniformly bounded from above and from below by positive numbers. Moreover, (6-1) implies

$$\|W_n(t)\|_X = \frac{\|(U_n(t))\|_X}{\alpha_n} \geq C\frac{E(u_n(t))^{1/2}}{\alpha_n} \geq C\frac{(E(u_n)(0) - \alpha_n^2/n)^{1/2}}{\alpha_n} \geq \frac{C}{2} > 0 \tag{6-5}$$

for any $t \in [0, T_n]$ and n large enough.

We set $f_n = 1/\alpha_n R(u_n)$ and $F_n = (0, f_n)$. The stability estimate of Theorem 2.2 implies that $\|u_n\|_{L^q([k,k+1],L^r)} \leq C\alpha_n$ uniformly for $n, k \in \mathbb{N}$. In particular, combined with (6-2), this gives

$$\|f_n\|_{L^1([k,k+1],L^2)} = \left\|\frac{1}{\alpha_n}R(\alpha_n w_n)\right\|_{L^1([k,k+1],L^2)} \leq C(\alpha_n + \alpha_n^{p-1}).$$

We can argue as in Proposition 4.3 and write

$$\begin{aligned} W_n(T_n) = e^{\widetilde{A}T_n}W_n(0) + \sum_{k=0}^{\lfloor T_n\rfloor - 1} e^{\widetilde{A}(T_n-k)} \int_0^1 e^{-\widetilde{A}s} F_n(k+s)\,ds \\ + e^{\widetilde{A}(T_n-\lfloor T_n\rfloor)} \int_0^{T_n-\lfloor T_n\rfloor} e^{-\widetilde{A}s} F_n(\lfloor T_n\rfloor + s)\,ds, \end{aligned} \tag{6-6}$$

where $\widetilde{A}$ is the modified damped wave operator

$$\widetilde{A} = \begin{pmatrix} 0 & \mathrm{Id} \\ \Delta - \beta - f'(0) & -\gamma \end{pmatrix}.$$

Notice that $e^{\widetilde{A}t}$ decays exponentially, like e^{At} in Proposition 2.3, since (1-2) implies $f'(0) \geq 0$. By summing up as in Proposition 4.3, we get

$$\|W_n(T_n)\|_X \leq Ce^{-\lambda T_n} + C(\alpha_n + \alpha_n^{p-1}),$$

which goes to zero, in a contradiction with (6-5). □

As a direct consequence of Proposition 6.1, we obtain a unique continuation property for nonlinear wave equations. Notice that the time of observation T required for the unique continuation is not explicit. Thus, this result is not so convenient as a unique continuation property. But it may be useful for other nonlinear stabilization problems as $\Box u + \gamma(x)g(\partial_t u) + f(u) = 0$.

Corollary 6.2. *Let $f \in \mathscr{C}^1(\mathbb{R})$ satisfy* (1-2) *and let $E_0 \geq 0$. Assume that f is analytic in $\mathbb{R}$ and that ω is an open subset of Ω satisfying* (*GCC*). *Then there exist $T > 0$ such that the only solution u of*

$$\begin{cases} \Box u + \beta u + f(u) = 0 & \text{on } [-T, T] \times \Omega, \\ \partial_t u \equiv 0 & \text{on } [-T, T] \times \omega, \end{cases} \tag{6-7}$$

with $E(u)(0) \leq E_0$ is $u \equiv 0$.

Proof. Corollary 6.2 is a straightforward consequence of Proposition 6.1 since we can easily construct a smooth damping γ supported in ω and such that $\operatorname{supp} \gamma$ satisfies (GCC). We only have to remark that a solution u of (6-7) is also solution of (1-1). □

7. Proof of Theorem 1.2

Before starting the proof of Theorem 1.2 itself, we prove that $\mathfrak{C}^1(\mathbb{R})$ is a Baire space, that is, that any countable intersection of open dense sets is dense. This legitimizes the genericity in $\mathfrak{C}^1(\mathbb{R})$ as a good notion of large subsets of $\mathfrak{C}^1(\mathbb{R})$. We recall that $\mathfrak{C}^1(\mathbb{R})$ is defined by (1-6) and endowed by the Whitney topology, the open sets of which are generated by the neighborhoods $\mathcal{N}_{f,\delta}$ defined by (1-7).

Proposition 7.1. *The space $\mathfrak{C}^1(\mathbb{R})$ endowed with the Whitney topology is a Baire space.*

Proof. The set $\mathfrak{C}^1(\mathbb{R})$ is not an open set of $\mathscr{C}^1(\mathbb{R})$, and neither a submanifold. It is a closed subset of $\mathscr{C}^1(\mathbb{R})$, but $\mathscr{C}^1(\mathbb{R})$ endowed with the Whitney topology is not a completely metrizable space, since it is not even metrizable (the neighborhoods of a function f are not generated by a countable subset of them). Therefore, we have to go back to the basic proof of Baire property as in [Golubitsky and Guillemin 1973].

Let $\mathcal{U}$ be an open set of $\mathfrak{C}^1(\mathbb{R})$ and let $(\mathcal{O}_n)_{n\in\mathbb{N}}$ be a sequence of open dense sets of $\mathfrak{C}^1(\mathbb{R})$. By density, there exists a function $f_0 \in \mathfrak{C}^1(\mathbb{R})$ in $\mathcal{U} \cap \mathcal{O}_0$ and by openness, there exists a positive continuous function δ_0 such that the neighborhood $\mathcal{N}_{f_0,\delta_0}$ is contained in $\mathcal{U} \cap \mathcal{O}_0$. By choosing δ_0 small enough, one can also assume that $\mathcal{N}_{f_0,2\delta_0} \subset \mathcal{U} \cap \mathcal{O}_0$ and that $\sup_{u\in\mathbb{R}} |\delta_0(u)| \leq 1/2^0$. By recursion, one constructs similar balls $\mathcal{N}_{f_n,\delta_n} \subset \mathcal{N}_{f_{n-1},\delta_{n-1}} \subset \mathcal{N}_{f_0,\delta_0}$ such that $\mathcal{N}_{f_n,2\delta_n} \subset \mathcal{U} \cap \mathcal{O}_n$ and that $\sup_{u\in\mathbb{R}} |\delta_n(u)| \leq 1/2^n$. Since $\mathscr{C}^1([-m, m], \mathbb{R})$ endowed with the uniform convergence topology is a complete metric space, the sequence (f_i) converges to a function $f \in \mathscr{C}^1(\mathbb{R}, \mathbb{R})$ uniformly in any compact set of $\mathbb{R}$. By construction, the limit f satisfies

$$\max\bigl(|f(u) - f_n(u)|, |f'(u) - f_n'(u)|\bigr) \leq \delta_n(u) < 2\delta_n(u) \quad \text{for all } n \in \mathbb{N} \text{ and all } u \in \mathbb{R}, \tag{7-1}$$

as well as $f(0)=0$ and $uf(u)\geq 0$ since any f_n satisfies (1-2). Moreover, there exist $C>0$ and $p\in[1,5)$ such that f_0 satisfies

$$|f_0(u)|\leq C(1+|u|)^p \quad \text{and} \quad |f_0'(u)|\leq C(1+|u|)^{p-1}. \tag{7-2}$$

Since $\max(|f(u)-f_0(u)|,|f'(u)-f_0'(u)|)\leq \delta_0(u)\leq 1$, f also satisfies (7-2) with a constant $C'=C+1$. Therefore, f satisfies (1-2) and thus belongs to $\mathfrak{C}^1(\mathbb{R})$. In addition, f satisfying (7-1) and $\mathcal{N}_{f_n,2\delta_n}$ being contained in $\mathcal{U}\cap\mathcal{O}_n$, we get $f\in\mathcal{U}\cap\mathcal{O}_n$ for all n. This shows that $\cap_{n\in\mathbb{N}}\mathcal{O}_n$ intersects any open set $\mathcal{U}$ and therefore is dense in $\mathfrak{C}^1(\mathbb{R})$. ☐

Proof of Theorem 1.2. We denote by $\mathfrak{G}_n$ the set of functions $f\in\mathfrak{C}^1(\mathbb{R})$ such that the exponential decay property (ED) holds for $E_0=n$. Obviously, $\mathfrak{G}=\cap_{n\in\mathbb{N}}\mathfrak{G}_n$ and hence it is sufficient to prove that $\mathfrak{G}_n$ is an open dense subset of $\mathfrak{C}^1(\mathbb{R})$. We sketch here the main arguments to prove this last property.

$\mathfrak{G}_n$ is a dense subset. Let $\mathcal{N}$ be a neighborhood of $f_0\in\mathfrak{C}^1(\mathbb{R})$. Up to choosing $\mathcal{N}$ smaller, we can assume that the constant in (1-2) is independent of $f\in\mathcal{N}$. Due to Propositions 4.4 and 4.5, there exist constants $\mathcal{K}$ and M such that, for all $f\in\mathcal{N}$, all the global nondissipative trajectories u of (1-1) with $E(u)\leq n$ are such that $\|u\|_{L^\infty(\Omega\times\mathbb{R})}\leq \mathcal{K}M$. We claim that we can choose $f\in\mathcal{N}$ as close to f_0 as wanted such that f is analytic on $[-4\mathcal{K}M,4\mathcal{K}M]$ and still satisfies (1-2). Then Proposition 6.1 shows that f satisfies (ED) with $E_0=n$, i.e., that $f\in\mathfrak{G}_n$.

To obtain this suitable function f, we proceed as follows. First, we set $a=4\mathcal{K}M$ and notice that it is sufficient to explain how we construct f in $[-a,a]$. Indeed, one can easily extend a perturbation f of f_0 in $[-a,a]$ satisfying $f(s)s\geq 0$ to a perturbation $\tilde{f}$ of f_0 in $\mathbb{R}$, equal to f_0 outside of $[-a-1,a+1]$ and such that $f(s)s\geq 0$ in $[-a-1,a+1]$. We construct f in $[-a,a]$ as follows. Since $f_0(s)s\geq 0$, we have that $f_0'(0)\geq 0$. We perturb f_0 to f_1 such that $f_1(0)=0$, $f_1'(s)\geq\varepsilon>0$ in a small interval $[-\eta,\eta]$ and $sf_1(s)\geq 2\varepsilon$ in $[-a,-\eta]\cup[\eta,a]$, where ε could be chosen as small as needed. Then we perturb f_1 to obtain a function f_2 which is analytic in $[-a,a]$ and satisfies $f_2'(s)>0$ in $[-\eta,\eta]$, $sf_2(s)\geq\varepsilon$ in $[-a,-\eta]\cup[\eta,a]$ and $|f_2(0)|<\varepsilon/a$. Finally, we set $f(s)=f_2(s)-f_2(0)$ and check that f is analytic and satisfies $sf(s)\geq 0$ in $[-a,a]$. Moreover, up to choosing ε very small, f is as close to f_0 as wanted.

$\mathfrak{G}_n$ is an open subset. Let $f_0\in\mathfrak{G}_n$. Proposition 2.3 shows the existence of a constant C and a time T such that for all solution u of (1-1),

$$E(u(0))\leq E_0 \quad\Longrightarrow\quad E(u(0))\leq C\int_0^T\int_\Omega \gamma(x)|\partial_t u(x,t)|^2dxdt. \tag{7-3}$$

The continuity of the trajectories in X with respect to $f\in\mathfrak{C}^1(\mathbb{R})$ is not difficult to obtain: using the strong control of f given by Whitney topology, the arguments are the same as the ones of the proof of the continuity with respect to the initial data, stated in Theorem 2.2. Thus, (7-3) holds also for any f in a neighborhood $\mathcal{N}$ of f_0, replacing the constant C by a larger one. Therefore, Proposition 2.3 shows that $\mathcal{N}\subset\mathfrak{G}_n$ and hence that $\mathfrak{G}_n$ is open. ☐

8. A proof of compactness and regularity with the usual arguments of control theory

In this section, we give an alternative proof of the compactness and regularity properties of Propositions 4.3 and 4.4. We only give its outline since it is redundant in light of the previous results of the article. Moreover, it is quite similar to the arguments of [Dehman et al. 2003]. Yet the arguments of this section are interesting because they do not require any asymptotic arguments and they show a regularization effect through an observability estimate with a finite time T, which can be explicit. However, for the moment, it seems impossible to obtain an analytic regularity similar to Proposition 5.2 with these kind of arguments.

Instead of using a Duhamel formula with an infinite interval of time $(-\infty, t)$ as in (4-3), the main idea is to use as a black box an observability estimate for T large enough, T being the time of geometric control condition,

$$\|U_0\|^2_{X^s} \leq C\|Be^{tA}U_0\|^2_{L^2([0,T],X^s)}, \tag{8-1}$$

where

$$A = \begin{pmatrix} 0 & \mathrm{Id} \\ \Delta-\beta & -\gamma \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 \\ 0 & -\gamma \end{pmatrix}.$$

The first aim is to prove that a solution of (5-1), globally bounded in energy, is also globally bounded in X^s for $s \in [0, 1]$. We proceed step by step. First, let us show that it is bounded in X^ε.

- We fix $T >$ large enough to get the observability estimate (8-1). By the existence theory on each $[t_0, t_0+T]$, $u_{|[t_0,t_0+T]}$ is bounded in Strichartz norms, uniformly for $t_0 \in \mathbb{R}$. Since the nonlinearity is subcritical, Corollary 4.2 gives that $f(u)$ is globally bounded in $L^1([t_0, t_0+T], H^{1+\varepsilon})$.

- We decompose the solution into its linear and nonlinear part by the Duhamel formula,

$$U(t) = e^{A(t-t_0)}U(t_0) + \int_{t_0}^{t} e^{A(t_0-\tau)} f(U(\tau))\,d\tau = U_{lin} + U_{Nlin}.$$

Since $f(u)$ is bounded in $L^1([t_0, t_0+T], H^{1+\varepsilon})$, U_{Nlin} is uniformly bounded in $C([t_0, t_0+T], X^\varepsilon)$.

- We will now use the linear observability estimate (8-1) with $s = \varepsilon$, applying it to U_{lin}:

$$\|U(t_0)\|^2_{X^\varepsilon} = \|U_{lin}(t_0)\|^2_{X^\varepsilon} \leq C \int_{t_0}^{t_0+T} \|\gamma(x)\partial_t u_{lin}\|^2_{H^\varepsilon}. \tag{8-2}$$

Then, using the triangular inequality, we get

$$\begin{aligned}\int_{t_0}^{t_0+T} \|\gamma(x)\partial_t u_{lin}\|^2_{H^\varepsilon} &\leq 2\int_{t_0}^{t_0+T} \|\gamma(x)\partial_t u\|^2_{H^\varepsilon} + 2\int_{t_0}^{t_0+T} \|\gamma(x)\partial_t u_{Nlin}\|^2_{H^\varepsilon} \\ &\leq 2\int_{t_0}^{t_0+T} \|\gamma(x)\partial_t u_{Nlin}\|^2_{H^\varepsilon} \leq C,\end{aligned}$$

where we have used that $\partial_t u \equiv 0$ on ω and that U_{Nlin} is bounded in $C([t_0, t_0+T], X^\varepsilon)$. Combining this with (8-2) for any $t_0 \in \mathbb{R}$, we obtain that U is uniformly bounded in X^ε on $\mathbb{R}$.

Repeating the arguments, we show that u is bounded in $X^{2\varepsilon}$, $X^{3\varepsilon}$ and so on, until X^1. Similar ideas allow us to prove a theorem of propagation of compactness in finite time, replacing the asymptotic compactness property of Proposition 4.3.

As said above, an advantage of this method, compared to the one used in Propositions 4.3 and 4.4, is that it allows us to propagate the regularity or the compactness on some finite interval of fixed length. Yet, it seems that such propagation results are not available in the analytic setting. Indeed, it seems that, for nonlinear equations, the propagation of analytic regularity or of nullity in finite time is much harder to prove. We can for instance refer to the weaker (with respect to the geometry) result of Alinhac and Métivier [1984] or the negative result of Métivier [1993].

9. Applications

Control of the nonlinear wave equation. In this subsection, we give a short proof of Theorem 1.3, which states the global controllability of the nonlinear wave equation. The first step consists in a local control theorem.

Theorem 9.1 (local control). *Let ω satisfying the geometric control condition for a time T. Then there exists δ such that for any (u_0, u_1) in $H_0^1(\Omega) \times L^2(\Omega)$ with*

$$\|(u_0, u_1)\|_{H_0^1 \times L^2} \leq \delta$$

there exists $g \in L^\infty([0, T], L^2)$ supported in $[0, T] \times \omega$ such that the unique strong solution of

$$\begin{cases} \Box u + \beta u + f(u) = g & \text{on} \quad [0, T] \times \Omega, \\ (u(0), \partial_t u(0)) = (u_0, u_1), \end{cases}$$

satisfies $(u(T), \partial_t u(T)) = (0, 0)$.

Proof. The proof is exactly the same as that of Theorem 3 of [Dehman et al. 2003] or Theorem 3.2 of [Laurent 2011]. The main argument consists in seeing the problem as a perturbation of the linear controllability, which is known to be true in our setting. □

Now, as is very classical, we can combine the local controllability with our stabilization theorem to get global controllability.

Sketch of the proof of Theorem 1.3. In a first step, we choose as a control $g = -\gamma(x)\partial_t \tilde{u}$, where $\tilde{u}$ is solution of (1-1) with initial data (u_0, u_1). By uniqueness of solutions, we have $u = \tilde{u}$. Therefore, thanks to Theorem 1.1, for a large time T_1, only depending on R_0, we have $\|(u(T_1), \partial_t u(T_1))\|_{H^1 \times L^2} \leq \delta$. Then Theorem 9.1 allows to find a control that brings $(u(T_1), \partial_t u(T_1))$ to 0. In other words, we have found a control g supported in ω that brings (u_0, u_1) to 0. We obtain the same result for $(\tilde{u}_0, \tilde{u}_1)$ and conclude, by reversibility of the equation, that we can also bring 0 to $(\tilde{u}_0, \tilde{u}_1)$. □

Existence of a compact global attractor. In this subsection, we give the modification of the proofs of this paper necessary to get Theorem 1.4 about the existence of a global attractor.

The energy associated to (1-9) in $X = H_0^1(\Omega) \times L^2(\Omega)$ is given by

$$E(u, v) = \int_\Omega \tfrac{1}{2}(|\nabla u|^2 + |v|^2) + V(x, u)\, dx,$$

where $V(x, u) = \int_0^u f(x, \xi)\, d\xi$.

The existence of a compact global attractor for (1-9) is well known for the Sobolev subcritical case $p < 3$. The first proofs in this case go back to 1985 [Hale 1988; Haraux 1985b]; see [Raugel 2002] for other references. The case $p = 3$ as been studied in [Babin and Vishik 1992; Arrieta et al. 1992]. For $p \in (3, 5)$, Kapitanski [1995] proved the existence of a compact global attractor for (1-9) if Ω is a compact manifold without boundary and if $\gamma(x) = \gamma$ is a constant damping. Using the same arguments as in the proof of our main result, we can partially deal with the case $p \in (3, 5)$ with a localized damping $\gamma(x)$ and with unbounded manifold with boundaries.

Assume that f satisfies the assumption of Theorem 1.4. Then the arguments of this paper show the following properties.

(i) *The positive trajectories of bounded sets are bounded.* Indeed, (1-12) implies that for $x \notin B(x_0, R)$, we have $V(x, u) = \int_0^u f(x, \xi) d\xi \geq 0$. Moreover, for $x \in B(x_0, R)$, $V(x, \cdot)$ is nonincreasing on $(-\infty, -R)$ and nondecreasing on (R, ∞). Thus, $V(x, u)$ is bounded from below for $x \in B(x_0, R)$ and

$$E(u, v) \geq \tfrac{1}{2}\|(u, v)\|_X^2 + \operatorname{vol}(B(x_0, R)) \inf V \quad \text{for all } (u, v) \in X$$

The Sobolev embeddings $H^1(\Omega) \hookrightarrow L^{p+1}(\Omega)$ show that the bounded sets of X have a bounded energy. Since the energy E is nonincreasing along the trajectories of (1-9), we get that the trajectory of a bounded set is bounded.

(ii) *The dynamical system is asymptotically smooth.* The asymptotic compactness exactly corresponds to the statement of Proposition 4.3. Let us briefly explain why it can be extended to the case where f depends on x. The key point is the extension of Corollary 4.2. First notice that we assumed $f(x, 0) = 0$ on $\partial\Omega$ in order to guarantee the Dirichlet boundary condition for $f(x, u)$ if $u \in H_0^1(\Omega)$. Then it is not difficult to see that the discussion following Theorem 4.1 can be extended to the case f depending on x by using estimates (1-10). Corollary 4.2 follows then, except for a small change: since it is possible that $f(x, 0) \neq 0$ for some $x \in \Omega$, the conclusion of Corollary 4.2 should be replaced by

$$\|f(x, v)\|_{L^1([0,T], H_0^{s+\varepsilon}(\Omega))} \leq C\big(1 + \|v\|_{L^\infty([0,T], H^{s+1}(\Omega) \cap H_0^1(\Omega))}\big).$$

Then the proof of Proposition 4.3 is based on Corollary 4.2, the boundedness of the positive trajectories of bounded sets (both could be extended to the case where f depends on x as noticed above) and an application of Proposition 2.6 outside of a large ball. We conclude by noticing that, for x large, $f(x, u)u \geq 0$ and $\gamma(x) \geq \alpha > 0$ and thus Proposition 2.6 can still be applied exactly as in the proof of Proposition 4.3.

(iii) *The dynamical system generated by* (1-9) *is gradient.* That is, that the energy E is nonincreasing in time and is constant on a trajectory u if and only if u is an equilibrium point of (1-9). This last property is shown in Proposition 5.4 for f independent of x but can be easily generalized for $f = f(x, u)$. Notice that the proof of this property is the one where the analyticity of f is required since the unique continuation property of Section 3 is used. Finally, we remark that the gradient structure of (1-9) is interesting from the dynamical point of view since it implies that any trajectory $u(t)$ converges when t goes to $+\infty$ to the set of equilibrium points.

(iv) *The set of equilibrium points is bounded.* The argument is similar to the one of Corollary 5.5: if e is an equilibrium point of (1-9) then (1-12) implies that

$$\int \tfrac{1}{2}|\nabla e|^2+\beta|e|^2=-\int_\Omega f(x,e)e\,dx\le -\operatorname{vol}(B(x_0,R))\inf\{f(x,u)u \mid (x,u)\in\overline{\Omega}\times\mathbb{R}\},$$

where we have bounded $f(x,u)u$ from below exactly as we have done for $V(x,u)$ in (i).

It is well known (see [Hale 1988] or Theorem 4.6 of [Raugel 2002]) that properties (i)–(iv) yield the existence of a compact global attractor. Hence, we obtain the conclusion of Theorem 1.4.

Acknowledgements

Both authors benefited from the fruitful atmosphere of the conference Partial differential equations, optimal design and numerics in Benasque. We also would like to thank Mathieu Léautaud for his remarks about the optimality of hypothesis (GCC) in the nonlinear context.

References

[Alinhac and Métivier 1984] S. Alinhac and G. Métivier, “Propagation de l’analyticité des solutions de systèmes hyperboliques non-linéaires”, *Invent. Math.* **75**:2 (1984), 189–204. MR 86f:35010 Zbl 0545.35063

[Aloui and Khenissi 2002] L. Aloui and M. Khenissi, “Stabilisation pour l’équation des ondes dans un domaine extérieur”, *Rev. Mat. Iberoamericana* **18**:1 (2002), 1–16. MR 2003h:35150 Zbl 1016.35044

[Aloui et al. 2011] L. Aloui, S. Ibrahim, and K. Nakanishi, “Exponential energy decay for damped Klein–Gordon equation with nonlinearities of arbitrary growth”, *Comm. Partial Differential Equations* **36**:5 (2011), 797–818. MR 2012f:35348 Zbl 1243.35122

[Arrieta et al. 1992] J. Arrieta, A. N. Carvalho, and J. K. Hale, “A damped hyperbolic equation with critical exponent”, *Comm. Partial Differential Equations* **17**:5-6 (1992), 841–866. MR 93f:35145 Zbl 0815.35067

[Babin and Vishik 1992] A. V. Babin and M. I. Vishik, *Attractors of evolution equations*, Studies in Mathematics and its Applications **25**, North-Holland, Amsterdam, 1992. MR 93d:58090 Zbl 0778.58002

[Bardos et al. 1988] C. Bardos, G. Lebeau, and J. Rauch, “Un exemple d’utilisation des notions de propagation pour le contrôle et la stabilisation de problèmes hyperboliques”, pp. 11–31 in *Nonlinear hyperbolic equations in applied sciences* (Turin, Italy, 1987), Universitá e Politecnico, Turin, 1988. MR 90h:35030 Zbl 0673.93037

[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 94b:93067 Zbl 0786.93009

[Blair et al. 2009] M. D. Blair, H. F. Smith, and C. D. Sogge, “Strichartz estimates for the wave equation on manifolds with boundary”, *Ann. Inst. H. Poincaré Anal. Non Linéaire* **26**:5 (2009), 1817–1829. MR 2010k:35500 Zbl 1198.58012

[Burq 1997a] N. Burq, “Contrôlabilité exacte des ondes dans des ouverts peu réguliers”, *Asymptot. Anal.* **14**:2 (1997), 157–191. MR 98e:93005 Zbl 0892.93009

[Burq 1997b] N. Burq, “Mesures semi-classiques et mesures de défaut”, pp. 167–195 in *Séminaire Bourbaki* 1996/1997 (Exposé 826), Astérisque **245**, Société Mathématique de France, Paris, 1997. MR 99f:35011 Zbl 0954.35102

[Burq and Gérard 1997] N. Burq and P. Gérard, “Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes”, *C. R. Acad. Sci. Paris Sér. I Math.* **325**:7 (1997), 749–752. MR 98j:93052 Zbl 0906.93008

[Burq et al. 2008] N. Burq, G. Lebeau, and F. Planchon, “Global existence for energy critical waves in 3-D domains”, *J. Amer. Math. Soc.* **21**:3 (2008), 831–845. MR 2009f:35225 Zbl 1204.35119

[Dehman 2001] B. Dehman, “Stabilisation pour l’équation des ondes semi-linéaire”, *Asymptot. Anal.* **27**:2 (2001), 171–181. MR 2002g:35026 Zbl 1007.35005

[Dehman and Gérard 2002] B. Dehman and P. Gérard, “Stabilization for the nonlinear Klein Gordon equation with critical exponent”, preprint, 2002.

[Dehman et al. 2003] B. Dehman, G. Lebeau, and E. Zuazua, “Stabilization and control for the subcritical semilinear wave equation”, *Ann. Sci. École Norm. Sup.* (4) **36**:4 (2003), 525–551. MR 2004i:35223 Zbl 1036.35033

[Dehman et al. 2006] B. Dehman, P. Gérard, and G. Lebeau, “Stabilization and control for the nonlinear Schrödinger equation on a compact surface”, *Math. Z.* **254**:4 (2006), 729–749. MR 2007g:93007 Zbl 1127.93015

[Gérard 1991] P. Gérard, “Microlocal defect measures”, *Comm. Partial Differential Equations* **16**:11 (1991), 1761–1794. MR 92k:35027 Zbl 0770.35001

[Gérard 1996] P. Gérard, “Oscillations and concentration effects in semilinear dispersive wave equations”, *J. Funct. Anal.* **141**:1 (1996), 60–98. MR 97k:35171 Zbl 0868.35075

[Gérard and Leichtnam 1993] P. Gérard and É. Leichtnam, “Ergodic properties of eigenfunctions for the Dirichlet problem”, *Duke Math. J.* **71**:2 (1993), 559–607. MR 94i:35146 Zbl 0788.35103

[Ginibre and Velo 1985] J. Ginibre and G. Velo, “The global Cauchy problem for the nonlinear Klein–Gordon equation”, *Math. Z.* **189**:4 (1985), 487–505. MR 86f:35149 Zbl 0549.35108

[Ginibre and Velo 1989] J. Ginibre and G. Velo, “The global Cauchy problem for the nonlinear Klein–Gordon equation. II”, *Ann. Inst. H. Poincaré Anal. Non Linéaire* **6**:1 (1989), 15–35. MR 90b:35153 Zbl 0694.35153

[Golubitsky and Guillemin 1973] M. Golubitsky and V. Guillemin, *Stable mappings and their singularities*, Graduate Texts in Mathematics **14**, Springer, New York, 1973. MR 49 #6269 Zbl 0294.58004

[Hale 1988] J. K. Hale, *Asymptotic behavior of dissipative systems*, Mathematical Surveys and Monographs **25**, American Mathematical Society, Providence, RI, 1988. MR 89g:58059 Zbl 0642.58013

[Hale and Raugel 2003] J. K. Hale and G. Raugel, “Regularity, determining modes and Galerkin methods”, *J. Math. Pures Appl.* (9) **82**:9 (2003), 1075–1136. MR 2004k:37166 Zbl 1043.35048

[Haraux 1985a] A. Haraux, “Stabilization of trajectories for some weakly damped hyperbolic equations”, *J. Differential Equations* **59**:2 (1985), 145–154. MR 87b:35014 Zbl 0535.35006

[Haraux 1985b] A. Haraux, “Two remarks on hyperbolic dissipative problems”, pp. 6, 161–179 in *Nonlinear partial differential equations and their applications* (Paris, 1983–1984), edited by H. Brézis and J.-L. Lions, Res. Notes in Math. **122**, Pitman, Boston, MA, 1985. MR 879461 Zbl 0579.35057

[Hörmander 1997] L. Hörmander, “On the uniqueness of the Cauchy problem under partial analyticity assumptions”, pp. 179–219 in *Geometrical optics and related topics* (Cortona, Italy, 1996), edited by F. Colombini and N. Lerner, Progr. Nonlinear Differential Equations Appl. **32**, Birkhäuser, Boston, MA, 1997. MR 2033496 Zbl 0907.35002

[Ivanovici 2012] O. Ivanovici, “Counterexamples to the Strichartz inequalities for the wave equation in general boundary domains”, *J. Eur. Math. Soc.* (*JEMS*) **14**:5 (2012), 1357–1388. MR 2966654 Zbl 06095885

[Joly 2007] R. Joly, “New examples of damped wave equations with gradient-like structure”, *Asymptot. Anal.* **53**:4 (2007), 237–253. MR 2008k:35322 Zbl 05541561

[Kapitanski 1995] L. Kapitanski, “Minimal compact global attractor for a damped semilinear wave equation”, *Comm. Partial Differential Equations* **20**:7-8 (1995), 1303–1323. MR 96c:35120 Zbl 0829.35014

[Kapitanskiĭ 1990] L. V. Kapitanskiĭ, “Some generalizations of the Strichartz–Brenner inequality”, *Leningrad Math. J.* **1**:3 (1990), 693–726. MR 90h:46063 Zbl 0732.35118

[Khenissi 2003] M. Khenissi, “Équation des ondes amorties dans un domaine extérieur”, *Bull. Soc. Math. France* **131**:2 (2003), 211–228. MR 2005b:35171 Zbl 1024.35061

[Laurent 2010] C. Laurent, “Global controllability and stabilization for the nonlinear Schrödinger equation on some compact manifolds of dimension 3”, *SIAM J. Math. Anal.* **42**:2 (2010), 785–832. MR 2011g:93014 Zbl 1208.93035

[Laurent 2011] C. Laurent, “On stabilization and control for the critical Klein–Gordon equation on a 3-D compact manifold”, *J. Funct. Anal.* **260**:5 (2011), 1304–1368. MR 2011m:93075 Zbl 1244.35012

[Laurent et al. 2010] C. Laurent, L. Rosier, and B.-Y. Zhang, “Control and stabilization of the Korteweg-de Vries equation on a periodic domain”, *Comm. Partial Differential Equations* **35**:4 (2010), 707–744. MR 2012d:93024 Zbl 1213.93015

[Lebeau 1996] G. Lebeau, “Équation des ondes amorties”, pp. 73–109 in *Algebraic and geometric methods in mathematical physics* (Kaciveli, Albania, 1993), edited by A. Boutet de Monvel and V. Marchenko, Math. Phys. Stud. **19**, Kluwer Acad. Publ., Dordrecht, 1996. MR 97i:58173 Zbl 0863.58068

[Métivier 1993] G. Métivier, “Counterexamples to Hölmgren’s uniqueness for analytic nonlinear Cauchy problems”, *Invent. Math.* **112**:1 (1993), 217–222. MR 94b:35010 Zbl 0794.35026

[Miller 2002] L. Miller, “Escape function conditions for the observation, control, and stabilization of the wave equation”, *SIAM J. Control Optim.* **41**:5 (2002), 1554–1566. MR 2004b:93115 Zbl 1032.35117

[Miranda 1970] C. Miranda, *Partial differential equations of elliptic type*, vol. 2, Ergebnisse der Mathematik und ihrer Grenzgebiete (2), Springer, New York, 1970. MR 44 #1924 Zbl 0198.14101

[Pazy 1983] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, Applied Mathematical Sciences **44**, Springer, New York, 1983. MR 85g:47061 Zbl 0516.47023

[Rauch and Taylor 1974] J. Rauch and M. Taylor, “Exponential decay of solutions to hyperbolic equations in bounded domains”, *Indiana Univ. Math. J.* **24** (1974), 79–86. MR 50 #13906 Zbl 0281.35012

[Raugel 2002] G. Raugel, “Global attractors in partial differential equations”, pp. 885–982 in *Handbook of dynamical systems*, vol. 2, edited by B. Fiedler, North-Holland, Amsterdam, 2002. MR 2003f:37151 Zbl 1005.35001

[Robbiano and Zuily 1998] L. Robbiano and C. Zuily, “Uniqueness in the Cauchy problem for operators with partially holomorphic coefficients”, *Invent. Math.* **131**:3 (1998), 493–539. MR 99e:35004 Zbl 0909.35004

[Strichartz 1977] R. S. Strichartz, “Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations”, *Duke Math. J.* **44**:3 (1977), 705–714. MR 58 #23577 Zbl 0372.35001

[Tartar 1990] L. Tartar, “H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations”, *Proc. Roy. Soc. Edinburgh Sect. A* **115**:3-4 (1990), 193–230. MR 91h:35042 Zbl 0774.35008

[Tartar 2007] L. Tartar, *An introduction to Sobolev spaces and interpolation spaces*, Lecture Notes of the Unione Matematica Italiana **3**, Springer, Berlin, 2007. MR 2008g:46055 Zbl 1126.46001

[Tataru 1995] D. Tataru, “Unique continuation for solutions to PDE’s; between Hörmander’s theorem and Holmgren’s theorem”, *Comm. Partial Differential Equations* **20**:5-6 (1995), 855–884. MR 96e:35019 Zbl 0846.35021

[Tataru 1999] D. Tataru, “Unique continuation for operators with partially analytic coefficients”, *J. Math. Pures Appl.* (9) **78**:5 (1999), 505–521. MR 2000e:35005 Zbl 0936.35038

[Zuazua 1990] E. Zuazua, “Exponential decay for the semilinear wave equation with locally distributed damping”, *Comm. Partial Differential Equations* **15**:2 (1990), 205–235. MR 91b:35076 Zbl 0716.35010

[Zuazua 1991] E. Zuazua, “Exponential decay for the semilinear wave equation with localized damping in unbounded domains”, *J. Math. Pures Appl.* (9) **70**:4 (1991), 513–529. MR 93a:35111 Zbl 0765.35010

Received 9 May 2012. Revised 17 Aug 2012. Accepted 27 Sep 2012.

ROMAIN JOLY: `romain.joly@ujf-grenoble.fr`
Institut Fourier – UMR 5582, CNRS/Université de Grenoble, 100, rue des Maths, BP 74, 38402 St.-Martin-d’Hères, France

CAMILLE LAURENT: `laurent@ann.jussieu.fr`
Laboratoire Jacques-Louis Lions – UMR 7598, CNRS/Université Pierre et Marie Curie (Univ. Paris 6), Boîte courrier 187, 75005 Paris, France

ANALYSIS AND PDE
Vol. 6, No. 5, 2013
dx.doi.org/10.2140/apde.2013.6.1121

INSTABILITY THEORY OF THE NAVIER–STOKES–POISSON EQUATIONS

JUHI JANG AND IAN TICE

The stability question of the Lane–Emden stationary gaseous star configurations is an interesting problem arising in astrophysics. We establish both linear and nonlinear dynamical instability results for the Lane–Emden solutions in the framework of the Navier–Stokes–Poisson system with adiabatic exponent $\frac{6}{5} < \gamma < \frac{4}{3}$.

1. Introduction and formulation

One of the simplest fundamental hydrodynamical models to describe the motion of self-gravitating viscous gaseous stars is the compressible Navier–Stokes–Poisson system, which can be written in Eulerian coordinates as

$$\begin{cases} \partial_t \varrho + \operatorname{div}(\varrho \boldsymbol{u}) = 0, \\ \partial_t (\varrho \boldsymbol{u}) + \operatorname{div}(\varrho \boldsymbol{u} \otimes \boldsymbol{u}) + \operatorname{div} S = -\varrho \nabla \Phi, \\ \Delta \Phi = 4\pi \varrho, \end{cases} \tag{1-1}$$

where $(\boldsymbol{x}, t) \in \mathbb{R}^3 \times \mathbb{R}^+$, $\varrho(\boldsymbol{x}, t) \geq 0$ is the density, $\boldsymbol{u}(\boldsymbol{x}, t) \in \mathbb{R}^3$ is the velocity vector field of the gas, $\Phi(\boldsymbol{x}, t) \in \mathbb{R}$ is the potential function of the self-gravitational force, and the stress tensor S is given by

$$S = P I_{3\times3} - \varepsilon\big(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^t - \tfrac{2}{3}(\operatorname{div} \boldsymbol{u}) I_{3\times3}\big) - \delta(\operatorname{div} \boldsymbol{u}) I_{3\times3}, \tag{1-2}$$

where P is the pressure of the gas, $\varepsilon > 0$ is the shear viscosity, $\delta \geq 0$ is the bulk viscosity, and $\nabla \boldsymbol{u}^t$ denotes the transpose of $\nabla \boldsymbol{u}$. We consider polytropic gases for which the equation of state is given by

$$P = P(\varrho) = K \varrho^{\gamma}, \tag{1-3}$$

where K is an entropy constant and $\gamma > 1$ is an adiabatic exponent. Values of γ have their own physical significance [Chandrasekhar 1939]; for instance, $\gamma = \frac{5}{3}$ corresponds to a monatomic gas and $\gamma = \frac{7}{5}$ to a diatomic gas, and $\gamma \to 1^+$ for heavier molecules.

In the simplest setting, which we consider, solutions to (1-1) are spherically symmetric. For $r = |\boldsymbol{x}|$, this allows us to write

$$\boldsymbol{u}(\boldsymbol{x}, t) = u(r, t) \frac{\boldsymbol{x}}{r} \quad \text{for } u : [0, \infty) \times [0, \infty) \to \mathbb{R} \tag{1-4}$$

Jang was supported by NSF grants DMS-0908007 and DMS-1212142. Tice was supported by an NSF Postdoctoral Research Fellowship.

MSC2010: 35Q30, 35R35, 76E20, 85A30.

Keywords: Navier–Stokes–Poisson, free boundary problems, gaseous stars, hydrodynamic instability, nonlinear instability.

and

$$\varrho(\boldsymbol{x},t)=\varrho(r,t). \tag{1-5}$$

The equations (1-1) then reduce to the pair

$$\partial_t\varrho+u\partial_r\varrho+\frac{\varrho}{r^2}\partial_r(r^2u)=0 \tag{1-6}$$

and

$$\varrho(\partial_t u+u\partial_r u)+\partial_r P=-\frac{4\pi\varrho}{r^2}\int_0^r\varrho(s,t)s^2\,ds+\partial_r\biggl(\frac{\frac{4}{3}\varepsilon+\delta}{r^2}\partial_r(r^2u)\biggr). \tag{1-7}$$

The integral term on the right side of (1-7) corresponds to the gravitational force. Stationary solutions $\varrho=\varrho_0(r)$ and $u=0$, which correspond to nonmoving gaseous spheres in hydrostatic equilibrium, satisfy the following equation for $P_0=K\varrho_0^\gamma$:

$$\partial_r P_0(r)+\frac{4\pi\varrho_0(r)}{r^2}\int_0^r\varrho_0(s)s^2\,ds=0. \tag{1-8}$$

This equation can be solved by transforming it into the well-known Lane–Emden equation [Chandrasekhar 1939]. The solutions to (1-8) are positive and decreasing and can be characterized by the values of γ in the following fashion [Lin 1997]: for given finite total mass $M>0$, if $\gamma\in\bigl(\frac{6}{5},2\bigr)$, there exists at least one compactly supported solution ϱ_0. For $\gamma\in\bigl(\frac{4}{3},2\bigr)$, every solution is compactly supported and unique. If $\gamma=\frac{6}{5}$, the unique solution admits an analytic expression, and it has infinite support. On the other hand, for $\gamma\in\bigl(1,\frac{6}{5}\bigr)$, there are no solutions with finite total mass.

The stability of the Lane–Emden steady star configurations has been a question of great interest, and it has been conjectured by astrophysicists that stationary solutions for $\gamma<\frac{4}{3}$ are unstable. The linear stability theory of the above stationary solutions was studied in [Lin 1997] in the inviscid case, namely the Euler–Poisson system, by studying the eigenvalue problem associated to the linearized Euler–Poisson system: any stationary solution is linearly stable when $\gamma\in\bigl(\frac{4}{3},2\bigr)$ and unstable when $\gamma\in\bigl(1,\frac{4}{3}\bigr)$. In accordance with the linear stability theory, a nonlinear stability for $\gamma>\frac{4}{3}$ was established in [Rein 2003] by using a variational approach. In the case $\gamma=\frac{4}{3}$, the analysis in [Deng et al. 2002] identified an instability in which any small perturbation can cause part of the system to go off to infinity. In [Jang 2008], a nonlinear instability of the Lane–Emden steady star for $\gamma=\frac{6}{5}$ was proved based on the bootstrap argument, as pioneered in [Guo and Strauss 1995]. The stability question for the Euler–Poisson system with $\frac{6}{5}<\gamma<\frac{4}{3}$ remains an open problem.

The same stability question can also be asked in the presence of viscosity. There have been interesting studies on the stabilization effect of viscosity in the Navier–Stokes–Poisson system for $\gamma>\frac{4}{3}$ under various assumptions [Ducomet and Zlotnik 2005; Zhang and Fang 2009]. On the other hand, to our knowledge, no rigorous stability theories are available for $\gamma<\frac{4}{3}$, the instability regime in the inviscid case. In this regime for viscous gaseous stars, a particularly interesting problem is to investigate whether or not the viscosity would dominate the gravitational force and stabilize the whole system. The purpose of this article is to establish the instability theory of the Lane–Emden steady stars whose dynamics are governed by the Navier–Stokes–Poisson system for $\frac{6}{5}<\gamma<\frac{4}{3}$.

We now formulate the problem. We begin by introducing a vacuum free boundary.

1A. ***Vacuum free boundary.*** When $\gamma > \frac{6}{5}$, letting $R > 0$ be the radius of the steady star, it is well known [Lin 1997] that

$$\varrho_0(r) \sim (R-r)^{1/(\gamma-1)} \quad \text{for } r \text{ near } R. \tag{1-9}$$

This boundary behavior near a vacuum causes a degeneracy in (1-6) and (1-7), and it is not trivial to deal with such a degeneracy even for the local-in-time existence question; we refer, for instance, to [Jang 2010; Matusu-Necasova et al. 1997; Okada and Makino 1993] and also [Jang and Masmoudi 2009; 2010] for the compressible Euler case. It turns out that in order to capture boundary behavior such as (1-9) in the dynamical setting, one has to consider a free boundary problem associated to (1-6) and (1-7) as in [Jang 2010; Matusu-Necasova et al. 1997; Okada and Makino 1993]. We are interested in the evolution of compactly supported stars with a free boundary where the star meets a vacuum. This is implemented by assuming there is a radius $R = R(t) > 0$ such that

$$\varrho(r,t) > 0 \quad \text{for } r \in [0, R(t)) \quad \text{and} \quad \varrho(R(t),t) = 0. \tag{1-10}$$

At the free boundary we impose the kinematic condition

$$\frac{d}{dt} R(t) = u(R(t),t), \tag{1-11}$$

as well as the continuity of the normal stress, $S\nu = 0$ at the surface $r = R(t)$. The latter condition reduces to

$$P - \frac{4\varepsilon}{3}\left(\partial_r u - \frac{u}{r}\right) - \delta\left(\partial_r u + \frac{2u}{r}\right) = 0 \quad \text{for } r = R(t),\ t \geq 0. \tag{1-12}$$

Note that $P(R(t),t) = K\varrho^\gamma(R(t),t) = 0$, so this can be reduced to a relationship between $\partial_r u$ and u at $r = R(t)$. Finally, in order for $\boldsymbol{u} = u(r,t)\boldsymbol{x}/r$ to be continuous, we require $u(0,t) = 0$ for $t \geq 0$.

Since the boundary $R(t)$ is free to move in time in Eulerian coordinates, it is convenient to introduce Lagrangian coordinates so that the boundary becomes fixed. Following the framework used in [Jang 2010; Matusu-Necasova et al. 1997; Okada and Makino 1993], we study our instability problem in Lagrangian mass coordinates.

1B. ***Formulation in Lagrangian mass coordinates.*** We now reformulate the problem in Lagrangian mass coordinates. We set

$$x(r,t) = \int_0^r 4\pi s^2 \varrho(s,t)\,ds = \int_{B(0,r)} \varrho(y,t)\,dy \tag{1-13}$$

for the mass contained in an Eulerian ball of radius r at time t. Note that

$$\partial_r x(r,t) = 4\pi r^2 \varrho(r,t) \tag{1-14}$$

and that

$$\partial_t x(r,t) = \int_{B(0,r)} \partial_t \varrho(y,t)\,dy = -\int_{B(0,r)} \operatorname{div}(\varrho \boldsymbol{u})\,dy = -\int_{\partial B(0,r)} \varrho \boldsymbol{u}\cdot\nu = -4\pi r^2 \varrho(r,t) u(r,t). \tag{1-15}$$

In particular, this implies that $\partial_t x(R(t), t) = 0$, which means that the total mass $M > 0$ is preserved in time. The domain of x is then $[0, M]$. Switching to Lagrangian mass coordinates $(x, t) \in [0, M] \times [0, \infty)$ and letting the unknowns be

$$\rho(x, t) = \varrho(r, t) \quad \text{and} \quad v(x, t) = u(r, t), \tag{1-16}$$

we get the equations

$$\partial_t \rho + 4\pi\rho^2 \partial_x (r^2 v) = 0 \tag{1-17}$$

and

$$\partial_t v + 4\pi r^2 \partial_x P + \frac{x}{r^2} = 16\pi^2 r^2 \partial_x \left(\left(\frac{4\varepsilon}{3} + \delta \right) \rho \partial_x (r^2 v) \right). \tag{1-18}$$

In Lagrangian coordinates, our boundary conditions reduce to

$$v(0, t) = 0, \quad \rho(M, t) = 0, \tag{1-19}$$

and

$$P - \frac{4\varepsilon}{3} \left(4\pi r^2 \rho \partial_x v - \frac{v}{r} \right) - \delta \left(4\pi r^2 \rho \partial_x v + \frac{2v}{r} \right) = 0 \quad \text{at } x = M \quad \text{for all } t \geq 0. \tag{1-20}$$

In each of these equations, we have written

$$r(x, t) = \left(\frac{3}{4\pi} \int_0^x \frac{dy}{\rho(y, t)} \right)^{1/3}, \tag{1-21}$$

which inverts (1-13) by way of integrating (1-14). A simple computation, employing (1-17), shows that $\partial_t r(x, t) = v(x, t)$.

A stationary solution $\rho = \rho_0(x)$, $v = 0$, $P_0 = K\rho_0^\gamma$ to (1-17) and (1-18) satisfies the equation

$$4\pi r_0^2(x) \partial_x P_0(x) + \frac{x}{r_0^2(x)} = 0, \tag{1-22}$$

where

$$r_0(x) = \left(\frac{3}{4\pi} \int_0^x \frac{dy}{\rho_0(y)} \right)^{1/3}. \tag{1-23}$$

This is the Lagrangian version of (1-8). We denote such a Lane–Emden solution in Lagrangian mass coordinates by ρ_0 with pressure $P_0 = K\rho_0^\gamma$. Note that $\rho_0(x) > 0$ for $x \in [0, M)$ and that ρ_0 decreases until it vanishes at $x = M$. In Lagrangian x coordinates, the boundary behavior (1-9) is expressed as

$$\rho_0(x) \sim (M - x)^{1/\gamma} \quad \text{for } x \text{ near } M, \tag{1-24}$$

which can be also seen from (1-22). In particular, when $\gamma \in \left(\frac{6}{5}, \frac{4}{3}\right)$, this implies that $\dfrac{1}{\rho_0(x)}$ is integrable, so that $R = r_0(M) < \infty$, which corresponds to a star of finite radius.

The existence and uniqueness of strong solutions to the vacuum free boundary problem of the Navier–Stokes–Poisson system (1-17) and (1-18) featuring the behavior (1-24) of Lane–Emden solutions was established in [Jang 2010] when $\delta = 2\varepsilon/3 > 0$. The same methodology can be applied to our current setting as long as $\varepsilon > 0$ and $\delta > 0$, and we will take those strong solutions for granted in proving our nonlinear instability result. A well-posedness result in our energy space can be also proved based on our new a priori energy estimates for the fully nonlinear Navier–Stokes–Poisson system, described in Section 4.

1C. *Main results.* Throughout the paper, we assume that

$$\varepsilon > 0, \quad \delta > 0, \quad K > 0, \quad \text{and} \quad \tfrac{6}{5} < \gamma < \tfrac{4}{3} \tag{1-25}$$

are all fixed. Note that although the only physical requirement on the bulk viscosity is $\delta \geq 0$, the assumption $\delta > 0$ is critical for both our linear and nonlinear analysis. We will also write $M, R > 0$ for the mass and radius of a stationary solution to (1-22).

To state the main results, we first write the system in a perturbation form. For small perturbed solutions $\sigma := \rho - \rho_0$ and v around the steady states satisfying (1-22), the Navier–Stokes–Poisson system (1-17) and (1-18) can be written as

$$\begin{gathered}\partial_t \sigma + 4\pi\rho^2 \partial_x (r^2 v) = 0, \\ \partial_t v + 4\pi r^2 \partial_x P - 4\pi r_0^2 \partial_x P_0 + \frac{x}{r^2} - \frac{x}{r_0^2} = 16\pi^2 r^2 \partial_x \left(\left(\frac{4\varepsilon}{3} + \delta \right) \rho \partial_x (r^2 v) \right),\end{gathered} \tag{1-26}$$

with boundary conditions (1-19) and (1-20).

Our first main result concerns the existence of the largest growing mode of the linearized Navier–Stokes–Poisson system around Lane–Emden solutions, which shows a linear instability in the sense of Lin's stability criteria [1997].

Theorem 1.1. *Suppose* (1-25). *There exist* $\lambda > 0$ *and* $\sigma(x)$, $v(x)$ *such that* $\sigma(x)e^{\lambda t}$ *and* $v(x)e^{\lambda t}$ *solve the linearized Navier–Stokes–Poisson system* (2-1) *and* (2-2) *with the linearized boundary conditions* (2-3) *and* (2-4). *Moreover, this growing mode yields the largest possible growth rate to the linearized system.*

Remark 1.2. The growth rate $\lambda > 0$ produced in Theorem 1.1 clearly depends on the values of the viscosity parameters ε, δ. It is natural to consider the asymptotics of λ for large and small viscosities. In Proposition 2.11 below, we show that λ converges to the largest growth rate for the inviscid problem (identified by Lin [1997]) as $(\delta, \varepsilon) \to 0$. We also show that $\lambda \to 0$ as $\delta \to \infty$, which demonstrates that viscosity delays the onset of instability, since the escape time T^ι (given below in (1-27)) is inversely proportional to λ.

The precise statement of Theorem 1.1 with the estimates is given in Theorems 2.1 and 3.2. Our second main result establishes the fully nonlinear dynamical instability of the Lane–Emden solutions to the Navier–Stokes–Poisson system. In the statement of the theorem, for any given $\iota > 0$ and $\theta > \iota$, we write

$$T^\iota := \frac{1}{\lambda} \ln \frac{\theta}{\iota}, \tag{1-27}$$

where λ is the sharp linear growth rate obtained in Theorem 1.1.

Theorem 1.3. *Suppose* (1-25). *There exist function spaces* X *and* Y *as well as constants* $\theta > 0$ *and* $C > 0$ *such that for any sufficiently small* $\iota > 0$, *there exist solutions* $(\sigma^\iota(t), v^\iota(t))$ *to* (1-26) *for* $t \in [0, T)$ *with*

$T > T^\iota$ such that

$$\left\|(\sigma^\iota(0), v^\iota(0))\right\|_Y \le C\iota, \quad \text{but} \quad \sup_{0\le t\le T^\iota} \left\|(\sigma^\iota(t), v^\iota(t))\right\|_X \ge \theta. \tag{1-28}$$

The precise statement of Theorem 1.3 is given in Theorem 5.4, and the spaces X and Y will be clarified in Sections 4 and 5.

Remark 1.4. Our results show that regardless of how large the viscosity parameters ε, δ are, and no matter how small smooth initial perturbed data are taken to be, the system remains unstable. We conclude from this that all Lane–Emden steady star configurations for $\frac{6}{5} < \gamma < \frac{4}{3}$ are unstable, regardless of viscosity.

Remark 1.5. The escape time T^ι is determined through (1-27) by the linear growth rate λ. We note that the instability occurs before the possible breakdown or any collapse of strong solutions. We also remark that the instability occurs in the X norm, which when rewritten in Eulerian coordinates, is equivalent to

$$\int_{\mathbb{R}^3} \left(\frac{1}{2}\varrho|\boldsymbol{u}|^2 + \frac{\gamma}{2}\frac{\varrho_0}{\varrho}P_0\left|\frac{\sigma}{\varrho_0}\right|^2\right) d\boldsymbol{x}, \tag{1-29}$$

which is related to the positive part of the physical energy:

$$\int_{\mathbb{R}^3} \left(\frac{1}{2}\varrho|\boldsymbol{u}|^2 + \frac{1}{\gamma-1}P\right) d\boldsymbol{x}. \tag{1-30}$$

Of course, this is not a coincidence: the Lane–Emden solutions for $\gamma < \frac{4}{3}$ do not minimize the physical energy functional,

$$\begin{aligned} &\int_0^M \left(\frac{1}{2}|v|^2 + \frac{1}{\gamma-1}\frac{P}{\rho} - \frac{x}{r}\right) dx \text{ in Lagrangian mass coordinates, or} \\ &\int_{\mathbb{R}^3} \left(\frac{1}{2}\varrho|\boldsymbol{u}|^2 + \frac{1}{\gamma-1}P\right) d\boldsymbol{x} - \frac{1}{2}\int_{\mathbb{R}^3\times\mathbb{R}^3} \frac{\varrho(\boldsymbol{x})\varrho(\boldsymbol{y})}{|\boldsymbol{x}-\boldsymbol{y}|}\, d\boldsymbol{x}\, d\boldsymbol{y} \text{ in Eulerian coordinates,} \end{aligned} \tag{1-31}$$

and thus one might expect some kind of instability. They do minimize for $\gamma > \frac{4}{3}$ (see, for instance, [Jang 2008; Rein 2003]).

The presence of viscosity and the nonlinear boundary condition (1-20) for the Navier–Stokes–Poisson system make the problem distinguishable and interesting not only from a physical point of view, but also from a mathematical point of view. What follows now are some of the main mathematical difficulties we encounter in analyzing the system, and a brief discussion of our methods for resolving them.

The proof of Theorem 1.1 is based on a variational analysis of equations obtained by linearizing (1-26). The main difficulty that arises in constructing growing-mode solutions is that, due to the viscous terms, the growth rate (eigenvalue) appears in the problem with two different homogeneities. This breaks the natural variational structure used in [Lin 1997] to construct growing modes in the inviscid case. To get around this difficulty, we employ a technique introduced in [Guo and Tice 2010]: we introduce a relaxed parameter that allows us to remove one of the eigenvalue homogeneities, study the resulting modified eigenvalue problem (which has a nice variational structure), and finally return to the original formulation through a fixed point argument. While the solutions constructed in this manner are definitely growing modes, it is not clear a priori that they grow at the largest possible rate. To verify this, we carry out a careful analysis, paying particular attention to the boundary behavior of the growing mode, which will be crucially used in the subsequent nonlinear bootstrap argument.

The proof of Theorem 1.3 is based on a bootstrap argument from linear instability to nonlinear dynamical instability. Passing from a linearized instability to nonlinear instability requires much effort in the PDE context since the spectrum of the linear part is fairly complicated and the unboundedness of the nonlinear part usually yields a loss in derivatives. In order to get around these difficulties and to find the right space Y, we employ careful nonlinear energy estimates for the whole system so that, first, the nonlinear estimates can be closed, and second, their interplay with the linear analysis can complete the argument. For this particular problem, the space Y is minimally chosen so that the viscosity disturbance near the vacuum boundary can be controlled within Y.

We note that in Lagrangian mass coordinates, the continuity equation interacts well with the viscosity term, which allows us to derive nice estimates for σ/ρ_0 and its temporal and spatial derivatives. This plays an important role in closing our nonlinear energy estimates. The main technical difficulty is to derive Proposition 5.1, a key estimate for the bootstrap argument. The idea is to find an energy $\mathscr{E}$ that satisfies an inequality of the form

$$\frac{d}{dt}\mathscr{E} \leq \eta\mathscr{E} + \text{lower derivative terms}, \tag{1-32}$$

where η is smaller than the sharp linear growth rate. However, (1-32) is too good to hold in general due to the degeneracy of vacuum boundary and the complexity of the system near Lane–Emden stars. To overcome this difficulty, we introduce a collection of energy terms: some of them satisfy (1-32) under certain conditions, which we quantify; some of them are bootstrapped energies, the estimates of which are obtained by improved weighted energy estimates that exploit the structure of the equations; and others are auxiliary energies, the estimates of which are directly obtained from the equations. The gravitational potential has a smoothing effect, behaves well with the necessary weights, and does not create further difficulty in the nonlinear estimates. In fact, Section 4 is devoted to the introduction of those energy terms and the derivation of the estimates. Combining the estimates of the various energy terms, we can complete the bootstrap argument.

Another delicate and important issue is the nonlinear boundary condition (1-20). In order to carry out higher-order energy estimates that require integration by parts, we can only employ differential operators that respect the boundary conditions, namely temporal derivatives. This forces us to carefully use the structure of the equations in order to gain bounds on spatial derivatives. A second difficulty with the boundary arises because we use Duhamel's principle to study the nonlinear problem with the linearized evolution operator. The linearized boundary condition is homogeneous, but the nonlinear boundary condition is certainly not. This forces us to introduce a corrector function that removes the boundary inhomogeneity. While the construction of this function is not particularly delicate, the regularity required to do so dictates that we close our energy estimates at a higher order than we would otherwise.

The paper proceeds as follows. The first half is devoted to the development of the linear theory and the proof of Theorem 1.1. In Section 2, we formulate a variational problem to find a growing-mode solution to the linearized Navier–Stokes–Poisson system. In Section 3, we show that our growing-mode solution grows at the largest possible rate. In the second half of the paper, we carry out our nonlinear analysis. In Section 4, we derive high-order nonlinear energy inequalities. Based on the linear growth and the nonlinear estimates, we then prove the bootstrap argument and Theorem 1.3 in Section 5.

2. Construction of a growing mode solution to the linearized equations

2A. ***Linearization around a stationary solution.*** We now linearize the equations in Lagrangian mass coordinates around the stationary solution $v=0$, $\rho=\rho_0$, $r=r_0$ (as defined by (1-23)). We will write σ for the linearized density, and (by abuse of notation) v for the linearized velocity. Then the linearized equations are given by

$$\partial_t\sigma+4\pi\rho_0^2\partial_x(r_0^2v)=0 \tag{2-1}$$

and

$$\partial_t v+4\pi r_0^2\partial_x\tilde{P}+\frac{x}{\pi r_0^5}\int_0^x\frac{\sigma(y,t)}{\rho_0^2(y)}\,dy=16\pi^2r_0^2\partial_x\biggl(\biggl(\frac{4\varepsilon}{3}+\delta\biggr)\rho_0\partial_x(r_0^2v)\biggr), \tag{2-2}$$

where we have written $\tilde{P}=\gamma K\rho_0^{\gamma-1}\sigma$. The linearized boundary conditions are

$$v(0,t)=0,\quad \sigma(M,t)=0 \tag{2-3}$$

and

$$\tilde{P}-\frac{4\varepsilon}{3}\biggl(4\pi r_0^2\rho_0\partial_x v-\frac{v}{r_0}\biggr)-\delta\biggl(4\pi r_0^2\rho_0\partial_x v+\frac{2v}{r_0}\biggr)=0\quad\text{at } x=M\quad\text{for all } t\geq 0. \tag{2-4}$$

Again, we can view (2-4) as a boundary condition only for v since $\tilde{P}=\gamma K\rho_0^{\gamma-1}\sigma=0$ at $x=M$ for each $t\geq 0$.

It will often be useful for us to analyze a variant of this system, where we analyze the unknowns σ and $w:=r_0^2v$. For these unknowns, Equations (2-1)–(2-4) become

$$\begin{aligned}&\partial_t\sigma+4\pi\rho_0^2\partial_x w=0,\\&\partial_t w+4\pi r_0^4\partial_x(\gamma K\rho_0^{\gamma-1}\sigma)-4r_0\partial_x P_0\int_0^x\frac{\sigma(y,t)}{\rho_0^2(y)}\,dy=16\pi^2r_0^4\partial_x\biggl[\biggl(\frac{4\varepsilon}{3}+\delta\biggr)\rho_0\partial_x w\biggr],\end{aligned} \tag{2-5}$$

along with the boundary conditions

$$\frac{w}{r_0^2}(0,t)=\sigma(M,t)=0\quad\text{and}\quad\frac{4\varepsilon}{3}\biggl(4\pi r_0^3\rho_0\partial_x\biggl(\frac{w}{r_0^3}\biggr)\biggr)+\delta(4\pi\rho_0\partial_x w)=0\quad\text{at } x=M. \tag{2-6}$$

Note that $w/r_0^2=u$ is well-defined at $x=0$, but one may also view the first boundary condition in (2-6) in the sense of traces or limits.

