Analysis & PDE

Volume 7 No. 3 2014

msp

Analysis & PDE

msp.org/apde

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

The subscription price for 2014 is US $180/year for the electronic version, and $355/year (+$50, 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/

ANALYSIS AND PDE
Vol. 7, No. 3, 2014
dx.doi.org/10.2140/apde.2014.7.529

PRESCRIPTION DU SPECTRE DE STEKLOV DANS UNE CLASSE CONFORME

PIERRE JAMMES

Sur toute variété compacte de dimension $n \geq 3$ à bord, on prescrit toute partie finie du spectre de Steklov dans une classe conforme donnée. En particulier, on prescrit la multiplicité des valeurs propres. Sur une surface compacte à bord donnée, on montre que la multiplicité de la k-ième valeur propre est bornée indépendamment de la métrique. Sur le disque, on donne des résultats plus précis : la multiplicité de la première et la deuxième valeurs propres non nulles sont au plus 2 et 3 respectivement. Pour le problème de Steklov–Neumann sur le disque, on montre que la multiplicité de la k-ième valeur propre non nulle est au plus $k+1$.

On any compact manifold of dimension $n \geq 3$ with boundary, we prescribe any finite part of the Steklov spectrum within a given conformal class. In particular, we prescribe the multiplicity of the first eigenvalues. On a compact surface with boundary, we show that the multiplicity of the k-th eigenvalue is bounded independently of the metric. On the disk, we give more precise results: the multiplicity of the first and second positive eigenvalues are at most 2 and 3 respectively. For the Steklov–Neumann problem on the disk, we prove that the multiplicity of the k-th positive eigenvalue is at most $k+1$.

1. Introduction

Étant donnée une variété riemannienne (M, g) compacte à bord et une fonction strictement positive $\rho \in C^0(\partial M)$, le spectre de Steklov de M est l'ensemble des réels σ tels que le système

$$\begin{cases} \Delta f = 0 & \text{dans } M, \\ \partial f/\partial \nu = \sigma \rho f & \text{sur } \partial M, \end{cases} \tag{1.1}$$

où ν est un vecteur normal unitaire sortant le long de ∂M, admette des solutions non triviales. Ce spectre est formé de valeurs propres positives notées $0 = \sigma_0(M, g, \rho) < \sigma_1(M, g, \rho) \leq \sigma_2(M, g, \rho) \leq \cdots \to +\infty$. Si $\rho \equiv 1$, alors c'est aussi le spectre de l'opérateur Dirichlet-to-Neumann sur M.

Un grand nombre de travaux récents visent à borner ces valeurs propres sous des contraintes géométriques, par exemple avec des hypothèses conformes [Fraser et Schoen 2011; Hassannezhad 2011], ou en fonction d'une constante isopérimétrique [Colbois et al. 2011]. Le but de cet article vise au contraire à mettre en évidence le fait que ce spectre possède une certaine souplesse et que si la dimension de M est au moins 3, on peut en prescrire toute partie finie, c'est-à-dire que si on se donne une suite finie de réels strictement positifs, il existe une métrique sur M telle que cette suite soit le début du spectre. On va en fait montrer un résultat plus fort, à savoir qu'on peut prescrire simultanément le début du spectre, la multiplicité des valeurs propres, la classe conforme de la variété et la fonction densité ρ sur le bord.

MSC2010 : 35P15, 58J50.
Mots-clefs : Steklov eigenvalues, prescription of eigenvalues, conformal geometry.

On étend ainsi au spectre de Steklov un résultat obtenu dans [1987] pour le laplacien et généralisé aux formes différentielles dans [Jammes 2011].

Théorème 1.2. *Soit (M^n, g) une variété riemannienne compacte à bord de dimension $n \geq 3$. Étant donnée une fonction strictement positive $\rho \in C^0(\partial M)$, un entier $N \geq 1$ et une suite finie de réels strictement positifs $0 < a_1 \leq a_2 \leq \cdots \leq a_N$, il existe une métrique $\tilde{g}$ conforme à g telle que*

$$\sigma_k(M, \rho, \tilde{g}) = a_k$$

pour tout $k \in [1, N]$.

Remarque 1.3. On ne peut pas prescrire simultanément le spectre, le volume et la classe conforme. En effet, si on fixe le volume et la classe conforme, les valeurs propres ne peuvent pas être arbitrairement grandes (cf. [Fraser et Schoen 2011] et [Hassannezhad 2011]). Des obstructions semblables existent pour le laplacien usuel [El Soufi et Ilias 1986], le laplacien de Hodge en restriction aux formes différentielles de certains degrés [Jammes 2007; Jammes 2008] et l'opérateur de Dirac [Ammann 2003]. Le théorème 1.2 montre en revanche que même en fixant la classe conforme, on ne peut pas majorer le rapport σ_k/σ_l pour $k > l$.

Remarque 1.4. On sait que la prescription de multiplicité est possible pour les opérateurs de Schrödinger en dimension $n \geq 3$ [Colin de Verdière 1986 ; 1987] et les opérateurs agissant sur les formes différentielles en dimension $n \geq 4$ [Jammes 2011 ; 2012]. Mais ce problème n'est toujours pas résolu pour les formes différentielles en dimension 3, ni pour l'opérateur de Dirac, dont on ne sait actuellement prescrire le spectre que si les valeurs propres sont simples [Dahl 2005].

La principale difficulté consiste à prescrire la multiplicité des valeurs propres. On utilisera pour cela les techniques introduites dans [Colin de Verdière 1986] (voir [Jammes 2009] pour une présentation plus détaillée de ce sujet). Les principaux ingrédients sont des théorèmes de convergence spectrale (en particulier de convergence du spectre vers celui d'un domaine de la variété : théorème 3.8), un modèle de valeur propre multiple fourni par un laplacien combinatoire sur des graphes (paragraphe 4B).

La démonstration du théorème 1.2 échoue en dimension 2, entre autres à cause de l'invariance conforme de la norme L^2 du gradient d'une fonction. On va montrer que cette difficulté ne peut pas être contournée et qu'il existe en fait une obstruction à la prescription de multiplicité en dimension 2. La démonstration suit celles de S. Y. Cheng [1976] et G. Besson [1980] pour majorer la multiplicité des valeurs propres du laplacien.

Théorème 1.5. *Sur toute surface riemannienne compacte orientable à bord (M, g) de genre γ et toute fonction strictement positive $\rho \in C^0(\partial M)$, la multiplicité de $\sigma_k(M, \rho, g)$ est majorée par $4\gamma + 2k + 1$. Si M est non orientable et qu'on note l le nombre de composantes connexes de ∂M, alors la multiplicité de $\sigma_k(M, \rho, g)$ est majorée par $4p + 4k + 1$, où p est l'invariant topologique $1 - \chi(M) - l$.*

Remarque 1.6. Lors de la finalisation de cet article, la démonstration de bornes sur la multiplicité est apparu simultanément dans deux prépublications. A. Fraser et R. Schoen [2012] ont montré indépendamment le même théorème, avec une démonstration presque identique. Ils montrent aussi que la borne obtenue

pour $\sigma_1(S^1 \times [0,1])$, à savoir 3, est optimale. Simultanément, M. Karpukhin, G. Kokarev, I. Polterovich ont démontré dans [Karpukhin et al. 2012] une de ces bornes avec des techniques différentes : ils montrent que la multiplicité de σ_k est majorée par $2p+2k+1$ et $2p+2l+k$, que la surface soit orientable ou non.

Les bornes données par le théorème 1.5 sont les mêmes que celles obtenues par G. Besson pour les valeurs propres de laplacien. Il s'avère que le spectre de Steklov possède des rigidités supplémentaires qu'on va illustrer dans le cas du disque :

Théorème 1.7. *Sur le disque* $\mathbb{D}$, *la multiplicité de* $\sigma_1(\mathbb{D}, \rho, g)$ *est au plus* 2 *et celle de* $\sigma_2(\mathbb{D}, \rho, g)$ *est au plus* 3.

Remarque 1.8. En utilisant les résultats de [Colin de Verdière 1988], on peut facilement construire (par excision d'un petit disque sur la sphère) une métrique sur $\mathbb{D}$ telle que la première valeur propre non nulle du laplacien avec condition de Neumann (ou la seconde pour la condition de Dirichlet) soit de multiplicité 3. En outre, la borne sur la multiplicité de σ_1 est optimale puisque pour la métrique canonique, toutes les valeurs propres non nulles sont doubles.

Remarque 1.9. L'article [Alessandrini et Magnanini 1994], qui traite de la multiplicité des σ_k sur le disque, contient comme cas particulier le fait que la multiplicité de σ_1 est au plus 2 ; on en donnera une démonstration un peu plus directe. En revanche, la borne sur la multiplicité de σ_2 ne semble pas être apparue auparavant dans la littérature.

On va aussi montrer une autre borne sur la multiplicité dans le cas du disque, mais pour une variante du problème de Steklov, à savoir le problème de Steklov–Neumann. Ce problème est défini de la manière suivante : on partitionne le bord ∂M en deux sous-variétés (pas nécessairement connexes) $\partial M = \partial M_S \cup \partial M_N$ et pour une fonction $\rho \in C^0(\partial M_S)$, on pose la condition $\partial f/\partial \nu = \sigma \rho f$ sur ∂M_S et on demande à f de vérifier la condition de Neumann sur ∂M_N (voir paragraphe 2B pour plus de détails).

Théorème 1.10. *Étant donnée une partition* (*non triviale*) $\partial\mathbb{D}_S \cup \partial\mathbb{D}_N$ *du bord du disque* $\mathbb{D}$, *la multiplicité de* $\sigma_k(\mathbb{D}, \rho, g)$ *pour le problème de Steklov–Neumann relativement à cette partition est au plus* $k+1$.

Remarque 1.11. Pour le laplacien, les meilleures bornes connues sont asymptotiquement de l'ordre de $2k$ quand $k \to +\infty$. Par exemple, pour le laplacien de Dirichlet sur le disque, il et montré dans [Hoffmann-Ostenhof et al. 1999] que la multiplicité de la k-ième valeur propre est au plus $2k-3$. Dans [Karpukhin et al. 2012], la meilleure borne donnée pour le problème de Steklov sur le disque est $k+2$.

Remarque 1.12. Dans le cas particulier du problème hydrodynamique de ballottement (voir [Kopachevsky et Krein 2001] ou les rappels du paragraphe 2B), on sait que la première valeur propre non nulle est simple (cf. [Kozlov et al. 2004]). Il est conjecturé que les autres sont simples aussi, mais cette question reste ouverte.

Colin de Verdière a conjecturé que la multiplicité maximale de la deuxième valeur propre d'un opérateur de Schrödinger sur une surface M est $\mathrm{Chr}\, M - 1$, où $\mathrm{Chr}\, M$ est le nombre chromatique de M, c'est-à-dire le nombre de sommets du plus grand graphe complet plongeable dans M. Comme la démonstration du théorème 1.2 repose sur des graphes plongées dans M dont les sommets sont sur le bord de la variété, on peut envisager de transposer cette conjecture au problème de Steklov sous la forme suivante :

Conjecture 1.13. Soit M une surface compacte à bord M, et soit $\mathrm{Chr}(M, \partial M)$ le nombre de sommets du plus grand graphe complet qu'on peut plonger dans M en plaçant les sommets sur ∂M. Alors la multiplicité maximale de $\sigma_1(M)$ est $\mathrm{Chr}(M, \partial M) - 1$.

D'après ce qui précède, cette conjecture est vérifiée sur le disque $\mathbb{D}$ et le cylindre $S^1 \times [0, 1]$.

La section 2 rappellera quelques propriétés du spectre de Steklov et de l'opérateur Dirichlet-to-Neumann dont nous auront besoin. Nous montrerons dans la section 3 les théorèmes de convergence spectrale que nous utiliserons, et nous les appliquerons dans la section 4 pour démontrer le théorème 1.2. Enfin, la section 5 sera consacrée au cas de la dimension 2 et à la démonstration des théorèmes 1.5, 1.7 et 1.10.

2. Le problème de Steklov

2A. ***Définition du spectre de Steklov.*** On se donne une variété riemannienne (M, g) compacte à bord telle que ∂M soit C^1 par morceau (dans la suite, g désignera indifféremment la métrique sur M ou la métrique induite sur ∂M). Le problème des valeurs propres de Steklov consiste à résoudre l'équation

$$\begin{cases} \Delta f = 0 & \text{dans } M, \\ \partial f / \partial \nu = \sigma \rho f & \text{sur } \partial M, \end{cases} \tag{2.1}$$

où ν est un vecteur unitaire sortant normal au bord et $\rho \in C^0(\partial M)$ un fonction densité fixée. L'ensemble des réels σ solutions du problème forme un spectre discret positif noté

$$0 = \sigma_0(M, g, \rho) < \sigma_1(M, g, \rho) \leq \sigma_2(M, g, \rho) \leq \cdots . \tag{2.2}$$

Le problème de Steklov, déjà étudié à la fin du XIX$^{\text{e}}$ siècle et au début du XX$^{\text{e}}$ (voir [Stekloff 1899 ; 1902] et les références qui y sont données), apparaît dans divers problèmes physiques. Par exemple il permet de modéliser l'évolution d'une membrane libre dont la masse se concentre sur son bord, et il intervient dans certains problèmes de tomographie. On verra au paragraphe qui suit qu'il apparaît aussi en hydrodynamique.

Notre principal outil sera la caractérisation variationnelle suivante du spectre de Steklov (cf. [Bandle 1980]) :

$$\sigma_k(M, g, \rho) = \inf_{V_{k+1} \in H^1(M)} \sup_{f \in V_{k+1} \setminus \{0\}} \frac{\int_M |\mathrm{d}f|^2 \, \mathrm{d}v_g}{\int_{\partial M} f^2 \rho \, \mathrm{d}v_g}, \tag{2.3}$$

où V_k parcours les sous-espaces de dimension k de l'espace de Sobolev $H^1(M)$.

Il faut prendre garde au fait que $|f|^2 = \int_{\partial M} f^2 \rho \, \mathrm{d}v_g$ ne définit pas une norme de Hilbert sur $L^2(M)$ (elle est nulle sur les fonctions vérifiant la condition de Dirichlet). En revanche, on peut utiliser les techniques usuelles de min-max en considérant l'espace de Hilbert $L^2(\partial M)$ muni de la métrique $|\cdot|$ qu'on vient de définir, et la forme quadratique $Q(f) = \int_M |\mathrm{d}\tilde{f}|^2 \, \mathrm{d}v_g$, où $\tilde{f}$ est le prolongement harmonique de f. Il sera parfois commode de redéfinir la forme quadratique Q par

$$Q(f) = \inf_{\substack{\tilde{f} \in H^1(M) \\ \tilde{f}|_{\partial M} = f}} \int_M |\mathrm{d}\tilde{f}|^2 \, \mathrm{d}v_g. \tag{2.4}$$

Cette définition sera en particulier applicable dans les situations où on considère une métrique singulière sur M (voir paragraphe 2C).

Dans le cas homogène, c'est-à-dire quand $\rho \equiv 1$, le spectre de Steklov est aussi connu comme étant le spectre de l'opérateur Dirichlet-to-Neumann, qu'on notera $\Lambda : C^\infty(\partial M) \to C^\infty(\partial M)$, défini comme suit : étant donné une fonction $f \in C^\infty(\partial M)$, on prolonge harmoniquement f dans M et on pose

$$\Lambda f(x) = \frac{\partial f}{\partial \nu}(x). \tag{2.5}$$

Le spectre de Λ est bien celui de Q car pour une fonction harmonique, on a $\int_M |\mathrm{d}f|^2 \,\mathrm{d}v_g = \int_{\partial M} f \frac{\partial f}{\partial \nu} \,\mathrm{d}v_g$.

L'opérateur Λ n'est pas un opérateur différentiel sur ∂M (ce n'est même pas un opérateur local), mais c'est un opérateur pseudo-différentiel elliptique d'ordre 1 (cf. [Taylor 1996b, Chapter 7]). En particulier, nous utiliserons le fait qu'il vérifie une inégalité elliptique :

$$\|f\|^2_{H^1(\partial M)} \leq c \int_{\partial M} f \Lambda f \,\mathrm{d}v_g + c' \|f\|^2_{L^p(\partial M)}, \tag{2.6}$$

où $p \in [1, +\infty]$, les constante c, c' dépendant de p et de la métrique g sur M mais pas de f.

Pour finir, nous auront besoin d'une propriété d'unique prolongement des fonctions propres en dimension 2 :

Théorème 2.7. *Soit f une fonction propre du problème de Steklov sur une surface. Si f s'annule sur un ouvert du bord, alors $f \equiv 0$.*

Démonstration. Soit I un intervalle du bord sur lequel f s'annule. On peut déformer conformément la surface de manière à ce que I devienne géodésique et que la métrique reste inchangée sur le reste du bord. Par invariance conforme de l'harmonicité et de la condition $\partial f/\partial \nu = 0$, f est toujours fonction propre. En notant x un paramètre sur I, on a $\partial^2 f/\partial x^2 = 0$, donc aussi $\partial^2 f/\partial \nu^2 = 0$ puisque f est harmonique. Enfin, comme $\partial f/\partial \nu = 0$ sur I, on a aussi $\partial^2 f/\partial x\, \partial \nu = 0$ et donc le développement à l'ordre 2 de f est nul le long de I.

Par conséquent, au voisinage d'un point de I, on peut prolonger f par 0 en dehors de M et obtenir une fonction $\tilde{f}$ qui est C^2 et vérifie $\Delta f = 0$. Par unique prolongement des fonctions harmoniques, on a $f \equiv 0$ sur M. □

2B. ***Le problème de Steklov–Neumann.*** Étant donné une variété compacte à bord M, on se donne un domaine (ou une union de domaines disjoints) à bord C^1 par morceaux de ∂M qu'on notera ∂M_S, et on pose $\partial M_N = \partial M \backslash \partial M_S$. Si ρ est une fonction sur ∂M_S, le problème de Steklov–Neumann se pose ainsi :

$$\begin{cases} \Delta f = 0 & \text{dans } M, \\ \partial f/\partial \nu = \sigma \rho f & \text{sur } \partial M_S, \\ \partial f/\partial \nu = 0 & \text{sur } \partial M_N, \end{cases} \tag{2.8}$$

c'est-à-dire qu'on demande à la fonction harmonique f de vérifier la condition de Neumann sur ∂M_N. On appellera respectivement bord de Steklov et bord de Neumann les ensembles ∂M_S et ∂M_N. Les solutions de ce problème interviennent dans l'étude du phénomène hydrodynamique de ballottement (*sloshing*

problem) : si on considère un fluide parfait incompressible contenu dans un récipient M avec une surface libre ∂M_S, les petites oscillations périodiques du fluide correspondent aux solutions de (2.8) pour une fonction ρ constante (voir, par exemple, [Kopachevsky et Krein 2001]).

Le problème de Steklov–Neumann possède un spectre discret et positif qu'on notera

$$0 = \sigma_0(M, \partial M_S, g, \rho) < \sigma_1(M, \partial M_S, g, \rho) \leq \sigma_2(M, \partial M_S, g, \rho) \leq \cdots \tag{2.9}$$

Le spectre de Steklov–Neumann possède la même caractérisation variationnelle que le spectre de Steklov, à condition de restreindre l'intégrale sur le bord au bord de Steklov :

$$\sigma_k(M, \partial M_S, g, \rho) = \inf_{V_{k+1} \in H^1(M)} \sup_{f \in V_{k+1} \setminus \{0\}} \frac{\int_M |\mathrm{d}f|^2 \, \mathrm{d}v_g}{\int_{\partial M_S} f^2 \rho \, \mathrm{d}v_g}, \tag{2.10}$$

où V_k parcours les sous-espaces de dimension k de $H^1(M)$.

L'opérateur Dirichlet-to-Neumann est bien défini sur ∂M_S en considérant des fonctions harmoniques vérifiant la condition de Neumann sur ∂M_N et vérifie toujours l'inégalité elliptique (2.6).

On aura besoin du fait que si on se donne une fonction f sur ∂M_S et qu'on la prolonge en une fonction harmonique (toujours notée f), sa norme L^2 sur ∂M_N est contrôlée par sa norme sur ∂M_S, c'est-à-dire qu'il existe une constante $c > 0$ ne dépendant que de g et ρ telle que $\int_{\partial M_N} f^2 \leq c \int_{\partial M_S} f^2 \rho$. Cela découle du fait que la norme $L^2(\partial M_N)$ de f est contrôlée par sa norme $H^{1/2}(M)$, elle même contrôlée par sa norme $L^2(\partial M_S)$ (cf. [Taylor 1996a, §4.4]).

On utilisera aussi un bref usage du spectre de Steklov–Dirichlet, défini en considérant des fonctions harmoniques qui vérifient la condition $f = 0$ sur $\partial M \setminus \partial M_S$. La propriété de ce spectre qui nous intéressera est qu'il est strictement positif (cf. [Agranovich 2006]).

2C. ***Fonctions harmoniques et métriques singulières.*** Dans la section suivante, on aura à manipuler des métriques discontinues. Si U est un domaine de (M, g) et $\varepsilon \in [0, 1]$ un réel fixé, elles seront de la forme

$$\begin{cases} g_\varepsilon = \varepsilon^2 g & \text{sur } U, \\ g_\varepsilon = g & \text{sur } M \setminus U. \end{cases} \tag{2.11}$$

Comme les normes L^2 et de Sobolev pour les métriques g et g_ε sont équivalentes, la théorie spectrale de la forme quadratique $\|\mathrm{d}\cdot\|^2_{g_\varepsilon}$ sur $H^1(M)$ est donc similaire à celle de $\|\mathrm{d}\cdot\|^2_g$. On peut donc définir le prolongement harmonique d'une fonction $f \in C^\infty(\partial M)$ comme étant le prolongement minimisant cette forme quadratique. Cette définition est cohérente avec celle de la forme quadratique Q donnée par (2.4).

Un tel prolongement minimisera en particulier $\|\mathrm{d}\cdot\|^2_{g_\varepsilon}$ en restriction à chacun des domaines U et $M \setminus U$, il sera donc harmonique au sens usuel sur ces deux domaines.

Les mêmes remarques s'appliquent au problème de Steklov–Neumann.

3. Théorèmes de convergence spectrale

3A. ***Rappels.*** Dans cette section, nous allons montrer plusieurs théorèmes de convergence spectrale dont nous aurons besoin pour prescrire le spectre de Steklov. Nous utiliserons pour cela les techniques développées dans [Colin de Verdière 1986]. Pour prescrire la multiplicité des valeurs propres, il nous

faudra montrer la convergence des espaces propres, et nous aurons aussi besoin d'une certaine uniformité de la convergence, nous reprendrons pour cela les notations de Colin de Verdière :

Soit E_0 et E_1 sont deux sous-espaces vectoriels de même dimension N d'un espace de Hilbert, munis respectivement des formes quadratiques q_0 et q_1. Si E_0 et E_1 sont suffisamment proches, il existe une isométrie naturelle ψ entre les deux (voir la section I de [Colin de Verdière 1986] pour les détails de la construction), on définit alors l'écart entre q_0 et q_1 par $\|q_1 \circ \psi - q_0\|$. Pour deux formes quadratiques Q_0 et Q_1 sur l'espace de Hilbert, on appellera *N-écart spectral entre Q_0 et Q_1* l'écart entre les deux formes quadratiques restreintes à la somme des espaces propres associés aux N premières valeurs propres. Si cet écart est petit, alors les N premières valeurs propres de Q_0 et leurs espaces propres sont proches de ceux de Q_1.

On veut montrer que la convergence spectrale est uniforme pour une certaine famille de spectres limites. D'après Colin de Verdière, on dira donc qu'une forme quadratique vérifie l'hypothèse $(*)$ si ses valeurs propres vérifient

$$\lambda_1 \leq \cdots \leq \lambda_N < \lambda_N + \eta \leq \lambda_{N+1} \leq M \qquad (*)$$

pour un entier N et des réels η, $M > 0$ fixés une fois pour toutes. Dans les énoncés suivants, les constantes N, M et η auront ces valeurs préalablement fixées.

Lemme 3.1 [Colin de Verdière 1986, théorème I.7]. *Soit Q une forme quadratique positive sur un espace de Hilbert $\mathcal{H}$ dont le domaine admet la décomposition Q-orthogonale* $\mathrm{dom}(Q) = \mathcal{H}_0 \oplus \mathcal{H}_\infty$. *Pour tout $\varepsilon > 0$, il existe une constante $C(\eta, M, N, \varepsilon) > 0$ (grande) telle que si $Q_0 = Q|_{\mathcal{H}_0}$ vérifie l'hypothèse $(*)$ et si $Q(x) \geq C|x|^2$ pour tout $x \in \mathcal{H}_\infty$, alors Q et Q_0 ont un N-écart spectral inférieur à ε.*

Lemme 3.2 [ibid., théorème I.8]. *Soit $(\mathcal{H}, |\cdot|)$ un espace de Hilbert muni d'une forme quadratique positive Q. On se donne en outre une suite de métriques $|\cdot|_n$ sur $\mathcal{H}$ et une suite de formes quadratiques Q_n de même domaine que Q telles que* :

(i) *il existe $C_1, C_2 > 0$ tels que $C_1|x| \leq |x|_n \leq C_2|x|$ pour tout $x \in \mathcal{H}$* ;

(ii) $|x|_n \to |x|$ *pour tout* $x \in \mathrm{dom}(Q)$;

(iii) $Q(x) \leq Q_n(x)$ *pour tout* $x \in \mathrm{dom}(Q)$;

(iv) $Q_n(x) \to Q(x)$ *pour tout* $x \in \mathrm{dom}(Q)$.

Si Q vérifie l'hypothèse $(*)$, alors à partir d'un certain rang (dépendant de η, M et N), Q et Q_n ont un N-écart spectral inférieur à ε.

Remarque 3.3. Comme on l'a remarqué dans [Jammes 2011], dans le lemme 3.2, on peut affaiblir l'hypothèse $C_1|x| \leq |x|_n \leq C_2|x|$ en $C_1|x| \leq |x|_n \leq C_2|x| + \varepsilon_n Q_n(x)^{1/2}$ avec $\varepsilon_n \to 0$, la démonstration restant exactement la même (on peut aussi remplacer Q_n par Q dans cette dernière inégalité). En particulier, il n'est pas nécessaire que l'espace de Hilbert $(\mathcal{H}, |\cdot|)$ soit complet pour $|\cdot|_n$.

Remarque 3.4. On peut aussi remplacer l'hypothèse $x \in \mathrm{dom}(Q) \Rightarrow Q(x) \leq Q_n(x)$ par $Q(x) \leq M \Rightarrow Q(x) \leq (1 + \varepsilon_n) Q_n(x)$ avec $\varepsilon_n \to 0$.

Remarque 3.5. Pour déduire la convergence du spectre et des espaces propres de la convergence des formes quadratiques, on doit en principe se ramener à une norme de Hilbert fixe. Ça ne sera pas nécessaire dans la suite car les étapes de la démonstration où la norme varie seront traitées à l'aide du lemme 3.2.

3B. ***Densité et convergence de spectre.*** Notre premier résultat de convergence sera de montrer qu'avec une densité fixée sur le bord, on peut déformer conformément la métrique de manière à faire tendre le spectre de Steklov vers le spectre correspondant à une autre densité. On peut en outre faire tendre la métrique déformée vers la métrique initiale dans l'intérieur de la variété. On se restreindra au cas où la densité initiale est plus petite que la densité du spectre limite, ce qui sera suffisant pour les applications dans la section 4.

Théorème 3.6. *Soit (M, g) une variété riemannienne compacte à bord, et $\rho, \bar\rho \in C^0(\partial M)$ deux fonctions sur le bord de M telles que $\bar\rho \geq \rho$.*

Il existe une famille g_ε de métriques conformes à g sur M telle que :

(i) *$\sigma_k(M, g_\varepsilon, \rho)$ tend vers $\sigma_k(M, g, \bar\rho)$ quand $\varepsilon \to 0$ pour tout $k \geq 0$, avec convergence des espaces propres.*

(ii) *$g_\varepsilon = (\bar\rho/\rho)^{2/(n-1)} g$ sur ∂M.*

(iii) *g_ε tend vers g uniformément sur tout compact dans l'intérieur de M.*

En outre, si les $\sigma_k(M, g, \bar\rho)$ vérifient l'hypothèse (∗), alors le N-écart spectral entre (M, g_ε, ρ) et $(M, g, \bar\rho)$ tend vers 0.

Démonstration. On définit une famille $h_\varepsilon \in C^\infty(M)$ de facteurs conformes de la manière suivante : on fixe $h_\varepsilon(x) = (\bar\rho/\rho)^{1/(n-1)}$ pour $x \in \partial M$ et on étend h_ε de manière lisse de sorte que la famille (h_ε) tende simplement vers 1 dans l'intérieur de M, et uniformément sur tout compact ne rencontrant pas le bord. On pose alors $g_\varepsilon = h_\varepsilon^2 g$ pour tout ε.

La famille de métriques g_ε induit les familles de normes et de formes quadratiques

$$Q_\varepsilon(f) = \inf_{\substack{\tilde f \in H^1(M) \\ \tilde f|_{\partial M} = f}} \int_M h_\varepsilon^{n-2} |\mathrm{d}\tilde f|^2 \,\mathrm{d}v_g \quad \text{et} \quad |f|_\varepsilon = \int_{\partial M} f^2 \bar\rho \,\mathrm{d}v_g. \tag{3.7}$$

Comme $\bar\rho \geq \rho$, on peut choisir une suite (h_ε) décroissante, les suites Q_ε et $|\cdot|_\varepsilon$ vérifient alors les hypothèses du lemme 3.2, ce qui suffit pour conclure. □

3C. ***Convergence vers le spectre d'un domaine.*** Le second théorème consiste à faire converger le spectre de Steklov d'une variété à bord M vers le spectre de Steklov–Neumann d'un domaine U de M, avec la condition de Steklov sur $\partial U_S = \partial U \cap \partial M$ et la condition de Neumann sur le reste du bord de U. Ce résultat étend au spectre de Steklov de théorèmes analogues concernant le laplacien agissant sur les fonctions [Colin de Verdière 1986] et sur les formes différentielles [Jammes 2011].

Théorème 3.8. *Soit (M^n, g) une variété riemannienne compacte à bord de dimension $n \geq 3$, $\rho \in C^\infty(\partial M)$ et U un domaine de M à bord C^1 par morceau tel que $\partial U_S = \partial U \cap \partial M$ soit non vide. Il existe une famille g_ε de métriques sur M conformes à g telle que* :

(i) $g = g_\varepsilon$ *en restriction à* U.

(ii) $\mathrm{Vol}(M, g_\varepsilon) \to \mathrm{Vol}(U, g)$ *quand* $\varepsilon \to 0$.

(iii) $\sigma_k(M, \rho, g_\varepsilon) \to \sigma_k(U, \partial U_S, \rho|_{\partial U_S}, g|_U)$ *quand* $\varepsilon \to 0$ *pour tout* $k \geq 0$, *avec convergence des espaces propres.*

En outre, si les $\sigma_k(U, \partial U_S, \rho|_{\partial U_S}, g|_U)$ *vérifient l'hypothèse* (∗), *alors le* N*-écart spectral entre* (M, ρ, g_ε) *et* $(U, \partial U_S, \rho|_{\partial U_S}, g|_U)$ *tend vers* 0.

Démonstration. La démonstration est similaire à celle du théorème III.1 de [Colin de Verdière 1986] et passe par l'intermédiaire, pour un réel $\eta > 0$ petit donné, de la métrique singulière g_η définie par $g_\eta = g$ sur U et $g_\eta = \eta^2 g$ sur $M \backslash U$. Elle se déroule en deux étapes : d'abord, on montre la convergence spectre pour la famille de métriques singulières, puis on approche ces métriques singulières par des métriques lisses. On conclut en se donnant, pour un $\varepsilon > 0$ donné, une métrique g_η tel que l'écart spectral avec le spectre de (U, g) soit inférieur à ε, puis une métrique lisse g_ε tel que l'écart spectral avec g_η soit lui aussi inférieur à ε.