2B. ***Growing mode solution.*** We want to construct a growing mode solution to the linearized equations. We do so by looking for a solution of the form

$$\sigma(x,t)=\sigma(x)e^{\lambda t} \quad \text{and} \quad v(x,t)=v(x)e^{\lambda t} \tag{2-7}$$

for some $\lambda > 0$. If we can find such a solution, then we say the solution is a growing mode since $|e^{\lambda t}| \to \infty$ as $t \to \infty$. Plugging the ansatz (2-7) into the linearized equations (2-1)–(2-4) and eliminating the time exponentials, we arrive at a pair of equations for $\sigma(x)$ and $v(x)$:

$$\lambda\sigma + 4\pi\rho_0^2\partial_x(r_0^2 v) = 0 \tag{2-8}$$

and

$$\lambda v + 4\pi r_0^2\partial_x\tilde{P} + \frac{x}{\pi r_0^5}\int_0^x \frac{\sigma(y)}{\rho_0^2(y)}\,dy = 16\pi^2 r_0^2\partial_x\left(\left(\frac{4\varepsilon}{3}+\delta\right)\rho_0\partial_x(r_0^2 v)\right), \tag{2-9}$$

along with boundary conditions

$$v(0)=\sigma(M)=0 \quad \text{and} \quad -\frac{4\varepsilon}{3}\left(4\pi r_0^2\rho_0\partial_x v - \frac{v}{r_0}\right) - \delta\left(4\pi r_0^2\rho_0\partial_x v + \frac{2v}{r_0}\right) = 0 \quad \text{at } x = M. \tag{2-10}$$

Our main result of this section establishes the existence of such a growing mode.

Theorem 2.1. *There exist* $\lambda > 0$ *and* $\sigma, v : (0, M) \to \mathbb{R}$ *that solve* (2-8)–(2-10) *and satisfy the following.*

(1) σ *and* v *are smooth on* $(0, M)$ *and satisfy* (2-8)–(2-9) *classically for* $x \in (0, M)$.

(2) *It holds that*

$$\limsup_{x\to 0}\frac{|v(x)|}{r_0(x)} + \limsup_{x\to 0}|\sigma(x)| + \limsup_{x\to 0}|\partial_x(r_0^2 v)(x)| < \infty. \tag{2-11}$$

In particular, $v(0) = 0$.

(3) *Let* $\mathfrak{D}$ *denote the linear operator* $\mathfrak{D} = \rho_0\partial_x$. *Then* $\mathfrak{D}^k v$ *and* $\mathfrak{D}^k(\sigma/\rho_0)$ *have well-defined traces at* $x = M$ *for every integer* $k \geq 0$. *In particular*, $\sigma(M) = 0$.

(4) $\lambda > 0$ *satisfies the variational characterization*

$$\lambda\int_0^M\left(\delta\rho_0|\partial_x\theta|^2 + \frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{\theta}{r_0^3}\right)\right|^2\right)dx + \int_0^M\left(\frac{\gamma P_0\rho_0}{2}|\partial_x\theta|^2 + \frac{\partial_x P_0}{2\pi r_0^3}|\theta|^2\right)dx \geq -\lambda^2\int_0^M\frac{|\theta|^2}{16\pi^2 r_0^4}\,dx \tag{2-12}$$

for every θ *satisfying* $\sqrt{\rho_0}\partial_x\theta \in L^2((0,M))$ *and* $\theta/(r_0^2\sqrt{\rho_0}) \in L^2((0,M))$. *Note that for such* θ, *it holds that* $\theta/(r_0^3\sqrt{\rho_0}) \in L^2((0,M))$, *which means that all of the integrals in* (2-12) *are well-defined.*

(5) *It holds that*

$$\int_0^M\left(\left|\frac{\sigma}{\rho_0}\right|^2 + r_0^2\left|\partial_x\frac{\sigma}{\rho_0}\right|^2\right)dx + \int_0^M\left(\frac{|r_0^2 v|^2}{r_0^6\rho_0} + \rho_0|\partial_x(r_0^2 v)|^2 + r_0^2\left|\partial_x(\rho_0\partial_x(r_0^2 v))\right|^2\right)dx < \infty. \tag{2-13}$$

The proof of Theorem 2.1 will be completed in Section 2F. Throughout the rest of the section, we develop the tools needed in the proof. First we reformulate (2-8)–(2-10) to involve a single unknown function, ϕ. The resulting problem for ϕ does not possess a standard variational structure since λ appears both linearly and quadratically. To construct a solution using variational methods (required for proving (2-12), which is essential for the linear estimates of Section 3), we employ the technique of Guo and Tice [2010], which proceeds as follows. We modify the problem by replacing the linear appearance of λ by an arbitrary parameter $s > 0$. The resulting family (every $s > 0$) of problems is amenable to solution by the constrained minimization of an energy functional, and for a range of s we show that $\lambda = \lambda(s) > 0$. We then study the behavior of $\lambda(s)$ as a function of s and show that it is possible to find a unique fixed point such that $\lambda(s) = s > 0$. This then yields the desired solution ϕ, which in turn yields the solution to (2-8)–(2-10).

We begin by reducing to the study of a single unknown by introducing the function

$$\phi(x) := \int_0^x \frac{\sigma(y)}{\rho_0^2(y)}\,dy. \tag{2-14}$$

We may then use (2-8)–(2-10) to compute

$$v = -\frac{\lambda}{4\pi r_0^2}\phi, \quad \sigma = \rho_0^2 \partial_x \phi, \quad \text{and} \quad \partial_x \tilde{P} = \partial_x(\gamma \rho_0 P_0 \partial_x \phi), \tag{2-15}$$

where $P_0 = K\rho_0^\gamma$. Using these and replacing in (2-9), we arrive at a second-order equation for ϕ:

$$-\partial_x\left(\left(\frac{4\lambda\varepsilon}{3} + \lambda\delta + \gamma P_0\right)\rho_0 \partial_x \phi\right) + \frac{\partial_x P_0}{\pi r_0^3}\phi = -\frac{\lambda^2}{16\pi^2 r_0^4}\phi. \tag{2-16}$$

The corresponding boundary conditions are

$$\frac{\phi}{r_0^2}(0) = 0 \quad \text{and} \quad \frac{4\varepsilon}{3}\lambda\left(4\pi r_0^3 \rho_0 \partial_x\left(\frac{\phi}{r_0^3}\right)\right) + \delta\lambda(4\pi\rho_0\partial_x\phi) = 0 \quad \text{at } x = M. \tag{2-17}$$

2C. ***Modification of the problem.*** Note that Theorem 2.1 is phrased in Lagrangian mass coordinates. This is because we will use these coordinates in our nonlinear analysis later in the paper. However, constructing the solution to (2-16)–(2-17) is somewhat easier if we make a change of variables back to the Eulerian radial coordinates associated to the stationary solution. To avoid confusion with the Eulerian radial coordinate for the nonlinear problem, we will call our new variable $z = r_0(x)$, where r_0 is given by (1-23). If $x \in (0, M)$ for M the mass of the stationary star, then $z \in (0, R)$ for $R > 0$ its radius. We will write $\varrho_0(z) = \rho_0(x)$ for the stationary density, $P_0 = K\varrho^\gamma$, and $\varphi(z) = \phi(x)$ for the new unknown in z coordinates. Then

$$\partial_x = \frac{1}{4\pi z^2 \varrho_0}\partial_z. \tag{2-18}$$

In these coordinates, (2-16) becomes

$$-\partial_z\left(\left(\frac{4\lambda\varepsilon}{3} + \lambda\delta + \gamma P_0\right)\frac{\partial_z\varphi}{z^2}\right) + 4\frac{\partial_z P_0}{z^3}\varphi = -\frac{\lambda^2\varrho_0}{z^2}\varphi. \tag{2-19}$$

For the boundary condition at $z = R$, we use (2-17) to see that

$$\lambda\delta \frac{\partial_z \varphi(R)}{R^2} + \frac{4\lambda\varepsilon}{3}\left(\frac{\partial_z\varphi(R)}{R^2} - 3\frac{\varphi(R)}{R^3}\right) = 0. \tag{2-20}$$

At $z = 0$ we enforce the boundary condition $\varphi(0) = 0$. Once we have a solution in hand, we will show that, in fact, $\varphi(z)/z^2 \to 0$ as $z \to 0$, which allows us to switch back to the boundary condition $(\phi/r_0^2)(0) = 0$.

There is a difficulty in viewing (2-19)–(2-20) in a variational or Sturm–Liouville framework because of the appearance of λ with two different homogeneities. To get around this issue, we temporarily modify the problem in order to restore the variational structure. Ultimately we will undo the modification and return to the proper formulation.

Fix $s > 0$ and define

$$\tilde{\varepsilon} = s\varepsilon \quad \text{and} \quad \tilde{\delta} = s\delta. \tag{2-21}$$

Instead of (2-19), we will analyze the equation

$$-\partial_z\left(\left(\frac{4\tilde{\varepsilon}}{3} + \tilde{\delta} + \gamma P_0\right)\frac{\partial_z\varphi}{z^2}\right) + 4\frac{\partial_z P_0}{z^3}\varphi = -\frac{\lambda^2\varrho_0}{z^2}\varphi \tag{2-22}$$

for arbitrary $s > 0$. We couple this equation to the boundary conditions $\varphi(0) = 0$ and

$$\tilde{\delta}\frac{\partial_z\varphi(R)}{R^2} + \frac{4\tilde{\varepsilon}}{3}\left(\frac{\partial_z\varphi(R)}{R^2} - 3\frac{\varphi(R)}{R^3}\right) = 0. \tag{2-23}$$

Modifying the problem in this way restores the variational structure. Indeed, in (2-22) the λ^2 term can be viewed as an eigenvalue. Thinking of the principal eigenvalue λ as a function of s, that is, $\lambda = \lambda(s)$, we will show that it is possible to choose s such that $\lambda(s) > 0$ and $s = \lambda(s)$, which returns us to the original problem and yields a growing-mode solution.

2D. ***Constrained minimization.*** In order to construct solutions to (2-22)–(2-23), we will employ a constrained minimization. To begin, we define the function space on which the energy functionals will be defined. For $\tau > 0$, we define the weighted Sobolev space $H^1_\tau((0, R))$ as the completion of $\{u \in C^\infty([0, R]) \mid u(0) = 0\}$ with respect to the norm

$$\|u\|^2_{H^1_\tau} = \int_0^R \frac{|u'(z)|^2 + |u(z)|^2}{z^\tau}\,dz, \tag{2-24}$$

where $' = d/dz$. This weighted Sobolev space possesses the same sort of embedding (continuous and compact) properties as the usual space H^1. Since these results are not widely available in the literature, we record them in the following lemma.

Lemma 2.2. (1) *For $u \in H^1_\tau((0, R))$, we have the inequalities*

$$\sup_{0\le z\le R}\left|u(z)z^{-(\tau+1)/2}\right| \le \frac{1}{\sqrt{1+\tau}}\left(\int_0^R \frac{|u'(z)|^2}{z^\tau}\,dz\right)^{1/2} \tag{2-25}$$

and

$$\int_0^R \frac{|u(z)|^2}{z^{\tau+2}}\,dz \le \frac{4}{(1+\tau)^2}\int_0^R \frac{|u'(z)|^2}{z^\tau}\,dz. \tag{2-26}$$

(2) *Let* $0 \le \alpha < 1$. *We have the compact embedding* $H^1_\tau((0,R)) \Subset L^2_{\tau+1+\alpha}((0,R))$, *where the latter space is the weighted* L^2 *space with norm*

$$\|u\|^2_{L^2_{\tau+1+\alpha}} = \int_0^R \frac{|u(z)|^2}{z^{\tau+1+\alpha}}\,dz. \tag{2-27}$$

Proof. We begin with the inequalities in item (1). By approximation, we may assume that u is smooth and $u(0)=0$. Then

$$\begin{aligned}|u(z)| = |u(z)-u(0)| \le \int_0^z |u'(t)|\,dt &\le \left(\int_0^z t^\tau\,dt\right)^{1/2}\left(\int_0^z \frac{|u'(t)|^2}{t^\tau}\,dt\right)^{1/2}\\ &\le \left(\frac{z^{\tau+1}}{\tau+1}\right)^{1/2}\left(\int_0^R \frac{|u'(t)|^2}{t^\tau}\,dt\right)^{1/2},\end{aligned} \tag{2-28}$$

which yields the first inequality. To get the second, we recall an inequality due to G. H. Hardy:

$$\left(\int_0^\infty\left(\int_0^z |f(t)|\,dt\right)^p \frac{dz}{z^{b+1}}\right)^{1/p} \le \frac{p}{b}\left(\int_0^\infty |f(z)|^p z^{p-b-1}\,dz\right)^{1/p}, \tag{2-29}$$

for $1 \le p < \infty$ and $0 < b < \infty$, which follows immediately from Young's inequality on the multiplicative group $(0,\infty)$ with measure dt/t by convolving $|f(t)|t^{1-b/p}$ with $t^{-b/p}\chi_{(1,\infty)}(t)$. Then $|u(z)| \le \int_0^z |u'(t)|\,dt$ implies that

$$\int_0^R \frac{|u(z)|^2}{z^{\tau+2}}\,dz \le \int_0^R\left(\int_0^z |u'(t)|\,dt\right)^2 \frac{dz}{z^{\tau+2}}. \tag{2-30}$$

Applying Hardy's inequality to the right side with $f = u'\chi_{(0,R)}$, $b=\tau+1$, and $p=2$ yields

$$\int_0^R \frac{|u(z)|^2}{z^{\tau+2}}\,dz \le \frac{4}{(\tau+1)^2}\int_0^R \frac{|u'(z)|^2}{z^\tau}\,dz, \tag{2-31}$$

which is the desired inequality.

We now prove the compactness result. Assume that $\|u_n\|_{H^1_\tau} \le C$ for $n \in \mathbb{N}$. Fix $\kappa > 0$. We claim that there exists a subsequence $\{u_{n_i}\}$ such that

$$\sup_{i,j} \|u_{n_i} - u_{n_j}\|_{L^2_{\tau+1+\alpha}} \le \kappa. \tag{2-32}$$

To prove the claim, let $z_0 \in (0,R)$ be chosen such that

$$z_0^{1-\alpha}\frac{C^2}{(1+\tau)(1-\alpha)} \le \frac{\kappa}{2}. \tag{2-33}$$

Then since the subinterval (z_0,R) avoids the singularity of $1/z^\tau$, $u_n|_{(z_0,R)}$ is uniformly bounded in $H^1((z_0,R))$. By the compact embedding $H^1((z_0,R)) \Subset C^0((z_0,R))$, we may extract a subsequence $\{u_{n_i}\}$ that converges in $L^\infty((z_0,R))$. We are free to restrict the subsequence to large enough values of i that

$$\|u_{n_i} - u_{n_j}\|^2_{L^\infty((z_0,R))} \le \frac{\kappa {z_0}^{\tau+1+\alpha}}{2(R-z_0)} \quad \text{for all } i,j. \tag{2-34}$$

Then along this subsequence we can apply the first inequality in item (1) to get

$$\begin{aligned}\int_0^R \frac{|u_{n_i}(z)-u_{n_j}(z)|^2}{z^{\tau+1+\alpha}}\,dz &= \int_0^{z_0} \frac{|u_{n_i}(z)-u_{n_j}(z)|^2}{z^{\tau+1+\alpha}}\,dz + \int_{z_0}^R \frac{|u_{n_i}(z)-u_{n_j}(z)|^2}{z^{\tau+1+\alpha}}\,dz \\ &\le \frac{C^2}{1+\tau}\int_0^{z_0}\frac{dz}{z^{\alpha}} + \frac{R-z_0}{{z_0}^{\tau+1+\alpha}}\|u_{n_i}-u_{n_j}\|^2_{L^\infty((z_0,R))} \le \kappa, \end{aligned} \tag{2-35}$$

which proves the claim. Now we may use the claim with $\kappa = 1/k$, $k \in \mathbb{N}$ and employ a standard diagonal argument to extract a subsequence converging in $L^2_{\tau+1+\alpha}((0, R))$. □

Remark 2.3. The inequality (2-26) implies that we can take the norm on H^1_τ to be

$$\|u\|^2_{H^1_\tau} = \int_0^R \frac{|u'(z)|^2}{z^\tau}\,dz. \tag{2-36}$$

We can now define the energy functionals to use in the constrained minimization. Let

$$E(\varphi) = \int_0^R \left[(\tilde{\delta}+\gamma P_0)\frac{|\partial_z\varphi|^2}{z^2} + \frac{4\tilde{\varepsilon}}{3z^2}\left|\partial_z\varphi - 3\frac{\varphi}{z}\right|^2 + 4\frac{\partial_z P_0}{z^3}|\varphi|^2\right] dz \tag{2-37}$$

and

$$J(\varphi) = \int_0^R \frac{\varrho_0}{z^2}|\varphi|^2\,dz. \tag{2-38}$$

By (2-26) in Lemma 2.2, both E and J are well-defined on the space $H^1_2((0, R))$. Note, though, that E is not positive definite since $\partial_z P_0 < 0$. Define the set

$$\mathscr{A} := \{\varphi \in H^1_2((0, R)) \mid J(\varphi) = 1\}. \tag{2-39}$$

We will build solutions to (2-22) by minimizing E over $\mathscr{A}$. First we show that such a minimizer exists.

Proposition 2.4. *E achieves its infimum on the set $\mathscr{A}$.*

Proof. To begin, we show that E is coercive on $\mathscr{A}$, which amounts to controlling the last term in E. Recall that by (1-9), $\varrho_0(z) \sim (R-z)^{1/(\gamma-1)}$ for z near R. This implies that

$$\frac{\partial_z P_0}{\varrho_0} = \gamma K \varrho_0^{\gamma-2}\partial_z\varrho_0 = \frac{\gamma K}{\gamma-1}\partial_z(\varrho_0^{\gamma-1}) \tag{2-40}$$

is bounded near $z = R$. Since ϱ_0 and $P_0 = K\varrho_0^\gamma$ are smooth and bounded below away from $z = R$, this implies that

$$\left\|\frac{\partial_z P_0}{\varrho_0}\right\|_{L^\infty((0,R))} < \infty. \tag{2-41}$$

Then for any $z_0 \in (0, R)$, we have the bound

$$\begin{aligned}\int_0^R |\partial_z P_0| \frac{|\varphi|^2}{z^3}\,dz &= \int_0^{z_0} |\partial_z P_0| \frac{z|\varphi|^2}{z^4}\,dz + \int_{z_0}^R \frac{|\partial_z P_0|}{z\varrho_0} \frac{\varrho_0|\varphi|^2}{z^2}\,dz \\ &\le z_0 \|\partial_z P_0\|_{L^\infty} \int_0^{z_0} \frac{|\varphi|^2}{z^4}\,dz + \frac{1}{z_0} \left\| \frac{\partial_z P_0}{\varrho_0} \right\|_{L^\infty} \int_{z_0}^R \frac{\varrho_0 |\varphi|^2}{z^2}\,dz \\ &\le z_0 \frac{4}{9} \|\partial_z P_0\|_{L^\infty} \int_0^R \frac{|\partial_z\varphi|^2}{z^2}\,dz + \frac{1}{z_0} \left\| \frac{\partial_z P_0}{\varrho_0} \right\|_{L^\infty}. \end{aligned} \tag{2-42}$$

For the second inequality we have used Lemma 2.2 and the fact that $\varphi \in \mathscr{A}$. Then by choosing z_0 sufficiently small, we have that

$$E(\varphi) \ge -C_{z_0} + \int_0^R \left[\left(\frac{\tilde\delta}{2} + \gamma P_0 \right) \frac{|\partial_z\varphi|^2}{z^2} + \frac{4\tilde\varepsilon}{3z^2} \left| \partial_z\varphi - 3\frac{\varphi}{z} \right|^2 \right] dz \tag{2-43}$$

for a constant $C_{z_0} > 0$ depending on the choice of z_0, which immediately yields the desired coercivity since $\tilde\delta > 0$.

With the coercivity in hand, we may deduce the existence of a minimizer by using the standard direct methods, employing Lemma 2.2 for compactness. ☐

Since a minimizer exists, we can now define the function $\mu : (0, \infty) \to \mathbb{R}$ by

$$\mu(s) = \inf_{\varphi \in \mathscr{A}} E(\varphi; s), \tag{2-44}$$

where we write $E(\varphi) = E(\varphi; s)$ to emphasize the dependence of E on the parameter $s > 0$, that is,

$$E(\varphi; s) = s \int_0^R \left[\delta \frac{|\partial_z\varphi|^2}{z^2} + \frac{4\varepsilon}{3z^2} \left| \partial_z\varphi - 3\frac{\varphi}{z} \right|^2 \right] dz + \int_0^R \left[\gamma P_0 \frac{|\partial_z\varphi|^2}{z^2} + 4\frac{\partial_z P_0}{z^3} |\varphi|^2 \right] dz. \tag{2-45}$$

The minimizer we have constructed satisfies Euler–Lagrange equations of the form (2-22).

Proposition 2.5. *Let $\varphi \in \mathscr{A}$ be the minimizer of E constructed in Proposition 2.4. Let $\mu := E(\varphi)$. Then φ is smooth on $(0, R]$ and satisfies*

$$-\partial_z \left(\left(\frac{4\tilde\varepsilon}{3} + \tilde\delta + \gamma P_0 \right) \frac{\partial_z\varphi}{z^2} \right) + 4\frac{\partial_z P_0}{z^3}\varphi = \frac{\mu\varrho_0}{z^2}\varphi \tag{2-46}$$

along with the boundary conditions $\varphi(0) = 0$ and

$$\tilde\delta \frac{\partial_z\varphi(R)}{R^2} + \frac{4\tilde\varepsilon}{3} \left(\frac{\partial_z\varphi(R)}{R^2} - 3\frac{\varphi(R)}{R^3} \right) = 0. \tag{2-47}$$

Proof. Fix $\varphi_0 \in H_2^1((0, R))$. Define

$$j(t, \tau) = J(\varphi + t\varphi_0 + \tau\varphi) \tag{2-48}$$

and note that $j(0, 0) = 1$. Moreover, j is smooth and

$$\frac{\partial j}{\partial t}(0, 0) = 2\int_0^R \varrho_0 \frac{\varphi_0\varphi}{z^2}\,dz \quad \text{and} \quad \frac{\partial j}{\partial \tau}(0, 0) = 2\int_0^R \varrho_0 \frac{\varphi^2}{z^2}\,dz = 2. \tag{2-49}$$

So, by the inverse function theorem, we can solve for $\tau = \tau(t)$ in a neighborhood of 0 as a C^1 function of t such that $\tau(0) = 0$ and $j(t, \tau(t)) = 1$. We may differentiate the last equation to find

$$\frac{\partial j}{\partial t}(0,0) + \frac{\partial j}{\partial \tau}(0,0)\tau'(0) = 0, \tag{2-50}$$

and hence

$$\tau'(0) = -\frac{1}{2}\frac{\partial j}{\partial t}(0,0) = -\int_0^R \varrho_0 \frac{\varphi_0\varphi}{z^2}\,dz. \tag{2-51}$$

Since φ is a minimizer over $\mathscr{A}$, we then have

$$0 = \frac{d}{dt}\bigg|_{t=0} E(\varphi + t\varphi_0 + \tau(t)\varphi), \tag{2-52}$$

which implies that

$$\begin{aligned} 0 = \int_0^R \frac{\tilde{\delta} + \gamma P_0}{z^2}\partial_z\varphi\big(\partial_z\varphi_0 + \tau'(0)\partial_z\varphi\big)\,dz + \int_0^R 4\frac{\partial_z P_0}{z^3}\varphi\big(\varphi_0 + \tau'(0)\varphi\big)\,dz \\ + \int_0^R \frac{4\tilde{\varepsilon}}{3z^2}\left(\partial_z\varphi - 3\frac{\varphi}{z}\right)\left(\partial_z\varphi_0 - 3\frac{\varphi_0}{z} + \tau'(0)\left(\partial_z\varphi - 3\frac{\varphi}{z}\right)\right)dz. \end{aligned} \tag{2-53}$$

Rearranging and plugging in the value of $\tau'(0)$, we may rewrite this equation as

$$\begin{aligned} \mu \int_0^R \frac{\varrho_0}{z^2}\varphi_0\varphi\,dz \\ = \int_0^R \frac{\tilde{\delta} + \gamma P_0}{z^2}\partial_z\varphi\partial_z\varphi_0\,dz + \int_0^R 4\frac{\partial_z P_0}{z^3}\varphi\varphi_0\,dz + \int_0^R \frac{4\tilde{\varepsilon}}{3z^2}\left(\partial_z\varphi - 3\frac{\varphi}{z}\right)\left(\partial_z\varphi_0 - 3\frac{\varphi_0}{z}\right)dz, \end{aligned} \tag{2-54}$$

where the eigenvalue is $\mu = E(\varphi)$.

By making variations with φ_0 compactly supported in $(0, R)$, we find that φ satisfies (2-46) in a weak sense in $(0, R)$. Standard bootstrapping arguments then show that $\varphi \in H^k((z_0, R))$ for all $k \geq 0$ and $0 < z_0 < R$, and hence φ is smooth in $(0, R]$. This implies that the equations are also classically satisfied. Since $\varphi \in H^2((R/2, R))$, the traces of φ, $\partial_z\varphi$ are well-defined at the endpoint $z = R$. Making variations with respect to arbitrary $\varphi_0 \in C_c^\infty((0, R])$, we find that the boundary condition (2-47) is satisfied. The condition $\varphi(0) = 0$ is satisfied by virtue of Lemma 2.2. □

We now want to show that the minimizers, which are solutions to (2-46), satisfy the asymptotic condition $|\varphi(z)|/z^2 \to 0$ as $z \to 0$. As a preliminary step, we record an asymptotic result for solutions to a more generic ODE.

Lemma 2.6 [Lin 1997, Proposition A.1]. *Suppose that $\psi(\tau)$ solves*

$$\psi''(\tau) + \big(\alpha\tau^{-1} + g(\tau)\big)\psi'(\tau) + \tau^{-1}f(\tau)\psi(\tau) = 0 \quad \textit{if}\ 0 < \tau < \tau_0, \tag{2-55}$$

where $' = d/d\tau$ and $f, g \in C^0([0, \tau_0])$. If $\alpha < 0$, then either $\psi(0) \neq 0$ or

$$|\psi(\tau)| \leq \frac{C}{\tau^{\alpha-1}} \quad \textit{and} \quad |\psi'(\tau)| \leq \frac{C}{\tau^{\alpha}}. \tag{2-56}$$

Proof. The case $\alpha > 2$ is the content of Proposition A.1 of [Lin 1997], but the proof of the proposition also shows the result when $\alpha < 0$. ☐

Next we use this lemma to establish the asymptotics at $z = 0$ for solutions to (2-46).

Lemma 2.7. *Suppose φ is a solution of* (2-46). *Then $|\varphi(z)| \le Cz^3$ and $|\varphi'(z)| \le Cz^2$ near $z = 0$.*

Proof. We begin by rewriting (2-46). Define $X = \tilde{\varepsilon} + \tilde{\delta} + \gamma P_0 = \tilde{\varepsilon} + \tilde{\delta} + \gamma K \varrho_0^\gamma$ and $X_0 = \tilde{\varepsilon} + P_0 = \tilde{\varepsilon} + K \varrho_0^\gamma$. Then (2-46) is equivalent to the equation

$$\varphi'' + \left(\frac{X'}{X} - \frac{2}{z} \right) \varphi' - 4 \frac{X_0'}{zX} \varphi = -\mu \frac{\varrho_0}{X} \varphi. \tag{2-57}$$

Note that X'/X, X_0'/X, and ϱ_0/X are all continuous at $z = 0$, so we may apply Lemma 2.6 with $\alpha = -2$ to deduce that either $\varphi(0) \ne 0$ or $|\varphi(z)| \le Cz^3$ and $|\varphi'(z)| \le Cz^2$ near $z = 0$. By Lemma 2.2, the former condition cannot hold, so the latter conditions must be the case. ☐

2E. *Properties of the eigenvalue $\mu(s)$.*

It is convenient to decompose E according to

$$E(\varphi; s) = E_0(\varphi) + s E_1(\varphi) \tag{2-58}$$

for

$$E_0(\varphi) := \int_0^R \left(\gamma P_0 \frac{|\partial_z \varphi|^2}{z^2} + 4 \frac{\partial_z P_0}{z^3} |\varphi|^2 \right) dz \tag{2-59}$$

and

$$E_1(\varphi) := \int_0^R \left(\delta \frac{|\partial_z \varphi|^2}{z^2} + \frac{4\varepsilon}{3z^2} \left| \partial_z \varphi - 3 \frac{\varphi}{z} \right|^2 \right) dz \ge 0. \tag{2-60}$$

Notice that since E_0 does not involve either δ or ε, we may view it as the "inviscid" part of E. Because $\partial_z P_0$ lacks a sign, E_0 fails to be nonnegative. However, an easy modification of the argument used in Proposition 2.4 shows that $\inf_{\mathcal{A}} E_0 > -\infty$. As a consequence of the analysis of the inviscid problem, carried out by Lin [1997], we have that this infimum is actually negative and is achieved, and its value characterizes the fastest growing mode for the inviscid problem. Indeed, there exists a $\varphi_0 \in \mathcal{A}$ such that

$$0 > -\chi_0^2 := \inf_{\varphi \in \mathcal{A}} E_0(\varphi) = E_0(\varphi_0), \tag{2-61}$$

with $\chi_0 > 0$ the fastest growth rate and φ_0 the corresponding growing mode solution for the linearized inviscid problem.

We are ultimately concerned with finding $\mu = -\lambda^2$ for some $\lambda > 0$. This requires us to work in a range of s such that $\mu(s) < 0$. Our next result shows that $\mu(s) < 0$ for s sufficiently small.

Lemma 2.8. *There exist constants $C_1, C_2 \ge 0$ depending on φ_0 such that*

$$\mu(s) \le s(\delta C_1 + \varepsilon C_2) - \chi_0^2. \tag{2-62}$$

In particular, $\mu(s) < 0$ for s sufficiently small.

Proof. Let $\varphi_0 \in \mathscr{A}$ be the minimizer of E_0 from (2-61). Then, using the decomposition (2-58), we find

$$\mu(s) = \inf_{\mathscr{A}} E \le E(\varphi_0) = E_0(\varphi_0) + sE_1(\varphi_0) = sE_1(\varphi_0) - \chi_0^2. \tag{2-63}$$

Then $E_1(\varphi_0) = \delta C_1 + \varepsilon C_2$, where

$$C_1 = \int_0^R \frac{|\partial_z \varphi_0|^2}{z^2}\, dz > 0 \quad \text{and} \quad C_2 = \int_0^R \frac{4}{3z^2} \left| \partial_z \varphi_0 - 3\frac{\varphi_0}{z} \right|^2 dz > 0. \tag{2-64}$$

Strict inequality holds for C_1 since $\varphi_0 \in \mathscr{A}$, while it holds for C_2 since $C_2 = 0$ if and only if $\varphi_0(z) = \alpha z^3$ for some $\alpha \in \mathbb{R}$, but one can check that this is not a solution to (2-46) with $\tilde{\varepsilon}, \tilde{\delta} = 0$. □

The next proposition proves some crucial monotonicity and continuity properties of $\mu(s)$ for $s > 0$.

Proposition 2.9. (1) *$\mu(s)$ is strictly increasing in s.*

(2) *There exists a constant $C_3 > 0$ such that*

$$\mu(s) \ge -\chi_0^2 + sC_3\delta, \tag{2-65}$$

where $\chi_0 > 0$ is given in (2-61).

(3) *μ is locally Lipschitz on* $(0, \infty)$, *and in particular, μ is continuous on* $(0, \infty)$.

Proof. We begin by establishing some notation. According to Proposition 2.4, for each $s \in (0, \infty)$ we can find $\varphi_s \in \mathscr{A}$ such that

$$E(\varphi_s; s) = \inf_{\varphi \in \mathscr{A}} E(\varphi; s) = \mu(s). \tag{2-66}$$

Next, we recall the decomposition of E given in (2-58) and note that the nonnegativity of E_1 implies that E is nondecreasing in s with $\varphi \in \mathscr{A}$ kept fixed.

To prove the first assertion, note that if $s_1, s_2 \in (0, \infty)$ with $s_1 \le s_2$, then the minimality of φ_{s_i} and the nonnegativity of E_1 imply that

$$\mu(s_1) = E(\varphi_{s_1}; s_1) \le E(\varphi_{s_2}; s_1) \le E(\varphi_{s_2}; s_2) = \mu(s_2). \tag{2-67}$$

This shows that μ is nondecreasing in s. Suppose by way of contradiction that $\mu(s_1) = \mu(s_2)$ for $s_1 \neq s_2$. Then the last inequality implies that

$$s_1 E_1(\varphi_{s_2}) = s_2 E_1(\varphi_{s_2}), \tag{2-68}$$

which means that $E_1(\varphi_{s_2}) = 0$. The vanishing of $E_1(\varphi_{s_2})$ implies that $\varphi_{s_2} = 0$, which is impossible since $\varphi_{s_2} \in \mathscr{A}$. Hence equality cannot be achieved, and μ is strictly increasing in s.

Now note that (2-58), the nonnegativity of E_1, and (2-61) imply that

$$\mu(s) \ge \inf_{\varphi \in \mathscr{A}} E_0(\varphi) + s \inf_{\varphi \in \mathscr{A}} E_1(\varphi) = -\chi_0^2 + s \inf_{\varphi \in \mathscr{A}} E_1(\varphi). \tag{2-69}$$

It is a simple matter to see that

$$\inf_{\varphi \in \mathscr{A}} E_1(\varphi) \ge \delta \inf_{\varphi \in \mathscr{A}} \int_0^R \frac{|\partial_z \varphi|^2}{z^2}\, dz := C_3\delta > 0. \tag{2-70}$$

The second assertion follows.

Now fix $Q = [a, b] \Subset (0, \infty)$, and fix any $\psi \in \mathcal{A}$. Again by the nonnegativity of E_1 and the minimality of φ_s, we deduce that

$$E(\psi; b) \geq E(\psi; s) \geq E(\varphi_s; s) \geq a E_1(\varphi_s) - \chi_0^2 \tag{2-71}$$

for all $s \in Q$. This implies that there exists a constant $0 < C = C(a, b, \psi, \gamma, K) < \infty$ such that

$$\sup_{s \in Q} E_1(\varphi_s) \leq C. \tag{2-72}$$

Let $s_1, s_2 \in Q$. Using the minimality of φ_{s_1} compared to φ_{s_2}, we know that

$$\mu(s_1) = E(\varphi_{s_1}; s_1) \leq E(\varphi_{s_2}; s_1), \tag{2-73}$$

but from our decomposition (2-58), we may bound

$$E(\varphi_{s_2}; s_1) \leq E(\varphi_{s_2}; s_2) + |s_1 - s_2| E_1(\varphi_{s_2}) = \mu(s_2) + |s_1 - s_2| E_1(\varphi_{s_2}). \tag{2-74}$$

Chaining these two inequalities together and employing (2-72), we find that

$$\mu(s_1) \leq \mu(s_2) + C|s_1 - s_2|. \tag{2-75}$$

Reversing the role of the indices 1 and 2 in the derivation of this inequality gives the same bound with s_1 switched with s_2. We deduce that

$$|\mu(s_1) - \mu(s_2)| \leq C|s_1 - s_2|, \tag{2-76}$$

which proves item (3). ☐

Now we know that the eigenvalue $\mu(s)$ is negative as long as $s < \dfrac{\chi_0^2}{\delta C_1 + \varepsilon C_2}$ and that μ is continuous on $(0, \infty)$. We can then define the nonempty open set

$$\Omega = \mu^{-1}((-\infty, 0)) \subset (0, \infty), \tag{2-77}$$

on which we can calculate $\lambda(s) = \sqrt{-\mu(s)} > 0$.

It turns out that the set Ω is sufficiently large to find $s > 0$ such that $\lambda(s) = s$. This inversion will then allow us to solve the original growing-mode equations.

Proposition 2.10. *There exists a unique* $s \in \Omega$ *such that* $\lambda(s) = \sqrt{-\mu(s)} > 0$ *and* $\lambda(s) = s$.

Proof. According to Lemma 2.8, we know that $\mu(s) < 0$ for $s \in \big[0, \chi_0^2/(\delta C_1 + \varepsilon C_2)\big)$. Moreover, the lower bound (2-65) in Proposition 2.9 implies that $\mu(s) \to +\infty$ as $s \to \infty$. This implies the existence of $s_0 \in (0, \infty)$ such that $\Omega = (0, s_0)$, which means that $\lambda(s_0) = 0$. Define the function $\Psi : (0, s_0) \to (0, \infty)$ by $\Psi(s) = s/\lambda(s)$. The monotonicity and continuity properties of μ are inherited by Ψ, that is, Ψ is continuous on $(0, s_0)$ and strictly increases from 0 to $+\infty$ as $s \to s_0$. As such, we may apply the intermediate value theorem to find a unique $s \in (0, s_0)$ such that $\Psi(s) = 1$. For this s, we then have that $s = \lambda(s)$, the desired result. ☐

Up to now we have viewed the viscosity parameters ε, δ as being fixed. With the unique fixed point $\lambda(s) = \sqrt{-\mu(s)} = s > 0$ in hand, we can now consider the behavior of λ with respect to the viscosity parameters, ε, δ. To this end, let us write $\lambda = \lambda(\delta, \varepsilon)$ in the following.

Proposition 2.11. *Write* $\lambda = \lambda(\delta, \varepsilon) > 0$ *for the unique* λ *produced by Proposition 2.10 for a given* $\varepsilon, \delta > 0$. *Then*

$$\lim_{(\delta,\varepsilon)\to 0} \lambda(\delta, \varepsilon) = \chi_0 \tag{2-78}$$

and

$$\lim_{\delta\to\infty} \lambda(\delta, \varepsilon) = 0. \tag{2-79}$$

Proof. Combining the estimate from Lemma 2.8 with (2-65) from Proposition 2.9 and employing Proposition 2.10, we find that

$$\lambda(\delta, \varepsilon)^2 + \delta C_3 \lambda(\delta, \varepsilon) - \chi_0^2 \le 0 \le \lambda(\delta, \varepsilon)^2 + (\delta C_1 + \varepsilon C_2)\lambda(\delta, \varepsilon) - \chi_0^2, \tag{2-80}$$

for constants $C_1, C_2, C_3 > 0$ independent of ε, δ. The first inequality in (2-80) implies that

$$\lambda(\delta, \varepsilon) \le \tfrac{1}{2}\bigl(-\delta C_3 + \sqrt{\delta^2 C_3^2 + 4\chi_0^2}\bigr), \tag{2-81}$$

while the second and the fact that $\lambda(\delta, \varepsilon) > 0$ imply that

$$\lambda(\delta, \varepsilon) \ge \tfrac{1}{2}\bigl(-(\delta C_1 + \varepsilon C_2) + \sqrt{(\delta C_1 + \varepsilon C_2)^2 + 4\chi_0^2}\bigr). \tag{2-82}$$

Sending $(\delta, \varepsilon) \to 0$ and chaining together (2-81) and (2-82) then yields (2-78). On the other hand, expanding the right side of (2-81) for large δ shows that

$$0 \le \lambda(\delta, \varepsilon) \le \frac{\chi_0^2}{\delta C_3} + o(1), \quad \text{for } o(1) \to 0 \quad \text{as } \delta \to \infty, \tag{2-83}$$

which implies (2-79). □

Remark 2.12. Proposition 2.11 has two important consequences. The first is that the fastest inviscid growth rate is recovered in the inviscid limit $(\delta, \varepsilon) \to 0$. This can be understood as a continuity result. The second is that large bulk viscosity suppresses the viscous growth rate, and for sufficiently large δ, the growth rate is very slow. This causes the delay of the instability occurrence time.

2F. *Proof of Theorem 2.1.*

We now combine our above analysis to deduce the existence of a solution φ, $\lambda > 0$ to (2-19)–(2-20).

Theorem 2.13. *There exist* $\lambda > 0$ *and* $\varphi \in H_2^1((0, R))$, *smooth on* $(0, R]$, *that solve* (2-19) *along with the boundary condition* (2-20). *The solution satisfies the asymptotics* $|\varphi(z)| \le C z^3$ *and* $|\partial_z \varphi(z)| \le C z^2$ *as* $z \to 0$.

Proof. Combining Propositions 2.5 and 2.10, we see that there exists a solution to (2-22) and (2-23) for $\lambda(s) = \sqrt{-\mu(s)} > 0$, satisfying $s = \lambda(s)$. This implies that the solution is actually a solution to (2-19) and (2-20). The asymptotics at $z = 0$ follow from Lemma 2.7. □

An immediate consequence of Theorem 2.13 is the existence of a solution to (2-16)–(2-17).

Corollary 2.14. *There exist* $\lambda > 0$ *and* $\phi(x) = \varphi(r_0(x))$, *smooth on* $(0, M)$, *that solve* (2-16)–(2-17). *The solution satisfies*

$$\limsup_{x\to 0} \frac{|\phi(x)|}{r_0^3(x)} + \limsup_{x\to 0} |\partial_x \phi(x)| < \infty. \tag{2-84}$$

Let $\mathfrak{D}$ *denote the linear operator* $\mathfrak{D}\phi(x) = \rho_0(x)\partial_x\phi(x)$. *The solution satisfies the property that* $\mathfrak{D}^k\phi$ *has a well-defined trace at* $x = M$ *for every integer* $k \geq 0$.

Proof. All of the conclusions, except those concerning $\mathfrak{D}$, follow directly from Theorem 2.13. When $k = 0$, the trace $\mathfrak{D}^0\phi(M) = \phi(M)$ is well-defined, since $\varphi(R) = \varphi(r_0(M))$ is well-defined. Note that since $\partial_x r_0(x) = 1/(4\pi\rho_0(x)r_0^2(x))$, we have

$$\mathfrak{D}\phi(x) = \rho_0(x)\partial_x\phi(x) = \frac{\partial_z\varphi(r_0(x))}{4\pi r_0^2(x)} \implies \mathfrak{D}\phi(M) = \frac{\partial_z\varphi(R)}{4\pi R^2}, \tag{2-85}$$

so that $\mathfrak{D}\phi(M)$ is well-defined. In other words, the multiplication by ρ_0 in the operator $\mathfrak{D}$ removes the potential singularity in $\partial_x\phi$ near $x = M$. We may argue similarly, using the fact that $\partial_z^k\varphi(R)$ is well-defined for all $k \geq 0$, to deduce that $\mathfrak{D}^k\phi(M)$ is well-defined for all $k \geq 0$ as well. □

Now, with Corollary 2.14 in hand, we are ready to present:

Proof of Theorem 2.1. Let $\lambda > 0$ and $\phi(x)$ be the solution to (2-16)–(2-17) given in Corollary 2.14. Let us then define v and σ according to

$$v = -\frac{\lambda}{4\pi r_0^2}\phi \quad \text{and} \quad \sigma = \rho_0^2\partial_x\phi. \tag{2-86}$$

Using these definitions of v and σ in conjunction with the properties of ϕ recorded in Corollary 2.14, we easily deduce items (1)–(3).

To prove the variational characterization of item (4), we return to the variational characterization of λ in $z = r_0(x)$ coordinates. According to Theorem 2.13, $\lambda > 0$ satisfies

$$\lambda \int_0^R \left(\delta\frac{|\partial_z\vartheta|^2}{z^2} + \frac{4\varepsilon}{3z^2}\left|\partial_z\vartheta - 3\frac{\vartheta}{z}\right|^2 \right) dz + \int_0^R \left(\gamma P_0 \frac{|\partial_z\vartheta|^2}{z^2} + 4\frac{\partial_z P_0}{z^3}|\vartheta|^2 \right) dz \geq -\lambda^2 \int_0^R \frac{\rho_0}{z^2}|\vartheta|^2\, dz \tag{2-87}$$

for every $\vartheta \in H_2^1((0, R))$. Then the variational characterization in (2-12) follows by making a change of coordinates $\theta(x) = \vartheta(z) = \vartheta(r_0(x))$. Note that $\vartheta \in H_2^1((0, R))$ if and only if $\sqrt{\rho_0}\partial_x\theta \in L^2((0, M))$ and $\theta/(r_0^2\sqrt{\rho_0}) \in L^2((0, M))$. Also, changing coordinates in (2-26) of Lemma 2.2 shows that $\theta/(r_0^3\sqrt{\rho_0}) \in L^2((0, M))$, which means that all of the integrals in (2-12) are well-defined.

We now turn to the proof of (2-13). Using the inclusion $\varphi \in H_2^1((0, R))$, the above analysis implies that $\sqrt{\rho_0}\partial_x\phi, \phi/(r_0^3\sqrt{\rho_0}) \in L^2((0, M))$. From this and (2-16), we may then deduce that

$$\int_0^M \left(\frac{|\phi|^2}{r_0^6\rho_0} + \rho_0|\partial_x\phi|^2 + r_0^2|\partial_x(\rho_0\partial_x\phi)|^2 \right) dx < \infty. \tag{2-88}$$

This and (2-86) then imply (2-13). □

3. Linear estimates

Due to the indirect way in which we constructed growing mode solutions in Section 2, it is not immediately obvious that the $\lambda > 0$ of Theorem 2.1 is the largest possible growth rate. However, because of the inequality (2-12), we can show that no solution to the linearized problem (2-1)–(2-4) can grow in time at a rate faster than $e^{\lambda t}$. Hence the growing mode constructed in Theorem 2.1 actually does grow in time at the fastest possible rate. The proof of this result and its implications for solutions to the inhomogeneous linearized problem are the subject of this section.

3A. *Estimates in the second-order formulation.*

First we will prove estimates for solutions to the following second-order problem.

$$-\frac{\partial_t^2 \phi}{16\pi^2 r_0^4} = \frac{\partial_x P_0}{\pi r_0^3}\phi - \partial_x\left[\left(\frac{4\varepsilon}{3}+\delta\right)\rho_0 \partial_x \partial_t \phi + \gamma P_0 \rho_0 \partial_x \phi\right] \quad \text{for } x \in (0, M), \tag{3-1}$$

with boundary conditions

$$\phi(0,t) = 0 \quad \text{and} \quad \frac{4\varepsilon}{3}\left(4\pi r_0^3 \rho_0 \partial_x\left(\frac{\phi}{r_0^3}\right)\right) + \delta(4\pi\rho_0\partial_x\phi) = 0 \quad \text{at } x = M, \tag{3-2}$$

and initial conditions $\phi(x,0)$ and $\partial_t\phi(x,0)$ given. We will assume throughout that ϕ satisfies $\sqrt{\rho_0}\partial_x\phi \in L^2((0,M))$ and $\phi/(r_0^2\sqrt{\rho_0}) \in L^2((0,M))$.

Solutions to this linear problem obey an energy evolution equation related to the inequality (2-12). We record this now.

Proposition 3.1. *Suppose ϕ is a solution to* (3-1)–(3-2). *Then*

$$\partial_t \int_0^M \frac{|\partial_t\phi|^2}{32\pi^2 r_0^4}\,dx + \int_0^M \left(\delta\rho_0|\partial_x\partial_t\phi|^2 + \frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{\partial_t\phi}{r_0^3}\right)\right|^2\right)dx = -\partial_t \int_0^M \left(\frac{\gamma P_0\rho_0}{2}|\partial_x\phi|^2 + \frac{\partial_x P_0}{2\pi r_0^3}|\phi|^2\right)dx. \tag{3-3}$$

Proof. Multiply (3-1) by $\partial_t\phi$ and integrate over $x \in (0, M)$. An integration by parts, an application of the boundary conditions (3-2), and some simple algebra yield the desired equality. ☐

We can use this and the variational characterization of λ given in Theorem 2.1 to deduce some estimates.

Theorem 3.2. *Let ϕ solve* (3-1)–(3-2). *Then we have the following estimates*:

$$\int_0^M \frac{|\phi(t)|^2}{16\pi^2 r_0^4}\,dx + \int_0^t\int_0^M \left(\delta\rho_0|\partial_x\phi(s)|^2 + \frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{\phi(s)}{r_0^3}\right)\right|^2\right)dx\,ds \le e^{2\lambda t}\int_0^M \frac{|\phi(0)|^2}{16\pi^2 r_0^4}dx + \frac{K_1}{2\lambda}(e^{2\lambda t}-1), \tag{3-4}$$

$$\frac{1}{\lambda}\int_0^M \frac{|\partial_t\phi(t)|^2}{16\pi^2 r_0^4}\,dx+\int_0^M\left(\delta\rho_0|\partial_x\phi(t)|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{\phi(t)}{r_0^3}\right)\right|^2\right)dx \\ \le e^{2\lambda t}\left(2\lambda\int_0^M\frac{|\phi(0)|^2}{16\pi^2 r_0^4}\,dx+K_1\right), \tag{3-5}$$

and

$$\frac{1}{2}\int_0^M \gamma P_0\rho_0|\partial_x\phi(t)|^2\,dx\le K_0+C_0\left[e^{2\lambda t}\int_0^M\frac{|\phi(0)|^2}{16\pi^2 r_0^4}\,dx+\frac{K_1}{2\lambda}(e^{2\lambda t}-1)\right]. \tag{3-6}$$

Here

$$K_0=\int_0^M\frac{|\partial_t\phi(0)|^2}{16\pi^2 r_0^4}\,dx+\frac{1}{2}\int_0^M\gamma P_0\rho_0|\partial_x\phi(0)|^2\,dx, \tag{3-7}$$

$$K_1=\frac{2K_0}{\lambda}+2\int_0^M\left(\delta\rho_0|\partial_x\phi(0)|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{\phi(0)}{r_0^3}\right)\right|^2\right)dx, \tag{3-8}$$

and

$$C_0=2\sup_{x\in(0,M)}\frac{x}{r_0^3(x)}<\infty. \tag{3-9}$$

Proof. We integrate the result of Proposition 3.1 in time from 0 to t to see that

$$\int_0^M\frac{|\partial_t\phi(t)|^2}{32\pi^2 r_0^4}\,dx+\int_0^t\int_0^M\left(\delta\rho_0|\partial_x\partial_t\phi(s)|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{\partial_t\phi(s)}{r_0^3}\right)\right|^2\right)dx\,ds \\ =K_0+\int_0^M\frac{\partial_x P_0}{2\pi r_0^3}|\phi(0)|^2\,dx-\int_0^M\left(\frac{\gamma P_0\rho_0}{2}|\partial_x\phi(t)|^2+\frac{\partial_x P_0}{2\pi r_0^3}|\phi(t)|^2\right)dx. \tag{3-10}$$

Note that since

$$\partial_x P_0=-\frac{x}{4\pi r_0^4},$$

we have

$$\int_0^M\frac{\partial_x P_0}{2\pi r_0^3}|\phi(0)|^2dx=-\int_0^M\frac{x}{8\pi^2 r_0^7}|\phi(0)|^2dx\le 0. \tag{3-11}$$

The variational characterization of λ given in (2-12) of Theorem 2.1 allows us to estimate

$$-\frac{1}{2}\int_0^M\left(\frac{\gamma P_0\rho_0}{2}|\partial_x\phi(t)|^2+\frac{\partial_x P_0}{2\pi r_0^3}|\phi(t)|^2\right)dx-\frac{\lambda}{2}\int_0^M\left(\delta\rho_0|\partial_x\phi(t)|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{\phi(t)}{r_0^3}\right)\right|^2\right)dx \\ \le\frac{\lambda^2}{2}\int_0^M\frac{|\phi(t)|^2}{16\pi^2 r_0^4}\,dx. \tag{3-12}$$

We may then combine (3-10)–(3-12) to see that

$$\int_0^M \frac{|\partial_t\phi(t)|^2}{32\pi^2 r_0^4}\,dx+\int_0^t\int_0^M\left(\delta\rho_0|\partial_x\partial_t\phi(s)|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{\partial_t\phi(s)}{r_0^3}\right)\right|^2\right)dx\,ds \\ \le K_0+\frac{\lambda^2}{2}\int_0^M\frac{|\phi(t)|^2}{16\pi^2 r_0^4}\,dx+\frac{\lambda}{2}\int_0^M\left(\delta\rho_0|\partial_x\phi(t)|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{\phi(t)}{r_0^3}\right)\right|^2\right)dx. \quad (3\text{-}13)$$

For the sake of brevity in the rest of the proof, we now rewrite (3-13) as

$$\frac{1}{2}\|\partial_t\phi(t)\|_1^2+\int_0^t\|\partial_t\phi(s)\|_2^2\,ds\le K_0+\frac{\lambda^2}{2}\|\phi(t)\|_1^2+\frac{\lambda}{2}\|\phi(t)\|_2^2 \quad (3\text{-}14)$$

for the two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ given by

$$\|\psi\|_1^2:=\int_0^M\frac{|\psi|^2}{16\pi^2 r_0^4}\,dx, \quad (3\text{-}15)$$

$$\|\psi\|_2^2:=\int_0^M\left(\delta\rho_0|\partial_x\psi|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{\psi}{r_0^3}\right)\right|^2\right)dx. \quad (3\text{-}16)$$

Both of these norms are clearly generated by inner products, which we will write as $\langle\cdot,\cdot\rangle_i$ for $i=1,2$.

Integrating in time and using Cauchy's inequality, we may write the bound

$$\lambda\|\phi(t)\|_2^2=\lambda\|\phi(0)\|_2^2+\lambda\int_0^t 2\langle\phi(s),\partial_t\phi(s)\rangle_2\,ds \\ \le\lambda\|\phi(0)\|_2^2+\int_0^t\|\partial_t\phi(s)\|_2^2\,ds+\lambda^2\int_0^t\|\phi(s)\|_2^2\,ds. \quad (3\text{-}17)$$

On the other hand,

$$\lambda\partial_t\|\phi(t)\|_1^2=\lambda 2\langle\partial_t\phi(t),\phi(t)\rangle_1\le\lambda^2\|\phi(t)\|_1^2+\|\partial_t\phi(t)\|_1^2. \quad (3\text{-}18)$$

We may combine these two inequalities with (3-14) to derive the differential inequality

$$\partial_t\|\phi(t)\|_1^2+\|\phi(t)\|_2^2\le K_1+2\lambda\|\phi(t)\|_1^2+2\lambda\int_0^t\|\phi(s)\|_2^2\,ds, \quad (3\text{-}19)$$

for K_1 as defined in the hypotheses. An application of Gronwall's lemma then shows that

$$\|\phi(t)\|_1^2+\int_0^t\|\phi(s)\|_2^2\,ds\le e^{2\lambda t}\|\phi(0)\|_1^2+\frac{K_1}{2\lambda}(e^{2\lambda t}-1) \quad (3\text{-}20)$$

for all $t\ge 0$, which is the bound (3-4).

To derive the estimate (3-5), we return to (3-14) and plug in (3-17) and (3-20) to see that

$$\frac{1}{\lambda}\|\partial_t\phi(t)\|_1^2+\|\phi(t)\|_2^2\le K_1+\lambda\|\phi(t)\|_1^2+2\lambda\int_0^t\|\phi(s)\|_2^2\,ds\le e^{2\lambda t}\big(2\lambda\|\phi(0)\|_1^2+K_1\big). \quad (3\text{-}21)$$

Finally, for (3-6), we return to (3-10) and employ (3-11) to see that

$$\frac{1}{2}\int_0^M \gamma P_0\rho_0|\partial_x\phi(t)|^2\,dx \le K_0-\int_0^M \frac{\partial_x P_0}{2\pi r_0^3}|\phi(t)|^2\,dx = K_0+\int_0^M \frac{x}{8\pi^2 r_0^7}|\phi(t)|^2\,dx. \tag{3-22}$$

Since L'Hospital's theorem implies that

$$\lim_{x\to 0}\frac{x}{r_0^3(x)}=\lim_{x\to 0}\frac{4\pi\rho_0(x)}{3}=\frac{4\pi\rho_0(0)}{3}<\infty, \tag{3-23}$$

we may deduce that

$$\sup_{x\in(0,M)}\frac{x}{r_0^3(x)}<\infty. \tag{3-24}$$

The estimate (3-6) then follows directly from (3-22), (3-24), and the estimate of $\|\phi(t)\|_1^2$ in (3-4). □

3B. ***Estimates in the first-order formulation.*** Now consider σ and w to be solutions to the first-order linear system (2-5) with boundary conditions (2-6) and initial conditions $\sigma(x,0)$ and $w(x,0)$. A simple calculation shows that if we apply ∂_t to the second equation in (2-5) and then eliminate $\partial_t\sigma$ by using the first equation in (2-5), then we arrive at the second-order formulation (3-1)–(3-2) for $\phi=w$. Then Theorem 3.2 yields various estimates for $w=\phi$. We now seek to rewrite these estimates for w and to use them to derive a similar estimate for σ.

Theorem 3.3. *Let* σ, w *solve the linear system* (2-5)–(2-6). *Then*

$$\begin{aligned}\int_0^M \frac{|w(t)|^2}{r_0^4}\,dx+\int_0^M \gamma K\rho_0^{\gamma-1}\left|\frac{\sigma(t)}{\rho_0}\right|^2 dx+16\pi^2\int_0^M\left(\delta\rho_0|\partial_x w(t)|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{w(t)}{r_0^3}\right)\right|^2\right)dx\\ \le Ce^{2\lambda t}\left[\int_0^M \frac{|w(0)|^2}{r_0^4}\,dx+\int_0^M\left(\gamma K\rho_0^{\gamma-1}\left|\frac{\sigma(0)}{\rho_0}\right|^2+\gamma K\rho_0^{\gamma+1}|\partial_x w(0)|^2+\frac{|\partial_t w(0)|^2}{r_0^4}\right)dx\right.\\ \left.+\int_0^M\left(\delta\rho_0|\partial_x w(0)|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{w(0)}{r_0^3}\right)\right|^2\right)dx\right].\end{aligned} \tag{3-25}$$

Proof. We switch to the second-order formulation for $\phi=w$. Then the estimates (3-4)–(3-6) of Theorem 3.2 imply that

$$\begin{aligned}\int_0^M \frac{|w(t)|^2}{r_0^4}\,dx+\int_0^M \gamma K\rho_0^{\gamma+1}|\partial_x w(t)|^2\,dx+16\pi^2\int_0^M\left(\delta\rho_0|\partial_x w(t)|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{w(t)}{r_0^3}\right)\right|^2\right)dx\\ \le Ce^{2\lambda t}\left[\int_0^M \frac{|w(0)|^2}{r_0^4}\,dx+\int_0^M\left(\gamma K\rho_0^{\gamma+1}|\partial_x w(0)|^2+\frac{|\partial_t w(0)|^2}{r_0^4}\right)dx\right.\\ \left.+\int_0^M\left(\delta\rho_0|\partial_x w(0)|^2+\frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{w(0)}{r_0^3}\right)\right|^2\right)dx\right].\end{aligned} \tag{3-26}$$

Let us call the term in the brackets on the right side of this equation $\mathscr{X}_0$. Since $\partial_t\sigma=-4\pi\rho_0^2\partial_x w$, we then have that

$$\int_0^M \gamma K \rho_0^{\gamma+1} |\partial_x w(t)|^2 \, dx = \int_0^M \frac{\gamma K \rho_0^{\gamma-1}}{16\pi^2 \rho_0^2} |\partial_t \sigma(t)|^2 \, dx. \tag{3-27}$$

The right side of (3-27) defines the square of a norm $\|\cdot\|$ in a Hilbert space, and in this case Cauchy–Schwarz and the chain rule imply that $\partial_t \|\psi(t)\| \le \|\partial_t \psi(t)\|$ for a one-parameter family $\psi(t)$ in the space. Using this, we then have that

$$\partial_t \left(\int_0^M \frac{\gamma K \rho_0^{\gamma-1}}{16\pi^2 \rho_0^2} |\sigma(t)|^2 \, dx \right)^{1/2} \le \left(\int_0^M \frac{\gamma K \rho_0^{\gamma-1}}{16\pi^2 \rho_0^2} |\partial_t \sigma(t)|^2 \, dx \right)^{1/2} \le \sqrt{C \mathcal{Z}_0} e^{\lambda t}. \tag{3-28}$$

Integrating this in time, we then find that

$$\begin{aligned} \left(\int_0^M \frac{\gamma K \rho_0^{\gamma-1}}{16\pi^2 \rho_0^2} |\sigma(t)|^2 \, dx \right)^{1/2} &\le \left(\int_0^M \frac{\gamma K \rho_0^{\gamma-1}}{16\pi^2 \rho_0^2} |\sigma(0)|^2 \, dx \right)^{1/2} + \frac{\sqrt{C \mathcal{Z}_0}}{\lambda} (e^{\lambda t} - 1) \\ &\le C e^{\lambda t} \sqrt{ \int_0^M \frac{\gamma K \rho_0^{\gamma-1}}{16\pi^2 \rho_0^2} |\sigma(0)|^2 \, dx + \mathcal{Z}_0 }. \end{aligned} \tag{3-29}$$

The estimate (3-25) then follows directly from (3-26), (3-27), and (3-29). ☐

3C. ***Estimates for the inhomogeneous first-order problem.*** Consider the linear operators

$$\mathcal{L}_1 w = 4\pi \rho_0^2 \partial_x w, \tag{3-30}$$

$$\mathcal{L}_2 \sigma = 4\pi r_0^4 \partial_x (\gamma K \rho_0^{\gamma-1} \sigma) - 4 r_0 \partial_x P_0 \int_0^x \frac{\sigma(y)}{\rho_0^2(y)} \, dy, \tag{3-31}$$

$$\mathcal{L}_3 w = -16\pi^2 r_0^4 \partial_x \left[\left(\frac{4\varepsilon}{3} + \delta \right) \rho_0 \partial_x w \right], \tag{3-32}$$

and the corresponding matrix of operators

$$\mathcal{L} = \begin{pmatrix} 0 & -\mathcal{L}_1 \\ -\mathcal{L}_2 & -\mathcal{L}_3 \end{pmatrix}. \tag{3-33}$$

We also consider the boundary operator

$$\mathcal{B}(w) = -\frac{4\varepsilon}{3} \left(4\pi r_0^3 \rho_0 \partial_x \left(\frac{w}{r_0^3} \right) \right) - \delta (4\pi \rho_0 \partial_x w). \tag{3-34}$$

Notice that the first-order equations (2-5)–(2-6) are equivalent to the equation

$$\partial_t \begin{pmatrix} \sigma \\ w \end{pmatrix} - \mathcal{L} \begin{pmatrix} \sigma \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{3-35}$$

with homogeneous boundary conditions

$$\frac{w}{r_0^2}(0,t) = \sigma(M,t) = 0 \quad \text{and} \quad \mathcal{B}(w) = 0 \quad \text{at } x = M. \tag{3-36}$$

Let us by denote $e^{t\mathscr{L}}$ the solution operator to (3-35)–(3-36), that is,

$$e^{t\mathscr{L}}\begin{pmatrix}\sigma(0)\\ w(0)\end{pmatrix}=\begin{pmatrix}\sigma(t)\\ w(t)\end{pmatrix}, \tag{3-37}$$

where σ and w solve (3-35)–(3-36) with initial data $\sigma(0)$ and $w(0)$. Note that below in (3-47) we show this operator is bounded.

Suppose now that σ and w solve the inhomogeneous problem

$$\partial_t\begin{pmatrix}\sigma\\ w\end{pmatrix}-\mathscr{L}\begin{pmatrix}\sigma\\ w\end{pmatrix}=\begin{pmatrix}N_1\\ N_2\end{pmatrix} \tag{3-38}$$

along with the boundary conditions

$$\frac{w}{r_0^2}(0,t)=\sigma(M,t)=0 \quad\text{and}\quad \mathscr{B}(w)=N_{\mathscr{B}} \quad\text{at } x=M. \tag{3-39}$$

Here we assume that $N_1=N_1(x,t)$, $N_2=N_2(x,t)$, but that $N_{\mathscr{B}}=N_{\mathscr{B}}(t)$, that is, the boundary inhomogeneity only depends on time. In order to use the linear theory we have developed, we must rewrite this as a system with homogeneous boundary conditions. To accomplish this, we will utilize the following lemma.

Lemma 3.4. *Let*

$$\psi(x,t)=-\frac{N_{\mathscr{B}}(t)}{3\delta}r_0^3(x). \tag{3-40}$$

Then for each t, $\psi(t)$ *satisfies* $\mathscr{L}_3\psi(t)=0$ *for* $x\in(0,M)$ *and* $\mathscr{B}\psi(t)=N_{\mathscr{B}}(t)$ *at* $x=M$. *Also,* $\mathscr{L}_1\psi(t)=-N_{\mathscr{B}}(t)\rho_0(x)/\delta$.

Proof. The results follow from simple computations. □

With this ψ in hand, we can reformulate (3-38)–(3-39) so that the resulting problem has homogeneous boundary conditions. Let $w=\psi+\bar{w}$. Then Lemma 3.4 implies that

$$\partial_t\begin{pmatrix}\sigma\\ \bar{w}\end{pmatrix}-\mathscr{L}\begin{pmatrix}\sigma\\ \bar{w}\end{pmatrix}=\begin{pmatrix}N_1\\ N_2\end{pmatrix}+\begin{pmatrix}-\mathscr{L}_1\psi\\ -\partial_t\psi\end{pmatrix}=\begin{pmatrix}N_1+\dfrac{N_{\mathscr{B}}\rho_0}{\delta}\\ N_2+\dfrac{\partial_t N_{\mathscr{B}}r_0^3}{3\delta}\end{pmatrix}, \tag{3-41}$$

along with the boundary conditions

$$\frac{\bar{w}}{r_0^2}(0,t)=\sigma(M,t)=0 \quad\text{and}\quad \mathscr{B}(\bar{w})=0 \quad\text{at } x=M. \tag{3-42}$$

Employing the variation of parameters, we can then solve (3-41)–(3-42) via

$$\begin{pmatrix}\sigma(t)\\ \bar{w}(t)\end{pmatrix}=e^{t\mathscr{L}}\begin{pmatrix}\sigma(0)\\ \bar{w}(0)\end{pmatrix}+\int_0^t e^{(t-s)\mathscr{L}}\begin{pmatrix}N_1(s)\\ N_2(s)\end{pmatrix}ds+\frac{1}{\delta}\int_0^t e^{(t-s)\mathscr{L}}\begin{pmatrix}N_{\mathscr{B}}(s)\rho_0\\ \frac{1}{3}\partial_t N_{\mathscr{B}}(s)r_0^3\end{pmatrix}ds. \tag{3-43}$$

We can then go back to $w = \psi + \bar{w}$:

$$\begin{pmatrix}\sigma(t)\\ w(t)\end{pmatrix} = e^{t\mathscr{L}}\begin{pmatrix}\sigma(0)\\ \bar{w}(0)\end{pmatrix} - \frac{1}{\delta}\begin{pmatrix}0\\ \frac{1}{3}N_{\mathscr{B}}(t)r_0^3\end{pmatrix} + \int_0^t e^{(t-s)\mathscr{L}}\begin{pmatrix}N_1(s)\\ N_2(s)\end{pmatrix} ds + \frac{1}{\delta}\int_0^t e^{(t-s)\mathscr{L}}\begin{pmatrix}N_{\mathscr{B}}(s)\rho_0\\ \frac{1}{3}\partial_t N_{\mathscr{B}}(s)r_0^3\end{pmatrix} ds. \tag{3-44}$$

Now let us define a norm for the pair σ, w given by

$$\left\|\begin{pmatrix}\sigma\\ w\end{pmatrix}\right\|_0^2 := \frac{1}{2}\int_0^M \gamma K\rho_0^{\gamma-1}\left|\frac{\sigma}{\rho_0}\right|^2 dx + \frac{1}{2}\int_0^M \frac{|w|^2}{r_0^4}\,dx + \frac{1}{2}\int_0^M 16\pi^2\left(\delta\rho_0|\partial_x w|^2 + \frac{4\varepsilon}{3}\rho_0\left|r_0^3\partial_x\left(\frac{w}{r_0^3}\right)\right|^2\right) dx. \tag{3-45}$$

We also define

$$\mathfrak{E}(\sigma, w) := \left\|\begin{pmatrix}\sigma\\ w\end{pmatrix}\right\|_0^2 + \frac{1}{2}\int_0^M \gamma K\rho_0^{\gamma-1}\left|\frac{\partial_t\sigma}{\rho_0}\right|^2 dx + \frac{1}{2}\int_0^M \frac{|\partial_t w|^2}{r_0^4}\,dx. \tag{3-46}$$

We can then recast the result of Theorem 3.3 as

$$\left\|e^{t\mathscr{L}}\begin{pmatrix}\sigma(0)\\ w(0)\end{pmatrix}\right\|_0^2 \le Ce^{2\lambda t}\,\mathfrak{E}\big(\sigma(0), w(0)\big). \tag{3-47}$$

Using these quantities and estimate (3-47), we can record estimates for solutions to (3-38)–(3-39).

Theorem 3.5. *Suppose that σ and w solve the inhomogeneous linear problem* (3-38)–(3-39). *Let ψ be given by Lemma 3.4 and $\bar{w} = w - \psi$. Let $\|\cdot\|_0$ and $\mathfrak{E}(\cdot\,,\cdot)$ be given by* (3-45) *and* (3-46), *respectively. Then*

$$\left\|\begin{pmatrix}\sigma(t)\\ w(t)\end{pmatrix} - e^{t\mathscr{L}}\begin{pmatrix}\sigma(0)\\ \bar{w}(0)\end{pmatrix}\right\|_0 \le \int_0^t Ce^{\lambda(t-s)}\sqrt{\mathfrak{E}(N_1(s), N_2(s))}\,ds + \frac{C}{\delta}|N_{\mathscr{B}}(t)| + \frac{C}{\delta}\int_0^t e^{\lambda(t-s)}\big(|N_{\mathscr{B}}(s)| + |\partial_t N_{\mathscr{B}}(s)| + |\partial_t^2 N_{\mathscr{B}}(s)|\big)\,ds. \tag{3-48}$$

Proof. From the above analysis, we know that σ and w are given by (3-44), where $e^{t\mathscr{L}}$ is the homogeneous solution operator given by (3-37). Hence (3-47) implies that

$$\left\|\begin{pmatrix}\sigma(t)\\ w(t)\end{pmatrix} - e^{t\mathscr{L}}\begin{pmatrix}\sigma(0)\\ \bar{w}(0)\end{pmatrix}\right\|_0 \le \int_0^t Ce^{\lambda(t-s)}\sqrt{\mathfrak{E}(N_1(s), N_2(s))}\,ds + \frac{1}{\delta}\left\|\begin{pmatrix}0\\ \frac{1}{3}N_{\mathscr{B}}(t)r_0^3\end{pmatrix}\right\|_0 + \frac{1}{\delta}\int_0^t Ce^{\lambda(t-s)}\sqrt{\mathfrak{E}\big(N_{\mathscr{B}}(s)\rho_0, \tfrac{1}{3}\partial_t N_{\mathscr{B}}(s)r_0^3\big)}\,ds. \tag{3-49}$$

Then, since $N_{\mathscr{B}}(t)$ is only a function of time, not of x, we can easily estimate

$$\left\|\begin{pmatrix}0\\ \frac{1}{3}N_{\mathscr{B}}(t)r_0^3\end{pmatrix}\right\|_0 \le C|N_{\mathscr{B}}(t)| \tag{3-50}$$

and

$$\sqrt{\mathfrak{E}\big(N_{\mathcal{B}}(s)\rho_0,\, \tfrac{1}{3}\partial_t N_{\mathcal{B}}(s) r_0^3\big)} \le C\big(|N_{\mathcal{B}}(s)| + |\partial_t N_{\mathcal{B}}(s)| + |\partial_t^2 N_{\mathcal{B}}(s)|\big), \tag{3-51}$$

where $C > 0$ in (3-50)–(3-51) is a constant depending on various (finite) integrals of ρ_0 and r_0. The estimate (3-48) then follows by combining (3-49)–(3-51). ☐

4. Nonlinear energy estimates

4A. ***Definitions.*** We are interested in small perturbations σ, v around the stationary solution $\rho = \rho_0$, $r = r_0$, and $v = 0$. In particular, we assume that

$$\tfrac{9}{10}\rho_0 \le \rho_0 + \sigma \le \tfrac{11}{10}\rho_0. \tag{4-1}$$

This assumption will be justified later when we close the nonlinear energy estimates. For such small solutions, the Navier–Stokes–Poisson system (1-17) and (1-18) can be written as follows:

$$\begin{gathered}\partial_t \sigma + 4\pi\rho^2 \partial_x (r^2 v) = 0,\\ \partial_t v + 4\pi r^2 \partial_x P - 4\pi r_0^2 \partial_x P_0 + \frac{x}{r^2} - \frac{x}{r_0^2} = 16\pi^2 r^2 \partial_x \left(\left(\frac{4\varepsilon}{3} + \delta \right) \rho \partial_x (r^2 v) \right).\end{gathered} \tag{4-2}$$

The dynamics of r are determined by

$$r(x,t) = \left(\frac{3}{4\pi} \int_0^x \frac{dy}{\rho_0(y) + \sigma(y,t)} \right)^{1/3} \quad \text{and} \quad \partial_t r(x,t) = v(x,t). \tag{4-3}$$

It turns out that it is convenient to analyze $\dfrac{\sigma}{\rho_0}$ rather than σ itself, so we rewrite the continuity equation as

$$\frac{\rho_0}{\rho} \partial_t \left(\frac{\sigma}{\rho_0} \right) + 4\pi\rho \partial_x (r^2 v) = 0. \tag{4-4}$$

We will also rewrite the momentum equation. To do so, we first note that

$$4\pi r^2 \partial_x P - 4\pi r_0^2 \partial_x P_0 + \frac{x}{r^2} - \frac{x}{r_0^2} = 4\pi r^2 \partial_x (P - P_0) + x \left(\frac{1}{r^2} - \frac{r^2}{r_0^4} \right), \tag{4-5}$$

and then note that for small perturbations satisfying (4-1), $P - P_0 = K(\rho^\gamma - \rho_0^\gamma)$ can be written as

$$P - P_0 = \rho_0^\gamma \left\{ K\gamma \frac{\sigma}{\rho_0} + a_* \left(\frac{\sigma}{\rho_0} \right)^2 \right\}, \tag{4-6}$$

where a_* is the smooth bounded remainder from the Taylor's theorem. We then rewrite the momentum equation as

$$\partial_t v + 4\pi r^2 \partial_x \left\{ K\gamma \rho_0^\gamma \frac{\sigma}{\rho_0} + a_* \rho_0^\gamma \left(\frac{\sigma}{\rho_0} \right)^2 \right\} + x \left(\frac{1}{r^2} - \frac{r^2}{r_0^4} \right) = \mathcal{V}, \tag{4-7}$$

where

$$\mathcal{V} := 16\pi^2 r^2 \partial_x\biggl(\Bigl(\frac{4\varepsilon}{3}+\delta\Bigr)\rho\,\partial_x(r^2 v)\biggr).$$

We give an equivalent expression for $\mathcal{V}$ so that it appreciates the boundary condition (1-20) in energy estimates:

$$\mathcal{V} = 16\pi^2 r^2 \partial_x \mathcal{W} + \frac{4\varepsilon}{3} 12\pi r^2 \partial_x\Bigl(\frac{v}{r}\Bigr), \tag{4-8}$$

where

$$\mathcal{W} = \delta\rho\,\partial_x(r^2 v) + \frac{4\varepsilon}{3}\rho r^3 \partial_x\Bigl(\frac{v}{r}\Bigr) \tag{4-9}$$

satisfies $\mathcal{W}(M) = 0$ because of the boundary condition (1-20). We use ν to denote the minimal viscosity coefficient:

$$\nu := \min\Bigl\{\delta, \frac{4\varepsilon}{3}\Bigr\}. \tag{4-10}$$

We now define instant energy functionals for σ and v. In what follows, all of the integrals are understood to be over the interval $[0, M]$.

$$\begin{aligned}
\mathcal{E}^0 &:= \frac{1}{2}\int |v|^2\,dx + \frac{1}{2}\int \frac{K\gamma\rho_0^{\gamma-1}}{\Bigl(1+\dfrac{\sigma}{\rho_0}\Bigr)^2}\Bigl|\frac{\sigma}{\rho_0}\Bigr|^2 dx + \frac{1}{2}\int \nu\Bigl|1-\frac{r_0}{r}\Bigr|^2 dx \\
&=: \mathcal{E}^{0,v} + \mathcal{E}^{0,\sigma} + \mathcal{E}^{0,r}, \\
\mathcal{E}^1 &:= \frac{1}{2}\biggl[\delta\int 16\pi^2\rho|\partial_x(r^2 v)|^2\,dx + \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\Bigl|\partial_x\Bigl(\frac{v}{r}\Bigr)\Bigr|^2 dx\biggr], \\
&\qquad + \frac{1}{2}\int\Bigl(\delta+\frac{4\varepsilon}{3}\Bigr)16\pi^2 r^4 \frac{1}{1+\dfrac{\sigma}{\rho_0}}\Bigl|\partial_x\Bigl(\frac{\sigma}{\rho_0}\Bigr)\Bigr|^2 dx \\
&=: \mathcal{E}^{1,v} + \mathcal{E}^{1,\sigma}, \qquad\qquad (4\text{-}11) \\
\mathcal{E}^2 &:= \frac{1}{2}\int |\partial_t v|^2\,dx + \frac{1}{2}\int \frac{K\gamma\rho_0^{\gamma-1}}{\Bigl(1+\dfrac{\sigma}{\rho_0}\Bigr)^2}\Bigl|\partial_t\Bigl(\frac{\sigma}{\rho_0}\Bigr)\Bigr|^2 dx \\
&=: \mathcal{E}^{2,v} + \mathcal{E}^{2,\sigma}, \\
\mathcal{E}^3 &:= \frac{1}{2}\biggl[\delta\int 16\pi^2\rho|\partial_x(r^2\partial_t v)|^2 dx + \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\Bigl|\partial_x\Bigl(\frac{\partial_t v}{r}\Bigr)\Bigr|^2 dx\biggr], \\
\mathcal{E}^4 &:= \frac{1}{2}\int\Bigl(\delta+\frac{4\varepsilon}{3}\Bigr)4\pi\rho_0\Bigl|\partial_x\Bigl(r^4\partial_x\Bigl(\frac{\sigma}{\rho_0}\Bigr)\Bigr)\Bigr|^2 dx.
\end{aligned}$$

The corresponding dissipations are given by

$$\begin{aligned}
\mathscr{D}^0 &:= \delta\int 16\pi^2\rho|\partial_x(r^2v)|^2\,dx + \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\left|\partial_x\left(\frac{v}{r}\right)\right|^2 dx,\\
\mathscr{D}^1 &:= \int|\partial_t v|^2\,dx + \int 16\pi^2K\gamma r^4\rho_0^\gamma\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx := \mathscr{D}^{1,v}+\mathscr{D}^{1,\sigma},\\
\mathscr{D}^2 &:= \delta\int 16\pi^2\rho|\partial_x(r^2\partial_t v)|^2\,dx + \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\left|\partial_x\left(\frac{\partial_t v}{r}\right)\right|^2 dx,\\
\mathscr{D}^3 &:= \int|\partial_t^2 v|^2\,dx,\\
\mathscr{D}^4 &:= \int 4\pi K\gamma r^2\rho\rho_0^\gamma\left|\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)\right|^2 dx.
\end{aligned}\tag{4-12}$$

We note that $\mathscr{E}^0$ in (4-11) corresponds to the physical energy given in (1-31) and $\mathscr{D}^0$ is the corresponding dissipation. $\mathscr{E}^1$ is the energy for the first spatial derivatives of v and σ and its structure comes from the viscosity term (for instance, see (4-8)), and $\mathscr{D}^1$ is the corresponding dissipation. $\mathscr{E}^2$ and $\mathscr{E}^3$ are the temporally higher-order energies of $\mathscr{E}^0$ and $\mathscr{E}^1$. $\mathscr{E}^4$ is the energy for the second derivative of σ and its form is closely related to the structure of the Navier–Stokes–Poisson system (4-2), which can be seen in (4-79).

In addition, we introduce various bootstrapped and auxiliary energies and dissipations (denoted by subscripts b and a, respectively) that can be controlled with the above instant energies and dissipations:

$$\begin{aligned}
\mathscr{E}_b^{0,r} &:= \int\frac{\nu}{\rho}\left|1-\frac{r_0}{r}\right|^2 dx,\\
\mathscr{E}_b^{1,\sigma} &:= \frac{1}{2}\int\left(\delta+\frac{4\varepsilon}{3}\right)16\pi^2\frac{r^2}{\rho}\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx,\\
\mathscr{D}_b^{1,\sigma} &:= \int 16\pi^2K\gamma r^2\rho_0^{\gamma-1}\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx.
\end{aligned}\tag{4-13}$$

We note that these bootstrapped energies and dissipations have similar structure to the ones without subscript b, but have the stronger weights $1/\rho$ because ρ vanishes at $x=M$. The control of them will allow us to have the estimates with improved weights, and it will also be helpful to obtain the higher-order estimates. The following auxiliary energies are motivated by the structure of the higher-order derivatives of the equations in (4-2); for instance, see (4-45), (4-75) and (4-79).

$$\begin{aligned}
\mathscr{E}_a^{3,\sigma} &:= \int\left(\delta+\frac{4\varepsilon}{3}\right)^2 16\pi^2\frac{r^2}{\rho}\left|\partial_x\partial_t\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx,\\
\mathscr{E}_{a_1}^{3,v} &:= \int\frac{r^2}{\rho}\left|\partial_x(\rho\partial_x(r^2v))\right|^2 dx,\\
\mathscr{E}_{a_2}^{3,v} &:= \int\rho r^6\left|\partial_x\left(\rho r^3\partial_x\left(\frac{v}{r}\right)\right)\right|^2 dx,
\end{aligned}\tag{4-14}$$

$$\begin{aligned} \mathscr{E}^4_{a_1} &:= \int \Big(\delta + \frac{4\varepsilon}{3}\Big) 4\pi\rho_0 \left|\partial_x\Big(r^4\partial_t\partial_x\Big(\frac{\sigma}{\rho_0}\Big)\Big)\right|^2 dx, \\ \mathscr{E}^4_{a_2} &:= \int 16\pi^2\rho_0 \big|\partial_x\big(r^4\partial_x(\rho\partial_x(r^2 v))\big)\big|^2\, dx. \end{aligned} \tag{4-15}$$

Finally, we introduce some bootstrap energies that depend on a parameter $\beta \in \mathbb{R}$:

$$\mathscr{E}^{0,\sigma}_{\beta} := \int \frac{\rho_0^{\beta+1}}{\rho}\left|\frac{\sigma}{\rho_0}\right|^2 dx, \quad \mathscr{E}^{2,\sigma}_{\beta} := \int \frac{\rho_0^{\beta+1}}{\rho}\left|\partial_t\Big(\frac{\sigma}{\rho_0}\Big)\right|^2 dx. \tag{4-16}$$

For the proof of our instability in Section 5, we will need to invoke higher-order energy functionals and dissipations, which are the higher-order generalizations of the above energies and dissipations. For $i = 2$ and 3, let

$$\begin{aligned} \mathscr{E}^{1+2i} &:= \frac{1}{2}\int |\partial_t^i v|^2 dx + \frac{1}{2}\int \frac{K\gamma\rho_0^{\gamma-1}}{\Big(1+\dfrac{\sigma}{\rho_0}\Big)^2}\left|\partial_t^i\Big(\frac{\sigma}{\rho_0}\Big)\right|^2 dx, \\ \mathscr{D}^{1+2i} &:= \delta\int 16\pi^2\rho|\partial_x(r^2\partial_t^i v)|^2\, dx + \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\left|\partial_x\Big(\frac{\partial_t^i v}{r}\Big)\right|^2 dx, \\ \mathscr{E}^{2+2i} &:= \frac{1}{2}\bigg[\delta\int 16\pi^2\rho|\partial_x(r^2\partial_t^i v)|^2\, dx + \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\left|\partial_x\Big(\frac{\partial_t^i v}{r}\Big)\right|^2 dx\bigg], \\ \mathscr{D}^{2+2i} &:= \int |\partial_t^{i+1} v|^2\, dx. \end{aligned} \tag{4-17}$$

Next we define bootstrapped energies and auxiliary energies for $i = 2$ and 3:

$$\begin{aligned} \mathscr{E}^{1+2i,\sigma}_{-1} &:= \int \frac{1}{\rho_0}\left|\partial_t^i\Big(\frac{\sigma}{\rho_0}\Big)\right|^2 dx, \quad \mathscr{E}^{1+2i,v}_{a} := \int r^4\big|\partial_x\big(\rho\partial_x(\partial_t^{i-1}[r^2 v])\big)\big|^2\, dx, \\ \mathscr{E}^{2+2i,\sigma}_{a} &:= \int \Big(\delta + \frac{4\varepsilon}{3}\Big)^2 16\pi^2 r^2\left|\partial_x\partial_t^i\Big(\frac{\sigma}{\rho_0}\Big)\right|^2 dx. \end{aligned} \tag{4-18}$$

We then define the total energy by

$$\begin{aligned} \mathscr{E} := \sum_{i=0}^{8}\mathscr{E}^i + \mathscr{E}^{0,r}_b + \mathscr{E}^{1,\sigma}_b + \mathscr{E}^{0,\sigma}_{-1} + \mathscr{E}^{2,\sigma}_{-1} + \mathscr{E}^{5,\sigma}_{-1} + \mathscr{E}^{7,\sigma}_{-1} \\ + \mathscr{E}^{3,\sigma}_a + \mathscr{E}^{3,v}_{a_1} + \mathscr{E}^{3,v}_{a_2} + \mathscr{E}^4_{a_1} + \mathscr{E}^4_{a_2} + \mathscr{E}^{5,\sigma}_a + \mathscr{E}^{6,\sigma}_a + \mathscr{E}^{7,\sigma}_a + \mathscr{E}^{8,\sigma}_a. \end{aligned} \tag{4-19}$$

The introduction of the above notation for the energies and dissipations is lengthy, but at each level they capture the complex structure of the Navier–Stokes–Poisson system with degeneracy of ρ at $x = M$ and r at $x = 0$, and lead to successful energy estimates. We have separated the energies from one another because the estimate of each energy term in $\mathscr{E}$ will be derived by a different strategy and method.

Throughout the rest of the section, we assume that

$$\begin{aligned}\left\|\frac{\sigma}{\rho_0}\right\|_{L^\infty}+\left\|\partial_t\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty}+\left\|\partial_t^2\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty}+\left\|\partial_t^3\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty}+\left\|1-\frac{r_0}{r}\right\|_{L^\infty}&\\ +\left\|\rho r^3\partial_x\left(\frac{v}{r}\right)\right\|_{L^\infty}+\left\|\rho r^3\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty}+\left\|\frac{v}{r}\right\|_{L^\infty}+\left\|\frac{\partial_t v}{r}\right\|_{L^\infty}+\left\|\frac{\partial_t^2 v}{r}\right\|_{L^\infty}&\le\theta_1\end{aligned}\tag{4-20}$$

for sufficiently small constant θ_1, where the norm $\|\cdot\|_{L^\infty}$ is over the spatial region $[0, M]$. The validity of this assumption within the total energy $\mathscr{E}$ will be justified in Lemma 4.9.

Since r is determined through an integral of σ as in (4-3), for small perturbations satisfying (4-1) we may use Taylor's theorem to write r_0/r as

$$\frac{r_0}{r}=1+\frac{1}{4\pi r_0^3}\int_0^x\frac{\sigma}{\rho_0^2}\,dy+\frac{c_1}{r_0^3}\int_0^x\frac{1}{\rho_*}\left(\frac{\sigma}{\rho_0}\right)^2dy+\frac{c_2}{r_0^6}\left(\int_0^x\frac{\sigma}{\rho_0^2}\,dy\right)^2,\tag{4-21}$$

where $\rho_*/\rho_0\sim 1$ is a bounded smooth function of σ/ρ_0. Hence the $1-r_0/r$ estimate (up to a constant) in (4-20) can actually be guaranteed by the smallness of the other terms in (4-20).

The relation (4-21) will be useful in various places. We now record a couple other useful identities.

Dynamics of r_0/r. From (4-3), we have

$$\begin{aligned}\partial_t\left(\frac{r_0}{r}\right)&=-\frac{r_0 v}{r^2}=-\left(\frac{v}{r}\right)\left(\frac{r_0}{r}\right),\\ \partial_x\left(\frac{r_0}{r}\right)&=\frac{1}{4\pi\rho_0 r_0^2 r}-\frac{r_0}{4\pi\rho r^4}=\frac{1-\left(\dfrac{r_0}{r}\right)^3+\dfrac{\sigma}{\rho_0}}{4\pi\rho r_0^2 r}.\end{aligned}\tag{4-22}$$

Some useful inequalities and identities. For any v (not just solutions),

$$\begin{aligned}&\frac{v}{r}=\frac{4\pi}{3}\left\{\rho\partial_x(r^2v)-\rho r^3\partial_x\left(\frac{v}{r}\right)\right\}\implies\frac{v^2}{\rho r^2}\le\frac{32\pi^2}{9}\left\{\rho|\partial_x(r^2v)|^2+\rho r^6\left|\partial_x\left(\frac{v}{r}\right)\right|^2\right\},\\ &\rho\partial_x(rv^2)=\rho\partial_x\left[(r^2v)^2\cdot\frac{1}{r^3}\right]=2\frac{v}{r}\rho\partial_x(r^2v)-\frac{3}{4\pi}\frac{v^2}{r^2}=\frac{v}{r}\left\{\rho\partial_x(r^2v)+\rho r^3\partial_x\left(\frac{v}{r}\right)\right\}.\end{aligned}\tag{4-23}$$

4B. ***Estimates.*** Throughout the rest of the section, we use C to denote a generic constant that may differ from line to line, and η to denote a sufficiently small fixed constant which will be determined later. The constants C are allowed to depend on η, which presents no trouble in our ultimate analysis since first we will fix an η, which then fixes the constants.

In the following series of lemmas, we provide the energy inequalities for $\mathscr{E}$. We present them in the order that we use for the bootstrap argument in Section 5A. Here is the flowchart for the estimates:

$$\begin{aligned}\mathscr{E}^0 &\to \mathscr{E}_\beta^{0,\sigma} \to \mathscr{E}^1 \to \mathscr{E}^2 \to \mathscr{E}_\beta^{2,\sigma} \to \mathscr{E}^3\\ &\to \mathscr{E}_b^{0,r} \to \mathscr{E}_b^{1,\sigma} \to \mathscr{E}_a^{3,\sigma} \to \mathscr{E}_{a_1}^{3,v} \to \mathscr{E}_{a_2}^{3,v}\\ &\to \mathscr{E}^4 \to \mathscr{E}_{a_1}^4 \to \mathscr{E}_{a_2}^4 \to \mathscr{E}^5 \to \mathscr{E}_a^5 \to \mathscr{E}_{-1}^{5,\sigma}\\ &\to \mathscr{E}^6 \to \mathscr{E}_a^{6,\sigma} \to \mathscr{E}^7 \to \mathscr{E}_a^7 \to \mathscr{E}_{-1}^{7,\sigma} \to \mathscr{E}^8 \to \mathscr{E}_a^{8,\sigma}.\end{aligned} \tag{4-24}$$

We start with $\mathscr{E}^0$ and $\mathscr{D}^0$.

Lemma 4.1.
$$\frac{d}{dt}\mathscr{E}^0 + \mathscr{D}^0 \le C(1+\theta_1)\mathscr{E}^0 + \frac{1}{2}\mathscr{D}^0. \tag{4-25}$$

Proof. Multiply (4-7) by v and integrate to get

$$\frac{1}{2}\frac{d}{dt}\int |v|^2\,dx \underbrace{-\int 4\pi\,\partial_x(r^2v)\left\{K\gamma\rho_0^\gamma\frac{\sigma}{\rho_0} + a_*\rho_0^\gamma\left(\frac{\sigma}{\rho_0}\right)^2\right\}dx}_{\text{(i)}} + \underbrace{\int v\frac{x(r_0^4-r^4)}{r^2r_0^4}\,dx}_{\text{(ii)}} = \underbrace{\int v\mathscr{V}\,dx}_{\text{(iii)}}. \tag{4-26}$$

For (i), we use (4-4) to see that

$$\begin{aligned}\text{(i)} &= \int \frac{\rho_0}{\rho^2}\partial_t\left(\frac{\sigma}{\rho_0}\right)\left\{K\gamma\rho_0^\gamma\frac{\sigma}{\rho_0} + a_*\rho_0^\gamma\left(\frac{\sigma}{\rho_0}\right)^2\right\}dx\\ &= \frac{1}{2}\frac{d}{dt}\int \frac{K\gamma\rho_0^{\gamma-1}}{\left(1+\frac{\sigma}{\rho_0}\right)^2}\left|\frac{\sigma}{\rho_0}\right|^2 dx + \int \frac{\left(K\gamma + \left(1+\frac{\sigma}{\rho_0}\right)a_*\right)\rho_0^{\gamma-1}}{\left(1+\frac{\sigma}{\rho_0}\right)^3}\left|\frac{\sigma}{\rho_0}\right|^2\partial_t\left(\frac{\sigma}{\rho_0}\right)dx.\end{aligned} \tag{4-27}$$

However,

$$\left|\int \frac{\left(K\gamma + \left(1+\frac{\sigma}{\rho_0}\right)a_*\right)\rho_0^{\gamma-1}}{\left(1+\frac{\sigma}{\rho_0}\right)^3}\left|\frac{\sigma}{\rho_0}\right|^2\partial_t\left(\frac{\sigma}{\rho_0}\right)dx\right| \le C(1+\theta_1)\mathscr{E}^0. \tag{4-28}$$

For (ii), the Cauchy–Schwarz inequality yields

$$|\text{(ii)}| \le \nu\int\frac{|v|^2}{\rho r^2}\,dx + \frac{1}{\nu}\int\rho\left|\frac{x}{r_0^4}(r^2+r_0^2)(r+r_0)\right|^2\left|1-\frac{r_0}{r}\right|^2dx \le \frac{2}{9}\mathscr{D}^0 + C\mathscr{E}^{0,r}, \tag{4-29}$$

where we have used (4-23) at the second inequality. From (4-8) and the boundary condition $\mathscr{W}(M,t)=0$, we get

$$\text{(iii)} = -\delta\int 16\pi^2\rho|\partial_x(r^2v)|^2\,dx - \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\left|\partial_x\left(\frac{v}{r}\right)\right|^2dx = -\mathscr{D}^0. \tag{4-30}$$

Next, from (4-22),

$$\begin{aligned}\frac{\nu}{2}\frac{d}{dt}\int\left|1-\frac{r_0}{r}\right|^2 dx &= -\nu\int\frac{v}{r}\frac{r_0}{r}\left(1-\frac{r_0}{r}\right)dx\\ &\le \nu\int\frac{|v|^2}{\rho r^2}\,dx+\nu\int\rho\left|\frac{r_0}{r}\right|^2\left|1-\frac{r_0}{r}\right|^2dx \le \frac{2}{9}\mathscr{D}^0+C\mathscr{E}^{0,r}.\end{aligned} \tag{4-31}$$

The desired estimate then follows by combining these estimates. □

With Lemma 4.1, we can bootstrap to control $\frac{\sigma}{\rho_0}$ with an improved weight. Multiply (4-4) by $\rho_0^{\beta}\frac{\sigma}{\rho_0}$ and integrate to get

$$\int\frac{\rho_0^{\beta+1}}{\rho}\frac{\sigma}{\rho_0}\partial_t\left(\frac{\sigma}{\rho_0}\right)dx = -\int\rho^{1/2}\rho_0^{\beta}\frac{\sigma}{\rho_0}\cdot 4\pi\rho^{1/2}\partial_x(r^2v)\,dx. \tag{4-32}$$

Thus

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int\frac{\rho_0^{\beta+1}}{\rho}\left|\frac{\sigma}{\rho_0}\right|^2dx\\ &\qquad\le\frac{C}{\eta}\int 16\pi^2\rho|\partial_x(r^2v)|^2\,dx+\eta\int\rho\rho_0^{2\beta}\left|\frac{\sigma}{\rho_0}\right|^2dx-\frac{1}{2}\int\frac{\rho_0^{\beta+1}}{\rho^2}\partial_t\sigma\left|\frac{\sigma}{\rho_0}\right|^2dx,\end{aligned} \tag{4-33}$$

which means that

$$\frac{d}{dt}\mathscr{E}_{\beta}^{0,\sigma}\le\frac{C}{\eta}\mathscr{D}^0+\eta\mathscr{E}_{2\beta+1}^{0,\sigma}+C\theta_1\mathscr{E}_{\beta}^{0,\sigma}. \tag{4-34}$$

Next we consider $\mathscr{E}^1$ and $\mathscr{D}^1$.

Lemma 4.2. *We have*

$$\frac{d}{dt}\mathscr{E}^1+\mathscr{D}^1\le(\eta+C\theta_1)\mathscr{E}^1+\frac{1}{2}\mathscr{D}^1+C(\mathscr{E}^0+\mathscr{E}_0^{0,\sigma})+q\mathscr{E}^{2,v}, \tag{4-35}$$

where

$$q:=q_1+q_2:=\left\|16\pi^2\left(K\gamma+\frac{4}{K\gamma}a_*^2\left(\frac{\sigma}{\rho_0}\right)^2\right)\rho_0^{\gamma}\right\|_{L^\infty}+\left\|\frac{\left(1+\frac{\sigma}{\rho_0}\right)}{\eta\left(\delta+\frac{4\varepsilon}{3}\right)}\right\|_{L^\infty} \tag{4-36}$$

is bounded due to (4-20).

Proof. We divide the proof into steps.

Step 1 ($\mathscr{E}^{1,v}$ and $\mathscr{D}^{1,v}$). Multiply (4-7) by $\partial_t v$ and integrate to get

$$\int|\partial_t v|^2\,dx+\underbrace{\int 4\pi r^2\partial_t v\,\partial_x\left\{K\gamma\rho_0^{\gamma}\frac{\sigma}{\rho_0}+a_*\rho_0^{\gamma}\left(\frac{\sigma}{\rho_0}\right)^2\right\}dx}_{\text{(iv)}}+\underbrace{\int\partial_t v\frac{x(r_0^4-r^4)}{r^2r_0^4}\,dx}_{\text{(v)}}-\underbrace{\int\partial_t v\mathscr{V}\,dx}_{\text{(vi)}}=0.$$

For (iv), we first expand

$$\begin{aligned}(\text{iv}) &= \int 4\pi r^2 \partial_t v\left\{K\gamma\partial_x(\rho_0^\gamma)\frac{\sigma}{\rho_0} + K\gamma\rho_0^\gamma\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\}dx \\ &\quad + \int 4\pi r^2\partial_t v\left\{\partial_x(a_*\rho_0^\gamma)\left(\frac{\sigma}{\rho_0}\right)^2 + 2a_*\rho_0^\gamma\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\}dx \\ &=: (\text{iv})_1 + (\text{iv})_2 + (\text{iv})_3 + (\text{iv})_4, \end{aligned} \tag{4-37}$$

and then estimate

$$\left|(\text{iv})_1 + (\text{iv})_3\right| \le \frac{1}{4}\int|\partial_t v|^2\,dx + C(1+\theta_1^2)\int\left|\frac{\sigma}{\rho_0}\right|^2 dx \le \frac{1}{4}\mathscr{D}^{1,v} + C\mathscr{E}_0^{0,\sigma} \tag{4-38}$$

and

$$\begin{aligned}\left|(\text{iv})_2 + (\text{iv})_4\right| &\le \int 8\pi^2 K\gamma r^4\rho_0^\gamma\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx + \int 8\pi^2\left(K\gamma + \frac{4}{K\gamma}a_*^2\left(\frac{\sigma}{\rho_0}\right)^2\right)\rho_0^\gamma|\partial_t v|^2\,dx \\ &\le \tfrac{1}{2}\mathscr{D}^{1,\sigma} + q_1\mathscr{E}^{2,v}. \end{aligned} \tag{4-39}$$

For (v), we get

$$|(\text{v})| \le \frac{1}{4}\int|\partial_t v|^2\,dx + \int\left|\frac{x}{r^2 r_0^4}(r^2+r_0^2)(r+r_0)\right|^2|r-r_0|^2\,dx \le \tfrac{1}{4}\mathscr{D}^{1,v} + C\mathscr{E}^{0,r}. \tag{4-40}$$

The term (vi) forms the energy $\mathscr{E}^{1,v}$ and nonlinear commutators:

$$\begin{aligned}(\text{vi}) &= -\frac{1}{2}\frac{d}{dt}\left[\delta\int 16\pi^2\rho|\partial_x(r^2v)|^2\,dx + \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\left|\partial_x\left(\frac{v}{r}\right)\right|^2 dx\right] \\ &\quad + \frac{1}{2}\left[\delta\int 16\pi^2\partial_t\sigma|\partial_x(r^2v)|^2\,dx + \frac{4\varepsilon}{3}\int 16\pi^2\partial_t(\rho r^6)\left|\partial_x\left(\frac{v}{r}\right)\right|^2 dx\right] \\ &\quad + \delta\int 16\pi^2\partial_x(v\cdot 2rv)\rho\partial_x(r^2v)\,dx + \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\partial_x\left(v\cdot\left(-\frac{v}{r^2}\right)\right)\partial_x\left(\frac{v}{r}\right)dx. \end{aligned} \tag{4-41}$$

Using (4-20) and the fact that

$$\partial_x(v\cdot 2rv) = 2\left(\frac{v}{r}\partial_x(r^2v) + r^2v\partial_x\left(\frac{v}{r}\right)\right) \tag{4-42}$$

and

$$\partial_x\left(v\cdot\left(-\frac{v}{r^2}\right)\right)\partial_x\left(\frac{v}{r}\right) = -2\frac{v}{r}\left|\partial_x\left(\frac{v}{r}\right)\right|^2, \tag{4-43}$$

the absolute values of the second and third lines may be bounded by $C\theta_1\mathscr{E}^{1,v}$.

We may now combine the above to deduce that

$$\frac{d}{dt}\mathscr{E}^{1,v} + \mathscr{D}^{1,v} \le \tfrac{1}{2}\mathscr{D}^1 + C\theta_1\mathscr{E}^{1,v} + C\mathscr{E}^{0,r} + C\mathscr{E}_0^{0,\sigma} + q_1\mathscr{E}^{2,v}. \tag{4-44}$$

Step 2 ($\mathscr{E}^{1,\sigma}$ and $\mathscr{D}^{1,\sigma}$). For the estimate of $\partial_x(\sigma/\rho_0)$, we first rewrite (4-7) by replacing $\rho\partial_x(r^2v)$ in $\mathscr{V}$ by $\partial_t(\sigma/\rho_0)$ through the continuity equation (4-4):

$$\begin{aligned}&\left(\delta+\frac{4\varepsilon}{3}\right)4\pi r^2\left\{\frac{\rho_0}{\rho}\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right)+\partial_x\left(\frac{\rho_0}{\rho}\right)\partial_t\left(\frac{\sigma}{\rho_0}\right)\right\}+\partial_t v+\frac{x(r_0^4-r^4)}{r^2r_0^4}\\&\qquad+4\pi r^2\left\{K\gamma\rho_0^\gamma\partial_x\left(\frac{\sigma}{\rho_0}\right)+K\gamma\partial_x\left(\rho_0^\gamma\right)\frac{\sigma}{\rho_0}+\partial_x(a_*\rho_0^\gamma)\left(\frac{\sigma}{\rho_0}\right)^2+2a_*\rho_0^\gamma\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\}=0.\end{aligned}\tag{4-45}$$

Note that

$$\partial_x\left(\frac{\rho_0}{\rho}\right)=-\left(1+\frac{\sigma}{\rho_0}\right)^{-2}\partial_x\left(\frac{\sigma}{\rho_0}\right).$$

Multiplying (4-45) by $4\pi r^2\partial_x\left(\frac{\sigma}{\rho_0}\right)$ and integrating, we are led to the estimate

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int\left(\delta+\frac{4\varepsilon}{3}\right)16\pi^2r^4\frac{1}{1+\dfrac{\sigma}{\rho_0}}\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2dx+\int 16\pi^2K\gamma r^4\rho_0^\gamma\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2dx\\&\qquad\qquad\le(\eta+C\theta_1)\mathscr{E}^{1,\sigma}+C\left(\mathscr{E}^{0,r}+\mathscr{E}_0^{0,\sigma}\right)+\int\frac{1+\dfrac{\sigma}{\rho_0}}{2\eta\left(\delta+\dfrac{4\varepsilon}{3}\right)}|\partial_t v|^2\,dx.\end{aligned}\tag{4-46}$$

Note that the last term in (4-46) may be bounded by $q_2\mathscr{E}^{2,v}$. We then obtain (4-35) by combining (4-44) and (4-46). ☐

The estimate (4-35) is not of a closed form by itself. Its use will be apparent when it is coupled with the result of the following lemma.

Lemma 4.3.
$$\frac{d}{dt}\mathscr{E}^2+\mathscr{D}^2\le(\eta+C\theta_1)\mathscr{E}^2+C\theta_1\mathscr{D}^0+\left(\frac{1}{4}+C\theta_1\right)\mathscr{D}^2+C\left(\theta_1\mathscr{E}_0^{0,\sigma}+\mathscr{E}^0+\mathscr{E}^1\right).\tag{4-47}$$

Proof. We take ∂_t of (4-7) to see that

$$\begin{aligned}&\partial_t^2v+4\pi r^2\partial_x\left\{K\gamma\rho_0^\gamma\partial_t\left(\frac{\sigma}{\rho_0}\right)+2a_*\rho_0^\gamma\frac{\sigma}{\rho_0}\partial_t\left(\frac{\sigma}{\rho_0}\right)+\partial_t a_*\rho_0^\gamma\left(\frac{\sigma}{\rho_0}\right)^2\right\}\\&\qquad+8\pi rv\left\{K\gamma\rho_0^\gamma\partial_x\left(\frac{\sigma}{\rho_0}\right)+K\gamma\partial_x\left(\rho_0^\gamma\right)\frac{\sigma}{\rho_0}+\partial_x(a_*\rho_0^\gamma)\left(\frac{\sigma}{\rho_0}\right)^2+2a_*\rho_0^\gamma\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\}\\&\qquad\qquad-2\frac{xv(r_0^4-r^4)}{r^3r_0^4}-4\frac{xrv}{r_0^4}=\partial_t\mathscr{V}.\end{aligned}\tag{4-48}$$

The energy estimate (4-47) may be derived from (4-48) as in Lemma 4.1: we multiply (4-48) by $\partial_t v$, integrate over $x\in[0,M]$, and integrate various terms by parts in order to identify $d\mathscr{E}^2/dt$, $\mathscr{D}^2$, and some error (lower-order or commutator) terms, the latter of which may be estimated by the right side of (4-47). Since the argument is essentially the same as that of Lemma 4.1, we present only a sketch.

The product of $\partial_t v$ with the first two terms in the first line in (4-48) forms the energy term $\partial_t\mathscr{E}^2$ and some error terms:

$$\int \partial_t v\left[\partial_t^2 v + 4\pi r^2\partial_x\left(K\gamma\rho_0^{\gamma}\partial_t\left(\frac{\sigma}{\rho_0}\right)\right)\right]dx = \frac{1}{2}\frac{d}{dt}\left\{\int |\partial_t v|^2\,dx + \int \frac{K\gamma\rho_0^{\gamma-1}}{\left(1+\frac{\sigma}{\rho_0}\right)^2}\left|\partial_t\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx\right\} + \mathscr{X},$$

where $\mathscr{X}$ is a term whose absolute value may be estimated by the right side of (4-47). Here we have used the continuity equation (4-4) and an integration by parts on the second term.

Next, we compute

$$\partial_t\mathcal{V} = 16\pi^2 r^2\partial_x\partial_t\mathcal{W} + 16\pi^2(2rv)\partial_x\mathcal{W} + \frac{4\varepsilon}{3}12\pi r^2\partial_x\left(\frac{\partial_t v}{r}\right) \tag{4-49}$$

and note that the boundary condition $\mathcal{W}(M,t)=0$ implies that $\partial_t\mathcal{W}(M,t)=0$ as well. This allows us to integrate by parts without introducing boundary terms:

$$\int 16\pi^2 r^2\partial_x\partial_t\mathcal{W}\partial_t v\,dx = -\int 16\pi^2(r^2\partial_t v)\partial_t\mathcal{W}\,dx. \tag{4-50}$$

Using this, we find that

$$\int \partial_t\mathcal{V}\partial_t v\,dx = -\delta\int 16\pi^2\rho|\partial_x(r^2\partial_t v)|^2\,dx + \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\left|\partial_x\left(\frac{\partial_t v}{r}\right)\right|^2 dx + \mathscr{X}, \tag{4-51}$$

where again $\mathscr{X}$ is an error term with the property that $|\mathscr{X}|$ is bounded by the right side of (4-47).

Finally, all of the remaining terms that arise when we multiply (4-48) by $\partial_t v$ can also be estimated by the right side of (4-47). For example, the second term in the third line can be estimated by noting that xr/r_0^4 is bounded, which means that

$$-\int 4\frac{xrv}{r_0^4}\partial_t v\,dx \le \eta\int |\partial_t v|^2\,dx + C\int |v|^2\,dx \le \eta\mathscr{E}^2 + C\mathscr{E}^0. \tag{4-52}$$

Combining all of this, we find that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}&\left\{\int |\partial_t v|^2\,dx + \int \frac{K\gamma\rho_0^{\gamma-1}}{\left(1+\frac{\sigma}{\rho_0}\right)^2}\left|\partial_t\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx\right\}\\ &+ \delta\int 16\pi^2\rho|\partial_x(r^2\partial_t v)|^2\,dx + \frac{4\varepsilon}{3}\int 16\pi^2\rho r^6\left|\partial_x\left(\frac{\partial_t v}{r}\right)\right|^2 dx\\ &\le \left(\tfrac{1}{4}+C\theta_1\right)\mathscr{D}^2 + \eta\mathscr{E}^2 + C\theta_1(\mathscr{E}^2+\mathscr{D}^0) + C\theta_1\mathscr{E}_0^{0,\sigma} + C(\mathscr{E}^0+\mathscr{E}^1),\end{aligned} \tag{4-53}$$

which yields (4-47). ☐

We now derive bootstrapped estimates for $\partial_t\left(\frac{\sigma}{\rho_0}\right)$. We take ∂_t of (4-4) to get

$$\frac{\rho_0}{\rho}\partial_t^2\left(\frac{\sigma}{\rho_0}\right) = -4\pi\rho\partial_x(r^2\partial_t v) - 8\pi\rho\partial_x(rv^2) - 4\pi\,\partial_t\sigma\,\partial_x(r^2 v) + \frac{\rho_0^2}{\rho^2}\left(\partial_t\left(\frac{\sigma}{\rho_0}\right)\right)^2. \tag{4-54}$$

Next, we multiply (4-54) by $\rho_0^{\beta}\partial_t\left(\frac{\sigma}{\rho_0}\right)$ and integrate to see that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int\frac{\rho_0^{\beta+1}}{\rho}\left|\partial_t\left(\frac{\sigma}{\rho_0}\right)\right|^2dx &\le \frac{C}{\eta}\int 16\pi^2\rho|\partial_x(r^2\partial_t v)|^2\,dx+\eta\int\rho\rho_0^{2\beta}\left|\partial_t\left(\frac{\sigma}{\rho_0}\right)\right|^2dx\\ &\quad-\int 8\pi\frac{v}{r}\left\{\rho\partial_x(r^2v)+\rho r^3\partial_x\left(\frac{v}{r}\right)\right\}\rho_0^{\beta}\partial_t\left(\frac{\sigma}{\rho_0}\right)dx+\frac{3}{2}\int\frac{\rho_0^{\beta+2}}{\rho^2}\left(\partial_t\left(\frac{\sigma}{\rho_0}\right)\right)^3dx\\ &\quad-4\pi\int\partial_t\sigma\,\partial_x(r^2v)\rho_0^{\beta}\partial_t\left(\frac{\sigma}{\rho_0}\right)dx.\end{aligned}\tag{4-55}$$

Then we estimate

$$-\int 8\pi\frac{v}{r}\left\{\rho\partial_x(r^2v)+\rho r^3\partial_x\left(\frac{v}{r}\right)\right\}\rho_0^{\beta}\partial_t\left(\frac{\sigma}{\rho_0}\right)dx-4\pi\int\partial_t\sigma\,\partial_x(r^2v)\rho_0^{\beta}\partial_t\left(\frac{\sigma}{\rho_0}\right)dx\le C\theta_1\mathscr{D}^0+C\theta_1\mathscr{E}^{2,\sigma}_{2\beta+1}$$

to obtain

$$\frac{d}{dt}\mathscr{E}^{2,\sigma}_{\beta}\le\frac{1}{4}\mathscr{D}^2+C\theta_1\mathscr{D}^0+(\eta+C\theta_1)\mathscr{E}^{2,\sigma}_{2\beta+1}+C\theta_1\mathscr{E}^{2,\sigma}_{\beta}.\tag{4-56}$$

Next we estimate $\mathscr{E}^3$ and $\mathscr{D}^3$.

Lemma 4.4. *There exists an energy* $\mathscr{F}^3$ *such that*

$$\frac{d}{dt}[\mathscr{E}^3+\mathscr{F}^3]+\mathscr{D}^3\le C\theta_1\mathscr{E}^3+\left(\frac{3}{8}+\frac{\theta_1}{4}\right)\mathscr{D}^3+C\big(\mathscr{E}^{2,\sigma}_0+\mathscr{E}^2+\mathscr{E}^1+\mathscr{E}^0\big).\tag{4-57}$$

Moreover, *we have the estimate* $|\mathscr{F}^3|\le C\theta_1(\mathscr{E}^3+\mathscr{E}^1)$.

Proof. First recall (4-48) and rewrite it as

$$\begin{aligned}\partial_t^2v&+\underbrace{4\pi r^2\left\{K\gamma\rho_0^{\gamma}\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right)+K\gamma\partial_x(\rho_0^{\gamma})\partial_t\left(\frac{\sigma}{\rho_0}\right)\right\}}_{(\mathrm{a}_1)}\\ &+\underbrace{4\pi r^2\left\{\partial_t\left[\partial_x(a_*\rho_0^{\gamma})\left(\frac{\sigma}{\rho_0}\right)^2+2(a_*\rho_0^{\gamma})\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right]\right\}}_{(\mathrm{a}_2)}\\ &+\underbrace{8\pi rv\left\{K\gamma\rho_0^{\gamma}\partial_x\left(\frac{\sigma}{\rho_0}\right)+K\gamma\partial_x(\rho_0^{\gamma})\frac{\sigma}{\rho_0}+\partial_x(a_*\rho_0^{\gamma})\left(\frac{\sigma}{\rho_0}\right)^2+2a_*\rho_0^{\gamma}\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\}}_{(\mathrm{b})}\\ &\underbrace{-2\frac{xv(r_0^4-r^4)}{r^3r_0^4}-4\frac{xrv}{r_0^4}}_{(\mathrm{c})}=\partial_t\mathscr{V},\end{aligned}\tag{4-58}$$

where $\partial_t\mathscr{V}$ is given in (4-49). To derive (4-57), we will multiply by ∂_t^2v and integrate over x. We divide the estimates into the following steps.

Step 1 We begin with an estimate of the product of $\partial_t^2 v$ with the terms (a_1), (a_2), (b), and (c). First, we use (4-45) to replace $\partial_t\partial_x(\sigma/\rho_0)$ by lower-order terms:

$$
\begin{aligned}
(\mathrm{a}_1)+(\mathrm{a}_2) &= 4\pi r^2\left\{K\gamma\rho_0^\gamma\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right)+K\gamma\partial_x(\rho_0^\gamma)\partial_t\left(\frac{\sigma}{\rho_0}\right)\right\}\\
&\qquad +4\pi r^2\left\{\partial_t\left[\partial_x\big(a_*\rho_0^\gamma\big)\left(\frac{\sigma}{\rho_0}\right)^2+2\big(a_*\rho_0^\gamma\big)\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right]\right\}\\
&= -4\pi r^2\left(K\gamma+2a_*\frac{\sigma}{\rho_0}\right)\rho_0^\gamma\frac{\rho}{\rho_0}\partial_x\left(\frac{\rho_0}{\rho}\right)\partial_t\left(\frac{\sigma}{\rho_0}\right)\\
&\qquad -\left(K\gamma+2a_*\frac{\sigma}{\rho_0}\right)\rho_0^\gamma\frac{\rho}{\rho_0\left(\delta+\dfrac{4\varepsilon}{3}\right)}\left\{\partial_t v+\frac{x(r_0^4-r^4)}{r^2r_0^4}+4\pi r^2\left[K\gamma\rho_0^\gamma\partial_x\left(\frac{\sigma}{\rho_0}\right)\right.\right.\\
&\qquad\qquad \left.\left.+K\gamma\partial_x(\rho_0^\gamma)\frac{\sigma}{\rho_0}+\partial_x(a_*\rho_0^\gamma)\left(\frac{\sigma}{\rho_0}\right)^2+2a_*\rho_0^\gamma\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right]\right\}\\
&\qquad +4\pi r^2\left\{K\gamma\partial_x(\rho_0^\gamma)\partial_t\left(\frac{\sigma}{\rho_0}\right)+\partial_t\left[\partial_x(a_*\rho_0^\gamma)\left(\frac{\sigma}{\rho_0}\right)^2\right]+2\rho_0^\gamma\partial_t\left(a_*\frac{\sigma}{\rho_0}\right)\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\}\\
&=: (\mathrm{A}_1)+(\mathrm{A}_2)+(\mathrm{A}_3).
\end{aligned}
\tag{4-59}
$$

Then $\int\partial_t^2 v\cdot[(\mathrm{a}_1)+(\mathrm{a}_2)]\,dx$ can be estimated as follows:

$$
\int\partial_t^2 v\cdot(\mathrm{A}_1)\,dx\le\frac{\theta_1}{8}\int\left|\partial_t^2 v\right|^2dx+C\theta_1\int\frac{\rho_0^{2\gamma+2}}{\rho^2}r^4\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2dx\le\frac{\theta_1}{8}\mathscr{D}^3+C\theta_1\mathscr{E}^{1,\sigma},\tag{4-60}
$$

$$
\int\partial_t^2 v\cdot(\mathrm{A}_2)\,dx\le\frac{3}{32}\int|\partial_t^2 v|^2\,dx+C\big[\mathscr{E}^{2,v}+\mathscr{E}^{0,r}+(1+\theta_1)(\mathscr{E}^{1,\sigma}+\mathscr{E}^{0,\sigma})\big],\tag{4-61}
$$

$$
\int\partial_t^2 v\cdot(\mathrm{A}_3)\,dx\le\frac{3}{32}\int|\partial_t^2 v|^2\,dx+C\theta_1\mathscr{E}^{1,\sigma}+C(1+\theta_1)\mathscr{E}_0^{2,\sigma}.\tag{4-62}
$$

For (b) and (c), we may estimate

$$
\begin{aligned}
\int\partial_t^2 v\cdot(\mathrm{b})\,dx&\le\frac{\theta_1}{8}\int|\partial_t^2 v|^2\,dx+C\theta_1\big(\mathscr{E}^{1,\sigma}+\mathscr{E}_0^{0,\sigma}\big),\\
\int\partial_t^2 v\cdot(\mathrm{c})\,dx&\le\frac{3}{32}\int|\partial_t^2 v|^2\,dx+C\big(\mathscr{E}^{0,v}+\mathscr{E}^{0,r}\big).
\end{aligned}
\tag{4-63}
$$

Combining the above, we arrive at an estimate for $\int\partial_t^2 v\cdot\big[(\mathrm{a}_1)+(\mathrm{a}_2)+(\mathrm{b})+(\mathrm{c})\big]\,dx$.

Step 2 (the viscosity term). Now we consider the viscosity term, $\partial_t\mathscr{V}$. We claim that there exist $\mathscr{F}^3$, G such that

$$
\int\partial_t^2 v\cdot\partial_t\mathscr{V}\,dx=-\frac{d}{dt}\mathscr{E}^3-\frac{d}{dt}\mathscr{F}^3+G,\tag{4-64}
$$

where

$$|\mathscr{F}^3| \le C\theta_1(\mathscr{E}^3+\mathscr{E}^1) \quad \text{and} \quad |G| \le \tfrac{3}{32}\mathscr{D}^3 + C\theta_1\bigl(\mathscr{E}^3+\mathscr{E}_0^{2,\sigma}+\mathscr{E}^1\bigr). \tag{4-65}$$

Recall that $\partial_t\mathscr{V}$ may be computed as in (4-49), and that $\partial_t\mathscr{W}(M,t)=0$. Then a simple but lengthy computation, using integration by parts, reveals that

$$\int \partial_t^2 v\cdot\partial_t\mathscr{V}\,dx = -\frac{1}{2}\frac{d}{dt}\int \delta 16\pi^2\rho\bigl|\partial_x(r^2\partial_t v)\bigr|^2 dx + \frac{4\varepsilon}{3}16\pi^2\rho r^6\left|\partial_x\left(\frac{\partial_t v}{r}\right)\right|^2 dx + G_0 + Y, \tag{4-66}$$

where

$$\begin{aligned} G_0 = {}&\int\left(\frac{\partial_t\sigma}{2\rho}+\frac{2v}{r}\right)\left\{\delta 16\pi^2\rho\bigl|\partial_x(r^2\partial_t v)\bigr|^2 + \frac{4\varepsilon}{3}16\pi^2\rho r^6\left|\partial_x\left(\frac{\partial_t v}{r}\right)\right|^2\right\}dx \\ &+\int \rho r^3\partial_x\left(\frac{v}{r}\right)\left\{\delta 32\pi^2\frac{\partial_t v}{r}\partial_x(r^2\partial_t v) - \frac{4\varepsilon}{3}16\pi^2\frac{\partial_t v}{r}r^3\partial_x\left(\frac{\partial_t v}{r}\right)\right\}dx \end{aligned} \tag{4-67}$$

and $Y=Y_1+Y_2$ with

$$\begin{aligned} Y_1 &= -16\pi^2\int\bigl[\delta\partial_t\sigma\,\partial_x(r^2v)+\delta\rho\,\partial_x(2rv^2)\bigr]\partial_x(r^2\partial_t^2 v)\,dx \\ &\quad -16\pi^2\int\left[\frac{4\varepsilon}{3}\partial_t(\rho r^3)\partial_x\left(\frac{v}{r}\right)-\frac{4\varepsilon}{3}\rho r^3\partial_x\left(\frac{v^2}{r^2}\right)\right]\partial_x(r^2\partial_t^2 v)\,dx, \\ Y_2 &= 32\pi^2\int rv\partial_t^2 v\,\partial_x\left[\delta\rho\,\partial_x(r^2v)+\frac{4\varepsilon}{3}\rho r^3\partial_x\left(\frac{v}{r}\right)\right]dx. \end{aligned} \tag{4-68}$$

Let us define the quantity Q such that $Y_1 = -16\pi^2\int Q\,\partial_x(r^2\partial_t^2 v)\,dx$, that is, Q is the sum of the bracketed terms in the Y_1 integrand. Then we may compute

$$Y_1 = \frac{d}{dt}\int -16\pi^2\partial_x(r^2\partial_t v)Q\,dx + \int 16\pi^2\bigl(\partial_x(2rv\partial_t v)Q+\partial_x(r^2\partial_t v)\partial_t Q\bigr)\,dx := -\frac{d}{dt}\mathscr{F}_1^3 + G_1. \tag{4-69}$$

Similarly, we have that

$$\begin{aligned} Y_2 &= \frac{d}{dt}\int -16\pi^2\partial_x(r^2\partial_t v)\frac{2v}{r}\left[\delta\rho\,\partial_x(r^2v)+\frac{4\varepsilon}{3}\rho r^3\partial_x\left(\frac{v}{r}\right)\right]dx \\ &\quad + 16\pi^2\int\left[\partial_x(2rv\partial_t v)\frac{2v}{r}\mathscr{W} + \partial_x(r^2\partial_t v)\partial_t\left(\frac{2v}{r}\right)\mathscr{W} + \partial_x(r^2\partial_t v)\frac{2v}{r}\partial_t\mathscr{W}\right]dx \\ &\quad - 16\pi^2\int r^2\partial_t^2 v\,\mathscr{W}\,\partial_x\left(\frac{2v}{r}\right)dx \\ &= -\frac{d}{dt}\mathscr{F}_2^3 + G_2. \end{aligned} \tag{4-70}$$

Combining the above, we find that (4-64) holds with $\mathscr{F}^3=\mathscr{F}_1^3+\mathscr{F}_2^3$ and $G=G_0+G_1+G_2$. To complete the proof of the claim, we note that the estimates (4-65) follow from the definition of $\mathscr{F}^3$ and G, using (4-54) to replace $\partial_t^2\sigma$ by other terms.