On fera souvent appel à la forme quadratique définie en (2.4), en particulier quand on manipule des métriques singulières.

Étape 1. Un réel $\eta > 0$ étant donné, la métrique g_η induit sur $L^2(\partial M)$ la forme quadratique

$$Q_\eta(f) = \inf_{\tilde{f}|_{\partial M} = f} \left(\int_U |\mathrm{d}\tilde{f}|^2 \, \mathrm{d}v_g + \eta^{n-2} \int_{M \backslash U} |\mathrm{d}\tilde{f}|^2 \, \mathrm{d}v_g \right) \tag{3.9}$$

et la norme $|f|_{g_\eta} = \int_{\partial U_S} f^2 \rho \, \mathrm{d}v_g + \eta^{(n-1)} \int_{\partial M \backslash \partial U_S} f^2 \rho \, \mathrm{d}v_g$. On va utiliser le lemme 3.1 pour se ramener à un sous-domaine de la forme quadratique puis appliquer le lemme 3.2.

En notant $\mathcal{H}$ le domaine de la forme quadratique Q_η, on définit l'espace $\mathcal{H}_\infty = \{f \in \mathcal{H},\ f|_{\partial U_S} = 0\}$ et on note $\mathcal{H}_0$ son orthogonal pour la forme quadratique Q_η. Pour appliquer le lemme 3.1, on doit minorer la forme quadratique Q_η sur $\mathcal{H}_\infty$ en fonction de $|\cdot|_\eta$. Si $f \in \mathcal{H}_\infty$, alors

$$|f|_\eta^2 = \eta^{(n-1)} \int_{\partial M \backslash \partial U_S} f^2 \rho \, \mathrm{d}v_g = \eta^{(n-1)} |f|^2$$

et

$$Q_\eta(f) \geq \eta^{n-2} \inf_{\tilde{f}|_{\partial M} = f} \int_M |\mathrm{d}\tilde{f}|^2 \, \mathrm{d}v_g = \eta^{n-2} Q(f).$$

On est donc ramené à l'étude du spectre de la forme quadratique Q associée à la métrique initiale g en restriction à l'espace $\mathcal{H}_\infty$, c'est-à-dire à minorer le spectre de Steklov sur M avec condition de Dirichlet sur ∂U_S. Comme 0 n'est pas dans le spectre de Steklov–Dirichlet (cf. paragraphe 2B), il existe une constante $c > 0$ telle que $Q(f)/|f|^2 > c$ pour tout $f \in \mathcal{H}_\infty$. Par conséquent, $Q_\eta(f)/|f|_\eta \geq c \cdot \eta^{-2}$ pour tout $f \in \mathcal{H}_\infty$. Si η est suffisamment petit, on peut donc appliquer le lemme 3.1 et en déduire que le spectre pour la métrique g_η est proche du spectre de Q_η restreint à $\mathcal{H}_0$.

Il reste à montrer que la limite du spectre de $Q_\eta|_{\mathcal{H}_0}$ est le spectre de Steklov–Neumann du domaine U. On utilisera pour cela le lemme 3.2. Puisque $\mathcal{H}_0$ est défini comme le Q_η-orthogonal des fonctions de ∂M nulles sur ∂U_S, une fonction de $\mathcal{H}_0$ est entièrement déterminée par sa restriction à ∂U_S. Plus précisément,

parmi les fonctions f dont la valeur sur ∂U_S est fixée, celle qui est dans $\mathscr{H}_0$ est celle minimise la forme quadratique Q_η. C'est donc la restriction au bord du prolongement harmonique (tel qu'on l'a défini au paragraphe 2C) de $f_{\partial U_S}$ avec condition de Neumann sur $\partial M \backslash \partial U_S$. Dans la suite, on identifiera souvent une fonction sur ∂U_S avec le prolongement ainsi défini.

La norme $|\cdot|_\eta$ converge en décroissant vers la norme $|\cdot|$ définie par $|f| = \int_{\partial U_S} f^2 \rho \, \mathrm{d}v_g$. L'hypothèse (ii) et la première inégalité de l'hypothèse (i) du lemme 3.2 sont donc satisfaites. Les hypothèses (iii) et (iv) sont vérifiées pour les mêmes raisons.

Il reste à montrer que la deuxième inégalité de l'hypothèse (i) est vérifiée. Pour cela, on doit majorer $\int_{\partial M \backslash \partial U_S} f^2 \rho \, \mathrm{d}v_g$. Notons $\tilde{f}$ le prolongement de $f|_{\partial U_S}$ qui est harmonique au sens du paragraphe 2C, c'est-à-dire que

$$Q_\eta(f) = \int_U |\mathrm{d}\tilde{f}|^2 \, \mathrm{d}v_g + \eta^{n-2} \int_{M \backslash U} |\mathrm{d}\tilde{f}|^2 \, \mathrm{d}v_g.$$

Comme on l'a remarqué au paragraphe 2B, puisque $\tilde{f}$ est harmonique sur $M \backslash U$ avec condition de Neumann sur $\partial M \backslash \partial U_S$, la norme $L^2(\partial M \backslash \partial U_S, \rho)$ de $\tilde{f}$ est contrôlée par sa norme L^2 sur $\partial U \backslash \partial U_S$, c'est-à-dire que

$$\int_{\partial M \backslash \partial U_S} \tilde{f}^2 \rho \, \mathrm{d}v_g \leq c_1 \int_{\partial U \backslash \partial U_S} \tilde{f}^2 \rho \, \mathrm{d}v_g.$$

Notons que la constante c_1 est invariante par homothétie, donc indépendante de η, à condition de considérer sur $\partial(M \backslash U)$ la métrique induite par la métrique de $M \backslash U$. En considérant la métrique g_η sur $\partial M \backslash \partial U_S$ et la métrique g sur $\partial U \backslash \partial U_S$ on obtient

$$\int_{\partial M \backslash \partial U_S} \tilde{f}^2 \rho \mathrm{d}v_{g_\eta} \leq \eta^{n-1} c_1 \int_{\partial U \backslash \partial U_S} \tilde{f}^2 \, \mathrm{d}v_g \tag{3.10}$$

On majore le membre de droite à l'aide de l'inégalité elliptique de l'opérateur Dirichlet-to-Neumann sur ∂U.

$$\int_{\partial U \backslash \partial U_S} \tilde{f}^2 \, \mathrm{d}v_g \leq \|f\|^2_{L^2(\partial U)} \leq \|f\|^2_{H^1(\partial U)} \leq c_2 \int_{\partial U} f \frac{\partial f}{\partial \nu} \, \mathrm{d}v_g = c_2 \int_U |\mathrm{d}\tilde{f}|^2 \, \mathrm{d}v_g \leq c_2 Q_\eta(f). \tag{3.11}$$

On a donc finalement $|f|^2_\eta \leq |f|^2 + \eta^{n-1} c_1 c_2 Q_\eta(f)$ ce qui permet d'appliquer le lemme 3.2 et la remarque 3.3.

Étape 2. On doit montrer que pour tout $\eta > 0$, le spectre de Q_η peut être approché par le spectre de métriques lisses conformes à g.

Le paramètre η état fixé, on définit une suite de facteurs conformes h_i tels que la suite (h_i) converge en décroissant vers la fonction $\chi_U + \eta \chi_{M \backslash U}$ et on pose $g_i = h_i^2 g$. Les suites de norme de Hilbert $|\cdot|_i$ et de formes quadratiques Q_i associées à g_i convergent vers $|\cdot|_\eta$ et Q_η en vérifiant les hypothèses du lemme 3.2, ce qui assure la convergence du spectre et des espaces propres. □

3D. ***Convergence vers le spectre du bord.*** Enfin, nous allons montrer qu'on peut faire tendre le spectre de Steklov homogène (c'est-à-dire que $\rho \equiv 1$) d'une variété à bord M vers le spectre du laplacien sur ∂M, la métrique sur ∂M restant homothétique à la métrique initiale. Bien que ce théorème ne soit pas

indispensable pour démontrer le théorème 1.2, on peut l'utiliser si la dimension de ∂M est au moins 3. Il semble aussi intéressant en lui-même et fournit un exemple d'application du théorème démontré au paragraphe précédent.

Théorème 3.12. *Soit (M^n, g) une variété riemannienne compacte à bord de dimension $n \geq 3$. Il existe une famille g_ε de métriques sur M conformes à g et homothétiques à g le long de ∂M telle que pour tout $k \geq 0$ on ait $\sigma_k(M, g_\varepsilon) \to \lambda_k(M, g)$ quand $\varepsilon \to 0$.*

En outre, si les $\lambda_k(M, g)$ vérifient l'hypothèse (∗), alors le N-écart spectral entre le spectre de Steklov de (M, g_ε) et le spectre du laplacien de $(\partial M, g)$ tend vers 0.

Démonstration. Le principe de la démonstration consiste à se ramener au cas d'un voisinage collier du bord (avec la condition mixte Steklov–Neumann) en utilisant le théorème 3.8. On va procéder en trois étapes : d'abord montrer la convergence du spectre d'un voisinage collier de ∂M muni d'une métrique produit, puis traiter le cas de la restriction de métrique g à ce voisinage collier, et enfin montrer la convergence du spectre de M.

Étape 1. On va déterminer l'asymptotique (quand $\eta \to 0$) du spectre de la variété produit $\partial M \times [0, \eta]$ pour un métrique produit, avec la condition de Steklov sur $\partial M \times \{0\}$ et la condition de Neumann sur $\partial M \times \{\eta\}$.

On peut déduire ce spectre du spectre de Steklov de $\partial M \times [0, 2\eta]$ (avec condition de Steklov sur les deux bords) par symétrie : en effet, on peut partitionner les valeurs propres de $\partial M \times [0, 2\eta]$ en deux, selon que les fonctions propres sont symétriques ou antisymétriques. Ces fonctions propres vérifient la condition de Neumann sur $\partial M \times \{\eta\}$ dans le premier cas et la condition de Dirichlet dans le second cas. Le spectre de $\partial M \times [0, \eta]$ avec condition mixte est donc le spectre de Steklov de $\partial M \times [0, 2\eta]$ restreint aux fonctions symétriques.

Le spectre de Steklov de $\partial M \times [0, 2\eta]$ a été calculé explicitement dans lemme 6.1 de [Colbois et al. 2011] en fonction du spectre du laplacien sur ∂M : si ∂M est de volume 1 et si on note λ_k ses valeurs propres et u_k ses fonctions propres, alors le spectre non nul de $\partial M \times [0, 2\eta]$ restreint aux fonctions symétriques est $\sqrt{\lambda_k}\tanh(\eta\sqrt{\lambda_k})$ les fonctions propres associées étant $\cosh(\sqrt{\lambda_k}t)u_k(x)$, où x désigne un point de ∂M et t la coordonnée sur l'intervalle. Le spectre de Steklov–Neumann de $\partial M \times [0, \eta]$ se comporte donc asymptotiquement comme $\eta\lambda_k$ quand $\eta \to 0$. En pratiquant une homothétie sur $\partial M \times [0, \eta]$, on peut donc faire tendre son spectre vers λ_k. On peut facilement vérifier à l'aide de l'expression des fonctions propres qu'il y a bien convergence des espaces propres.

Étape 2. Étant donnée la variété à bord (M, g) et un réel $\eta > 0$ petit, on considère le η-voisinage collier M_η de ∂M, c'est-à-dire que $M_\eta = \{x \in M, d(x, \partial M) \leq \eta\}$. Pour η suffisamment petit, M_η est difféomorphe au produit de ∂M avec un intervalle. On considère alors le problème de Steklov–Neumann sur M_η comme dans l'étape 1.

Quand η tend vers 0, la métrique g restreinte à M_η est de plus en plus proche d'une métrique produit. Plus précisément, il existe une famille de réels $\tau_\eta > 1$ telle que $\tau_\eta \to 1$ quand $\eta \to 0$ et $1/\tau_\eta g_\eta \leq g|_{M_\eta} \leq \tau_\eta g_\eta$, où g_η désigne la métrique produit sur $\partial M \times [0, \eta]$. Comme la constante τ_η contrôle aussi l'écart entre les normes de Hilbert et les formes quadratiques pour le problème de Steklov–Neumann sur (M_η, g) et

(M_η, g_η), on peut appliquer le lemme 3.2 et la remarque 3.4 pour obtenir la convergence spectrale comme dans l'étape 1.

Étape 3. Pour pouvoir conclure, il on aura besoin de faire tendre le spectre de Steklov de M vers celui de M_η en restant dans la classe conforme de g. On va utilise pour cela le théorème 3.8 :

Un réel $\varepsilon > 0$ étant donné, on choisit $\eta > 0$ et un rapport d'homothétie $r_\eta > 0$ tels que le spectre de $(M_\eta, r_\eta^2 g)$ soit ε-proche de celui de $(\partial M, g)$. Puis, en appliquant le théorème 3.8 avec $U = M_\eta$, on obtient une métrique g'_η sur M tel que le spectre de (M, g'_η) soit ε-proche de celui de $(M_\eta, r_\eta^2 g)$. Quand ε tend vers 0, on a ainsi convergence du spectre et des espaces propres de (M, ρ, g_ε) vers ceux du laplacien sur $(\partial M, g)$. □

4. Prescription du spectre

4A. ***L'hypothèse de transversalité d'Arnol'd.*** Pour prescrire la multiplicité des valeurs propres de Steklov nous utilisons, selon la méthode introduite par Colin de Verdière, trois ingrédients : les théorèmes de convergence spectrale démontrés dans les sections précédentes, des modèles de valeurs propres multiples déjà connus et une propriété de stabilité vérifiée par ces modèles. Nous allons commencer par rappeler cette dernière. On verra au paragraphe suivant comment utiliser des graphes complets comme modèles de spectre avec multiplicité.

On suppose qu'on a une famille d'opérateurs $(P_a)_{a\in B^k}$, où B^k est la boule unité de $\mathbb{R}^k$ (en pratique, P_a est l'opérateur Dirichlet-to-Neumann associé à une métrique g_a), tels que P_0 possède une valeur propre λ_0 d'espace propre E_0 et de multiplicité N. Pour les petites valeurs de a, P_a possède des valeurs propres proches de λ_0 dont la somme des espaces propres est de dimension N. Comme dans la définition de l'écart spectral, on identifie cette somme à E_0 et on note q_a la forme quadratique associée à P_a transportée sur E_0.

Définition 4.1 [Colin de Verdière 1988]. On dit que λ_0 vérifie l'hypothèse de transversalité d'Arnol'd si l'application $\Psi : a \mapsto q_a$ de B^k dans $\mathcal{Q}(E_0)$ est essentielle en 0, c'est-à-dire qu'il existe $\varepsilon > 0$ tel que si $\Phi : B^k \to \mathcal{Q}(E_0)$ vérifie $\|\Psi - \Phi\|_\infty \leq \varepsilon$, alors il existe $a_0 \in B^k$ tel que $\Phi(a_0) = q_0$.

Une propriété cruciale est que si Φ provient d'une famille (P'_a) d'opérateurs, alors λ_0 est valeur propre de P'_{a_0} de multiplicité N et vérifie la même propriété de transversalité, ce qui justifie qu'on parle de stabilité de la multiplicité. Comme remarqué par Colin de Verdière, on peut généraliser cette définition à une suite finie de valeurs propres.

4B. ***Voisinages tubulaires de graphes.*** Colin de Verdière a montré qu'un graphe complet muni d'un laplacien combinatoire et d'une métrique appropriée possède une (ou plusieurs) valeur propre multiple vérifiant la propriété de transversalité d'Arnol'd. Dans ce paragraphe, nous allons utiliser ce résultat pour construire une variété dont on prescrit le début du spectre Steklov–Neumann avec multiplicité.

On note Γ un graphe fini, S l'ensemble de ses sommets et A l'ensemble de ses arêtes. On se donne une métrique sur ce graphe en associant à chaque arête $a_i \in A$, une longueur $l_i > 0$. Le laplacien combinatoire sur Γ est l'opérateur agissant sur les fonctions $S \to \mathbb{R}$ induit par la forme quadratique

$q(f) = \sum_{a_i \in A} l_i \, \mathrm{d}_{a_i}(f)^2$, avec $\mathrm{d}_a(f) = (f(x) - f(y))/l_i$, x et y étant les extrémités de l'arête a_i. L'espace des fonctions $\mathbb{R}^S$ sur les sommets est muni de sa structure euclidienne canonique.

On utilise alors le résultat suivant :

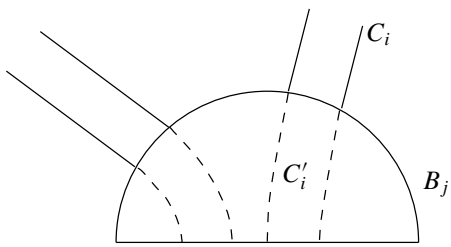
Théorème 4.2 [Colin de Verdière 1988, §4]. *Étant donné une suite $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$, il existe une métrique sur le graphe complet à $N+1$ sommets dont le spectre est la suite (λ_i). De plus, ce spectre vérifie la propriété de transversalité d'Arnol'd.*

Il reste à construire une famille de variétés dont le début du spectre ressemble à celui d'un graphe complet. Ces variétés, qu'on notera Ω_ε, seront localement des domaines euclidiens. Le graphe Γ sera plongeable isométriquement dans Ω_ε, les sommets étant situés sur le bord de Steklov de la variété (construire Ω_ε comme étant globalement un domaine euclidien nécessiterait d'imposer des contraintes sur les longueurs l_i, ce qu'on veut éviter).

Plus précisément, deux réels $c > 0$ et ε étant fixé, on se donne pour chaque sommet $s_j \in S$ une demi-boule B_j de rayon $c\varepsilon$ (dans la suite, utilisera l'indice j pour les sommets du graphe et on réservera l'indice i pour les arêtes). La constante c sera fixée plus loin. Pour chaque arête $a_i \in A$, on se donne ensuite un cylindre C_i de rayon ε et de longueur $l_i - 2c\varepsilon$. Si on note j et j' les indices des sommets extrémités de l'arête a_i, et qu'on place les boules B^{n-1} qui bordent le cylindre C_i tangentiellement aux demi-boules B_j et $B_{j'}$, on peut plonger isométriquement l'arête a_i dans la réunion de C_i, B_j et $B_{j'}$,en identifiant les sommets s_j et $s_{j'}$ avec les centres de B_j et $B_{j'}$. En répétant le procédé pour chaque arête, le graphe Γ se plonge isométriquement dans la réunion des demi-boules B_j et des cylindres C_i. Pour construire le domaine Ω_ε, on va prolonger le cylindre C_i en un cylindre C_i' dont les extrémités seront dans les demi-boules. Chaque boule B_j étant vue localement comme un domaine euclidien, on construit une application de $C_i' = B^{n-1} \times [0, l_i]$ dans la réunion des C_i et des voisinages des B_j telle que :

- $B^{n-1} \times [c\varepsilon, l_i - c\varepsilon]$ est envoyé isométriquement sur C_i ;
- pour $t \in [0, c\varepsilon]$ et $[l_i - c\varepsilon, l_i]$, chaque boule $B^{n-1} \times \{t\}$ est plongée isométriquement dans le voisinage de la demi-boule B_j correspondante ;
- $B^{n-1} \times \{0\}$ et $B^{n-1} \times \{1\}$ sont envoyé sur le bord équatorial de B_j ;
- l'application obtenue est 1-lipschitzienne.

Les extrémités des cylindres ne sont donc pas isométriques à la métrique produit mais légèrement tordus à l'intérieur des boules :



En outre, on fait en sorte que les images de chaque C_i' soient disjointes (on choisit c assez grand pour que ça soit possible). Ces précisions techniques faciliteront l'étude du spectre. On peut remarquer que quand ε tend vers 0, Ω_ε tend vers le graphe Γ pour la distance de Gromov–Hausdorff.

En considérant la condition de Steklov sur les boules équatoriales des demi-boules B_j et la condition de Neumann sur le reste du bord de Ω_ε, on va montrer que le début du spectre de Ω_ε tend vers le spectre du graphe, à une constante multiplicative près :

Théorème 4.3. *Le N-écart spectral entre le spectre de Steklov–Neumann de Ω_ε et le spectre de $c^{n-1}\Delta_\Gamma$, où Δ_Γ désigne le laplacien combinatoire sur Γ, tend vers 0 quand ε tend vers 0.*

Démonstration. La démonstration se déroule en deux étapes. D'abord, on décompose l'espace des fonctions harmoniques (pour le problème de Steklov–Neumann) sur Ω_ε en deux sous-espaces pour appliquer le lemme 3.1, puis on montre la convergence en se restreignant à l'un des sous-espaces. On note $\partial\Omega_{\varepsilon,S}$ le bord de Steklov de Ω_ε, $\partial\Omega^j_{\varepsilon,S}$, $1 \le j \le N$ ses composantes connexes et $\mathcal{H}$ l'espace des fonctions harmoniques sur Ω_ε vérifiant la condition de Neumann sur $\partial\Omega_\varepsilon \backslash \partial\Omega_{\varepsilon,S}$.

Étape 1. On définit l'espace $\mathcal{H}_0$ comme étant l'espace des fonctions harmoniques de Ω_ε constantes sur chacune des N composantes connexes du bord de Steklov. L'orthogonal de $\mathcal{H}_0$ pour la forme quadratique Q associée à l'opérateur Λ contient les fonction constantes, qui sont aussi dans $\mathcal{H}_0$. On définit donc $\mathcal{H}_\infty$, comme l'espace des fonctions Q-orthogonales à $\mathcal{H}_0$ et d'intégrale nulle sur $\partial\Omega_{\varepsilon,S}$. Si $f \in \mathcal{H}_0$ et si g est Q-orthogonale à H_0, alors on a, en notant f_j la valeur de f sur $\partial\Omega^j_{\varepsilon,S}$:

$$(f, \Lambda g) = \int_{\partial\Omega_{\varepsilon,S}} f \frac{\partial g}{\partial \nu} = \sum_j f_j \int_{\partial\Omega^j_{\varepsilon,S}} \frac{\partial g}{\partial \nu}.$$

Comme $(f, \Lambda g)$ est nul pour tout $f \in \mathcal{H}_0$, on en déduit que $\displaystyle\int_{\partial\Omega^j_{\varepsilon,S}} \frac{\partial g}{\partial \nu} = 0$ pour tout j. On a donc

$$\mathcal{H}_\infty = \left\{ f \in \mathcal{H} : \int_{\partial\Omega_{\varepsilon,S}} f = 0,\ \int_{\partial\Omega^j_{\varepsilon,S}} \frac{\partial f}{\partial \nu} = 0 \text{ pour tout } j \right\}. \tag{4.4}$$

On doit minorer la forme quadratique Q sur l'espace $\mathcal{H}_\infty$. Pour cela, on va passer par l'intermédiaire du domaine $D = \bigcup_j B_j$. Mais comme la restriction des formes harmoniques de Ω_ε ne vérifient pas la condition de Neumann sur les hémisphères qui bordent les B_j on va d'abord reformuler l'expression du bas du spectre de Q sur $\mathcal{H}_\infty$. On pose, pour toute fonction $\tilde{f} \in C^\infty(F)$ telle que $\int_{\partial\Omega_{\varepsilon,S}} \tilde{f} = 0$,

$$\tilde{Q}(\tilde{f}) = \inf\left\{ \int_{\Omega_\varepsilon} |\mathrm{d}f|^2 : f|_{\partial\Omega_{\varepsilon,S}} = \tilde{f},\ \int_{\partial\Omega^j_{\varepsilon,S}} \frac{\partial f}{\partial \nu} = 0 \right\} \tag{4.5}$$

On peut vérifier que la borne inférieure de $\tilde{Q}$ (pour $\|\tilde{f}\|_2 = 1$) coïncide bien avec le bas du spectre de Q sur $\mathcal{H}_\infty$.

On définit les espaces $\mathcal{H}^D$, $\mathcal{H}^D_\infty$ et les formes quadratiques Q^D, $\tilde{Q}^D$ en remplaçant Ω_ε par D dans les définitions de $\mathcal{H}$, $\mathcal{H}_\infty$, Q et $\tilde{Q}$. Comme $D \subset \Omega_\varepsilon$, $\tilde{Q}$ est minoré par la forme quadratique $\tilde{Q}^D$. La première valeur propre de $Q_{\mathcal{H}_\infty}$ est donc minorée par la première valeur propre de $Q^D_{\mathcal{H}_\infty}$. Le domaine D possède N composantes connexes, donc la multiplicité de 0 dans spectre de Steklov de D est N, les fonctions propres étant les fonctions constantes sur chaque B_j. La première valeur propre de $Q^D_{\mathcal{H}_\infty}$ est donc la $(N+1)$-ième valeur propre de D, qui est la première valeur propre non nulle $\sigma_1(B(\varepsilon))$ de la demi-boule de rayon $c\varepsilon$. Cette valeur propre se comporte comme ε^{-2} quand $\varepsilon \to 0$, ce qui permet d'appliquer que lemme 3.1.

Étape 2. On doit maintenant comparer les spectres de $Q_{\mathscr{H}_0}$ et de la forme quadratique q associée au laplacien combinatoire sur Γ. Les deux domaines des formes quadratiques sont en bijection de manière évidente, en identifiant une fonction sur les sommets s_j de Γ avec une fonction constante sur chaque F_j, prenant les mêmes valeurs. Les normes sur les deux espaces sont différentes. On notera $|\cdot|_\Gamma$ la norme euclidienne canonique sur $\mathbb{R}^S$, la norme sur $\mathscr{H}_0$ est alors $|\cdot| = (c\varepsilon)^{n-1}\omega_{n-1}|\cdot|_\Gamma$, où ω_{n-1} désigne le volume de la boule euclidienne canonique de dimension $n-1$.

Étant donnée f est une fonction sur S, on construit une fonction test $\bar{f}$ sur Ω_ε prenant les mêmes valeurs que f sur chaque F_j, constante sur chaque demi-boule B_j et prolongée de manière affine sur les cylindres C_i constituant le domaine Ω_ε. On a

$$Q(f) \leq Q(\bar{f}) = \int_{\Omega_\varepsilon} |\mathrm{d}\bar{f}|^2 = \varepsilon^{n-1}\omega_{n-1} \sum_{a_i \in A} \frac{1}{l_i - 2c\varepsilon}(f(x_i) - f(y_i))^2,$$

où x_i et x_i sont les extrémités de l'arête a_i, donc $\limsup_{\varepsilon\to 0} Q(f)/|f|^2 \leq c^{n-1}q(f)/|f|_\Gamma^2$.

Réciproquement, étant donné une fonction $f \in \mathscr{H}_0$, on construit une fonction test sur les arêtes du graphe Γ. À partir de la donnée de f sur l'image d'un cylindre $C_i' = B^{n-1} \times [0, \varepsilon, l_i]$, on définit une fonction $\tilde{f}$ sur l'intervalle $[0, l_i]$ par moyennation sur chaque boule B^{n-1}, c'est-à-dire que

$$\tilde{f}(t) = \frac{1}{\varepsilon^{n-1}\omega_{n-1}} \int_{B^{n-1}} f(x,t)\,\mathrm{d}x.$$

On a alors, en utilisant le fait que plongement de C_i' dans Ω_ε est 1-lipschitzien,

$$|\mathrm{d}\tilde{f}|^2 \leq \frac{1}{\varepsilon^{n-1}\omega_{n-1}} \int_{B^{n-1}} |\mathrm{d}f|^2\,\mathrm{d}x. \tag{4.6}$$

On obtient ainsi une fonction $\tilde{f}$ sur Γ qui est C^1, qui coïncide avec f sur les sommets et qui vérifie $\varepsilon^{n-1}\omega_{n-1}\|\mathrm{d}\tilde{f}\|^2 \leq \int_{\Omega_\varepsilon} |\mathrm{d}f|^2 = Q(f)$. Comme sur le graphe, on a $\|\mathrm{d}\tilde{f}\|^2 \geq q(f)$, on obtient que

$$\liminf_{\varepsilon\to 0} \frac{Q(f)}{|f|^2} \geq \frac{c^{n-1}q(f)}{|f|_\Gamma^2}.$$

On a finalement montré que $Q(f)/|f|^2$ converge simplement vers $c^{n-1}q(f)/|f|_\Gamma^2$ quand $\varepsilon \to 0$. Comme on travaille sur des espaces de dimension finie, cela suffit pour assurer la convergence du spectre et des espaces propres des deux opérateurs. □

4C. ***Application à la prescription de spectre.*** On a maintenant tous les ingrédients pour montrer le théorème 1.2. La méthode la plus directe serait d'utiliser le théorème 3.12 de convergence du spectre vers celui du bord et les résultats de prescription obtenus dans [Colin de Verdière 1987] (on peut les adapter de manière à prescrire la classe conforme). Cependant, cette méthode ne fonctionne que si la dimension du bord est au moins 3. On va donc procéder autrement en utilisant les plongements de graphes construits au paragraphe précédent.

Démonstration du théorème 1.2. D'après le théorème 4.2, il existe un graphe complet Γ ayant le spectre voulu, avec la propriété de stabilité. On va transplanter ce spectre dans la variété M en commençant par traiter le cas $\rho = 1$.

On commence par déformer la variété M en respectant la classe conforme et de manière à pouvoir plonger isométriquement le graphe Γ dans M en plaçant les sommets sur ∂M. Comme la dimension de M est plus grande que 3, on peut le faire sans que les arêtes se croisent. On note g la métrique obtenue sur M.

A priori, la métrique au voisinage du plongement de Γ n'est pas euclidienne, on ne peut donc pas plonger isométriquement un ouvert Ω_ε (construit au paragraphe précédent) au voisinage de Γ. Cependant, pour tout ε on peut déformer (de manière non conforme) la métrique g en une métrique g_ε telle que les graphes soient toujours plongés isométriquement et que Γ possède un voisinage isométrique au domaine Ω_ε. On peut de plus faire en sorte que $(1/\tau_\varepsilon)g_\varepsilon \leq g \leq \tau_\varepsilon g_\varepsilon$, pour une famille de réels $\tau_\varepsilon > 1$ telle que $\tau_\varepsilon \to 1$ quand $\varepsilon \to 0$.

On peut maintenant appliquer les résultats de convergence spectrale de la section précédente. Pour un $\delta > 0$ petit donné, on peut trouver un ε tel que le N-écart spectral entre Γ et Ω_ε soit inférieur à δ. En utilisant les arguments de la démonstration du théorème 3.12 (étape 2), on peut choisir ε suffisamment petit pour que le N-écart spectral entre $(\Omega_\varepsilon, g_\varepsilon)$ et (Ω_ε, g) soit inférieur à δ. Enfin, on peut faire converger le spectre de M vers celui de (Ω_ε, g) d'après le théorème 3.8, et en particulier déformer g de manière conforme de sorte que le N-écart spectral entre M et (Ω_ε, g) soit lui aussi inférieur à δ. On peut donc rentre le N-écart spectral entre M et Γ arbitrairement petit, et ce de manière conforme.

Traitons maintenant le cas où ρ varie. Quitte à multiplier les a_i par une constante, on peut supposer que $\rho \leq 1$. Il suffit d'ajouter une étape à la construction précédente et d'utiliser le théorème 3.6 pour faire tendre le spectre de la variété (M, ρ) vers celui de M sans densité. □

5. Multiplicité en dimension 2

5A. ***Lignes nodales des fonctions propres.*** On va montrer dans cette section les obstructions à la prescription de multiplicité en dimension 2 (théorèmes 1.5, 1.7 et 1.10). Comme dans le cas du laplacien, les deux principaux ingrédients sont le théorème nodal de Courant et le théorème de Cheng sur la structure local de l'ensemble nodal. Dans toute la suite du texte, les fonctions propres considérées seront les fonctions harmoniques sur M et pas leur restriction à ∂M. En particulier, les lignes et les domaines nodaux seront considérés sur M.