Step 3 (conclusion). The only term that remains is

$$\int \partial_t^2 v \partial_t^2 v \, dx = \mathscr{D}^3. \tag{4-71}$$

With this, all of the terms in (4-58) are accounted for. We may then combine the analysis of Steps 1 and 2 to deduce the estimate (4-57). □

We now bootstrap more estimates. First, we multiply (4-22) by $\frac{1}{\rho}\Big(1 - \frac{r_0}{r}\Big)$ and integrate to get

$$\frac{d}{dt}\mathscr{E}_b^{0,r} \le (\eta + C\theta_1)\mathscr{E}_b^{0,r} + C\mathscr{E}^{1,v}, \tag{4-72}$$

where we have used (4-23) to control $\int \frac{|v|^2}{r^2\rho}\, dx \le C\mathscr{E}^{1,v}$. By multiplying (4-45) by $\frac{1}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)$ and integrating, we get

$$\frac{d}{dt}\mathscr{E}_b^{1,\sigma} + \mathscr{D}_b^{1,\sigma} \le (\eta + C\theta_1)\mathscr{E}_b^{1,\sigma} + C\big(\mathscr{E}^3 + \mathscr{E}_{-1}^{0,\sigma} + \mathscr{E}_b^{0,r}\big). \tag{4-73}$$

Note that here we have again used (4-23) to control $\int |\partial_t v|^2/(r^2\rho)\, dx$, which is possible since (4-23) is valid for any choice of v, not just solutions. From (4-45) we also see that

$$\mathscr{E}_a^{3,\sigma} = \int \left(\delta + \frac{4\varepsilon}{3}\right)^2 16\pi^2 \frac{r^2}{\rho}\left|\partial_x\partial_t\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx \le C\big(\mathscr{E}^3 + \mathscr{E}_b^{1,\sigma} + \mathscr{E}_{-1}^{0,\sigma} + \mathscr{E}^0\big). \tag{4-74}$$

Next, by applying ∂_x to (4-4), we find that

$$\frac{\rho_0}{\rho}\partial_x\partial_t\left(\frac{\sigma}{\rho_0}\right) + \partial_x\left(\frac{\rho_0}{\rho}\right)\partial_t\left(\frac{\sigma}{\rho_0}\right) + 4\pi\,\partial_x\big(\rho\partial_x(r^2 v)\big) = 0. \tag{4-75}$$

We then use this to get

$$\mathscr{E}_{a_1}^{3,v} = \int \frac{r^2}{\rho}\big|\partial_x(\rho\partial_x(r^2 v))\big|^2 dx \le C\big(\mathscr{E}_a^{3,\sigma} + \mathscr{E}_b^{1,\sigma}\big). \tag{4-76}$$

Since $\int \rho r^6 \left|\partial_x\left(\frac{v}{r}\right)\right|^2 dx \le C\mathscr{E}^1$, (4-7) implies that

$$\mathscr{E}_{a_2}^{3,v} = \int \rho r^6 \left|\partial_x\left(\rho r^3 \partial_x\left(\frac{v}{r}\right)\right)\right|^2 dx \le C\big(\mathscr{E}^2 + \mathscr{E}^1 + \mathscr{E}^0\big). \tag{4-77}$$

We now illustrate how the higher-order energy estimates of spatial derivatives of $\partial_x(\sigma/\rho_0)$ and $\partial_x(\rho\partial_x(r^2 v))$ work. The following lemma concerns the estimate of $\partial_x(r^4\partial_x(\sigma/\rho_0))$.

Lemma 4.5. $$\frac{d}{dt}\mathscr{E}^4 + \mathscr{D}^4 \le (\eta + C\theta_1)\mathscr{E}^4 + C\big(\mathscr{E}^3 + \mathscr{E}^{0,r} + \mathscr{E}_b^{0,r} + \mathscr{E}_{-1}^{0,\sigma}\big) + C\theta_1\big(\mathscr{E}_b^{1,\sigma} + \mathscr{E}_a^{3,\sigma}\big). \tag{4-78}$$

Proof. First, we multiply (4-45) by r^2 and apply ∂_x to get

$$\begin{aligned}
&\left(\delta+\frac{4\varepsilon}{3}\right)4\pi\left\{\underbrace{\frac{\rho_0}{\rho}\partial_x\left(r^4\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)}_{\text{(i)}}+2\,\underbrace{\partial_x\left(\frac{\rho_0}{\rho}\right)r^4\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right)}_{\text{(ii)}}+\underbrace{\partial_x\left(r^4\partial_x\left(\frac{\rho_0}{\rho}\right)\right)\partial_t\left(\frac{\sigma}{\rho_0}\right)}_{\text{(iii)}}\right\}\\
&\qquad+\underbrace{\partial_x(r^2\partial_t v)}_{\text{(iv)}}+\underbrace{\frac{(r_0^4-r^4)}{r_0^4}}_{\text{(v)}}-\underbrace{\frac{xr^4}{\pi r_0^7\rho}\left(1-\left(\frac{r_0}{r}\right)^3+\frac{\sigma}{\rho_0}\right)}_{\text{(vi)}}+4\pi\,\underbrace{K\gamma\rho_0^\gamma\,\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)}_{\text{(vii)}}\\
&\qquad+4\pi\left\{2K\gamma r^4\partial_x(\rho_0^\gamma)\partial_x\left(\frac{\sigma}{\rho_0}\right)+K\gamma\,\partial_x\big(r^4\partial_x(\rho_0^\gamma)\big)\frac{\sigma}{\rho_0}\right\}\\
&\qquad+4\pi\left\{2r^4\partial_x(a_*\rho_0^\gamma)\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)+\partial_x\big(r^4\partial_x(a_*\rho_0^\gamma)\big)\left(\frac{\sigma}{\rho_0}\right)^2\right\}\\
&\qquad+4\pi\left\{2a_*\rho_0^\gamma\frac{\sigma}{\rho_0}\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)+2r^4\partial_x\left(a_*\rho_0^\gamma\frac{\sigma}{\rho_0}\right)\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\}\\
&=0. \qquad\qquad (4\text{-}79)
\end{aligned}$$

The energy inequality (4-78) can be derived as in Step 2 of Lemma 4.2 by multiplying (4-79) by $\rho\partial_x(r^4\partial_x(\sigma/\rho_0))$ and integrating over x. We provide the details on how (i)–(vii) can be treated; other terms can be estimated similarly.

$$\begin{aligned}
&\int\left(\delta+\frac{4\varepsilon}{3}\right)4\pi\text{(i)}\cdot\rho\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)dx\\
&\qquad=\frac{1}{2}\frac{d}{dt}\int\left(\delta+\frac{4\varepsilon}{3}\right)4\pi\rho_0\left|\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)\right|^2dx\\
&\qquad\qquad-\left(\delta+\frac{4\varepsilon}{3}\right)4\pi\underbrace{\int\rho_0\partial_x\left(\partial_t(r^4)\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)dx}_{(*)}. \qquad (4\text{-}80)
\end{aligned}$$

Since $\partial_t(r^4)=4\frac{v}{r}r^4$,

$$(*)=\int 4\frac{v}{r}\rho_0\left|\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)\right|^2dx+\int 4\partial_x\left(\frac{v}{r}\right)r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\rho_0\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)dx, \qquad (4\text{-}81)$$

and since $\left|\frac{v}{r}\right|$ and $\left|\rho r^3\partial_x\left(\frac{v}{r}\right)\right|$ are bounded by θ_1,

$$|(*)|\le C\theta_1\big(\mathscr{E}^4+\mathscr{E}_b^{1,\sigma}\big). \qquad (4\text{-}82)$$

For (ii), we write

$$\int\text{(ii)}\cdot\rho\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)dx=-\int\frac{\rho_0r^3\partial_x\left(\frac{\sigma}{\rho_0}\right)}{1+\frac{\sigma}{\rho_0}}\cdot\frac{r}{\sqrt{\rho}}\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right)\cdot\sqrt{\rho}\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)dx, \qquad (4\text{-}83)$$

and therefore

$$\left|\int (\mathrm{ii})\cdot\rho\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)dx\right| \le C\theta_1\big(\mathscr{E}_a^{3,\sigma}+\mathscr{E}^4\big). \tag{4-84}$$

It is easy to see that

$$\left|\int\big[(\mathrm{iii})+(\mathrm{iv})+(\mathrm{v})-(\mathrm{vi})\big]\cdot\rho\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)dx\right| \le \left(\frac{\eta}{2}+C\theta_1\right)\mathscr{E}^4+C\big(\mathscr{E}^3+\mathscr{E}^{0,r}+\mathscr{E}_b^{0,r}+\mathscr{E}_{-1}^{0,\sigma}\big). \tag{4-85}$$

Finally, (vii) forms the dissipation $\mathscr{D}^4$. ☐

We also get an estimate for $\partial_x\left(r^4\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)$ from (4-79):

$$\mathscr{E}_{a_1}^4=\int\left(\delta+\frac{4\varepsilon}{3}\right)4\pi\rho_0\left|\partial_x\left(r^4\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)\right|^2dx \le \theta_1^2\mathscr{E}_{a_1}^{3,\sigma}+\mathscr{E}^4+\mathscr{E}^3+\mathscr{E}^1+\mathscr{E}_b^{0,r}+\mathscr{E}_{-1}^{0,\sigma}. \tag{4-86}$$

To derive an estimate of the third spatial derivatives of v, we first multiply (4-75) by r^4 and then apply ∂_x:

$$\begin{aligned}\frac{\rho_0}{\rho}\partial_x\left(r^4\partial_x\partial_t\left(\frac{\sigma}{\rho_0}\right)\right)+2r^4\partial_x\left(\frac{\rho_0}{\rho}\right)\partial_x\partial_t\left(\frac{\sigma}{\rho_0}\right)+\partial_x\left(r^4\partial_x\left(\frac{\rho_0}{\rho}\right)\right)\partial_t\left(\frac{\sigma}{\rho_0}\right)&\\ +4\pi\,\partial_x\big(r^4\partial_x(\rho\partial_x(r^2v))\big)=0.&\end{aligned} \tag{4-87}$$

Thus, we obtain

$$\mathscr{E}_{a_2}^4=\int 16\pi^2\rho_0\big|\partial_x\big(r^4\partial_x(\rho\partial_x(r^2v))\big)\big|^2dx \le \mathscr{E}_{a_1}^4+\theta_1^2\big(\mathscr{E}_a^{3,\sigma}+\mathscr{E}^4\big). \tag{4-88}$$

We now present the higher-order energy estimates. We start with $\mathscr{E}^5$ and $\mathscr{E}_a^5$.

Lemma 4.6.

$$\frac{d}{dt}\mathscr{E}^5+\mathscr{D}^5 \le (\eta+C\theta_1)\mathscr{E}^5+\big(\tfrac{1}{2}+C\theta_1\big)\mathscr{D}^5+C\theta_1^2\mathscr{E}^3+C\theta_1\big(\mathscr{E}_a^{3,\sigma}+\mathscr{E}^3+\mathscr{E}^2+\mathscr{E}^1\big)+C\mathscr{E}^2. \tag{4-89}$$

Proof. We apply ∂_t to (4-48) to see that

$$\begin{aligned}&\partial_t^3v+4\pi r^2\partial_x\left\{\left(K\gamma+2a_*\frac{\sigma}{\rho_0}\right)\rho_0^\gamma\partial_t^2\left(\frac{\sigma}{\rho_0}\right)+\rho_0^\gamma\left[a_*\left(\partial_t\left(\frac{\sigma}{\rho_0}\right)\right)^2+4\partial_ta_*\frac{\sigma}{\rho_0}\partial_t\left(\frac{\sigma}{\rho_0}\right)+\partial_t^2a_*\left(\frac{\sigma}{\rho_0}\right)^2\right]\right\}\\ &\quad+16\pi vr\left\{K\gamma\rho_0^\gamma\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right)+K\gamma\partial_x(\rho_0^\gamma)\partial_t\left(\frac{\sigma}{\rho_0}\right)\right\}+16\pi vr\partial_t\left[\partial_x(a_*\rho_0^\gamma)\left(\frac{\sigma}{\rho_0}\right)^2+2(a_*\rho_0^\gamma)\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right]\\ &\quad+8\pi(r\partial_tv+v^2)\left\{K\gamma\rho_0^\gamma\partial_x\left(\frac{\sigma}{\rho_0}\right)+K\gamma\partial_x(\rho_0^\gamma)\frac{\sigma}{\rho_0}+\partial_x(a_*\rho_0^\gamma)\left(\frac{\sigma}{\rho_0}\right)^2+2a_*\rho_0^\gamma\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\}\\ &\quad-\frac{2x(r_0^4-r^4)\partial_tv}{r^3r_0^4}-\frac{4xr\partial_tv}{r_0^4}-\frac{12xv^2}{r_0^4}+\frac{6x(r_0^4-r^4)v^2}{r^4r_0^4}\\ &=\partial_t^2\mathscr{V}.\end{aligned} \tag{4-90}$$

We derive the energy estimate of (4-89) from (4-48) by proceeding as in the proofs of Lemmas 4.1 and 4.3. That is, we multiply the resulting equation by $\partial_t^2 v$ and integrate over x, integrating by parts in some terms to recover $d\mathscr{E}^5/dt$, $\mathscr{D}^5$, and some error terms that can be estimated by the right side of (4-89). Since the method of proof is already recorded in Lemmas 4.1 and 4.3, we omit further details. □

An estimate of $\partial_x\big(\rho\partial_x(\partial_t[r^2 v])\big)$ can be obtained through (4-48):

$$\mathscr{E}_a^{5,v} \le C\big(\mathscr{E}^5 + \mathscr{E}_a^{3,\sigma} + \mathscr{E}^1\big) + C\theta_1\big(\mathscr{E}_{a_1}^{3,v} + \mathscr{E}_a^{3,\sigma} + \mathscr{E}^1 + \mathscr{E}^0\big). \tag{4-91}$$

We now bootstrap to control $\partial_t^2\left(\dfrac{\sigma}{\rho_0}\right)$. We apply ∂_t to (4-54) to get

$$\partial_t^3\left(\frac{\sigma}{\rho_0}\right) = -4\pi\frac{\rho}{\rho_0}\rho\partial_x(r^2\partial_t^2 v) - 24\pi\frac{\rho}{\rho_0}\underbrace{\rho\partial_x(rv\partial_t v)}_{\text{(a)}} - 8\pi\frac{\rho}{\rho_0}\underbrace{\rho\partial_x(v^3)}_{\text{(b)}} + 6\frac{\partial_t\sigma}{\rho}\partial_t^2\left(\frac{\sigma}{\rho_0}\right) - 6\frac{(\partial_t\sigma)^3}{\rho_0\rho^2}. \tag{4-92}$$

Note that

$$\text{(a)} = \frac{v}{r}\rho\partial_x(r^2\partial_t v) + \rho r^3\partial_x\left(\frac{v}{r}\right)\frac{\partial_t v}{r}$$

and

$$\text{(b)} = 3r^3\rho\left(\frac{v}{r}\right)^2\partial_x\left(\frac{v}{r}\right) + \frac{3}{4\pi}\left(\frac{v}{r}\right)^3,$$

and thus by multiplying (4-92) by $\dfrac{1}{\rho_0}\partial_t^2\left(\dfrac{\sigma}{\rho_0}\right)$ and integrating, we obtain

$$\frac{d}{dt}\mathscr{E}_{-1}^{5,\sigma} \le (\eta + C\theta_1)\mathscr{E}_{-1}^{5,\sigma} + C\mathscr{D}^5 + C\theta_1^2\big(\mathscr{E}^3 + \mathscr{E}^1 + \mathscr{E}_{-1}^{2,\sigma}\big). \tag{4-93}$$

Next, we take ∂_t of (4-45) to see that

$$\begin{aligned}
&\left(\delta + \frac{4\varepsilon}{3}\right)4\pi r^2\left\{\frac{\rho_0}{\rho}\partial_t^2\partial_x\left(\frac{\sigma}{\rho_0}\right) + \partial_x\left(\frac{\rho_0}{\rho}\right)\partial_t^2\left(\frac{\sigma}{\rho_0}\right) + \partial_t\left(\frac{\rho_0}{\rho}\right)\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right) + \partial_t\partial_x\left(\frac{\rho_0}{\rho}\right)\partial_t\left(\frac{\sigma}{\rho_0}\right)\right\}\\
&\qquad + \left(\delta + \frac{4\varepsilon}{3}\right)\frac{v}{r}8\pi r^2\left\{\frac{\rho_0}{\rho}\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right) + \partial_x\left(\frac{\rho_0}{\rho}\right)\partial_t\left(\frac{\sigma}{\rho_0}\right)\right\}\\
&\qquad + \partial_t^2 v - 2\frac{v}{r}\left(\frac{x}{r^2} + \frac{xr^2}{r_0^4}\right) + 4\pi r^2\left\{K\gamma\rho_0^\gamma\partial_t\partial_x\left(\frac{\sigma}{\rho_0}\right) + K\gamma\partial_x(\rho_0^\gamma)\partial_t\left(\frac{\sigma}{\rho_0}\right)\right\}\\
&\qquad + 4\pi r^2\partial_t\left[\partial_x(a_*\rho_0^\gamma)\left(\frac{\sigma}{\rho_0}\right)^2 + 2a_*\rho_0^\gamma\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right]\\
&\qquad + \frac{v}{r}8\pi r^2\left\{K\gamma\rho_0^\gamma\partial_x\left(\frac{\sigma}{\rho_0}\right) + K\gamma\partial_x(\rho_0^\gamma)\frac{\sigma}{\rho_0} + \partial_x(a_*\rho_0^\gamma)\left(\frac{\sigma}{\rho_0}\right)^2 + 2a_*\rho_0^\gamma\frac{\sigma}{\rho_0}\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\}\\
&= 0.
\end{aligned} \tag{4-94}$$

Therefore, by squaring (4-94) and integrating, we find that

$$\int \left(\delta+\frac{4\varepsilon}{3}\right)^2 16\pi^2 r^4 \left|\partial_t^2\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx \le C\big(\mathscr{E}^5+\mathscr{E}^1\big)+C\theta_1^2\big(\mathscr{E}_{-1}^{5,\sigma}+\mathscr{E}_a^{3,\sigma}+\mathscr{E}^1+\mathscr{E}_0^{0,\sigma}\big). \tag{4-95}$$

Also, by first dividing (4-94) by r and then squaring, we obtain

$$\begin{aligned}\mathscr{E}_a^{6,\sigma} &= \int \left(\delta+\frac{4\varepsilon}{3}\right)^2 16\pi^2 r^2 \left|\partial_t^2\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx \\ &\le C\big(\mathscr{E}^6+\mathscr{E}_a^{3,\sigma}+\mathscr{E}_{-1}^{2,\sigma}+\mathscr{E}^1\big)+C\theta_1^2\big(\mathscr{E}_b^{1,\sigma}+\mathscr{E}_a^{3,\sigma}+\mathscr{E}^1+\mathscr{E}_{-1}^{0,\sigma}\big).\end{aligned} \tag{4-96}$$

Now we record an estimate of $\mathscr{E}^6$.

Lemma 4.7. *There exists an* $\mathscr{F}^6$ *such that*

$$\frac{d}{dt}\big[\mathscr{E}^6+\widetilde{F}\big]+\mathscr{D}^6 \le \big(\eta+C\theta_1\big)\mathscr{E}^6+\big(\tfrac12+C\theta_1\big)\mathscr{D}^6+C\big(\mathscr{E}^5+\mathscr{E}_{-1}^{5,\sigma}+\mathscr{E}_b^{1,\sigma}+\mathscr{E}_0^{2,\sigma}+\mathscr{E}_0^{0,\sigma}+\mathscr{E}^3+\mathscr{E}^2+\mathscr{E}^1\big). \tag{4-97}$$

Moreover, $|\mathscr{F}^6| \le C\theta_1\big(\mathscr{E}^6+\mathscr{E}^3+\mathscr{E}^1\big)$.

Proof. The energy inequality (4-97) can be obtained by multiplying (4-90) by $\partial_t^3 v$ and integrating over x as done in Lemma 4.4. We omit further details. □

As seen in the previous estimates in Lemmas 4.3, 4.4, 4.6, and 4.7, the time differentiation of the equation keeps the main structure of the highest-order terms as well as the boundary condition. Using the time differentiated equations (4-90) and (4-92), we can follow the line of analysis presented in these four lemmas to derive energy inequalities for $\mathscr{E}^7$, $\mathscr{E}_a^7$, $\mathscr{E}_{-1}^{7,\sigma}$, $\mathscr{E}^8$ and $\mathscr{E}_a^{8,\sigma}$. We record these in the following lemma but omit a proof.

Lemma 4.8. *Let* $\mathscr{E}$ *be given by* (4-19). *We have the following estimates.*

$$\frac{d}{dt}\mathscr{E}^7+\mathscr{D}^7 \le (\eta+C\theta_1)\mathscr{E}^7+\big(\tfrac12+C\theta_1\big)\mathscr{D}^7+C\theta_1\big(\mathscr{E}_a^{6,\sigma}+\mathscr{E}^6+\mathscr{E}^5+\mathscr{E}^3+\mathscr{E}^2+\mathscr{E}^1\big)+C\mathscr{E}^5, \tag{4-98}$$

$$\mathscr{E}_a^7 \le C\big(\mathscr{E}^7+\mathscr{E}_a^{6,\sigma}+\mathscr{E}^2\big)+C\theta_1\big(\mathscr{E}_a^{6,\sigma}+\mathscr{E}_a^{3,\sigma}+\mathscr{E}^1+\mathscr{E}^2+\mathscr{E}_{a_1}^{3,v}+\mathscr{E}_a^5\big), \tag{4-99}$$

$$\frac{d}{dt}\mathscr{E}_{-1}^{7,\sigma} \le \big(\eta+C\theta_1\big)\mathscr{E}_{-1}^{7,\sigma}+C\mathscr{D}^7+C\theta_1^2\big(\mathscr{E}^6+\mathscr{E}^3+\mathscr{E}_{-1}^{5,\sigma}\big), \tag{4-100}$$

$$\begin{aligned}\frac{d}{dt}\big[\mathscr{E}^8+\mathscr{F}^8\big]+\mathscr{D}^8 \le \big(\eta+C\theta_1\big)\mathscr{E}^8+\big(\tfrac12+C\theta_1\big)\mathscr{D}^8+C\big(\mathscr{E}-\mathscr{E}^8-\mathscr{E}_a^{8,\sigma}\big)+C|\mathscr{E}|^2 \\ \textit{where } |\mathscr{F}^8| \le C\theta_1\mathscr{E}+C|\mathscr{E}|^2,\end{aligned} \tag{4-101}$$

$$\mathscr{E}_a^{8,\sigma} \le C\big(\mathscr{E}^8+\mathscr{E}_a^{6,\sigma}+\mathscr{E}_{-1}^{5,\sigma}+\mathscr{E}^3\big)+C\theta_1^2\big(\mathscr{E}_b^{1,\sigma}+\mathscr{E}_a^{3,\sigma}+\mathscr{E}^1+\mathscr{E}_{-1}^{0,\sigma}\big). \tag{4-102}$$

The next lemma ensures that the assumption (4-20) is valid within our energy $\mathscr{E}$.

Lemma 4.9. *There exists a constant* $\kappa > 0$ *such that if* $\mathscr{E} \le \kappa$, *then*

$$\begin{aligned}&\left\|\frac{\sigma}{\rho_0}\right\|_{L^\infty}+\left\|\partial_t\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty}+\left\|\partial_t^2\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty}+\left\|\partial_t^3\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty} \\ &\qquad+\left\|1-\frac{r_0}{r}\right\|_{L^\infty}+\left\|\rho r^3\partial_x\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty}+\left\|\frac{v}{r}\right\|_{L^\infty}+\left\|\frac{\partial_t v}{r}\right\|_{L^\infty}+\left\|\frac{\partial_t^2 v}{r}\right\|_{L^\infty} \le C\sqrt{\mathscr{E}},\end{aligned} \tag{4-103}$$

for some constant $C > 0$. Here $\mathscr{E}$ is given by (4-19).

Proof. The proof proceeds in four steps.

Step 1 ($\partial_t^k(\sigma/\rho_0)$ estimates). We begin by estimating σ/ρ_0 in $W^{1,1}((0,M))$. First, we use Hölder's inequality to estimate

$$\int \left|\frac{\sigma}{\rho_0}\right| dx \le \sqrt{M}\left(\int \left|\frac{\sigma}{\rho_0}\right| dx\right)^{1/2} \le C\sqrt{\mathscr{E}_{-1}^{0,\sigma}} \le C\sqrt{\mathscr{E}}. \tag{4-104}$$

On the other hand, we may estimate

$$\int \left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right| dx \le \left(\int \frac{r^2}{\rho}\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx\right)^{1/2}\left(\int \frac{\rho}{r^2} dx\right)^{1/2} \le C\left(\mathscr{E}_b^{1,\sigma}\right)^{1/2} \le C\sqrt{\mathscr{E}}. \tag{4-105}$$

Here we have used the fact that $r^2(x) \sim x^{2/3}$ for $x \sim 0$, which follows from the definition of $r(x)$ and L'Hospital's rule, to see that $\int(\rho/r^2)\,dx < \infty$. Combining these estimates with the usual one-dimensional Sobolev embedding $W^{1,1}((0,M)) \hookrightarrow C^0((0,M))$, we find that $\dfrac{\sigma}{\rho_0} \in C^0$ and

$$\left\|\frac{\sigma}{\rho_0}\right\|_{L^\infty} \le C\sqrt{\mathscr{E}}. \tag{4-106}$$

Now to control $\dfrac{\partial_t\sigma}{\rho_0}$, we argue similarly to estimate

$$\int \left|\frac{\partial_t\sigma}{\rho_0}\right| + \left|\partial_x\left(\frac{\partial_t\sigma}{\rho_0}\right)\right| dx \le C\sqrt{\mathscr{E}_{-1}^{2,\sigma}} + \sqrt{\mathscr{E}_a^{3,\sigma}} \le C\sqrt{\mathscr{E}}. \tag{4-107}$$

Then $\dfrac{\partial_t\sigma}{\rho_0} \in C^0$ and

$$\left\|\frac{\partial_t\sigma}{\rho_0}\right\|_{L^\infty} \le C\sqrt{\mathscr{E}}. \tag{4-108}$$

A similar argument, employing $\mathscr{E}_{-1}^{1+2i,\sigma}$ and $\mathscr{E}_a^{2+2i,\sigma}$ for $i = 1, 2$, then implies that

$$\frac{\partial_t^2\sigma}{\rho_0}, \frac{\partial_t^3\sigma}{\rho_0} \in C^0 \quad \text{and} \quad \left\|\frac{\partial_t^2\sigma}{\rho_0}\right\|_{L^\infty} + \left\|\frac{\partial_t^3\sigma}{\rho_0}\right\|_{L^\infty} \le C\sqrt{\mathscr{E}}. \tag{4-109}$$

We thus deduce from (4-106) and (4-108)–(4-109) that

$$\left\|\frac{\sigma}{\rho_0}\right\|_{L^\infty} + \left\|\partial_t\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty} + \left\|\partial_t^2\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty} + \left\|\partial_t^3\left(\frac{\sigma}{\rho_0}\right)\right\|_{L^\infty} \le C\sqrt{\mathscr{E}}. \tag{4-110}$$

Step 2 ($1 - r_0/r$ estimate). Let us now suppose that $\mathscr{E} \le \kappa$ with κ small enough that $C\sqrt{\mathscr{E}} \le \frac{1}{2}$, where $C > 0$ is the constant appearing on the right side of (4-110). In particular, this implies that $\|\sigma/\rho_0\|_{L^\infty} \le \frac{1}{2} < 1$. With this estimate in hand, we can derive an estimate for r_0/r. Indeed, the Taylor expansion (4-21) easily implies the estimate

$$\left\|1 - \frac{r_0}{r}\right\|_{L^\infty} \le C\left\|\frac{\sigma}{\rho_0}\right\|_{L^\infty}^{1+k} \le C\sqrt{\mathscr{E}} \tag{4-111}$$

for some $k \ge 0$. This is the $1 - r_0/r$ estimate in (4-103).

Step 3 ($\partial_t^k v/r$ estimates). We now turn to estimates for $\partial_t^k v/r$, $k = 0, 1, 2$. From Step 1, we know that σ/ρ_0 and $\partial_t(\sigma/\rho_0)$ are continuous and bounded, while from Step 2, we know that $\|\sigma/\rho_0\|_{L^\infty} \leq \frac{1}{2}$, so that $1+\sigma/\rho_0$ is also continuous and bounded. From the boundary conditions at $x = 0$, we also have that $r^2 v(0, t) = 0$. Hence we may spatially integrate the continuity equation (4-4) to see that

$$(r^2 v)(x, t) = \frac{-1}{4\pi} \int_0^x \frac{1}{\rho_0(y)\left(1+\dfrac{\sigma(y, t)}{\rho_0(y)}\right)^2} \frac{\partial_t \sigma(y, t)}{\rho_0(y)} dy. \tag{4-112}$$

Due to the asymptotics (1-24), we now have that

$$\int_0^M \frac{dy}{\rho_0(y)} < \infty. \tag{4-113}$$

This and the estimates (4-110) then imply that $\dfrac{v}{r} \in C^0$ and

$$\|r^2 v\|_{L^\infty} \leq C \left\| \partial_t \frac{\sigma}{\rho_0} \right\|_{L^\infty} \leq C\sqrt{\mathscr{E}}. \tag{4-114}$$

On the other hand, due to L'Hospital, we have that

$$\frac{1}{r^3(x, t)} \int_0^x \frac{dy}{\rho_0(y)} \sim \frac{4\pi\rho(x, t)}{3\rho_0(x)} = \frac{4\pi}{3}\left(1+\frac{\sigma(x, t)}{\rho_0(y)}\right) < \infty \quad \text{for } x \sim 0, \tag{4-115}$$

which means that

$$\sup_{x\in(0,M)} \frac{1}{r^3(x, t)} \int_0^x \frac{dy}{\rho_0(y)} < \infty. \tag{4-116}$$

We may then deduce that $\dfrac{v}{r} \in C^0$ and

$$\left\| \frac{v}{r} \right\|_{L^\infty} \leq C \left\| \partial_t \frac{\sigma}{\rho_0} \right\|_{L^\infty} \sup_{x\in(0,M)} \frac{1}{r^3(x, t)} \int_0^x \frac{dy}{\rho_0(y)} \leq C\sqrt{\mathscr{E}}. \tag{4-117}$$

Now we apply ∂_t to (4-4) and argue as above to see that

$$\begin{aligned}\big(r^2 \partial_t v\big)(x, t) = &-\int_0^x \frac{1}{4\pi\rho_0(y)\left(1+\dfrac{\sigma(y, t)}{\rho_0(y)}\right)^2} \frac{\partial_t^2 \sigma(y, t)}{\rho_0(y)} dy \\ &+ \int_0^x \frac{1}{2\pi\rho_0(y)\left(1+\dfrac{\sigma(y, t)}{\rho_0(y)}\right)^3} \left| \frac{\partial_t \sigma(y, t)}{\rho_0(y)} \right|^2 dy - \int_0^x 2(r^2 v)(y, t) \frac{v(y, t)}{r(y, t)} dy.\end{aligned} \tag{4-118}$$

Using this, we may argue as above (using estimates (4-114) and (4-117)) to deduce $r^2 \partial_t v$, $\dfrac{\partial_t v}{r} \in C^0$ and

$$\|r^2 \partial_t v\|_{L^\infty} + \left\| \frac{\partial_t v}{r} \right\|_{L^\infty} \leq C\sqrt{\mathscr{E}}. \tag{4-119}$$

An iterative argument, using ∂_t^2 applied to (4-4) in conjunction with the estimates (4-119), then allows us to see that $r^2\partial_t^2 v$, $\partial_t^2 v/r \in C^0$ with

$$\|r^2\partial_t^2 v\|_{L^\infty} + \left\| \frac{\partial_t^2 v}{r} \right\|_{L^\infty} \le C\sqrt{\mathscr{E}}. \tag{4-120}$$

Then (4-114), (4-117), and (4-119)–(4-120) may be combined to derive the $\dfrac{\partial_t^k v}{r}$ estimates recorded in (4-103).

Step 4 ($\rho r^3 \partial_x(\sigma/\rho_0)$ estimate). Since $\|\sigma/\rho_0\|_{L^\infty} \le \frac{1}{2}$, to prove the $\rho r^3 \partial_x(\sigma/\rho_0)$ estimate listed in (4-103), it suffices to estimate this term with ρ replaced by ρ_0. We claim that

$$\left\| \rho_0 r^3 \partial_x\left(\frac{\sigma}{\rho_0}\right) \right\|_{L^\infty} \le C\left(\sqrt{\mathscr{E}_b^{1,\sigma}} + \sqrt{\mathscr{E}^4}\right) \le C\sqrt{\mathscr{E}}. \tag{4-121}$$

To prove (4-121), we will use the one-dimensional Sobolev embedding $W^{1,1} \hookrightarrow C^0$. First note that

$$\int \rho_0 r^3 \left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right| dx \le \left(\int \rho\rho_0^2 r^2 dx\right)^{1/2} \left(\int \frac{r^2}{\rho}\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2 dx\right)^{1/2} \le C\sqrt{\mathscr{E}_b^{1,\sigma}}. \tag{4-122}$$

On the other hand, we may compute

$$\begin{aligned} \partial_x\left(\rho_0 r^3 \partial_x\left(\frac{\sigma}{\rho_0}\right)\right) &= \frac{\rho_0}{r}\partial_x\left(r^4 \partial_x\left(\frac{\sigma}{\rho_0}\right)\right) + \partial_x\rho_0 r^3 \partial_x\left(\frac{\sigma}{\rho_0}\right) - \frac{\rho_0}{4\pi\rho}\partial_x\left(\frac{\sigma}{\rho_0}\right), \\ \partial_x \rho_0 &= -\frac{x}{4\pi K\gamma\rho_0^{\gamma-1} r_0^4}. \end{aligned} \tag{4-123}$$

Then since $\dfrac{\rho_0}{\rho_0^{2\gamma-2}} \le \dfrac{C}{\rho_0}$ as long as $\gamma < 2$, we may estimate

$$\begin{aligned} &\int \rho_0 r_0^2 \left|\partial_x\left(\rho_0 r^3 \partial_x\left(\frac{\sigma}{\rho_0}\right)\right)\right|^2 dx \\ &\quad \le C\int \rho_0 r_0^2 \left[\frac{\rho_0^2}{r^2}\left|\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho_0}\right)\right)\right|^2 + \frac{\rho_0^2}{\rho^2}\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2 + \frac{x^2 r^6}{\rho_0^{2\gamma-2} r_0^8}\left|\partial_x\left(\frac{\sigma}{\rho_0}\right)\right|^2\right] dx \\ &\quad \le C\mathscr{E}^4 + \mathscr{E}_b^{1,\sigma}. \end{aligned} \tag{4-124}$$

Then from this and Hölder's inequality, we get

$$\begin{aligned} \int \left|\partial_x\left(\rho_0 r^3 \partial_x\left(\frac{\sigma}{\rho_0}\right)\right)\right| dx &\le \left(\int \frac{dx}{\rho_0 r_0^2}\right)^{1/2} \left(\int \rho_0 r_0^2 \left|\partial_x\left(\rho_0 r^3 \partial_x\left(\frac{\sigma}{\rho_0}\right)\right)\right|^2 dx\right)^{1/2} \\ &\le C\sqrt{\mathscr{E}^4} + \sqrt{\mathscr{E}_b^{1,\sigma}}. \end{aligned} \tag{4-125}$$

Together, the estimates (4-123) and (4-125) constitute a $W^{1,1}$ estimate for $\rho_0 r^3 \partial_x\left(\dfrac{\sigma}{\rho_0}\right)$, so we then obtain (4-121) via the Sobolev embedding. □

5. Nonlinear instability

5A. ***The bootstrap argument.*** Based on the nonlinear estimates in the previous section, we now establish a bootstrap argument that allows us to control the growth of $\mathscr{E}$ in terms of the linear growth rate λ, constructed in Theorem 2.1. The idea is to assume small data and that the lowest-order energy, $\mathscr{E}^0$, grows no faster than the linear growth rate; then the inequalities in the last section allow for a bootstrap argument that shows that all of $\mathscr{E}$ grows no faster than the linear growth rate.

Proposition 5.1. *Let σ and v be a solution to the Navier–Stokes–Poisson system* (4-2). *Assume that*

$$\sqrt{\mathscr{E}}(0) \le C_0\iota \quad \text{and} \quad \sqrt{\mathscr{E}^0}(t) \le C_0\iota e^{\lambda t} \quad \text{for } 0 \le t \le T, \tag{5-1}$$

where $\mathscr{E}^0$ and $\mathscr{E}$ are as defined in (4-11) *and* (4-19). *Then there exist $C_\star$ and $\theta_\star > 0$ such that if $0 \le t \le \min\{T, T(\iota, \theta_\star)\}$, then*

$$\sqrt{\mathscr{E}}(t) \le C_\star \iota e^{\lambda t} \le C_\star \theta_\star, \tag{5-2}$$

where we have written $T(\iota, \theta_\star) = \dfrac{1}{\lambda} \ln \dfrac{\theta_\star}{\iota}$.

Proof. To prove the result, we will employ a bootstrap argument using all of the nonlinear energy estimates derived in the previous section. We now choose θ_1 and η sufficiently small in all of these estimates that $C\theta_1 + \eta \le \lambda/2$ and $C\theta_1 \le \frac{1}{8}$ in all of the energy inequalities. Throughout this proof, we will write $\tilde{C}$ for a generic constant; we write this in place of C to distinguish the constants from those appearing in the nonlinear energy estimates.

To begin the bootstrapping, we show that the estimate (5-1) allows us to control an integral of the $\mathscr{D}^0$ dissipation. Indeed, we use (4-25) and (5-1) along with Gronwall's inequality to see that for $0 \le t \le T$,

$$\frac{d}{dt}\mathscr{E}^0 + \frac{1}{2}\mathscr{D}^0 \le C\mathscr{E}^0 \le CC_0^2(\iota e^{\lambda t})^2 + \frac{\lambda}{2}\mathscr{E}^0 \implies \frac{1}{2}\int_0^t e^{\lambda/2(t-s)}\mathscr{D}^0(s)\,ds \le \tilde{C}\iota^2 e^{2\lambda t}. \tag{5-3}$$

Then we employ (4-34) with $\beta = -1$ in conjunction with (5-3) to see that

$$\begin{aligned}\frac{d}{dt}\big(e^{-t\lambda/2}\mathscr{E}^{0,\sigma}_{-1}(t)\big) \le Ce^{-t\lambda/2}\mathscr{D}^0(t) &\implies \mathscr{E}^{0,\sigma}_{-1}(t) \le \mathscr{E}^{0,\sigma}_{-1}(0)e^{t\lambda/2} + C\int_0^t e^{\lambda/2(t-s)}\mathscr{D}^0(s)\,ds \\ &\implies \mathscr{E}^{0,\sigma}_{-1}(t) \le \tilde{C}\iota^2 e^{2\lambda t} \implies \mathscr{E}^{0,\sigma}_{0}(t) \le \tilde{C}\mathscr{E}^{0,\sigma}_{-1}(t) \le \tilde{C}\iota^2 e^{2\lambda t}.\end{aligned} \tag{5-4}$$

Next, let $q > 0$ be the constant from estimate (4-35) and choose $k = (4q)/\lambda$. Then (4-35) and (4-47), together with (4-25) and the above estimates, imply that for $0 \le t \le T$,

$$\begin{aligned}\frac{d}{dt}\big[k\mathscr{E}^2 + k\mathscr{E}^0 + \mathscr{E}^1\big] + \frac{k}{2}\mathscr{D}^2 &\le (\eta + C\theta_1)\big(\mathscr{E}^1 + k\mathscr{E}^2\big) + q\mathscr{E}^2 + C\big(\mathscr{E}^0 + \mathscr{E}^{0,\sigma}_0\big) \\ &\le \frac{\lambda}{2}\big(\mathscr{E}^1 + k\mathscr{E}^2 + k\mathscr{E}^0\big) + \tilde{C}\iota^2 e^{2\lambda t}.\end{aligned} \tag{5-5}$$

Using Gronwall's inequality again, we obtain from this that for $0 \le t \le T$,

$$\mathscr{E}^2(t) + \mathscr{E}^1(t) + \frac{1}{2}\int_0^t e^{\lambda/2(t-s)}\mathscr{D}^2(s)\,ds \le \tilde{C}\iota^2 e^{2\lambda t}. \tag{5-6}$$

Here we have used the fact that k is bounded and nonzero to absorb it into the constant $\tilde{C}$. We then employ (4-56) with $\beta = -1$ to see that

$$\frac{d}{dt}\big(e^{-t\lambda/2}\mathscr{E}^{2,\sigma}_{-1}(t)\big) \le Ce^{-t\lambda/2}\big(\mathscr{D}^2(t)+\mathscr{D}^0(t)\big)$$
$$\implies \mathscr{E}^{2,\sigma}_{-1}(t) \le \mathscr{E}^{2,\sigma}_{-1}(0)e^{t\lambda/2} + C\int_0^t e^{\lambda/2(t-s)}(\mathscr{D}^2(s)+\mathscr{D}^0(s))\,ds$$
$$\implies \mathscr{E}^{2,\sigma}_{-1}(t) \le \tilde{C}\iota^2 e^{2\lambda t} \implies \mathscr{E}^{2,\sigma}_{0}(t) \le \tilde{C}\mathscr{E}^{2,\sigma}_{-1}(t) \le \tilde{C}\iota^2 e^{2\lambda t}. \tag{5-7}$$

Bootstrapping further, (4-57) gives rise to

$$\mathscr{E}^3(t) \le \tilde{C}\iota^2 e^{2\lambda t}. \tag{5-8}$$

Similarly, from (4-72), (4-73), (4-74), (4-76), and (4-77), we also obtain, for $0 \le t \le T$,

$$\mathscr{E}^{0,r}_b(t) + \mathscr{E}^{1,\sigma}_b(t) + \mathscr{E}^{3,\sigma}_a(t) + \mathscr{E}^{3,v}_{a_1}(t) + \mathscr{E}^{3,v}_{a_2}(t) \le \tilde{C}\iota^2 e^{2\lambda t}. \tag{5-9}$$

Next, from (4-78), we get

$$\frac{d}{dt}\mathscr{E}^4 \le (\eta + C\theta_1)\mathscr{E}^4 + C\iota^2 e^{2\lambda t} \implies \mathscr{E}^4(t) \le \tilde{C}\iota^2 e^{2\lambda t}. \tag{5-10}$$

In turn, from (4-86) and (4-88), we find that

$$\mathscr{E}^4_{a_1}(t) + \mathscr{E}^4_{a_2}(t) \le \tilde{C}\iota^2 e^{2\lambda t}. \tag{5-11}$$

Similarly, the energy inequalities (4-89), (4-91), (4-93) yield

$$\mathscr{E}^5(t) + \mathscr{E}^5_a(t) + \mathscr{E}^{5,\sigma}_{-1}(t) \le \tilde{C}\iota^2 e^{2\lambda t}, \tag{5-12}$$

and (4-97) and (4-96) yield

$$\mathscr{E}^6(t) + \mathscr{E}^{6,\sigma}_a(t) \le \tilde{C}\iota^2 e^{2\lambda t}. \tag{5-13}$$

Successively, (4-98), (4-99), and (4-100) imply

$$\mathscr{E}^7(t) + \mathscr{E}^7_a(t) + \mathscr{E}^{7,\sigma}_{-1}(t) \le \tilde{C}\iota^2 e^{2\lambda t}. \tag{5-14}$$

To get the bound of $\mathscr{E}^8$, we first note that $\mathscr{E}^8$ satisfies the following inequality from (4-101) and (4-102):

$$\frac{d}{dt}\big[\mathscr{E}^8 + \widetilde{F}_*\big] \le (\eta + C\theta_1)\big(\mathscr{E}^8 + \widetilde{F}_*\big) + C_1|\mathscr{E}|^2 + \widetilde{C}\iota^2 e^{2\lambda t}, \tag{5-15}$$

for some constants $C_1 > 0$ and $\widetilde{C} > 0$. We now define T^* by

$$T^* := \sup\left\{t \;\middle|\; \mathscr{E}(s) \le \min\left\{\theta_1, \frac{\lambda}{4C_1}\right\} \text{ for } s \in [0,t]\right\}. \tag{5-16}$$

Let $0 \le t \le \min\{T, T^*\}$. Then by the Gronwall inequality, (5-15) implies that

$$\mathscr{E}^8(t) + \widetilde{F}_*(t) \le \tilde{C}\iota^2 e^{2\lambda t} \implies \mathscr{E}^8(t) + \mathscr{E}^{8,\sigma}_a(t) \le \tilde{C}\iota^2 e^{2\lambda t} \quad \text{for } 0 \le t \le \min\{T, T^*\}. \tag{5-17}$$

Thus, combining all of the above analysis, we finally obtain

$$\mathscr{E}(t) \le C_2 \iota^2 e^{2\lambda t} \quad \text{for } 0 \le t \le \min\{T, T^*\} \tag{5-18}$$

for a constant $C_2 > 0$ independent of ι.

We now choose $\theta_\star$ such that $C_2(\theta_\star)^2 < \min\{\theta_1, \lambda/4C_1\}$. We consider the following two cases.

(i) $T(\iota, \theta_\star) \le \min\{T, T^*\}$. In this case, the conclusion follows without any additional work.

(ii) $T(\iota, \theta_\star) > \min\{T, T^*\}$. We claim that it must hold that $T \le T^* < T(\iota, \theta_*)$, in which case the conclusion directly follows. To prove the claim, we note that otherwise we would have $T^* < T < T(\iota, \theta_\star)$. Letting $t = T^*$, from (5-18), we get

$$\mathscr{E}(T^*) \le C_2 \iota^2 e^{2\lambda T^*} < C_2 \iota^2 e^{2\lambda T^\iota} = C_2(\theta_\star)^2 \quad \text{by the definition of } T(\iota, \theta_\star), \tag{5-19}$$

but this is impossible due to our choice of $\theta_\star$ since it would then contradict the definition of T^*. Since we then find our desired estimate in both cases, this concludes the proof of the proposition. ☐

5B. ***Further nonlinear estimates.*** As preparation for the proof of our main theorem, we recall that the Navier–Stokes–Poisson system (4-2) can be written in perturbed form as in (3-38) and (3-39) in terms of σ and $w := r^2 v$:

$$\partial_t \begin{pmatrix} \sigma \\ w \end{pmatrix} - \mathscr{L}\begin{pmatrix} \sigma \\ w \end{pmatrix} = \begin{pmatrix} N_1 \\ N_2 \end{pmatrix}, \tag{5-20}$$

with the boundary conditions

$$\left(\frac{w}{r^2}\right)(0, t) = 0 \quad \text{and} \quad \sigma(M, t) = 0, \quad \mathscr{B}(w) = N_{\mathscr{B}} \quad \text{at} \quad x = M, \tag{5-21}$$

where the boundary operator $\mathscr{B}(w)$ is defined by (3-34), $N_{\mathscr{B}}$ is given as

$$N_{\mathscr{B}} = \left\{ \left(\delta + \frac{4\varepsilon}{3}\right) 4\pi \sigma \partial_x w - 4\varepsilon \left[\left(\frac{r_0}{r}\right)^3 - 1 \right] \frac{w}{r_0^3} \right\} \Bigg|_{x=M}, \tag{5-22}$$

and N_1 and N_2 become

$$\begin{aligned} N_1 &= -4\pi(2\rho_0 + \sigma)\sigma \partial_x w, \\ N_2 &= \frac{2w^2}{r^3} - 4\pi\left(r^4 - r_0^4\right)\partial_x\left(K\gamma\rho_0^\gamma \frac{\sigma}{\rho_0}\right) - 4\pi r^4 \partial_x\left(\rho_0^\gamma a_*\left(\frac{\sigma}{\rho_0}\right)^2\right) - M_1 - M_2, \end{aligned} \tag{5-23}$$

where

$$\begin{aligned} M_1 &= \frac{x}{r_0^4}\left(r_0^4 - r^4 - \frac{r_0}{\pi}\int_0^x \frac{\sigma}{\rho_0^2}\,dy\right) \\ &= x\left\{\frac{c_1}{r_0^3}\int_0^x \frac{1}{\rho_*}\left(\frac{\sigma}{\rho_0}\right)^2 dy + \frac{c_2}{r_0^6}\left(\int_0^x \frac{\sigma}{\rho_0^2}dy\right)^2\right\} \quad \text{by Taylor expansion,} \\ &\qquad \text{where } \rho_*/\rho_0 \sim 1 \text{ is a bounded smooth function of } \frac{\sigma}{\rho_0}, \end{aligned} \tag{5-24}$$

and

$$\begin{aligned} M_2 &= 16\pi^2\left(\delta+\frac{4\varepsilon}{3}\right)\left\{r_0^4\partial_x(\rho_0\partial_x w)-r^4\partial_x(\rho\partial_x w)\right\} \\ &= 16\pi^2\left(\delta+\frac{4\varepsilon}{3}\right)\left\{(r_0^4-r^4)\partial_x(\rho_0\partial_x w)-r^4\partial_x(\sigma\partial_x w)\right\}. \end{aligned} \tag{5-25}$$

It is possible to estimate these nonlinearities in terms of the energy $\mathscr{E}$ given by (4-19). We present these estimates now.

Lemma 5.2. *For each t,*

$$\mathfrak{E}(N_1,N_2)\le C|\mathscr{E}|^2 \quad \textit{and} \quad |N_{\mathscr{B}}|+|\partial_t N_{\mathscr{B}}|+|\partial_t^2 N_{\mathscr{B}}|\le C|\mathscr{E}|^2, \tag{5-26}$$

where $\mathfrak{E}$ is defined in (3-46).

Proof. The second inequality follows directly from Lemma 4.9. For the first inequality, we only provide the details for the highest-order nonlinear term M_2 in N_2. Lower-order terms may be estimated similarly. Throughout the proof, we will write θ_1 to denote the left side of estimate (4-103); Lemma 4.9 then implies that $\theta_1\le C\sqrt{\mathscr{E}}$.

By rewriting M_2 as

$$\begin{aligned} &\frac{M_2}{16\pi^2\left(\delta+\frac{4\varepsilon}{3}\right)} \\ &\quad= (r_0^4-r^4)\partial_x(\rho_0\partial_x w)-r^4\partial_x(\sigma\partial_x w) \\ &\quad= \left[\left(\frac{r_0}{r}\right)^4-1\right]\left\{\frac{\rho_0}{\rho}r^4\partial_x(\rho\partial_x w)+r^4\partial_x\left(\frac{\rho_0}{\rho}\right)\rho\partial_x w\right\}-\frac{\sigma}{\rho}r^4\partial_x(\rho\partial_x w)-r^4\partial_x\left(\frac{\sigma}{\rho}\right)\rho\partial_x w, \end{aligned} \tag{5-27}$$

it is easy to see that

$$\int\frac{|N_2|^2}{r_0^4}\,dx\le C\theta_1^2\mathscr{E}\le C|\mathscr{E}|^2. \tag{5-28}$$

Next,

$$\begin{aligned} \frac{\partial_x M_2}{16\pi^2\left(\delta+\frac{4\varepsilon}{3}\right)} &= \left(\frac{r_0}{r}\right)^3\frac{1-\left(\frac{r_0}{r}\right)^3+\frac{\sigma}{\rho_0}}{\pi\rho r_0^2 r}\left\{\frac{\rho_0}{\rho}r^4\partial_x(\rho\partial_x w)+r^4\partial_x\left(\frac{\rho_0}{\rho}\right)\rho\partial_x w\right\} \\ &\quad+\left[\left(\frac{r_0}{r}\right)^4-1\right]\left\{\frac{\rho_0}{\rho}\partial_x\big(r^4\partial_x(\rho\partial_x w)\big)+2r^4\partial_x\left(\frac{\rho_0}{\rho}\right)\partial_x(\rho\partial_x w)+\partial_x\left(r^4\partial_x\left(\frac{\rho_0}{\rho}\right)\right)\rho\partial_x w\right\} \\ &\quad-\frac{\sigma}{\rho}\partial_x\big(r^4\partial_x(\rho\partial_x w)\big)-2r^4\partial_x\left(\frac{\sigma}{\rho}\right)\partial_x(\rho\partial_x w)-\partial_x\left(r^4\partial_x\left(\frac{\sigma}{\rho}\right)\right)\rho\partial_x w. \end{aligned} \tag{5-29}$$

Hence, from the definition of the energies and from the estimates in the previous section,

$$\int\rho_0|\partial_x M_2|^2\,dx\le C\theta_1^2\big(\mathscr{E}_{a_2}^4+\mathscr{E}_{a_1}^{3,v}+\mathscr{E}_b^{1,\sigma}\big)+C\theta_1^4\big(\mathscr{E}^4+\mathscr{E}_b^{1,\sigma}\big)+C\big(\mathscr{E}_b^{1,\sigma}+\mathscr{E}^4\big)\mathscr{E}_{a_1}^{3,v}\le C|\mathscr{E}|^2. \tag{5-30}$$

On the other hand, $\partial_t M_2$ reads as

$$
\begin{aligned}
\frac{\partial_t M_2}{16\pi^2\left(\delta+\dfrac{4\varepsilon}{3}\right)} &= -4r^3 v\partial_x(\rho\partial_x w) - r^4\partial_x(\partial_t\sigma\,\partial_x w) - r^4\partial_x(\sigma\,\partial_x\partial_t w) + (r_0^4 - r^4)\partial_x(\rho_0\partial_x\partial_t w) \\
&= -4\frac{v}{r}r^4\partial_x(\rho\partial_x w) - \frac{\partial_t\sigma}{\rho}r^4\partial_x(\rho\partial_x w) - r^4\partial_x\left(\frac{\partial_t\sigma}{\rho}\right)\rho\partial_x w - r^4\frac{\sigma}{\rho}\partial_x(\rho\partial_x\partial_t w) \\
&\qquad - r^4\partial_x\left(\frac{\sigma}{\rho}\right)\rho\partial_x\partial_t w + (r_0^4 - r^4)\frac{\rho_0}{\rho}\partial_x(\rho\partial_x\partial_t w) + (r_0^4 - r^4)\partial_x\left(\frac{\rho_0}{\rho}\right)\rho\partial_x\partial_t w. \qquad (5\text{-}31)
\end{aligned}
$$

Thus

$$
\int \frac{|\partial_t M_2|^2}{r_0^4}\,dx \le C\theta_1^2\big(\mathscr{E}_{a_1}^{3,v} + \mathscr{E}_a^{3,\sigma} + \mathscr{E}_a^5\big) + C\big(\mathscr{E}^4 + \mathscr{E}_b^{1,\sigma}\big)\mathscr{E}_a^{6,\sigma} \le C|\mathscr{E}|^2. \qquad \square
$$

5C. ***Data analysis.*** In order to prove our nonlinear instability result, we want to use the linear growing mode solutions constructed in Theorem 2.1 to construct small initial data for the nonlinear problem, written in the perturbation formulation (5-20). Small data in the perturbation formulation correspond to initial data for (1-17)–(1-20) that are close to the stationary solutions $\rho = \rho_0$, $v = 0$, $r = r_0$. Unfortunately, due to the regularity framework (given by $\mathscr{E}$ as in (4-19)) in which we have proved our nonlinear estimates, we cannot simply set the initial data for the nonlinear problem (5-20) to be a small constant times the linear growing modes. The reason for this is that the initial data for the nonlinear problem must satisfy certain nonlinear compatibility conditions in order for us to guarantee local existence in the energy space defined by $\mathscr{E}$. Until now, we have taken the local well-posedness theory for the nonlinear problem for granted, but we must now say a few words about the compatibility conditions in order to construct our desired initial data.

Recall that we can rewrite the nonlinear problem (1-17)–(1-20) in the form (5-20)–(5-21) with nonlinearities given by (5-22)–(5-23). Let us concisely rewrite (5-20) as

$$
\partial_t\mathfrak{X} + \mathscr{L}\mathfrak{X} = \mathfrak{N}(\mathfrak{X}) \quad \text{for } \mathfrak{X} = \begin{pmatrix}\sigma \\ \bar{w}\end{pmatrix}, \qquad (5\text{-}32)
$$

where $\mathfrak{N}(\mathfrak{X})$ is the nonlinearity given in terms of N_1 and N_2 by the right side of (5-20). We will also rewrite the boundary conditions (5-21) as

$$
\mathfrak{C}(\mathfrak{X}) := \begin{pmatrix}(w/r_0^2)|_{x=0} \\ \sigma|_{x=M} \\ \mathscr{B}(w)|_{x=M}\end{pmatrix} = \begin{pmatrix}w(r_0^{-2} - r^{-2})|_{x=0} \\ 0 \\ N_{\mathscr{B}}\end{pmatrix} := \mathfrak{N}_{\mathscr{B}}(\mathfrak{X}). \qquad (5\text{-}33)
$$

Here r is determined as a nonlinear function of σ as usual.

Rewriting the nonlinear problem as (5-32)–(5-33) now allows us to easily describe the compatibility conditions for the initial data. Given $\mathfrak{X}(0)$ as initial data for $\mathfrak{X}$ at $t = 0$, we can use (5-32) to iteratively solve for $\partial_t^j \mathfrak{X}(0)$ for $j \geq 1$:

$$\begin{aligned}\partial_t \mathfrak{X}(0) &= -\mathscr{L}\mathfrak{X}(0) + \mathfrak{N}(\mathfrak{X}(0)),\\ \partial_t^2 \mathfrak{X}(0) &= -\mathscr{L}\partial_t\mathfrak{X}(0) + D\mathfrak{N}(\mathfrak{X}(0))\cdot\partial_t\mathfrak{X}(0)\\ &= -\mathscr{L}\bigl(-\mathscr{L}\mathfrak{X}(0) + \mathfrak{N}(\mathfrak{X}(0))\bigr) + D\mathfrak{N}(\mathfrak{X}(0))\cdot\bigl(-\mathscr{L}\mathfrak{X}(0) + \mathfrak{N}(\mathfrak{X}(0))\bigr),\end{aligned} \tag{5-34}$$

and so on for higher derivatives, where D is the derivative of the nonlinearity. We may similarly compute $\partial_t^j \mathfrak{N}_{\mathscr{B}}(\mathfrak{X})(0)$:

$$\partial_t \mathfrak{N}_{\mathscr{B}}(\mathfrak{X})(0) = D\mathfrak{N}_{\mathscr{B}}(\mathfrak{X}(0))\cdot\partial_t X(0) = D\mathfrak{N}_{\mathscr{B}}(\mathfrak{X}(0))\cdot\bigl[-\mathscr{L}\mathfrak{X}(0) + \mathfrak{N}(\mathfrak{X}(0))\bigr]\big|_{x=M}, \tag{5-35}$$

continuing as above for higher derivatives. This procedure may be carried out indefinitely as long as $\mathfrak{X}(0)$ is sufficiently smooth. However, we may also differentiate the boundary condition (5-33) with respect to time and then set $t = 0$ to see that the data must satisfy the boundary conditions

$$\mathfrak{C}(\partial_t^j \mathfrak{X}(0)) = \partial_t^j \mathfrak{N}_{\mathscr{B}}(\mathfrak{X})(0) \quad \text{for } j \geq 0. \tag{5-36}$$

Since the terms $\partial_t^j \mathfrak{X}(0)$ and $\partial_t \mathfrak{N}_{\mathscr{B}}(\mathfrak{X})(0)$ constructed in (5-34)–(5-35) are determined entirely by $\mathfrak{X}(0)$, we then find that the data $\mathfrak{X}(0)$ must satisfy the nonlinear compatibility conditions given by substituting (5-34)–(5-35) into (5-36).

For completely smooth solutions to the nonlinear problem, the compatibility conditions would have to hold for all $j \geq 0$. In our case, we only require solutions to remain in the energy space defined by $\mathscr{E}$, and as such, we must only solve for $\partial_t^j \mathfrak{X}(0)$ for $j = 1, 2, 3$, given $\mathfrak{X}(0)$. This then requires the compatibility condition from (5-36) only for $0 \leq j \leq 3$. Of course, in order to guarantee that $\mathscr{E}(0)$ is finite, we must have that $\partial_t^j \mathfrak{X}(0)$, $0 \leq j \leq 3$, satisfies the integrability conditions in the definition of $\mathscr{E}(0)$. This in turn gives us a natural Hilbert function space $\mathbb{H}$ with the following three properties. First, if $\mathfrak{X}(0) \in \mathbb{H}$, then we have the trace estimates needed to make sense of the boundary conditions in (5-36) for $0 \leq j \leq 3$. Second, if $\|\mathfrak{X}(0)\|_{\mathbb{H}}$ is sufficiently small, then

$$\mathscr{E}(0) \leq C\|\mathfrak{X}(0)\|_{\mathbb{H}}^2$$

for some $C > 0$. Here the smallness assumption is needed to deal with the nonlinearities in (5-34)–(5-35) and the r terms in $\mathscr{E}$. Third, the linear growing modes produced in Theorem 2.1 are in $\mathbb{H}$. It is straightforward to extract the proper definition of $\mathbb{H}$ from $\mathscr{E}$ and to work out the details of the estimate of $\mathscr{E}(0)$; as such, for the sake of brevity, we omit these. With $\mathbb{H}$ defined in this way, it is then easy to use estimate (2-13) of Theorem 2.1 in conjunction with (2-8)–(2-10) to see that the growing modes are in $\mathbb{H}$.

Now that we have stated the nonlinear compatibility conditions, we see why we cannot simply set $\mathfrak{X}(0) = \iota\mathfrak{X}_0$ with

$$\mathfrak{X}_0 = \begin{pmatrix}\sigma_\star\\ \bar{w}_\star\end{pmatrix} \tag{5-37}$$

for $\sigma_\star$ and $v_\star = \bar{w}_\star / r_0^2$ the growing mode solution constructed in Theorem 2.1 and $\iota > 0$ a small parameter. Indeed, these solve

$$\lambda\mathfrak{X}_0 + \mathscr{L}\mathfrak{X}_0 = 0 \ \text{ and } \ \mathfrak{C}(\mathfrak{X}_0) = 0 \qquad \Longrightarrow \qquad \mathfrak{C}(\mathscr{L}^j\mathfrak{X}_0) = 0 \quad \text{for all } j \geq 0, \tag{5-38}$$

which in particular means that $\mathfrak{X}(0) = \iota\mathfrak{X}_0$ does not satisfy the nonlinear compatibility condition (5-36) for $j \geq 1$.

To get around this obstacle, we will use the implicit function theorem to produce a curve of initial data satisfying the compatibility conditions, close to the linear growing modes. To this end, let us define the map $F : \mathbb{H} \to \mathbb{R}^{12}$ via

$$F(\mathfrak{X}) = \begin{pmatrix} \mathfrak{C}(\mathfrak{X}) \\ \mathfrak{C}(\partial_t \mathfrak{X}) \\ \mathfrak{C}(\partial_t^2 \mathfrak{X}) \\ \mathfrak{C}(\partial_t^3 \mathfrak{X}) \end{pmatrix} - \begin{pmatrix} \mathfrak{N}_{\mathscr{B}}(\mathfrak{X}) \\ \partial_t \mathfrak{N}_{\mathscr{B}}(\mathfrak{X}) \\ \partial_t^2 \mathfrak{N}_{\mathscr{B}}(\mathfrak{X}) \\ \partial_t^3 \mathfrak{N}_{\mathscr{B}}(\mathfrak{X}) \end{pmatrix}, \tag{5-39}$$

where we understand that $\partial_t^j \mathfrak{X}$ and $\partial_t^j \mathfrak{N}_{\mathscr{B}}(\mathfrak{X})$ for $j = 1, 2, 3$ are computed in terms of $\mathfrak{X}$ as in (5-34)–(5-35). Let $\mathfrak{X}_0$ be the linear growing modes as above and let $\mathfrak{X}_i \in \mathbb{H}$, $i = 1, \dots, 12$, be arbitrary for now, with exact values to be chosen later. We then define $f : \mathbb{R}^{1+12} \to \mathbb{R}^{12}$ via

$$f(t, \tau) = F\left(t\mathfrak{X}_0 + \sum_{i=1}^{12} \tau_i \mathfrak{X}_i \right) \quad \text{for } t \in \mathbb{R} \text{ and } \tau \in \mathbb{R}^{12}. \tag{5-40}$$

Given the structure of the nonlinearities $\mathfrak{N}(\,\cdot\,)$ and $\mathfrak{N}_{\mathscr{B}}(\,\cdot\,)$, one easily sees that $f \in C^2(\mathbb{R}^{1+12}; \mathbb{R}^{12})$. Also, $f(0, 0) = 0$ and

$$\frac{\partial f}{\partial t}(0, 0) = \begin{pmatrix} \mathfrak{C}(\mathfrak{X}_0) \\ \mathfrak{C}(\lambda \mathfrak{X}_0) \\ \mathfrak{C}(\lambda^2 \mathfrak{X}_0) \\ \mathfrak{C}(\lambda^3 \mathfrak{X}_0) \end{pmatrix} = 0 \quad \text{and} \quad \frac{\partial f}{\partial \tau_i}(0, 0) = \begin{pmatrix} \mathfrak{C}(\mathfrak{X}_i) \\ \mathfrak{C}(-\mathscr{L}\mathfrak{X}_i) \\ \mathfrak{C}(\mathscr{L}^2 \mathfrak{X}_i) \\ \mathfrak{C}(-\mathscr{L}^3 \mathfrak{X}_i) \end{pmatrix}. \tag{5-41}$$

From this it is then straightforward to choose the $\mathfrak{X}_i$ for $i = 1, \dots, 12$ such that the 12×12 matrix $(\partial f/\partial \tau)(0, 0)$ is invertible. The implicit function theorem then provides a small constant $\iota_0 > 0$ and a function $\xi : (-\iota_0, \iota_0) \to \mathbb{R}^{12}$ such that $f(t, \xi(t)) = 0$ for all $t \in (-\iota_0, \iota_0)$ and such that $\xi \in C^2$ and $\xi(0) = 0$. We may then differentiate the equation $f(t, \xi(t)) = 0$ with respect to t, set $t = 0$, and use the first equation in (5-41) to see that

$$0 = \frac{\partial f}{\partial t}(0, 0) + \frac{\partial f}{\partial \tau}(0, 0) \frac{d\xi(0)}{dt} = \frac{\partial f}{\partial \tau}(0, 0) \frac{d\xi}{dt}(0) \implies \frac{d\xi}{dt}(0) = 0, \tag{5-42}$$

since the matrix $(\partial f/\partial \tau)(0, 0)$ is invertible. Then $\xi \in C^2$ with $\xi(0) = \dot{\xi}(0) = 0$ such that $\xi(t)/t^2$ is well-defined and continuous on $(-\iota_0, \iota_0)$. Using this, we may then deduce the existence of a small parameter $\iota_0 > 0$ and a curve $\mathfrak{Y} : (-\iota_0, \iota_0) \to \mathbb{H}$ given by

$$\mathfrak{Y}(\iota) = \iota\mathfrak{X}_0 + \iota^2 \sum_{i=1}^{12} \mathfrak{X}_i \frac{\xi_i(\iota)}{\iota^2} := \iota\mathfrak{X}_0 + \iota^2 \bar{\mathfrak{Y}}(\iota), \tag{5-43}$$

such that for all $\iota \in (-\iota_0, \iota_0)$,

$$\begin{aligned} &F(\mathfrak{Y}(\iota)) = 0, \text{ that is, } \mathfrak{Y}(\iota) \text{ satisfies the nonlinear compatibility conditions,} \\ &\sqrt{\mathscr{E}(\mathfrak{Y}(\iota))} \le C\|\mathfrak{Y}(\iota)\|_{\mathbb{H}} \le C\iota, \text{ and} \\ &\mathfrak{E}(\bar{\mathfrak{Y}}(\iota)_1, \bar{\mathfrak{Y}}(\iota)_2) \le C, \end{aligned} \tag{5-44}$$

where the norm $\|\cdot\|_0 \le \|\cdot\|_{\mathbb{H}}$ is given by (3-45), the term $\mathfrak{E}$ is defined by (3-46), and in the second line we have written $\mathscr{E}(\mathfrak{Y}(\iota))$ for $\mathscr{E}(0)$ computed from the initial data $\mathfrak{X}(0) = \mathfrak{Y}(\iota)$.

We now recast the above discussion as a lemma.

Lemma 5.3. *Let $\sigma_\star$, $v_\star$ be the growing mode solution constructed in Theorem 2.1, write $\bar{w}_\star = r_0^2 v_\star$, and assume the normalization*

$$\left\| \begin{pmatrix} \sigma_\star \\ \bar{w}_\star \end{pmatrix} \right\|_0 = 1, \tag{5-45}$$

for $\|\cdot\|_0$ the norm defined by (3-45). *Then there exist a number $\iota_0 > 0$ and a family of initial data*

$$\begin{pmatrix} \sigma^\iota(0) \\ w^\iota(0) \end{pmatrix} = \mathfrak{X}(\iota) = \iota \begin{pmatrix} \sigma_\star \\ \bar{w}_\star \end{pmatrix} + \iota^2 \begin{pmatrix} \sigma_0(\iota) \\ w_0(\iota) \end{pmatrix} \tag{5-46}$$

for $\iota \in [0, \iota_0)$ such that the following hold.

(1) *$\mathfrak{X}(\iota)$ satisfies the nonlinear compatibility conditions required for a solution to the nonlinear problem* (5-32) *to exist in the energy space defined by $\mathscr{E}$.*

(2) *If $\mathscr{E}(0)$ denotes the value of $\mathscr{E}$ determined at $t = 0$ from the data $\mathfrak{X}(\iota)$, then $\mathscr{E}(0) \le C\iota^2$ for a constant $C > 0$.*

(3) *For all $\iota \in [0, \iota_0)$, we have*

$$\left\| \begin{pmatrix} \sigma_0(\iota) \\ w_0(\iota) \end{pmatrix} \right\|_0^2 \le \mathfrak{E}(\sigma_0(\iota), \quad w_0(\iota)) \le C \tag{5-47}$$

for a constant $C > 0$ independent of ι, where $\mathfrak{E}$ is given by (3-46).

(4) *Let ψ^ι denote the function given by* (3-40), *with $N_{\mathscr{B}} = N_{\mathscr{B}}(\mathfrak{X}(\iota))$ determined by the data $\mathfrak{X}(\iota)$ at $t = 0$. Then $\bar{w}^\iota(0) = w^\iota(0) - \psi^\iota$ satisfies the homogeneous boundary condition $\mathscr{B}(\bar{w}^\iota(0)) = 0$ and*

$$\left\| \begin{pmatrix} 0 \\ \psi^\iota \end{pmatrix} \right\|_0^2 \le \mathfrak{E}(0, \psi^\iota) \le C\iota^4 \tag{5-48}$$

for a constant $C > 0$ independent of ι.

Proof. Everything except for the last item is proved above. The last item follows from Lemma 3.4 and the fact that $N_{\mathscr{B}}$ is at least a quadratic nonlinearity. □

5D. *Instability.*

We are now ready to prove our main result.

Theorem 5.4. *There exist* $\theta_0 > 0$, $C > 0$, *and* $0 < \iota_0 < \theta_0$ *such that for any* $0 < \iota \le \iota_0$, *there exists a family of solutions* $\sigma^\iota(t)$ *and* $v^\iota(t)$ *to the Navier–Stokes–Poisson system* (4-2) *such that*

$$\sqrt{\mathscr{E}}(0) \le C\iota, \quad \textit{but} \quad \sup_{0\le t\le T^\iota} \sqrt{\mathscr{E}^0}(t) \ge \sup_{0\le t\le T^\iota} \sqrt{\mathscr{E}^{0,\sigma^\iota}(T^\iota)+\mathscr{E}^{0,v^\iota}(T^\iota)} \ge \theta_0. \tag{5-49}$$

Here T^ι *is given by* $T^\iota = \dfrac{1}{\lambda}\ln\dfrac{\theta_0}{\iota}$ *and* $\mathscr{E}^{0,\sigma^\iota}$ *and* $\mathscr{E}^{0,v^\iota}$ *are defined in the first line of* (4-11).

Proof. We divide the proof into steps. At several points in the proof we will restrict the size of θ. Whenever we do so, we assume that ι is also restricted such that $0 < \iota \le \iota_0 \le \theta$. We will choose the value of θ_0 in the final step of the proof.

Step 1 (data and the solutions). Let us assume that ι_0 is as small as the ι_0 appearing in Lemma 5.3, and then let $\mathfrak{X}(\iota)$ for $\iota \le \iota_0$ be the family of initial data for the nonlinear problem (5-32)–(5-33) given in the lemma. For $0 < \iota \le \iota_0$, we now let $\binom{\sigma^\iota}{w^\iota}$ be solutions to the Navier–Stokes–Poisson system (5-32)–(5-33) with a family of initial data

$$\begin{pmatrix}\sigma^\iota\\ w^\iota\end{pmatrix}\Bigg|_{t=0} = \begin{pmatrix}\sigma^\iota(0)\\ w^\iota(0)\end{pmatrix} = \mathfrak{X}(\iota) = \iota\begin{pmatrix}\sigma_\star\\ \bar{w}_\star\end{pmatrix} + \iota^2\begin{pmatrix}\sigma_0(\iota)\\ w_0(\iota)\end{pmatrix}. \tag{5-50}$$

The solution satisfies $\sqrt{\mathscr{E}}(0) \le C\iota$.

Note that since

$$r^\iota(x,0) = \left(\frac{3}{4\pi}\int_0^x \frac{dy}{\rho_0(y)+\iota\sigma_*(y)+\iota^2\sigma_0(\iota)(y)}\right)^{1/3}, \tag{5-51}$$

a Taylor expansion and item (2) of Lemma 5.3 allow us to estimate

$$\left\|1-\frac{r_0(x)}{r^\iota(x,0)}\right\|_{L^\infty}^2 + \frac{\nu}{2}\left\|1-\frac{r_0(x)}{r^\iota(x,0)}\right\|_{L^2}^2 \le A_1\iota^2 \tag{5-52}$$

for a constant $A_1 > 0$ independent of ι. From this, the normalization (5-45), and the estimate (5-47), we may assume that $\iota < \iota_0$ with ι_0 small enough that

$$\frac{\iota}{2} \le \sqrt{\mathscr{E}^{0,\sigma^\iota}(0)+\mathscr{E}^{0,v^\iota}(0)+\mathscr{E}^{1,v^\iota}(0)} + \sqrt{\mathscr{E}^{0,r^\iota}(0)} \le 2\iota. \tag{5-53}$$

Throughout the rest of the proof we will let $\mathscr{E}(t)$ denote the total energy, defined by (4-19), associated to the solutions σ^ι and w^ι at time t.

Step 2 (control of the energy). Let us define the constant

$$B_0 := \left(2 + \frac{27}{8\sqrt{2}\lambda}\|\rho_0\|_{L^\infty}^{1/2}\right). \tag{5-54}$$

It will be useful in determining the time-scale in which instability begins. Indeed, we define T by

$$T := \sup\left\{s \;\middle|\; \sqrt{\mathscr{E}^{0,\sigma^\iota}(t)+\mathscr{E}^{0,v^\iota}(t)+\mathscr{E}^{1,v^\iota}(t)} + \sqrt{\mathscr{E}^{0,r^\iota}(t)} \le (4+B_0)\iota e^{\lambda t} \text{ for } 0 \le t \le s\right\}. \tag{5-55}$$

The estimate (5-53) guarantees that $T > 0$. Then by Proposition 5.1 and (5-53), there exist $C_\star$ and $\theta_\star > 0$ such that for $0 \le t \le \min\{T, T(\iota, \theta_\star)\}$ (with $T(\iota, \theta_\star)$ given in the proposition),

$$\sqrt{\mathscr{E}}(t) \le C_\star \iota e^{\lambda t}. \tag{5-56}$$

Let us assume that $\theta \le \theta_\star$, which means that $T^\iota \le T(\iota, \theta_\star)$, and hence that the estimate (5-56) also holds for $0 \le t \le \min\{T, T^\iota\}$. Let us further assume that θ is small enough that $\sqrt{\mathscr{E}}(t) \le C_\star \theta$ is small enough that the right side of the estimate in Lemma 4.9 is smaller than $\frac{1}{2}$. In particular, this implies that

$$\left\| \frac{\sigma^\iota(t)}{\rho_0} \right\|_{L^\infty} + \left\| 1 - \frac{r_0}{r^\iota(t)} \right\|_{L^\infty} \le \frac{1}{2} \tag{5-57}$$

for all $0 \le t \le \min\{T, T^\iota\}$. By further restricting θ to decrease the bound of the terms in (5-57), and using the identities in (4-22) and (4-23), we can also bound

$$\tfrac{1}{4}\big(\mathscr{E}^{0,\sigma^\iota}(t) + \mathscr{E}^{0,v^\iota}(t) + \mathscr{E}^{1,v^\iota}(t)\big) \le \left\| \begin{pmatrix} \sigma^\iota(t) \\ w^\iota(t) \end{pmatrix} \right\|_0^2 \le 2\big(\mathscr{E}^{0,\sigma^\iota}(t) + \mathscr{E}^{0,v^\iota}(t) + \mathscr{E}^{1,v^\iota}(t)\big) \tag{5-58}$$

for $0 \le t \le T^\iota = \min\{T, T^\iota\}$.

Step 3 (linear estimates for σ^ι and w^ι). Because of the estimate (5-57), the boundary condition $w^\iota/(r^\iota)^2|_{x=0}$ is equivalent to $w^\iota/r_0^2|_{x=0}$. We can then modify the problem (5-32)–(5-33) to have the form (3-41)–(3-42), the latter of which has the homogeneous boundary conditions (3-42). This leads us to consider $\bar{w}^\iota(0) = w^\iota(0) - \psi^\iota$ as in Lemma 5.3, which satisfies $\mathscr{B}(\bar{w}^\iota(0)) = 0$ at $x = M$. We then have that

$$e^{t\mathscr{L}} \begin{pmatrix} \sigma^\iota(0) \\ \bar{w}^\iota(0) \end{pmatrix} = \iota e^{\lambda t} \begin{pmatrix} \sigma_\star \\ \bar{w}_\star \end{pmatrix} + \iota^2 e^{t\mathscr{L}} \begin{pmatrix} \sigma_0(\iota) \\ w_0(\iota) \end{pmatrix} - e^{t\mathscr{L}} \begin{pmatrix} 0 \\ \psi^\iota \end{pmatrix}. \tag{5-59}$$

Then the solutions $\begin{pmatrix} \sigma^\iota \\ w^\iota \end{pmatrix}$ to (5-32) can be written as in (3-44):

$$\begin{aligned} \begin{pmatrix} \sigma^\iota(t) \\ w^\iota(t) \end{pmatrix} = \iota e^{\lambda t} \begin{pmatrix} \sigma_\star \\ \bar{w}_\star \end{pmatrix} + \iota^2 e^{t\mathscr{L}} \begin{pmatrix} \sigma_0(\iota) \\ w_0(\iota) \end{pmatrix} - e^{t\mathscr{L}} \begin{pmatrix} 0 \\ \psi^\iota \end{pmatrix} - \frac{1}{\delta} \begin{pmatrix} 0 \\ N^\iota_{\mathscr{B}}(t) r_0^3/3 \end{pmatrix} \\ + \int_0^t e^{(t-s)\mathscr{L}} \begin{pmatrix} N^\iota_1(s) \\ N^\iota_2(s) \end{pmatrix} ds + \frac{1}{\delta} \int_0^t e^{(t-s)\mathscr{L}} \begin{pmatrix} N^\iota_{\mathscr{B}}(s)\rho_0 \\ \partial_t N^\iota_{\mathscr{B}}(s) r_0^3/3 \end{pmatrix} ds. \end{aligned} \tag{5-60}$$

Here the nonlinear terms $N^\iota_{\mathscr{B}}$, N^ι_1, and N^ι_2 are defined in terms of w^ι and σ^ι via (5-22) and (5-23).

Theorem 3.5, together with the nonlinear estimates of Lemma 5.2, imply that if the inequality $t \le \min\{T, T(\iota, \theta_\star)\}$ holds, then

$$\begin{aligned} \left\| \begin{pmatrix} \sigma^\iota(t) \\ w^\iota(t) \end{pmatrix} - \iota e^{t\lambda} \begin{pmatrix} \sigma_\star \\ \bar{w}_\star \end{pmatrix} - \iota^2 e^{t\mathscr{L}} \begin{pmatrix} \sigma_0(\iota) \\ w_0(\iota) \end{pmatrix} + e^{t\mathscr{L}} \begin{pmatrix} 0 \\ \psi^\iota \end{pmatrix} \right\|_0 &\le C|\mathscr{E}(t)|^2 + C \int_0^t e^{\lambda(t-s)} \mathscr{E}(s)\, ds \\ &\le C(\iota e^{\lambda t})^2 + C \int_0^t e^{\lambda(t-s)} \iota^2 e^{2\lambda s}\, ds \\ &\le A_2 (\iota e^{\lambda t})^2 \end{aligned} \tag{5-61}$$

for a constant $A_2 > 0$ independent of ι. On the other hand, because of the estimates (3-47) and (5-47)–(5-48), we may estimate

$$\begin{aligned}\left\| \iota^2 e^{t\mathscr{L}} \begin{pmatrix}\sigma_0(\iota)\\ w_0(\iota)\end{pmatrix}\right\|_0 + \left\| e^{t\mathscr{L}}\begin{pmatrix}0\\ \psi^\iota\end{pmatrix}\right\|_0 &\le \iota^2 C e^{\lambda t}\sqrt{\mathfrak{E}(\sigma_0(\iota), w_0(\iota))} + Ce^{\lambda t}\sqrt{\mathfrak{E}(0,\psi^\iota)} \\ &\le \iota^2 A_3 e^{\lambda t}\end{aligned} \tag{5-62}$$

for a constant $A_3 > 0$ independent of ι. Then we may then deduce from (5-61)–(5-62) and the normalization (5-45) that

$$\left\| \begin{pmatrix}\sigma^\iota(t)\\ w^\iota(t)\end{pmatrix}\right\|_0 \le \iota e^{\lambda t} + \iota^2 A_3 e^{\lambda t} + A_2(\iota e^{\lambda t})^2. \tag{5-63}$$

Step 4 (control of the r energy). We now turn to control of the term $\mathscr{E}^{0,r^\iota}$. First note that

$$\frac{d}{dt}\int \frac{\nu}{2}\left|1-\frac{r_0}{r^\iota}\right|^2 dx \le \left(\int \nu\left|1-\frac{r_0}{r^\iota}\right|^2 dx\right)^{1/2}\left(\int \nu\left|\frac{w^\iota}{r_0^3}\right|^2 dx\right)^{1/2}\left\|\frac{r_0}{r^\iota}\right\|_{L^\infty}^4, \tag{5-64}$$

which together with (5-57) implies that

$$\frac{d}{dt}\sqrt{\mathscr{E}^{0,r^\iota}(t)} \le \frac{81}{16\sqrt{2}}\left(\int \nu\left|\frac{w^\iota}{r_0^3}\right|^2 dx\right)^{1/2}. \tag{5-65}$$

We may then argue as in (4-23) to see that

$$\int \frac{\nu}{\rho_0}\left|\frac{w^\iota}{r_0^3}\right|^2 dx \le \int \frac{32\pi^2}{9}\left(\delta\rho_0|\partial_x w^\iota|^2 + \frac{4\varepsilon}{3}\rho_0 r_0^6\left|\partial_x\left(\frac{w^\iota}{r_0^3}\right)\right|^2\right) dx \le \frac{4}{9}\left\|\begin{pmatrix}\sigma^\iota(t)\\ w^\iota(t)\end{pmatrix}\right\|_0^2. \tag{5-66}$$

Combining (5-65) and (5-66) with (5-63), we find that

$$\frac{d}{dt}\sqrt{\mathscr{E}^{0,r^\iota}(t)} \le \frac{27}{8\sqrt{2}}\|\rho_0\|_{L^\infty}^{1/2}\left\|\begin{pmatrix}\sigma^\iota(t)\\ w^\iota(t)\end{pmatrix}\right\|_0 \le \frac{27}{8\sqrt{2}}\|\rho_0\|_{L^\infty}^{1/2}\bigl(\iota e^{\lambda t} + A_3\iota^2 e^{\lambda t} + A_2(\iota e^{\lambda t})^2\bigr) \tag{5-67}$$

for $0 \le t \le \min\{T, T^\iota\}$. Integrating this from 0 to $t \le \min\{T, T^\iota\}$ and employing (5-53) then yields the estimate

$$\begin{aligned}\sqrt{\mathscr{E}^{0,r^\iota}(t)} &\le \left(2 + \frac{27}{8\sqrt{2}\lambda}\|\rho_0\|_{L^\infty}^{1/2}\right)\bigl(\iota e^{\lambda t} + A_3\iota^2 e^{\lambda t}\bigr) + \frac{27A_2}{16\sqrt{2}\lambda}\|\rho_0\|_{L^\infty}^{1/2}(\iota e^{\lambda t})^2 \\ &= B_0(\iota e^{\lambda t}) + (\iota A_4)(\iota e^{\lambda t}) + A_5(\iota e^{\lambda t})^2\end{aligned} \tag{5-68}$$

for $0 \le t \le \min\{T, T^\iota\}$, where B_0 is the constant defined above in (5-54) and A_4, A_5 are constants independent of ι.

Step 5 (the bound $T^\iota \le T$). We now claim that if θ is taken to be small enough, then $T^\iota = (1/\lambda)\ln(\theta/\iota) \le T$. Suppose by way of contradiction that $T^\iota > T$. Then the first bounds in (5-58), (5-63), and (5-68) imply that

$$\begin{aligned}\sqrt{\mathscr{E}^{0,\sigma^\iota}(t) + \mathscr{E}^{0,v^\iota}(t) + \mathscr{E}^{1,v^\iota}(t)} + \sqrt{\mathscr{E}^{0,r^\iota}(t)} &\le (2 + B_0 + \iota A_4)(\iota e^{\lambda t}) + (2A_2 + A_5)(\iota e^{\lambda t})^2 \\ &\le \bigl[2 + B_0 + \iota A_4 + (2A_2 + A_5)\theta\bigr](\iota e^{\lambda t}) \\ &\le [3 + B_0](\iota e^{\lambda t})\end{aligned} \tag{5-69}$$

for $t \le T^\iota$, if we assume that θ is small enough that $\theta(2A_2 + A_5) \le \frac{1}{2}$ and $\iota_0 A_4 \le \frac{1}{2}$. For this choice of θ, we then find from the definition of T that $T \ge T^\iota$, a contradiction. Hence, $T^\iota \le T$ for θ sufficiently small.

Step 6 (conclusion: instability). We now define the L^2 part of the norm $\| \cdot \|_0$ by

$$\left\| \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_{00}^2 := \frac{1}{2} \int K\gamma\rho_0^{\gamma-1} \left| \frac{\sigma}{\rho_0} \right|^2 dx + \frac{1}{2} \int \left| \frac{w}{r_0^2} \right|^2 dx. \tag{5-70}$$

Note that $\| \cdot \|_{00} \le \| \cdot \|_0$ and that by the normalization (5-45), we have that the data satisfy

$$\left\| \begin{pmatrix} \sigma_\star \\ \bar{w}_\star \end{pmatrix} \right\|_{00}^2 := C_{00} \in (0, 1). \tag{5-71}$$

Also, we may argue as in the derivation of (5-58) to see that

$$\sqrt{\mathscr{E}^{0,\sigma^\iota}(t) + \mathscr{E}^{0,v^\iota}(t)} \ge \frac{1}{\sqrt{2}} \left\| \begin{pmatrix} \sigma^\iota \\ w^\iota \end{pmatrix} \right\|_{00} \tag{5-72}$$

for $0 \le t \le T^\iota = \min\{T, T^\iota\}$.

Let us now further assume that θ is small enough that $A_2\theta \le \dfrac{C_{00}}{4}$ and $\iota_0 A_3 \le \dfrac{C_{00}}{4}$. We can then combine (5-71), (5-72), (5-61), and (5-62) to deduce that

$$\begin{aligned}
\sqrt{\mathscr{E}^{0,\sigma^\iota}(T^\iota) + \mathscr{E}^{0,v^\iota}(T^\iota)} &\ge \frac{1}{\sqrt{2}} \left\| \begin{pmatrix} \sigma^\iota(T^\iota) \\ w^\iota(T^\iota) \end{pmatrix} \right\|_{00} \ge \frac{1}{\sqrt{2}} \iota e^{\lambda T^\iota} \left\| \begin{pmatrix} \sigma_\star \\ \bar{w}_\star \end{pmatrix} \right\|_{00} \\
&\quad - \frac{1}{\sqrt{2}} \left\| \begin{pmatrix} \sigma^\iota(T^\iota) \\ w^\iota(T^\iota) \end{pmatrix} - \iota e^{\lambda T^\iota} \begin{pmatrix} \sigma_\star \\ \bar{w}_\star \end{pmatrix} - \iota^2 e^{T^\iota \mathscr{L}} \begin{pmatrix} \sigma_0(\iota) \\ w_0(\iota) \end{pmatrix} + e^{T^\iota \mathscr{L}} \begin{pmatrix} 0 \\ \psi^\iota \end{pmatrix} \right\|_{00} \\
&\quad - \frac{1}{\sqrt{2}} \left\| \iota^2 e^{T^\iota \mathscr{L}} \begin{pmatrix} \sigma_0(\iota) \\ w_0(\iota) \end{pmatrix} \right\|_{00} - \frac{1}{\sqrt{2}} \left\| e^{T^\iota \mathscr{L}} \begin{pmatrix} 0 \\ \psi^\iota \end{pmatrix} \right\|_{00} \\
&\ge \frac{1}{\sqrt{2}} \iota e^{\lambda T^\iota} C_{00} - \frac{1}{\sqrt{2}} \left\| \iota^2 e^{T^\iota \mathscr{L}} \begin{pmatrix} \sigma_0(\iota) \\ w_0(\iota) \end{pmatrix} \right\|_0 - \frac{1}{\sqrt{2}} \left\| e^{T^\iota \mathscr{L}} \begin{pmatrix} 0 \\ \psi^\iota \end{pmatrix} \right\|_0 \\
&\quad - \frac{1}{\sqrt{2}} \left\| \begin{pmatrix} \sigma^\iota(T^\iota) \\ w^\iota(T^\iota) \end{pmatrix} - \iota e^{\lambda T^\iota} \begin{pmatrix} \sigma_\star \\ \bar{w}_\star \end{pmatrix} - \iota^2 e^{T^\iota \mathscr{L}} \begin{pmatrix} \sigma_0(\iota) \\ w_0(\iota) \end{pmatrix} + e^{T^\iota \mathscr{L}} \begin{pmatrix} 0 \\ \psi^\iota \end{pmatrix} \right\|_0 \\
&\ge \frac{1}{\sqrt{2}} \bigl(\iota e^{\lambda T^\iota} C_{00} - A_3 \iota^2 e^{\lambda T^\iota} - A_2 (\iota e^{\lambda T^\iota})^2 \bigr) \\
&= \frac{1}{\sqrt{2}} \bigl(\theta C_{00} - \theta \iota A_3 - A_2 \theta^2 \bigr) \ge \frac{C_{00}}{2\sqrt{2}} \theta.
\end{aligned} \tag{5-73}$$

Setting $\theta_0 = \dfrac{\theta C_{00}}{2\sqrt{2}}$, we find that (5-49) holds. This completes the proof of the theorem. ☐

References

[Chandrasekhar 1939] S. Chandrasekhar, *An introduction to the study of stellar structure*, University of Chicago, Chicago, 1939. Reprinted Dover, New York, 1957. MR 19,1142b Zbl 0022.19207

[Deng et al. 2002] Y. Deng, T.-P. Liu, T. Yang, and Z.-a. Yao, “Solutions of Euler–Poisson equations for gaseous stars”, *Arch. Ration. Mech. Anal.* **164**:3 (2002), 261–285. MR 2004h:35187 Zbl 1038.76036

[Ducomet and Zlotnik 2005] B. Ducomet and A. Zlotnik, "Stabilization and stability for the spherically symmetric Navier–Stokes–Poisson system", *Appl. Math. Lett.* **18**:10 (2005), 1190–1198. MR 2006b:35259 Zbl 1125.35071

[Guo and Strauss 1995] Y. Guo and W. A. Strauss, "Instability of periodic BGK equilibria", *Comm. Pure Appl. Math.* **48**:8 (1995), 861–894. MR 96j:35252 Zbl 0840.45012

[Guo and Tice 2010] Y. Guo and I. Tice, "Linear Rayleigh–Taylor instability for viscous, compressible fluids", *SIAM J. Math. Anal.* **42**:4 (2010), 1688–1720. MR 2011j:76084 Zbl 05895002

[Jang 2008] J. Jang, "Nonlinear instability in gravitational Euler–Poisson systems for $\gamma = \frac{6}{5}$", *Arch. Ration. Mech. Anal.* **188**:2 (2008), 265–307. MR 2009a:85003 Zbl 1192.85003

[Jang 2010] J. Jang, "Local well-posedness of dynamics of viscous gaseous stars", *Arch. Ration. Mech. Anal.* **195**:3 (2010), 797–863. MR 2011c:35453 Zbl 1197.35294

[Jang and Masmoudi 2009] J. Jang and N. Masmoudi, "Well-posedness for compressible Euler equations with physical vacuum singularity", *Comm. Pure Appl. Math.* **62**:10 (2009), 1327–1385. MR 2010j:35384 Zbl 1213.35298

[Jang and Masmoudi 2010] J. Jang and N. Masmoudi, "Well-posedness of compressible Euler equations in a physical vacuum", preprint, 2010. arXiv 1005.4441

[Lin 1997] S.-S. Lin, "Stability of gaseous stars in spherically symmetric motions", *SIAM J. Math. Anal.* **28**:3 (1997), 539–569. MR 98b:85001 Zbl 0871.35012

[Matusu-Necasova et al. 1997] S. Matusu-Necasova, M. Okada, and T. Makino, "Free boundary problem for the equation of spherically symmetric motion of viscous gas, III", *Japan J. Indust. Appl. Math.* **14**:2 (1997), 199–213. MR 98e:35177

[Okada and Makino 1993] M. Okada and T. Makino, "Free boundary problem for the equation of spherically symmetric motion of viscous gas", *Japan J. Indust. Appl. Math.* **10**:2 (1993), 219–235. MR 94f:35153 Zbl 0783.76082

[Rein 2003] G. Rein, "Non-linear stability of gaseous stars", *Arch. Ration. Mech. Anal.* **168**:2 (2003), 115–130. MR 2005d:35214 Zbl 1044.76026

[Zhang and Fang 2009] T. Zhang and D. Fang, "Global behavior of spherically symmetric Navier–Stokes–Poisson system with degenerate viscosity coefficients", *Arch. Ration. Mech. Anal.* **191**:2 (2009), 195–243. MR 2010f:35447 Zbl 1161.76051

Received 5 Jul 2012. Revised 4 Jan 2013. Accepted 28 Feb 2013.

JUHI JANG: `juhijang@math.ucr.edu`
Department of Mathematics, University of California, Riverside, Riverside, CA 92521, United States

IAN TICE: `iantice@andrew.cmu.edu`
Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, United States

ANALYSIS AND PDE
Vol. 6, No. 5, 2013
dx.doi.org/10.2140/apde.2013.6.1183

DYNAMICAL IONIZATION BOUNDS FOR ATOMS

ENNO LENZMANN AND MATHIEU LEWIN

We study the long-time behavior of the 3-dimensional repulsive nonlinear Hartree equation with an external attractive Coulomb potential $-Z/|x|$, which is a nonlinear model for the quantum dynamics of an atom. We show that, after a sufficiently long time, the average number of electrons in any finite ball is always smaller than $4Z$ ($2Z$ in the radial case). This is a time-dependent generalization of a celebrated result by E.H. Lieb on the maximum negative ionization of atoms in the stationary case. Our proof involves a novel positive commutator argument (based on the cubic weight $|x|^3$) and our findings are reminiscent of the RAGE theorem.

In addition, we prove a similar universal bound on the local kinetic energy. In particular, our main result means that, in a weak sense, any solution is attracted to a bounded set in the energy space, whatever the size of the initial datum. Moreover, we extend our main result to Hartree–Fock theory and to the linear many-body Schrödinger equation for atoms.

1. Introduction and main result

Rigorous attempts to answer the question *How many electrons can a nucleus bind?* have appeared in the literature over the last decades [Ruskai 1981; 1982; Sigal 1982; 1984; Lieb 1984; Lieb et al. 1988; Solovej 1991; 2003; Nam 2012]. So far, the question has only been addressed in a time-independent setting, that is, the absence of bound states was shown when the number of electrons in the atom is too large. In the present paper we shall rigorously formulate and provide an answer to a similar question in the time-dependent setting: *How many electrons can a nucleus keep in its neighborhood for a long time?*

Our main purpose is therefore the rigorous understanding of the long-time behavior of atoms. We shall prove, for instance, that, in the Hartree approximation, a nucleus of charge Z cannot bind in a time-averaged sense more than $4Z$ electrons ($2Z$ in the radial case). In particular, we will recover some of the known time-independent results (nonexistence of bound states), but by different arguments. One key ingredient in our paper turns out to be a new commutator estimate leading to a novel monotonicity formula, which may be of independent interest for both linear and nonlinear Schrödinger equations.

As a model for the quantum dynamics of an atom, let us first consider the time-dependent nonlinear Hartree equation with an external Coulomb potential:

Lenzmann was partially supported through a Steno fellowship from the Danish Research Council (FNU). Lewin acknowledges financial support from the French Ministry of Research (ANR-10-BLAN-0101) and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 grant agreement MNIQS 258023).
MSC2010: 35Q55, 81Q05, 81Q10, 35Q41.
Keywords: Hartree equation, RAGE theorem, ionization bound, positive commutator.

$$\begin{cases} i\dfrac{\partial}{\partial t}u(t,x) = \left(-\Delta - \dfrac{Z}{|x|} + |u|^2 * \dfrac{1}{|x|}\right)u(t,x), \\ u(0,x) = u_0(x) \in H^1(\mathbb{R}^3). \end{cases} \tag{1-1}$$

Here $u(t,x)$ describes the quantum state of the electrons (which are treated as bosons for simplicity) in an atom [Hartree 1928a; 1928b; Slater 1930]. The terms in the parentheses are, respectively, the kinetic energy operator of the electrons, the electrostatic attractive interaction with the nucleus of charge Z, and the mutual repulsion between the electrons themselves (in units such that $m = 2$ and $\hbar = e = 1$). The total number of electrons in the system is a conserved quantity, which is given by

$$\int_{\mathbb{R}^3} |u(t,x)|^2\,dx = \int_{\mathbb{R}^3} |u_0(x)|^2\,dx =: N.$$

In physical applications, the number N is an integer, but it is convenient to allow any positive real number here. Note that, in Section 4 below, we will also consider the physically more accurate Hartree–Fock model as well as the full many-body Schrödinger equation describing atoms. But for the time being, we deal with the Hartree equation.

The nonlinear equation (1-1) and many variations thereof have been studied extensively in the literature. The existence of a unique strong global-in-time solution to (1-1) with an initial datum $u_0 \in H^1(\mathbb{R}^3)$ goes back to Chadam and Glassey [1975]. Their argument is based on a fixed point argument combined with the conservation of the Hartree energy, defined by

$$\mathscr{E}_Z(u) := \int_{\mathbb{R}^3} |\nabla u(x)|^2\,dx - Z\int_{\mathbb{R}^3} \frac{|u(x)|^2}{|x|}\,dx + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{|u(x)|^2\,|u(y)|^2}{|x-y|}\,dx\,dy. \tag{1-2}$$

In fact, the global well-posedness result for (1-1) can be extended to initial data in $L^2(\mathbb{R}^3)$; see, for instance, [Hayashi and Ozawa 1989; Castella 1997]. However, in what follows, we will always assume that u_0 lies in the energy space $H^1(\mathbb{R}^3)$ so that its corresponding energy is well-defined.

When $Z \leqslant 0$, the solution $u(t)$ to the Hartree equation (1-1) exhibits a purely dispersive behavior, which has been studied by many authors. Here, some works were devoted to the understanding of the dispersive effects for any initial datum [Glassey 1977a; Dias and Figueira 1981; Hayashi and Ozawa 1987; Hayashi 1988; Gasser et al. 1998; Sánchez and Soler 2004], whereas several others dealt with the construction of (modified) scattering [Ginibre and Velo 1980; Ginibre and Ozawa 1993; Hayashi et al. 1998; Hayashi and Naumkin 1998; Ginibre and Velo 2000a; 2000b; López and Soler 2000; Wada 2001; Nakanishi 2002].

In this paper, we are interested in the physically more relevant case when $Z > 0$ holds, which corresponds to having an external attractive long-range potential due to the presence of a positively charged atomic nucleus. The electrons can (and will) now be bound by the nucleus, and the problem of understanding the long-time behavior of solutions is much more delicate. For instance, it was already noticed by Chadam and Glassey [1975, Theorem 4.1] that the solution $u(t)$ cannot tend to zero in $L^\infty(\mathbb{R}^3)$ as $t \to \infty$ for negative energies $\mathscr{E}_Z(u_0) < 0$, which can occur if $Z > 0$ holds.

When $Z > 0$, there exist nonlinear bound states that are solutions of (1-1) taking the simple form

$u(x)e^{-it\lambda}$, where $u \in H^1(\mathbb{R}^3)$ solves the nonlinear eigenvalue equation

$$\left(-\Delta - \frac{Z}{|x|} + |u|^2 * \frac{1}{|x|}\right)u = \lambda u. \tag{1-3}$$

For any fixed $0 < N \leqslant Z$, it is known that (1-3) has infinitely many solutions such that $\int_{\mathbb{R}^3} |u|^2 = N$. Moreover, there is a unique positive solution, which minimizes the Hartree energy (1-2) [Lieb and Simon 1977; Bader 1978] subject to N fixed, and the other (sign-changing) solutions can be constructed by min-max methods [Wolkowisky 1972/73; Stuart 1973; Lions 1981]. The interpretation of the condition $0 < N \leqslant Z$ is that the atom is neutral (if $N = Z$) or positively ionized (if $N < Z$). In this situation, it is not energetically favorable to send a positive fraction of L^2-mass $\mu > 0$, say, to spatial infinity, since the remaining charge is $Z - (N - \mu) > 0$ positive and thus attractive far away from the origin. A more precise mathematical statement is that the Palais–Smale sequences with a bounded Morse index cannot exhibit a lack of compactness when $N \leqslant Z$, and this implies the existence of infinitely many critical points [Berestycki and Lions 1983; Lions 1987; Ghoussoub 1993].

It is known that there are bound states in the case of negative ionization, that is, when $N > Z$ holds. By [Lieb 1981, Theorem 7.19] (see also [Benguria 1979; Benguria et al. 1981]), there is a minimizer of the Hartree functional for N slightly larger than Z. However, it is physically clear that there should not be any bound state when N is too large compared to Z, because a given nucleus is not expected to bind too many electrons compared to its nuclear charge. In [Benguria 1979; Lieb 1981; 1984], it was proved that there exists a universal critical constant $1 < \gamma_c < 2$ such that (1-3) has *no solution* for $N > \gamma_c Z$, but has at least one for $N \leqslant \gamma_c Z$. That γ_c is independent of Z follows from a simple scaling argument.

Let us now collect some basic facts about the set of solutions of the time-independent problem (1-3). For any $u \in H^1(\mathbb{R}^3)$, the self-adjoint operator

$$-\Delta - \frac{Z}{|x|} + |u|^2 * \frac{1}{|x|}$$

has no positive eigenvalue, by the Kato–Agmon–Simon theorem [Reed and Simon 1978, Theorem XIII.58]. This shows that, necessarily, $\lambda \leqslant 0$ in (1-3). Furthermore, we can derive an upper bound on $\|\nabla u\|_{L^2}$ which only depends on Z as follows. If $u \in H^1(\mathbb{R}^3)$ solves (1-3), then, by taking the scalar product with u, we find that

$$\int_{\mathbb{R}^3} |\nabla u(x)|^2 \, dx \leqslant Z \int_{\mathbb{R}^3} \frac{|u(x)|^2}{|x|} \, dx \leqslant Z \|\nabla u\|_{L^2} \|u\|_{L^2}.$$

Here we have used the inequality

$$\int_{\mathbb{R}^3} \frac{|u(x)|^2}{|x|} \, dx \leqslant \min_{z \geqslant 0} \left(\frac{z}{2} \int_{\mathbb{R}^3} |u|^2 + \frac{1}{2z} \int_{\mathbb{R}^3} |\nabla u|^2 \right) = \|u\|_{L^2(\mathbb{R}^3)} \|\nabla u\|_{L^2(\mathbb{R}^3)},$$

which follows from the value of the hydrogen ground state energy, $\inf \operatorname{Spec}(-\Delta/2 - z|x|^{-1}) = -z^2/2$. We conclude that any solution $u \in H^1(\mathbb{R}^3)$ to (1-3) must satisfy the bound

$$\int_{\mathbb{R}^3} |\nabla u|^2 \leqslant \gamma_c Z^3.$$

Recalling that $\int_{\mathbb{R}^3}|u|^2 \leqslant \gamma_c Z$, we conclude that the set of all stationary states

$$\mathcal{A}_Z := \{u \in H^1(\mathbb{R}^3) : u \text{ solves (1-3) for some } \lambda \leqslant 0\} \tag{1-4}$$

is bounded in $H^1(\mathbb{R}^3)$. Elementary arguments show that $\mathcal{A}_Z$ is weakly compact in $H^1(\mathbb{R}^3)$. But we note that the set $\mathcal{A}_Z$ is not compact in the strong H^1-topology.

Supported by physical reasoning and rigorous results in linear scattering theory about asymptotic completeness (see Remark 2 below), it is common belief for infinite-dimensional Hamiltonian systems such as (1-1) that any of its solutions should behave for large times as a superposition of one or several states getting closer to the global attractor $\mathcal{A}_Z$, plus a dispersive part. This is what has already been shown for $Z \leqslant 0$, in which case $\mathcal{A}_Z = \{0\}$. Not much is known in this direction for nonlinear Schrödinger equations [Tao 2007; 2008], and solving this problem (also known as soliton resolution) constitutes a major mathematical challenge. For the Hartree equation (1-1) studied in this paper, the situation is even less clear because of possible modified scattering due to the long-range effects of the Coulomb potential. We can, however, formulate a simpler (but weaker) conjecture as follows.

Conjecture 1 (the global attractor). *Let $u(t)$ be the unique solution to the Hartree equation (1-1) for some $u_0 \in H^1(\mathbb{R}^3)$. Take any sequence of times $t_n \to \infty$ such that $u(t_n) \rightharpoonup u_*$ weakly in $H^1(\mathbb{R}^3)$. Then $u_* \in \mathcal{A}_Z$.*

Remark 2 (the many-body Schrödinger case). Let us recall that the Hartree equation (1-1) is a nonlinear approximation of the linear many-body Schrödinger equation

$$\begin{cases} i\dfrac{\partial}{\partial t}\Psi(t) = \left(\displaystyle\sum_{j=1}^{N}\left(-\Delta_{x_j} - \frac{Z}{|x_j|}\right) + \sum_{1\leqslant k<\ell\leqslant N}\frac{1}{|x_k - x_\ell|}\right)\Psi(t), \\ \Psi(0) = \Psi_0 \in H^1((\mathbb{R}^3)^N). \end{cases} \tag{1-5}$$

Contrary to the Hartree case where we can allow $N = \int_{\mathbb{R}^3}|u|^2$ to take any positive real value, the number N of electrons must of course be an integer for (1-5). The Hartree equation (1-1) is obtained by constraining the solution $\Psi(t)$ to stay on the manifold of product states of the form $\Psi(t, x_1, \dots, x_N) = \psi(t, x_1) \times \cdots \times \psi(t, x_N)$ and using the Dirac–Frenkel principle. Then $u(t) = \sqrt{N}\psi(t)$ solves (1-1). Let us remark that (1-5) can be rewritten after a simple rescaling as

$$\begin{cases} i\dfrac{1}{Z^2}\dfrac{\partial}{\partial t}\Psi(t) = \left(\displaystyle\sum_{j=1}^{N}\left(-\Delta_{x_j} - \frac{1}{|x_j|}\right) + \frac{1}{Z}\sum_{1\leqslant k<\ell\leqslant N}\frac{1}{|x_k - x_\ell|}\right)\Psi(t), \\ \Psi(0) = \Psi_0 \in H^1((\mathbb{R}^3)^N). \end{cases} \tag{1-6}$$

Thus the limit of large $N \to \infty$ with N/Z fixed corresponds to the usual mean-field limit. In this regime, Hartree's theory is known to properly describe (bosonic) atoms, both for ground states [Benguria and Lieb 1983] and in the time-dependent case [Erdős and Yau 2001; Bardos et al. 2000]. See also [Schlein 2008; Fröhlich and Lenzmann 2004] for a review on mean-field limits and the Hartree approximation.

The many-body equation (1-5) looks complicated, but it has the advantage of being linear. In particular, the RAGE theorem tells us that the only possible nonzero weak limits of $\Psi(t)$ when $t \to \infty$ are bound states of the Hamiltonian $H(N)$ in the parentheses [Ruelle 1969; Amrein and Georgescu 1973/74; Enss 1978; Reed and Simon 1979]. This is not a very precise description of the solution for large times because if some particles stay close to the nucleus while other escape to infinity, we will always get $\Psi(t) \rightharpoonup 0$ weakly in $H^1(\mathbb{R}^{3N})$; see [Lewin 2011]. However, asymptotic completeness is known to hold for the linear evolution equation (1-5). This exactly says that any solution $\Psi(t)$ is, in an appropriate sense, a superposition of bound states of the operators $H(k)$ with $1 \leqslant k \leqslant N$ and of scattering states [Dereziński 1993; Sigal and Soffer 1994; Hunziker and Sigal 2000]. Because of the behavior of the underlying many-body system, it is reasonable to believe that the same should be true for the Hartree equation (1-1).

A somewhat weaker property that would follow from Conjecture 1 (at least for (1-7)) is that, for large times, the local mass of any solution has to be smaller than $\gamma_c Z$.

Conjecture 3 (asymptotic number of electrons and kinetic energy). Let $u(t)$ be the unique solution to the Hartree equation (1-1) for some $u_0 \in H^1(\mathbb{R}^3)$. Then

$$\limsup_{t\to\infty} \int_{|x|\leqslant r} |u(t,x)|^2\,dx \leqslant \sup_{u\in\mathcal{A}_Z} \int_{\mathbb{R}^3} |u|^2 = \gamma_c Z \tag{1-7}$$

and

$$\limsup_{t\to\infty} \int_{|x|\leqslant r} |\nabla u(t,x)|^2\,dx \leqslant \sup_{u\in\mathcal{A}_Z} \int_{\mathbb{R}^3} |\nabla u|^2 \leqslant \gamma_c Z^3 \tag{1-8}$$

for all $r > 0$.

The upper bound $\gamma_c Z^3$ is certainly not optimal here. In physical terms, the conjecture says that whatever the number of electrons we start with (and whatever their kinetic energy), we will always end up with at most $\gamma_c Z$ electrons having a universally bounded total kinetic energy. The other electrons have to scatter because the attraction of the nucleus with positive charge Z is not strong enough to keep all the electrons in its neighborhood. It could be that proving the weaker Conjecture 3 is not much easier than proving the stronger Conjecture 1. We actually have very little information on γ_c.

In this paper, we are interested in Conjecture 3. We will prove a time-averaged version of (1-7), with γ_c replaced by 2 in the radial case, and by 4 in the general case. Our main result is as follows.

Theorem 4 (long-time behavior of atoms in Hartree theory). *Suppose $Z > 0$, let u_0 be an arbitrary initial datum in $H^1(\mathbb{R}^3)$, and denote by $u(t)$ the unique solution of* (1-1). *Then, for any $R > 0$, we have the estimate*

$$\frac{1}{T}\int_0^T dt \int_{\mathbb{R}^3} dx \frac{|u(t,x)|^2}{1+|x|^2/R^2} \leqslant 4Z + \frac{3}{R} + \frac{2\sqrt{KN}R^2}{ZT} \tag{1-9}$$

with

$$N := \int_{\mathbb{R}^3} |u_0|^2$$

and

$$K := \sup_{t\geqslant 0} \int_{\mathbb{R}^3} |\nabla u(t)|^2 \leqslant Z^2 N + 2\|\nabla u_0\|^2_{L^2(\mathbb{R}^3)} + N^3 \|\nabla u_0\|_{L^2(\mathbb{R}^3)}. \tag{1-10}$$

In particular, we have

$$\limsup_{T\to\infty}\frac{1}{T}\int_0^T dt\int_{|x|\leqslant r} dx|u(t,x)|^2\leqslant 4Z \tag{1-11}$$

for every $r>0$. *Similarly, we have the following estimate on the local kinetic energy*:

$$\frac{1}{T}\int_0^T dt\int_{\mathbb{R}^3} dx\frac{|\nabla u(t,x)|^2}{(1+|x|/R)^2}\leqslant\left(\frac{Z^2}{4}+\frac{2Z}{R}+\frac{3Z}{R^2}\right)\frac{1}{T}\int_0^T dt\int_{\mathbb{R}^3} dx\frac{|u(t,x)|^2}{1+|x|^2/R^2}+\frac{2R\sqrt{K}\sqrt{N}}{T}. \tag{1-12}$$

Therefore

$$\limsup_{T\to\infty}\frac{1}{T}\int_0^T dt\int_{|x|\leqslant r} dx|\nabla u(t,x)|^2\leqslant Z^3 \tag{1-13}$$

for every $r>0$.

If the initial datum $u_0=u_0(|x|)$ *is radial,* $u(t)$ *is radial for all times and the same estimate* (1-9) *holds true with* $4Z$ *replaced by* $2Z$. *Similarly, the estimate* (1-13) *holds true with* Z^3 *replaced by* $Z^3/2$.

Note that we do not exactly get that the limiting mass is $\leqslant 4Z$ for large times, but we only know it in the sense of *time averages* of the form $\langle f\rangle_T=T^{-1}\int_0^T f\,dt$. Such a statement is reminiscent of the celebrated RAGE theorem [Ruelle 1969; Amrein and Georgescu 1973/74; Enss 1978; Reed and Simon 1979] for linear time evolutions generated by self-adjoint operators. The constants in the error terms of (1-9) and (1-12) are probably not optimal at all, but they are displayed here to emphasize that our method can provide simple and explicit bounds. However, we have not tried to optimize these constants too much.

In the radial case, we are able to get the same numerical value of 2 as the best known estimate on γ_c. However, we use a virial-type argument that seems to be quite different from Lieb's celebrated proof [1984] in the stationary case (which, for radial solutions, goes back to [Benguria 1979]). In particular, our approach provides an alternative proof of the fact that $\gamma_c<2$ in the stationary radial case.

Strategy of the proof. Now we explain the main ideas used in the proof of Theorem 4. To this end, we start by quickly recalling Lieb's proof [1984] that $\gamma_c<2$ holds. His idea is to take the scalar product of the stationary Hartree equation (1-3) with $|x|u(x)$, leading to the estimate

$$\left\langle u,\frac{|x|(-\Delta)+(-\Delta)|x|}{2}u\right\rangle-ZN+\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{(|x|+|y|)|u(x)|^2|u(y)|^2}{2|x-y|}\,dx\,dy\leqslant 0,$$

using that $\lambda\leqslant 0$ holds. To conclude, it suffices to notice that we have

$$\frac{|x|(-\Delta)+(-\Delta)|x|}{2}=|x|^{1/2}\left(-\Delta-\frac{1}{4|x|^2}\right)|x|^{1/2}>0,$$

by Hardy's inequality, and that

$$\frac{|x|+|y|}{|x-y|}\geqslant 1,$$

by the triangle inequality. Combining these estimates, we obtain that $-ZN + N^2/2 < 0$, which implies the bound $N < 2Z$ for the stationary problem (1-3). (Note that the inequality is strict, since there is no optimizer in Hardy's inequality.)

In view of Lieb's argument for the stationary problem (1-3), it appears to be a viable strategy in the time-dependent setting to consider the quantity $M(t) = \int |x|\,|u(t,x)|^2\,dx$ (or some spatially localized version thereof). Indeed, if we take the second time derivative of $M(t)$, we are (formally) led to the well-known *Morawetz–Lin–Strauss estimate* for nonlinear Schrödinger (NLS) equations, which has proved to be of enormous value in the setting of NLS equations with purely repulsive interactions. However, due to presence of the attractive term $-Z/|x|$ with $Z > 0$ in the Hartree equation (1-1), the use of the classical Morawetz–Lin–Strauss bounds does not yield any dispersive information about $u(t,x)$, even in the case when N is large compared to Z.

In our situation, it turns out that it is more natural to study the time evolution of the third moment $M(t) = \int |x|^3 |u(t,x)|^2\,dx$. If we compute its second time derivative, we obtain

$$\begin{aligned}\frac{1}{3}\frac{d^2}{dt^2}\int_{\mathbb{R}^3} |x|^3 |u(t,x)|^2\,dx &= \frac{d}{dt}\langle u(t), Au(t)\rangle = 2\Re\Big\langle \frac{\partial}{\partial t}u(t), Au(t)\Big\rangle \\ &= \langle u(t), i[-\Delta, A]u(t)\rangle + \langle u(t), i[V_u, A]u(t)\rangle\end{aligned}$$

with $A := -i(\nabla\cdot x|x| + |x|\,x\cdot\nabla)$ and $V_u = -Z|x|^{-1} + |u|^2 * |x|^{-1}$. This is the same as multiplying the time-dependent equation (1-1) by $\overline{Au(t)}$ and taking the imaginary part. Our key observation is the positivity of the commutator

$$i[-\Delta, A] = -\tfrac{1}{3}[\Delta, [\Delta, |x|^3]] \geqslant 0 \tag{1-14}$$

(see also (2-6) below), combined with the fact that

$$\begin{aligned}\langle u(t), i[V_u, A]u(t)\rangle &= -2\int_{\mathbb{R}^3} |x|x\cdot\nabla V_u(x)|u(t,x)|^2 \\ &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} (|x|x - |y|y)\cdot\frac{x-y}{|x-y|^3}|u(t,x)|^2|u(t,y)|^2\,dx\,dy - 2ZN \\ &\geqslant \kappa N^2 - 2ZN,\end{aligned}$$

where $\kappa = 1$ if $u(t)$ is radial and $\kappa = \frac{1}{2}$ otherwise (see Lemma 9 below). Hence, when $N > 2Z/\kappa$, we deduce the lower bound

$$\frac{1}{3}\frac{d^2}{dt^2}\int_{\mathbb{R}^3} |x|^3|u(t,x)|^2\,dx = \frac{d}{dt}\langle u(t), Au(t)\rangle \geqslant N(\kappa N - 2Z) > 0. \tag{1-15}$$

Therefore the quantity $\int_{\mathbb{R}^3} |x|^3|u(t,x)|^2\,dx$ grows at least like t^2 for large t and in particular $\langle u(t), Au(t)\rangle$ is a monotone increasing quantity. This growth is a strong indication that some dispersion takes place and some particles have to escape to infinity. (A regularized version of the previous estimate will indeed show this claim for any H^1-solution.) Note also that, in the time-independent case when u is a nonlinear bound state (and hence the left side in (1-15) must be zero), this is also a new proof of Lieb's inequality $\gamma_c < 2$ in the radial setting, since $\kappa = 1$ holds under this symmetry assumption.

Let us generally remark that virial or positive commutator arguments are very common in the literature [Killip and Visan 2008; Colliander et al. 2003]. When $|x|^3$ is replaced by $|x|$, this leads to the famous Morawetz inequalities [1968], as already mentioned, whereas the case of $|x|^2$ gives the virial identity used by Glassey [1977b] to prove finite-time blowup for NLS equations. Tao [2008] advocated the use of $|x|^4$ for some nonlinear Schrödinger equations in dimension $d \geqslant 7$ in order to get a universal bound on the mass of the solution. We are not aware of any use of the multiplier $|x|^3$ in the literature.

In fact, using the cubic weight $|x|^3$ is rather natural from a dimensional point of view in our situation: if the potential term $[V_u, A]$ should be $O(1)$, the virial function must behave like the third power of a length to compensate the Laplacian and the Coulomb potential.

For the proof of our main result, we will in fact derive a whole class of double commutator estimates of the same kind as (1-14), which we think is of independent interest too. In particular, we will show in (2-6) below that, in any dimension $d \geqslant 1$, we have the commutator bound

$$-\big[\Delta, [\Delta, |x|^{\beta}]\big] \geqslant \beta(\beta+d-4)(d-\beta)|x|^{\beta-4}, \tag{1-16}$$

provided that $\beta \geqslant \max(1, 4-d)$. Note that the right side is $\geqslant 0$ when $\beta \leqslant d$. In spite of the fact that (1-16) turns out to be equivalent to a general version of Hardy's inequality, we have not found it explicitly written (let alone systematically treated) in the literature. Notice that the bound (1-16) contains the usual inequalities for $\beta = 1, 2$, as well as Tao's estimate for $\beta = 4$. In the present application, we shall use (1-16) in dimension $d = 3$ with $\beta = 3$, or rather a regularized version thereof. However, the positivity of this commutator does not directly follow as in the "classical" cases when $\beta = 1, 2$. To wit, for $d = \beta = 3$, a calculation (which will be detailed below) yields the identity

$$-\big[\Delta, [\Delta, |x|^3]\big] = -\Delta\Delta|x|^3 - \nabla\cdot(\mathrm{Hess}_{|x|^3})\nabla = -\frac{24}{|x|} - 12\nabla\cdot[|x|(1+\omega_x\omega_x^T)]\nabla,$$

where $\omega_x = x/|x|$ denotes the unit vector in direction $x \in \mathbb{R}^3$. Obviously, the first term on the right side is negative definite. Nevertheless, when combined with the second term, the generalized Hardy's inequality (see (2-7) below) shows that the whole right-hand side is indeed nonnegative, and hence the estimate (1-16) follows in the particular case $d = \beta = 3$.