Avec ces précisions, le théorème nodal de Courant est valide pour les problèmes de Steklov et Steklov–Neumann, quelle que soit la dimension :

Théorème 5.1. *Le nombre de domaines nodaux de la k-ième fonction propre du problème de Steklov* (*ou de Steklov–Neumann*) *est au plus égal à $k + 1$.*

La démonstration (essentiellement la même que dans le cas du laplacien) est donnée dans [Kuttler et Sigillito 1969] pour la dimension 2, et elle se généralise immédiatement en toute dimension.

Contrairement aux fonctions propres du laplacien, les fonctions propres de Steklov ont la particularité que leurs domaines nodaux rencontrent toujours le bord. Cette propriété a déjà été utilisée, par exemple, dans [Bañuelos et al. 2010], et nous y feront appel pour démontrer le théorème 1.7 :

Lemme 5.2. *Tout domaine nodal rencontre le bord de la variété. Dans le cas du problème de Steklov–Neumann, tout domaine nodal rencontre le bord de Steklov.*

Démonstration. Soit f une fonction harmonique non nulle et D un domaine nodal de f ne rencontrant pas le bord de Steklov de la variété. Comme f est harmonique et nulle sur le bord de D (ou vérifie la condition de Neumann le long du bord de Neumann de la variété), elle est uniformément nulle dans D. Par conséquent, elle est nulle partout. □

S. Y. Cheng [1976] a décrit la structure locale de l'ensemble nodal des fonctions propres du laplacien en dimension 2. On peut les appliquer aux fonctions harmoniques, et le lemme qui précède permet de préciser certaines propriétés topologiques des domaines et des lignes nodales, en particulier leur incompressibilité (une partie d'une surface est dite incompressible si son groupe fondamental s'injecte dans celui de la surface). L'énoncé qui suit rassemble ces résultats :

Théorème 5.3. *Supposons que M est de dimension 2, et soit f une fonction propre du problème de Steklov. Alors* :

(1) *Les domaines nodaux de f sont incompressibles.*

(2) *L'ensemble nodal de f intérieur à M est la réunion d'un nombre fini de courbes C^2 qui sont soit des cercles immergés, soit des arcs immergés dont les extrémités sont sur ∂M.*

(3) *La réunion de ces courbes forme un graphe fini dont les composantes connexes sont incompressibles.*

(4) *Soit p un point intérieur à M. Si p est un point critique de f situé sur l'ensemble nodal et que l'ordre d'annulation de f en p est k, alors au voisinage de p l'ensemble nodal est la réunion de k courbes s'intersectant en p, de courbure géodésique nulle en p et formant un système équiangulaire (en particulier, les sommets du graphe nodal intérieur à M sont de degré pair).*

(5) *Tout point du bord où f s'annule est l'extrémité d'une ligne nodale intérieure à M.*

(6) *Chaque composante connexe du bord contient un nombre pair d'extrémités du graphe nodal.*

(7) *Dans le cas du problème de Steklov–Neumann, si p est un point du bord de Neumann où f s'annule, l'ordre d'annulation k de f en p est fini et le point p est un zéro isolé en restriction à ∂M_N. Au voisinage de p dans M, l'ensemble nodal est la réunion de k arcs partant de p, de courbure géodésique nulle en p et dont l'extension par réflexion par rapport au bord forme un système équiangulaire.*

On appliquera en particulier les propriétés d'incompressibilité au cas du disque. On obtient alors :

Corollaire 5.4. *Si M est homéomorphe à un disque, alors les domaines nodaux sont homéomorphes à des disques et les composantes connexes du graphe nodal sont des arbres.*

Démonstration du théorème 5.3. Comme $\Delta f = 0$ dans l'intérieur de M, on peut appliquer le théorème 2.5 de [Cheng 1976]. En particulier, l'ensemble nodal est la réunion de courbes immergées (les lignes nodales) qui sont localement en nombre fini, cette réunion étant homéomorphe à un graphe localement fini. Cependant, comme l'intérieur de la surface n'est pas compact, on doit vérifier la finitude globale du graphe nodal, qui découle des deux points suivants :

(i) le nombre de lignes nodales est fini ;

(ii) les points d'intersection des lignes nodales sont en nombre fini.

Le point (i) se déduit du théorème de Courant : le nombre de domaines délimités par un ensemble de lignes nodales est au moins égal au nombre de ces lignes ; par conséquent le nombre total de lignes nodales est majoré par le nombre de domaines nodaux, en particulier il est fini. On montre le point (ii) à l'aide de la formule d'Euler–Poincaré appliquée à la surface : comme les domaines nodaux sont en nombre fini et que leur caractéristique d'Euler est majorée par 1, la caractéristique d'Euler du graphe nodal est minorée en fonction de la topologie de la surface et du nombre de domaines. Or, les sommets du graphe sont de deux types : d'une part les sommets situés sur le bord de la surface aux extrémités des lignes nodales, qui sont en nombre fini et de degré fini car il n'y a qu'un nombre fini de les lignes nodales ; d'autre part les intersections de lignes, qui sont de degré au moins 4. Si les sommets intérieurs sont en nombre infini, la caractéristique d'Euler du graphe serait donc $-\infty$, ce qui contredit la formule d'Euler–Poincaré.

On en déduit de ce qui précède les points (2) et (4) du théorème et le fait que le graphe nodal est fini.

Soit D un domaine nodal et γ une courbe de D non contractile dans D. Si γ est contractile dans M, alors elle entoure un domaine nodal D' distinct de D. En outre, γ sépare D' de ∂M, ce qui contredit le lemme 5.2. Par conséquent, D est incompressible. Le même argument montre l'incompressibilité du graphe nodal. On obtient ainsi les points (1) et (3).

Montrons le point (5). Supposons que p est un point du bord qui n'est pas l'extrémité d'une ligne nodale. Le point p n'est donc pas situé à la frontière entre deux domaines nodaux, il est contenu dans un domaine nodal D sur lequel on supposera que f est positive. D'après la propriété d'unique prolongement (théorème 2.7), il n'y a pas de ligne nodale le long du bord, on peut donc trouver un petit voisinage U de p délimité par une courbe de niveau $f(x) = \varepsilon$ avec $\varepsilon > 0$ petit. En restriction à D, f est la première fonction propre du problème de Steklov–Dirichlet avec condition de Dirichlet sur les ligne nodales qui bordent D à l'intérieur de M. Or, si on définit la fonction test $\tilde{f}$ par $\tilde{f} = \varepsilon$ sur U et $\tilde{f} = f$ sur $D\backslash U$, le quotient de Rayleigh de $\tilde{f}$ est strictement plus petit que celui de f, ce qui contredit que f soit la première fonction propre sur D.

Le fait qu'un nombre pair de lignes nodales rejoigne chaque composante du bord découle du fait que le signe de la fonction propre change chaque fois qu'on traverse une ligne nodale.

Reste à traiter le cas du problème de Steklov–Neumann. On considère deux copies de la variété M qu'on recolle de manière symétrique le long du bord de Neumann et on note M' la surface obtenue. Comme le problème de Steklov est conformément invariant en dimension 2, on peut lisser la métrique le long du recollement de manière conforme et symétrique. Les fonctions propres sur M correspondent alors aux fonctions propres sur M' qui sont symétriques. On peut en particulier leur appliquer les résultats

de Cheng (point (4)). La symétrie de la fonction implique la symétrie des lignes nodales (sur M') au voisinage du bord de Neumann de M. On doit encore montrer que le bord de Neumann ne contient pas de ligne nodale : une fonction harmonique f sur M' est la partie réelle d'une fonction holomorphe g (en munissant M' de la structure complexe induite par la structure conforme). Si f est une fonction propre symétrique, alors la condition de Neumann et l'équation de Cauchy–Riemann implique que $\mathrm{Im}(g)$ est constante le long du bord de Neumann de M. On peut choisir g de sorte que cette constante soit nulle, les zéros de f sur le bord de Neumann sont donc les zéros d'une fonction holomorphe. Par conséquent ils sont isolés. □

Démonstration du corollaire 5.4. Si M est un disque, l'incompressibilité des domaines nodaux implique qu'ils sont simplement connexes, donc que ce sont des disques.

Les composantes connexes du graphe nodal sont planaires, et leur incompressibilité signifie qu'ils sont sans cycle. Donc ce sont des arbres. □

5B. ***Bornes sur la multiplicité.*** On peut maintenant démontrer les théorèmes 1.5, 1.7 et 1.10. En ce qui concerne le théorème 1.5, on reprendra les arguments de [Cheng 1976] et [Besson 1980], qui sont moins précis que ceux de [Nadirashvili 1987] mais plus faciles à adapter au problème de Steklov.

Démonstration du théorème 1.5. Supposons que la surface M soit orientable. On note E_k l'espace propre associé à la valeur propre $\sigma_k(M)$ et m_k sa multiplicité. Selon [Besson 1980], si $m_k > 4\gamma + 2k + 1$, il existe un point x dans l'intérieur de M et une fonction propre $f \in E_k$ telle que l'ordre d'annulation de f en x soit strictement supérieur à $2\gamma + k$. Localement, il existe donc au moins $4\gamma + 2k + 2$ arcs nodaux partant de p.

Si on « ferme » la surface en quotientant chaque composante du bord sur un point, tous les arcs nodaux se referment, et il existe donc au moins $2\gamma + k + 1$ lacets distincts C^1 par morceaux dans l'ensemble nodal. Or, Cheng [1976, Lemma 3.1] a montré que ces lacets décomposent la surface en au moins $k + 2$ composantes connexes. La fonction f possède donc au moins $k + 2$ domaines nodaux, ce qui contredit le théorème de Courant.

Comme dans [Besson 1980], le cas des surfaces non orientables se traite par passage à un revêtement à deux feuillets. La surface obtenue en quotientant les bords est de caractéristique d'Euler $p = (1 - \chi(M) - l)$. Les arguments des [ibid.] donnent alors la majoration $m_k \leq 4p + 4k + 3$. □

Démonstration du théorème 1.7. On note E l'espace propre associé à la valeur propre $\sigma_i(M, \rho, g)$ pour $i = 1$ ou 2.

Soit p un point intérieur au disque. Si E est de dimension au moins 4, il existe une fonction propre non nulle $f \in E$ telle que f et $\mathrm{d}f$ soient nuls en p. Le point p est donc un sommet du graphe nodal de f et il en part au moins quatre arêtes. Comme le graphe nodal est un arbre dont les feuilles sont sur le bord, il délimite au moins quatre domaines nodaux. Il y a donc contradiction avec le théorème de Courant.

Supposons maintenant que $i = 1$ et que E soit de dimension 3. Si p_0 est un point du bord, le sous-espace des fonctions $f \in E$ telles que $f(p_0) = 0$ est de dimension au moins 2. Pour tout point p du bord distinct de p_0, il existe donc une fonction f_p, qu'on supposera de norme 1, telle que $f(p) = f(p_0) = 0$. Comme

chacun de ces points est nécessairement l'extrémité d'un ligne nodale et que la fonction f_p a exactement deux domaines nodaux, p_0 et p sont les extrémités de l'unique ligne nodale de f_p.

Le bord est donc partagé en deux intervalles, I_p^+ et I_p^-, d'extrémités p_0 et p, sur lesquels la fonction f_p est respectivement positive et négative. En faisant tendre p vers p_0, on peut faire tendre la longueur de I_p^- vers 0. Comme les fonctions f_p sont normées et que E est de dimension finie, la famille f_p admet une limite f (quitte à extraire une sous-famille). La fonction f est alors positive ou nulle sur la totalité du bord, puisque l'évaluation en un point est une forme linéaire continue sur E. Par conséquent, f est de signe constant, ce qui est impossible puisque c'est une fonction propre de la valeur propre σ_1. □

Démonstration du théorème 1.10. On note I une composante connexe du bord de Neumann $\partial\mathbb{D}_N$ et on choisit $k+1$ points distincts $x_1, \dots, x_{k+1}$ dans I. Supposons que la multiplicité de $\sigma_k(\mathbb{D}, \partial\mathbb{D}_S, \rho, g)$ soit supérieure ou égale à $k+2$. On peut alors trouver une fonction propre f associée à cette valeur propre qui s'annule en tous les points x_i, $i \in [1, k+1]$. En vertu du point 6 du théorème 5.3, chaque x_i appartient à une composante connexe du graphe nodal qui joint x_i à un point du bord de Steklov, ces composantes étant distinctes. L'ensemble nodal sépare donc $\mathbb{D}$ en au moins $k+2$ composantes connexes, ce qui contredit le théorème de Courant. □

Bibliographie

[Agranovich 2006] M. S. Agranovich, "On a mixed Poincaré–Steklov type spectral problem in a Lipschitz domain", *Russ. J. Math. Phys.* **13**:3 (2006), 239–244. MR 2007j:35038 Zbl 1162.35351

[Alessandrini et Magnanini 1994] G. Alessandrini et R. Magnanini, "Elliptic equations in divergence form, geometric critical points of solutions, and Stekloff eigenfunctions", *SIAM J. Math. Anal.* **25**:5 (1994), 1259–1268. MR 95f:35180 Zbl 0809.35070

[Ammann 2003] B. Ammann, "A spin-conformal lower bound of the first positive Dirac eigenvalue", *Differential Geom. Appl.* **18**:1 (2003), 21–32. MR 2004e:58052 Zbl 1030.58020

[Bandle 1980] C. Bandle, *Isoperimetric inequalities and applications*, Monographs and Studies in Mathematics **7**, Pitman, Boston, MA, 1980. MR 81e:35095 Zbl 0436.35063

[Bañuelos et al. 2010] R. Bañuelos, T. Kulczycki, I. Polterovich et B. Siudeja, "Eigenvalue inequalities for mixed Steklov problems", pp. 19–34 dans *Operator theory and its applications*, édité par M. Levitin et D. Vassiliev, Amer. Math. Soc. Transl. Ser. 2 **231**, Amer. Math. Soc., Providence, RI, 2010. MR 2012e:35176 Zbl 1217.35127

[Besson 1980] G. Besson, "Sur la multiplicité de la première valeur propre des surfaces riemanniennes", *Ann. Inst. Fourier* (*Grenoble*) **30**:1 (1980), 109–128. MR 81h:58059 Zbl 0417.30033

[Cheng 1976] S. Y. Cheng, "Eigenfunctions and nodal sets", *Comment. Math. Helv.* **51**:1 (1976), 43–55. MR 53 #1661 Zbl 0334.35022

[Colbois et al. 2011] B. Colbois, A. El Soufi et A. Girouard, "Isoperimetric control of the Steklov spectrum", *J. Funct. Anal.* **261**:5 (2011), 1384–1399. MR 2012m:35328 Zbl 1235.58020

[Colin de Verdière 1986] Y. Colin de Verdière, "Sur la multiplicité de la première valeur propre non nulle du laplacien", *Comment. Math. Helv.* **61**:2 (1986), 254–270. MR 88b:58140 Zbl 0607.53028

[Colin de Verdière 1987] Y. Colin de Verdière, "Construction de laplaciens dont une partie finie du spectre est donnée", *Ann. Sci. École Norm. Sup.* (4) **20**:4 (1987), 599–615. MR 90d:58156 Zbl 0636.58036

[Colin de Verdière 1988] Y. Colin de Verdière, "Sur une hypothèse de transversalité d'Arnol′d", *Comment. Math. Helv.* **63**:2 (1988), 184–193. MR 90c:58183 Zbl 0672.58046

[Dahl 2005] M. Dahl, "Prescribing eigenvalues of the Dirac operator", *Manuscripta Math.* **118**:2 (2005), 191–199. MR 2006h:58037 Zbl 1081.58021

[El Soufi et Ilias 1986] A. El Soufi et S. Ilias, "Immersions minimales, première valeur propre du laplacien et volume conforme", *Math. Ann.* **275**:2 (1986), 257–267. MR 87j:53088 Zbl 0675.53045

[Fraser et Schoen 2011] A. Fraser et R. Schoen, "The first Steklov eigenvalue, conformal geometry, and minimal surfaces", *Adv. Math.* **226**:5 (2011), 4011–4030. MR 2012f:58054 Zbl 1215.53052

[Fraser et Schoen 2012] A. Fraser et R. Schoen, "Eigenvalue bounds and minimal surfaces in the ball", preprint, 2012. arXiv 1209.3789v1

[Hassannezhad 2011] A. Hassannezhad, "Conformal upper bounds for the eigenvalues of the Laplacian and Steklov problem", *J. Funct. Anal.* **261**:12 (2011), 3419–3436. MR 2012i:58024 Zbl 1232.58023

[Hoffmann-Ostenhof et al. 1999] T. Hoffmann-Ostenhof, P. W. Michor et N. Nadirashvili, "Bounds on the multiplicity of eigenvalues for fixed membranes", *Geom. Funct. Anal.* **9**:6 (1999), 1169–1188. MR 2001i:35066 Zbl 0949.35102

[Jammes 2007] P. Jammes, "Minoration conforme du spectre du laplacien de Hodge-de Rham", *Manuscripta Math.* **123**:1 (2007), 15–23. MR 2008a:58029 Zbl 1127.35027

[Jammes 2008] P. Jammes, "Prescription du spectre du laplacien de Hodge-de Rham dans une classe conforme", *Comment. Math. Helv.* **83**:3 (2008), 521–537. MR 2009f:58046 Zbl 1155.58305

[Jammes 2009] P. Jammes, "Sur la multiplicité des valeurs propres d'une variété compacte", pp. 1–11 dans *Actes du Séminaire de Théorie Spectrale et Géométrie*, Sémin. Théor. Spectr. Géom. **26**, Institut Fourier, St. Martin-d'Héres, 2009. MR 2011d:58075 Zbl 1235.58022

[Jammes 2011] P. Jammes, "Prescription de la multiplicité des valeurs propres du laplacien de Hodge–de Rham", *Comment. Math. Helv.* **86**:4 (2011), 967–984. MR 2851874 Zbl 1248.58016

[Jammes 2012] P. Jammes, "Sur la multiplicité des valeurs propres du laplacien de Witten", *Trans. Amer. Math. Soc.* **364**:6 (2012), 2825–2845. MR 2888230 Zbl 1242.58015

[Karpukhin et al. 2012] M. Karpukhin, G. Kokarev et I. Polterovich, "Multiplicity bounds for Steklov eigenvalues on Riemannian surfaces", preprint, 2012. arXiv 1209.4869

[Kopachevsky et Krein 2001] N. D. Kopachevsky et S. G. Krein, *Operator approach to linear problems of hydrodynamics, I: Self-adjoint problems for an ideal fluid*, Operator Theory : Advances and Applications **128**, Birkhäuser, Basel, 2001. MR 2003e:76003 Zbl 0979.76002

[Kozlov et al. 2004] V. Kozlov, N. Kuznetsov et O. Motygin, "On the two-dimensional sloshing problem", *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **460**:2049 (2004), 2587–2603. MR 2005d:76004 Zbl 1068.35107

[Kuttler et Sigillito 1969] J. R. Kuttler et V. G. Sigillito, "An inequality of a Stekloff eigenvalue by the method of defect", *Proc. Amer. Math. Soc.* **20** (1969), 357–360. MR 38 #3633 Zbl 0176.09901

[Nadirashvili 1987] N. S. Nadirashvili, "Multiple eigenvalues of the Laplace operator", *Mat. Sb.* (*N.S.*) **133(175)**:2 (1987), 223–237. In Russian ; translated in *Math. USSR, Sb.* **61** :1 (1988), 225–238. MR 89a:58113 Zbl 0672.35049

[Stekloff 1899] W. Stekloff, "Sur l'existence des fonctions fondamentales", *C. R. Acad. Sci. Paris* **128**:1 (1899), 808–810. JFM 30.0374.03

[Stekloff 1902] W. Stekloff, "Sur les problèmes fondamentaux de la physique mathématique (suite et fin)", *Ann. Sci. École Norm. Sup.* (3) **19** (1902), 455–490. MR 1509018 JFM 33.0800.01

[Taylor 1996a] M. E. Taylor, *Partial differential equations, I: Basic theory*, Applied Mathematical Sciences **115**, Springer, New York, 1996. MR 98b:35002b Zbl 0869.35002

[Taylor 1996b] M. E. Taylor, *Partial differential equations, II: Qualitative studies of linear equations*, Applied Mathematical Sciences **116**, Springer, New York, 1996. MR 98b:35003 Zbl 0869.35003

Received 25 Sep 2012. Revised 20 Oct 2013. Accepted 13 Nov 2013.

PIERRE JAMMES: pjammes@unice.fr
Laboratoire J. A. Dieudonné, UMR no. 7351 CNRS UNS, Université de Nice Sophia Antipolis, 06108 Nice Cedex 02, France
et
Département de Mathématiques, Université de Nice Sophia Antipolis, Parc Valrose, 06108 Nice Cedex 02, France

ANALYSIS AND PDE
Vol. 7, No. 3, 2014
dx.doi.org/10.2140/apde.2014.7.551

SEMILINEAR GEOMETRIC OPTICS WITH BOUNDARY AMPLIFICATION

JEAN-FRANCOIS COULOMBEL, OLIVIER GUÈS AND MARK WILLIAMS

We study weakly stable semilinear hyperbolic boundary value problems with highly oscillatory data. Here weak stability means that exponentially growing modes are absent, but the so-called uniform Lopatinskii condition fails at some boundary frequency β in the hyperbolic region. As a consequence of this degeneracy there is an amplification phenomenon: outgoing waves of amplitude $O(\varepsilon^2)$ and wavelength ε give rise to reflected waves of amplitude $O(\varepsilon)$, so the overall solution has amplitude $O(\varepsilon)$. Moreover, the reflecting waves emanate from a radiating wave that propagates in the boundary along a characteristic of the Lopatinskii determinant.

An approximate solution that displays the qualitative behavior just described is constructed by solving suitable profile equations that exhibit a loss of derivatives, so we solve the profile equations by a Nash–Moser iteration. The exact solution is constructed by solving an associated singular problem involving singular derivatives of the form $\partial_{x'} + \beta\partial_{\theta_0}/\varepsilon$, x' being the tangential variables with respect to the boundary. Tame estimates for the linearization of that problem are proved using a first-order (wavetrain) calculus of singular pseudodifferential operators constructed in a companion article ("Singular pseudodifferential calculus for wavetrains and pulses", arXiv 1201.6202, 2012). These estimates exhibit a loss of one singular derivative and force us to construct the exact solution by a separate Nash–Moser iteration.

The same estimates are used in the error analysis, which shows that the exact and approximate solutions are close in L^∞ on a fixed time interval independent of the (small) wavelength ε. The approach using singular systems allows us to avoid constructing high-order expansions and making small divisor assumptions. Our analysis of the exact singular system applies with no change to the case of pulses, provided one substitutes the pulse calculus from the companion paper for the wavetrain calculus.

1. Introduction and main results 552
2. Exact oscillatory solutions on a fixed time interval 573
3. Profile equations 591
4. Error analysis 599
5. Nash–Moser schemes 603
Appendix A. A calculus of singular pseudodifferential operators 615
Appendix B. An example derived from the Euler equations 619
References 624

Coulombel and Guès were supported by the French Agence Nationale de la Recherche, contract ANR-08-JCJC-0132-01. Williams was partially supported by NSF grants number DMS-0701201 and DMS-1001616.
MSC2010: 35L50.
Keywords: hyperbolic systems, boundary conditions, weak stability, geometric optics.

1. Introduction and main results

In this paper we study weakly stable semilinear hyperbolic boundary value problems with oscillatory data. The problems are weakly stable in the sense that exponentially growing modes are absent, but the uniform Lopatinskii condition fails at a boundary frequency β in the hyperbolic region $\mathcal{H}$.[1] As a consequence of this degeneracy in the boundary conditions, there is an amplification phenomenon: boundary data of wavelength ε and amplitude $O(\varepsilon^2)$ in problem (1-1) below gives rise to a response of amplitude $O(\varepsilon)$. In the meantime, resonance may occur between distinct oscillations. In the situation studied below, a resonant quadratic interaction between two incoming waves of amplitude $O(\varepsilon)$ may produce an outgoing wave of amplitude $O(\varepsilon^2)$. When reflected and amplified on the boundary, this oscillation gives rise to incoming waves of amplitude $O(\varepsilon)$. Hence the $O(\varepsilon)$ amplitude regime appears as the natural weakly nonlinear regime.

Let us now introduce some notation. On $\overline{\mathbb{R}}^{d+1}_+ = \{x = (x', x_d) = (t, y, x_d) = (t, x'') : x_d \geq 0\}$, consider the $N \times N$ semilinear hyperbolic boundary problem for $v = v_\varepsilon(x)$, where $\varepsilon > 0$:[2]

$$\begin{aligned} &\text{(a)} \quad L_0(\partial)v + f_0(v) = 0, \\ &\text{(b)} \quad \phi(v) = \varepsilon^2 G\left(x', \frac{x' \cdot \beta}{\varepsilon}\right) \quad \text{on } x_d = 0, \\ &\text{(c)} \quad v = 0 \text{ and } G = 0 \quad\quad \text{in } t < 0, \end{aligned} \tag{1-1}$$

where $L_0(\partial) = \partial_t + \sum_{j=1}^d B_j \partial_j$, the matrix B_d is invertible, and both $f_0(v)$ and $\phi(v)$ vanish at $v = 0$. The function $G(x', \theta_0)$ is assumed to be periodic in θ_0, and the frequency $\beta \in \mathbb{R}^d \setminus \{0\}$ is taken to be a boundary frequency at which the so-called uniform Lopatinskii condition fails. A consequence of this failure is that the choice of the factor ε^2 in (1-1)(b) corresponds to the weakly nonlinear regime for this problem. The leading profile is nonlinearly coupled to the next-order profile in the nonlinear system (1-35)–(1-36) derived below. We also refer to Appendix B for a detailed specific example which illustrates the nonlinear feature of the leading profile equation.

Before proceeding, we write the problem in an equivalent form that is better adapted to the boundary. After multiplying (1-1)(a) by $(B_d)^{-1}$, we obtain

$$\begin{aligned} &L(\partial)v + f(v) = 0, \\ &\phi(v) = \varepsilon^2 G\left(x', \frac{x' \cdot \beta}{\varepsilon}\right) \quad \text{on } x_d = 0, \\ &v = 0 \text{ and } G = 0 \quad\quad \text{in } t < 0, \end{aligned} \tag{1-2}$$

where we have set

$$L(\partial) = \partial_d + \sum_{j=0}^{d-1} A_j \partial_j \quad \text{with } A_j := B_d^{-1} B_j \text{ for } j = 0, \ldots, d-1.$$

[1] See Definition 1.4 and Assumption 1.6 for precise statements.

[2] We usually suppress the subscript ε.

Setting $v = \varepsilon u$ and writing $f(v) = D(v)v$, $\phi(v) = \psi(v)v$, we get the problem for $u = u_\varepsilon(x)$

$$\begin{aligned} &\text{(a)} \quad L(\partial)u + D(\varepsilon u)u = 0, \\ &\text{(b)} \quad \psi(\varepsilon u)u = \varepsilon G\left(x', \frac{x'\cdot\beta}{\varepsilon}\right) \quad \text{on } x_d = 0, \\ &\text{(c)} \quad u = 0 \qquad\qquad\qquad\quad \text{in } t < 0. \end{aligned} \tag{1-3}$$

For problem (1-3) we pose the two basic questions of rigorous nonlinear geometric optics:

(1) Does an exact solution u_ε of (1-3) exist for $\varepsilon \in (0, 1]$ on a fixed time interval $[0, T_0]$ independent of ε?

(2) Suppose the answer to the first question is yes. If we let $u_\varepsilon^{\text{app}}$ denote an approximate solution on $[0, T_0]$ constructed by the methods of nonlinear geometric optics (that is, solving eikonal equations for phases and suitable transport equations for profiles), how well does $u_\varepsilon^{\text{app}}$ approximate u_ε for ε small? For example, is it true that[3]

$$\lim_{\varepsilon\to 0} |u_\varepsilon - u_\varepsilon^{\text{app}}|_{L^\infty} \to 0? \tag{1-4}$$

The amplification phenomenon was studied in a formal way for several different *quasilinear* problems [Artola and Majda 1987; Majda and Artola 1988; Majda and Rosales 1983]. The last of these papers studied amplification in connection with Mach stem formation in reacting shock fronts, while [Artola and Majda 1987] explored a connection to the formation of instabilities in compressible vortex sheets. Both papers derived equations for profiles using an ansatz that exhibited amplification; however, neither of the two questions posed above were addressed. The first rigorous amplification results were proved in [Coulombel and Guès 2010] for *linear* problems. That article provided positive answers to the above questions (question (1) is trivial for linear problems) by making use of approximate solutions of high-order, and showed in particular that the limit (1-4) holds.

In this paper we give positive answers to the above questions for the *semilinear* system (1-3). As is typical in nonlinear geometric optics problems involving several phases, difficulties with small divisors rule out the construction of high-order approximate solutions.[4] Instead of constructing the exact solution u_ε as a small perturbation of a high-order approximate solution, we construct u_ε in the form

$$u_\varepsilon(x) = U_\varepsilon(x, \theta_0)|_{\theta_0 = \beta\cdot x'/\varepsilon},$$

where $U_\varepsilon(x, \theta_0)$ is an exact solution of the singular system (1-18). The singular system is solved using symmetrization and diagonalization arguments [Williams 2002], modified and supplemented with methods [Coulombel 2004] for deriving linear estimates for weakly stable hyperbolic boundary problems. In deriving the basic estimate (2-4) for the singular linear problem, a loss of derivatives[5] forces us to use a

[3] Let us observe that by the amplification phenomenon, we expect the solution v to (1-1) to have amplitude $O(\varepsilon)$, so the solution u to (1-3) should have amplitude $O(1)$. Hence the limit (1-4) deals with the difference between two $O(1)$ quantities.

[4] Such difficulties are sometimes avoided by assuming that small divisors do not occur; see, for example, [Joly et al. 1993]. But we do not want to make this assumption.

[5] In fact, the basic L^2 estimate for the singular system (1-18) exhibits loss of a single "singular derivative" $\partial_{x'} + \beta\partial_{\theta_0}/\varepsilon$, which is optimal according to the analysis in [Coulombel and Guès 2010].

new tool, namely, a substantial refinement, given in the companion paper [Coulombel et al. 2012], of the calculus of singular pseudodifferential operators constructed in [Williams 2002]. In the new version of the calculus, residual operators have better smoothing properties than previously realized and can therefore be considered as remainders in our problem. The loss of derivatives in the linear estimate presents a serious difficulty in the application to our semilinear problem. Picard iteration appears to be out of the question, so in Section 5B we use a Nash–Moser iteration scheme adapted to the scale of spaces (1-19) to construct the solution $U_\varepsilon(x, \theta_0)$ to the semilinear singular problem.

If problem (1-3) satisfied the uniform Lopatinskii condition, then, because of the factor ε in the boundary data εG, the equations for the leading profile, $\mathcal{V}^0$ in (1-15), would be linear; and in fact $\mathcal{V}^0$ would vanish. The weakly nonlinear regime would correspond to a source term G (and not εG) in (1-3); see [Williams 1996; 2000]. Under our weak stability assumption, it turns out that $\mathcal{V}^0$ is nonlinearly coupled to the second-order profile $\mathcal{V}^1$ in the profile equations (1-35) and (1-36). To solve these equations, we first isolate a "key subsystem" (1-42) that decouples from the full system. The basic L^2 estimate for the linearization of the key subsystem still exhibits a loss of one derivative, and we are again forced to use Nash–Moser iteration in order to solve this subsystem. Once the key subsystem is solved, the solution of the full profile system (1-35)–(1-36) follows easily. It appears in our analysis that the leading-order amplitude equation shares the weak well-posedness of the original nonlinear problem, but we have not checked whether the loss of derivative for the amplitude equation is optimal (we conjecture that it is).