Ultimately, we are interested in general H^1-solutions $u(t)$ without imposing any spatial weight condition. Therefore, the strategy of proving Theorem 4 explained above needs to be further refined. In particular, the desired bound (1-9) on a ball of radius R cannot be obtained by only looking at the second derivative of the third moment as we have just explained. Our method to extend (1-9) to any H^1-valued solution $u(t)$ is to replace the function $|x|^3$ by a radial function $f_R(|x|)$ which behaves like $|x|^3$ on the ball of radius R and like $|x|$ at infinity. This will imply that $A_{f_R} = -i[\Delta, f_R]$ defines a bounded operator from $H^1(\mathbb{R}^3)$ to $L^2(\mathbb{R}^3)$. Furthermore, we will need to derive a sufficiently good lower bound on the double commutator $-[\Delta, [\Delta, f_R]]$ in order to imitate the previous argument on the ball only. In Section 2, we explain how to do this for a general function f. Finally, the bound (1-12) on the local kinetic energy is itself obtained by considering another virial function g_R which behaves like $|x|^2$ on the ball of radius R and like $|x|$ at infinity. The complete proof of Theorem 4 is given in Section 3.

Extensions: Hartree–Fock and many-body Schrödinger theory. In physical reality electrons are fermions, which means that the many-body wave function $\Psi = \Psi(t, x_1, \dots, x_N)$ in (1-5) must be *antisymmetric* with respect to exchanges of its spatial variables $x_1, \dots, x_N$. The Hartree state $\psi(t, x_1) \cdots \psi(t, x_N)$ is symmetric, and it is therefore not allowed for physical electrons. This is why one speaks about *bosonic* atoms. The simplest product-like antisymmetric wave function is a *Hartree–Fock state*, sometimes also called a Slater determinant:

$$\Psi(t, x_1, \dots, x_N) = \frac{1}{\sqrt{N!}} \sum_{\sigma \in \mathcal{S}_N} \varepsilon(\sigma) u_1(t, x_{\sigma(1)}) \cdots u_N(t, x_{\sigma(N)}), \quad \langle u_j, u_k \rangle_{L^2} = \delta_{jk}.$$

In Section 4.1 below, we extend Theorem 4 to the corresponding time-dependent Hartree–Fock equations; see Theorem 13 for a precise statement. Finally, we also consider the full many-body Schrödinger equation (1-5) in Section 4.2 below, where our findings are summarized in Theorem 15.

2. Estimating the commutator $-[\Delta, [\Delta, f(x)]]$

Throughout this section, we use the convenient notation

$$p := -i\nabla,$$

and, in particular, we have $p^2 = -\Delta$ in what follows. In this section, we investigate how to get lower bounds for a double commutator of the form $-[p^2, [p^2, f(x)]]$ in general space dimensions $d \geqslant 1$. Such a double commutator always arises when computing the second derivative of the expectation value of $f(x)$ in a nonrelativistic system based on the Laplacian. We always assume that f is smooth enough (possibly only outside of the origin) such that the double commutator can be at least properly interpreted as a quadratic form on $C_c^\infty(\mathbb{R}^d)$ or on $C_c^\infty(\mathbb{R}^d \setminus \{0\})$.

Our starting point is the well-known formula for the double commutator, which follows from a tedious but simple calculation:

$$-[p^2, [p^2, f(x)]] = -(\Delta\Delta f)(x) + 4p \cdot (\operatorname{Hess} f(x))p. \tag{2-1}$$

Since the Hessian of f appears on the right side, it is natural to restrict to convex functions f. Then the second term is nonnegative in the sense of operators. One can use this term to control the bi-Laplacian of f by resorting to Hardy's trick, which is based on writing

$$\begin{aligned} p \cdot (\operatorname{Hess} f(x))p &= (p + iF(x)) \cdot (\operatorname{Hess} f(x))(p - iF(x)) \\ &\quad + i(p \cdot (\operatorname{Hess} f(x))F(x) - F(x) \cdot (\operatorname{Hess} f(x))p) - F(x) \cdot (\operatorname{Hess} f(x))F(x) \\ &\geqslant \operatorname{div}(\operatorname{Hess} f(x)F(x)) - F(x) \cdot (\operatorname{Hess} f(x))F(x) \end{aligned} \tag{2-2}$$

for any sufficiently smooth real vector field $F : \mathbb{R}^3 \to \mathbb{R}^3$. Here we have only used that $(p + iF(x)) \cdot (\operatorname{Hess} f(x))(p - iF(x)) \geqslant 0$ holds, which simply follows from the assumed convexity $\operatorname{Hess} f(x) \geqslant 0$ and the self-adjointness $(p + iF(x))^* = p - iF(x)$. For dimensional reasons, it is natural to take F of the form $F(x) = \alpha x|x|^{-2}$ with some constant $\alpha \in \mathbb{R}$. We thus obtain the lower bound

$$\begin{aligned}
-[p^2, [p^2, f(x)]] &= 4\left(p + i\alpha\frac{x}{|x|^2}\right) \cdot (\operatorname{Hess} f(x))\left(p - i\alpha\frac{x}{|x|^2}\right) \\
&\qquad + 4\alpha \operatorname{div}\left(\operatorname{Hess} f(x)\frac{x}{|x|^2}\right) - 4\alpha^2\frac{x^T(\operatorname{Hess} f(x))x}{|x|^4} - (\Delta\Delta f)(x) \\
&\geqslant 4\alpha \operatorname{div}\left(\operatorname{Hess} f(x)\frac{x}{|x|^2}\right) - 4\alpha^2\frac{x^T(\operatorname{Hess} f(x))x}{|x|^4} - (\Delta\Delta f)(x)
\end{aligned} \tag{2-3}$$

for a sufficiently smooth convex function f and any $\alpha \in \mathbb{R}$. By using Hardy's trick we are able to obtain a lower bound which does not contain the differential operator p. Our estimate only involves a multiplication operator. By varying α, we can try to make the negative part of this function as small as possible.

Let us now restrict ourselves to a radial function $f(|x|)$ and use the notation $r = |x|$ and $\omega_x := x/|x|$ for simplicity. Some tedious calculations show that

$$\begin{aligned}
\operatorname{Hess} f(|x|) &= (1 - \omega_x\omega_x^T)\frac{f'(r)}{r} + \omega_x\omega_x^T f''(r), \\
\operatorname{div}\left((\operatorname{Hess} f(x))\frac{x}{|x|^2}\right) &= \operatorname{div}\left(\frac{f''(r)}{r}\omega_x\right) = \frac{f^{(3)}(r)}{r} + (d-2)\frac{f''(r)}{r^2}, \\
\frac{x^T(\operatorname{Hess} f(x))x}{|x|^4} &= \frac{f''(r)}{r^2}.
\end{aligned}$$

Moreover, we recall the formula for the Bi-Laplacian of a radial function:

$$\Delta\Delta f(|x|) = f^{(4)}(r) + 2(d-1)\frac{f^{(3)}(r)}{r} + (d-1)(d-3)\frac{f^{(2)}(r)}{r^2} - (d-1)(d-3)\frac{f'(r)}{r^3}.$$

Therefore we can rewrite the equality in (2-3) for a radial function f as

$$\begin{aligned}
&-\big[p^2, [p^2, f(|x|)]\big] \\
&\quad = 4\left(p + i\alpha\frac{\omega_x}{r}\right) \cdot \left((1 - \omega_x\omega_x^T)\frac{f'(r)}{r} + \omega_x\omega_x^T f''(r)\right)\left(p - i\alpha\frac{\omega_x}{r}\right) - f^{(4)}(r) \\
&\qquad + 4\left(\alpha - \frac{d-1}{2}\right)\frac{f^{(3)}(r)}{r} + 4\left(\alpha(d-2) - \alpha^2 - \frac{(d-1)(d-3)}{4}\right)\frac{f''(r)}{r^2} + (d-1)(d-3)\frac{f'(r)}{r^3}.
\end{aligned} \tag{2-4}$$

The operator on the first line is $\geqslant 0$ when $x \mapsto f(|x|)$ is convex. In dimension $d = 3$, we already get a simple estimate.

Lemma 5 (a lower bound for $d = 3$). *Let* $f : [0, \infty) \to \mathbb{R}$ *be a convex nondecreasing function such that* $x \mapsto f^{(4)}(|x|) \in L^1_{\mathrm{loc}}(\mathbb{R}^3)$. *Then we have*

$$\begin{aligned}
-[p^2,[p^2,f(|x|)]] &= 4\left(p + i\frac{\omega_x}{r}\right) \cdot \left((1 - \omega_x\omega_x^T)\frac{f'(r)}{r} + \omega_x\omega_x^T f''(r)\right)\left(p - i\frac{\omega_x}{r}\right) - f^{(4)}(|x|) \\
&\geqslant -f^{(4)}(|x|)
\end{aligned} \tag{2-5}$$

in the sense of quadratic forms on $C_c^\infty(\mathbb{R}^3)$.

Proof. Take $\alpha = 1$ in (2-4). ☐

Coming back to (2-4) and taking now the convex function $f(|x|) = |x|^\beta$ with $\beta \geqslant 1$, we obtain the following general result.

Lemma 6 (estimate on $-[p^2, [p^2, |x|^\beta]]$). *For all* $\beta \geqslant \max(1, 4-d)$, *we have*

$$-[p^2, [p^2, |x|^\beta]] \geqslant \beta(\beta+d-4)(d-\beta)|x|^{\beta-4} \tag{2-6}$$

in the sense of quadratic forms on $C_c^\infty(\mathbb{R}^d)$ (*or on* $C_c^\infty(\mathbb{R}^d \setminus \{0\})$ *if* $\beta = 4-d$). *The right side of* (2-6) *is nonnegative for* $\max(1, 4-d) \leqslant \beta \leqslant d$.

Proof. Take $f(r) = r^\beta$ in (2-4) and optimize with respect to α (the optimum is $\alpha = (\beta+d-4)/2$). We need $\beta \geqslant 1$ to make sure that f is nondecreasing and convex, and $\beta > 4-d$ to ensure that all the terms are in $L^1_{\mathrm{loc}}(\mathbb{R}^d)$. For $\beta = 4-d \geqslant 1$, the right side of (2-4) vanishes and the bound stays correct by a simple limit argument. We remark that, in the borderline case $\beta = 4-d$, there is a positive δ-measure occurring at the origin $x = 0$, which we do not see when using functions of $C_c^\infty(\mathbb{R}^d \setminus \{0\})$. ☐

Remark 7. From (2-1) we immediately get the special formula $-[p^2, [p^2, |x|^2]] = 8p^2 \geqslant 0$, valid in any dimension $d \geqslant 1$. For $d \geqslant 3$ and $\beta = 2$, the lower bound given in Lemma 6 is then a direct consequence of Hardy's inequality $4p^2 \geqslant (d-2)^2|x|^{-2}$. In fact, we shall see below that the bound in Lemma 6 is equivalent to a generalized version of Hardy's inequality.

We conclude this section with some general observations. First, we note that Lemma 6 gives a nonnegative lower bound in (2-6) in dimension $d = 2$ for the choice $\beta = 2$ only. In higher dimensions $d \geqslant 3$, the right side is nonnegative for any $1 \leqslant \beta \leqslant d$. When $\beta = 4$, we get the simple lower bound

$$-[p^2, [p^2, |x|^4]] \geqslant 4d(d-4) \quad \text{for } d \geqslant 4,$$

which was used for the first time by Tao [2008].

As we have seen, the bound (2-6) is equivalent to the operator inequality

$$\left(p + i\alpha\frac{\omega_x}{r}\right) \cdot \left((1-\omega_x\omega_x^T)\frac{f'(r)}{r} + \omega_x\omega_x^T f''(r)\right)\left(p - i\alpha\frac{\omega_x}{r}\right) \geqslant 0$$

with $f(r) = r^\beta$. This can also be written for the optimal $\alpha = (\beta+d-4)/2$ as

$$\int_{\mathbb{R}^d} |x|^{\beta-2}\left(|P_x^\perp \nabla u(x)|^2 + (\beta-1)\left|\omega_x \cdot \nabla u(x) + \frac{\beta+d-4}{2|x|}u(x)\right|^2\right) dx \geqslant 0,$$

where $P_x^\perp = 1 - \omega_x\omega_x^T$ is the projection on the two-dimensional space orthogonal to ω_x. Saying that the second term is nonnegative is equivalent, for $\beta > 1$, to the (generalized) Hardy inequality

$$\int_{\mathbb{R}^d} |x|^{\beta-2}|\omega_x \cdot \nabla u(x)|^2\, dx \geqslant \frac{(\beta+d-4)^2}{4}\int_{\mathbb{R}^d} |x|^{\beta-4}|u(x)|^2\, dx. \tag{2-7}$$

Hence we see that (2-6) is nothing else but a reformulation of Hardy's inequality (2-7).

Remark 8 (fractional Laplacians). Using the integral representation

$$x^\theta = \frac{\sin(\pi\theta)}{\pi}\int_0^\infty \frac{x}{x+s}s^{\theta-1}\,ds \quad \text{for } 0<\theta<1,$$

we can easily transpose most of our estimates to fractional powers $|p|^{2\theta}=(-\Delta)^\theta$ and $\langle p\rangle^{2\theta}=(|p|^2+1)^\theta$ with $\theta\in(0,1)$. For instance, for the pseudorelativistic kinetic energy operator $\sqrt{p^2+1}$, we have, at least formally,

$$\begin{aligned}&-\big[\sqrt{1+p^2},\big[\sqrt{1+p^2},f(x)\big]\big]\\&\quad=\frac{1}{\pi^2}\int_0^\infty\sqrt{s}\,ds\int_0^\infty\sqrt{t}\,dt\,\frac{1}{(1+p^2+s)(1+p^2+t)}(-[p^2,[p^2,f(x)]])\frac{1}{(1+p^2+s)(1+p^2+t)}.\end{aligned}$$

In particular, we find

$$-\big[\sqrt{1+p^2},\big[\sqrt{1+p^2},|x|^\beta\big]\big]\geqslant 0$$

for $\max(1,4-d)\leqslant\beta\leqslant d$. For a general convex radial function f and in $d=3$ dimensions, we obtain the estimate

$$-\big[\sqrt{1+p^2},\big[\sqrt{1+p^2},f(|x|)\big]\big]\geqslant-\tfrac{1}{4}\|f_+^{(4)}\|_{L^\infty(\mathbb{R}^3)}$$

with $f_+^{(4)}$ denoting the positive part of $f^{(4)}$.

3. Proof of Theorem 4

In this section, we provide the proof of our main result given by Theorem 4. We always assume that the initial datum u_0 is smooth and decays fast enough, such that our calculations are justified. As we will see below, our estimates only involve the $H^1(\mathbb{R}^3)$ norm of u_0, and thus the general case can be obtained by a simple limiting argument, which we do not detail here.

Proof of Theorem 4. Step 1: *the virial identity.* Consider a smooth radial convex function f. We define the corresponding virial operator

$$A_f:=p\cdot\nabla f+\nabla f\cdot p=p\cdot\omega_x f'(|x|)+f'(|x|)\omega_x\cdot p. \tag{3-1}$$

Using (2-5), we get

$$\begin{aligned}&\frac{d}{dt}\langle u(t),A_fu(t)\rangle\\&\quad=4\int_{\mathbb{R}^3}\frac{f'(|x|)}{|x|}|P_x^\perp\nabla u(t,x)|^2\,dx+4\int_{\mathbb{R}^3}f''(|x|)\left|\omega_x\cdot\nabla u(t,x)+\frac{u(t,x)}{|x|}\right|^2dx\\&\qquad-\int_{\mathbb{R}^3}f^{(4)}(|x|)|u(t,x)|^2\,dx-2\int_{\mathbb{R}^3}f'(|x|)|u(t,x)|^2\omega_x\cdot\nabla V_u(t,x)\,dx,\end{aligned}\tag{3-2}$$

with

$$V_u(t,x)=-\frac{Z}{|x|}+|u(t)|^2*|x|^{-1}:=-\frac{Z}{|x|}+W_u(t,x).$$

The first potential term is just

$$-2\int_{\mathbb{R}^3} f'(|x|)|u(t,x)|^2\omega_x\cdot\nabla\Bigl(-\frac{Z}{|x|}\Bigr)\,dx = -2Z\int_{\mathbb{R}^3}\frac{f'(|x|)}{|x|^2}|u(t,x)|^2.$$

The second potential term can be expressed as

$$\begin{aligned}-2\int_{\mathbb{R}^3} & f'(|x|)|u(t,x)|^2\omega_x\cdot\nabla W_u(t,x)\,dx\\ &= 2\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} f'(|x|)\omega_x\cdot\frac{x-y}{|x-y|^3}|u(t,x)|^2|u(t,y)|^2\,dx\,dy\\ &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{(f'(|x|)\omega_x - f'(|y|)\omega_y)\cdot(x-y)}{|x-y|^3}|u(t,x)|^2|u(t,y)|^2\,dx\,dy,\end{aligned}$$

where in the last line we have just exchanged the role of x and y. Inserting in (3-2), we arrive at the expression

$$\begin{aligned}\frac{d}{dt}\langle u(t), A_f u(t)\rangle = 4\int_{\mathbb{R}^3}\frac{f'(|x|)}{|x|}|P_x^{\perp}\nabla u(t,x)|^2\,dx + 4\int_{\mathbb{R}^3} f''(|x|)\left|\omega_x\cdot\nabla u(t,x)+\frac{u(t,x)}{|x|}\right|^2 dx\\ -\int_{\mathbb{R}^3} f^{(4)}(|x|)|u(t,x)|^2\,dx - 2Z\int_{\mathbb{R}^3}\frac{f'(|x|)}{|x|^2}|u(t,x)|^2\,dx\\ +\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{(f'(|x|)\omega_x - f'(|y|)\omega_y)\cdot(x-y)}{|x-y|^3}|u(t,x)|^2|u(t,y)|^2\,dx\,dy.\end{aligned}\tag{3-3}$$

For dimensional reasons, it is natural to take $f(|x|)=|x|^3/3$. The following lemma allows us to deal with the last potential term in this special case.

Lemma 9 (lower bound on the nonlinear term for $f(r)=r^3/3$). *We have*

$$\frac{(|x|^2\omega_x - |y|^2\omega_y)\cdot(x-y)}{|x-y|^3}\geqslant\frac{1}{2}\tag{3-4}$$

for all $x\neq y\in\mathbb{R}^3$. In the radial case we have

$$⨍_{S^2}⨍_{S^2}\frac{(|x|^2\omega_x - |y|^2\omega_y)\cdot(x-y)}{|x-y|^3}\,d\omega_x\,d\omega_y = 1\tag{3-5}$$

where $⨍_{S^2} d\omega_x = (4\pi)^{-1}\int_{S^2} d\omega_x$ denotes the (normalized) angular integration.

Proof. We compute

$$\frac{(|x|^2\omega_x - |y|^2\omega_y)\cdot(x-y)}{|x-y|^3} = \frac{r^3+s^3-(r^2s+s^2r)\omega_x\cdot\omega_y}{(r^2+s^2-2rs\omega_x\cdot\omega_y)^{3/2}} = \frac{1+u^3-(u+u^2)\theta}{(1+u^2-2u\theta)^{3/2}}$$

with $x=r\omega_x$, $y=s\omega_y$, $u:=\min(r,s)\max(r,s)^{-1}\in[0,1]$, and $\theta:=\omega_x\cdot\omega_y\in[-1,1]$. Differentiating with respect to θ, we find

$$\frac{d}{d\theta}\left(\frac{1+u^3-(u+u^2)\theta}{(1+u^2-2u\theta)^{3/2}}\right) = \frac{u(1+u)(u^2+(1-u)(2-u))-\theta u^2(1+u)}{(1+u^2-2u\theta)^{5/2}}.$$

We have $u^2+(1-u)(2-u)=u+2(u-1)^2\geqslant u$ and therefore the numerator is nonnegative for $u>0$ and $\theta\in[-1,1]$. We conclude that the minimum is attained for $\theta=-1$. The value is

$$\frac{1+u^3+u+u^2}{(1+u)^3}=1-\frac{2u}{(1+u)^2}\geqslant\tfrac{1}{2},$$

where the minimum is attained for $u=1$. All in all, we find that

$$\frac{(|x|^2\omega_x-|y|^2\omega_y)\cdot(x-y)}{|x-y|^3}\geqslant\tfrac{1}{2},$$

as was claimed. In the radial case we find by explicit integration

$$⨍_{S^2}⨍_{S^2}\frac{(|x|^2\omega_x-|y|^2\omega_y)\cdot(x-y)}{|x-y|^3}\,d\omega_x\,d\omega_y=\frac{1}{2}\int_{-1}^{1}\frac{1+u^3-(u+u^2)\theta}{(1+u^2-2u\theta)^{3/2}}\,d\theta=1.\qquad\square$$

For $f(r)=r^3/3$, the previous estimates give

$$\frac{d}{dt}\langle u(t),A_f u(t)\rangle\\ =4\int_{\mathbb{R}^3}|x|\,|P_x^{\perp}\nabla u(t,x)|^2\,dx+8\int_{\mathbb{R}^3}|x|\left|\omega_x\cdot\nabla u(t,x)+\frac{u(t,x)}{|x|}\right|^2dx+\kappa N^2-2ZN,\tag{3-6}$$

where $\kappa=1$ in the radial case and $\kappa=\frac{1}{2}$ otherwise. If u is a stationary state, the left side is independent of t and this is a new proof that $N<4Z$ ($N<2Z$ in the radial case) for bound states. Equation (3-6) is a new *monotonicity formula* for the Coulombic Hartree equation, when $N\geqslant 4Z$ ($N\geqslant 2Z$ in the radial case).

Step 2: *the localized virial estimate.* We now use a localized virial estimate, which means that we choose a virial function f_R which behaves like $|x|^3/3$ on a ball of radius R and like $|x|$ at infinity. We will take f_R of the form

$$f_R(|x|)=R^3 f(|x|/R)$$

for

$$f(r)=r-\arctan r,\tag{3-7}$$

which we have chosen to have

$$f'(r)=\frac{r^2}{1+r^2}=1-\frac{1}{1+r^2}.\tag{3-8}$$

Clearly, the first derivative f' is nondecreasing and positive. Hence $x\mapsto f(|x|)$ is a convex function on $\mathbb{R}^3$. The following lemma gathers some important properties of f, which are the 'localized' equivalent of Lemma 9 above.

Lemma 10 (the virial function f). *Let f be as in* (3-7). *We have*

$$\frac{(f'(|x|)\omega_x-f'(|y|)\omega_y)\cdot(x-y)}{|x-y|^3}\geqslant\frac{1}{2}\frac{f'(|x|)}{|x|^2}\frac{f'(|y|)}{|y|^2}\tag{3-9}$$

for all $x \neq y \in \mathbb{R}^3$. In the radial case, we get

$$\begin{aligned}\fint_{S^2}\fint_{S^2} \frac{(f'(|x|)\omega_x - f'(|y|)\omega_y)\cdot(x-y)}{|x-y|^3}\,d\omega_x\,d\omega_y \\ = \frac{f'(\max(|x|,|y|))}{\max(|x|,|y|)^2} = \frac{1}{1+\max(|x|^2,|y|^2)} \geqslant \frac{f'(|x|)}{|x|^2}\frac{f'(|y|)}{|y|^2}.\end{aligned} \tag{3-10}$$

Proof. As in Lemma 9, we write

$$\frac{(f'(|x|)\omega_x - f'(|y|)\omega_y)\cdot(x-y)}{|x-y|^3} = \frac{rf'(r)+sf'(s)-\theta(sf'(r)+rf'(s))}{(r^2+s^2-2rs\theta)^{3/2}} \tag{3-11}$$

with $r = |x|$, $s = |y|$, and $\theta = \omega_x \cdot \omega_y \in [-1, 1]$. Differentiating with respect to θ, we find

$$\frac{r(2r^2-s^2)f'(r)+s(2s^2-r^2)f'(s)-\theta rs(sf'(r)+rf'(s))}{(r^2+s^2-2rs\theta)^{5/2}}.$$

Since $f' > 0$, the numerator is positive for $\theta \leqslant \theta_c$ and negative for $\theta \geqslant \theta_c$. Regardless of whether $\theta_c \in [-1, 1]$ or not, the minimum of the function in (3-11) is attained at $\theta = \pm 1$. For $\theta = -1$, we find

$$\frac{f'(r)+f'(s)}{(r+s)^2} = \frac{r^2+s^2+2r^2s^2}{(1+r^2)(1+s^2)(r+s)^2}.$$

Now we remark that

$$\frac{r^2+s^2+2r^2s^2}{(r+s)^2} = \frac{1}{2} + \frac{(r-s)^2+4r^2s^2}{2(r+s)^2} \geqslant \frac{1}{2},$$

and therefore

$$\frac{f'(r)+f'(s)}{(r+s)^2} \geqslant \frac{1}{2(1+r^2)(1+s^2)}.$$

For $\theta = 1$, we find

$$\frac{|f'(r)-f'(s)|}{(r-s)^2} = \frac{|r^2(1+s^2)-s^2(1+r^2)|}{(1+r^2)(1+s^2)(r-s)^2} = \frac{r+s}{|r-s|}\frac{f'(r)}{r^2}\frac{f'(s)}{s^2}.$$

We have, with $u = \min(r,s)\max(r,s)^{-1}$,

$$\frac{r+s}{|r-s|} = \frac{1+u}{1-u} = 1 + \frac{2u}{1-u} \geqslant 1.$$

We conclude that

$$\frac{(f'(|x|)\omega_x - f'(|y|)\omega_y)\cdot(x-y)}{|x-y|^3} \geqslant \frac{1}{2}\frac{f'(|x|)}{|x|^2}\frac{f'(|y|)}{|y|^2}$$

for all $x \neq y \in \mathbb{R}^3$, as was stated.

In the radial case we have to compute the integral over the angle explicitly. We use the notation $r_< := \min(r, s)$ and $r_> := \max(r, s)$, and we get

$$
\begin{aligned}
&\fint_{S^2}\fint_{S^2} \frac{(f'(|x|)\omega_x - f'(|y|)\omega_y)\cdot(x-y)}{|x-y|^3}\,d\omega_x\,d\omega_y \\
&\qquad = \frac{1}{2}\int_{-1}^{1} \frac{rf'(r)+sf'(s)-\theta(sf'(r)+rf'(s))}{(r^2+s^2-2rs\theta)^{3/2}}\,d\theta \\
&\qquad = \frac{rf'(r)+sf'(s)}{r_>^3}\frac{1}{2}\int_{-1}^{1} d\theta \frac{1}{(1+u^2-2u\theta)^{3/2}} - \frac{(rf'(s)+sf'(r))}{r_>^3}\frac{1}{2}\int_{-1}^{1} d\theta \frac{\theta}{(1+u^2-2u\theta)^{3/2}} \\
&\qquad = \frac{rf'(r)+sf'(s)}{r_>^3}\frac{1}{1-u^2} - \frac{(rf'(s)+sf'(r))}{r_>^3}\frac{u}{1-u^2} \\
&\qquad = \frac{1}{r_>(r_>^2-r_<^2)}(r_>f'(r_>)+r_<f'(r_<)-(r_<f'(r_>)+r_>f'(r_<))r_</r_>) \\
&\qquad = \frac{1}{r_>(r_>^2-r_<^2)}(r_>f'(r_>)-r_<^2f'(r_>)/r_>) = \frac{f'(r_>)}{r_>^2}.
\end{aligned}
$$

This calculation is valid for an arbitrary radial differentiable function f, not just the specific f chosen above. The proof of Lemma 10 is now complete. ☐

We apply (3-3) for $f_R = R^3 f(\cdot/R)$ with f given by (3-7). We get the expression

$$
\begin{aligned}
&\frac{d}{dt}\langle u(t), A_{f_R}u(t)\rangle \\
&\quad = 4R\int_{\mathbb{R}^3} \frac{Rf'(|x|/R)}{|x|}|P_x^\perp \nabla u(t,x)|^2\,dx + 4R\int_{\mathbb{R}^3} f''(|x|/R)\left|\omega_x\cdot\nabla u(t,x)+\frac{u(t,x)}{|x|}\right|^2 dx \\
&\qquad - \frac{1}{R}\int_{\mathbb{R}^3} f^{(4)}(|x|/R)|u(t,x)|^2\,dx - 2Z\int_{\mathbb{R}^3}\frac{R^2f'(|x|/R)}{|x|^2}|u(t,x)|^2\,dx \\
&\qquad + R^2\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{(f'(|x|/R)\omega_x - f'(|y|/R)\omega_y)\cdot(x-y)}{|x-y|^3}|u(t,x)|^2|u(t,y)|^2\,dx\,dy. \qquad (3\text{-}12)
\end{aligned}
$$

We now define the localized mass by

$$
M_R(t) := \int_{\mathbb{R}^3}\frac{f_R'(|x|)}{|x|^2}|u(t,x)|^2\,dx = \int_{\mathbb{R}^3}\frac{1}{1+R^{-2}|x|^2}|u(t,x)|^2\,dx. \qquad (3\text{-}13)
$$

Using (3-9) (or (3-10) in the radial case), we get the lower bound

$$
\frac{d}{dt}\langle u(t), A_{f_R}u(t)\rangle \geqslant -\frac{1}{R}\int_{\mathbb{R}^3} f^{(4)}(|x|/R)|u(t,x)|^2\,dx - 2ZM_R(t) + \kappa M_R(t)^2 \qquad (3\text{-}14)
$$

with $\kappa = 1$ in the radial case and $\kappa = \frac{1}{2}$ otherwise. Finally we remark that

$$
-f^{(4)}(r) = 24r\frac{1-r^2}{(1+r^2)^4} \geqslant -24\frac{r^3}{(1+r^2)^4}\mathbb{1}(r\geqslant 1) \geqslant -\frac{3}{1+r^2} = -3\frac{f'(r)}{r^2}
$$

(the best numerical constant is 1.33 instead of 3) and we get our final lower bound

$$\frac{d}{dt}\langle u(t),\, A_{f_R}u(t)\rangle \geqslant -\left(2Z+\frac{3}{R}\right)M_R(t)+\kappa M_R(t)^2. \tag{3-15}$$

To conclude our proof of (1-9), we average (3-15) over a time interval $[0,\,T]$ and use Jensen's inequality,

$$\frac{1}{T}\int_0^T M_R(t)^2\,dt \geqslant \left(\frac{1}{T}\int_0^T M_R(t)\,dt\right)^2,$$

to get

$$\frac{\langle u(T),\, A_{f_R}u(T)\rangle-\langle u(0),\, A_{f_R}u(0)\rangle}{T} \geqslant \kappa\left(\frac{1}{T}\int_0^T M_R(t)\,dt\right)^2-(2Z+3/R)\left(\frac{1}{T}\int_0^T M_R(t)\,dt\right). \tag{3-16}$$

Note that

$$|\langle u(t),\, A_{f_R}u(t)\rangle| = \left|\big\langle u(t),\,(p\cdot\nabla f_R(|x|)+\nabla f_R(|x|)\cdot p)u(t)\big\rangle\right| \leqslant 2\sqrt{K}\sqrt{N}\|f_R'\|_{L^\infty}=2\sqrt{K}\sqrt{N}\,R^2,$$

since $\sup_{r\geqslant 0} f'(r)=1$, and where we recall that $K=\sup_t\|\nabla u(t)\|_{L^2}$. In summary, we conclude that

$$\kappa\left(\frac{1}{T}\int_0^T M_R(t)\,dt\right)^2-(2Z+3/R)\left(\frac{1}{T}\int_0^T M_R(t)\,dt\right)\leqslant 4\sqrt{KN}\frac{R^2}{T}.$$

Using $\sqrt{1+u}\leqslant 1+u/2$, this implies

$$\begin{aligned}\frac{1}{T}\int_0^T M_R(t)\,dt &\leqslant \frac{2Z+3/R}{2\kappa}+\frac{2Z+3/R}{2\kappa}\sqrt{1+\frac{16\kappa\sqrt{KN}R^2}{(2Z+3/R)^2T}}\\ &\leqslant \frac{2Z}{\kappa}+\frac{3}{R}+\frac{4\sqrt{KN}R^2}{(2Z+3/R)T}\leqslant \frac{2Z}{\kappa}+\frac{3}{R}+\frac{2\sqrt{KN}R^2}{ZT},\end{aligned}$$

which ends the proof of (1-9).

Remark 11. Our proof works unchanged for a more general time average based on a positive function μ such that $\int_0^\infty\mu=1$ and μ' is a bounded Borel measure. More precisely, we have the estimate

$$\int_0^\infty\frac{\mu(t/T)}{T}\,dt\int_{\mathbb{R}^3}\frac{|u(t,x)|^2}{1+|x|^2/R^2}\,dx\leqslant\frac{2Z}{\kappa}+\frac{3}{R}+\frac{\sqrt{KN}R^2}{ZT}\int_0^\infty|\mu'|.$$

For instance, one could take $\mu(t)=e^{-t}$.

Step 3: *estimate on the local kinetic energy*. We show here that the kinetic energy also has a universal upper bound in average, on any ball of radius R. This time, we use a localized virial identity based on the function

$$g_R(|x|)=R^2g(|x|/R),$$

which behaves like $|x|^2$ on B_R and like $|x|$ at infinity. More precisely, we take

$$g(r)=r-\log(1+r) \tag{3-17}$$

which is such that

$$g'(r)=\frac{r}{1+r}=1-\frac{1}{1+r}.$$

Clearly g' is positive and nondecreasing, therefore $x\mapsto g(|x|)$ is convex on $\mathbb{R}^3$.

We use the lower bound (2-4) with $\alpha=0$ and we get, by the same calculations as before,

$$\begin{aligned}\frac{d}{dt}\langle u(t),A_{g_R}u(t)\rangle=4&\int_{\mathbb{R}^3}\left(\frac{Rg'(|x|/R)}{|x|}|P_x^{\perp}\nabla u(t,x)|^2+g''(|x|/R)|\omega_x\cdot\nabla u(t,x)|^2\right)dx\\&-\frac{1}{R^2}\int_{\mathbb{R}^3}\left(g^{(4)}(|x|/R)+4\frac{Rg^{(3)}(|x|/R)}{|x|}\right)|u(t,x)|^2\,dx\\&+R\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{(\nabla g(x/R)-\nabla g(y/R))\cdot(x-y)}{|x-y|^3}|u(t,x)|^2|u(t,y)|^2\,dx\,dy\\&-2Z\int_{\mathbb{R}^3}\frac{Rg'(|x|/R)}{|x|^2}|u(t,x)|^2\,dx.\end{aligned}\tag{3-18}$$

We denote by

$$K_R(t):=\int_{\mathbb{R}^3}g''(|x|/R)|\nabla u(t,x)|^2\,dx=\int_{\mathbb{R}^3}\frac{|\nabla u(t,x)|^2}{(1+R^{-1}|x|)^2}\,dx$$

the local kinetic energy. Since $x\mapsto g(|x|)$ is convex,

$$(\nabla g(x)-\nabla g(y))\cdot(x-y)\geqslant 0$$

for all $x,y\in\mathbb{R}^3$. Also, we notice that

$$g''(r)=\frac{1}{(1+r)^2}\leqslant\frac{1}{1+r}=\frac{g'(r)}{r}.$$

Finally, we compute

$$g^{(3)}(r)=-\frac{2}{(1+r)^3}\leqslant 0$$

and

$$g^{(4)}(r)=\frac{6}{(1+r)^4}\leqslant\frac{6}{1+r^2}=6\frac{f'(r)}{r^2}.$$

So we arrive at the estimate

$$\frac{d}{dt}\langle u(t),A_{g_R}u(t)\rangle\geqslant 4K_R(t)-\frac{6}{R^2}M_R(t)-2Z\int_{\mathbb{R}^3}\frac{Rg'(|x|/R)}{|x|^2}|u(t,x)|^2\,dx.\tag{3-19}$$

In order to control the negative term, we again use Hardy's trick:

$$\begin{aligned}0 &\leqslant \int_{\mathbb{R}^3} g''(|x|/R)|\nabla u(t,x)+\alpha\omega_x u(t,x)|^2\,dx\\ &= \int_{\mathbb{R}^3} g''(|x|/R)|\nabla u(t,x)|^2\,dx - \alpha\int_{\mathbb{R}^3}\operatorname{div}(\omega_x g''(|x|/R))|u(t,x)|^2\,dx + \alpha^2\int_{\mathbb{R}^3} g''(|x|/R)|u(t,x)|^2\,dx\\ &= \int_{\mathbb{R}^3} g''(|x|/R)|\nabla u(t,x)|^2\,dx + \alpha^2\int_{\mathbb{R}^3} g''(|x|/R)|u(t,x)|^2\,dx\\ &\qquad\qquad - 2\alpha\int_{\mathbb{R}^3}\frac{g''(|x|/R)}{|x|}|u(t,x)|^2\,dx - \frac{\alpha}{R}\int_{\mathbb{R}^3} g^{(3)}(|x|/R)|u(t,x)|^2\,dx.\end{aligned}$$

Therefore, using that $-g^{(3)}(r) = 2(1+r)^{-3} \leqslant 2(1+r^2)^{-1} = 2f'(r)r^{-2}$ and that $g''(r) = (1+r)^{-2} \leqslant (1+r^2)^{-1} = f'(r)r^{-2}$, we find

$$\int_{\mathbb{R}^3}\frac{g''(|x|/R)}{|x|}|u(t,x)|^2\,dx \leqslant \frac{1}{2\alpha}K_R(t) + \left(\frac{\alpha}{2}+\frac{1}{R}\right)M_R(t).$$

Coming back to the negative term in (3-18), we write

$$\begin{aligned}\int_{\mathbb{R}^3}\frac{Rg'(|x|/R)}{|x|^2}|u(t,x)|^2\,dx &= \int_{\mathbb{R}^3}\frac{1}{|x|(1+|x|/R)}|u(t,x)|^2\,dx\\ &= \int_{\mathbb{R}^3}\frac{1}{|x|(1+|x|/R)^2}|u(t,x)|^2\,dx + \frac{1}{R}\int_{\mathbb{R}^3}\frac{1}{(1+|x|/R)^2}|u(t,x)|^2\,dx\\ &\leqslant \frac{1}{2\alpha}K_R(t) + \left(\frac{\alpha}{2}+\frac{2}{R}\right)M_R(t).\end{aligned}$$

Inserting in (3-19) gives

$$\frac{d}{dt}\langle u(t), A_{g_R}u(t)\rangle \geqslant \left(4-\frac{Z}{\alpha}\right)K_R(t) - Z\left(\alpha+\frac{4}{R}+\frac{6}{R^2}\right)M_R(t). \tag{3-20}$$

Taking $\alpha = Z/2$ leads to

$$\frac{d}{dt}\langle u(t), A_{g_R}u(t)\rangle \geqslant 2K_R(t) - Z\left(\frac{Z}{2}+\frac{4}{R}+\frac{6}{R^2}\right)M_R(t). \tag{3-21}$$

To conclude our proof, we average over t in an interval $[0, T]$ using that

$$|\langle u(t), A_{g_R}u(t)\rangle| \leqslant 2\sqrt{K}\sqrt{N}\|g_R'\|_{L^\infty} = 2R\sqrt{K}\sqrt{N},$$

and we get

$$\frac{1}{T}\int_0^T K_R(t)\,dt \leqslant Z\left(\frac{Z}{4}+\frac{2}{R}+\frac{3}{R^2}\right)\frac{1}{T}\int_0^T M_R(t)\,dt + \frac{2R\sqrt{K}\sqrt{N}}{T}, \tag{3-22}$$

which concludes the proof of (1-12).

Step 4: *Estimate on K*. We end the proof of Theorem 4 by estimating the maximal value K of the kinetic energy of $u(t)$ in terms of $\|u_0\|_{H^1}$, using the conservation of energy.

Lemma 12 (kinetic energy estimate). *We have*, *for all $t \in \mathbb{R}$*,

$$\|\nabla u(t)\|^2_{L^2(\mathbb{R}^3)} \leqslant Z^2\|u_0\|^2_{L^2(\mathbb{R}^3)} + \|\nabla u_0\|^2_{L^2(\mathbb{R}^3)} + \tfrac{1}{2}\|u_0\|^3_{L^2(\mathbb{R}^3)}\|\nabla u_0\|_{L^2(\mathbb{R}^3)}. \tag{3-23}$$

Proof of Lemma 12. By conservation of energy and mass, we find

$$\mathscr{E}_Z(u_0)=\mathscr{E}_Z(u)\geqslant\left\langle u,\left(-\frac{\Delta}{2}-\frac{Z}{|x|}\right)u\right\rangle+\tfrac{1}{2}\|\nabla u\|^2_{L^2(\mathbb{R}^3)}\geqslant-\frac{Z^2}{2}\int_{\mathbb{R}^3}|u_0|^2+\tfrac{1}{2}\|\nabla u\|^2_{L^2(\mathbb{R}^3)},$$

since $-\Delta/2-Z|x|^{-1}\geqslant-Z^2/2$ (hydrogen atom). Next, for $x\in\mathbb{R}^3$ and $u\in H^1(\mathbb{R}^3)$, we note the bound

$$\int_{\mathbb{R}^3}\frac{|u(y)|^2}{|x-y|}\,dy\leqslant\min_{z\geqslant 0}\left(\frac{z}{2}\int_{\mathbb{R}^3}|u|^2+\frac{1}{2z}\int_{\mathbb{R}^3}|\nabla u|^2\right)=\|u\|_{L^2(\mathbb{R}^3)}\|\nabla u\|_{L^2(\mathbb{R}^3)},\tag{3-24}$$

which gives us

$$\mathscr{E}_Z(u_0)\leqslant\|\nabla u_0\|^2_{L^2(\mathbb{R}^3)}+\tfrac{1}{2}\|u_0\|^3_{L^2(\mathbb{R}^3)}\|\nabla u_0\|_{L^2(\mathbb{R}^3)}.$$

Hence,

$$\|\nabla u\|^2_{L^2(\mathbb{R}^3)}\leqslant Z^2\|u_0\|^2_{L^2(\mathbb{R}^3)}+2\|\nabla u_0\|^2_{L^2(\mathbb{R}^3)}+\|u_0\|^3_{L^2(\mathbb{R}^3)}\|\nabla u_0\|_{L^2(\mathbb{R}^3)}.$$ ☐

This concludes the proof of Theorem 4. ☐

4. Extensions: Hartree–Fock and many-body Schrödinger theories

4.1. ***Hartree–Fock theory.*** The Hartree–Fock equations describe the nonlinear evolution of a wave function taking the form of a Slater determinant, that is,

$$\Psi(t)=\frac{1}{\sqrt{N!}}\sum_{\sigma\in\mathfrak{S}_N}\operatorname{sgn}(\sigma)u_1(t,x_{\sigma(1)})\cdots u_N(t,x_{\sigma(N)}),$$

where the functions $u_1,\dots,u_N$ model the states of the N electrons. The physical fact that electrons are fermions is expressed in the Pauli principle given by the orthonormality condition

$$\langle u_j,u_k\rangle_{L^2}=\delta_{jk}.$$

The Hartree–Fock equations [Lieb and Simon 1977; Chadam 1976; Bove et al. 1976] form a system of N coupled nonlinear equations similar to (1-1):

$$\begin{cases} i\dfrac{\partial}{\partial t}u_j=H_u u_j,\\ H_u v=-\Delta v-Z|x|^{-1}v+\displaystyle\sum_{k=1}^N|u_k|^2*|x|^{-1}v-\sum_{k=1}^N(\overline{u_j}v)*|x|^{-1}u_k.\end{cases}\tag{4-1}$$

One simple way to write the same equation is to introduce the one-body density matrix

$$\gamma(t):=\sum_{k=1}^N|u_k\rangle\langle u_k|,$$

which is the orthogonal projection onto the space spanned by the functions $u_1,\dots,u_N$. Then (4-1) is equivalent to the so-called von Neumann equation,

$$\begin{cases} i\dfrac{\partial}{\partial t}\gamma=[H_\gamma,\gamma],\\ H_\gamma v=-\Delta v-Z|x|^{-1}v+\rho_{\gamma(t)}*|x|^{-1}v-\displaystyle\int_{\mathbb{R}^3}\frac{\gamma(t,x,y)}{|x-y|}v(y)\,dy.\end{cases}\tag{4-2}$$

Here $\rho_\gamma(x) := \gamma(x,x)$ is the density associated with the matrix γ. The time-dependent equation (4-2) does in fact make sense for any trace-class operator γ such that

$$0 \leqslant \gamma \leqslant 1 \quad \text{and} \quad \mathrm{Tr}(\gamma) = N,$$

which corresponds to *generalized Hartree–Fock states* [Bach et al. 1994]. Note that the infinite-rank case $\operatorname{rank}\gamma = +\infty$ is also allowed here. We refer to [Chadam 1976; Bove et al. 1976] for the proof of global well-posedness for (4-2) with initial data such that $\mathrm{Tr}(1-\Delta)\gamma_0 < +\infty$.

The following result is the equivalent of Theorem 4 in the Hartree–Fock case.

Theorem 13 (long-time behavior of atoms in Hartree–Fock theory). *Suppose $Z > 0$ and let γ_0 be an arbitrary initial datum such that*

$$\mathrm{Tr}(1-\Delta)\gamma_0 < \infty.$$

Denote by $\gamma(t)$ the unique solution of (4-2). *Then we have the estimate*

$$\frac{1}{T}\int_0^T dt \int_{\mathbb{R}^3} dx \frac{\rho_{\gamma(t)}(x)}{1+|x|^2/R^2} \leqslant 4Z + 1 + \frac{3}{R} + \frac{2\sqrt{KN}R^2}{ZT} \tag{4-3}$$

with

$$N := \mathrm{Tr}(\gamma_0)$$

and

$$K := \sup_{t\geqslant 0} \mathrm{Tr}(-\Delta)\gamma(t) \leqslant Z^2 N + 2\,\mathrm{Tr}(-\Delta)\gamma_0 + N^3\sqrt{\mathrm{Tr}(-\Delta)\gamma_0}. \tag{4-4}$$

In particular, we have

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T dt \int_{|x|\leqslant r} dx \rho_{\gamma(t)}(x) \leqslant 4Z + 1 \tag{4-5}$$

for every $r > 0$. Similarly, we have the following estimate on the local kinetic energy:

$$\frac{1}{T}\int_0^T dt \int_{\mathbb{R}^3} dx \frac{\tau_{\gamma(t)}(x)}{(1+|x|/R)^2} \leqslant \left(\frac{Z^2}{4} + \frac{2Z}{R} + \frac{3Z}{R^2}\right)\frac{1}{T}\int_0^T dt \int_{\mathbb{R}^3} \frac{\rho_{\gamma(t)}(x)}{1+|x|^2/R^2}\,dx + \frac{2R\sqrt{K}\sqrt{N}}{T}, \tag{4-6}$$

where $\tau_\gamma(x) = -\sum_{k=1}^3(\partial_k\gamma\partial_k)(x,x)$ is the density of kinetic energy, and therefore

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T dt \int_{|x|\leqslant r} dx \tau_{\gamma(t)}(x) \leqslant \frac{Z^2}{4}(4Z+1) \tag{4-7}$$

for every $r > 0$.

If the initial datum γ_0 is radial in the sense that

$$\gamma_0(\mathcal{R}x, \mathcal{R}y) = \gamma_0(x,y) \quad \text{for all } x, y \in \mathbb{R}^3 \text{ and all } \mathcal{R} \in \mathrm{SO}(3),$$

then $\gamma(t)$ is radial for all times and the same estimates (4-3) *and* (4-7) *hold true with $4Z+1$ replaced by $2Z+1$.*

The proof of Theorem 13 is very similar to that of Theorem 4, the main new difficulty being the control of the exchange term. Thus we only explain how to deal with it.

Sketch of the proof of Theorem 13. First, we consider a sufficiently smooth radial function $f = f(|x|)$. (Below we will take $f = f_R$, the same as in the proof of Theorem 4.) Differentiating with respect to t, we find

$$\begin{aligned}\frac{d}{dt}\operatorname{Tr}(A_f\gamma) &= i\operatorname{Tr}([H_\gamma, A_f]\gamma)\\ &= -\operatorname{Tr}([p^2,[p^2,f]]\gamma) + i\operatorname{Tr}([V_\gamma, A_f]\gamma) - i\operatorname{Tr}([X_\gamma, A_f]\gamma),\end{aligned}\tag{4-8}$$

where $V_\gamma = -Z|x|^{-1} + |x|^{-1} * \rho_\gamma$ and X_γ is the *exchange term* defined by

$$(X_\gamma u)(x) = \int_{\mathbb{R}^3} \frac{\gamma(x,y)}{|x-y|} u(y)\,dy.$$

Note that

$$i[V_\gamma, A_f] = -2\nabla f\cdot\nabla V_\gamma$$

is a function (that is, a multiplication operator). Analogous to the Hartree case, we thus obtain

$$\begin{aligned}i\operatorname{Tr}([V_\gamma, A_f]\gamma) &= -2\int_{\mathbb{R}^3}\rho_\gamma(x)\nabla f(x)\cdot\nabla V_\gamma(x)\,dx\\ &= -2Z\int_{\mathbb{R}^3}\frac{f'(|x|)}{|x|^2}\rho_\gamma(t,x)\,dx + \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{(\nabla f(x)-\nabla f(y))\cdot(x-y)}{|x-y|^3}\rho_\gamma(x)\rho_\gamma(y)\,dx\,dy.\end{aligned}$$

The exchange term is controlled using the following fact.

Lemma 14 (exchange term). *Let* $\operatorname{Tr}(1-\Delta)\gamma < +\infty$ *and suppose* $f : \mathbb{R}^d \to \mathbb{R}$ *satisfies* $\nabla f \in L^\infty(\mathbb{R}^d)$. *Then we have*

$$i\operatorname{Tr}([X_\gamma, A_f]\gamma) = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{(\nabla f(x)-\nabla f(y))\cdot(x-y)}{|x-y|^3}|\gamma(x,y)|^2\,dx\,dy.\tag{4-9}$$

Proof. The proof is an explicit computation:

$$\begin{aligned}&i\operatorname{Tr}([X_\gamma, A_f]\gamma)\\ &\quad= i\operatorname{Tr}([X_\gamma,(p\cdot(\nabla f)+(\nabla f)\cdot p)]\gamma)\\ &\quad= +\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}X_\gamma(x,y)\nabla_y\cdot(\nabla f)(y)\gamma(y,x)\,dx\,dy - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\gamma(y,x)\nabla_x\cdot(\nabla f)(x)X_\gamma(x,y)\,dx\,dy\\ &\qquad+\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}X_\gamma(x,y)(\nabla f)(y)\cdot\nabla_y\gamma(y,x)\,dx\,dy - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\gamma(y,x)(\nabla f)(x)\cdot\nabla_x X_\gamma(x,y)\,dx\,dy.\end{aligned}$$

Integrating by parts for the first two terms, we find

$$\begin{aligned}&i\operatorname{Tr}([X_\gamma, A_f]\gamma)\\ &\quad= -\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\gamma(y,x)(\nabla f)(y)\cdot\nabla_y X_\gamma(x,y)\,dx\,dy + \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}X_\gamma(x,y)(\nabla f)(x)\cdot\nabla_x\gamma(y,x)\,dx\,dy\\ &\qquad+\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}X_\gamma(x,y)(\nabla f)(y)\cdot\nabla_y\gamma(y,x)\,dx\,dy - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\gamma(y,x)(\nabla f)(x)\cdot\nabla_x X_\gamma(x,y)\,dx\,dy.\end{aligned}$$

Now we use that

$$\nabla_y X_\gamma(x,y) = \frac{1}{|x-y|}\nabla_y\gamma(x,y) + \gamma(x,y)\nabla_y\frac{1}{|x-y|}$$

and we exchange x and y in the second and fourth integrals. The final result is

$$\begin{aligned} i\,\mathrm{Tr}([X_\gamma, A_f]\gamma) &= -\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} |\gamma(x,y)|^2 \Big((\nabla f)(y)\cdot\nabla_y \frac{1}{|x-y|} + (\nabla f)(x)\cdot\nabla_x \frac{1}{|x-y|} \Big)\,dx\,dy \\ &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{(\nabla f(x)-\nabla f(y))\cdot(x-y)}{|x-y|^3} |\gamma(x,y)|^2\,dx\,dy. \end{aligned}$$ □

Inserting this in (4-8) gives the following value for the derivative of the expectation value of A_f:

$$\begin{aligned} \frac{d}{dt}\mathrm{Tr}(A_f\gamma) = -\mathrm{Tr}([p^2,[p^2,f]]\gamma) - 2Z\int_{\mathbb{R}^3} \frac{f'(|x|)}{|x|^2}\rho_\gamma(t,x)\,dx \\ + \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{(\nabla f(x)-\nabla f(y))\cdot(x-y)}{|x-y|^3}(\rho_\gamma(x)\rho_\gamma(y)-|\gamma(x,y)|^2)\,dx\,dy. \end{aligned} \tag{4-10}$$

Since f is convex, we have the operator bound

$$-[p^2,[p^2,f]] \geqslant -f^{(4)}(|x|),$$

which gives

$$-\mathrm{Tr}([p^2,[p^2,f]]\gamma) \geqslant -\mathrm{Tr}(f^{(4)}\gamma) = -\int_{\mathbb{R}^3} f^{(4)}(|x|)\rho_{\gamma(t)}(x)\,dx$$

because $\gamma \geqslant 0$. Thus we can argue exactly as in the Hartree case. We start by taking f_R given by (3-7) and define the local mass by

$$M_R(t) := \int_{\mathbb{R}^3} \frac{f'(|x|)}{|x|^2}\rho_{\gamma(t)}(x)\,dx.$$

Then we use the bound (3-9), that is,

$$\frac{(\nabla f_R(x)-\nabla f_R(y))\cdot(x-y)}{|x-y|^3} \geqslant \frac{1}{2}\,\frac{R^2 f'(|x|/R)}{|x|^2}\,\frac{R^2 f'(|y|/R)}{|y|^2},$$

as well as the fact that $\rho_\gamma(x)\rho_\gamma(y) \geqslant |\gamma(x,y)|^2$ for a.e. $x, y \in \mathbb{R}^3$ (by the Cauchy–Schwarz inequality and the eigenfunction expansion for γ.) This gives

$$\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{(\nabla f_R(x)-\nabla f_R(y))\cdot(x-y)}{|x-y|^3}(\rho_\gamma(x)\rho_\gamma(y)-|\gamma(x,y)|^2)\,dx\,dy \geqslant \frac{M_R(t)^2}{2} - \tfrac{1}{2}\mathrm{Tr}(h_R\gamma h_R\gamma),$$

with $h_R := R^2 f'(|x|/R)|x|^{-2}$. Since $0 \leqslant \gamma \leqslant 1$ and $0 \leqslant h_R \leqslant 1$, we have $h_R\gamma h_R \leqslant (h_R)^2 \leqslant h_R$, and therefore

$$\mathrm{Tr}(h_R\gamma h_R\gamma) \leqslant \mathrm{Tr}(h_R\gamma) = M_R(t).$$

We conclude that

$$\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{(\nabla f(x)-\nabla f(y))\cdot(x-y)}{|x-y|^3}(\rho_\gamma(x)\rho_\gamma(y)-|\gamma(x,y)|^2)\,dx\,dy \geqslant \frac{M_R(t)^2-M_R(t)}{2}.$$

The additional term is responsible for the change of $4Z$ into $4Z+1$. In the radial case, we use (3-10) instead and we get rid of the factor of $\frac{1}{2}$ on the left side. The rest of the proof is exactly the same as in the Hartree case. □

4.2. ***Many-body Schrödinger equation.*** Our method also applies to the linear many-body Schrödinger equation

$$\begin{cases} i\dfrac{\partial}{\partial t}\Psi(t)=H(N,Z)\Psi(t),\\ H(N,Z)=\displaystyle\sum_{j=1}^{N}\left(-\Delta_{x_j}-\frac{Z}{|x_j|}\right)+\frac{1}{2}\sum_{1\leqslant k\neq\ell\leqslant N}\frac{1}{|x_k-x_\ell|},\\ \Psi(0)=\Psi_0\in H^1((\mathbb{R}^3)^N),\end{cases}\tag{4-11}$$

of which the Hartree and Hartree–Fock models are nonlinear approximations.

The Hamiltonian $H(N,Z)$ is self-adjoint and bounded from below on $L^2((\mathbb{R}^3)^N)$. Its operator domain is $H^2((\mathbb{R}^3)^N)$ and its quadratic form domain is $H^1((\mathbb{R}^3)^N)$. Of particular interest are its restrictions to the symmetric (bosonic) and antisymmetric (fermionic) subspaces. These are also self-adjoint operators, denoted, respectively, by $H_s(N,Z)$ and $H_a(N,Z)$. In either of these two subspaces, the essential spectrum of $H_{a/s}(N,Z)$ is a half line $[\Sigma_{a/s}(N,Z),\infty)$ where

$$\Sigma_{a/s}(N,Z)=\inf\operatorname{Spec}(H_{a/s}(N-1,Z)),$$

by the HVZ Theorem [Reed and Simon 1978; Cycon et al. 1987]. It is known that there are no positive eigenvalues [Froese and Herbst 1982], but there might be embedded eigenvalues in $[\Sigma_{a/s}(N,Z),0]$. There exists a critical number of particles $N^c_{a/s}(Z)$ such that $H_{a/s}(N,Z)$ has no eigenvalues below $\Sigma_{a/s}(N,Z)$ for $N>N^c_{a/s}(Z)$; see [Ruskai 1982; Sigal 1982; 1984]. For bosons, it is known that

$$\lim_{Z\to\infty}\frac{N^c_s(Z)}{Z}=\tilde{\gamma}_c,$$

where $\tilde{\gamma}_c\simeq 1.21\leqslant\gamma_c$ is the largest number of electrons that ground states can have in Hartree theory [Benguria and Lieb 1983; Baumgartner 1984; Solovej 1990]. For fermions, it was proved [Lieb et al. 1988] that

$$\lim_{Z\to\infty}\frac{N^c_a(Z)}{Z}=1.$$

The best bound valid for all N goes back to [Lieb 1984] and it holds both for bosons and fermions: $N^c_{a/s}(Z)<2Z+1$. For fermions, it was improved to

$$N^c_a(Z)<1.22Z+3Z^{1/3}$$

by Nam [2012].

All the previous authors seem to have only studied when the Hamiltonian $H_{a/s}(N,Z)$ ceases to have eigenvalues below its essential spectrum. The question of the existence of embedded eigenvalues in $[\Sigma_{a/s}(N,Z),0]$ does not seem to have been addressed so far. But this is a relevant problem in the context of the time-dependent equation. Our method allows us to prove that there are no eigenvalue at all when $N\geqslant 4Z+1$.

Theorem 15 (linear many-body Schrödinger equation). *The Hamiltonian* $H(N,Z)$ *has no eigenvalue when* $N\geqslant 4Z+1$.

Here we do not distinguish between the different particle statistics. Thus our result applies to all of $L^2((\mathbb{R}^3)^N)$ and it deals with all possible symmetries. We, however, conjecture that the largest N such that $H_{a/s}(N, Z)$ can have eigenvalues behaves like $N^c_{a/s}(Z)$ for large Z.

Proof. Let $\Psi \in H^2((\mathbb{R}^3)^N)$ be an eigenfunction of $H(N, Z)$ and let $f_R(|x|) = R^3 f(|x|/R)$ be as in (3-7). Then we write

$$\begin{aligned}
0 &= \Bigl\langle \Psi, i\Bigl(H(N, Z) \sum_{j=1}^{N} (A_{f_R})_{x_j} - \sum_{j=1}^{N} (A_{f_R})_{x_j} H(N, Z)\Bigr)\Psi\Bigr\rangle \\
&= \sum_{j=1}^{N} \langle \Psi, i[p_j^2, (A_{f_R})_{x_j}]\Psi\rangle - 2\sum_{j=1}^{N} \Bigl\langle \Psi, \nabla f_R(x_j)\cdot \nabla_{x_j}\Bigl(-\frac{Z}{|x_j|} + \frac{1}{2}\sum_{k\neq j} \frac{1}{|x_j - x_k|}\Bigr)\Psi\Bigr\rangle \\
&> -\frac{1}{R}\int_{\mathbb{R}^3} f^{(4)}\Bigl(\frac{|x|}{R}\Bigr)\rho_\Psi(x)\,dx - 2Z\int_{\mathbb{R}^3} \frac{R^2 f'(|x|/R)}{|x|^2}\rho_\Psi(x)\,dx \\
&\qquad\qquad + \Bigl\langle \Psi, \Bigl(\sum_{1\leqslant j\neq k\leqslant N} \frac{(\nabla f_R(x_j) - \nabla f_R(x_k))\cdot (x_j - x_k)}{|x_j - x_k|^3}\Bigr)\Psi\Bigr\rangle.
\end{aligned}$$

Using (3-9), we get

$$\begin{aligned}
&\Bigl\langle \Psi, \Bigl(\sum_{1\leqslant j\neq k\leqslant N} \frac{(\nabla f_R(x_j) - \nabla f_R(x_k))\cdot (x_j - x_k)}{|x_j - x_k|^3}\Bigr)\Psi\Bigr\rangle \\
&\qquad \geqslant \frac{1}{2}\Bigl\langle \Psi, \Bigl(\sum_{1\leqslant j\neq k\leqslant N} \frac{R^2 f_R'(|x_j|)}{|x_j|^2}\frac{R^2 f_R'(|x_k|)}{|x_k|^2}\Bigr)\Psi\Bigr\rangle \\
&\qquad = \frac{1}{2}\Bigl\langle \Psi, \Bigl(\sum_{j=1}^{N} \frac{R^2 f_R'(|x_j|)}{|x_j|^2}\Bigr)^2 \Psi\Bigr\rangle - \frac{1}{2}\Bigl\langle \Psi, \Bigl(\sum_{j=1}^{N} \Bigl(\frac{R^2 f_R'(|x_j|)}{|x_j|^2}\Bigr)^2\Bigr)\Psi\Bigr\rangle \\
&\qquad \geqslant \frac{1}{2}\Bigl\langle \Psi, \Bigl(\sum_{j=1}^{N} \frac{R^2 f_R'(|x_j|)}{|x_j|^2}\Bigr)\Psi\Bigr\rangle^2 - \frac{1}{2}\Bigl\langle \Psi, \Bigl(\sum_{j=1}^{N} \frac{R^2 f_R'(|x_j|)}{|x_j|^2}\Bigr)\Psi\Bigr\rangle \\
&\qquad = \frac{1}{2}\Bigl(\int_{\mathbb{R}^3} \frac{R^2 f'(|x|/R)}{|x|^2}\rho_\Psi(x)\,dx\Bigr)^2 - \frac{1}{2}\int_{\mathbb{R}^3} \frac{R^2 f'(|x|/R)}{|x|^2}\rho_\Psi(x)\,dx.
\end{aligned}$$

In the last line we have used Jensen's inequality as well as the fact that $f'(r)/r^2 = 1/(1+r^2) \leqslant 1$. Passing to the limit as $R \to \infty$ gives $N < 4Z + 1$. □

Since $H(N, Z)$ has no eigenvalue when $N \geqslant 4Z + 1$, it follows from the known existence of scattering and the asymptotic completeness [Dereziński 1993; Sigal and Soffer 1994; Hunziker and Sigal 2000] that any solution $\Psi(t)$ of the time-dependent equation (4-11) behaves (in an appropriate sense) as a superposition of bound states of $H(k, Z)$ with $k < 4Z + 1$ plus a scattering part. In particular, it is possible to prove that

$$\limsup_{t\to\infty} \int_{|x|\leqslant r} \rho_\Psi(t, x)\,dx \leqslant 4Z + 1.$$

By using the argument in the proof of Theorem 15 and following step by step the method of Section 3, one can get a simple proof of the weaker result

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T dt \int_{|x|\leqslant r} \rho_\Psi(t,x)\,dx \leqslant 4Z+1.$$

References

[Amrein and Georgescu 1973/74] W. O. Amrein and V. Georgescu, "On the characterization of bound states and scattering states in quantum mechanics", *Helv. Phys. Acta* **46** (1973/74), 635–658. MR 50 #15705

[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 96a:82003 Zbl 0839.60095

[Bader 1978] P. Bader, "Variational method for the Hartree equation of the helium atom", *Proc. Roy. Soc. Edinburgh Sect. A* **82**:1-2 (1978), 27–39. MR 80d:81037 Zbl 0403.45021

[Bardos et al. 2000] C. Bardos, F. Golse, and N. J. Mauser, "Weak coupling limit of the N–particle Schrödinger equation", *Methods Appl. Anal.* **7**:2 (2000), 275–293. MR 2003c:81215 Zbl 1003.81027

[Baumgartner 1984] B. Baumgartner, "On Thomas–Fermi–von Weizsäcker and Hartree energies as functions of the degree of ionisation", *J. Phys. A* **17**:8 (1984), 1593–1601. MR 85k:81192 Zbl 0541.49020

[Benguria 1979] R. Benguria, *The von Weizsacker and exchange corrections in the Thomas–Fermi theory*, Ph.D. thesis, Princeton University, 1979, Available at http://www.fis.puc.cl/∼rbenguri/PHDthesis.pdf.

[Benguria and Lieb 1983] R. Benguria and E. H. Lieb, "Proof of the stability of highly negative ions in the absence of the Pauli principle", *Phys. Rev. Lett.* **50**:22 (1983), 1771–1774.

[Benguria et al. 1981] R. Benguria, H. Brézis, and E. H. Lieb, "The Thomas–Fermi–von Weizsäcker theory of atoms and molecules", *Comm. Math. Phys.* **79**:2 (1981), 167–180. MR 83m:81114 Zbl 0478.49035

[Berestycki and Lions 1983] H. Berestycki and P.-L. Lions, "Nonlinear scalar field equations, II: Existence of infinitely many solutions", *Arch. Rational Mech. Anal.* **82**:4 (1983), 347–375. MR 84h:35054b Zbl 0556.35046

[Bove et al. 1976] A. Bove, G. Da Prato, and G. Fano, "On the Hartree–Fock time-dependent problem", *Comm. Math. Phys.* **49**:1 (1976), 25–33. MR 56 #14297

[Castella 1997] F. Castella, "L^2 solutions to the Schrödinger–Poisson system: Existence, uniqueness, time behaviour, and smoothing effects", *Math. Models Methods Appl. Sci.* **7**:8 (1997), 1051–1083. MR 99f:82044 Zbl 0892.35141

[Chadam 1976] J. M. Chadam, "The time-dependent Hartree–Fock equations with Coulomb two-body interaction", *Comm. Math. Phys.* **46**:2 (1976), 99–104. MR 53 #15173 Zbl 0322.35043

[Chadam and Glassey 1975] J. M. Chadam and R. T. Glassey, "Global existence of solutions to the Cauchy problem for time-dependent Hartree equations", *J. Mathematical Phys.* **16** (1975), 1122–1130. MR 54 #1957 Zbl 0299.35084

[Colliander et al. 2003] J. Colliander, M. Keel, H. Staffilani, H. Takaoka, and T. Tao, "Viriel, Morawetz, and interaction Morawetz inequalities", preprint, 2003, Available at http://www.math.ucla.edu/∼tao/preprints/Expository/morawetz-expository.dvi. Notes from T. Tao's website.

[Cycon et al. 1987] H. L. Cycon, R. G. Froese, W. Kirsch, and B. Simon, *Schrödinger operators with application to quantum mechanics and global geometry*, Springer, Berlin, 1987. MR 88g:35003 Zbl 0619.47005

[Dereziński 1993] J. Dereziński, "Asymptotic completeness of long-range N–body quantum systems", *Ann. of Math.* (2) **138**:2 (1993), 427–476. MR 94g:81209 Zbl 0844.47005

[Dias and Figueira 1981] J. a.-P. Dias and M. Figueira, "Conservation laws and time decay for the solutions of some nonlinear Schrödinger–Hartree equations and systems", *J. Math. Anal. Appl.* **84**:2 (1981), 486–508. MR 83d:35013 Zbl 0481.35057

[Enss 1978] V. Enss, "Asymptotic completeness for quantum mechanical potential scattering, I: Short range potentials", *Comm. Math. Phys.* **61**:3 (1978), 285–291. MR 58 #25583 Zbl 0389.47005

[Erdős and Yau 2001] L. Erdős and H.-T. Yau, "Derivation of the nonlinear Schrödinger equation from a many body Coulomb system", *Adv. Theor. Math. Phys.* **5**:6 (2001), 1169–1205. MR 2004c:82075 Zbl 1014.81063 arXiv 0111042

[Froese and Herbst 1982] R. Froese and I. Herbst, "Exponential bounds and absence of positive eigenvalues for N–body Schrödinger operators", *Comm. Math. Phys.* **87**:3 (1982), 429–447. MR 85g:35091 Zbl 0509.35061

[Fröhlich and Lenzmann 2004] J. Fröhlich and E. Lenzmann, "Mean-field limit of quantum Bose gases and nonlinear Hartree equation", pp. XIX–1–26 in *Séminaire: Équations aux Dérivées Partielles, 2003–2004*, École Polytech., Palaiseau, 2004. MR 2005m:81079

[Gasser et al. 1998] I. Gasser, R. Illner, P. A. Markowich, and C. Schmeiser, "Semiclassical, $t \to \infty$ asymptotics and dispersive effects for Hartree–Fock systems", *RAIRO Modél. Math. Anal. Numér.* **32**:6 (1998), 699–713. MR 99k:35164 Zbl 0924.35125

[Ghoussoub 1993] N. Ghoussoub, *Duality and perturbation methods in critical point theory*, Cambridge Tracts in Mathematics **107**, Cambridge University Press, 1993. MR 95a:58021 Zbl 0790.58002

[Ginibre and Ozawa 1993] J. Ginibre and T. Ozawa, "Long range scattering for nonlinear Schrödinger and Hartree equations in space dimension $n \geqslant 2$", *Comm. Math. Phys.* **151**:3 (1993), 619–645. MR 93m:35168 Zbl 0776.35070

[Ginibre and Velo 1980] J. Ginibre and G. Velo, "On a class of nonlinear Schrödinger equations with nonlocal interaction", *Math. Z.* **170**:2 (1980), 109–136. MR 82c:35018 Zbl 0407.35063

[Ginibre and Velo 2000a] J. Ginibre and G. Velo, "Long range scattering and modified wave operators for some Hartree type equations, I", *Rev. Math. Phys.* **12**:3 (2000), 361–429. MR 2001c:35177 Zbl 1044.35041

[Ginibre and Velo 2000b] J. Ginibre and G. Velo, "Long range scattering and modified wave operators for some Hartree type equations, II", *Ann. Henri Poincaré* **1**:4 (2000), 753–800. MR 2001i:35230 Zbl 1024.35084

[Glassey 1977a] R. T. Glassey, "Asymptotic behavior of solutions to certain nonlinear Schrödinger–Hartree equations", *Comm. Math. Phys.* **53**:1 (1977), 9–18. MR 58 #6638 Zbl 0339.35013

[Glassey 1977b] R. T. Glassey, "On the blowing up of solutions to the Cauchy problem for nonlinear Schrödinger equations", *J. Math. Phys.* **18**:9 (1977), 1794–1797. MR 57 #842 Zbl 0372.35009

[Hartree 1928a] D. R. Hartree, "The wave mechanics of an atom with a non-Coulomb central field, I: Theory and methods", *Proc. Camb. Phil. Soc.* **24**:1 (1928), 89–110. JFM 54.0966.05

[Hartree 1928b] D. R. Hartree, "The wave mechanics of an atom with a non-Coulomb central field, II: Some results and discussion", *Proc. Camb. Phil. Soc.* **24**:1 (1928), 111–132. JFM 54.0966.05

[Hayashi 1988] N. Hayashi, "Asymptotic behavior of solutions to time-dependent Hartree equations", *Nonlinear Anal.* **12**:3 (1988), 313–319. MR 89d:35021 Zbl 0647.35013

[Hayashi and Naumkin 1998] N. Hayashi and P. I. Naumkin, "Asymptotics for large time of solutions to the nonlinear Schrödinger and Hartree equations", *Amer. J. Math.* **120**:2 (1998), 369–389. MR 99f:35190 Zbl 0917.35128

[Hayashi and Ozawa 1987] N. Hayashi and T. Ozawa, "Time decay of solutions to the Cauchy problem for time-dependent Schrödinger–Hartree equations", *Comm. Math. Phys.* **110**:3 (1987), 467–478. MR 88e:35161 Zbl 0648.35078

[Hayashi and Ozawa 1989] N. Hayashi and T. Ozawa, "Smoothing effect for some Schrödinger equations", *J. Funct. Anal.* **85**:2 (1989), 307–348. MR 90j:35172 Zbl 0681.35079

[Hayashi et al. 1998] N. Hayashi, P. I. Naumkin, and T. Ozawa, "Scattering theory for the Hartree equation", *SIAM J. Math. Anal.* **29**:5 (1998), 1256–1267. MR 99c:35222 Zbl 0916.35110

[Hunziker and Sigal 2000] W. Hunziker and I. M. Sigal, "Time-dependent scattering theory of N–body quantum systems", *Rev. Math. Phys.* **12**:8 (2000), 1033–1084. MR 2001k:81347 Zbl 0978.47008

[Killip and Visan 2008] R. Killip and M. Visan, "Nonlinear Schrödinger equations at critical regularity", Lecture notes, Clay Mathematics Institute, 2008, Available at http://www.math.ucla.edu/~visan/ClayLectureNotes.pdf.

[Lewin 2011] M. Lewin, "Geometric methods for nonlinear many-body quantum systems", *J. Funct. Anal.* **260**:12 (2011), 3535–3595. MR 2012d:81101 Zbl 1216.81180

[Lieb 1981] E. H. Lieb, "Thomas–Fermi and related theories of atoms and molecules", *Rev. Modern Phys.* **53**:4 (1981), 603–641. MR 83a:81080a Zbl 1049.81679

[Lieb 1984] E. H. Lieb, "Bound on the maximum negative ionization of atoms and molecules", *Phys. Rev. A* **29**:6 (1984), 3018–3028.

[Lieb and Simon 1977] E. H. Lieb and B. Simon, "The Hartree–Fock theory for Coulomb systems", *Comm. Math. Phys.* **53**:3 (1977), 185–194. MR 56 #10566

[Lieb et al. 1988] E. H. Lieb, I. M. Sigal, B. Simon, and W. Thirring, "Approximate neutrality of large–Z ions", *Comm. Math. Phys.* **116**:4 (1988), 635–644. MR 89h:81269

[Lions 1981] P.-L. Lions, "Some remarks on Hartree equation", *Nonlinear Anal.* **5**:11 (1981), 1245–1256. MR 83h:58025 Zbl 0472.35074

[Lions 1987] P.-L. Lions, "Solutions of Hartree–Fock equations for Coulomb systems", *Comm. Math. Phys.* **109**:1 (1987), 33–97. MR 88e:35170 Zbl 0618.35111

[López and Soler 2000] J. L. López and J. Soler, "Asymptotic behavior to the 3–D Schrödinger/Hartree–Poisson and Wigner–Poisson systems", *Math. Models Methods Appl. Sci.* **10**:6 (2000), 923–943. MR 2002b:81076 Zbl 1012.81032

[Morawetz 1968] C. S. Morawetz, "Time decay for the nonlinear Klein–Gordon equations", *Proc. Roy. Soc. Ser. A* **306** (1968), 291–296. MR 38 #2455 Zbl 0157.41502

[Nakanishi 2002] K. Nakanishi, "Modified wave operators for the Hartree equation with data, image and convergence in the same space", *Commun. Pure Appl. Anal.* **1**:2 (2002), 237–252. MR 2003j:35301 Zbl 1013.35056

[Nam 2012] P. T. Nam, "New bounds on the maximum ionization of atoms", *Comm. Math. Phys.* **312**:2 (2012), 427–445. MR 2917171 Zbl 1259.81085

[Reed and Simon 1978] M. Reed and B. Simon, *Methods of modern mathematical physics, IV: Analysis of operators*, Academic Press, New York, 1978. MR 58 #12429c Zbl 0401.47001

[Reed and Simon 1979] M. Reed and B. Simon, *Methods of modern mathematical physics, III: Scattering theory*, Academic Press, New York, 1979. MR 80m:81085 Zbl 0405.47007

[Ruelle 1969] D. Ruelle, "A remark on bound states in potential-scattering theory", *Nuovo Cimento A* (10) **61** (1969), 655–662. MR 39 #7907

[Ruskai 1981] M. B. Ruskai, "Absence of discrete spectrum in highly negative ions", *Comm. Math. Phys.* **82**:4 (1981), 457–469. MR 83a:81087

[Ruskai 1982] M. B. Ruskai, "Absence of discrete spectrum in highly negative ions, II: Extension to fermions", *Comm. Math. Phys.* **85**:2 (1982), 325–327. MR 84b:81124

[Sánchez and Soler 2004] Ó. Sánchez and J. Soler, "Asymptotic decay estimates for the repulsive Schrödinger–Poisson system", *Math. Methods Appl. Sci.* **27**:4 (2004), 371–380. MR 2004m:35219 Zbl 1042.35076

[Schlein 2008] B. Schlein, "Derivation of effective evolution equations from microscopic quantum dynamics", Lecture notes, ETH Zurich, 2008. arXiv 0807.4307

[Sigal 1982] I. M. Sigal, "Geometric methods in the quantum many-body problem: nonexistence of very negative ions", *Comm. Math. Phys.* **85**:2 (1982), 309–324. MR 83m:81117 Zbl 0503.47041

[Sigal 1984] I. M. Sigal, "How many electrons can a nucleus bind?", *Ann. Physics* **157**:2 (1984), 307–320. MR 86a:81097

[Sigal and Soffer 1994] I. M. Sigal and A. Soffer, "Asymptotic completeness of N–particle long-range scattering", *J. Amer. Math. Soc.* **7**:2 (1994), 307–334. MR 94k:81327 Zbl 0811.35091

[Slater 1930] J. C. Slater, "Note on Hartree's method", *Phys. Rev.* **35**:2 (1930), 210–211.

[Solovej 1990] J. P. Solovej, "Asymptotics for bosonic atoms", *Lett. Math. Phys.* **20**:2 (1990), 165–172. MR 91m:81328 Zbl 0712.35075

[Solovej 1991] J. P. Solovej, "Proof of the ionization conjecture in a reduced Hartree–Fock model", *Invent. Math.* **104**:2 (1991), 291–311. MR 92f:81238 Zbl 0732.35066

[Solovej 2003] J. P. Solovej, "The ionization conjecture in Hartree–Fock theory", *Ann. of Math.* (2) **158**:2 (2003), 509–576. MR 2005g:81358 Zbl 1106.81081

[Stuart 1973] C. A. Stuart, "Existence theory for the Hartree equation", *Arch. Rational Mech. Anal.* **51** (1973), 60–69. MR 49 #11998 Zbl 0287.34032

[Tao 2007] T. Tao, "A (concentration-)compact attractor for high-dimensional non-linear Schrödinger equations", *Dyn. Partial Differ. Equ.* **4**:1 (2007), 1–53. MR 2007k:35479 Zbl 1142.35088

[Tao 2008] T. Tao, "A global compact attractor for high-dimensional defocusing non-linear Schrödinger equations with potential", *Dyn. Partial Differ. Equ.* **5**:2 (2008), 101–116. MR 2009h:35413 Zbl 1156.35095

[Wada 2001] T. Wada, “Asymptotic expansion of the solution to the nonlinear Schrödinger equation with nonlocal interaction”, *J. Funct. Anal.* **180**:1 (2001), 11–30. MR 2002f:35201 Zbl 0998.35008

[Wolkowisky 1972/73] J. H. Wolkowisky, “Existence of solutions of the Hartree equations for N electrons: an application of the Schauder–Tychonoff theorem”, *Indiana Univ. Math. J.* **22** (1972/73), 551–568. MR 47 #2300 Zbl 0237.34006

Received 20 Nov 2012. Accepted 29 Apr 2013.

ENNO LENZMANN: `enno.lenzmann@unibas.ch`
Mathematisches Institut, Universität Basel, Rheinsprung 21, CH-4051 Basel, Switzerland

MATHIEU LEWIN: `Mathieu.Lewin@math.cnrs.fr`
CNRS, Laboratoire de Mathématiques, Université de Cergy-Pontoise, 95000 Cergy-Pontoise, France

ANALYSIS AND PDE
Vol. 6, No. 5, 2013
dx.doi.org/10.2140/apde.2013.6.1213

NODAL COUNT OF GRAPH EIGENFUNCTIONS VIA MAGNETIC PERTURBATION

GREGORY BERKOLAIKO

We establish a connection between the stability of an eigenvalue under a magnetic perturbation and the number of zeros of the corresponding eigenfunction. Namely, we consider an eigenfunction of discrete Laplacian on a graph and count the number of edges where the eigenfunction changes sign (has a “zero”). It is known that the n-th eigenfunction has $n-1+s$ such zeros, where the “nodal surplus” s is an integer between 0 and the first Betti number of the graph.

We then perturb the Laplacian with a weak magnetic field and view the n-th eigenvalue as a function of the perturbation. It is shown that this function has a critical point at the zero field and that the Morse index of the critical point is equal to the nodal surplus s of the n-th eigenfunction of the unperturbed graph.

1. Introduction

Studying zeros of eigenfunctions is a question with rich history. While experimental observations have been mentioned by Leonardo da Vinci [MacCurdy 1938], Galileo [1638] and Hooke [Birch 1756], and greatly systematized by Chladni [1787], the first mathematical result is probably due to Sturm [1836]. The Oscillation Theorem of Sturm states that the number of internal zeros of the n-th eigenfunction of a Sturm–Liouville operator on an interval is equal to $n-1$. Equivalently, the zeros of the n-th eigenfunction divide the interval into n parts. In higher dimensions, the latter equality becomes a one-sided inequality: Courant [1923] (see also [Courant and Hilbert 1953]) proved that the zero curves (surfaces) of the n-th eigenfunction of the Laplacian divide the domain into at most n parts (called the “nodal domains”).

Recently, there has been a resurgence of interest in counting the nodal domains of eigenfunctions, with many exciting conjectures and rigorous results. The nodal count seems to have universal features [Blum et al. 2002; Bogomolny and Schmit 2002; Nazarov and Sodin 2009], is conjectured to resolve isospectrality [Gnutzmann et al. 2006], and has connections to minimal partitions of the domain [Helffer et al. 2009; Berkolaiko et al. 2012a], to name but a few. For a selection of research articles and historical reviews, see [Smilansky and Stöckmann 2007].

On graphs, the question can be formulated regarding the signs of the eigenfunctions of the operator

$$H : \mathbb{R}^{|V|} \to \mathbb{R}^{|V|}, \qquad H = Q - C, \tag{1}$$

where V is the set of the vertices of the graph, Q is an arbitrary real diagonal matrix, and C is the adjacency matrix of the graph. The operator H is a discrete analogue of the Schrödinger operator with

---

The author is supported by the NSF grant DMS-0907968.
MSC2010: 05C50, 58J50, 81Q10, 81Q35.
Keywords: discrete Laplace operator, nodal count, discrete magnetic Schrödinger operator.

electric potential. In this discrete setting, by a “zero” we understand an edge on which the eigenfunction changes sign, and not the exceptional (with respect to perturbation of Q) situation of an eigenfunction having a zero entry.

The subject of sign changes and nodal domains (connected components of the graph left after cutting the above edges) was addressed by, among others, Fiedler [1975], who showed the analogue of Sturm equality for *tree graphs* (see also [Bıyıkoğlu 2003]); Davies, Gladwell, Leydold and Stadler [Davies et al. 2001], who proved an analogue of the Courant (upper) bound for the number of nodal domains; Berkolaiko [2008], who proved a lower bound for graphs with cycles; and Oren [2007], who found a bound for the nodal domains in terms of the chromatic number of the graph. A number of predictions regarding the nodal count in regular graphs (assuming an adaptation of the random wave model) is put forward in [Elon 2008]. For more information, the interested reader is referred to [Bıyıkoğlu et al. 2007; Band et al. 2008].

The study of the magnetic Schrödinger operator on graphs has a similarly rich history. To give a sample, Harper [1955] used the tight-binding model (discrete Laplacian) to describe the effect of the magnetic field on conduction (see also [Hofstadter 1976]). In mathematical literature, the discrete magnetic Schrödinger operator was introduced by Lieb and Loss [1993] and Sunada [1993; 1994], and studied in [Shubin 1994; Colin de Verdière 1998; Colin de Verdière et al. 2011], among other sources (see also [Sunada 2008] for a review).

In this paper, we present a surprising connection between the two topics, namely, the number of sign changes of the n-th eigenfunction and the behavior of the eigenvalue λ_n under the perturbation of the operator H by a magnetic field. To make a precise statement, we need to introduce some notation.

The eigenvalues of the operator H on a connected graph are ordered in increasing fashion,

$$\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_{|V|}.$$

We will only consider the eigenvalue–eigenfunction pairs $(\lambda_n, f^{(n)})$ such that the eigenvalue is simple and the eigenfunction is nonzero at the vertices of the graph. This situation is generic with respect to perturbations of the potential Q, and thus we will refer to the members of such a pair as a *generic eigenvalue* and a *generic eigenfunction* correspondingly. We denote by ϕ_n the number of *sign changes* (also called *sign flips*, hence the notation ϕ) which are defined as the edges of the graph at whose endpoints the eigenfunction $f^{(n)}$ has different signs. The combined results of [Fiedler 1975; Berkolaiko 2008; Berkolaiko et al. 2012b] bound the number ϕ_n by

$$n - 1 \leq \phi_n \leq n - 1 + \beta, \tag{2}$$

where $\beta := |E| - |V| + 1$ is the first Betti number (the number of independent cycles) of the graph. Here and throughout the manuscript, we assume that the graph is connected. We will call the quantity

$$\sigma_n = \phi_n - (n - 1), \qquad 0 \leq \sigma_n \leq \beta \tag{3}$$

the *nodal surplus*. This is the extra number of sign changes that an eigenfunction has due to the graph’s nontrivial topology.

A magnetic field on discrete graphs has been introduced in, among other sources, [Lieb and Loss 1993; Sunada 1994; Colin de Verdière 1998]. Up to unitary equivalence, it can be specified using β phases $\vec{\alpha} = (\alpha_j)_{j=1}^{\beta} \in (-\pi, \pi]^{\beta}$ that will be described in Section 2. We consider the eigenvalues of the graph as functions of the parameters $\vec{\alpha}$. The zero phases, $\vec{\alpha} = 0$, correspond to the graph Γ without the magnetic field. We are now ready to formulate our main result, which connects the behavior of the eigenvalue $\lambda_n(\vec{\alpha})$ as a function of the magnetic phases to the number of zeros of the eigenfunction at $\vec{\alpha} = 0$.

Theorem 1.1. *The point $\vec{\alpha} = 0$ is the critical point of the function $\lambda_n(\vec{\alpha})$. If $\lambda_n(0)$, the n-th eigenvalue of the nonmagnetic operator, is generic, then this critical point is nondegenerate and its Morse index — the number of negative eigenvalues of the Hessian — is equal to the nodal surplus σ_n of the eigenfunction $f^{(n)}$ of the nonmagnetic operator.*

An immediate consequence of this theorem is the following.

Corollary 1.2. *The generic n-th eigenvalue of the discrete Schrödinger operator is stable with respect to magnetic perturbation of the operator if and only if the corresponding eigenfunction has exactly $n - 1$ sign changes.* (By "stability" we mean that the eigenvalue has a local minimum at zero magnetic field.)

Other possible consequences of our result and links to several other questions are discussed in Section 6. The rest of the paper is structured as follows. In Section 2 we provide detailed definitions. Section 3 is devoted to a duality between the magnetic perturbation and a certain perturbation to the potential, coupled with removal of edges. This leads to an alternative proof of the result in the case $\beta = 1$ (Subsection 3.3), which, although unnecessary for the general proof, provides us with some important insights. Section 4 collects the tools necessary for the proof of Theorem 1.1, while Section 5 contains the proof itself, which is done by extending the magnetic phases into the complex plane and relating the purely imaginary phases to the edge-removal perturbation.

2. The magnetic Hamiltonian on discrete graphs

Let $\Gamma = (V, E)$ be a simple finite connected graph with vertex set V and edge set E. We define the Schrödinger operator with potential $q : V \to \mathbb{R}$ by

$$H : \mathbb{R}^{|V|} \to \mathbb{R}^{|V|}, \qquad (H\psi)_u = -\sum_{v \sim u} \psi_v + q_u \psi_u, \tag{4}$$

that is, the matrix H is

$$H = Q - C, \tag{5}$$

where Q is the diagonal matrix of site potentials q_u and C is the adjacency matrix of the graph. It is perhaps more usual (and physically motivated) to represent the Hamiltonian as $H = Q + L$, where the Laplacian L is given by $L = D - C$ with D being the diagonal matrix of vertex degrees. But since we will not be imposing any restrictions on the potential Q, we absorb the matrix D into Q.

The operator H has $|V|$ eigenvalues, which we number in increasing order:

$$\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_{|V|}.$$

We define the magnetic Hamiltonian (magnetic Schrödinger operator) on discrete graphs as

$$(H\psi)_u = -\sum_{v\sim u} e^{iA_{v,u}}\psi_v + q_u\psi_u, \tag{6}$$

with the convention that $A_{v,u} = -A_{u,v}$, which makes H self-adjoint. For further details, the reader should consult [Lieb and Loss 1993; Sunada 1994; Colin de Verdière 1998; Colin de Verdière et al. 2011].

A sequence of directed edges $C = [u_1, u_2, \ldots, u_n]$ is called a cycle if the terminus of edge u_j coincides with the origin of the edge u_{j+1} for all j (u_{n+1} is understood as u_1). The flux through the cycle C is defined as

$$\Phi_C = \bigl(A_{u_1,u_2} + \cdots + A_{u_{n-1},u_n} + A_{u_n,u_1}\bigr) \mod 2\pi. \tag{7}$$

Two operators which have the same flux through every cycle C are unitarily equivalent (by a gauge transformation). Therefore, the effect of the magnetic field on the spectrum is fully determined by β fluxes through a chosen set of basis cycles of the cycle space. We denote them by $\alpha_1, \ldots, \alpha_\beta$ and consider the n-th eigenvalue of the graph as a function of $\vec{\alpha}$.

More precisely, fix an arbitrary spanning tree of the graph and let S be the set of edges that do not belong to the chosen tree. Obviously, S contains exactly β edges.

Lemma 2.1. *Any magnetic Schrödinger operator on the graph Γ is unitarily equivalent to one of the operators of the type*

$$H_{u,v} = \begin{cases} q_u, & u = v, \\ -1, & (u,v) \in \mathscr{E} \setminus S, \\ -e^{\pm i\alpha_s}, & (u,v) = s \in S, \end{cases} \tag{8}$$

where the sign in the exponent is plus if $u < v$ and minus if $u > v$.

Example 2.2. Consider the triangle graph — a graph with three vertices and three edges connecting them. One of the equivalent forms of the magnetic Hamiltonian for this graph is

$$H(\Gamma^{\alpha}_{\text{mag}}) = \begin{pmatrix} q_1 & -e^{i\alpha} & -1 \\ -e^{-i\alpha} & q_2 & -1 \\ -1 & -1 & q_3 \end{pmatrix}.$$

The spectrum of $H(\Gamma^{\alpha}_{\text{mag}})$ as a function of $\alpha \in (-\pi, \pi]$ is shown in Figure 1. The eigenfunctions of $H(\Gamma) = H(\Gamma^{\alpha=0}_{\text{mag}})$ have $\phi_1 = 0$, $\phi_2 = 2$ and $\phi_3 = 2$ sign changes correspondingly (these are the only choices consistent with (2) and the topology of the graph). The nodal surpluses are $\sigma_1 = 0$, $\sigma_2 = 1$ and $\sigma_3 = 0$, which agrees with $\alpha = 0$ being the point of minimum, maximum and minimum of the corresponding curves.

3. A duality between a magnetic phase and a cut

In this section, we explore a simple result which shows a connection between two types of perturbations of the operator H that will be used to prove the main theorem. It illustrates the duality between the perturbation of a discrete Schrödinger operator by a magnetic phase on a cycle and the operation of

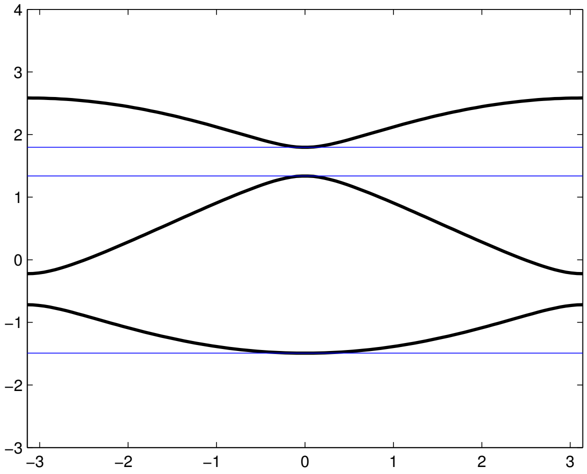

Figure 1. The eigenvalues of the triangle graph as functions of a magnetic phase α (bold curves) and the eigenvalues of the unperturbed graph (horizontal lines).

removing ("cutting") an edge that lies on the cycle. The latter operation was used to prove the lower bound on the number of nodal domains in [Berkolaiko 2008] and to study partitions on discrete graphs in [Berkolaiko et al. 2012b].

Tools used. The result of this section (Theorem 3.3 below) is based on the following version of Weyl's inequality of linear algebra that can be obtained using the variational characterization of the eigenvalues (see [Horn and Johnson 1985, Chapter 4] for similar results).

Theorem 3.1. *Let A be a self-adjoint matrix and B be a rank-one positive semidefinite self-adjoint matrix. Then*

$$\lambda_n(A-B) \le \lambda_n(A) \le \lambda_{n+1}(A-B), \tag{9}$$

where λ_n is the n-th eigenvalue, numbered in increasing order, of the corresponding matrix. Moreover, the inequalities are strict if and only if $\lambda_n(A)$ is simple and its eigenvector is not in the null-space of B.

Similarly, when B is negative definite, we have

$$\lambda_{n-1}(A-B) \le \lambda_n(A) \le \lambda_n(A-B), \tag{10}$$

with an analogous condition for strict inequalities.

Another useful result is the first term in the perturbation expansion of a parameter-dependent eigenvalue. Let $A(x)$ be a Hermitian matrix-valued analytic function of x. Let $\lambda(x)$ be an eigenvalue of the matrix A that is simple in a neighborhood of a point x_0. We know from standard perturbation theory [Kato 1976] that $\lambda(x)$ is an analytic function. Denote by $u(x)$ the normalized eigenvector corresponding to the eigenvalue λ. Then we have the following formula for the derivative of λ evaluated at the point $x = x_0$:

$$\frac{\partial}{\partial x}\lambda = \left\langle u, \frac{\partial A}{\partial x} u \right\rangle. \tag{11}$$

Two operations on a graph. Let λ_n be a simple eigenvalue and let the corresponding eigenfunction f be nonzero on vertices. Let (u_1, u_2) be an edge that belongs to one of the cycles of the graph. We allow the graph to have magnetic phases on some edges, but assume that there is no phase on the edge (u_1, u_2). Then the operator $H = Q - C$ has the following subblock corresponding to vertices u_1 and u_2:

$$H(\Gamma)_{[u_1,u_2]} = \begin{pmatrix} q_{u_1} & -1 \\ -1 & q_{u_2} \end{pmatrix}. \tag{12}$$

We consider two modifications of the original graph. The first modification of the graph is a cut: we remove the edge (u_1, u_2) and change the potential at sites u_1 and u_2. Namely, we change the $[u_1, u_2]$ subblock to

$$H(\Gamma_\gamma^{\mathrm{cut}})_{[u_1,u_2]} = \begin{pmatrix} q_{u_1} - \gamma & 0 \\ 0 & q_{u_2} - 1/\gamma \end{pmatrix}, \tag{13}$$

and leave the rest of the matrix H intact. We denote this modification by $H(\Gamma_\gamma^{\mathrm{cut}})$. Note that this modification is a rank-one perturbation of the original operator $H(\Gamma)$. Namely, $H(\Gamma_\gamma^{\mathrm{cut}}) = H(\Gamma) - B^c$, where the matrix B^c has the $[u_1, u_2]$ subblock

$$B^c_{[u_1,u_2]} = \begin{pmatrix} \gamma & -1 \\ -1 & 1/\gamma \end{pmatrix}, \tag{14}$$

and the rest of the elements are zero. Then B^c is positive definite if $\gamma > 0$ and negative definite if $\gamma < 0$. Note that the cases $\gamma = \infty$ and $\gamma = 0$ can also be given the meaning of removing (or imposing the Dirichlet condition at) the vertex u_1 or the vertex u_2 correspondingly. However, we will not dwell on this issue, and exclude these cases from our consideration.

Notably, if f is an eigenfunction of $H(\Gamma)$ and $\gamma = f_{u_2}/f_{u_1} \in \mathbb{R}$, where f_u is the value of f at the vertex u, then f is also an eigenfunction of $H(\Gamma_\gamma^{\mathrm{cut}})$. Equivalently, f is in the null-space of the perturbation B^c.

The second modification of the original graph is the introduction of a magnetic phase on the edge (u_1, u_2). The $[u_1, u_2]$ subblock of the new operator $H(\Gamma_{\mathrm{mag}}^\alpha)$ is

$$H(\Gamma_{\mathrm{mag}}^\alpha)_{[u_1,u_2]} = \begin{pmatrix} q_{u_1} & -e^{i\alpha} \\ -e^{-i\alpha} & q_{u_2} \end{pmatrix}, \tag{15}$$

while other entries coincide with those of $H(\Gamma)$. Note that $H(\Gamma_{\mathrm{mag}}^\alpha)$ is *not* a rank-one perturbation of $H(\Gamma)$. However, it is a rank-one perturbation of the cut graph $H(\Gamma_\gamma^{\mathrm{cut}})$ for any values of α and γ. Namely, $H(\Gamma_\gamma^{\mathrm{cut}}) = H(\Gamma_{\mathrm{mag}}^\alpha) - B^{mc}$, where

$$B^{mc}_{[u_1,u_2]} = \begin{pmatrix} \gamma & -e^{i\alpha} \\ -e^{-i\alpha} & 1/\gamma \end{pmatrix}, \tag{16}$$

and all other entries of B^{mc} are zero. Also, the spectra of $H(\Gamma_{\mathrm{mag}}^\alpha)$ and $H(\Gamma)$ coincide when $\alpha = 0$ since the operators coincide.

A duality between the two operations. We now want to apply Theorem 3.1 to the spectra of Γ, $\Gamma_\gamma^{\text{cut}}$ and $\Gamma_{\text{mag}}^\alpha$. However, we must take care to distinguish the two cases that correspond to equations (9) and (10) ($\gamma > 0$ and $\gamma < 0$ correspondingly).

Definition 3.2. The eigenvalues of Γ, $\Gamma_\gamma^{\text{cut}}$ and $\Gamma_{\text{mag}}^\alpha$ will be numbered in increasing order starting from 1. We will also use the convention

$$\lambda_j(\Gamma) = \begin{cases} -\infty, & j < 1, \\ \infty, & j > n, \end{cases}$$

for the cases when the index of λ happens to be out of bounds.

Theorem 3.3. *Let* $p(\gamma)$ *be* 1 *if* $\gamma < 0$ *and* 0 *otherwise. Then the following inequalities hold*:

$$\lambda_{n-p(\gamma)}(\Gamma_\gamma^{\text{cut}}) \le \lambda_n(\Gamma_{\text{mag}}^\alpha) \le \lambda_{n-p(\gamma)+1}(\Gamma_\gamma^{\text{cut}}), \tag{17}$$

for all values of α *and* γ. *Furthermore*, *for any fixed* n,

$$\max_\gamma \lambda_{n-p(\gamma)}(\Gamma_\gamma^{\text{cut}}) = \min_\alpha \lambda_n(\Gamma_{\text{mag}}^\alpha) =: M_1 \tag{18}$$

and

$$\max_\alpha \lambda_n(\Gamma_{\text{mag}}^\alpha) = \min_\gamma \lambda_{n-p(\gamma)+1}(\Gamma_\gamma^{\text{cut}}) =: M_2. \tag{19}$$

Finally, *if there are no magnetic phases on the graph* Γ (*that is*, *all entries of* $H(\Gamma)$ *are real*), *then one of the extremal values* M_1 *or* M_2 *is equal to* $\lambda_n(\Gamma) = \lambda_n(\Gamma_{\text{mag}}^{\alpha=0})$, *while the other is equal to* $\lambda_n(\hat{\Gamma}) := \lambda_n(\Gamma_{\text{mag}}^{\alpha=\pi})$.

Remark 3.4. Note that at this point we don't know which extremum, M_1 or M_2, is equal to $\lambda_n(\Gamma)$. In other words, $\alpha = 0$ may be either a maximum or a minimum of $\lambda_n(\Gamma_{\text{mag}}^\alpha)$; see Figure 1. This information is related to the nodal surplus. The point $\alpha = \pi$ will then be a minimum or a maximum, correspondingly. Also, if the graph Γ had some magnetic phases on it before we added a phase α on the edge (u_1, u_2), the extrema with respect to α do not have to occur at 0 and π.

Note that we have also defined yet another modification of the graph Γ, the graph $\hat{\Gamma}$ whose adjacency matrix has -1 in place of 1 for the entries C_{u_1,u_2} and C_{u_1,u_2}.

Remark 3.5. Let $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ be the extended real line and $\widehat{\mathbb{R}} = \overline{\mathbb{R}}/[-\infty = \infty]$ be its projective ("wrapped") version. The eigenvalue $\lambda_{n-p(\gamma)}(\Gamma_\gamma^{\text{cut}})$ is then a continuous function of γ, considered as a function from $\widehat{\mathbb{R}}$ to $\overline{\mathbb{R}}$; see Figure 2 for an example. Note that by our definitions, $\lambda_{n-p(\gamma)}(\Gamma_\gamma^{\text{cut}}) = -\infty$ for $n = 1$ and $\gamma < 0$.

Proof of Theorem 3.3. The inequalities follow directly from Theorem 3.1, since for any α, the graph $\Gamma_{\text{mag}}^\alpha$ is a rank-one perturbation of $\Gamma_\gamma^{\text{cut}}$. Whether it is positive or negative definite depends on the sign of γ, and results in the shift by p.

We get the properties of the extrema as follows. Observe that if $\max \lambda_{n-p}(\Gamma_\gamma^{\text{cut}}) = \min \lambda_{n-p+1}(\Gamma_\gamma^{\text{cut}})$, then $\lambda_n(\Gamma_{\text{mag}}^\alpha)$ is constant and equal to the common value of $\lambda_{n-p}(\Gamma_\gamma^{\text{cut}})$ and $\lambda_{n-p+1}(\Gamma_\gamma^{\text{cut}})$.

Let now $\max \lambda_{n-p}(\Gamma_\gamma^{\text{cut}}) < \min \lambda_{n-p+1}(\Gamma_\gamma^{\text{cut}})$. The eigenvalues of a one-parameter family can always be represented as a set of analytic functions (that can intersect). Let $\lambda'(\Gamma_\gamma^{\text{cut}})$ be the analytic function that

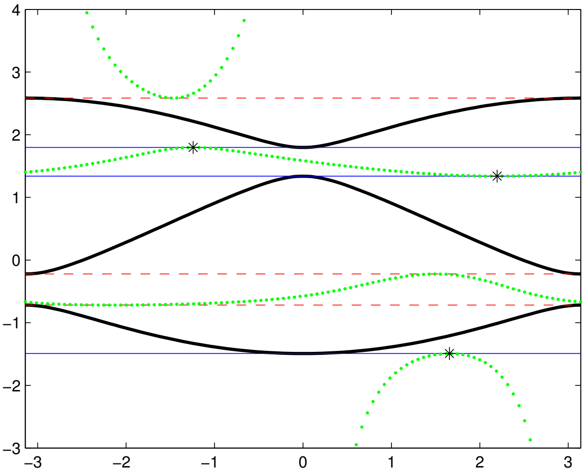

Figure 2. The duality between a magnetic field on one side and cut edge with added potential on the other. The graph is a triangle. The bold curves correspond to the eigenvalues as functions of the magnetic phase. The dotted curves correspond to varying the potential parameter γ after cutting the edge. The x-axis ranges from $-\pi/2$ to $\pi/2$ with the magnetic phase taken as $\alpha = 2x$ and the potential parameter $\gamma = \tan(x)$. The horizontal solid lines are the eigenvalues of the original graph, while the horizontal dashed lines are the eigenvalues of the graph with the magnetic phase π.

achieves the maximum $\max \lambda_{n-p}(\Gamma_\gamma^{\mathrm{cut}})$ and f be the corresponding eigenfunction. We will differentiate $\lambda'(\Gamma_\gamma^{\mathrm{cut}})$ using (11). At the maximum point $\gamma = \gamma^\circ$, we have, by (13),

$$0 = \frac{d\lambda'}{d\gamma} = \left\langle f, \frac{dB^c}{d\gamma} f \right\rangle = -|f_{u_1}|^2 + \frac{|f_{u_2}|^2}{(\gamma^\circ)^2}. \tag{20}$$

From here it follows that

$$\gamma^\circ = \pm \frac{|f_{u_2}|}{|f_{u_1}|} \qquad \text{or, equivalently,} \qquad \left| \frac{\gamma^\circ f_{u_1}}{f_{u_2}} \right| = 1. \tag{21}$$

Let $\widetilde{\alpha}$ be the solution of $e^{i\alpha} = \gamma^\circ f_{u_1}/f_{u_2}$. Direct calculation shows that the eigenfunction f is in the null-space of the perturbation B^{mc} of (16) with $\alpha = \widetilde{\alpha}$, and therefore f is both in the spectrum of $\Gamma_{\gamma^\circ}^{\mathrm{cut}}$ and in the spectrum of $\Gamma_{\mathrm{mag}}^{\widetilde{\alpha}}$, so (18) follows. The proof of (19) is completely analogous.

Note that we could instead differentiate the eigenvalue of $\Gamma_{\mathrm{mag}}^{\alpha}$, leading to the condition

$$f_{u_2} \overline{f_{u_1}} e^{i\widetilde{\alpha}} \in \mathbb{R}, \tag{22}$$

instead of (20). One then sets $\gamma^\circ = e^{i\widetilde{\alpha}} f_{u_2}/f_{u_1} \in \mathbb{R}$, to the same effect.

Finally, when the matrix $H(\Gamma)$ is real, the eigenfunctions of $\Gamma_\gamma^{\mathrm{cut}}$, $\Gamma_{\mathrm{mag}}^{\alpha=0}$ and $\Gamma_{\mathrm{mag}}^{\alpha=\pi}$ are real-valued. When $\alpha = 0$, we can verify directly that the eigenfunction f of $\Gamma_{\mathrm{mag}}^{\alpha=0}$ is also an eigenfunction of $\Gamma_{\gamma^\circ}^{\mathrm{cut}}$ by setting $\gamma^\circ = f_{u_2}/f_{u_1}$. When $\alpha = \pi$, we also set $\gamma = -f_{u_2}/f_{u_1}$ and do the same. ☐

Theorem 3.3 highlights a sort of duality between the two modifications of the graph Γ. The spectra of the graphs with a magnetic phase form bands (as the phase is varied), while the spectra of the graphs with the cut fill the gaps between these bands. Minima of one correspond to maxima of the other, and in half of the cases correspond to eigenvalues of the original graph.

We now explain how the $\beta = 1$ case of Theorem 1.1 follows from Theorem 3.3. While for general β, the proof is significantly different (it bypasses the interlacing inequalities and goes straight to the quadratic form), some key features are the same as in this simple case.

Starting with the eigenvalue λ_n of Γ and the corresponding eigenfunction f, we cut an edge on the only cycle of Γ to obtain a family of trees $\Gamma_\gamma^{\mathrm{cut}}$. For $\gamma = \gamma^\circ := f_{u_2}/f_{u_1}$, we have either

$$\max_\gamma \lambda_{n-p(\gamma)}(\Gamma_\gamma^{\mathrm{cut}}) = \lambda_{n-p(\gamma^\circ)}(\Gamma_{\gamma^\circ}^{\mathrm{cut}}) = \lambda_n(\Gamma) = \min_\alpha \lambda_n(\Gamma_{\mathrm{mag}}^\alpha)$$

or

$$\max_\alpha \lambda_n(\Gamma_{\mathrm{mag}}^\alpha) = \lambda_n(\Gamma) = \lambda_{n-p(\gamma^\circ)+1}(\Gamma_{\gamma^\circ}^{\mathrm{cut}}) = \min_\gamma \lambda_{n-p(\gamma)+1}(\Gamma_\gamma^{\mathrm{cut}}).$$

In the first case, according to Fiedler's theorem (Equation (2) with $\beta = 0$), the function f has $n - p(\gamma^\circ) - 1$ sign changes *with respect to* the tree $\Gamma_\gamma^{\mathrm{cut}}$. Adding back the removed edge (u_1, u_2) adds another sign change if $\gamma^\circ < 0$, and doesn't change the number of sign changes otherwise. In other words, it adds $p(\gamma^\circ)$ sign changes. Thus, with respect to Γ, the function f has $n - 1$ sign changes and $\sigma_n = 0$. In the second case, we similarly conclude that f has $n - p(\gamma^\circ)$ sign changes with respect to $\Gamma_\gamma^{\mathrm{cut}}$, and n sign changes with respect to Γ. The nodal surplus is $\sigma_n = 1$.

On the other hand, in the first case, $\lambda_n(\Gamma)$ is a minimum of $\lambda_n(\Gamma_{\mathrm{mag}}^\alpha)$ (Morse index 0), while in the second, it is a maximum of $\lambda_n(\Gamma_{\mathrm{mag}}^\alpha)$ (Morse index 1), which shows that the Morse index coincides with σ_n in the case $\beta = 1$.

Remark 3.6. In the $\beta = 1$ case, the spectrum of the cut graph $\Gamma_\gamma^{\mathrm{cut}}$ completely fills the gaps in the magnetic spectrum (see Theorem 3.3 and Figure 2). This is not the case for $\beta > 1$, although an interesting relationship persists, as will become apparent in Section 5.

4. Tools of the main proof

In this section, we collect some basic facts that will be repeatedly used in the proof of Theorem 1.1.

Critical points of the quadratic form.

Definition 4.1. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. If c is a critical point (that is, $\nabla F(c) = 0$), the *inertia* of c is the triple (n_-, n_0, n_+) that counts the number of negative, zero and positive eigenvalues correspondingly of the Hessian (the matrix of second derivatives) at the point c. The number n_- is called the *Morse index* (or simply *index*).

The next lemma is a reminder that the eigenvectors of a symmetric matrix are critical points of the quadratic form on the unit sphere.

Lemma 4.2. *Let A be a $d \times d$ real symmetric matrix and let $h(x) = \langle x, Ax\rangle$, $x \in \mathbb{R}^d$, be the associated quadratic form. Then the* (*real*) *eigenvectors of the matrix A are critical points of the function $h(x)$ on the unit sphere $\|x\| = 1$.*

Let λ_n be the n-th eigenvalue of A and let $f^{(n)}$ be the corresponding normalized eigenfunction. Define

$$n_- = \#\{\lambda_m < \lambda_n\}, \qquad n_0 = \#\{\lambda_m = \lambda_n,\ m \neq n\}, \qquad n_+ = \#\{\lambda_m > \lambda_n\}, \tag{23}$$

with $n_- + n_0 + n_+ = d - 1$. Then the inertia of the critical point $x = f^{(n)}$ is (n_-, n_0, n_+). In particular, if λ_n is a simple eigenvalue, the inertia is $(n-1, 0, d-n)$.

Remark 4.3. The value of the quadratic form h at the critical point $f^{(n)}$ is λ_n.

Proof. The idea is intuitively clear: n_- — which is the Morse index — counts the number of directions in which the quadratic form decreases relative to the value at $x = f^{(n)}$. These directions are the eigenvectors corresponding to the eigenvalues that are less than λ_n. Similar characterizations are valid for n_0 and n_+.

We note that by Sylvester's law of inertia, the inertia is invariant under the change of variables. Making the orthogonal change of coordinates to the eigenbasis of the matrix A, the quadratic form $h(a)$ becomes

$$h(a) = \lambda_1 a_1^2 + \lambda_2 a_2^2 + \cdots + \lambda_d a_d^2,$$

while the sphere is given by the equations

$$a_1^2 + a_2^2 + \cdots + a_d^2 = 1.$$

Thus, on the sphere, the quadratic form in terms of variables $a_1, \ldots, a_{n-1}, a_{n+1}, \ldots, a_d$ is given by

$$h = \lambda_n + \sum_{j \neq n} (\lambda_j - \lambda_n) a_j^2,$$

and the Hessian is a diagonal matrix with $\lambda_j - \lambda_n$, $j = 1, \ldots, d$, $j \neq n$. The statement of the lemma follows immediately. ☐

Reduction to the critical manifold. The tool introduced in this section is a simple idea already used in [Band et al. 2012; Berkolaiko et al. 2012a; 2012b]. If we have a function $f(x_1, \ldots, x_n)$ with a critical point c, then under some general conditions, there is an $(n-1)$-dimensional manifold around the point c on which the local minimum of f is achieved when we vary the variable x_1 and keep the others fixed. Then the Morse index of f restricted to this manifold is the same as the Morse index of the unrestricted function. On the other hand, if the manifold is the locus of local maxima with respect to the variable x_1, the Morse index on the manifold is one less than the unrestricted Morse index. The following lemma is a simple generalization of this idea. The proof is a simplified finite-dimensional adaptation of the proof in [Berkolaiko et al. 2012a].

Lemma 4.4 (reduction lemma). *Let $X = Y \oplus Y'$ be a direct decomposition of a finite-dimensional vector space. Let $f : X \to \mathbb{R}$ be a smooth functional such that* $(0, 0) \in X$ *is its critical point with inertia $\mathcal{I}_X$. Further, for every $y \in Y$ locally around* 0, *let the functional $f(y, y')$ considered as a function of y' have a* *critical point at* $y' = 0$ *with inertia* $\mathcal{I}_{Y'}$, *that* (*locally*) *does not depend on* y. *Then the Hessian of* f *is reduced by the decomposition* $X = Y \oplus Y'$, *and the inertia of* f *with respect to the space* Y *is*

$$\mathcal{I}_Y = \mathcal{I}_X - \mathcal{I}_{Y'}. \tag{24}$$

Proof. We calculate the mixed derivative of f with respect to one variable from Y and the other from Y'. In a slight abuse of notation, we denote these variables simply by y and y'. We have

$$\frac{\partial^2 f}{\partial y\, \partial y'}(0, 0) = \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial y'}(y, 0)\right]\Bigg|_{y=0} = 0, \tag{25}$$

since $y' = 0$ is the critical point of $f(y, y')$ as a function of y' for every y. Thus the Hessian of f has a block-diagonal form with two blocks that correspond to Y and Y'. The spectrum of the Hessian is the union of the spectra of the blocks and the inertia is the sum of the inertias of the blocks,

$$\mathcal{I}_X = \mathcal{I}_Y + \mathcal{I}_{Y'}.$$

Equation (24) follows immediately. □

Remark 4.5. Lemma 4.4 can be simply extended to the case when, for every fixed y, the critical point with respect to y' is located at $y' = q(y)$ (rather than $y' = 0$). The function $q(y)$ defines the critical manifold $\mathfrak{Q} = (y, q(y))$. If $q(y)$ is a smooth function of y and $q(0) = 0$, the change of variables

$$y \mapsto y, \qquad y' \mapsto y' - q(y)$$

is nondegenerate (its Jacobian is a triangular matrix with 1s on the diagonal) and makes f satisfy the assumptions of Lemma 4.4. By Sylvester's law of inertia, the conclusion of the lemma is invariant under the change of variables. Therefore, the inertia of $f\big|_{\mathfrak{Q}}$ at point 0 is

$$\mathcal{I}_{\mathfrak{Q}} = \mathcal{I}_X - \mathcal{I}_{Y'}.$$

5. Proof of the main theorem

We prove the main result in three steps. First we show by an explicit computation that the point 0 is a critical point of the function $\lambda_n(\vec{\alpha})$, where $\vec{\alpha} = (\alpha_1, \dots, \alpha_\beta) \in (-\pi, \pi]^\beta$ are the magnetic phases.

Then we fix an eigenpair $\lambda = \lambda_n(\Gamma)$ and f. We cut β edges of the graph, turning it into a tree T, but modifying the potentials so that the eigenfunction f is also an eigenfunction of the tree T. It now corresponds to an eigenvalue number m, that is, $\lambda_m(T) = \lambda$. Considering the eigenvalue $\lambda_m(T)$ as a function of the potentials, we find its inertia by two applications of the reduction lemma to the corresponding quadratic form. The result of this step is related to the results on critical equipartitions [Berkolaiko et al. 2012b].

Finally, we relate the inertia of the function $\lambda_m(T)$ to the inertia of the function $\lambda_n(\vec{\alpha})$ at the corresponding critical points. This is done by complexifying $\vec{\alpha}$ and relating the function $\lambda_n(\vec{\alpha})$ on the imaginary axis to the function $\lambda_m(T)$ by a change of variables.

We recall that S is a set of β edges whose removal turns the graph Γ into a tree. By $\Gamma^{\vec{\alpha}}_{\mathrm{mag}}$, we denote the graph obtained from Γ by introducing magnetic phases $\vec{\alpha} = (\alpha_1, \dots, \alpha_\beta)$ on the edges from the set S. Similarly, by $\Gamma^{\mathrm{cut}}_{\vec{\gamma}}$ we denote the tree graph obtained by cutting every edge from S in the manner already described (see Equation (13) and around it). For future reference, we list the quadratic forms of the original graph and the graph $\Gamma^{\mathrm{cut}}_{\vec{\gamma}}$, grouping the terms to highlight the differences between the two forms:

$$h(\vec{x}) = \sum_u q_u x_u^2 - \sum_{(u,v)\in E\setminus S} 2x_u x_v - \sum_{(u,v)\in S} 2x_u x_v, \tag{26}$$

$$h^{\mathrm{cut}}_{\vec{\gamma}}(\vec{x}) = \sum_u q_u x_u^2 - \sum_{(u,v)\in E\setminus S} 2x_u x_v - \sum_{e_j=(u,v)\in S} \left(\gamma_j x_u^2 + \frac{x_v^2}{\gamma_j}\right). \tag{27}$$

Critical points. Let f be an eigenfunction of the graph Γ. We have seen in Theorem 3.3 and its proof that the points $\alpha = 0$ and $\gamma = \gamma^\circ$ (see Equation (21)) are special: at these points, f is an eigenfunction of the graphs $\Gamma^{\alpha}_{\mathrm{mag}}$ and $\Gamma^{\mathrm{cut}}_{\gamma}$. Moreover, they are critical points of the corresponding eigenvalues considered as functions of the parameters α and γ, respectively. The result of this section generalizes this observation.

Theorem 5.1. *Let f be an eigenfunction of $H(\Gamma)$ that corresponds to a simple eigenvalue $\lambda = \lambda_n(\Gamma)$. Assume f is nonzero on vertices of the graph Γ. For every edge $(u_j, v_j) \in S$, $j = 1, \dots, \beta$, let*

$$\gamma_j^\circ = \frac{f_{v_j}}{f_{u_j}}. \tag{28}$$

Let p denote the number of negatives among the values γ_j°:

$$p = \#\{\gamma_j^\circ < 0,\ j = 1, \dots, \beta\}.$$

Then

$$\lambda_n(\Gamma) = \lambda_{\phi_n - p + 1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}^\circ}), \tag{29}$$

where ϕ_n is the number of sign changes of f with respect to the graph Γ. The eigenvalue λ_{ϕ_n-p+1} of the tree $\Gamma^{\mathrm{cut}}_{\vec{\gamma}^\circ}$ is simple. Moreover, the point $\vec{\gamma}^\circ = (\gamma_1^\circ, \dots, \gamma_\beta^\circ)$ is a critical point of the function $\lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}})$.

Similarly for $\Gamma^{\vec{\alpha}}_{\mathrm{mag}}$,

$$\lambda_n(\Gamma) = \lambda_n(\Gamma^{0,\dots,0}_{\mathrm{mag}}) \tag{30}$$

and $(0, \dots, 0)$ is a critical point of the function $\lambda_n(\Gamma^{\vec{\alpha}}_{\mathrm{mag}})$.

Proof. It can be verified directly that f is an eigenfunction of the graph $\Gamma^{\mathrm{cut}}_{\vec{\gamma}^\circ}$. The nodal bound (2) with $\beta = 0$ (proved by Fiedler [1975]; see also [Berkolaiko 2008]) shows that the eigenvalue corresponding to the function f has number $\mu' + 1$ in the spectrum of the tree $\Gamma^{\mathrm{cut}}_{\vec{\gamma}^\circ}$, where μ' is the number of sign changes of f with respect to the tree. In general, this number is different from ϕ_n because we might have cut some of the edges on which f was changing sign. However, according to (28), these edges gave rise to *negative* values of γ_j°, and therefore $\mu' = \phi_n - p$, proving (29). The eigenvalue that corresponds to a nonzero eigenvector on a tree is simple [Fiedler 1975], establishing simplicity of $\lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}^\circ})$. Equation (30) is trivial since $\Gamma^{0,\dots,0}_{\mathrm{mag}} = \Gamma$.

To prove criticality of the points, we calculate the derivatives. Because the eigenvalues in question are simple, they are analytic functions of the parameters and can be differentiated.

The derivative of $\lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}})$ with respect to γ_j has been calculated in (20), resulting in

$$\frac{\partial}{\partial \gamma_j}\lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}})\big|_{(\gamma_1^\circ,\dots,\gamma_\beta^\circ)} = -|f_{u_j}|^2 + \frac{|f_{v_j}|^2}{\gamma_j^{\circ 2}} = 0, \tag{31}$$

where we used the definition of γ_j° from (28).

The derivative of $\lambda_n(\Gamma^{\vec{\alpha}}_{\mathrm{mag}})$ can be evaluated similarly using (11), leading to

$$\frac{\partial}{\partial \alpha_j}\lambda_n(\Gamma^{\vec{\alpha}}_{\mathrm{mag}})\big|_{(0,\dots,0)} = -i\,\overline{f_{u_j}}\,f_{v_j} + i f_{u_j}\overline{f_{v_j}} = \mathrm{Im}(\overline{f_{u_j}}\,f_{v_j}) = 0, \tag{32}$$

since the eigenfunction f is real-valued. Alternatively, we can observe that $\lambda_n(\Gamma^{\vec{\alpha}}_{\mathrm{mag}})$ is invariant with respect to reflection $\alpha \mapsto -\alpha$. □

Index of the eigenvalue on the tree. In this section, we elaborate on the first part of the result of Theorem 5.1, namely that $(\gamma_1^\circ,\dots,\gamma_\beta^\circ)$ is a critical point of the function $\lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}})$.