The error analysis used to answer question (2) above is based on the estimate for the singular system (1-18) (see Proposition 2.2) and is discussed in more detail in Section 1E.

This paper can be read independently of [Coulombel et al. 2012]; for the reader's convenience, we have gathered all the necessary material on the singular calculus in Appendix A. Before discussing this more fully, we provide some definitions, notation, and a precise statement of assumptions.

Remark 1.1. We emphasize that our approach for constructing exact highly oscillating solutions for the system (1-1) can be used without any modification for constructing exact amplified pulses. More precisely, the estimates and well-posedness argument of Sections 2A, 2B, and 2C for the linearized singular system (2-1), and the Nash–Moser argument of Section 5B for the nonlinear singular system (1-18) have all been written so as to carry over verbatim to the case of pulses. Amplification of pulses is treated in [Coulombel and Williams 2013], where we consider a function G in (1-1) that has suitable decay properties with respect to its additional variable $\theta_0 \in \mathbb{R}$ (this functional framework is relevant for applications to lasers). We refer to [Coulombel and Williams 2013] for the precise statements in the pulse case. The main difference between the analysis of wavetrains and pulses lies in the leading-order profile equation and in the construction and estimation of correctors needed in the error analysis. The novelty is that we can get a rate of convergence for (1-4) while this seems out of reach for wavetrains.

1A. ***Assumptions.*** We make the following hyperbolicity assumption on the system (1-1):

Assumption 1.2. There exists an integer $q \geq 1$, some real functions $\lambda_1, \dots, \lambda_q$ that are analytic on $\mathbb{R}^d \setminus \{0\}$ and homogeneous of degree 1, and there exist some positive integers $\nu_1, \dots, \nu_q$ such that

$$\det\Big[\tau I + \sum_{j=1}^{d} \xi_j B_j\Big] = \prod_{k=1}^{q} (\tau + \lambda_k(\xi))^{\nu_k} \quad \text{for all } \xi = (\xi_1, \dots, \xi_d) \in \mathbb{R}^d \setminus \{0\}.$$

Moreover the eigenvalues $\lambda_1(\xi), \dots, \lambda_q(\xi)$ are semisimple (their algebraic multiplicity equals their geometric multiplicity) and satisfy $\lambda_1(\xi) < \cdots < \lambda_q(\xi)$ for all $\xi \in \mathbb{R}^d \setminus \{0\}$.

For simplicity, we restrict our analysis to noncharacteristic boundaries, and therefore make the following assumption.

Assumption 1.3. The matrix B_d is invertible and the matrix $B := \psi(0)$ has maximal rank, its rank p being equal to the number of positive eigenvalues of B_d (counted with their multiplicity). Moreover, the integer p satisfies $1 \le p \le N-1$.

In the normal modes analysis for (1-3), one first performs a Laplace transform in the time variable t and a Fourier transform in the tangential space variables y. We let $\tau - i\gamma \in \mathbb{C}$ and $\eta \in \mathbb{R}^{d-1}$ denote the dual variables of t and y. We introduce the symbol

$$\mathcal{A}(\zeta) := -iB_d^{-1}\Big((\tau - i\gamma)I + \sum_{j=1}^{d-1} \eta_j B_j\Big), \quad \zeta := (\tau - i\gamma, \eta) \in \mathbb{C} \times \mathbb{R}^{d-1}.$$

For future use, we also define the following sets of frequencies:

$$\Xi := \{(\tau - i\gamma, \eta) \in \mathbb{C} \times \mathbb{R}^{d-1} \setminus (0,0) : \gamma \ge 0\}, \qquad \Sigma := \{\zeta \in \Xi : \tau^2 + \gamma^2 + |\eta|^2 = 1\},$$
$$\Xi_0 := \{(\tau, \eta) \in \mathbb{R} \times \mathbb{R}^{d-1} \setminus (0,0)\} = \Xi \cap \{\gamma = 0\}, \quad \Sigma_0 := \Sigma \cap \Xi_0.$$

Two key objects in our analysis are the hyperbolic region and the glancing set, defined as follows.

Definition 1.4.

- The hyperbolic region $\mathcal{H}$ is the set of all $(\tau, \eta) \in \Xi_0$ such that the matrix $\mathcal{A}(\tau, \eta)$ is diagonalizable with purely imaginary eigenvalues.
- Let $\boldsymbol{G}$ denote the set of all $(\tau, \xi) \in \mathbb{R} \times \mathbb{R}^d$ such that $\xi \neq 0$ and there exists an integer $k \in \{1, \dots, q\}$ satisfying
$$\tau + \lambda_k(\xi) = \frac{\partial \lambda_k}{\partial \xi_d}(\xi) = 0.$$
If $\pi(\boldsymbol{G})$ denotes the projection of $\boldsymbol{G}$ on the d first coordinates (that is, $\pi(\tau, \xi) = (\tau, \xi_1, \dots, \xi_{d-1})$ for all (τ, ξ)), the glancing set $\mathcal{G}$ is $\mathcal{G} := \pi(\boldsymbol{G}) \subset \Xi_0$.

We recall the following result, proved in [Kreiss 1970] in the strictly hyperbolic case (when all integers ν_j in Assumption 1.2 equal 1) and [Métivier 2000] in our more general framework.

Proposition 1.5 [Kreiss 1970; Métivier 2000]. *Let Assumptions 1.2 and 1.3 be satisfied. Then, for all $\zeta \in \Xi \setminus \Xi_0$, the matrix $\mathcal{A}(\zeta)$ has no purely imaginary eigenvalue and its stable subspace $\mathbb{E}^s(\zeta)$ has dimension p. Furthermore, $\mathbb{E}^s$ defines an analytic vector bundle over $\Xi \setminus \Xi_0$ that can be extended as a continuous vector bundle over Ξ.*

For all $(\tau, \eta) \in \Xi_0$, we let $\mathbb{E}^s(\tau, \eta)$ denote the continuous extension of $\mathbb{E}^s$ to the point (τ, η). The analysis in [Métivier 2000] shows that away from the glancing set $\mathcal{G} \subset \Xi_0$, $\mathbb{E}^s(\zeta)$ depends analytically on ζ, and the hyperbolic region $\mathcal{H}$ does not contain any glancing point.

To treat the case when the boundary operator in (1-3)(b) is independent of u, which is to say $\psi(\varepsilon u) \equiv \psi(0) =: B$, we make the following *weak stability assumption* on the problem $(L(\partial), B)$.

Assumption 1.6.
- For all $\zeta \in \Xi \setminus \Xi_0$, $\ker B \cap \mathbb{E}^s(\zeta) = \{0\}$.
- The set $\Upsilon_0 := \{\zeta \in \Sigma_0 : \ker B \cap \mathbb{E}^s(\zeta) \neq \{0\}\}$ is nonempty and included in the hyperbolic region $\mathcal{H}$.
- For all $\underline{\zeta} \in \Upsilon_0$, there exists a neighborhood $\mathcal{V}$ of $\underline{\zeta}$ in Σ, a real valued $\mathcal{C}^\infty$ function σ defined on $\mathcal{V}$, a basis $E_1(\zeta), \dots, E_p(\zeta)$ of $\mathbb{E}^s(\zeta)$ that is of class $\mathcal{C}^\infty$ with respect to $\zeta \in \mathcal{V}$, and a matrix $P(\zeta) \in \mathrm{GL}_p(\mathbb{C})$ that is of class $\mathcal{C}^\infty$ with respect to $\zeta \in \mathcal{V}$, such that

$$\text{for all } \zeta \in \mathcal{V}, \quad B\begin{pmatrix} E_1(\zeta) & \cdots & E_p(\zeta)\end{pmatrix} = P(\zeta)\,\mathrm{diag}(\gamma + i\sigma(\zeta), 1, \dots, 1).$$

For comparison and later reference we recall the following definition.

Definition 1.7 [Kreiss 1970]. As before let p be the number of positive eigenvalues of B_d. The problem $(L(\partial), B)$ is said to be *uniformly stable* or to satisfy the *uniform Lopatinskii condition* if

$$B : \mathbb{E}^s(\zeta) \to \mathbb{C}^p$$

is an isomorphism for all $\zeta \in \Sigma$.

Remark 1.8. Observe that if $(L(\partial), B)$ satisfies the uniform Lopatinskii condition, continuity implies that this condition still holds for $(L(\partial), B + \dot{\psi})$, where $\dot{\psi}$ is any sufficiently small perturbation of B. Hence the uniform Lopatinskii condition is a convenient framework for nonlinear perturbation. The analogous statement may not be true when $(L(\partial), B)$ is only weakly stable. Remarkably, weak stability persists under perturbation in the so-called WR class exhibited in [Benzoni-Gavage et al. 2002], and Assumption 1.6 is a convenient equivalent definition of the WR class; see [Coulombel and Guès 2010, Appendix B]. In order to handle general nonlinear boundary conditions as in (1-3), we strengthen Assumption 1.6 in Assumption 1.12.

Boundary and interior phases. We consider a planar real phase ϕ_0 defined on the boundary:

$$\phi_0(t, y) := \underline{\tau} t + \underline{\eta} \cdot y, \quad (\underline{\tau}, \underline{\eta}) \in \Xi_0. \tag{1-5}$$

As follows from earlier works (see, for example, [Majda and Artola 1988]), oscillations on the boundary associated with the phase ϕ_0 give rise to oscillations in the interior associated with some planar phases ϕ_m. These phases are characteristic for the hyperbolic operator $L_0(\partial)$ and their trace on the boundary $\{x_d = 0\}$ equals ϕ_0. For now we make the following assumption.

Assumption 1.9. The phase ϕ_0 defined by (1-5) satisfies $(\underline{\tau}, \underline{\eta}) \in \Upsilon_0$. In particular $(\underline{\tau}, \underline{\eta}) \in \mathcal{H}$.

Thanks to Assumption 1.9, we know that the matrix $\mathcal{A}(\underline{\tau}, \underline{\eta})$ is diagonalizable with purely imaginary eigenvalues. These eigenvalues are denoted by $i\underline{\omega}_1, \dots, i\underline{\omega}_M$, where the $\underline{\omega}_m$s are real and pairwise distinct.

The $\underline{\omega}_m$s are the roots (and all the roots are real) of the dispersion relation

$$\det\Big[\underline{\tau} I + \sum_{j=1}^{d-1} \underline{\eta}_j B_j + \omega B_d\Big] = 0.$$

To each root $\underline{\omega}_m$ there corresponds a unique integer $k_m \in \{1, \dots, q\}$ such that $\underline{\tau} + \lambda_{k_m}(\underline{\eta}, \underline{\omega}_m) = 0$. We can then define the following real[6] phases and their associated group velocities:

$$\text{for all } m = 1, \dots, M, \quad \phi_m(x) := \phi_0(t, y) + \underline{\omega}_m x_d, \quad \boldsymbol{v}_m := \nabla \lambda_{k_m}(\underline{\eta}, \underline{\omega}_m). \tag{1-6}$$

Let us observe that each group velocity $\boldsymbol{v}_m$ is either incoming or outgoing with respect to the space domain $\mathbb{R}^d_+$: the last coordinate of $\boldsymbol{v}_m$ is nonzero. This property holds because $(\underline{\tau}, \underline{\eta})$ does not belong to the glancing set $\mathcal{G}$. We can therefore adopt the following classification.

Definition 1.10. The phase ϕ_m is incoming when the group velocity $\boldsymbol{v}_m$ is incoming (that is, when $\partial_{\xi_d} \lambda_{k_m}(\underline{\eta}, \underline{\omega}_m) > 0$), and it is outgoing when the group velocity $\boldsymbol{v}_m$ is outgoing ($\partial_{\xi_d} \lambda_{k_m}(\underline{\eta}, \underline{\omega}_m) < 0$).

In all that follows, we let $\mathcal{I}$ denote the set of indices $m \in \{1, \dots, M\}$ such that ϕ_m is an incoming phase, and $\mathcal{O}$ denote the set of indices $m \in \{1, \dots, M\}$ such that ϕ_m is an outgoing phase. If $p \geq 1$, $\mathcal{I}$ is nonempty, while if $p \leq N - 1$, $\mathcal{O}$ is nonempty (see Lemma 1.11). We will use the notation

$$L_0(\tau, \xi) := \tau I + \sum_{j=1}^{d} \xi_j B_j, \quad L(\beta, \underline{\omega}_m) := \underline{\omega}_m I + \sum_{k=0}^{d-1} \beta_k A_k,$$
$$\beta := (\underline{\tau}, \underline{\eta}), \quad x' = (t, y), \quad \phi_0(x') = \beta \cdot x'.$$

For each phase ϕ_m, $d\phi_m$ denotes the differential of the function ϕ_m with respect to its argument $x = (t, y, x_d)$. It follows from Assumption 1.2 that the eigenspace of $\mathcal{A}(\beta)$ associated with the eigenvalue $i\underline{\omega}_m$ coincides with the kernel of $L_0(d\phi_m)$ and has dimension ν_{k_m}. The following well-known lemma, whose proof is recalled in [Coulombel and Guès 2010], gives a useful decomposition of $\mathbb{E}^s$ in the hyperbolic region.

Lemma 1.11. *The stable subspace $\mathbb{E}^s(\beta)$ admits the decomposition*

$$\mathbb{E}^s(\beta) = \bigoplus_{m \in \mathcal{I}} \ker L_0(d\phi_m), \tag{1-7}$$

and each vector space in the decomposition (1-7) admits a basis of real vectors.

To formulate our last assumption we observe first that for every point $\underline{\zeta} \in \mathcal{H}$ there is a neighborhood $\mathcal{V}$ of $\underline{\zeta}$ in Σ and a C^∞ conjugator $Q_0(\zeta)$ defined on $\mathcal{V}$ such that

$$Q_0(\zeta) \mathcal{A}(\zeta) Q_0^{-1}(\zeta) = \begin{pmatrix} i\omega_1(\zeta) I_{n_1} & & 0 \\ & \ddots & \\ 0 & & i\omega_J(\zeta) I_{n_J} \end{pmatrix} =: -\mathbb{D}_1(\zeta), \tag{1-8}$$

[6] If $(\underline{\tau}, \underline{\eta})$ does not belong to the hyperbolic region $\mathcal{H}$, some of the phases ϕ_m may be complex; see, for example, [Williams 1996; 2000; Lescarret 2007; Marcou 2010]. Moreover, glancing phases introduce a new scale $\sqrt{\varepsilon}$ as well as boundary layers.

where the ω_j are real when $\gamma = 0$ and there is a constant $c > 0$ such that either

$$\mathrm{Re}(i\omega_j) \le -c\gamma \quad \text{or} \quad \mathrm{Re}(i\omega_j) \ge c\gamma \text{ for all } \zeta \in \mathcal{V}.$$

In view of Lemma 1.11, we can choose the first p columns of $Q_0^{-1}(\zeta)$ to be a basis of $\mathbb{E}^s(\zeta)$, and write

$$Q_0^{-1}(\zeta) = [Q_{\mathrm{in}}(\zeta)\, Q_{\mathrm{out}}(\zeta)].$$

Choose J' so that the first J' blocks of $-\mathbb{D}_1$ lie in the first p columns, and the remaining blocks in the remaining $N - p$ columns. Thus $\mathrm{Re}(i\omega_j) \le -c\gamma$ if and only if $1 \le j \le J'$.

Observing that the linearization of the boundary condition in (1-3) is

$$\dot u \mapsto \psi(\varepsilon u)\dot u + [d\psi(\varepsilon u)\dot u]\varepsilon u,$$

we define the operator

$$\mathcal{B}(v_1, v_2)\dot u := \psi(v_1)\dot u + [d\psi(v_1)\dot u]v_2, \tag{1-9}$$

which appears in Assumption 1.12. For later use we also define

$$\mathcal{D}(v_1, v_2)\dot u := D(v_1)\dot u + [dD(v_1)\dot u]v_2, \tag{1-10}$$

as well as

$$\mathcal{B}(v_1) := \mathcal{B}(v_1, v_1), \mathcal{D}(v_1) := \mathcal{D}(v_1, v_1). \tag{1-11}$$

We now state the weak stability assumption that we make when considering the general case of nonlinear boundary conditions in (1-3).

Assumption 1.12.

- There exists a neighborhood $\mathbb{O}$ of $(0, 0) \in \mathbb{R}^{2N}$ such that for all $(v_1, v_2) \in \mathbb{O}$ and all $\zeta \in \Xi \setminus \Xi_0$, $\ker \mathcal{B}(v_1, v_2) \cap \mathbb{E}^s(\zeta) = \{0\}$. For each $(v_1, v_2) \in \mathbb{O}$, the set

$$\Upsilon(v_1, v_2) := \{\zeta \in \Sigma_0 : \ker \mathcal{B}(v_1, v_2) \cap \mathbb{E}^s(\zeta) \ne \{0\}\}$$

  is nonempty and is included in the hyperbolic region $\mathcal{H}$. Moreover, if we set $\Upsilon := \bigcup_{(v_1,v_2)\in\mathbb{O}} \Upsilon(v_1, v_2)$, $\overline{\Upsilon} \subset \mathcal{H}$ (closure in Σ_0).

- For every $\underline{\zeta} \in \overline{\Upsilon}$, there exists a neighborhood $\mathcal{V}$ of $\underline{\zeta}$ in Σ and a C^∞ function $\sigma(v_1, v_2, \zeta)$ on $\mathbb{O} \times \mathcal{V}$ such that for all $(v_1, v_2, \zeta) \in \mathbb{O} \times \mathcal{V}$ we have $\ker \mathcal{B}(v_1, v_2) \cap \mathbb{E}^s(\zeta) \ne \{0\}$ if and only if $\zeta \in \Sigma_0$ and $\sigma(v_1, v_2, \zeta) = 0$.

  Moreover, there exist matrices $P_i(v_1, v_2, \zeta) \in \mathrm{GL}_p(\mathbb{C})$, $i = 1, 2$, of class C^∞ on $\mathbb{O} \times \mathcal{V}$ such that, for all $(v_1, v_2, \zeta) \in \mathbb{O} \times \mathcal{V}$,

$$P_1(v_1, v_2, \zeta)\mathcal{B}(v_1, v_2)Q_{\mathrm{in}}(\zeta)P_2(v_1, v_2, \zeta) = \mathrm{diag}(\gamma + i\sigma(v_1, v_2, \zeta), 1, \dots, 1). \tag{1-12}$$

For nonlinear boundary conditions, the phase ϕ_0 in (1-5) is assumed to satisfy $(\underline{\tau}, \underline{\eta}) \in \Upsilon(0, 0)$, or, in other words, the intersection $\ker B \cap \mathbb{E}^s(\underline{\tau}, \underline{\eta})$ is not reduced to $\{0\}$ (the set Υ_0 in Assumption 1.6 is a short notation for $\Upsilon(0, 0)$). The phases ϕ_m are still defined by (1-6) and thus only depend on $L(\partial)$ and B, and not on the nonlinear perturbations f_0 and $\psi(\varepsilon u) - \psi(0)$ added in (1-3).

Remark 1.13. (1) The properties stated in Assumption 1.12 are just a convenient description of the requirements for belonging to the WR class of [Benzoni-Gavage et al. 2002]. Like the uniform Lopatinskii condition, Assumption 1.12 can, in practice, be verified by hand via a "constant coefficient" computation. More precisely, for (v_1, v_2) near $(0, 0) \in \mathbb{R}^{2N}$ and $\zeta \in \Sigma$, one can define (see, for example, [Benzoni-Gavage and Serre 2007, chapter 4]) a Lopatinskii determinant $\Delta(v_1, v_2, \zeta)$ that is C^∞ in (v_1, v_2), analytic in $\zeta = (\tau - i\gamma, \eta)$ on $\Sigma \setminus \mathcal{G}$, and satisfies

$$\Delta(v_1, v_2, \zeta) = 0 \quad \text{if and only if } \ker \mathcal{B}(v_1, v_2) \cap \mathbb{E}^s(\zeta) \neq \{0\}.$$

In particular, $\Delta(v_1, v_2, \cdot)$ is real-analytic on $\mathcal{H}$.

Following [Benzoni-Gavage et al. 2002] (see also [Benzoni-Gavage and Serre 2007, chapter 8]), we claim that Assumption 1.12 holds provided

$$\varnothing \neq \{\zeta \in \Sigma : \Delta(0, 0, \zeta) = 0\} \subset \mathcal{H} \quad \text{and} \quad \Delta(0, 0, \underline{\zeta}) = 0 \Rightarrow \partial_\tau \Delta(0, 0, \underline{\zeta}) \neq 0, \tag{1-13}$$

and thus it only involves a weak stability property for the linearized problem at $(v_1, v_2) = (0, 0)$. Indeed, the implicit function theorem then implies that, for (v_1, v_2) near zero and (τ, η) near $\underline{\zeta}$, the set

$$\{(\tau, \eta) \in \Sigma_0 : \Delta(v_1, v_2, \tau, \eta) = 0\}$$

is a real-analytic hypersurface in $\mathcal{H}$. On the other hand, an application of the implicit function theorem to $\Delta(v_1, v_2, z, \eta)$, for $(z, \eta) \in \Sigma$, shows that the real dimension of the manifold

$$\{(z, \eta) \in \Sigma : \Delta(v_1, v_2, z, \eta) = 0\}$$

must be the same, that is, $d - 2$. The two zero sets must then coincide; there are no zeros in $\Sigma \setminus \Sigma_0$. The function σ and the neighborhoods $\mathcal{O}$ and $\mathcal{V}$ arise in a factorization of Δ given by the Weierstrass preparation theorem. The construction of the conjugating matrices P_i, $i = 1, 2$ follows from a construction in [Sablé-Tougeron 1988, Pages 268–270].

Instead of assuming (1-13), we have stated Assumption 1.12 in a form that is more directly applicable to the proof of Proposition 2.2 and to the error analysis of Theorem 4.1.

(2) To prove the basic estimate for the linearized singular system, Proposition 2.2, and to construct the exact solution U_ε to the singular system (1-18), it is enough to require that the analogue of Assumption 1.12 holds when $\mathcal{B}(v_1, v_2)$ is replaced by $\mathcal{B}(v_1) := \mathcal{B}(v_1, v_1)$. However, for the error analysis of Section 4 in the case of nonlinear boundary conditions, we need Assumption 1.12 as stated.

The next lemma, proved in [Coulombel and Guès 2010], gives a useful decomposition of $\mathbb{C}^N$ and introduces projectors needed later for formulating and solving the profile equations.

Lemma 1.14. *The space $\mathbb{C}^N$ admits the decomposition*

$$\mathbb{C}^N = \bigoplus_{m=1}^{M} \ker L_0(d\phi_m), \tag{1-14}$$

and each vector space in (1-14) admits a basis of real vectors. If we let $P_1, \dots, P_M$ denote the projectors associated with the decomposition (1-14), we have $\operatorname{Im} B_d^{-1} L_0(d\phi_m) = \ker P_m$ for all $m = 1, \dots, M$.

1B. ***Main results.*** For each $m \in \{1, \dots, M\}$ we let

$$r_{m,k}, \quad k = 1, \dots, \nu_{k_m},$$

denote a basis of $\ker L_0(d\phi_m)$ consisting of real vectors. In Section 4 we shall construct a "corrected" approximate solution u_ε^c of (1-3) of the form

$$u_\varepsilon^c(x) = \mathcal{V}^0\Big(x, \frac{\phi}{\varepsilon}\Big) + \varepsilon \mathcal{V}^1\Big(x, \frac{\phi}{\varepsilon}\Big) + \varepsilon^2 \mathcal{U}_p^2\Big(x, \frac{\phi_0}{\varepsilon}, \frac{x_d}{\varepsilon}\Big), \tag{1-15}$$

where $\phi := (\phi_1, \dots, \phi_M)$ denotes the collection of all phases,

$$\begin{aligned} \mathcal{V}^0\Big(x, \frac{\phi}{\varepsilon}\Big) &= \sum_{m \in \mathcal{I}} \sum_{k=1}^{\nu_{k_m}} \sigma_{m,k}\Big(x, \frac{\phi_m}{\varepsilon}\Big) r_{m,k}, \\ \mathcal{V}^1\Big(x, \frac{\phi}{\varepsilon}\Big) &= \underline{\mathcal{V}}^1(x) + \sum_{m=1}^{M} \sum_{k=1}^{\nu_{k_m}} \tau_{m,k}\Big(x, \frac{\phi_m}{\varepsilon}\Big) r_{m,k} + \mathcal{R}\mathcal{V}^0, \end{aligned} \tag{1-16}$$

and the $\sigma_{m,k}(x, \theta_m)$ and $\tau_{m,k}(x, \theta_m)$ are scalar C^1 functions periodic in θ_m with mean 0 which describe the propagation of oscillations with phase ϕ_m and group velocity $\boldsymbol{v}_m$. Here $\mathcal{R}$ denotes the nonlocal operator

$$\mathcal{R}\mathcal{V}^0 = -R[L(\partial_x)\mathcal{V}^0 + D(0)\mathcal{V}^0]$$

for R defined as in (1-32). The last corrector $\varepsilon^2 \mathcal{U}_p^2(x, \theta_0, \xi_d)$ in (1-15) is a trigonometric polynomial constructed in the error analysis of Section 4.

The next theorem, our main result, is an immediate corollary of the more precise Theorem 4.1. Here we let $\Omega_T := \{(x, \theta_0) = (t, y, x_d, \theta_0) \in \mathbb{R}^{d+1} \times \mathbb{T}^1 : x_d \geq 0, t < T\}$ and $b\Omega_T := \{(t, y, \theta_0) \in \mathbb{R}^d \times \mathbb{T}^1 : t < T\}$. The spaces E^s are defined in (1-19).

Theorem 1.15. *We make Assumptions 1.2, 1.3, 1.6, and 1.9 when the boundary condition in (1-3) is linear* ($\psi(\varepsilon u) \equiv \psi(0)$); *in the general case we substitute Assumption 1.12 for Assumption 1.6. Fix* $T > 0$, *set* $M_0 := 3d + 5$, *and let*

$$\mu := [(d+1)/2] + M_0 + 3 \quad \textit{and} \quad \tilde{\mu} := 2\mu - [(d+1)/2].$$

Consider the semilinear boundary problem (1-3), where $G(t, y, \theta_0) \in H^{\tilde{\mu}}(b\Omega_T)$. *There exists* $\varepsilon_0 > 0$ *such that if* $\langle G \rangle_{H^{\mu+2}(b\Omega_T)}$ *is small enough, there exists a unique function* $U_\varepsilon(x, \theta_0) \in E^{\mu-1}(\Omega_T)$ *satisfying the singular system (1-18) on* Ω_T *such that*

$$u_\varepsilon(x) := U_\varepsilon\Big(x, \frac{x' \cdot \beta}{\varepsilon}\Big)$$

is an exact solution of (1-3) on $(-\infty, T] \times \overline{\mathbb{R}}^d_+$ *for* $0 < \varepsilon \leq \varepsilon_0$. *In addition there exists a profile* $\mathcal{V}^0(x, \theta)$ *as in (1-16), whose components* $\sigma_{m,k}$ *lie in* $H^{\mu-1}(\Omega_T)$, *such that the approximate solution defined by*

$$u_\varepsilon^{\mathrm{app}} := \mathcal{V}^0\Big(x, \frac{\phi}{\varepsilon}\Big)$$

satisfies

$$\lim_{\varepsilon \to 0} |u_\varepsilon - u_\varepsilon^{\mathrm{app}}|_{L^\infty} = 0 \quad \textit{on } (-\infty, T] \times \overline{\mathbb{R}}^d_+.$$

Observe that although the boundary data in problem (1-3) is of size $O(\varepsilon)$, the approximate solution $u_\varepsilon^{\mathrm{app}}$ is of size $O(1)$, exhibiting an amplification due to the weak stability at frequency β. The main information provided by Theorem 1.15 is that this amplification does not rule out the existence of a smooth solution on a fixed time interval, that is, it does not trigger a violent instability, at least in this weakly nonlinear regime. As far as we know, the derivation of the leading-order amplitude equation (1-42) is also new in the general framework that we consider. This amplitude equation shares some features of the Burgers equation and we expect that its solutions may develop singularities in finite time; see similar discussions in [Majda and Rosales 1984]. We hope that the analysis developed in this article will be useful in justifying *quasilinear* amplification phenomena such as the Mach stems or kink modes formation [Artola and Majda 1987; Majda and Artola 1988; Majda and Rosales 1983], but there are still many obstacles along the way.

Remark 1.16. (a) In order to avoid some technicalities, we have stated our main result for a problem (1-3) where all data vanish for $t < 0$. This result easily implies a similar result in which outgoing waves defined in $t < 0$ of amplitude $O(\varepsilon)$ and wavelength ε give rise to reflected waves of amplitude $O(1)$. In either formulation, analysis of the profile equations (see Remark 1.28) shows that the waves of amplitude $O(1)$ emanate from a radiating wave that propagates in the boundary along a characteristic of the Lopatinskii determinant.

(b) We have decided to fix $T > 0$ at the start and choose data small enough so that a solution to the nonlinear problem exists up to time T. One can also (as discussed in Remark 3.7) fix the data in the problem (G in (1-3)) at the start, and then choose T small enough so that a solution to the nonlinear problem exists up to time T.

In the remainder of this introduction, we discuss the construction of exact solutions, the construction of the approximate solution $\mathcal{V}^0$, and the error analysis. Complete proofs are given in Sections 2, 3, 4, and 5.

1C. ***Exact solutions and singular systems.*** The theory of weakly stable hyperbolic initial boundary value problems fails to provide a solution of the system (1-3) that exists on a fixed time interval independent of ε.[7] In order to obtain such an exact solution to the system (1-3), we adopt the strategy of studying an associated singular problem first used in [Joly et al. 1995] for an initial value problem in free space. We look for a solution of the form

$$u_\varepsilon(x) = U_\varepsilon(x, \theta_0)|_{\theta_0 = \phi_0(x')/\varepsilon}, \tag{1-17}$$

where $U_\varepsilon(x, \theta_0)$ is periodic in θ_0 and satisfies the singular system derived by substituting (1-17) into problem (1-3). Recalling that $L(\partial) = \partial_d + \sum_{j=0}^{d-1} A_j \partial_j$ we obtain

$$\begin{gathered} \partial_d U_\varepsilon + \sum_{j=0}^{d-1} A_j \left(\partial_j + \frac{\beta_j \partial_{\theta_0}}{\varepsilon}\right) U_\varepsilon + D(\varepsilon U_\varepsilon) U_\varepsilon =: \partial_d U_\varepsilon + \mathbb{A}\left(\partial_{x'} + \frac{\beta \partial_{\theta_0}}{\varepsilon}\right) U_\varepsilon + D(\varepsilon U_\varepsilon) U_\varepsilon = 0, \\ \psi(\varepsilon U_\varepsilon) U_\varepsilon|_{x_d = 0} = \varepsilon G(x', \theta_0), \\ U_\varepsilon = 0 \quad \text{in } t < 0. \end{gathered} \tag{1-18}$$

[7] This would be true even for problems $(L(\partial), B)$ that are uniformly stable in the sense of Definition 1.7.

The special difficulties presented by such singular problems when there is a boundary are described in detail in the introductions to [Williams 1996; 2002; Coulombel et al. 2011]. In particular, we mention:

(a) Symmetry assumptions on the matrices B_j appearing in the problem (1-1) equivalent to (1-3) are generally of no help in obtaining an L^2 estimate for (1-18) (boundary conditions satisfying Assumption 1.6 cannot be maximally dissipative; see [Coulombel and Guès 2010]).