Theorem 5.2. *Let f be an eigenfunction of $H(\Gamma)$ that corresponds to a simple eigenvalue $\lambda = \lambda_n(\Gamma)$. Assume f is nonzero on vertices of the graph Γ and has ϕ_n sign changes. For every edge $(u_j, v_j) \in S$, $j = 1,\dots,\beta$, let*

$$\gamma_j^\circ = \frac{f_v}{f_u}. \tag{33}$$

As before, p denotes the number of negatives among the values γ_j°. Then the point $(\gamma_1^\circ,\dots,\gamma_\beta^\circ)$ as a critical point of the function $\lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}})$ is nondegenerate and has inertia

$$\bigl(n-1+\beta-\phi_n,\ 0,\ \phi_n-n+1\bigr).$$

Proof. Denote by d the number of vertices of the graph Γ. Consider $h^{\mathrm{cut}}_{\vec{\gamma}}(\vec{x})$, which is the quadratic form on the Hamiltonian of $\Gamma^{\mathrm{cut}}_{\vec{\gamma}}$, as a function of $d+\beta$ real variables $(x_1,\dots,x_d,\gamma_1,\dots,\gamma_\beta)$ on the manifold $x_1^2+\cdots+x_d^2=1$. We note that the point $(f_1,\dots,f_d,\gamma_1^\circ,\dots,\gamma_\beta^\circ)$ is a critical point of $h^{\mathrm{cut}}_{\vec{\gamma}}(\vec{x})$, as can be easily shown by explicit computation. Indeed, the value of the Lagrange multiplier is the eigenvalue λ_n and the gradient of

$$F(x_1,\dots,x_d,\gamma_1,\dots\gamma_\beta) = h^{\mathrm{cut}}_{\vec{\gamma}}(\vec{x}) - \lambda_n(x_1^2+\cdots+x_d^2)$$

is zero: the first d equations become the eigenvalue condition $Hf = \lambda_n f$ and the last β are the same as (31).

We now describe the outline of the proof. Denote the inertia of the point $(f_1,\dots,f_d,\gamma_1^\circ,\dots,\gamma_\beta^\circ)$ by $\mathcal{I}$. To calculate it, we will look for critical points of $h^{\mathrm{cut}}_{\vec{\gamma}}(\vec{x})$ as a function of $\gamma_1,\dots,\gamma_\beta$. These points will define a critical manifold to which we will apply Lemma 4.4 via Remark 4.5 (this reduction corresponds to the left arrow in Figure 3). On the critical manifold, the function $h^{\mathrm{cut}}_{\vec{\gamma}}(\vec{x})$ will coincide with $h(x)$, the quadratic form of the original graph, whose inertia we know by Lemma 4.2. Having found the inertia of the critical point at the top of Figure 3, we will apply minimax with respect to variables $x_1,\dots,x_d$ to

$$\begin{tikzcd}
 & h^{\mathrm{cut}}_{\vec{\gamma}}(\vec{x}) \arrow[dl, "{\min\limits_{\gamma_j<0}\max\limits_{\gamma_j>0},\ \beta - p}"'] \arrow[d, "{\text{eigenvector minimax},\ \phi_n - p}"] \\
h(\vec{x}) & \lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}})
\end{tikzcd}$$

Figure 3. Schematic diagram of the proof of Theorem 5.2. The reductions are indicated by arrows, with the description of the parameters that are being reduced and the index of the reduction. Since we know the index of the critical point of $h(\vec{x})$, we can follow the diagram, applying the reduction lemma, to calculate the index of $\lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}})$.

follow the vertical arrow of Figure 3. This will take us to the eigenvalue $\lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}})$, and we will be able to calculate its inertia applying Lemma 4.4 again.

Consider $\vec{x}$ varying locally around the point f, so that the elements of $\vec{x}$ remain bounded away from zero. For each fixed $\vec{x}$, we look for a critical point with respect to the variables $(\gamma_1, \dots, \gamma_\beta)$. The terms of $h^{\mathrm{cut}}_{\vec{\gamma}}(\vec{x})$ that depend on a given γ have the form

$$T(\gamma) = -\gamma x_u^2 - \frac{x_v^2}{\gamma}. \tag{34}$$

The critical point is $\gamma = g(\vec{x}) = x_v/x_u$, which is a smooth function of $\vec{x}$. The points $(x_1, \dots, x_d, g_1, \dots g_\beta)$ define the critical manifold to which we apply Lemma 4.4 (via Remark 4.5). Note that the critical manifold includes the point $(f_1, \dots, f_d, \gamma_1^\circ, \dots, \gamma_\beta^\circ)$. Moreover, the critical point with respect to a given γ is a maximum if $g(\vec{x}) > 0$ and a minimum if $g(\vec{x}) < 0$. Each point is nondegenerate and, moreover, the sign of g_j is locally the same as the sign of γ_j° for all j. Different variables γ_j are not coupled, and thus the Hessian is diagonal. Therefore, the inertia of the points on the critical manifold is $(\beta - p, 0, p)$ — it is a minimum with respect to p variables and maximum with respect to $\beta - p$. We remind the reader that p is the number of negatives among $\{\gamma_j^\circ\}$.

Consider now the function $h^{\mathrm{cut}}_{\vec{\gamma}}(\vec{x})$ on the critical manifold. When $\gamma = g$, the term (34) evaluates to

$$T(g) = -2x_u x_v,$$

and we find that, on the critical manifold, the function $h^{\mathrm{cut}}_{\vec{\gamma}}(\vec{x})$ coincides with the quadratic form of the original graph, $h(\vec{x})$. The point $\vec{x} = f$, being the n-th eigenfunction of the graph, is a nondegenerate critical point of $h(\vec{x})$ and has inertia $(n-1, 0, d-n)$. Applying Lemma 4.4, we obtain

$$\mathcal{I} = \big(n - 1 + (\beta - p),\ 0,\ d - n + p\big).$$

In particular, we conclude that the point $(f_1, \dots, f_d, \gamma_1^\circ, \dots, \gamma_\beta^\circ)$ is a nondegenerate critical point.

For every value of $(\gamma_1, \dots, \gamma_\beta)$ locally around the point $(\gamma_1^\circ, \dots, \gamma_\beta^\circ)$, consider the $(\phi_n - p + 1)$-th eigenvector $f_{\vec{\gamma}}^{\text{cut}}$ of $\Gamma_{\vec{\gamma}}^{\text{cut}}$. According to Lemma 4.2, it is a nondegenerate critical point of $h_{\vec{\gamma}}^{\text{cut}}(\vec{x})$ as a function of $\vec{x}$ with inertia $(\phi_n - p, 0, d + p - \phi_n - 1)$. At the critical point, the value of the $h_{\vec{\gamma}}^{\text{cut}}$ is

$$h_{\vec{\gamma}}^{\text{cut}}(f_{\vec{\gamma}}^{\text{cut}}) = \lambda_{\phi_n - p+1}(\Gamma_{\vec{\gamma}}^{\text{cut}}),$$

which is the function whose inertia we strive to evaluate.

According to standard perturbation theory (see [Kato 1976], for example), the eigenvector $f_{\vec{\gamma}}^{\text{cut}}$ is a smooth (indeed, analytic) function of $(\gamma_1, \dots, \gamma_\beta)$. This allows us to use Lemma 4.4 again, concluding that the critical point $(\gamma_1^\circ, \dots, \gamma_\beta^\circ)$ of $h_{\vec{\gamma}}^{\text{cut}}(f_{\vec{\gamma}}^{\text{cut}})$ has inertia

$$\mathcal{I} - (\phi_n - p, 0, d + p - \phi_n - 1) = (n - 1 + \beta - \phi_n, 0, \phi_n - n + 1). \qquad \square$$

Remark 5.3. In [Berkolaiko et al. 2012b], the eigenvalue of the tree graph $\Gamma_{\vec{\gamma}}^{\text{cut}}$ was interpreted as the energy of the "partition" with the given number of domains. Theorem 5.2 gives another route for the proof of the results of that paper.

Index of the eigenvalue as a function of the magnetic field. Now we move from the critical point on the tree to the critical point of the eigenvalue of the graph with magnetic phases. We can apply the same method, retracing our steps, but now considering the quadratic forms $h_{\vec{\gamma}}^{\text{cut}}(z)$ and

$$h_{\text{mag}}^{\vec{\alpha}}(\vec{z}) = \sum_u q_u |z_u|^2 - \sum_{(u,v)\in E\setminus S} 2\,\text{Re}(\bar{z}_u z_v) - \sum_{e_j=(u,v)\in S} 2\,\text{Re}(\bar{z}_u e^{i\alpha_j} z_v) \tag{35}$$

as functions of complex variables z. Considering the complex space as a real space of double dimension leads to the inertia in the Hermitian analogue of Lemma 4.2 being $(2n_-, 2n_0 + 1, 2n_+)$. Finding extrema of $h_{\text{mag}}^{\vec{\alpha}}(z)$ with respect to $\vec{\alpha}$ and of $h_{\vec{\gamma}}^{\text{cut}}(z)$ with respect to $\vec{\gamma}$ results in the same values, and thus we relate the indices of $\lambda_n(\Gamma_{\text{mag}}^{\vec{\alpha}})$ and $\lambda_{\phi_n-p+1}(\Gamma_{\vec{\gamma}}^{\text{cut}})$ through a chain of four applications of the reduction lemma (Lemma 4.4), illustrated in Figure 4.

However, instead of following the above plan, we present a simpler yet more insightful proof which can be summarized as follows: after a change of variables, the function $\lambda_{\phi_n-p+1}(\Gamma_{\vec{\gamma}}^{\text{cut}})$ coincides with the function $\lambda_n(\Gamma_{\text{mag}}^{\vec{\alpha}})$ with purely imaginary values of the magnetic phases $\vec{\alpha}$. This will give us full understanding of the quadratic term (the Hessian) of the analytic function $\lambda_n(\Gamma_{\text{mag}}^{\vec{\alpha}})$.

Theorem 5.4. *Let f be an eigenfunction of $H(\Gamma)$ that corresponds to a simple eigenvalue $\lambda = \lambda_n(\Gamma)$. Assume f is nonzero on vertices of the graph Γ and has ϕ_n sign changes. Let $\Gamma_{\text{mag}}^{\vec{\alpha}}$ be the graph with the magnetic phases $\vec{\alpha} = (\alpha_1, \dots, \alpha_\beta)$ introduced on the edges from the set S. Then the index of $(0, \dots, 0)$ as a critical point of the function $\lambda_n(\Gamma_{\text{mag}}^{\vec{\alpha}})$ is the nodal surplus $\sigma_n := \phi_n - (n - 1)$. The critical point is nondegenerate.*

Proof. First we remark that analyticity of $\lambda_n(\Gamma_{\text{mag}}^{\vec{\alpha}})$ is a consequence of standard perturbation theory applied to the simple eigenvalue $\lambda_n(0)$. Moreover, when $\alpha_j = i\xi_j$, with real ξ_j, the Hamiltonian $H(\Gamma_{\text{mag}}^{i\vec{\xi}})$

$$\begin{tikzcd}
h^{\vec{\alpha}}_{\mathrm{mag}}(z) \arrow[d, swap, "\begin{array}{l}\text{eigenvector}\\ \text{minimax,}\\ 2(n-1)\end{array}"] \arrow[dr, "\underset{\alpha_j}{\min}\,\underset{\alpha_k}{\max},\ p"] & & h^{\mathrm{cut}}_{\vec{\gamma}}(z) \arrow[dl, swap, "\underset{\gamma_j<0}{\min}\,\underset{\gamma_j>0}{\max},\ \beta-p"] \arrow[d, "\begin{array}{l}\text{eigenvector}\\ \text{minimax,}\\ 2(\phi_n-p)\end{array}"] \\
\lambda_n(\Gamma^{\vec{\alpha}}_{\mathrm{mag}}) & \hat{h}(z) & \lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}})
\end{tikzcd}$$

Figure 4. Schematic diagram of a possible proof of Theorem 5.4. From Theorem 5.2, we know the index of the critical point of $\lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}})$ (bottom right corner). We then apply the reduction lemma four times to get the index of $\lambda_n(\Gamma^{\vec{\alpha}}_{\mathrm{mag}})$ (bottom left corner).

is a matrix with real entries. It is no longer Hermitian, but its complex eigenvalues must come in conjugate pairs; therefore a simple eigenvalue $\lambda_n(0)$ remains real for $\vec{\xi}$ in a small neighborhood of 0.

Let $\psi = \psi(\vec{\xi})$ be the corresponding real eigenfunction. It is a perturbation of f; therefore it is nonzero locally around $\vec{\xi} = 0$. For every edge $(u_j, v_j) \in S$, we let

$$\gamma_j = \frac{e^{-\xi_j}\psi_v(\vec{\xi})}{\psi_u(\vec{\xi})}. \tag{36}$$

This defines a mapping

$$R : (\xi_1, \dots, \xi_\beta) \mapsto (\gamma_1, \dots \gamma_\beta), \tag{37}$$

which is smooth in a neighborhood of zero. We also have $R(0, \dots, 0) = (\gamma_1^\circ, \dots, \gamma_\beta^\circ) = \vec{\gamma}^\circ$, where the γ_j° are given by (33). The inverse of R, which can be directly calculated from (36), is also a smooth function in a neighborhood of the point $\vec{\gamma}^\circ$. Therefore R is a diffeomorphism.

Moreover, ψ is an eigenfunction of both $\Gamma^{i\vec{\xi}}_{\mathrm{mag}}$ (by construction) and $\Gamma^{\mathrm{cut}}_{\vec{\gamma}}$ with $\vec{\gamma} = R(\vec{\xi})$ (since ψ is in the null-space of the perturbation B^{mc} of (16)), and their eigenvalues coincide, with the appropriate shift in numbering (see (29)). Namely, we have

$$\lambda_n(\Gamma^{i\vec{\xi}}_{\mathrm{mag}}) = \lambda_{\phi_n-p+1}(\Gamma^{\mathrm{cut}}_{\vec{\gamma}}), \qquad \vec{\gamma} = R(\vec{\xi}).$$

By Sylvester's law of inertia, the index is not affected by the diffeomorphism R, and we get from Theorem 5.2 that $\vec{\xi} = 0$ is a nondegenerate critical point of $\lambda_n(\Gamma^{i\vec{\xi}}_{\mathrm{mag}})$ of inertia

$$\mathcal{I}_\xi = (n - 1 + \beta - \phi_n,\ 0,\ \phi_n - n + 1).$$

Finally, since $\vec{\alpha} = i\vec{\xi}$, the Hessian of λ_n with respect to $\vec{\alpha}$ is the Hessian with respect to $\vec{\xi}$ multiplied by $i^2 = -1$. The entries n_- and n_+ of the inertia get swapped; therefore the inertia of $\lambda_n(\Gamma^{\vec{\alpha}}_{\mathrm{mag}})$ is

$$\mathcal{I}_\alpha = (\phi_n - n + 1,\ 0,\ n - 1 + \beta - \phi_n). \qquad \square$$

6. Discussion

Some simple extensions. As already mentioned, the criticality of the point $(0, \dots, 0)$ can be easily obtained from the fact that $\lambda_n(\Gamma_{\text{mag}}^{\vec{\alpha}})$ is symmetric with respect to each variable α_j. In fact, there are 2^β points of symmetry of the function $\lambda_n(\Gamma_{\text{mag}}^{\vec{\alpha}})$, namely, the points where each α_j is equal to either 0 or π. Taking $\alpha_j = \pi$ makes the corresponding edge have the weight -1 (rather than 1) in the connectivity matrix. A statement similar to Theorem 1.1 can be proved about every point of symmetry, with the appropriate modification of the notion of a sign change: ϕ_n counts the number of edges (u, v) such that $H_{u,v} f_u f_v > 0$.

One can also easily extend the results to generalized Schrödinger operators on the graph, i.e., symmetric matrices H with the property that $H_{u,v} \neq 0$ if and only if the vertices u and v are connected. The magnetic field is introduced by multiplying off-diagonal matrix elements by phases. If $H_{u,v}$ is allowed to be positive, the notion of a "sign change" has to be modified to refer to the edges (u, v) with $H_{u,v} f_u f_v > 0$, as above. With this modification, the statement of Theorem 1.1 remains valid as stated.

The necessary modifications to the proofs are limited to having $H^2_{u_1,u_2}/\gamma$ in place of $1/\gamma$ in the definition of the "cut" Hamiltonian, Equation (13), and letting the critical value of γ_j be $\gamma_j^\circ = -H_{u_j,v_j} f_{v_j}/f_{u_j}$. All other considerations remain unchanged (in particular, Fiedler's theorem on tree eigenfunctions is already formulated in terms of "generalized sign changes").

Further consequences. Perhaps the most important feature of Theorem 1.1 is that it allows us to access some of the features of the eigenfunction via the behavior of the corresponding eigenvalue under perturbation. It is known that the eigenvalues of the Laplacian are connected to the statistics of the closed paths on the graph. The connection is given through the so-called "trace formulae", which can be obtained from a graph analogue of the Selberg zeta function, the Ihara zeta function [Ihara 1966; Bass 1992; Stark and Terras 1996]. An extension by Bartholdi [1999] (see also [Mizuno and Sato 2005]) was used in [Oren et al. 2009] to obtain a family of trace formulae including the ones for the magnetic Laplacian. Thus, the closed paths on the graph determine the spectrum of the magnetic Laplacian, which, in turn, determines the nodal count. This, in principle, establishes the existence of a general connection between the nodal count and the closed paths. However, we are not aware of any concrete general formulas. We note that such a connection has been earlier conjectured by Smilansky, with special cases reported in [Gnutzmann et al. 2006; Aronovitch and Smilansky 2010].

We would also like to mention that the result of this paper has already been used in an elegant proof by Band [2012] of the converse of Fiedler's theorem: if for all n, the n-th graph eigenfunction is generic and has $n - 1$ sign changes, the graph is a tree.

There is an interesting connection between the magnetic spectrum of a compact graph and the continuous spectrum of a periodic graph. Namely, the eigenvalue $\lambda_n(\Gamma_{\text{mag}}^{\vec{\alpha}})$ featured in this paper is the dispersion relation for the maximal Abelian cover of the graph Γ, a well studied object. One of the interesting questions regarding this object is the "full spectrum property" [Higuchi and Shirai 2004; Higuchi and Nomura 2009; Sunada 2008]: whether the continuous spectrum of the cover graph of a regular graph — in our terms, the union of ranges of the functions $\lambda_n(\Gamma_{\text{mag}}^{\vec{\alpha}})$ — contains no gaps. This question can be reformulated in terms of eigenfunctions of graphs $\Gamma_{\text{mag}}^{\vec{\alpha}}$ with all $\alpha_j = 0$ or π that have minimal and maximal number of sign changes.

This, in turn, is related to the question of whether the extrema of the dispersion relation are always achieved at the symmetry points described above. Examples to the contrary have been put forward in [Harrison et al. 2007; Exner et al. 2010]. However, an important question remains: how can one characterize the extremal points that are not points of symmetry? In this direction, the duality with the cut graphs (Section 3) might provide some answers. One can speculate that critical points of the dispersion relation correspond to critical points of the eigenvalues of the cut graph $\Gamma_{\vec{\gamma}}^{\text{cut}}$ that *do not* give rise to the eigenfunction of the graph Γ. Further, we conjecture that these "unclaimed" critical points correspond to eigenfunctions of Γ modified by enforcing Dirichlet conditions at some vertices.

The results of the present paper are derived under the assumption that the eigenvalue is nondegenerate. While this is the generic situation with respect to the change in the potential Q, it is also interesting to consider what happens in the degenerate case. The linear Zeeman effect (the magnetic perturbation splitting eigenvalues) suggests that the singularities of $\lambda_n(\Gamma_{\text{mag}}^{\vec{\alpha}})$ are conical. It should be possible to define the index of the singularity point that does not rely on differentiability.

Finally, it would be most interesting to generalize the results of the present paper to manifolds. However, we immediately encounter a conceptual problem — the "number" of zeros is infinite. Still, some measure of instability of the eigenvalue under magnetic perturbation should be related to some measure of the zero set of the corresponding eigenfunction. This can be intuitively visualized by approximating the domain eigenfunction as eigenfunctions of a discrete mesh. Moreover, the method of proof used in Section 5 might be appropriate for the manifolds as well: it is based on a connection between the magnetic spectrum and the energy of the equipartitions (see Remark 5.3), and on manifolds the equipartitions are well understood [Berkolaiko et al. 2012a].

After this manuscript had been submitted, the author was notified by Y. Colin de Verdière that he found an alternative proof Theorem 1.1, which appears in this issue [Colin de Verdiére 2013]. The proof is based on a direct application of the eigenvalue perturbation formulas and a clever choice of gauge that significantly simplifies the calculations. Colin de Verdière also succeeded in proving an analogue of Theorem 1.1 for continuous Schrödinger operators on a circle (also called Hill operators). An extension of Theorem 1.1 to general quantum graphs has been subsequently obtained in [Berkolaiko and Weyand 2012].

Acknowledgment

The impetus for considering magnetic perturbation was indirectly given to the author by two people: B. Helffer and P. Kuchment. B. Helffer, in a talk at "Selected topics in spectral theory" at the Erwin Schrödinger Institute (Vienna), January 13–15, 2011, described the conjecture that the minimal spectral partition of a domain can be obtained by finding the minimal energy of the Laplacian with a number of Aharonov–Bohm flux lines (the conjecture has since been proven by B. Helffer and T. Hoffmann-Ostenhof [2013]). P. Kuchment gave the author a preprint by Rueckriemen [2011], which discussed objects similar to the ones introduced in a study of minimal spectral partitions on graphs [Band et al. 2012]. We are also grateful to P. Kuchment for many subsequent enlightening discussions, in particular regarding the topics touched upon in Section 6. Interesting discussions with and corrections by R. Band, Y. Colin de Verdière, U. Smilansky and T. Weyand are gratefully acknowledged. The visit to the ESI in Vienna was possible through the generosity of the Institute.

References

[Aronovitch and Smilansky 2010] A. Aronovitch and U. Smilansky, "Trace formula for counting nodal domains on the boundaries of chaotic 2D billiards", preprint, 2010. arXiv 1006.5656

[Band 2012] R. Band, "The nodal count {0, 1, 2, 3, . . .} implies the graph is a tree", preprint, 2012. To appear in *Phil. Trans. Roy. Soc. A*. arXiv 1212.6710

[Band et al. 2008] R. Band, I. Oren, and U. Smilansky, "Nodal domains on graphs — how to count them and why?", pp. 5–27 in *Analysis on graphs and its applications*, edited by P. Exner et al., Proc. Sympos. Pure Math. **77**, Amer. Math. Soc., Providence, RI, 2008. MR 2010d:05088 Zbl 1194.05078

[Band et al. 2012] R. Band, G. Berkolaiko, H. Raz, and U. Smilansky, "The number of nodal domains on quantum graphs as a stability index of graph partitions", *Comm. Math. Phys.* **311**:3 (2012), 815–838. MR 2909765 Zbl 1244.81026

[Bartholdi 1999] L. Bartholdi, "Counting paths in graphs", *Enseign. Math.* (2) **45**:1-2 (1999), 83–131. MR 2000f:05047 Zbl 0961.05032

[Bass 1992] H. Bass, "The Ihara–Selberg zeta function of a tree lattice", *Internat. J. Math.* **3**:6 (1992), 717–797. MR 94a:11072 Zbl 0767.11025

[Berkolaiko 2008] G. Berkolaiko, "A lower bound for nodal count on discrete and metric graphs", *Comm. Math. Phys.* **278**:3 (2008), 803–819. MR 2008j:39019 Zbl 1171.05356

[Berkolaiko and Weyand 2012] G. Berkolaiko and T. Weyand, "Stability of eigenvalues of quantum graphs with respect to magnetic perturbation and the nodal count of the eigenfunctions", preprint, 2012. To appear in *Phil. Trans. Roy. Soc. A*. arXiv 1212.4475

[Berkolaiko et al. 2012a] G. Berkolaiko, P. Kuchment, and U. Smilansky, "Critical partitions and nodal deficiency of billiard eigenfunctions", *Geom. Funct. Anal.* **22**:6 (2012), 1517–1540. MR 3000497 Zbl 06132219

[Berkolaiko et al. 2012b] G. Berkolaiko, H. Raz, and U. Smilansky, "Stability of nodal structures in graph eigenfunctions and its relation to the nodal domain count", *J. Phys. A* **45**:16 (2012), 165203, 16. MR 2910498 Zbl 1242.05159

[Birch 1756] T. Birch, *History of the Royal Society of London*, Millar, London, 1756.

[Bıyıkoğlu 2003] T. Bıyıkoğlu, "A discrete nodal domain theorem for trees", *Linear Algebra Appl.* **360** (2003), 197–205. MR 2003j:05082 Zbl 1012.05114

[Bıyıkoğlu et al. 2007] T. Bıyıkoğlu, J. Leydold, and P. F. Stadler, *Laplacian eigenvectors of graphs*, Lecture Notes in Mathematics **1915**, Springer, Berlin, 2007. MR 2009a:05119 Zbl 1129.05001

[Blum et al. 2002] G. Blum, S. Gnutzmann, and U. Smilansky, "Nodal domains statistics: A criterion for quantum chaos", *Phys. Rev. Lett.* **88** (2002).

[Bogomolny and Schmit 2002] E. Bogomolny and C. Schmit, "Percolation model for nodal domains of chaotic wave functions", *Phys. Rev. Lett.* **88** (2002).

[Chladni 1787] E. Chladni, *Entdeckungen über die Theorie des Klanges*, Weidmanns Erben und Reich, Leipzig, 1787.

[Colin de Verdière 1998] Y. Colin de Verdière, *Spectres de graphes*, Cours Spécialisés **4**, Société Mathématique de France, Paris, 1998. MR 99k:05108 Zbl 0913.05071

[Colin de Verdiére 2013] Y. Colin de Verdiére, "Magnetic interpretation of the nodal defect on graphs", *Anal. PDE* **6**:5 (2013), 1235–1242.

[Colin de Verdière et al. 2011] Y. Colin de Verdière, N. Torki-Hamza, and F. Truc, "Essential self-adjointness for combinatorial Schrödinger operators, III: Magnetic fields", *Ann. Fac. Sci. Toulouse Math.* (6) **20**:3 (2011), 599–611. MR 2012m:81063 Zbl 1250.47025

[Courant 1923] R. Courant, "Ein allgemeiner Satz zur Theorie der Eigenfuktionen selbstadjungierter Differentialausdrücke", *Nachr. Ges. Wiss. Göttingen. Math Phys.* (1923), 81–84.

[Courant and Hilbert 1953] R. Courant and D. Hilbert, *Methods of mathematical physics*, vol. 1, Interscience, New York, 1953. MR 16,426a Zbl 0051.28802

[Davies et al. 2001] E. B. Davies, G. M. L. Gladwell, J. Leydold, and P. F. Stadler, "Discrete nodal domain theorems", *Linear Algebra Appl.* **336** (2001), 51–60. MR 2002j:39011 Zbl 0990.05093

[Elon 2008] Y. Elon, "Eigenvectors of the discrete Laplacian on regular graphs — a statistical approach", *J. Phys. A* **41**:43 (2008), 435203, 17. MR 2009j:81044 Zbl 1157.39314

[Exner et al. 2010] P. Exner, P. Kuchment, and B. Winn, "On the location of spectral edges in $\mathbb{Z}$-periodic media", *J. Phys. A* **43**:47 (2010), 474022, 8. MR 2011m:81113 Zbl 1204.81063

[Fiedler 1975] M. Fiedler, "Eigenvectors of acyclic matrices", *Czechoslovak Math. J.* **25(100)**:4 (1975), 607–618. MR 52 #8151 Zbl 0325.15014

[Galileo 1638] Galileo Galilei, *Discorsi e dimostrazioni matematiche, intorno a due nuove scienze*, Louis Elsevier, Leiden, 1638.

[Gnutzmann et al. 2006] S. Gnutzmann, P. D. Karageorge, and U. Smilansky, "Can one count the shape of a drum?", *Phys. Rev. Lett.* **97**:9 (2006), 090201, 4. MR 2007h:58056 Zbl 1228.58019

[Harper 1955] P. G. Harper, "Single band motion of conduction electrons in a uniform magnetic field", *Proc. Phys. Soc. Lon. A* **68** (1955), 874–878.

[Harrison et al. 2007] J. M. Harrison, P. Kuchment, A. Sobolev, and B. Winn, "On occurrence of spectral edges for periodic operators inside the Brillouin zone", *J. Phys. A* **40**:27 (2007), 7597–7618. MR 2008j:81039 Zbl 1141.82340

[Helffer and Hoffmann-Ostenhof 2013] B. Helffer and T. Hoffmann-Ostenhof, "On a magnetic characterization of spectral minimal partitions", *J. Eur. Math. Soc.* **15**:6 (2013), 2081–2092.

[Helffer et al. 2009] B. Helffer, T. Hoffmann-Ostenhof, and S. Terracini, "Nodal domains and spectral minimal partitions", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **26**:1 (2009), 101–138. MR 2010j:35092 Zbl 1171.35083

[Higuchi and Nomura 2009] Y. Higuchi and Y. Nomura, "Spectral structure of the Laplacian on a covering graph", *European J. Combin.* **30**:2 (2009), 570–585. MR 2010a:35028 Zbl 1247.05137

[Higuchi and Shirai 2004] Y. Higuchi and T. Shirai, "Some spectral and geometric properties for infinite graphs", pp. 29–56 in *Discrete geometric analysis*, edited by M. Kotani et al., Contemp. Math. **347**, Amer. Math. Soc., Providence, RI, 2004. MR 2006f:47040 Zbl 1079.47035

[Hofstadter 1976] D. R. Hofstadter, "Energy levels and wave functions of bloch electrons in rational and irrational magnetic fields", *Phys. Rev. B* **14** (1976), 2239–2249.

[Horn and Johnson 1985] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, 1985. MR 87e:15001 Zbl 0576.15001

[Ihara 1966] Y. Ihara, "On discrete subgroups of the two by two projective linear group over p-adic fields", *J. Math. Soc. Japan* **18** (1966), 219–235. MR 36 #6511 Zbl 0158.27702

[Kato 1976] T. Kato, *Perturbation theory for linear operators*, 2nd ed., Grundlehren der Mathematischen Wissenschaften **132**, Springer, Berlin, 1976. MR 53 #11389 Zbl 0342.47009

[Lieb and Loss 1993] E. H. Lieb and M. Loss, "Fluxes, Laplacians, and Kasteleyn's theorem", *Duke Math. J.* **71**:2 (1993), 337–363. MR 94j:81298 Zbl 0787.05083

[MacCurdy 1938] E. MacCurdy, *Notebooks of Leonardo da Vinci*, Jonathan Cape, London, 1938.

[Mizuno and Sato 2005] H. Mizuno and I. Sato, "A new proof of Bartholdi's theorem", *J. Algebraic Combin.* **22**:3 (2005), 259–271. MR 2006f:05119 Zbl 1143.05057

[Nazarov and Sodin 2009] F. Nazarov and M. Sodin, "On the number of nodal domains of random spherical harmonics", *Amer. J. Math.* **131**:5 (2009), 1337–1357. MR 2011b:60098 Zbl 1186.60022

[Oren 2007] I. Oren, "Nodal domain counts and the chromatic number of graphs", *J. Phys. A* **40**:32 (2007), 9825–9832. MR 2009g:05106 Zbl 1122.05060

[Oren et al. 2009] I. Oren, A. Godel, and U. Smilansky, "Trace formulae and spectral statistics for discrete Laplacians on regular graphs, I", *J. Phys. A* **42**:41 (2009), 415101, 20. MR 2010k:05181 Zbl 1179.81083

[Rueckriemen 2011] R. Rueckriemen, “Recovering quantum graphs from their Bloch spectrum”, preprint, 2011. To appear in *Ann. Inst. Fourier*. arXiv 1101.6002

[Shubin 1994] M. A. Shubin, “Discrete magnetic Laplacian”, *Comm. Math. Phys.* **164**:2 (1994), 259–275. MR 97b:82010 Zbl 0811.46079

[Smilansky and Stöckmann 2007] U. Smilansky and H.-J. Stöckmann (editors), *Nodal patterns in physics and mathematics* (Wittenberg, 2006), European Physical Journal Special Topics **145**, Springer, Heidelberg, 2007.

[Stark and Terras 1996] H. M. Stark and A. A. Terras, “Zeta functions of finite graphs and coverings”, *Adv. Math.* **121**:1 (1996), 124–165. MR 98b:11094 Zbl 0874.11064

[Sturm 1836] C. Sturm, “Mémoire sur une classe d’équations à différences partielles”, *J. Math. Pures Appl.* **1** (1836), 373–444.

[Sunada 1993] T. Sunada, “Generalized Harper operator on a graph”, Talk at “Zeta Functions in Number Theory and Geometric Analysis”, John Hopkins University, March 29–April 3, 1993.

[Sunada 1994] T. Sunada, “A discrete analogue of periodic magnetic Schrödinger operators”, pp. 283–299 in *Geometry of the spectrum* (Seattle, WA, 1993), edited by R. Brooks et al., Contemp. Math. **173**, Amer. Math. Soc., Providence, RI, 1994. MR 95i:58185 Zbl 0805.47028

[Sunada 2008] T. Sunada, “Discrete geometric analysis”, pp. 51–83 in *Analysis on graphs and its applications*, edited by P. Exner et al., Proc. Sympos. Pure Math. **77**, Amer. Math. Soc., Providence, RI, 2008. MR 2010h:05006 Zbl 1226.05001

Received 10 Dec 2012. Accepted 19 Jan 2013.

GREGORY BERKOLAIKO: `berko@math.tamu.edu`
Department of Mathematics, Texas A&M University, College Station, TX 77843-3368, United States

ANALYSIS AND PDE
Vol. 6, No. 5, 2013
dx.doi.org/10.2140/apde.2013.6.1235

MAGNETIC INTERPRETATION OF THE NODAL DEFECT ON GRAPHS

YVES COLIN DE VERDIÈRE

We present a natural proof of a recent and surprising result of Gregory Berkolaiko interpreting the Courant nodal defect as a Morse index. This proof is inspired by a nice paper of Miroslav Fiedler published in 1975.

1. Introduction

The "nodal defect" of an eigenfunction of a Schrödinger operator is closely related to the difference between the upper bound on the number of nodal domains given by Courant's theorem and the number of nodal domains. Berkolaiko [2013] has proved a nice formula for the nodal defect of an eigenfunction of a Schrödinger operator on a finite graph in terms of the Morse index of the corresponding eigenvalue as a function of a magnetic deformation of the operator. His proof remains mysterious and rather indirect. In order to get a better understanding in view of possible generalizations, it is desirable to have a more direct approach. This is what we do here, with a proof inspired by [Fiedler 1975].

After reviewing our notations, we state the main result, as well as a reinterpretation in terms of Hessians of a determinant, and give an informal description of the proof in Section 3. The proof itself is implemented in Sections 4 and 5 with an alternative view provided in Appendix A. The continuous Schrödinger operator on a circle was considered in the preprint version of this paper [Colin de Verdière 2012]. The case of quantum graphs, i.e., graphs as 1-dimensional simplicial complexes, is worked out in [Berkolaiko and Weyand 2012].

2. Notation

Let $G = (X, E)$ be a finite connected graph, where X is the set of vertices and E the set of unoriented edges. We denote by $\{x, y\}$ the edge linking the vertices x and y. We denote by $\vec{E}$ the set of oriented edges and by $[x, y]$ the edge from x to y; the set $\vec{E}$ is a 2-fold cover of E. A 1-form α on G is a map $\vec{E} \to \mathbb{R}$ such that $\alpha([y, x]) = -\alpha([x, y])$ for all $\{x, y\} \in E$. We denote by $\Omega^1(G)$ the vector space of dimension $\# E$ of 1-forms on G. The operator $d : \mathbb{R}^X \to \Omega^1(G)$ is defined by $df([x, y]) = f(y) - f(x)$. If Q is a nondegenerate, not necessarily positive, quadratic form on $\Omega^1(G)$, we denote by $d^\star$ the adjoint of d, where $\mathbb{R}^X$ carries the canonical Euclidean structure and $\Omega^1(G)$ is equipped with the symmetric inner product $\hat{Q}$ associated to Q. We have $\dim \ker d^\star = \beta$, where $\beta = 1 + \# E - \# X$ is the dimension

---

MSC2010: 05C50, 35P05.
Keywords: operator on graphs, Schrödinger operators with magnetic fields, Morse index, Courant nodal theorem, Hill's operator.

of the space of cycles of G. We will show later that, in our context, we have the Hodge decomposition $\Omega^1(G) = d\mathbb{R}^X \oplus \ker d^\star$, where both spaces are $\hat{Q}$-orthogonal.

Following [Colin de Verdière 1998], we denote by $\mathbb{O}_G$ the set of $X \times X$ real symmetric matrices H which satisfy $h_{x,y} < 0$ if $\{x, y\} \in E$ and $h_{x,y} = 0$ if $\{x, y\} \notin E$ and $x \neq y$. Note that the diagonal entries of H are arbitrary. An element H of $\mathbb{O}_G$ is called a *Schrödinger operator* on the graph G. It will be useful to write the quadratic form associated to H as

$$q_1(f) = -\sum_{\{x,y\}\in E} h_{x,y}(f(x) - f(y))^2 + \sum_{x\in X} V_x f(x)^2,$$

with $V_x = h_{x,x} + \sum_{y\sim x} h_{x,y}$. A *magnetic field* on G is a map $B : \vec{E} \to U(1)$ defined by $B([x, y]) = e^{i\alpha_{x,y}}$, where $[x, y] \mapsto \alpha_{x,y}$ is a 1-form on G. We denote by $\mathcal{B}_G = e^{i\Omega^1(G)}$ the manifold of magnetic fields on G. The magnetic Schrödinger operator H_B associated to $H \in \mathbb{O}_G$ and $B = e^{i\alpha}$ is defined by the quadratic form

$$q_B(f) = -\frac{1}{2} \sum_{[x,y]\in\vec{E}} h_{x,y}|f(x) - e^{i\alpha_{x,y}} f(y)|^2 + \sum_{x\in X} V_x |f(x)|^2$$

associated to a Hermitian form on $\mathbb{C}^X$. More explicitly, if $f \in \mathbb{C}^X$,

$$Hf(x) = h_{x,x} f(x) + \sum_{y\sim x} h_{x,y} e^{i\alpha_{x,y}} f(y). \tag{1}$$

We fix H and we denote by

$$\lambda_1(B) \le \lambda_2(B) \le \cdots \le \lambda_n(B) \le \cdots \le \lambda_{\#X}(B)$$

the eigenvalues of H_B. It will be important to notice that $\lambda_n(\bar{B}) = \lambda_n(B)$. Moreover, we have a gauge invariance: the operators H_B and $H_{B'}$ with $\alpha' = \alpha + df$ for some $f \in \mathbb{R}^X$ are unitarily equivalent. Hence they have the same eigenvalues. This implies that, if $\Omega^1(G) = d\mathbb{R}^X \oplus \ker d^\star$ (this is not always the case because Q is not positive), it is enough to consider 1-forms in the subspace $\ker d^\star$ of $\Omega^1(G)$ when studying the map $\Lambda_n : B \to \lambda_n(B)$. This holds in particular for investigations concerning the Hessian and the Morse index.

3. Statement of Berkolaiko's magnetic theorem

Before stating the main result, we recall:

Definition 1. The *Morse index* $j(q) \in \mathbb{N} \cup \{+\infty\}$ of a quadratic form q on a real vector space E is defined by $j(q) = \sup_F \dim F$, where F is a subspace of E such that $q_{|F\setminus 0}$ is less than 0. The *nullity* of q is the dimension of the kernel of q.

The *Morse index* of a smooth real-valued function f defined on a smooth manifold M at a *critical point* $x_0 \in M$ (i.e., a point satisfying $df(x_0) = 0$) is the Morse index of the Hessian of f, which is a canonically defined quadratic form on the tangent space $T_{x_0}M$. The critical point x_0 is called *nondegenerate* if the previous Hessian is nondegenerate. The *nullity* of the critical point x_0 of f is the nullity of the Hessian of f at the point x_0.

The aim of this note is to prove the following nice results due to Berkolaiko [2008; 2013]:

Theorem 1. *Let $G = (X, E)$ be a finite connected graph and β the dimension of the space of cycles of G. We suppose that the n-th eigenvalue λ_n of $H \in \mathcal{O}_G$ is simple. We assume moreover that an associated nonzero eigenfunction ϕ_n satisfies $\phi_n(x) \neq 0$ for all $x \in X$. Then, the number ν of edges along which ϕ_n changes sign satisfies $n - 1 \leq \nu \leq n - 1 + \beta$.*

Moreover $\Lambda_n : B \to \lambda_n(B)$ is smooth at $B \equiv 1$ which is a critical point of Λ_n and the nodal defect, *$\delta_n = \nu - (n - 1)$, is the Morse index of Λ_n at that point. If M is the manifold of dimension β of magnetic fields on G modulo the gauge transforms, the function $[B] \to \Lambda_n(B)$ has $[B = 1]$ as a nondegenerate critical point.*

Remark 1. The previous results can be extended by replacing the critical point $B \equiv 1$ by $B_{x,y} = \pm 1$ for all edges $\{x, y\} \in E$. The number ν is then the number of edges $\{x, y\} \in E$ satisfying $B_{x,y}\phi_n(x)\phi_n(y) < 0$ where ϕ_n is the corresponding eigenfunction.

Remark 2. The assumptions on H are satisfied for H in an open dense subset of $\mathcal{O}_G$.

The upper bound of ν in the first part of Theorem 1 is related to the Courant nodal theorem (see [Courant and Hilbert 1953, Section VI.6]) as follows: a nodal domain on a graph for the eigenfunction ϕ_n is a connected component of the subgraph G' of G obtained by removing the edges along which ϕ_n changes sign. Denoting by μ the number of nodal domains of ϕ_n, the Courant theorem for graphs (see [Colin de Verdière 1998, Theorem 2.4]) asserts that $\mu \leq n$; using the Euler formula for the graph G' and because $\mu = b_0(G')$, the number of connected components of the graph G', we get also a lower bound (see [Berkolaiko 2008]):

Corollary 1. *Under the assumptions of Theorem 1, we have $n - \beta \leq \mu \leq n$.*

Example 3.1 (bipartite graphs). Let $G = (V, E)$ be a bipartite graph: $V = Y \cup Z$ and all edges have one vertex in Y and the other in Z. Let U be the involution on $\mathbb{R}^V$ given by $Uf(x) = -f(x)$ if $x \in Y$ and $Uf(x) = f(x)$ if $x \in Z$ and let B be a magnetic field. Then $UH_BU = -H'_B$ with $H' \in O_G$, so that $\lambda_{|V|}(H_B) = -\lambda_1(H'_B)$. And hence it follows from the diamagnetic inequality that $B \to \lambda_{|V|}(H_B)$ has a maximum at $B \equiv 1$. And hence the Morse index of the Hessian of $B \to \lambda_{|V|}(H_B)$ at $B \equiv 1$ is the dimension of the manifold of magnetic fields, namely β. On the other hand the first eigenfunction ϕ_1 of H' is everywhere greater than 0 and the number of sign changes of $U\phi_1$ is $|E|$. So Berkolaiko's formula for $\lambda_{|V|}$ gives $(|V| - 1) + \beta = |E|$. This is the Euler formula.

Theorem 1 can be reinterpreted as follows:

Theorem 2. *Under the assumptions as in Theorem 1, consider the functional $D_n : B \mapsto \det(H_B - \lambda_n(1))$. Then $B \equiv 1$ is a nondegenerate critical point of D_n whose Morse index is δ_n if n is odd and $\beta - \delta_n$ if n is even.*

Proof. Under the assumptions of the theorem we have

$$\det(H_B - \lambda_n(1)) = (\lambda_n(B) - \lambda_n(1)) \det{}'(H_B - \lambda_n(1))$$

where $\det'(H_B) = F(B)$ is the product of the eigenvalues $\lambda_j - \lambda_n(1)$ for $j \neq n$. The following lemma is easy to check by direct computations of the second derivatives:

Lemma 1. *Let $F = fG$ where F, f, G are smooth real valued functions defined near a point x_0 on a smooth manifold. Let us assume that $f(x_0) = 0$ and $f'(x_0) = 0$; then the Hessian of F at the point x_0 is $G(x_0)$ times the Hessian of f at x_0.*

From the lemma, we get that the Hessian of D_n at $B \equiv 1$ is $F(1)$ times the Hessian of Λ_n. We have $(-1)^{n-1} F(1) > 0$. The conclusion follows. ☐

There is a formula for the characteristic polynomial of a magnetic Laplacian on graphs due to Robin Forman [1993] and reproved by Richard Kenyon [2012] and Yurii Burman [2012]. Using the gauge change $f \to f\phi_n$ as in [Colin de Verdière 1998] gives a Laplace type operator whose entries can be of any sign. Forman's formula extends to that case and it would be nice to relate Berkolaiko's formula to Forman's formula.

Important warning: Without loss of generality, we can and *will* assume in the rest of this note that $\lambda_n = \Lambda_n(1) = 0$. This implies that the Morse index of q_1 is $n - 1$.

In the course of the proof we will use a special choice of gauge in which we can compute the Hessian explicitly. More precisely, according to the classical perturbation formulae,

$$\ddot{\lambda} = (\phi, \ddot{H}\phi) + 2(\dot{H}\phi, \dot{\phi}),$$

where we assumed that λ is at a critical point: $\dot{\lambda} = 0$. The first term is easy to calculate explicitly; for perturbation in the direction of the 1-form ω it is

$$Q(\omega) = \frac{1}{2} \sum_{\vec{E}} a_{x,y} \omega([x, y])^2 \quad \text{with } a_{x,y} = -h_{x,y} \phi_n(x) \phi_n(y) = a_{y,x}. \tag{2}$$

Considered as a quadratic form in ω, Q is already in the diagonal form. Its index is clearly the number of negative values among $\{-h_{x,y}\phi_n(x)\phi_n(y)\}$, or, in other words, the number ν of edges where ϕ_n changes sign!

We will present an explicit choice of gauge in which the second term vanishes. The condition for this is $\dot{H}\phi = 0$ which, after explicit calculation, can be interpreted as $\omega \in \ker d^\star$, where $d^\star$ is the conjugate of d with respect to the inner product induced by (2).

Finally, we observe that the index of $Q(\omega)$ has been computed to be ν in the whole of $\Omega^1(G)$, whereas we should be restricting ourselves to our chosen gauge, $\omega \in \ker d^\star$. We will show that this restriction reduces the index precisely by $n - 1$. Indeed, the splitting $\Omega^1(G) = d\mathbb{R}^X \oplus \ker d^\star$ is orthogonal with respect to the form Q; therefore

$$\operatorname{ind}(Q) = \operatorname{ind}(Q|_{d\mathbb{R}^X}) + \operatorname{ind}(Q|_{\ker d^\star}).$$

We establish that $\operatorname{ind}(Q|_{d\mathbb{R}^X}) = n - 1$ by relating the form Q on $d\mathbb{R}^X$ to the quadratic form q_1 around the point ϕ_n.

4. The quadratic form Q

Lemma 2. *The set of forms* $f \to (f(x) - f(y))^2$ *where* $\{x, y\} \in \mathcal{P}_2(X)$, *the set of subsets with two elements of* X, *and* $f \to f(x)^2$ *with* $x \in X$ *is a basis of the set of quadratic forms on* $\mathbb{R}^X$.

Definition 2. A quadratic form q on $\mathbb{R}^X$ is said of Laplace type if for all $f \in \mathbb{R}^X$, $\hat{q}(1, f) \equiv 0$ where $\hat{q}$ is the symmetric bilinear form associated to q.

Lemma 3. *The set of forms* $f \to (f(x) - f(y))^2$, $\{x, y\} \in \mathcal{P}_2(X)$ *is a basis of the space of quadratic forms of Laplace type.*

The form $\tilde{q}_1 : f \to q_1(\phi_n f)$, where $\phi_n f$ is the pointwise product of ϕ_n and f, is of Laplace type because

$$\widehat{\tilde{q}_1}(1, g) = \langle H\phi_n | \phi_n g\rangle = \langle 0 | \phi_n g\rangle.$$

Hence $\widehat{\tilde{q}_1}(1, g) = 0$.

Moreover, $\tilde{q}_1(f) = Q(df)$. Indeed, because of Lemma 3, it is enough to compare the coefficients of the basis forms $f \to (f(x) - f(y))^2$. The form $f \to Q(df)$ is already expanded in this basis. To find the coefficient for the form $f \to \tilde{q}_1(f)$, we observe that (because we know it is of Laplace type) the coefficient in question is minus the coefficient in front of the term $f(x)f(y)$, divided by two. This evaluates to $a_{x,y}$ (see (2)).

In fact, we will need to use $\hat{Q}(df, dg) = \langle H(\phi_n f) | \phi_n g\rangle$.

Lemma 4. *The Morse index of* $Q_{|d\mathbb{R}^X}$ *is equal to* $n - 1$.

It is a general fact that the Morse index of the quadratic form $f \to Q(Af)$ is the same as the Morse index of the restriction of Q to the image of A. Hence, the Morse index of $Q_{|d\mathbb{R}^X}$ is the Morse index of $\tilde{q}_1$ on $\mathbb{R}^X$. Because $f \to \phi_n f$ is a linear isomorphism, this index is equal to the index of q_1 by the Sylvester theorem. Since $\lambda_n = 0$, the index of q_1 is $n - 1$ by elementary spectral theory.

Lemma 5. *Let us denote by* $d^\star$ *the adjoint of* d *where* $\mathbb{R}^X$ *is equipped with the canonical Euclidean structure and* $\Omega^1(G)$ *with the inner product associated to* Q. *The space* $\Omega^1(G)$ *splits as*

$$\Omega^1(G) = d\mathbb{R}^X \oplus \ker d^\star$$

(*Hodge type splitting*), *and this decomposition is* Q*-orthogonal.*

More explicitly $d^\star$ is given by

$$d^\star\omega(x) = \sum_{y \sim x} a_{x,y}\omega([y, x]).$$

If $\omega = df$ satisfies $d^\star\omega = 0$, we have $d^\star df = 0$. Hence $\hat{Q}(df, dg) = 0$ for all g and $\langle H(\phi_n f) | \phi_n g)\rangle = 0$. Because λ_n is of multiplicity 1, this implies that f is constant and hence $df = 0$. So $d\mathbb{R}^X \cap \ker d^\star = \{0\}$ and the conclusions follow.

At this point, we know that the nodal defect is the Morse index of the restriction of Q to the space $\ker d^\star$ of dimension β. The first part of Theorem 1 follows.

5. The magnetic Hessian

We need one more fact to complete the proof: to identify the Hessian of Λ_n on $e^{i\ker d^\star}$ at $B \equiv 1$ with the restriction of Q to $\ker d^\star$.

Let us denote by $S \subset \mathbb{C}^X$ the set of unit vectors f normalized so that $f(x_0)$ is real and $f(x_0) > 0$ where x_0 is chosen in X.

Lemma 6. *The point $B \equiv 1$ is a critical point of Λ_n. If $\phi_n(B) \in S$ is the eigenfunction of H_B corresponding to the eigenvalue $\lambda_n(B)$, the differential of $B \to \phi_n(B)$ vanishes at $B \equiv 1$ on* $\ker d^\star$.

The first property comes from the fact that $\Lambda_n(\bar{B}) = \Lambda_n(B)$. We can compute, for any variation $e^{it\alpha}$, t close to 0, of $B \equiv 1$, that $\dot{H}_B\phi_n + H\dot{\phi}_n = 0$. The condition $d^\star\alpha = 0$ can be written as

$$\sum_{y\sim x} h_{x,y}\phi_n(y)\alpha_{x,y} = 0 \quad \text{for all } x \in X.$$

From (1), this is equivalent to $\dot{H}_B\phi_n = 0$. Hence $H(\dot{\phi}_n) = 0$ and $\dot{\phi}_n = c\phi_n$ since λ_n is simple. From the normalization $\|\phi_n(B)\| = 1$, we get $c \in i\mathbb{R}$ and, since $\dot{\phi}_n(x_0) \in \mathbb{R}$, the number c is real. We deduce that $\dot{\phi}_n = 0$.

Lemma 7. *The function $F : S \times e^{i\ker d^\star} \to \mathbb{R}$ defined by $F(f, e^{i\alpha}) = \langle H_{e^{i\alpha}} f \,|\, f\rangle$ admits $(\phi_n, 0)$ as a critical point and the Hessian of $(\Lambda_n)_{|e^{i\ker d^\star}}$ at the point $B \equiv 1$ is the form Q.*

The differential of F with respect to f vanishes because f is an eigenfunction of H. The differential with respect to $\ker d^\star$ vanishes, because $F(f, e^{i\alpha}) = F(f, e^{-i\alpha})$. The Hessian of F at $(\phi_n, 0)$ is well defined. Because the differential at $B = 1$ of $B \to \phi_n(B)$ vanishes on $e^{i\ker d^\star}$, the Hessians of $\Lambda_n : B \to F(\phi_n(B), B)$ and $M_n : B \to F(\phi_n(1), B)$ agree. A simple calculation of the Hessian of M_n gives the result:

$$\begin{aligned} M_n(e^{i\alpha}) &= -\frac{1}{2}\sum_{[x,y]\in\vec{E}} h_{x,y}|\phi_n(x) - e^{i\alpha_{x,y}}\phi_n(y)|^2 + \sum_{x\in X} V_x|\phi_n(x)|^2 \\ &= -\sum_{[x,y]\in E} h_{x,y}\big(\phi_n(x)^2 + \phi_n(y)^2 - 2\cos\alpha_{x,y}\phi_n(x)\phi_n(y)\big) + \sum_{x\in X} V_x|\phi_n(x)|^2. \end{aligned}$$

Computing the second derivative with respect to α at $\alpha = 0$ gives $\mathrm{Hessian}(M_n) = Q(\alpha)$.

Appendix A: A pedestrian approach to the calculus of the Hessian of Λ_n in Section 5

We will derive a direct approach to the calculus of the second derivative of an eigenvalue which could be used directly in the proof of Lemma 7. Let $t \to A(t)$ be a C^2 curve defined near $t = 0$ in the space of Hermitian matrices on a finite-dimensional Hilbert space $(\mathcal{H}, \langle\cdot|\cdot\rangle)$. Let us assume that $\lambda(0)$ is an eigenvalue of $A(0)$ of multiplicity one with a normalized eigenvector $\phi(0)$. Then, for t close to 0, $A(t)$ has a simple eigenvalue $\lambda(t)$ of multiplicity one which is a C^2 function of t. We can choose an associated eigenfunction $\phi(t)$ which is C^2 with respect to t. The following assertions give the values of the first and second derivatives of $\lambda(t)$ at $t = 0$:

Proposition 1. *Under the previous assumptions, we have*

$$\lambda'(0) = \langle A'(0)\phi(0) | \phi(0)\rangle.$$

If $\lambda'(0) = 0$, *we have*

$$\lambda''(0) = \langle A''(0)\phi(0) | \phi(0)\rangle + 2\langle \phi'(0) | A'(0)\phi(0)\rangle,$$

where $\phi'(0)$ *is any solution of* $(A(0) - \lambda(0))\phi'(0) = -A'(0)\phi(0)$.

In particular, if $A'(0)\phi(0) = 0$,

$$\lambda''(0) = \langle A''(0)\phi(0) | \phi(0)\rangle.$$

Proof. We start with $(A(t) - \lambda(t))\phi(t) = 0$ where $\phi(t)$ is an eigenfunction of $A(t)$ which depends in a C^2 way on t. Taking the first derivative, we get

$$(A'(t) - \lambda'(t))\phi(t) + (A(t) - \lambda(t))\phi'(t) = 0. \tag{3}$$

Putting $t = 0$ and taking the scalar product with $\phi(0)$, we get the formula for $\lambda'(0)$. Similarly, the t-derivative of (3) is

$$(A''(t) - \lambda''(t))\phi(t) + 2(A'(t) - \lambda'(t))\phi'(t) + (A(t) - \lambda(t))\phi''(t) = 0. \tag{4}$$

Putting $t = 0$, taking the scalar product with $\phi(0)$ and using $\lambda'(0) = 0$, we get the result. ☐

We can apply this to $A(t) := H_{e^{it\alpha}}$ with $\alpha \in \ker d^\star$ in order to get the Hessian of Λ_n in Section 5. The condition $A'(0)\phi(0) = 0$ is exactly $d^\star\alpha = 0$!

Appendix B: The case where the eigenfunction vanishes at some vertex

In this appendix, we take $H \in \mathbb{O}_G$ and assume that $\lambda_n = 0$ is nondegenerate eigenvalue of H with a normalized eigenfunction ϕ. We have:

Proposition 2. *Let us assume that, for all vertices x satisfying $\phi(x) = 0$, there exists a vertex $y \sim x$ so that $\phi(y) \neq 0$. Then, for any $\psi \in \mathbb{R}^X$ orthogonal to ϕ, there exists a smooth deformation $H_t \in \mathbb{O}_G$ of H so that $\dot\phi = \psi$.*

It is enough to check that the space of $\dot H\phi$ is $\mathbb{R}^X$ and to use the first variation formulae given in Appendix A.

Theorem 3. *Let us assume that the function ϕ vanishes at the unique vertex x_0. Then, the nullity of the Hessian of the "magnetic variation" of H is at least $|n_+ - n_-|$ where $n_\pm$ is the number of vertices $x \sim x_0$ so that $\pm\phi(x) > 0$.*

Proof. Choose a smooth variation H_t of H so that $\dot\phi(x_0) = 1$. Let ν be the number of sign changes of ϕ away from x_0. Then, for $t > 0$ small enough, the number of sign changes of ϕ_t is $\nu + n_-$ while, for $t < 0$ small enough, it is $\nu + n_+$. We see from Theorem 1 that the magnetic Morse index is $\nu + n_- - (n-1)$ for $t > 0$ and $\nu + n_+ - (n-1)$. The discontinuity of the Morse index at $t = 0$ is $|n_+ - n_-|$. This gives the lower bound on the nullity. ☐

Corollary 2. *If $|n_+ - n_-| > \beta$, the eigenvalue 0 is degenerate.*

Let us remark that this lower bound is not always sharp. In the following example, we have $n_+ = n_-$, $\beta = 2$ and the nullity of the Hessian is 2.

Example B.1. The graph G is made of 2 cycles of length 3 with a common vertex. The matrix of H is chosen as follows:

$$[H] = -\begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 2 & 0 & 0 \\ 1 & 2 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 1 & 1 \end{pmatrix}.$$

Using the fact that the graph has a symmetry of order 2 exchanging the 2 cycles, one can split $\mathbb{R}^X$ and the matrix H into the even and odd parts. This allows us to check that $\lambda_4 = 0$ is nondegenerate. In order to compute the magnetic Hessian, we check that it is possible to build a decomposition $\Omega^1(G) = d\mathbb{R}^X \oplus K$ which is Q-orthogonal and with $K \subset \ker d^\star$. It is then easy to check that the magnetic Hessian evaluated on K vanishes.

Acknowledgements

Thanks to Roland Bacher, Gregory Berkolaiko and Françoise Truc for carefully checking the manuscript.

References

[Berkolaiko 2008] G. Berkolaiko, "A lower bound for nodal count on discrete and metric graphs", *Comm. Math. Phys.* **278**:3 (2008), 803–819. MR 2008j:39019 Zbl 1171.05356

[Berkolaiko 2013] G. Berkolaiko, "Nodal count of graph eigenfunctions via magnetic perturbation", *Anal. PDE* **6**:5 (2013), 1213–1233.

[Berkolaiko and Weyand 2012] G. Berkolaiko and T. Weyand, "Stability of eigenvalues of quantum graphs with respect to magnetic perturbation and the nodal count of the eigenfunctions", preprint, 2012. arXiv 1212.4475

[Burman 2012] Y. M. Burman, "Operators of rank 1, discrete path integration and graph Laplacians", preprint, 2012. arXiv 1205.1123

[Colin de Verdière 1998] Y. Colin de Verdière, *Spectres de graphes*, Cours Spécialisés **4**, Société Mathématique de France, Paris, 1998. MR 99k:05108 Zbl 0913.05071

[Colin de Verdière 2012] Y. Colin de Verdière, "Magnetic interpretation of the nodal defect on graphs", version 1, preprint, 2012. arXiv 1201.1110v1

[Courant and Hilbert 1953] R. Courant and D. Hilbert, *Methods of mathematical physics, I*, Interscience Publishers, New York, 1953. MR 16,426a Zbl 0051.28802

[Fiedler 1975] M. Fiedler, "Eigenvectors of acyclic matrices", *Czechoslovak Math. J.* **25**:4 (1975), 607–618. MR 52 #8151 Zbl 0325.15014

[Forman 1993] R. Forman, "Determinants of Laplacians on graphs", *Topology* **32**:1 (1993), 35–46. MR 94g:58247 Zbl 0780.05041

[Kenyon 2012] R. Kenyon, "The Laplacian on planar graphs and graphs on surfaces", preprint, 2012. arXiv 1203.1256

Received 13 Dec 2012. Revised 15 Jan 2013. Accepted 14 Feb 2013.

YVES COLIN DE VERDIÈRE: `yves.colin-de-verdiere@ujf-grenoble.fr`
Unité Mixte de Recherche CNRS-UJF 5582, Institut Fourier, BP 74, 38402 Saint Martin d'Hères cedex, France
http://www-fourier.ujf-grenoble.fr/~ycolver/

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at msp.berkeley.edu/apde.

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 in APDE are usually in English, 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 make any reference to the bibliography. 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 subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should 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 details about how your graphics were generated.

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.

Analysis & PDE

Volume 6 No. 5 2013

A Lichnerowicz estimate for the first eigenvalue of convex domains in Kähler manifolds 1001
Vincent Guedj, Boris Kolev and Nader Yeganefar

Sharp modulus of continuity for parabolic equations on manifolds and lower bounds for the first eigenvalue 1013
Ben Andrews and Julie Clutterbuck

Some minimization problems in the class of convex functions with prescribed determinant 1025
Nam Q. Le and Ovidiu Savin

On the spectrum of deformations of compact double-sided flat hypersurfaces 1051
Denis Borisov and Pedro Freitas

Stabilization for the semilinear wave equation with geometric control condition 1089
Romain Joly and Camille Laurent

Instability theory of the Navier–Stokes–Poisson equations 1121
Juhi Jang and Ian Tice

Dynamical ionization bounds for atoms 1183
Enno Lenzmann and Mathieu Lewin

Nodal count of graph eigenfunctions via magnetic perturbation 1213
Gregory Berkolaiko

Magnetic interpretation of the nodal defect on graphs 1235
Yves Colin de Verdière

2157-5045(2013)6:5;1-D