(b) One cannot control L^∞ norms just by estimating tangential derivatives $\partial^\alpha_{(x',\theta_0)} U_\varepsilon$ because (1-18) is not a hyperbolic problem in the x_d direction;[8] moreover, even if one has estimates of tangential derivatives uniform with respect to ε, because of the factors $1/\varepsilon$ in (1-18), one cannot just use the equation to control $\partial_d U_\varepsilon$ and thereby control L^∞ norms.

To deal with these difficulties, Williams [2002] introduced a class of singular pseudodifferential operators, acting on functions $U(x', \theta_0)$ that are 2π-periodic in θ_0 and having the form

$$p_D U(x', \theta_0) = \frac{1}{(2\pi)^d} \sum_{k\in\mathbb{Z}} \int_{\mathbb{R}^d} e^{ix'\cdot\xi' + i\theta_0 k}\, p\left(\varepsilon V(x', \theta_0), \xi' + \frac{k\beta}{\varepsilon}, \gamma\right) \widehat{U}(\xi', k)\, d\xi', \gamma \geq 1.$$

Observe that the differential operator $\mathbb{A}$ appearing in (1-18) can be expressed in this form. Kreiss-type symmetrizers $r_s(D_{x',\theta_0})$ in the singular calculus were constructed in [Williams 2002] for (quasilinear systems similar to) (1-18) under the assumption that $(L(\partial), \psi(0))$ is uniformly stable in the sense of Definition 1.7. With these, one can prove $L^2(x_d, H^s(x', \theta_0))$ estimates uniform in ε for (1-18), even when εG is replaced by G in the boundary condition. To progress further and control L^∞ norms, the boundary frequency β is restricted to lie in the complement of the glancing set. With this extra assumption, the singular calculus was used in [Williams 2002] to block-diagonalize the singular operator $\mathbb{A}\left(\varepsilon U_\varepsilon, \partial_{x'} + \beta\partial_{\theta_0}/\varepsilon\right)$ microlocally near the β direction and thereby prove estimates uniform with respect to ε in the spaces

$$E^s := C(x_d, H^s(x', \theta_0)) \cap L^2(x_d, H^{s+1}(x', \theta_0)). \tag{1-19}$$

These spaces are Banach algebras and are contained in L^∞ for $s > (d+1)/2$. For large enough s, as determined by the requirements of the calculus, existence of solutions to (1-18) in E^s on a time interval $[0, T]$ independent of $\varepsilon \in (0, \varepsilon_0]$ follows by Picard iteration in the uniformly stable case.

The singular calculus of [Williams 2002] was used again in [Coulombel et al. 2011] to rigorously justify leading-order geometric optics expansions for the quasilinear analogue of (1-3) in the uniformly stable case (with $\beta \in \mathcal{H}$ and the forcing term G in place of εG in the boundary condition). Under the assumptions made in the present paper, in particular assuming weak stability as in Assumptions 1.6 and 1.12, we face the additional difficulty that the basic L^2 estimate for the problem $(L(\partial), B)$ exhibits a loss of derivatives. A consequence of this is that the singular calculus of [Williams 2002] is no longer adequate for estimating solutions of (1-18). The main reason is that remainders in the calculus of [Williams 2002] are just bounded operators on L^2, while for energy estimates with a loss of derivative, remainders should be smoothing operators. We therefore need to use an improved version of the calculus constructed in

[8] For initial value problems in free space, one *can* control L^∞ norms just by estimating enough derivatives tangent to time slices $t = c$.

[Coulombel et al. 2012] in which residual operators are shown to have better smoothing properties than previously thought. With the improved calculus we are able in Section 2C to estimate solutions of (1-18) in E^s spaces (1-19), but of course there is a loss of one singular derivative in the estimates. This loss forces us in Section 5B to use Nash–Moser iteration on the scale of E^s spaces to obtain an exact solution of the singular system (1-18) on a fixed time interval independent of ε. Observe that one singular derivative costs a factor $1/\varepsilon$ and this is another reason why the scaling εG in (1-18) is crucial.

Remark 1.17. The main idea employed in proving the estimate for the linearized singular problem, Proposition 2.2, is to adapt the techniques of [Coulombel 2004] to the singular pseudodifferential framework. There is however one major obstacle along the way. While the error term in the composition of two zero-order operators (or in the composition of an operator of order -1 (on the left) with an operator of order 1, a $(-1, 1)$ composition) is smoothing of order 1 in the sense of (A-3), the same is unfortunately not true of the error term in $(1, -1)$ compositions (there are counterexamples for that). The properties of the $(1, -1)$ error terms that arise in our proof are described in Lemma 2.6.

1D. ***Derivation of the leading profile equations.*** We now derive the profile equations for the semilinear problem (1-3). We work with profiles $\mathcal{V}^j(x, \theta)$ periodic in $\theta = (\theta_1, \dots, \theta_M)$, where θ_j is a placeholder for ϕ_j/ε. Looking for an approximate solution of (1-3) of the form $u^a = (\mathcal{V}^0 + \varepsilon\mathcal{V}^1 + \varepsilon^2\mathcal{V}^2)|_{\theta=\phi/\varepsilon}$, where $\phi = (\phi_1, \dots, \phi_M)$, we get interior equations

$$\begin{aligned} &\text{(a)} \quad \mathcal{L}(\partial_\theta)\mathcal{V}^0 = 0, \\ &\text{(b)} \quad \mathcal{L}(\partial_\theta)\mathcal{V}^1 + L(\partial)\mathcal{V}^0 + D(0)\mathcal{V}^0 = 0, \\ &\text{(c)} \quad \mathcal{L}(\partial_\theta)\mathcal{V}^2 + L(\partial)\mathcal{V}^1 + D(0)\mathcal{V}^1 + (dD(0)\mathcal{V}^0)\mathcal{V}^0 = 0, \end{aligned} \tag{1-20}$$

by plugging u^a into (1-3)(a) and setting the coefficients of, respectively, ε^{-1}, ε^0, and ε equal to zero. The operator $\mathcal{L}(\partial_\theta)$ is defined by

$$\mathcal{L}(\partial_\theta) := \sum_{j=1}^{M} L(d\phi_j)\partial_{\theta_j}. \tag{1-21}$$

With $B := \psi(0)$, the boundary equations, obtained by plugging u^a into (1-3)(b) and setting the coefficients of ε^0 and ε equal to zero, are

$$\begin{aligned} &B\mathcal{V}^0(x', 0, \theta_0, \dots, \theta_0) = 0, \\ &B\mathcal{V}^1 + (d\psi(0)\mathcal{V}^0)\mathcal{V}^0 = G(x', \theta_0), \end{aligned} \tag{1-22}$$

where θ_0 is a placeholder for ϕ_0/ε. We will see that as a consequence of the weak stability at frequency β, the problem for the leading profile $\mathcal{V}^0$ is nonlinear and nonlocal. (See Appendix B for a concrete example.) Thus, the scaling in (1-2) *is* the weakly nonlinear scaling when the uniform Lopatinskii condition fails at a hyperbolic frequency β. To analyze these equations, we proceed to define appropriate function spaces and a pair of auxiliary operators E and R.

Functions $\mathcal{V}(x, \theta) \in L^2(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$ have Fourier series

$$\mathcal{V}(x, \theta) = \sum_{\alpha\in\mathbb{Z}^M} V_\alpha(x)e^{i\alpha\cdot\theta}. \tag{1-23}$$

Since only quadratic interactions appear in (1-20) and we anticipate that $\mathcal{V}^0$ will have the form in (1-16), for $k = 1, 2$ we let

$$\mathbb{Z}^{M;k} = \{\alpha \in \mathbb{Z}^M : \text{at most } k \text{ components of } \alpha \text{ are nonzero}\},$$

and we consider the subspace $H^{s;k}(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M) \subset H^s(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$ defined by

$$H^{s;k}(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M) = \left\{ \mathcal{V}(x,\theta) \in H^s(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M) : \mathcal{V}(x,\theta) = \sum_{\alpha \in \mathbb{Z}^{M;k}} V_\alpha(x) e^{i\alpha\cdot\theta} \right\}. \tag{1-24}$$

Thus multiplication defines a continuous map

$$H^{s;1}(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M) \times H^{s;1}(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M) \to H^{s;2}(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M) \tag{1-25}$$

for $s > (d+1+2)/2$.

Definition 1.18. Setting $\phi := (\phi_1, \dots, \phi_M)$, we say $\alpha \in \mathbb{Z}^{M;2}$ is a *characteristic mode* and write $\alpha \in \mathcal{C}$ if $\det L(d(\alpha\cdot\phi)) = 0$. Otherwise we call α a *noncharacteristic* mode. We decompose $\mathcal{C}$ as

$$\mathcal{C} = \bigcup_{m=1}^{M} \mathcal{C}_m, \quad \text{where } \mathcal{C}_m := \{\alpha \in \mathbb{Z}^{M;2} : \alpha\cdot\phi = n_\alpha \phi_m \text{ for some } n_\alpha \in \mathbb{Z}\}.$$

Observe that for $\alpha \in \mathcal{C}_m$, the integer n_α is necessarily equal to $\sum_{k=1}^M \alpha_k$. Since ϕ_i and ϕ_j are linearly independent for $i \neq j$, any $\alpha \in \mathbb{Z}^{M;2} \setminus 0$ belongs to at most one of the sets $\mathcal{C}_m$ and $n_\alpha \neq 0$ if $\alpha \neq 0$.

Elements $\alpha \in \mathcal{C}_m$ with two nonzero components correspond to *resonances*. Resonances are generated in products like $\sigma_{p,k}(x, \phi_p/\varepsilon)\sigma_{r,k'}(x, \phi_r/\varepsilon)$, which arise from the quadratic term in (1-20)(c), whenever there exists a relation of the form

$$n_m\phi_m = n_p\phi_p + n_r\phi_r, \quad \text{where } m \in \{1,\dots,M\}\setminus\{p,r\} \text{ and } n_m, n_p, n_r \in \mathbb{Z}.$$

We then refer to (ϕ_m, ϕ_p, ϕ_r) as a triple of resonant phases. This relation implies, for example, that ϕ_p oscillations interact with ϕ_r oscillations to produce ϕ_m oscillations.

Definition 1.19. We define the continuous projector[9] $E : H^{s;2}(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M) \to H^{s;1}(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$, $s \geq 0$, by

$$E = E_0 + \sum_{m=1}^{M} E_m, \quad \text{where } E_0\mathcal{V} := V_0 \text{ and } E_m\mathcal{V} := \sum_{\alpha \in \mathcal{C}_m \setminus 0} P_m V_\alpha(x) \mathrm{e}^{i n_\alpha \theta_m}, \tag{1-26}$$

for P_m as in Lemma 1.14.

For $\mathcal{L}(\partial_\theta)$ as in (1-21), we have that, for $\mathcal{V}^0 \in H^{s;2}(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$,

$$E\mathcal{V}^0 = \mathcal{V}^0 \quad \text{if and only if } \mathcal{V}^0 \in H^{s;1}(\bar{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M) \text{ and } \mathcal{L}(\partial_\theta)\mathcal{V}^0 = 0, \tag{1-27}$$

and (1-27) in turn is equivalent to the property that $\mathcal{V}^0$ has an expansion of the form

$$\mathcal{V}^0 = \underline{v}(x) + \sum_{m=1}^{M} \sum_{k=1}^{\nu_{k_m}} \sigma_{m,k}(x,\theta_m) r_{m,k}, \tag{1-28}$$

[9] The continuity of E is shown in [Coulombel et al. 2011, Remark 2.5].

for some real-valued functions $\sigma_{m,k}$. Moreover, since for any m,

$$L(d\phi_m) = \underline{\omega}_m I + \sum_{j=0}^{d-1} \beta_j A_j = \sum_{k\neq m} (\underline{\omega}_m - \underline{\omega}_k) P_k, \tag{1-29}$$

we have, for $\mathcal{V} \in H^{s;2}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$,

$$E\mathscr{L}(\partial_\theta)\mathcal{V} = \mathscr{L}(\partial_\theta)E\mathcal{V} = 0. \tag{1-30}$$

We also need to introduce a partial inverse R for $\mathscr{L}(\partial_\theta)$. We begin by defining

$$R_m := \sum_{k\neq m} \frac{1}{\underline{\omega}_m - \underline{\omega}_k} P_k,$$

which in view of (1-29) satisfies

$$L(d\phi_m)R_m = R_m L(d\phi_m) = I - P_m. \tag{1-31}$$

The operator R is defined formally at first on functions

$$\mathcal{V}(x,\theta) = \sum_{\alpha\in\mathbb{Z}^{M;2}} V_\alpha(x) e^{i\alpha\cdot\theta} \text{ of } H^{s;2}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$$

by

$$R\mathcal{V} := \sum_{\alpha\in\mathbb{Z}^{M;2}} R(\alpha) V_\alpha(x) e^{i\alpha\cdot\theta} \tag{1-32}$$

where

$$R(\alpha) := \begin{cases} R_m/(in_\alpha) & \text{if } \alpha \in \mathscr{C}_m \setminus \{0\}, \\ 0 & \text{if } \alpha = 0, \\ \mathscr{L}(i\alpha)^{-1} & \text{if } \alpha \notin \mathscr{C}, \end{cases} \tag{1-33}$$

and

$$\mathscr{L}(i\alpha) := i \sum_{m=1}^{M} \alpha_m L(d\phi_m) = iL(d(\alpha\cdot\phi)).$$

Remark 1.20. The operator R is well-defined on functions $\mathcal{V} \in H^{s;2}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$ whose spectrum contains only finitely many noncharacteristic modes, and then $R\mathcal{V}$ lies in the same space. Otherwise, there can be a problem with small divisors; the possibility of there being infinitely many noncharacteristic modes α for which $\det L(d(\alpha\cdot\phi))$ is close to zero can prevent convergence of (1-32) in $H^{t;2}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$ for any t.

It follows readily from (1-31) that, for $\mathcal{F} \in H^{s;1}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$, $s > 0$,

$$\mathscr{L}(\partial_\theta)R\mathcal{F} = R\mathscr{L}(\partial_\theta)\mathcal{F} = (I - E)\mathcal{F}. \tag{1-34}$$

Such $\mathcal{F}$ have no noncharacteristic modes. Along with (1-30), (1-34) implies the following.

Proposition 1.21. *Suppose $\mathcal{F} \in H^{s;1}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$, $s \geq 0$. Then the equation $\mathcal{L}(\partial_\theta)\mathcal{V} = \mathcal{F}$ has a solution $\mathcal{V} \in H^{s;1}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$ if and only if $E\mathcal{F} = 0$.*

By applying the operators E and R to the equations (1-20) and using (1-27), (1-30), and (1-34), we obtain

$$\begin{aligned} &\text{(a)} \quad E\mathcal{V}^0 = \mathcal{V}^0, \\ &\text{(b)} \quad E(L(\partial)\mathcal{V}^0 + D(0)\mathcal{V}^0) = 0, \\ &\text{(c)} \quad B\mathcal{V}^0 = 0 \quad \text{on } x_d = 0, \theta = (\theta_0, \dots, \theta_0), \\ &\text{(d)} \quad \mathcal{V}^0 = 0 \quad \text{in } t < 0 \end{aligned} \tag{1-35}$$

and

$$\begin{aligned} &\text{(a)} \quad (I - E)\mathcal{V}^1 + R(L(\partial)\mathcal{V}^0 + D(0)\mathcal{V}^0) = 0, \\ &\text{(b)} \quad E\left(L(\partial)\mathcal{V}^1 + D(0)\mathcal{V}^1 + (dD(0)\mathcal{V}^0)\mathcal{V}^0\right) = 0, \\ &\text{(c)} \quad B\mathcal{V}^1 + (d\psi(0)\mathcal{V}^0)\mathcal{V}^0 = G \quad \text{on } x_d = 0, \theta = (\theta_0, \dots, \theta_0), \\ &\text{(d)} \quad \mathcal{V}^1 = 0 \quad \text{in } t < 0. \end{aligned} \tag{1-36}$$

Remark 1.22. (a) Since $E\mathcal{V}^0 = \mathcal{V}^0$, the function $L(\partial)\mathcal{V}^0 + D(0)\mathcal{V}^0$ in (1-36)(a) has *no* noncharacteristic modes so the action of R on this function is well-defined.

(b) It is easy to check that functions $\mathcal{V}^0$, $\mathcal{V}^1$ belonging to $H^{s;1}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$, $s > (d+3)/2$, and satisfying (1-35) and (1-36)(a) also satisfy (1-20)(a)–(b) and (1-22). Equation (1-36)(b) and Proposition 1.21 suggest that we might obtain a solution of (1-20)(c) by taking

$$(I - E)\mathcal{V}^2 = -R(L(\partial)\mathcal{V}^1 + D(0)\mathcal{V}^1 + (dD(0)\mathcal{V}^0)\mathcal{V}^0).$$

There are two problems with this. First, the quadratic term $(\mathrm{d}D(0)\mathcal{V}^0)\mathcal{V}^0$ generally has *infinitely* many noncharacteristic modes, so one should expect a problem with small divisors. Second, the statement (1-34) and Proposition 1.21 are both *not* true when $\mathcal{F} \in H^{s;2}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$, even if $\mathcal{F}$ has finitely many noncharacteristic modes.[10] These difficulties affect the error analysis and are discussed further in Section 1E.

To determine the equations satisfied by the individual profiles $\underline{v}(x)$, $\sigma_{m,k}(x, \theta_m)$ in the expansion (1-28) of $\mathcal{V}^0$, we first refine the decomposition of the projector E in (1-26). For each $m \in \{1, \dots, M\}$ we let

$$\ell_{m,k}, k = 1, \dots, \nu_{k_m}$$

denote a basis of real vectors for the left eigenspace of the real matrix

$$i\mathcal{A}(\beta) = \underline{\tau} A_0 + \sum_{j=1}^{d-1} \underline{\eta}_j A_j \tag{1-37}$$

[10] This is because of the fact that for any $k \in \mathbb{Z} \setminus \{0\}$, there can be many $\alpha \in (\mathcal{C}_m \setminus 0) \cap \mathbb{Z}^{M;2}$ such that $n_\alpha = k$. See the proof of Proposition 1.29.

associated to the eigenvalue $-\underline{\omega}_m$, chosen to satisfy

$$\ell_{m,k} \cdot r_{m',k'} = \begin{cases} 1 & \text{if } m = m' \text{ and } k = k', \\ 0 & \text{otherwise.} \end{cases}$$

For $v \in \mathbb{C}^N$ set

$$P_{m,k} v := (\ell_{m,k} \cdot v) r_{m,k} \quad \text{(no complex conjugation here).}$$

We can now write

$$E = E_0 + \sum_{m=1}^{M} \sum_{k=1}^{v_{k_m}} E_{m,k},$$

where $E_{m,k} := P_{m,k} E_m$. When the multiplicity $k = 1$, we write E_m instead of $E_{m,1}$ and do similarly for $\ell_{m,k}$, $r_{m,k}$ and so on.

The following lemma, which is a slight variation on a well-known result [Lax 1957], is included for the sake of completeness.

Lemma 1.23. *Suppose $E\mathring{V}^0 = \mathring{V}^0$ and that $\mathring{V}^0$ has the expansion (1-28). Then*

$$E_{m,k}(L(\partial)\mathring{V}^0) = (X_{\phi_m} \sigma_{m,k}) r_{m,k}$$

where X_{ϕ_m} is the characteristic vector field associated to ϕ_m:[11]

$$X_{\phi_m} := \partial_d + \sum_{j=0}^{d-1} -\partial_{\xi_j} \omega_m(\beta) \partial_j.$$

Proof. For $\xi' \in \mathscr{H}$ near β, let $-\omega_m(\xi')$ be the eigenvalues $i\mathscr{A}(\xi')$ — see (1-37) — and let $P_m(\xi')$ be the corresponding projectors; these objects depend smoothly on ξ' near β thanks to the analysis of [Métivier 2000]. Differentiate the equation

$$\left(\omega_m(\xi') I + \sum_{j=0}^{d-1} A_j \xi_j\right) P_m(\xi') = 0$$

with respect to ξ_j, evaluate at β, and apply P_m on the left to obtain

$$P_m A_j P_m = -\partial_{\xi_j} \omega_m(\beta) P_m,$$

from which the lemma readily follows. □

By Assumption 1.6 we know that the vector space $\ker B \cap \mathbb{E}^s(\beta)$ is one-dimensional; moreover, it admits a real basis because B has real coefficients and $\mathbb{E}^s(\beta)$ has a real basis. This vector space is therefore spanned by some $e \in \mathbb{R}^N \setminus \{0\}$ that we can decompose in a unique way by using Lemma 1.11:

$$\ker B \cap \mathbb{E}^s(\beta) = \operatorname{Span}\{e\}, \quad e = \sum_{m \in \mathcal{J}} e_m, \quad P_m e_m = e_m. \tag{1-38}$$

[11] The vector field X_{ϕ_m} is a constant multiple of the vector field $\partial_t + \boldsymbol{v}_m \cdot \nabla_{x''}$ computed by Lax for the Cauchy problem, where $\boldsymbol{v}_m$ is the group velocity defined in Definition 1.10.

Each vector e_m in (1-38) has real components. We also know that the vector space $B\mathbb{E}^s(\beta)$ is $(p-1)$-dimensional. We can therefore write it as the kernel of a real linear form:

$$B\mathbb{E}^s(\beta) = \{X \in \mathbb{C}^p,\ b \cdot X = 0\}, \tag{1-39}$$

for a suitable vector $b \in \mathbb{R}^p \setminus \{0\}$.

Any function $\mathcal{V}(x,\theta) \in H^{s;2}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$ can be decomposed:

$$\mathcal{V} = \underline{\mathcal{V}} + \mathcal{V}_{\mathrm{inc}} + \mathcal{V}_{\mathrm{out}} + \mathcal{V}_{\mathrm{nonch}} = \underline{\mathcal{V}} + \mathcal{V}^*,$$

where the terms correspond respectively to the parts of the Fourier series (1-23) with $\alpha = 0$, α incoming, α outgoing, and α noncharacteristic.[12]

Proposition 1.24. *Suppose $\mathcal{V}^0 \in H^{s;2}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$, $s \geq 1$, is a solution of (1-35). Then*

$$\underline{\mathcal{V}}^0 = 0, \quad \mathcal{V}^0_{\mathrm{out}} = 0, \quad \mathcal{V}^0_{\mathrm{nonch}} = 0, \quad \text{and so } \mathcal{V}^0 = \mathcal{V}^0_{\mathrm{inc}} = E\mathcal{V}^0_{\mathrm{inc}},$$
$$\mathcal{V}^0(x', 0, \theta_0, \dots, \theta_0) = a(x', \theta_0)e \quad \text{for some unknown periodic function } a \text{ with mean } 0.$$

Proof. Since $E\mathcal{V}^0 = \mathcal{V}^0$, we have $\mathcal{V}^0_{\mathrm{nonch}} = 0$. Applying E_0 to problem (1-35), we find that the mean value $\underline{\mathcal{V}}^0$ satisfies the weakly stable boundary problem

$$\begin{aligned} &L(\partial)\underline{\mathcal{V}}^0 + D(0)\underline{\mathcal{V}}^0 = 0, \\ &B\underline{\mathcal{V}}^0 = 0 \quad \text{on } x_d = 0, \\ &\underline{\mathcal{V}}^0 = 0 \quad \text{in } t < 0. \end{aligned}$$

By the well-posedness result of [Coulombel 2005] we have $\underline{\mathcal{V}}^0 = 0$.

Lemma 1.23 implies that outgoing profiles $\sigma_{m,k}$, $m \in \mathcal{O}$, in the expansion (1-28) of $\mathcal{V}^0$ satisfy problems of the form

$$X_{\phi_m}\sigma_{m,k} + \sum_{k'=1}^{\nu_{k_m}} (\ell_{m,k} \cdot D(0) r_{m,k'})\sigma_{m,k'} = 0,$$
$$\sigma_{m,k} = 0 \quad \text{in } t < 0,$$

where X_{ϕ_m} is an outgoing vector field. Thus $\sigma_{m,k} = 0$ for all $k = 1, \dots, \nu_{k_m}$.

The last statement of Proposition 1.24 follows immediately from the boundary condition in (1-35) and (1-38). □

Since $\mathcal{V}^0 = \mathcal{V}^0_{\mathrm{inc}}$, we obtain from (1-36)(a)

$$(I - E)\mathcal{V}^1 = (I - E)\mathcal{V}^1_{\mathrm{inc}} = -R(L(\partial)\mathcal{V}^0 + D(0)\mathcal{V}^0),$$

so

$$\mathcal{V}^1 = \underline{\mathcal{V}}^1 + \mathcal{V}^1_{\mathrm{inc}} + \mathcal{V}^1_{\mathrm{out}} \in H^{s;1}, \quad \text{where } E\mathcal{V}^1_{\mathrm{out}} = \mathcal{V}^1_{\mathrm{out}}.$$

[12] Here we say α is incoming if $\alpha \in \mathcal{C}_m \setminus 0$ for an index m such that ϕ_m is an incoming phase.

Next decompose the boundary condition (1-36)(c):

$$\begin{aligned} BE\mathcal{V}^1_{\mathrm{inc}} &= G^* - [(d\psi(0)\mathcal{V}^0)\mathcal{V}^0)]^* - B\mathcal{V}^1_{\mathrm{out}} - B(I-E)\mathcal{V}^1_{\mathrm{inc}} \\ &= G^* - [(d\psi(0)\mathcal{V}^0)\mathcal{V}^0)]^* - B\mathcal{V}^1_{\mathrm{out}} + BR(L(\partial)\mathcal{V}^0 + D(0)\mathcal{V}^0). \end{aligned} \tag{1-40}$$

Remark 1.25. (a) If $\mathcal{V}^1_{\mathrm{out}}|_{x_d=0,\theta_j=\theta_0}$ were known, one could write down a transport equation for $a(x',\theta_0)$ which is determined by the solvability condition for (1-40) implied by (1-39):

$$b\cdot(G^* - [(d\psi(0)\mathcal{V}^0)\mathcal{V}^0]^* - B\mathcal{V}^1_{\mathrm{out}} + BR(L(\partial)\mathcal{V}^0 + D(0)\mathcal{V}^0)) = 0. \tag{1-41}$$

However, the presence of the term $E((dD(0)\mathcal{V}^0)\mathcal{V}^0)$ in (1-36)(b) implies that two incoming modes in $\mathcal{V}^0_{\mathrm{inc}}$ (which is still unknown) can resonate to produce an outgoing mode that will affect $\mathcal{V}^1_{\mathrm{out}}$. Thus we do not know $\mathcal{V}^1_{\mathrm{out}}|_{x_d=0,\theta_j=\theta_0}$, and we see that the nonlinear boundary equation (1-41) is coupled to the nonlinear interior equation (1-36).

(b) If the phases are such that an outgoing mode can never be produced by a product of two incoming modes, $\mathcal{V}^1_{\mathrm{out}}$ can be determined from (1-36) to be 0, and one can proceed as in [Coulombel and Guès 2010] to solve for a without having to use Nash–Moser iteration.

The key subsystem to focus on now is (recalling $\mathcal{V}^0 = E\mathcal{V}^0 = \mathcal{V}^0_{\mathrm{inc}}$ and writing with obvious notation $E = E_0 + E_{\mathrm{inc}} + E_{\mathrm{out}}$)

$$\begin{aligned} &\text{(a)} \quad E_{\mathrm{inc}}(L(\partial)\mathcal{V}^0_{\mathrm{inc}} + D(0)\mathcal{V}^0_{\mathrm{inc}}) = 0, \\ &\text{(b)} \quad E_{\mathrm{out}}(L(\partial)\mathcal{V}^1_{\mathrm{out}} + D(0)\mathcal{V}^1_{\mathrm{out}} + (dD(0)\mathcal{V}^0_{\mathrm{inc}})\mathcal{V}^0_{\mathrm{inc}}) = 0, \\ &\text{(c)} \quad b\cdot(G^* - [(d\psi(0)\mathcal{V}^0_{\mathrm{inc}})\mathcal{V}^0_{\mathrm{inc}}]^* - B\mathcal{V}^1_{\mathrm{out}} + BR(L(\partial)\mathcal{V}^0_{\mathrm{inc}} + D(0)\mathcal{V}^0_{\mathrm{inc}})) = 0, \\ &\text{(d)} \quad \mathcal{V}^0_{\mathrm{inc}}(x',0,\theta_0,\dots,\theta_0) = a(x',\theta_0)e, \end{aligned} \tag{1-42}$$

where $\mathcal{V}^0_{\mathrm{inc}}$ and $\mathcal{V}^1_{\mathrm{out}}$ both vanish in $t<0$.

A formula for $\mathcal{V}^0_{\mathrm{inc}}$ in terms of $a(x',\theta_0)$ can be determined by solving transport equations using (1-42)(a), and that formula can be plugged into (1-42)(b) to get $\mathcal{V}^1_{\mathrm{out}}$ in terms of a. Thus the subsystem (1-42) can be expressed as a very complicated nonlinear, nonlocal equation for the single unknown a. This is done in Appendix B for a strictly hyperbolic example with only one resonance. However, that is not the way we solve (1-42); instead we solve the subsystem in its above form by iteration. Picard iteration does not work; there is a loss of derivatives from one iterate to the next (because of R), so we use a Nash–Moser scheme. An essential point is to take advantage of the smoothing property of the interaction integrals that pick out resonances in $E_{\mathrm{out}}((dD(0)\mathcal{V}^0_{\mathrm{inc}})\mathcal{V}^0_{\mathrm{inc}})$;[13] that property allows us to get tame estimates in Section 3.

An important tool in solving the subsystem (1-42) is the following result from [Coulombel and Guès 2010], which will allow us to write the boundary equation (1-42)(c) as a transport equation for $a(x',\theta_0)$.

Proposition 1.26 [Coulombel and Guès 2010, Proposition 3.5]. *Let the vectors b and e_m be as in (1-39) and (1-38), and let $\sigma(\zeta)$ be the function appearing in Assumption 1.6. There exists a nonzero real number*

[13] Interaction integrals are similar to convolution integrals.

κ such that

$$R_m P_m = 0 \quad \text{for all } m \in \{1, \dots, M\},$$

$$b \cdot B \sum_{m \in \mathcal{J}} R_m A_0 e_m = \kappa \partial_\tau \sigma(\underline{\tau}, \underline{\eta}) \quad \text{and} \quad \partial_\tau \sigma(\underline{\tau}, \underline{\eta}) = 1,$$

$$b \cdot B \sum_{m \in \mathcal{J}} R_m A_j e_m = \kappa \partial_{\eta_j} \sigma(\underline{\tau}, \underline{\eta}),\ j = 1, \dots, d-1,$$

and thus

$$b \cdot B \sum_{m \in \mathcal{J}} R_m L(\partial) e_m = \kappa \left(\partial_\tau \sigma(\underline{\tau}, \underline{\eta}) \partial_t + \sum_{j=1}^{d-1} \partial_{\eta_j} \sigma(\underline{\tau}, \underline{\eta}) \partial_{x_j} \right) =: X_{\mathrm{Lop}}.$$

Taking note of the denominator in_α in the definition (1-33) of R, we immediately obtain:

Corollary 1.27. *The boundary term $b \cdot BRL(\partial)\mathcal{V}^0_{\mathrm{inc}}$ in (1-42) may be written*

$$b \cdot BRL(\partial)\mathcal{V}^0_{\mathrm{inc}} = X_{\mathrm{Lop}} \mathcal{A},$$

where $\mathcal{A}(x', \theta_0)$ is the unique function with mean 0 in θ_0 such that $\partial_{\theta_0} \mathcal{A} = a$.

Remark 1.28. Proposition 1.26 shows that propagation in the boundary, which is described by $a(x', \theta_0)$, is governed by the (x-projection of the) Hamiltonian vector field associated to the Lopatinskii determinant. Since $\mathcal{V}^0(x', 0, \theta_0, \dots, \theta_0) = a(x', \theta_0)e$, this shows that waves of amplitude $O(1)$ emanate from the radiating boundary wave defined by a.

After (1-42) is solved, $\mathcal{V}^0$ is known, so $\underline{\mathcal{V}}^1$, $\mathcal{V}^1_{\mathrm{out}}$, and $(I - E)\mathcal{V}^1_{\mathrm{inc}}$ can now be determined by returning to the full system (1-36). The trace of $E\mathcal{V}^1_{\mathrm{inc}}$ is not yet determined; one should make a *choice* of $E\mathcal{V}^1_{\mathrm{inc}}|_{x_d=0, \theta_j=\theta_0}$ such that (1-40) holds, and then solve for $E\mathcal{V}^1_{\mathrm{inc}}$ using (1-36)(b). A precise description of the regularity of $\mathcal{V}^0$ and $\mathcal{V}^1$ is given in Theorem 5.11. The last piece of the corrected approximate solution, $\varepsilon^2 \mathcal{U}^2_p$ in (1-15), is discussed next.

1E. ***Error analysis.*** Given a periodic function $f(x, \theta)$, where $\theta = (\theta_1, \dots, \theta_M)$, let us denote

$$f(x, \theta)|_{\theta \to (\theta_0, \xi_d)} := f(x, \theta_0 + \underline{\omega}_1 \xi_d, \dots, \theta_0 + \underline{\omega}_M \xi_d);$$

so we have

$$f(x, \theta)|_{\theta \to (\phi_0/\varepsilon, x_d/\varepsilon)} = f\left(x, \frac{\phi}{\varepsilon}\right).$$

Taking the profiles $\mathcal{V}^0$, $\mathcal{V}^1$ constructed in Theorem 5.11, if we define

$$\mathcal{U}^b_\varepsilon(x, \theta_0) := (\mathcal{V}^0(x, \theta) + \varepsilon \mathcal{V}^1(x, \theta))|_{\theta \to (\theta_0, x_d/\varepsilon)},$$

we find that $\mathcal{U}^b_\varepsilon$ satisfies the singular system

$$\begin{aligned}
&\text{(a)} \quad \mathbb{L}_\varepsilon(\mathcal{U}^b_\varepsilon) := \partial_d \mathcal{U}^b_\varepsilon + \mathbb{A}\left(\partial_{x'} + \frac{\beta \partial_{\theta_0}}{\varepsilon}\right)\mathcal{U}^b_\varepsilon + D(\varepsilon \mathcal{U}^b_\varepsilon)\mathcal{U}^b_\varepsilon = O(\varepsilon), \\
&\text{(b)} \quad \psi(\varepsilon \mathcal{U}^b_\varepsilon)\mathcal{U}^b_\varepsilon = \varepsilon G(x', \theta_0) + O(\varepsilon^2) \quad \text{on } x_d = 0, \\
&\text{(c)} \quad \mathcal{U}^b_\varepsilon = 0 \quad \text{in } t < 0,
\end{aligned} \tag{1-43}$$

where the error terms refer to norms in E^s and H^t spaces whose orders are made precise in Section 4. For example, (1-43) follows directly from the profile equations (1-20)(a)–(b), together with the identity

$$\mathbb{L}_\varepsilon(f(x,\theta)|_{\theta\to(\theta_0,x_d/\varepsilon)}) = \frac{1}{\varepsilon}(\mathcal{L}(\partial_\theta)f(x,\theta))|_{\theta\to(\theta_0,x_d/\varepsilon)} + (L(\partial)f(x,\theta))|_{\theta\to(\theta_0,x_d/\varepsilon)} + (D(\varepsilon f)f)|_{\theta\to(\theta_0,x_d/\varepsilon)}. \quad (1\text{-}44)$$

Since our basic estimate for the linearized singular system exhibits a loss of one singular derivative (basically, we lose a $1/\varepsilon$ factor), the accuracy in (1-43)(a) is not good enough to conclude that

$$|U_\varepsilon - {}^{\mathcal{O}}u^b_\varepsilon|_{L^\infty(x,\theta_0)}$$

is small (the error terms are only $O(\varepsilon)$). Thus, to improve the accuracy, we construct an additional corrector ${}^{\mathcal{O}}u^2_p(x,\theta_0,\xi_d)$ and replace ${}^{\mathcal{O}}u^b_\varepsilon$ by

$${}^{\mathcal{O}}u_\varepsilon(x,\theta_0) := ({}^{\mathcal{O}}V^0(x,\theta) + \varepsilon {}^{\mathcal{O}}V^1(x,\theta))|_{\theta\to(\theta_0,x_d/\varepsilon)} + \varepsilon^2 {}^{\mathcal{O}}u^2_p\left(x,\theta_0,\frac{x_d}{\varepsilon}\right). \quad (1\text{-}45)$$

In constructing ${}^{\mathcal{O}}u^2_p$, we deal with the first (small divisor) problem described in Remark 1.22(b) by approximating ${}^{\mathcal{O}}V^0$ and ${}^{\mathcal{O}}V^1$ by trigonometric polynomials ${}^{\mathcal{O}}V^0_p$ and ${}^{\mathcal{O}}V^1_p$ to within an accuracy $\delta > 0$ in appropriate Sobolev norms, and seek ${}^{\mathcal{O}}u^2_p$ in the form of a trigonometric polynomial.[14] To deal with the second (solvability) problem, we use the following proposition, which allows us to use the profile equation (1-36)(b) as a solvability condition, in spite of the failure of Proposition 1.21 when $\mathcal{F} \in H^{s;2}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$. We define

$$\mathcal{L}_0(\partial_{\theta_0},\partial_{\xi_d}) := L(d\phi_0)\partial_{\theta_0} + \partial_{\xi_d}.$$

Proposition 1.29. *Suppose $F(x,\theta) \in H^{s;2}(\overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^M)$ has a Fourier series which is a finite sum and that $EF = 0$. Then there exists a solution of the equation*

$$\mathcal{L}_0(\partial_{\theta_0},\partial_{\xi_d}){}^{\mathcal{O}}u(x,\theta_0,\xi_d) = F(x,\theta)|_{\theta\to(\theta_0,\xi_d)} \quad (1\text{-}46)$$

in the form of a trigonometric polynomial in (θ_0,ξ_d) of the form

$${}^{\mathcal{O}}u(x,\theta_0,\xi_d) = \sum_{(\kappa_0,\kappa_d)\in\mathcal{J}} U_{\kappa_0,\kappa_d}(x)e^{i\kappa_0\theta_0 + i\kappa_d\xi_d}, \quad (1\text{-}47)$$

where $\mathcal{J}$ is a finite subset of $\mathbb{Z}\times\mathbb{R}$ and the coefficients U_{κ_0,κ_d} lie in $H^s(\overline{\mathbb{R}}^{d+1}_+)$.

The proof is given in Section 4. Observe that ${}^{\mathcal{O}}u$ is periodic in θ_0 but almost periodic in (θ_0,ξ_d). Proposition 1.29 is applied to solve the equation

$$\mathcal{L}_0(\partial_{\theta_0},\partial_{\xi_d}){}^{\mathcal{O}}u^2_p = [-(I-E)(L(\partial){}^{\mathcal{O}}V^1_p + D(0){}^{\mathcal{O}}V^1_p + (\mathrm{d}D(0){}^{\mathcal{O}}V^0_p){}^{\mathcal{O}}V^0_p)]|_{\theta\to(\theta_0,\xi_d)}.$$

[14] Trigonometric polynomial approximations were already used to deal with small divisor problems in the error analysis of [Joly et al. 1995].

With this choice of $\mathcal{U}_p^2$ we show in Section 4 that the new approximate solution $\mathcal{U}_\varepsilon(x,\theta_0)$ in (1-45) satisfies instead of (1-43) the singular system

$$\begin{aligned} &\text{(a)} \quad \mathbb{L}_\varepsilon(\mathcal{U}_\varepsilon) = O(\varepsilon(K\delta + C(\delta)\varepsilon)),\\ &\text{(b)} \quad \psi(\varepsilon\mathcal{U}_\varepsilon)\mathcal{U}_\varepsilon - \varepsilon G(x',\theta_0) = O(\varepsilon^2 C(\delta)) \quad \text{on } x_d = 0,\\ &\text{(c)} \quad \mathcal{U}_\varepsilon = 0 \quad \text{in } t < 0, \end{aligned} \tag{1-48}$$

where the errors in (1-48)(a)–(b) are measured in appropriate norms. Now one can apply our basic estimate (2-41) for the linearized singular problem to conclude that the difference between exact and approximate solutions of the semilinear singular system (1-18) satisfies, for some constants $C(\delta)$ and K,

$$|U_\varepsilon(x,\theta_0) - \mathcal{U}_\varepsilon(x,\theta_0)|_{E^s} \le K\delta + C(\delta)\varepsilon, \quad \text{for some } s > \frac{d+1}{2}.$$

This estimate clearly implies the conclusion of Theorem 1.15 by choosing first $\delta > 0$ small enough and then letting ε tend to zero (this is the same final argument as in [Joly et al. 1995]).

1F. ***Remarks on quasilinear problems.*** In this article, we are able to rigorously justify a weakly nonlinear regime with amplification for *semilinear* hyperbolic initial boundary value problems. Our assumptions only deal with the principal part of the operators, meaning that we only assume a weak stability property for the problem $(L(\partial), B)$ obtained by linearizing at the origin and dropping the zero-order term in the hyperbolic system. The weak stability is of WR type in the terminology of [Benzoni-Gavage et al. 2002]. Despite the weak regime that we consider ($O(\varepsilon^2)$ source term at the boundary and $O(\varepsilon)$ solution), the leading profile equation displays some *quasilinear* features. We emphasize that the regime that we consider here is exactly one power of ε weaker than the weakly nonlinear regime for the semilinear Cauchy problem or for semilinear uniformly stable boundary value problems. As in [Coulombel and Guès 2010], this power of ε corresponds exactly to the loss of one derivative in the energy estimates.

We believe that the techniques developed here can be extended to give a rigorous justification of weakly nonlinear geometric optics with amplification for *quasilinear* hyperbolic initial boundary value problems of the form

$$\partial_t v + \sum_{j=1}^{d} B_j(v)\partial_j v + f_0(v) = 0, \tag{1-49}$$

$$\phi(v) = \varepsilon^3 G\left(x', \frac{x'\cdot\beta}{\varepsilon}\right) \qquad \text{on } x_d = 0, \tag{1-50}$$

$$v = 0 \quad \text{and} \quad G = 0 \qquad \text{in } t < 0. \tag{1-51}$$

The corresponding solution v_ε would be of amplitude $O(\varepsilon^2)$. In particular the arguments used in Section 2 to obtain uniform estimates with a loss of one singular derivative for the singular initial boundary value problem might be extended to the corresponding singular quasilinear problem. There are however several new obstacles along the way, one of which is to extend the singular pseudodifferential calculus of [Coulombel et al. 2012] in order to obtain a two-terms expansion of $(1,0)$ and $(0,1)$ compositions. The weaker scaling (ε^2 in place of ε) should be sufficient to obtain the appropriate results. Let us observe

that, for $O(\varepsilon^2)$ solutions, the principal part of the hyperbolic operator has coefficients that are uniformly bounded in $W^{2,\infty}$, which is precisely the regularity needed in [Coulombel 2004; 2005] to obtain a priori estimates and well-posedness. The leading profile equation obtained in this quasilinear framework is very similar to the one we have derived here, and we thus believe that a weak well-posedness result using Nash–Moser iteration should prove the existence of the leading profile. For all the above reasons, we thus believe that the ε^3 source term on the boundary is the relevant "weakly nonlinear regime with amplification" in the quasilinear case, and we postpone the verification of the many technical details to a future work. Unfortunately, this regime would still be beyond the one considered in [Artola and Majda 1987; Majda and Rosales 1983], so there would still be a new ingredient to incorporate in order to justify the calculations of these papers.

2. Exact oscillatory solutions on a fixed time interval

2A. ***The basic estimate for the linearized singular system.*** In this section, it is our goal to prove Proposition 2.2 and its time-localized version, that is, Proposition 2.9. These propositions provide the a priori estimates for the linearized singular system that form the basis for the Nash–Moser iteration of Section 5B and the error analysis of Section 4.

We begin by gathering some of the notation for spaces and norms that is needed below.

Notation 2.1. Here we take $s \in \mathbb{N} = \{0, 1, 2, \dots\}$.

(a) Let $\Omega := \overline{\mathbb{R}}^{d+1}_+ \times \mathbb{T}^1$, $\Omega_T := \Omega \cap \{-\infty < t < T\}$, $b\Omega := \mathbb{R}^d \times \mathbb{T}^1$, $b\Omega_T := b\Omega \cap \{-\infty < t < T\}$, and set $\omega_T := \overline{\mathbb{R}}^{d+1}_+ \cap \{-\infty < t < T\}$.

(b) Let $H^s \equiv H^s(b\Omega)$, the standard Sobolev space with norm $\langle V(x', \theta_0)\rangle_s$. For $\gamma \geq 1$ we set $H^s_\gamma := e^{\gamma t} H^s$ and $\langle V\rangle_{s,\gamma} := \langle e^{-\gamma t} V\rangle_s$.

(c) $L^2 H^s \equiv L^2(\overline{\mathbb{R}}_+, H^s(b\Omega))$ with norm $|U(x, \theta_0)|_{L^2 H^s} \equiv |U|_{0,s}$ given by

$$|U|^2_{0,s} = \int_0^\infty |U(x', x_d, \theta_0)|^2_{H^s(b\Omega)} dx_d.$$

The corresponding norm on $L^2 H^s_\gamma$ is denoted by $|V|_{0,s,\gamma}$.

(d) $CH^s \equiv C(\overline{\mathbb{R}}_+, H^s(b\Omega))$ denotes the space of continuous bounded functions of x_d with values in $H^s(b\Omega)$, with norm

$$|U(x, \theta_0)|_{CH^s} = |U|_{\infty,s} := \sup_{x_d \geq 0} |U(., x_d, .)|_{H^s(b\Omega_T)}$$

(note that $CH^s \subset L^\infty H^s$). The corresponding norm on CH^s_γ is denoted by $|V|_{\infty,s,\gamma}$.

(e) Let $M_0 := 3d+5$ and define $C^{0,M_0} := C(\overline{\mathbb{R}}_+, C^{M_0}(b\Omega))$ as the space of continuous bounded functions of x_d with values in $C^{M_0}(b\Omega)$, with norm $|U(x, \theta_0)|_{C^{0,M_0}} := |U|_{L^\infty W^{M_0,\infty}}$. Here $L^\infty W^{M_0,\infty}$ denotes the space $L^\infty(\overline{\mathbb{R}}_+; W^{M_0,\infty}(b\Omega))$.[15]

[15] The size of M_0 is determined by the requirements of the singular calculus described in Appendix A.

(f) The corresponding spaces on Ω_T are denoted by $L^2H^s_T$, $L^2H^s_{\gamma,T}$, CH^s_T, $CH^s_{\gamma,T}$ and C^{0,M_0}_T with norms $|U|_{0,s,T}$, $|U|_{0,s,\gamma,T}$, $|U|_{\infty,s,T}$, $|U|_{\infty,s,\gamma,T}$, and $|U|_{C^{0,M_0}_T}$, respectively. On $b\Omega_T$ we use the spaces H^s_T and $H^s_{\gamma,T}$ with norms $\langle U\rangle_{s,T}$ and $\langle U\rangle_{s,\gamma,T}$.

(g) All constants appearing in the estimates below are independent of ε, γ, and T unless such dependence is explicitly noted.

The linearization of the singular problem (1-18) at $U(x,\theta_0)$ has the form

$$\begin{aligned}
&\text{(a)}\quad \partial_d\dot U_\varepsilon + \mathbb{A}\Big(\partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\Big)\dot U_\varepsilon + \mathscr{D}(\varepsilon U)\dot U_\varepsilon = f(x,\theta_0) \quad \text{on } \Omega,\\
&\text{(b)}\quad \mathscr{B}(\varepsilon U)\dot U_\varepsilon|_{x_d=0} = g(x',\theta_0),\\
&\text{(c)}\quad \dot U_\varepsilon = 0 \quad \text{in } t<0,
\end{aligned} \tag{2-1}$$

where the matrices $\mathscr{B}(\varepsilon U)$, $\mathscr{D}(\varepsilon U)$ are defined in (1-11).[16] Instead of (2-1), consider the equivalent problem satisfied by $\dot U^\gamma := e^{-\gamma t}\dot U$:

$$\begin{aligned}
&\partial_d\dot U^\gamma + \mathbb{A}\Big((\partial_t+\gamma,\partial_{x''}) + \frac{\beta\partial_{\theta_0}}{\varepsilon}\Big)\dot U^\gamma + \mathscr{D}(\varepsilon U)\dot U^\gamma = f^\gamma(x,\theta_0),\\
&\mathscr{B}(\varepsilon U)\dot U^\gamma|_{x_d=0} = g^\gamma(x',\theta_0),\\
&\dot U^\gamma = 0 \quad \text{in } t<0.
\end{aligned} \tag{2-2}$$

Below we let Λ_D denote the singular Fourier multiplier (see (A-2)) associated to the symbol

$$\Lambda(X,\gamma) := \Big(\gamma^2 + \Big|\xi' + \frac{k\beta}{\varepsilon}\Big|^2\Big)^{1/2}, \quad X := \xi' + \frac{k\beta}{\varepsilon}. \tag{2-3}$$

The basic estimate for the linearized singular problem (2-2) is given in the next proposition. Observe that the estimate (2-4) exhibits a loss of one "singular derivative" Λ_D. In view of [Coulombel and Guès 2010, Theorem 4.1], there is strong evidence that the loss below is optimal.

Proposition 2.2 (main L^2 linear estimate). *We make the structural assumptions of Theorem 1.15 and recall $M_0 = 3d+5$. Fix $K>0$ and suppose $|\varepsilon\partial_d U|_{C^{0,M_0-1}} + |U|_{C^{0,M_0}} \le K$ for $\varepsilon\in(0,1]$. There exist positive constants $\varepsilon_0(K)>0$, $C(K)>0$, and $\gamma_0(K)\ge 1$ such that sufficiently smooth solutions $\dot U$ of the linearized singular problem (2-1) satisfy*[17]

$$|\dot U^\gamma|_{0,0} + \frac{\langle\dot U^\gamma\rangle_0}{\sqrt\gamma} \le C(K)\Big(\frac{|\Lambda_D f^\gamma|_{0,0} + |\varepsilon^{-1}f^\gamma|_{0,0}}{\gamma^2} + \frac{\langle\Lambda_D g^\gamma\rangle_0 + \langle\varepsilon^{-1}g^\gamma\rangle_0}{\gamma^{3/2}}\Big), \tag{2-4}$$

for $\gamma\ge\gamma_0(K)$, $0<\varepsilon\le\varepsilon_0(K)$.

The same estimate holds if $\mathscr{B}(\varepsilon U)$ in (2-1) is replaced by $\mathscr{B}(\varepsilon U,\varepsilon\mathfrak{U})$ and $\mathscr{D}(\varepsilon U)$ is replaced by $\mathscr{D}(\varepsilon U,\varepsilon\mathfrak{U})$, as long as $|\varepsilon\partial_d(U,\mathfrak{U})|_{C^{0,M_0-1}} + |U,\mathfrak{U}|_{C^{0,M_0}} \le K$ for $\varepsilon\in(0,1]$.

[16] Here and below we often suppress the subscript ε on $\dot U$.

[17] Note that the norms $|u|_{0,1}$ and $|\Lambda_D u|_{0,0}$ are not equivalent.

Corollary 2.3 (main H^1_{tan} linear estimate). *Under the same assumptions as in Proposition 2.2, smooth enough solutions $\dot{U}$ of the linearized singular problem (2-1) satisfy*

$$|\dot{U}^\gamma|_{\infty,0} + |\dot{U}^\gamma|_{0,1} + \frac{\langle \dot{U}^\gamma\rangle_1}{\sqrt{\gamma}} \le C(K)\left(\frac{|\Lambda_D f^\gamma|_{0,1} + |\varepsilon^{-1} f^\gamma|_{0,1}}{\gamma^2} + \frac{\langle \Lambda_D g^\gamma\rangle_1 + \langle \varepsilon^{-1} g^\gamma\rangle_1}{\gamma^{3/2}}\right), \tag{2-5}$$

for $\gamma \ge \gamma_0(K)$, $0 < \varepsilon \le \varepsilon_0(K)$.

Short guide to the proof. The proof of Proposition 2.2 is completed using the next two propositions, each of which has the same hypotheses as Proposition 2.2. In the first step of the proof of Proposition 2.2, we choose a partition of unity defined by frequency cutoffs $\chi_i(\zeta)$, $i = 1, \dots, N_1 + N_2$, such that for $i = 1, \dots, N_1$ the function χ_i is supported near a point of the "bad" set $\overline{\Upsilon}$, while for $i > N_1$ the function χ_i is supported away from $\overline{\Upsilon}$. The estimates of $\chi_{i,D}\dot{U}^\gamma$ for $i > N_1$ are done in Proposition 2.8. For such indices, Kreiss symmetrizers in the singular calculus are used to estimate $\chi_{i,D}\dot{U}^\gamma$ without loss.

Proof of Proposition 2.2. (I): *Partition of unity*. The compactness of $\overline{\Upsilon}$ (see Assumption 1.12) and Σ allows us to choose a finite open covering of Σ, $\mathscr{C} = \{\mathscr{V}_i\}_{i=1,\dots,N_1+N_2}$ such that $\{\mathscr{V}_i\}_{i=1,\dots,N_1}$ covers $\overline{\Upsilon}$ and such that $\bigcup_{N_1+1}^{N_1+N_2} \mathscr{V}_i$ is disjoint from a neighborhood of $\overline{\Upsilon}$. Since $\overline{\Upsilon} \subset \mathscr{H}$, we can arrange so that for each $i \in \{1, \dots, N_1\}$ there is a conjugator $Q_{0,i}(\zeta)$[18] and diagonal matrix $\mathbb{D}_{1,i}(\zeta)$ satisfying (1-8) in $\mathscr{V}_i$. Moreover, we can choose a neighborhood $\mathbb{O}$ of $(0, 0) \in \mathbb{R}^{2N}$ such that for each $i \le N_1$ there are functions σ_i, $P_{i,1}$, and $P_{i,2}$ on $\mathbb{O} \times \mathscr{V}_i$ with the properties described in Assumption 1.12. For these symbols, we shall use the substitution $(v_1, v_2) \to (\varepsilon U(x, \theta_0), \varepsilon U(x, \theta_0))$ to prescribe the space dependence.[19]

We let $\chi_i(\zeta)$, $i = 1, \dots, N_1 + N_2$ be a smooth partition of unity subordinate to $\mathscr{C}$, and extend the χ_i to all ζ as functions homogeneous of degree zero. We smoothly extend each $Q_{0,i}$ (as a matrix with bounded inverse) first to Σ, and then to all ζ as a function homogenous of degree zero. We take similar extensions in ζ of $P_{i,1}$, $P_{i,2}$, $\mathbb{D}_{1,i}$, and σ_i, but with homogeneity of degree 1 in the cases of $\mathbb{D}_{1,i}$ and σ_i. As with $Q_{0,i}$, the extensions of $P_{i,1}$ and $P_{i,2}$ are taken to have bounded inverses.[20] Of course, for a given $i \le N_1$, the property (1-12) is satisfied only for $\zeta/|\zeta| \in \mathscr{V}_i$.

(II): *Estimate near the bad set*. The first estimate deals with a piece of $\dot{U}^\gamma$ that is microlocalized near the bad set Υ.

Proposition 2.4. *Fix i such that $1 \le i \le N_1$, let $\dot{U}_1^\gamma := \chi_{i,D}\dot{U}^\gamma$ and write*

$$\dot{U}_1^\gamma = \dot{U}_{1,\text{in}}^\gamma + \dot{U}_{1,\text{out}}^\gamma,$$

where[21]

$$\dot{U}_{1,\text{in}}^\gamma := (Q_D)^{-1}(w_{\text{in}}, 0) \quad \text{and} \quad \dot{U}_{1,\text{out}}^\gamma := (Q_D)^{-1}(0, w_{\text{out}}).$$

[18] Recall the notation $\zeta = (\tau - i\gamma, \eta)$. Sometimes we also write $\zeta = (\xi', \gamma)$ to match the notation of [Coulombel et al. 2012].

[19] The substitution $(v_1, v_2) \to (\varepsilon U(x, \theta_0), \varepsilon^0 u(x, \theta_0))$ is also used at one point.

[20] Taking such extensions reduces the number of cutoff functions we need later.

[21] Here Q_D, $w_{\text{in}} \in \mathbb{C}^p$, and $w_{\text{out}} \in \mathbb{C}^{N-p}$ are defined by the diagonalization procedure explained in the proof.

Then we have

$$|\dot U^\gamma_{1,\mathrm{in}}|_{0,0}+\frac{|\dot U^\gamma_{1,\mathrm{in}}|_{\infty,0}}{\sqrt\gamma}+\frac{|(\Lambda_D,\varepsilon^{-1})\dot U^\gamma_{1,\mathrm{out}}|_{0,0}}{\gamma}+\frac{|(\Lambda_D,\varepsilon^{-1})\dot U^\gamma_{1,out}|_{\infty,0}}{\gamma^{3/2}}$$
$$+|(\varepsilon\Lambda_D)^{-1}\dot U^\gamma_{1,\mathrm{in}}|_{0,0}+\frac{\langle(\varepsilon\Lambda_D)^{-1}\dot U^\gamma_{1,\mathrm{in}}|_{x_d=0}\rangle_0}{\sqrt\gamma}$$
$$\le C\bigg(\frac{|(\Lambda_D,\varepsilon^{-1})f^\gamma|_{0,0}}{\gamma^2}+\frac{\langle(\Lambda_D,\varepsilon^{-1})g^\gamma\rangle_0}{\gamma^{3/2}}+\frac{|\dot U^\gamma|_{0,0}+|(\varepsilon\Lambda_D)^{-1}\dot U^\gamma|_{0,0}}{\gamma^2}$$
$$+\frac{\langle\dot U^\gamma|_{x_d=0}\rangle_0+\langle(\varepsilon\Lambda_D)^{-1}\dot U^\gamma|_{x_d=0}\rangle_0}{\gamma^{3/2}}\bigg). \tag{2-6}$$

Proof of Proposition 2.4. The loss of derivatives in the estimate prevents us from treating the zero-order term $\mathcal{D}(\varepsilon U)\dot U^\gamma$ as a forcing term, as we would in a uniformly stable problem. Thus we need to use an argument that simultaneously diagonalizes $\mathbb{A}$ and the lower-order term $\mathcal{D}(\varepsilon U)$.

We now set $\chi_i=\boldsymbol{\chi}$, $v:=\boldsymbol{\chi}_D\dot U^\gamma=\dot U^\gamma_1$, and estimate v. We let $\mathbb{A}(X,\gamma)=-\mathcal{A}(X,\gamma)$ denote the singular symbol such that

$$\mathbb{A}_D=\mathbb{A}\bigg((\partial_t+\gamma,\partial_{x''})+\frac{\beta\partial_{\theta_0}}{\varepsilon}\bigg).$$

Dropping superscripts γ, we see from (2-2) that v satisfies

$$\begin{aligned}&\partial_d v+\mathbb{A}_D v+\mathcal{D}(\varepsilon U)v=\boldsymbol{\chi}_D f+[\mathcal{D}(\varepsilon U),\boldsymbol{\chi}_D]\dot U=\boldsymbol{\chi}_D f+r_{-1,D}\dot U,\\&\mathcal{B}(\varepsilon U)v|_{x_d=0}=\boldsymbol{\chi}_D g+[\mathcal{B}(\varepsilon U),\boldsymbol{\chi}_D]\dot U|_{x_d=0}=\boldsymbol{\chi}_D g+r_{-1,D}\dot U|_{x_d=0}.\end{aligned} \tag{2-7}$$

Here and below $r_{-1,D}$ denotes a singular operator of order -1 (which can change from one occurrence to the next) computed using the singular calculus. Similarly, $r_{0,D}$ will denote an operator of order 0. In spite of the loss of the factor Λ_D in the estimate (2-4), we are able to treat $r_{-1,D}\dot U$ as a forcing term (see, for example, (2-16) below). A term like $r_{0,D}\dot U/\gamma$ would be too large to absorb.

The first several steps of the proof estimate the terms in the first line of (2-6).

Step 1: *Simultaneous diagonalization.* This diagonalization argument is similar to the one in [Coulombel 2004]. Let $Q_0(\zeta):=Q_{0,i}(\zeta)$ and $\mathbb{D}_1(\zeta):=\mathbb{D}_{1,i}(\zeta)$ be the matrices as in (1-8) such that

$$Q_0(\zeta)\mathbb{A}(\zeta)Q_0^{-1}(\zeta)=\mathbb{D}_1(\zeta)$$

in the conical extension of $\mathcal{V}_i$. We define

$$w:=Q_D v,$$

where $Q=Q_0(X,\gamma)+Q_{-1}(\varepsilon U,X,\gamma)$. Here the matrix $Q_{-1}(\varepsilon U,\zeta)$ is a symbol of order -1 defined for all ζ, but chosen so that, on the conical extension of $\mathcal{V}_i$, the matrix

$$\mathbb{D}_0(\varepsilon U,\zeta):=[Q_{-1}Q_0^{-1},\mathbb{D}_1]+Q_0\mathcal{D}(\varepsilon U)Q_0^{-1} \tag{2-8}$$

is block diagonal, necessarily of order 0, with blocks of the same dimensions $n_1,\dots,n_J$ as those of $\mathbb{D}_1$. Since the eigenvalues associated to the blocks of $\mathbb{D}_1$ are mutually distinct, a direct computation shows that $Q_{-1}Q_0^{-1}$, and thus Q_{-1}, can be chosen so that the commutator cancels the off-diagonal blocks of

$Q_0\mathcal{D}(\varepsilon U)Q_0^{-1}$. (The diagonal blocks of the commutator are all zero blocks and therefore cannot cancel those of $Q_0\mathcal{D}(\varepsilon U)Q_0^{-1}$.) Since $Q_0\mathbb{A} = \mathbb{D}_1 Q_0$ on $\mathcal{V}_i$, (2-8) implies the relation

$$Q\mathbb{A} + Q_0\mathcal{D} = \mathbb{D}_1 Q + [Q_{-1}Q_0^{-1}, \mathbb{D}_1]Q_0 + Q_0\mathcal{D} = \mathbb{D}_1 Q + \mathbb{D}_0 Q_0. \tag{2-9}$$

Remark 2.5. (1) The scalar entries of the matrix $Q_{-1,D}$ can be chosen to have the form

$$(Q_{-1,D})_{i,j} = c(\varepsilon U)a_{-1,D},$$

where $a_{-1}(\zeta)$ is of order -1 and independent of (x, θ), thus giving rise to a Fourier multiplier.

(2) Since $(Q_{0,D})^{-1}Q_{-1,D}$ has norm less than one as an operator on L^2 for γ large, we can define $(Q_D)^{-1}$ as an operator on L^2 using a Neumann series.

Noting that x-dependence is absent in $\mathbb{A}$ and Q_0 and using the commutation property (2-9), we have

$$\begin{aligned}
\partial_d w &= Q_D\partial_d v + (\partial_d Q_{-1})_D v = -Q_D(\mathbb{A} + \mathcal{D}(\varepsilon U))_D v + Q_D\boldsymbol{\chi}_D f + r_{-1,D}\dot{U} + (\partial_d Q_{-1})_D v \\
&= -(Q\mathbb{A} + Q_0\mathcal{D}(\varepsilon U))_D v + Q_D\boldsymbol{\chi}_D f + r_{-1,D}\dot{U} + (\partial_d Q_{-1})_D v \\
&= -(\mathbb{D}_1 Q + \mathbb{D}_0 Q_0)_D v + Q_D\boldsymbol{\chi}_D f + r_{-1,D}\dot{U} + (\partial_d Q_{-1})_D v \\
&= -(\mathbb{D}_1 + \mathbb{D}_0)_D w + r_{0,D} f + r_{-1,D}\dot{U} + R_D^a v.
\end{aligned} \tag{2-10}$$

In the final line of (2-10), the operator $r_{-1,D}$ is explicitly given by

$$Q_D[\mathcal{D}(\varepsilon U), \boldsymbol{\chi}_D]\dot{U} - (Q_{0,D}\mathcal{D}(\varepsilon U) - (Q_0\mathcal{D}(\varepsilon U))_D)\boldsymbol{\chi}_D\dot{U} - Q_{-1,D}\mathcal{D}(\varepsilon U)\boldsymbol{\chi}_D\dot{U} + \mathbb{D}_{0,D}Q_{-1,D}\boldsymbol{\chi}_D\dot{U}, \tag{2-11}$$

and the second remainder term is decomposed as $R_D^a = R_D^b + R_D^c$ with operators R_D^b, R_D^c defined by

$$\begin{aligned}
&\text{(a)} \quad R_D^b v := (\partial_d Q_{-1})_D v, \\
&\text{(b)} \quad R_D^c v := \mathbb{D}_{1,D}(Q_{-1})_D v - (\mathbb{D}_1 Q_{-1})_D v.
\end{aligned} \tag{2-12}$$

In view of Remark 2.5 the scalar entries of R_D^b and R_D^c have the form

$$(\partial_d c(\varepsilon U))a_{-1,D} \quad \text{and} \quad [\alpha_{1,D}, c(\varepsilon U)]a_{-1,D}, \tag{2-13}$$

respectively. In (2-13), $\alpha_1(\zeta)$ denotes one of the diagonal entries of $\mathbb{D}_1(\zeta)$. Here and below $a_{-1,D}$ denotes a singular operator of order -1 associated to a symbol $a_{-1}(\zeta)$ which may change from term to term.

The precise estimate of the above remainder terms is one of the keys to the proof of Proposition 2.4.

Lemma 2.6. *The remainder terms $r_{-1,D}\dot{U}$ and $R_D^a v$ in the last line of (2-10) satisfy estimates of the form*

$$\begin{aligned}
&|r_{-1,D}\dot{U}|_{0,0} \le C(K)|\Lambda_D^{-1}\dot{U}|_{0,0}, \quad |R_D^a v|_{0,0} \le C(K)|\Lambda_D^{-1}v|_{0,0}, \\
&|\Lambda_D r_{-1,D}\dot{U}|_{0,0} \le C(K)|\dot{U}|_{0,0}, \quad |\Lambda_D R_D^a v|_{0,0} \le C(K)(|v|_{0,0} + |(\varepsilon\Lambda_D)^{-1}v|_{0,0}),
\end{aligned}$$

with a constant $C(K)$ that is uniform with respect to ε and γ.

Proof of Lemma 2.6. • The estimate of $R_D^a v$ in L^2 comes from the expression (2-13) of the coefficients of R_D^b and R_D^c. For instance, the commutator $[\alpha_{1,D}, c(\varepsilon U)]$ is bounded on L^2 uniformly on ε, γ (see

Appendix A), and we can isolate a Fourier multiplier $a_{-1,D}$ on the right. In particular, we obtain the weaker estimate

$$\gamma|R_D^a v|_{0,0} \le C|v|_{0,0},$$

and we are now going to estimate the singular derivative $(\partial_{x'} + \beta\partial_{\theta_0}/\varepsilon)R_D^a v$. Let us deal with the operator R_D^c (the estimate involving R_D^b is similar). When applying the singular derivative, we need to estimate terms of the form

$$[\alpha_{1,D}, c(\varepsilon U)]\left(\partial_{x_j} + \frac{\beta_j\partial_{\theta_0}}{\varepsilon}\right)a_{-1,D}v + [\alpha_{1,D}, \partial_{x_j}c(\varepsilon U)]a_{-1,D}v + \frac{\beta_j}{\varepsilon}[\alpha_{1,D}, \partial_{\theta_0}c(\varepsilon U)]a_{-1,D}v.$$

The first term is estimated by v in L^2, while the second and, above all, the third term are estimated by $(\varepsilon\Lambda_D)^{-1}v$ in L^2.

• The estimate of $\Lambda_D r_{-1,D}\dot U$ is precisely the definition of the notation $r_{-1,D}$ and it follows from the rules of symbolic calculus; see Appendix A. We thus focus on the L^2 estimate of the remainder where we wish to gain a factor Λ_D^{-1} rather than a mere $1/\gamma$. Let us first consider the term $Q_{-1,D}\mathcal{D}(\varepsilon U)\boldsymbol{\chi}_D\dot U$ in (2-11). We write

$$Q_{-1,D}\mathcal{D}(\varepsilon U)\boldsymbol{\chi}_D\dot U = Q_{-1,D}(\mathcal{D}(\varepsilon U)\Lambda_D)\boldsymbol{\chi}_D\Lambda_D^{-1}\dot U = Q_{-1,D}(\mathcal{D}(\varepsilon U)\Lambda)_D\boldsymbol{\chi}_D\Lambda_D^{-1}\dot U = r_{0,D}\Lambda_D^{-1}\dot U,$$

where we have applied the symbolic calculus rule in the end for the $(-1, 1)$ product. Similarly, we can write the first commutator in (2-11) as

$$Q_D[\mathcal{D}(\varepsilon U), \boldsymbol{\chi}_D]\dot U = r_{0,D}[(\mathcal{D}(\varepsilon U)\Lambda)_D, \boldsymbol{\chi}_D]\Lambda_D^{-1}\dot U = r_{0,D}\Lambda_D^{-1}\dot U.$$

We leave to the reader the other two terms in (2-11) that can be treated in an analogous way. Eventually, we can write the term $r_{-1,D}\dot U$ in the last line of (2-10) as $r_{0,D}\Lambda_D^{-1}\dot U$ and the L^2 estimate follows. □

The estimates of Lemma 2.6 seem to be the best we can hope for in the case of the bad $(1, -1)$ product (2-12)(b), which is the reason for the need to estimate such terms as those on the left of inequality (2-6).

Step 2: *Outgoing modes.* Recall that $-\mathbb{D}_1$ and $-\mathbb{D}_0$ are block diagonal:

$$-\mathbb{D}_1(\zeta) = \begin{pmatrix} i\omega_1(\zeta)I_{n_1} & & 0 \\ & \ddots & \\ 0 & & i\omega_J(\zeta)I_{n_J} \end{pmatrix}, \quad -\mathbb{D}_0(\varepsilon U, \zeta) = \begin{pmatrix} C_1 & & 0 \\ & \ddots & \\ 0 & & C_J \end{pmatrix},$$

so the system (2-10) satisfied by $w = (w_1, \dots, w_J)$ can be written as a collection of J decoupled transport equations

$$\partial_d w_j = (i\omega_j)_D w_j + C_{j,D}w_j + r_{0,D}f + r_{-1,D}\dot U + R_D^a\dot U \tag{2-14}$$

with $\mathrm{Re}(i\omega_j) \le -c\gamma$ for $1 \le j \le J'$, and $\mathrm{Re}(i\omega_j) \ge c\gamma$ for $J'+1 \le j \le J$ ($c > 0$ denotes a constant).

Following the strategy of [Coulombel 2004], we now give two preliminary estimates of the outgoing modes w_j, $j \geq J'+1$. Taking the real part of the $L^2(\Omega)$ inner product of (2-14) with $-\Lambda_D^2 w_j$, we obtain

$$-\frac{\langle \Lambda_D w_j(0)\rangle_0^2}{2} = \mathrm{Re}(\Lambda_D (i\omega_j)_D w_j, \Lambda_D w_j)_{L^2(\Omega)} + \mathrm{Re}(\Lambda_D C_{j,D} w_j, \Lambda_D w_j)_{L^2(\Omega)} + \mathrm{Re}(\Lambda_D r_{0,D} f, \Lambda_D w_j)_{L^2(\Omega)} + \mathrm{Re}(\Lambda_D r_{-1,D} \dot{U}, \Lambda_D w_j)_{L^2(\Omega)} + \mathrm{Re}(\Lambda_D R_D^a \dot{U}, \Lambda_D w_j)_{L^2(\Omega)}.$$

Since $\mathrm{Re}(i\omega_j) \geq c\gamma$, we get, after absorbing some terms on the left,

$$\gamma |\Lambda_D w_{\mathrm{out}}|_{0,0}^2 + \langle \Lambda_D w_{\mathrm{out}}(0)\rangle_0^2 \leq \frac{C}{\gamma}(|\Lambda_D f|_{0,0}^2 + |\dot{U}|_{0,0}^2 + |(\varepsilon \Lambda_D)^{-1}\dot{U}|_{0,0}^2). \tag{2-15}$$

Here, for example, we have used Young's inequality and Lemma 2.6 and estimated

$$|\mathrm{Re}(\Lambda_D R_D^a \dot{U}, \Lambda_D w_j)_{L^2(\Omega)}| \leq \frac{C_\delta}{\gamma}(|\dot{U}|_{0,0}^2 + |(\varepsilon\Lambda_D)^{-1}\dot{U}|_{0,0}^2) + \delta\gamma |\Lambda_D w_j|_{0,0}^2. \tag{2-16}$$

Taking the real part of the L^2 inner product of (2-14) with w_j on $[x_d, \infty) \times b\Omega$ instead of Ω, we obtain, for all $x_d \geq 0$,

$$\gamma |w_j|_{0,0}^2 + \langle w_j(x_d)\rangle_0^2 \leq \frac{C}{\gamma}\left(|f|_{0,0}^2 + \frac{1}{\gamma^2}|\dot{U}|_{0,0}^2\right). \tag{2-17}$$

Finally, adding to (2-15) the estimate $\gamma^2 \times$(2-17) and the estimates we obtain in the same way by pairing (2-14) with w_j/ε^2 (here we use the L^2 estimate of the remainders given in Lemma 2.6), we obtain

$$\gamma |(\Lambda_D, \varepsilon^{-1}) w_{\mathrm{out}}|_{0,0}^2 + |(\Lambda_D, \varepsilon^{-1}) w_{\mathrm{out}}|_{\infty,0}^2 \leq \frac{C}{\gamma}\big(|(\Lambda_D, \varepsilon^{-1}) f|_{0,0}^2 + |\dot{U}|_{0,0}^2 + |(\varepsilon\Lambda_D)^{-1}\dot{U}|_{0,0}^2\big). \tag{2-18}$$

This completes the estimate of the outgoing terms in the first line of (2-6).

Step 3: *Incoming modes* I. Estimating w_j for $j \leq J'$ in a similar way, but now using $\mathrm{Re}(i\omega_j) \leq -c\gamma$ and pairing the corresponding transport equation with w_j, we obtain

$$\gamma^3 |w_{\mathrm{in}}|_{0,0}^2 + \gamma^2 |w_{\mathrm{in}}|_{\infty,0}^2 \leq C\gamma^2 \langle w_{\mathrm{in}}|_{x_d=0}\rangle_0^2 + \frac{C}{\gamma}(|\gamma f|_{0,0}^2 + |\dot{U}|_{0,0}^2). \tag{2-19}$$

This L^2 estimate does not cause any problem because we have a good L^2 control of the remainder $R_D^a v$ appearing on the right of (2-14); see Lemma 2.6 (we have even weakened the estimate of the remainders in Lemma 2.6 by simply estimating them in terms of $|\dot{U}|_{0,0}/\gamma$). Moreover the term $|w_{\mathrm{in}}|_{\infty,0}^2$ was estimated by considering the L^2 pairing on $[0, x_d] \times b\Omega$ instead of Ω.

Step 4: *Boundary estimate.* We observe that v can be expressed in terms of w as

$$v = (Q_0^{-1})_D w + r_{-1,D}\dot{U}.$$

Recalling the boundary condition in (2-7) and using the decomposition $Q_0^{-1}(\zeta) = [Q_{\mathrm{in}}(\zeta)\, Q_{\mathrm{out}}(\zeta)]$, we accordingly let $w = (w_{\mathrm{in}}, w_{\mathrm{out}})$ and rewrite the boundary condition in (2-7) as

$$\mathcal{B}(\varepsilon U) Q_{\mathrm{in},D} w_{\mathrm{in}}|_{x_d=0} = -\mathcal{B}(\varepsilon U) Q_{\mathrm{out},D} w_{\mathrm{out}}|_{x_d=0} + \chi_D g + r_{-1,D}\dot{U}|_{x_d=0}. \tag{2-20}$$

By (1-12) we have on $\mathcal{V}_i$

$$\mathcal{B}(\varepsilon U)Q_{\mathrm{in}} = P_1^{-1}(P_1\mathcal{B}(\varepsilon U)Q_{\mathrm{in}}P_2)P_2^{-1} = P_1^{-1}\begin{pmatrix}\Lambda^{-1}(\gamma+i\sigma) & \\ & I\end{pmatrix}P_2^{-1},$$

so using the rules of singular calculus, we get

$$\Lambda_D\mathcal{B}(\varepsilon U)Q_{\mathrm{in},D}w_{\mathrm{in}}|_{x_d=0} = (P_1^{-1})_D\begin{pmatrix}\gamma+i\sigma_D & \\ & \Lambda_D I\end{pmatrix}(P_2^{-1})_D w_{\mathrm{in}}|_{x_d=0} + r_{0,D}w_{\mathrm{in}}|_{x_d=0}.$$

With (2-20), this implies

$$\left\langle (P_1^{-1})_D\begin{pmatrix}\gamma+i\sigma_D & \\ & \Lambda_D I\end{pmatrix}(P_2^{-1})_D w_{\mathrm{in}}|_{x_d=0}\right\rangle_0 \le C(\langle \Lambda_D w_{\mathrm{out}}|_{x_d=0}\rangle_0 + \langle \Lambda_D g\rangle_0 + \langle \dot{U}|_{x_d=0}\rangle_0). \quad (2\text{-}21)$$

We have $P_{1,D}(P_1^{-1})_D = I + r_{-1,D}$ so up to choosing γ large (and absorbing the $r_{-1,D}$ term), the estimate (2-21) implies

$$\left\langle \begin{pmatrix}\gamma+i\sigma_D & \\ & \Lambda_D I\end{pmatrix}(P_2^{-1})_D w_{\mathrm{in}}|_{x_d=0}\right\rangle_0 \le C(\langle \Lambda_D w_{\mathrm{out}}|_{x_d=0}\rangle_0 + \langle \Lambda_D g\rangle_0 + \langle \dot{U}|_{x_d=0}\rangle_0). \quad (2\text{-}22)$$

Letting

$$\begin{pmatrix} w_1 \\ w'\end{pmatrix} := (P_2^{-1})_D w_{\mathrm{in}}|_{x_d=0},$$

we find, using the fact that σ is real and again choosing γ large enough,

$$\left\langle \begin{pmatrix}\gamma+i\sigma_D & \\ & \Lambda_D I\end{pmatrix}\begin{pmatrix} w_1 \\ w'\end{pmatrix}\right\rangle_0^2 \ge \frac{1}{C}(\gamma^2\langle w_1\rangle_0^2 + \langle \Lambda_D w'\rangle_0^2) \ge \frac{\gamma^2}{C}\langle w_1, w'\rangle_0^2.$$

Thus, from (2-22), we may conclude

$$\gamma\langle w_{\mathrm{in}}|_{x_d=0}\rangle_0 \le C(\langle \Lambda_D w_{\mathrm{out}}|_{x_d=0}\rangle_0 + \langle \Lambda_D g\rangle_0 + \langle \dot{U}|_{x_d=0}\rangle_0). \quad (2\text{-}23)$$

Combining the estimates (2-19) and (2-23), we have thus derived the bound

$$\gamma^3|w_{\mathrm{in}}|_{0,0}^2 + \gamma^2|w_{\mathrm{in}}|_{\infty,0}^2 \le \frac{C}{\gamma}(|\gamma f|_{0,0}^2 + |\dot{U}|_{0,0}^2) + C(\langle \Lambda_D g\rangle_0^2 + \langle \dot{U}|_{x_d=0}\rangle_0^2) + C\langle \Lambda_D w_{\mathrm{out}}|_{x_d=0}\rangle_0^2.$$

Together with (2-18) this completes the estimate of the terms in the first line of (2-6).

Remark 2.7. At this point we can see the need to estimate the remaining terms on the left in the estimate (2-6) as well as the similar terms on the left in the Kreiss estimate (2-29). We must estimate those terms in order to be able to absorb the terms involving $(\varepsilon\Lambda_D)^{-1}\dot{U}^\gamma$ on the right side of (2-6). Recall that such terms come from the bad $(1,-1)$ product and from the $\partial_d Q_{-1}$ term. This is one of the major differences between our analysis and that in [Coulombel 2004].

Step 5: *Incoming modes* II. Here we begin to estimate the terms in the second line of (2-6). We introduce the functions $\tilde v := \Lambda_D^{-1} v$ and $\tilde v' := \tilde v/\varepsilon$, and see that the function $\tilde v'$ satisfies

$$\begin{aligned} \partial_d \tilde v' + \mathbb{A}_D \tilde v' + \mathcal{D}(\varepsilon U)\tilde v' &= \frac{\Lambda_D^{-1}}{\varepsilon}\boldsymbol{\chi}_D f + \frac{\Lambda_D^{-1}}{\varepsilon}[\mathcal{D}(\varepsilon U), \boldsymbol{\chi}_D]\dot U + \big(\mathcal{D}(\varepsilon U) - \Lambda_D^{-1}\mathcal{D}(\varepsilon U)\Lambda_D\big)\tilde v', \\ \mathcal{B}(\varepsilon U)\tilde v'|_{x_d=0} &= \frac{\Lambda_D^{-1}}{\varepsilon}\boldsymbol{\chi}_D g + \frac{\Lambda_D^{-1}}{\varepsilon}[\mathcal{B}(\varepsilon U), \boldsymbol{\chi}_D]\dot U|_{x_d=0} + [\mathcal{B}(\varepsilon U), \Lambda_D^{-1}]\frac{v}{\varepsilon}|_{x_d=0}. \end{aligned} \tag{2-24}$$

We can thus diagonalize the problem for $\tilde v'$ with the *same* operator $Q_D = Q_{0,D} + Q_{-1,D}$ as before. Introducing the function $\tilde w' := Q_D \tilde v'$, we find that $\tilde w'$ satisfies

$$\begin{aligned} &\text{(a)} \quad \partial_d \tilde w' = -(\mathbb{D}_1 + \mathbb{D}_0)_D \tilde w' + Q_D \frac{\Lambda_D^{-1}}{\varepsilon}\boldsymbol{\chi}_D f + Q_D \frac{\Lambda_D^{-1}}{\varepsilon}[\mathcal{D}(\varepsilon U), \boldsymbol{\chi}_D]\dot U + \frac{1}{\gamma} r_{0,D}\tilde v', \\ &\text{(b)} \quad \mathcal{B}(\varepsilon U) Q_{\mathrm{in},D}\tilde w'_{\mathrm{in}} = -\mathcal{B}(\varepsilon U) Q_{\mathrm{out},D}\tilde w'_{\mathrm{out}} + \frac{\Lambda_D^{-1}}{\varepsilon}\boldsymbol{\chi}_D g + \frac{\Lambda_D^{-1}}{\varepsilon}[\mathcal{B}(\varepsilon U), \boldsymbol{\chi}_D]\dot U \\ &\qquad\qquad + \Lambda_D^{-1}[\Lambda_D, \mathcal{B}(\varepsilon U)]\tilde v' + r_{-1,D}\tilde v', \end{aligned} \tag{2-25}$$

where we have collected several terms into remainders of the form $\gamma^{-1} r_{0,D}\tilde v'$. For instance, we have used

$$R_D^a \tilde v' = \frac{1}{\gamma} r_{0,D}\tilde v', \quad Q_{-1,D}\mathcal{D}(\varepsilon U)\tilde v' = \frac{1}{\gamma} r_{0,D}\tilde v'.$$

Next we fix an index $j \in \{1, \dots, J'\}$. Taking the real part of the $L^2(\Omega)$ inner product of (2-25)(a) with $\tilde w'_j$, we obtain the standard L^2 estimate for incoming modes:

$$\begin{aligned} &\gamma |\tilde w'_{\mathrm{in}}|_{0,0}^2 \\ &\quad \le C\langle \tilde w'_{\mathrm{in}}|_{x_d=0}\rangle_0^2 + \frac{C}{\gamma}\Big(|(\varepsilon\Lambda_D)^{-1} f|_{0,0}^2 + \frac{1}{\gamma^2}|(\varepsilon\Lambda_D)^{-1}\dot U|_{0,0}^2 + |(\varepsilon\Lambda_D)^{-1}[\mathcal{D}(\varepsilon U), \boldsymbol{\chi}_D]\dot U|_{0,0}^2 \Big) \\ &\quad \le C\langle \tilde w'_{\mathrm{in}}|_{x_d=0}\rangle_0^2 + \frac{C}{\gamma^3}(|\varepsilon^{-1} f|_{0,0}^2 + |(\varepsilon\Lambda_D)^{-1}\dot U|_{0,0}^2). \end{aligned} \tag{2-26}$$

We thus wish to control the trace of $\tilde w'_{\mathrm{in}}$.

Step 6: *Control of the trace of* $\tilde w'_{\mathrm{in}}$. Using (2-25)(b) and arguing as in Step 4, we obtain the boundary estimate

$$\begin{aligned} &\gamma \langle \tilde w'_{\mathrm{in}}|_{x_d=0}\rangle_0 \\ &\quad \le C(\langle \Lambda_D \tilde w'_{\mathrm{out}}|_{x_d=0}\rangle_0 + \langle \varepsilon^{-1} g\rangle_0 + \langle \varepsilon^{-1}[\mathcal{B}(\varepsilon U), \boldsymbol{\chi}_D]\dot U|_{x_d=0}\rangle_0 + \langle \tilde v'|_{x_d=0}\rangle_0 + \langle \varepsilon^{-1}[\Lambda_D, \mathcal{B}(\varepsilon U)]\tilde v|_{x_d=0}\rangle_0) \\ &\quad \le C(\langle \Lambda_D \tilde w'_{\mathrm{out}}|_{x_d=0}\rangle_0 + \langle \varepsilon^{-1} g\rangle_0 + \langle (\varepsilon\Lambda_D)^{-1}\dot U|_{x_d=0}\rangle_0). \end{aligned}$$

Combining with (2-26), we have derived

$$\begin{aligned} &\gamma |\tilde w'_{\mathrm{in}}|_{0,0}^2 + \langle \tilde w'_{\mathrm{in}}|_{x_d=0}\rangle_0^2 \\ &\le \frac{C}{\gamma^2}\langle \Lambda_D \tilde w'_{\mathrm{out}}|_{x_d=0}\rangle_0^2 + \frac{C}{\gamma^3}(|\varepsilon^{-1} f|_{0,0}^2 + |(\varepsilon\Lambda_D)^{-1}\dot U|_{0,0}^2) + \frac{C}{\gamma^2}(\langle \varepsilon^{-1} g\rangle_0^2 + \langle (\varepsilon\Lambda_D)^{-1}\dot U|_{x_d=0}\rangle_0^2). \end{aligned} \tag{2-27}$$

We expect $\Lambda_D \tilde{w}'_{\mathrm{out}}$ to be comparable to $w_{\mathrm{out}}/\varepsilon$ and thus use (2-18); this is checked and made precise in the next and last step of the proof of Proposition 2.4.

Step 7: *Relation between* $\Lambda_D\tilde{w}'$ *and* w/ε*, and conclusion.* Using the definitions

$$\tilde{w}' = Q_D\tilde{v}' = Q_D(\varepsilon\Lambda_D)^{-1}v \quad \text{and} \quad w = Q_D v,$$

and the fact that Λ_D commutes with $Q_{0,D}$, we compute

$$\Lambda_D\tilde{w}' = \varepsilon^{-1}Q_D v + r_{0,D}\tilde{v}' = \varepsilon^{-1}w + r_{0,D}\tilde{v}'.$$

We have thus derived the bound from above

$$\frac{1}{\gamma^2}\langle \Lambda_D\tilde{w}'_{\mathrm{out}}|_{x_d=0}\rangle_0^2 \le \frac{C}{\gamma^2}(\langle \varepsilon^{-1}w_{\mathrm{out}}|_{x_d=0}\rangle_0^2 + \langle (\varepsilon\Lambda_D)^{-1}\dot{U}|_{x_d=0}\rangle_0^2),$$

which we combine with (2-27) and (2-18) to obtain

$$\begin{aligned}\gamma|\tilde{w}'_{\mathrm{in}}|_{0,0}^2 + \langle \tilde{w}'_{\mathrm{in}}|_{x_d=0}\rangle_0^2 \\ \le \frac{C}{\gamma^3}\big(|(\Lambda_D,\varepsilon^{-1})f|_{0,0}^2 + |(\varepsilon\Lambda_D)^{-1}\dot{U}|_{0,0}^2 + |\dot{U}|_{0,0}^2\big) + \frac{C}{\gamma^2}\big(\langle \varepsilon^{-1}g\rangle_0^2 + \langle (\varepsilon\Lambda_D)^{-1}\dot{U}|_{x_d=0}\rangle_0^2\big).\end{aligned} \tag{2-28}$$

It only remains to derive a bound from below to go from $\tilde{w}'_{\mathrm{in}}$ to $(\varepsilon\Lambda_D)^{-1}\dot{U}^\gamma_{1,\mathrm{in}}$. We first observe that estimating $(\varepsilon\Lambda_D)^{-1}\dot{U}^\gamma_{1,\mathrm{in}}$ as claimed in (2-6) amounts to estimating $Q_D(\varepsilon\Lambda_D)^{-1}\dot{U}^\gamma_{1,\mathrm{in}}$. We use the relation

$$Q_D(\varepsilon\Lambda_D)^{-1}\dot{U}^\gamma_{1,\mathrm{in}} = (\varepsilon\Lambda_D)^{-1}\begin{pmatrix} w_{\mathrm{in}} \\ 0\end{pmatrix} - [(\varepsilon\Lambda_D)^{-1}, Q_D]\dot{U}^\gamma_{1,\mathrm{in}} = (\varepsilon\Lambda_D)^{-1}\begin{pmatrix} w_{\mathrm{in}} \\ 0\end{pmatrix} - [(\varepsilon\Lambda_D)^{-1}, Q_{-1,D}]\dot{U}^\gamma_{1,\mathrm{in}},$$

and the special "decoupled" form of the coefficients of Q_{-1} to show that

$$[(\varepsilon\Lambda_D)^{-1}, Q_{-1,D}] = \frac{1}{\gamma^2}r_{0,D}(\varepsilon\Lambda_D)^{-1}.$$

Similarly, taking the "in" component of $\tilde{w}' = Q_D(\varepsilon\Lambda_D)^{-1}\dot{U}^\gamma_1$, we have

$$\tilde{w}'_{\mathrm{in}} = (\varepsilon\Lambda_D)^{-1}w_{\mathrm{in}} + \frac{1}{\gamma^2}r_{0,D}(\varepsilon\Lambda_D)^{-1}\dot{U}^\gamma_1,$$

so we obtain

$$Q_D(\varepsilon\Lambda_D)^{-1}\dot{U}^\gamma_{1,\mathrm{in}} = \begin{pmatrix}\tilde{w}'_{\mathrm{in}} \\ 0\end{pmatrix} + \frac{1}{\gamma^2}r_{0,D}(\varepsilon\Lambda_D)^{-1}\dot{U}.$$

We have therefore proved that (2-28) implies that the second line in (2-6) is controlled by the terms on the right of (2-6). This finishes the proof of Proposition 2.4. □

(III): *Estimate away from the bad set.* The next proposition provides a Kreiss-type estimate for the terms $\chi_{i,D}\dot{U}^\gamma$, where $i > N_1$.

Proposition 2.8. *Fix i such that $N_1+1 \le i \le N_2$ and let $\dot U_2^\gamma := \chi_{i,D}\dot U^\gamma$. We have*

$$|\dot U_2^\gamma|_{0,0} + \frac{\langle \dot U_2^\gamma|_{x_d=0}\rangle_0}{\sqrt\gamma} + |(\varepsilon\Lambda_D)^{-1}\dot U_2^\gamma|_{0,0} + \frac{\langle (\varepsilon\Lambda_D)^{-1}\dot U_2^\gamma|_{x_d=0}\rangle_0}{\sqrt\gamma}$$
$$\le C\bigg(\frac{|f^\gamma|_{0,0}+|(\varepsilon\Lambda_D)^{-1}f^\gamma|_{0,0}}{\gamma} + \frac{\langle g^\gamma\rangle_0 + \langle(\varepsilon\Lambda_D)^{-1}g^\gamma\rangle_0}{\sqrt\gamma}$$
$$+ \frac{|\dot U^\gamma|_{0,0}+|(\varepsilon\Lambda_D)^{-1}\dot U^\gamma|_{0,0}}{\gamma^2} + \frac{\langle \dot U^\gamma|_{x_d=0}\rangle_0 + \langle(\varepsilon\Lambda_D)^{-1}\dot U^\gamma|_{x_d=0}\rangle_0}{\gamma}\bigg). \quad (2\text{-}29)$$

Proof. Step 1: *L^2 estimate.* The first step is to prove the Kreiss-type estimate

$$|\dot U_2^\gamma|_{0,0} + \frac{\langle \dot U_2^\gamma|_{x_d=0}\rangle_0}{\sqrt\gamma} \le C\left(\frac{|f^\gamma|_{0,0}}{\gamma} + \frac{\langle g^\gamma\rangle_0}{\sqrt\gamma} + \frac{|\dot U^\gamma|_{0,0}}{\gamma^2} + \frac{\langle \dot U^\gamma|_{x_d=0}\rangle_0}{\gamma}\right). \quad (2\text{-}30)$$

For this we define the good set $G \subset \Sigma$ to be a neighborhood of the closure of $\bigcup_{i=N_1+1}^{N_2}\mathcal{V}_i$ such that G is disjoint from $\overline{\Upsilon}$; here the uniform Lopatinskii condition is satisfied. The classical construction of Kreiss symmetrizers [Kreiss 1970; Chazarain and Piriou 1982] provides us with an $N\times N$ symbol $R(\zeta)$, homogeneous of degree 0, such that, for some positive constants C, c, and $\zeta/|\zeta| \in G$, we have

$$\begin{aligned} &\text{(a)} \quad R(\zeta) = R(\zeta)^*,\\ &\text{(b)} \quad -\mathrm{Re}(R(\zeta)\mathbb{A}(\zeta)) \ge c\gamma I_N,\\ &\text{(c)} \quad R(\zeta) + C\mathcal{B}(0)^*\mathcal{B}(0) \ge cI_N. \end{aligned} \quad (2\text{-}31)$$

We take a smooth extension of R to all ζ as a symbol of order 0 such that (2-31)(a) holds. Observe that by continuity (2-31)(c) implies

$$R(\zeta) + C\mathcal{B}(\varepsilon U)^*\mathcal{B}(\varepsilon U) \ge cI_N \quad \text{for } \varepsilon \text{ small enough.} \quad (2\text{-}32)$$

As observed in [Williams 2002], we may now use R_D, the singular Fourier multiplier associated to the symbol $R(X,\gamma)$ as a Kreiss symmetrizer for the singular problem. Let $\chi_i = \boldsymbol{\chi}$, $v := \boldsymbol{\chi}_D\dot U^\gamma$, and denote by $\langle\cdot,\cdot\rangle$ the L^2 inner product on $b\Omega$. Using (2-7) to expand $\partial_d\langle v, R_D v\rangle$ and integrating in x_d over $[0,\infty)$, we obtain

$$-\langle v|_{x_d=0}, R_D v|_{x_d=0}\rangle$$
$$= -2\,\mathrm{Re}(R_D\mathbb{A}_D v, v) - 2\,\mathrm{Re}(R_D\mathcal{D}(\varepsilon U)v, v) + 2\,\mathrm{Re}(R_D\boldsymbol{\chi}_D f^\gamma, v) + O(|\dot U^\gamma|_{0,0}|v|_{0,0}/\gamma).$$

From (2-31)(b), (2-32), and the localized Gårding inequality (Proposition A.9),

$$\mathrm{Re}\langle (R + C\mathcal{B}(\varepsilon U)^*\mathcal{B}(\varepsilon U))_D v|_{x_d=0}, v|_{x_d=0}\rangle \ge c\langle v|_{x_d=0}\rangle_0^2 - C\frac{\langle \dot U^\gamma|_{x_d=0}\rangle_0^2}{\gamma}, \quad (2\text{-}33)$$

we easily derive the estimate (2-30).

Step 2: *Estimate of $(\varepsilon\Lambda_D)^{-1}\dot U_2^\gamma$.* Set $\tilde v := \Lambda_D^{-1}v$ and $\tilde v' = \tilde v/\varepsilon$. Then $\tilde v'$ satisfies the system (2-24), where the truncation function $\boldsymbol{\chi}$ has changed but the forcing terms have exactly the same expression. An

argument just like the one that gave the estimate (2-30) yields

$$|(\varepsilon\Lambda_D)^{-1}\dot U_2^\gamma|_{0,0}+\frac{\langle(\varepsilon\Lambda_D)^{-1}\dot U_2^\gamma|_{x_d=0}\rangle_0}{\sqrt\gamma}$$
$$\le C\Big(\frac{|(\varepsilon\Lambda_D)^{-1}f^\gamma|_{0,0}}{\gamma}+\frac{\langle(\varepsilon\Lambda_D)^{-1}g^\gamma\rangle_0}{\sqrt\gamma}+\frac{|\dot U^\gamma|_{0,0}+|(\varepsilon\Lambda_D)^{-1}\dot U^\gamma|_{0,0}}{\gamma^2}$$
$$+\frac{\langle\dot U^\gamma|_{x_d=0}\rangle_0+\langle(\varepsilon\Lambda_D)^{-1}\dot U^\gamma|_{x_d=0}\rangle_0}{\gamma}\Big).$$

Here instead of (2-33) we have used

$$\mathrm{Re}\langle(R+C\mathcal{B}(\varepsilon U)^*\mathcal{B}(\varepsilon U))_D\tilde v'|_{x_d=0},\tilde v'|_{x_d=0}\rangle\ge c\langle\tilde v'|_{x_d=0}\rangle_0^2-C\frac{\langle(\varepsilon\Lambda_D)^{-1}\dot U^\gamma|_{x_d=0}\rangle_0^2}{\gamma},$$

to recover the estimate of the trace of $\tilde v'$. The L^2 estimates of the forcing terms in the interior and on the boundary are exactly the same as in steps 4 and 5 of the previous proof. □

(IV): *Conclusion.* We use the previous propositions to complete the proof of Proposition 2.2. Summing the estimates (2-6) and (2-29) over $i\in\{1,\dots,N_2\}$ and absorbing error terms from the right by taking γ large, we derive

$$|\dot U^\gamma|_{0,0}+\frac{\langle\dot U^\gamma|_{x_d=0}\rangle_0}{\sqrt\gamma}+\frac{|\dot U_1^\gamma|_{\infty,0}}{\sqrt\gamma}\le C(K)\Big(\frac{|(\Lambda_D,\varepsilon^{-1})f^\gamma|_{0,0}}{\gamma^2}+\frac{\langle(\Lambda_D,\varepsilon^{-1})g^\gamma\rangle_0}{\gamma^{3/2}}\Big),\tag{2-34}$$

where we have "forgotten" on the left of the inequality the additional control of $(\varepsilon\Lambda_D)^{-1}\dot U^\gamma$ (this term has played its role, meaning that it was used to absorb some bad terms appearing on the right). This gives exactly (2-4) with the additional control of $\dot U_1^\gamma$ in $L^\infty(L^2)$. This additional property is used in the proof of Corollary 2.3. □

Proof of Corollary 2.3. It remains to estimate $|\dot U^\gamma|_{0,1}$ and $|\dot U^\gamma|_{\infty,0}$. We first estimate the first-order tangential derivatives. We can apply the a priori estimate (2-4) to the problem satisfied by $\partial_{(x',\theta_0)}\dot U^\gamma$, which is obtained by differentiating (2-2). This yields

$$|\dot U^\gamma|_{0,1}+\frac{\langle\dot U^\gamma|_{x_d=0}\rangle_1}{\sqrt\gamma}\le C(K)\Big(\frac{|(\Lambda_D,\varepsilon^{-1})f^\gamma|_{0,1}}{\gamma^2}+\frac{\langle(\Lambda_D,\varepsilon^{-1})g^\gamma\rangle_1}{\gamma^{3/2}}\Big),\tag{2-35}$$

which is the same as (2-5), except for the absence of $|\dot U^\gamma|_{\infty,0}$ on the left. Here we were able to treat commutators as forcing terms because, for example,

$$[\mathcal{D}(\varepsilon U),\partial_{(x',\theta_0)}]\dot U^\gamma=-(d\mathcal{D}(\varepsilon U)\cdot\varepsilon\partial_{(x',\theta_0)}U)\dot U^\gamma,$$

and the factor of ε coming out from the commutation allows us, for example, to estimate

$$|\Lambda_D[\mathcal{D}(\varepsilon U),\partial_{(x',\theta_0)}]\dot U^\gamma|_{0,0}\le C|\dot U^\gamma|_{0,1}.$$

It thus only remains to estimate the norm $|\dot{U}^\gamma|_{\infty,0}$. For $\delta_2 > 0$ to be chosen, we take $0 < \delta_1 < \delta_2$ and consider a symbol of order zero in the extended calculus, $\chi^e(\xi', k\beta/\varepsilon, \gamma)$, such that

$$0 \le \chi^e \le 1, \qquad \chi^e\left(\xi', \frac{k\beta}{\varepsilon}, \gamma\right) = 1 \quad \text{on } \left\{|\xi', \gamma| \le \delta_1 \frac{|k\beta|}{\varepsilon}\right\}, \qquad \operatorname{supp}\chi^e \subset \left\{|\xi', \gamma| \le \delta_2 \frac{|k\beta|}{\varepsilon}\right\}.$$

We then write $\dot{U}^\gamma = \chi_D^e \dot{U}^\gamma + (1 - \chi_D^e)\dot{U}^\gamma$ and begin by estimating $|(1 - \chi_D^e)\dot{U}^\gamma|_{0,\infty}$ by using the Sobolev-type estimate

$$|(1-\chi_D^e)\dot{U}^\gamma|_{\infty,0} \le C|(1-\chi_D^e)\partial_d \dot{U}^\gamma|_{0,0} + C|(1-\chi_D^e)\dot{U}^\gamma|_{0,0} \le C|(1-\chi_D^e)\partial_d \dot{U}^\gamma|_{0,0} + C|\dot{U}^\gamma|_{0,0}. \quad (2\text{-}36)$$

Using (2-2) and the fact that

$$|X, \gamma|\left(1 - \chi^e\left(\xi', \frac{k\beta}{\varepsilon}, \gamma\right)\right) \le C|\xi', \gamma|,$$

we obtain

$$\begin{aligned} |(1 - \chi_D^e)\partial_d \dot{U}^\gamma|_{0,0} &\le |\mathbb{A}_D(1 - \chi_D^e)\dot{U}^\gamma|_{0,0} + |(1 - \chi_D^e)\mathbb{D}\dot{U}^\gamma|_{0,0} + |(1 - \chi_D^e) f^\gamma|_{0,0} \\ &\le C(|\dot{U}^\gamma|_{0,1} + |f^\gamma|_{0,0}) \le C\left(|\dot{U}^\gamma|_{0,1} + \frac{|\Lambda_D f^\gamma|_{0,1}}{\gamma^2}\right), \end{aligned}$$

where the last inequality follows from $|f^\gamma|_{0,0} \le C|f^\gamma|_{0,1}/\gamma$. With (2-36) this gives

$$|(1 - \chi_D^e)\dot{U}^\gamma|_{\infty,0} \le C\left(|\dot{U}^\gamma|_{0,1} + \frac{|\Lambda_D(f^\gamma)|_{0,1}}{\gamma^2}\right). \quad (2\text{-}37)$$

To estimate $|\chi_D^e \dot{U}^\gamma|_{\infty,0}$ we observe that since $\beta \in \Upsilon$, we have, for $\delta_2 > 0$ chosen small enough,

$$\chi^e\left(\xi', \frac{k\beta}{\varepsilon}, \gamma\right) = \chi^e\left(\xi', \frac{k\beta}{\varepsilon}, \gamma\right) \sum_{i=1}^{N_1} \chi_i(X, \gamma),$$

for the χ_i chosen in Step I of the proof of Proposition 2.2. Thus

$$|\chi_D^e \dot{U}^\gamma|_{\infty,0} \le |\chi_D^e \dot{U}_1^\gamma|_{\infty,0} \le |\dot{U}_1^\gamma|_{\infty,0},$$

with $\dot{U}_1^\gamma$ defined in Proposition 2.4.[22] We can then apply the a priori estimate (2-34) and obtain

$$|\chi_D^e \dot{U}^\gamma|_{\infty,0} \le C\left(\frac{|(\Lambda_D, \varepsilon^{-1}) f^\gamma|_{0,0}}{\gamma^{3/2}} + \frac{\langle(\Lambda_D, \varepsilon^{-1}) g^\gamma\rangle_0}{\gamma}\right).$$

With (2-37) and (2-35), this completes the proof of Corollary 2.3. □

Let us quickly observe that the genuine Gårding inequality was used only once, in the proof of Proposition 2.2, namely in (2-33). In all other cases, we only used Plancherel's theorem for Fourier multipliers. This explains the slight difference between (2-29) and (2-6) for the powers of γ.

[22] More precisely, here $\dot{U}_1^\gamma$ is the sum of the similarly denoted functions in Proposition 2.4.

Next we "localize the estimate" to Ω_T. Since[23]

$$|\Lambda_D f^\gamma|_{0,1} \sim \left|\left(\gamma, \partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right) f^\gamma\right|_{0,1} \sim \left|\left(\gamma, \partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right) f\right|_{0,1,\gamma},$$

we can rewrite the a priori estimate (2-5) for solutions to the linearized system (2-1) as

$$|\dot U|_{\infty,0,\gamma} + |\dot U|_{0,1,\gamma} + \frac{\langle \dot U|_{x_d=0}\rangle_{1,\gamma}}{\sqrt\gamma} \le C(K)\left(\frac{|(\gamma, \partial_{x'} + \beta\partial_{\theta_0}/\varepsilon) f|_{0,1,\gamma} + |f/\varepsilon|_{0,1,\gamma}}{\gamma^2} + \frac{\langle(\gamma, \partial_{x'} + \beta\partial_{\theta_0}/\varepsilon) g\rangle_{1,\gamma} + \langle g/\varepsilon\rangle_{1,\gamma}}{\gamma^{3/2}}\right). \quad (2\text{-}38)$$

Suppose now that the singular problem (2-1) is posed on Ω_T instead of Ω. Given $f \in L^2 H^1_T$, one can define a Seeley extension $\tilde f \in L^2 H^1$ such that

$$\left|\left(\gamma, \partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right)\tilde f\right|_{0,1} + |\tilde f/\varepsilon|_{0,1} \le C\left(\left|\left(\gamma, \partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right) f\right|_{0,1,T} + |f/\varepsilon|_{0,1,T}\right),$$

where C is independent of γ, ε, and T. It is readily checked that the same extension satisfies

$$\left|\left(\gamma, \partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right)\tilde f\right|_{0,1,\gamma} + |\tilde f/\varepsilon|_{0,1,\gamma} \le C\left(\left|\left(\gamma, \partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right) f\right|_{0,1,\gamma,T} + |f/\varepsilon|_{0,1,\gamma,T}\right), \quad (2\text{-}39)$$

where again C is independent of γ, ε, and T. We claim that changing f, g, and U in $\{t > T\}$ does not affect the solution of (2-1) in $\{t < T\}$. (This causality principle is discussed further below together with the existence of solutions to the linearized system (2-1).) Hence the estimates (2-38) and (2-39) imply the following estimate for the singular problem on Ω_T:

$$|\dot U|_{\infty,0,\gamma,T} + |\dot U|_{0,1,\gamma,T} + \frac{\langle \dot U|_{x_d=0}\rangle_{1,\gamma,T}}{\sqrt\gamma} \le C(K)\left(\frac{|(\gamma, \partial_{x'} + \beta\partial_{\theta_0}/\varepsilon) f|_{0,1,\gamma,T} + |f/\varepsilon|_{0,1,\gamma,T}}{\gamma^2} + \frac{\langle(\gamma, \partial_{x'} + \beta\partial_{\theta_0}/\varepsilon) g\rangle_{1,\gamma,T} + \langle g/\varepsilon\rangle_{1,\gamma,T}}{\gamma^{3/2}}\right).$$

Let us now consider the linearized singular problem (2-1) on Ω_T with data of the form εf and εg instead of f and g. We note that

$$\left|\left(\gamma, \partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right)\varepsilon f\right|_{0,1,\gamma,T} \le C|f|_{0,2,\gamma,T} \quad \text{and} \quad \left\langle\left(\gamma, \partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right)\varepsilon g\right\rangle_{1,\gamma,T} \le C\langle g\rangle_{2,\gamma,T}.$$

Let us write the linearized operators on the left sides of (2-1)(a) and (b) as $\mathbb{L}'(\varepsilon U)\dot U$ and $\mathbb{B}'(\varepsilon U)\dot U$, respectively, and define

$$\mathscr{L}'_\varepsilon(U)\dot U := \frac{1}{\varepsilon}\mathbb{L}'(\varepsilon U)\dot U, \quad \mathscr{B}'_\varepsilon(U)\dot U := \frac{1}{\varepsilon}\mathbb{B}'(\varepsilon U)\dot U.$$

We have proved:

[23] Here "$\sim$" denotes equivalence of norms with constants independent of ε and γ.

Proposition 2.9. *Fix $K > 0$ and suppose $|\varepsilon\partial_d U|_{C_T^{0,M_0-1}} + |U|_{C_T^{0,M_0}} \le K$ for $\varepsilon \in (0,1]$. There exist positive constants $\varepsilon_0(K)$, $\gamma_0(K)$ such that solutions of the singular problem*

$$\begin{aligned} &\mathscr{L}'_\varepsilon(U)\dot U = f \quad \text{on } \Omega_T,\\ &\mathscr{B}'_\varepsilon(U)\dot U = g \quad \text{on } b\Omega_T,\\ &\dot U = 0 \qquad\quad \text{in } t<0 \end{aligned} \tag{2-40}$$

satisfy

$$|\dot U|_{\infty,0,\gamma,T} + |\dot U|_{0,1,\gamma,T} + \frac{\langle \dot U|_{x_d=0}\rangle_{1,\gamma,T}}{\sqrt{\gamma}} \le C(K)\left(\frac{|f|_{0,2,\gamma,T}}{\gamma^2} + \frac{\langle g\rangle_{2,\gamma,T}}{\gamma^{3/2}}\right) \tag{2-41}$$

for $0<\varepsilon\le\varepsilon_0(K)$, $\gamma\ge\gamma_0(K)$, and the constant $C(K)$ only depends on K.

The same estimate holds if $\mathscr{B}(\varepsilon U)$ in (2-1) is replaced by $\mathscr{B}(\varepsilon U, \varepsilon^{\mathcal{Q}}\mathcal{U})$ given in (1-9), and $\mathscr{D}(\varepsilon U)$ is replaced by $\mathscr{D}(\varepsilon U, \varepsilon^{\mathcal{Q}}\mathcal{U})$ given in (1-10), as long as $|\varepsilon\partial_d(U,{}^{\mathcal{Q}}\mathcal{U})|_{C_T^{0,M_0-1}} + |U,{}^{\mathcal{Q}}\mathcal{U}|_{C_T^{0,M_0}} \le K$ for $\varepsilon\in(0,1]$.

2B. ***Well-posedness of the linearized singular equations.*** In this short section, we explain why the analysis in [Coulombel 2005] gives existence and uniqueness of a solution to the linearized singular problem (2-40) for which the estimate (2-41) holds. First we define a dual problem for (2-1):

$$\begin{aligned} &\partial_d\dot U + \mathbb{A}^*\left(\partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right)\dot U + \tilde{\mathscr{D}}(\varepsilon U)\dot U = f(x,\theta_0) \quad \text{on } \Omega,\\ &\mathcal{M}(\varepsilon U)\dot U|_{x_d=0} = g(x',\theta_0), \end{aligned} \tag{2-42}$$

where $\mathbb{A}^*$ is obtained from $\mathbb{A}$ by first multiplying the system by the constant matrix B_d, then integrating by parts on Ω, and eventually multiplying by $(B_d^T)^{-1}$. The zero-order term is also changed accordingly. Following the standard procedure described for instance in [Benzoni-Gavage and Serre 2007, Chapter 4.4], the matrix $\mathcal{M}$ giving the adjoint boundary conditions is chosen such that

$$B_d = \mathscr{B}_1(v)^T\mathscr{B}(v) + \mathcal{M}(v)^T\mathcal{M}_1(v)$$

for all v sufficiently close to the origin, where $\mathscr{B}_1(v)$ and $\mathcal{M}_1(v)$ are additional matrices depending smoothly on v.

The expression of $\mathbb{A}^*$ shows that this singular operator coincides with the operator obtained by applying the substitution $\partial_{x'} \to \partial_{x'} + \beta\partial_{\theta_0}/\varepsilon$ to the dual operator

$$\partial_t + \sum_{j=1}^d B_j^T\partial_j = -L_0(\partial)^*.$$

It is known from the analysis in [Benzoni-Gavage and Serre 2007, Chapter 8.3] that the latter constant multiplicity hyperbolic operator with boundary conditions given by $\mathcal{M}(v)$ gives rise to a boundary value problem in the "backward" WR class (one just has to replace γ by $-\gamma$ for this dual problem). When we apply the singular transformation $\partial_{x'} \to \partial_{x'} + \beta\partial_{\theta_0}/\varepsilon$ to the boundary value problem defined by $(L_0(\partial)^*, \mathcal{M}(\varepsilon U))$, we can reproduce the analysis of the previous section and show that the same type of a priori estimate as in Proposition 2.2 holds for (2-42).

For all fixed $\varepsilon > 0$ small enough, we have thus proved that both the forward problem (2-1) and its dual problem (2-42) satisfy an a priori estimate with a loss of one tangential derivative. The estimates depend very badly on ε because the singular derivative $\partial_{x'} + \beta\partial_{\theta_0}/\varepsilon$ is estimated by $1/\varepsilon$ times the tangential H^1 norm with respect to (x', θ_0). Nevertheless, we can at this stage reproduce the arguments of [Coulombel 2005] to show the existence and uniqueness of L^2 solutions to (2-1) when the source terms f and g satisfy $f, \partial_{\theta_0} f, \partial_{x'} f \in L^2(\Omega_T)$, $g \in H^1(b\Omega_T)$. The analysis is actually much simpler than in [Coulombel 2005] because most of the technical difficulties there arise from commutations with the hyperbolic operator. Here the hyperbolic operator has constant coefficients so commutation with any scalar Fourier multiplier is exact. The analysis in [Coulombel 2005] also shows that weak solutions are limits of strong solutions when the hyperbolic operator has constant coefficients,[24] so we can show that weak solutions satisfy the energy estimate (2-4) with constants that are *uniform* with respect to the small parameter ε. Such global in time estimates imply the causality principle that "the future does not affect the past" and can be localized to Ω_T by the extension procedure previously described.

2C. ***Tame estimates.*** In this section we prove higher derivative estimates for the linearized singular problem (2-1), first in the "pretame" form of Proposition 2.11, and then in the final, "tame" form of Proposition 2.16, which is suitable for Nash–Moser iteration. Propositions 2.12 and 2.15 give pretame and tame estimates for second derivatives.

Notation 2.10. (a) Let $L^\infty W^{1,\infty} \equiv L^\infty(\overline{\mathbb{R}}_+, W^{1,\infty}(b\Omega))$ with norm $|U|_{L^\infty W^{1,\infty}} := |U|^*$. We also write $|U|_{L^\infty(\Omega)} = |U|_*$, $\langle V\rangle_{L^\infty(b\Omega)} = \langle V\rangle_*$, $\langle V\rangle_{W^{1,\infty}(b\Omega)} = \langle V\rangle^*$, $|U|_{L^\infty(\Omega_T)} = |U|_{*,T}$, etc.

(b) For $k \in \mathbb{N}$, let ∂^k denote the collection of tangential operators $\partial^\alpha_{(x',\theta_0)}$ with $|\alpha| = k$ (α is a multi-index). Sometimes ∂^k is used to denote a particular member of this collection. Set $\partial^0\phi = \phi$.

(c) For $k \in \{1, 2, 3, \dots\}$, denote by $\partial^{\langle k\rangle}\phi$ the set of products of the form $(\partial^{\alpha_1}\phi_{i_1})\cdots(\partial^{\alpha_r}\phi_{i_r})$ where $1 \le r \le k$, $\alpha_1 + \cdots \alpha_r = k$, $\alpha_i \ge 1$. Set $\partial^{\langle 0\rangle}\phi = 1$.

(d) For $r \ge 0$, let $[r]$ denote the smallest integer greater than r.

Our first goal is to prove the following "pretame" estimate for solutions to (2-40).

Proposition 2.11. *Fix $K > 0$ and suppose $|\varepsilon\partial_d U|_{C^{0,M_0-1}} + |U|_{C^{0,M_0}} \le K$ for $\varepsilon \in (0, 1]$. For $s \ge 0$ in any fixed finite interval, there exist positive constants $\varepsilon_0(K)$, $\gamma_0(K)$ such that the solution to the linearized singular problem (2-40) satisfies*

$$|\dot U|_{\infty,s,\gamma,T} + |\dot U|_{0,s+1,\gamma,T} + \frac{\langle \dot U|_{x_d=0}\rangle_{s+1,\gamma,T}}{\sqrt{\gamma}}$$
$$\le C(K)\left(\frac{|f|_{0,s+2,\gamma,T}}{\gamma^2} + \frac{\langle g\rangle_{s+2,\gamma,T}}{\gamma^{3/2}} + \frac{|U|_{0,s+2,\gamma,T}|\dot U|_{*,T}}{\gamma^2} + \frac{\langle U|_{x_d=0}\rangle_{s+2,\gamma,T}\langle \dot U|_{x_d=0}\rangle_{*,T}}{\gamma^{3/2}}\right), \quad (2\text{-}43)$$

for $0 < \varepsilon \le \varepsilon_0(K)$ and $\gamma \ge \gamma_0(K)$.

[24] Weak solutions are only "semistrong" solutions when the hyperbolic operator has variable coefficients.

Proof. The problem satisfied by $\partial^s \dot{U}$ is

$$\mathscr{L}'_\varepsilon(U)\partial^s \dot{U} = \partial^s f + \frac{1}{\varepsilon}[\mathscr{D}(\varepsilon U), \partial^s]\dot{U},$$
$$\mathscr{B}'_\varepsilon(U)\partial^s \dot{U} = \partial^s g + \frac{1}{\varepsilon}[\mathscr{B}(\varepsilon U), \partial^s]\dot{U}.$$

In applying the estimate (2-41) to this problem, we must, for example, compute $\partial^2([\mathscr{D}(\varepsilon U), \partial^s]\dot{U})$, which is a sum of terms of the form[25]

$$\tilde{\mathscr{D}}(\varepsilon U)\partial^{\langle j\rangle}(\varepsilon U)\partial^k \dot{U}, \quad \text{where } j+k = s+2,\ j \geq 1,$$

and $\tilde{\mathscr{D}}$ is some smooth function of its argument. Since $j \geq 1$, we can rewrite this as

$$\tilde{\mathscr{D}}(\varepsilon U)\partial^{\langle j-1\rangle}(\varepsilon U)\partial(\varepsilon U)\partial^k \dot{U}.$$

Using Moser estimates, we obtain

$$\left|\frac{1}{\varepsilon}\tilde{\mathscr{D}}(\varepsilon U)\partial^{\langle j-1\rangle}(\varepsilon U)\partial(\varepsilon U)\partial^k \dot{U}\right|_{0,\gamma,T} \leq C(K)|\dot{U}|_{*,T}|U|_{0,s+2,\gamma,T} + C(K)|\dot{U}|_{0,s+1,\gamma,T}.$$

The contribution from the final term on the right can be absorbed by taking γ large enough; thus this explains the third term on the right in (2-43). The final term on the right in (2-43) arises by the same argument applied to the boundary commutator. □

Next we prove estimates for the second derivatives

$$\mathscr{L}''_\varepsilon(U)(\dot{U}^a, \dot{U}^b) = d\mathscr{D}(\varepsilon U)(\dot{U}^a, \dot{U}^b),$$
$$\mathscr{B}''_\varepsilon(U)(\dot{U}^a, \dot{U}^b) = d\mathscr{B}(\varepsilon U)(\dot{U}^a, \dot{U}^b),$$

where we use the short notation

$$d\mathscr{D}(\varepsilon U)(\dot{U}^a, \dot{U}^b) := (d\mathscr{D}(\varepsilon U)\dot{U}^a)\dot{U}^b.$$

Proposition 2.12. *We have*

(a) $|\mathscr{L}''_\varepsilon(U)(\dot{U}^a, \dot{U}^b)|_{\infty,s,\gamma,T}$
$$\leq C(|U|_{*,T})(|\dot{U}^a|_{\infty,s,\gamma,T}|\dot{U}^b|_{*,T} + |\dot{U}^b|_{\infty,s,\gamma,T}|\dot{U}^a|_{*,T} + \varepsilon|U|_{\infty,s,\gamma,T}|\dot{U}^a|_{*,T}|\dot{U}^b|_{*,T}),$$

(b) $|\mathscr{L}''_\varepsilon(U)(\dot{U}^a, \dot{U}^b)|_{0,s+1,\gamma,T}$
$$\leq C(|U|_{*,T})(|\dot{U}^a|_{0,s+1,\gamma,T}|\dot{U}^b|_{*,T} + |\dot{U}^b|_{0,s+1,\gamma,T}|\dot{U}^a|_{*,T} + \varepsilon|U|_{0,s+1,\gamma,T}|\dot{U}^a|_{*,T}|\dot{U}^b|_{*,T}),$$

(c) $\langle\mathscr{B}''_\varepsilon(U)(\dot{U}^a, \dot{U}^b)\rangle_{s,\gamma,T}$
$$\leq C(\langle U\rangle_{*,T})(\langle\dot{U}^a\rangle_{s,\gamma,T}\langle\dot{U}^b\rangle_{*,T} + \langle\dot{U}^b\rangle_{s,\gamma,T}\langle\dot{U}^a\rangle_{*,T} + \varepsilon\langle U\rangle_{s,\gamma,T}\langle\dot{U}^a\rangle_{*,T}\langle\dot{U}^b\rangle_{*,T}).$$

Proof. For $t \leq s$ one computes $\partial^t(\mathscr{L}''_\varepsilon(U)(\dot{U}^a, \dot{U}^b))$, which is a sum of terms of the form

$$\tilde{\mathscr{D}}(\varepsilon U)\partial^{\langle k\rangle}(\varepsilon U)\partial^l \dot{U}^a \partial^m \dot{U}^b, \quad \text{where } k+l+m = t.$$

[25] More precisely, each component is a sum of such terms.

Thus, the first estimate follows directly from Moser estimates. The remaining estimates are proved the same way. □

In the iteration scheme of Section 5B we will use H^s_T spaces on the boundary, while in the interior we use the following spaces.

Definition 2.13. For $s \in \{0, 1, 2, \dots\}$ let

$$E^s_T = CH^s_T \cap L^2 H^{s+1}_T \quad \text{with the norm } |U(x,\theta_0)|_{E^s_T} := |U|_{\infty,s,T} + |U|_{0,s+1,T},$$
$$E^s_{\gamma,T} = CH^s_{\gamma,T} \cap L^2 H^{s+1}_{\gamma,T} \quad \text{with the norm } |U(x,\theta_0)|_{E^s_{\gamma,T}} := |U|_{\infty,s,\gamma,T} + |U|_{0,s+1,\gamma,T}.$$

Remark 2.14. By Sobolev embedding we have

$$s \ge [(d+1)/2] \quad \Rightarrow\ E^s_T \subset CH^s_T \subset L^\infty(\Omega_T),$$
$$s \ge [(d+1)/2]+1 \quad \Rightarrow\ E^s_T \subset CH^s_T \subset L^\infty(\bar{\mathbb{R}}_+, W^{1,\infty}(b\Omega_T)),$$
$$s \ge [(d+1)/2]+M_0 \ \Rightarrow\ E^s_T \subset CH^s_T \subset C^{0,M_0}_T.$$

Note that E^s_T is a Banach algebra for $s \ge [(d+1)/2]$.

By Proposition 2.12 and Remark 2.14 we immediately obtain:

Proposition 2.15 (tame estimates for second derivatives). *Let $\mu_0 = [(d+1)/2]$ and suppose $s \ge 0$ lies in some finite interval. Then*

(a) $$|\mathscr{L}''_\varepsilon(U)(\dot U^a, \dot U^b)|_{E^s_{\gamma,T}} \le C(|U|_{E^{\mu_0}_T})(|\dot U^a|_{E^s_{\gamma,T}}|\dot U^b|_{E^{\mu_0}_T} + |\dot U^b|_{E^s_{\gamma,T}}|\dot U^a|_{E^{\mu_0}_T} + \varepsilon|U|_{E^s_{\gamma,T}}|\dot U^a|_{E^{\mu_0}_T}|\dot U^b|_{E^{\mu_0}_T}),$$

(b) $$\langle \mathscr{B}''_\varepsilon(U)(\dot U^a, \dot U^b)\rangle_{s,\gamma,T} \le C(\langle U\rangle_{\mu_0,T})(\langle \dot U^a\rangle_{s,\gamma,T}\langle \dot U^b\rangle_{\mu_0,T} + \langle \dot U^b\rangle_{s,\gamma,T}\langle \dot U^a\rangle_{\mu_0,T} + \varepsilon\langle U\rangle_{s,\gamma,T}\langle \dot U^a\rangle_{\mu_0,T}\langle \dot U^b\rangle_{\mu_0,T}).$$

In order to obtain a tame estimate for the linearized system suitable for Nash–Moser iteration, we must recast estimate (2-43) without the L^∞ norms of $\dot U$ on the right. First of all, we fix the parameter $K > 0$. For instance, one may take $K = 1$. This choice is arbitrary because we are interested in a small data result.[26] We then choose constants $\varepsilon_0(K)$, $\gamma_0(K)$ as in Proposition 2.11 so that the estimate (2-43) holds for $s \in [0, \tilde\mu]$, where $\tilde\mu$ is defined in (5-57). *For the remainder of Section 2C and in Section 5B, the parameter K is fixed, and γ is also fixed as $\gamma = \gamma_0(K)$.*

Let

$$\kappa := |U|_{0,\mu_0+2,\gamma,T} + \langle U|_{x_d=0}\rangle_{\mu_0+2,\gamma,T}, \quad \text{where } \mu_0 := [(d+1)/2].$$

Applying (2-43) with $s = \mu_0$, we obtain for $0 < \varepsilon \le \varepsilon_0$

$$|\dot U|_{\infty,\mu_0,\gamma,T} + |\dot U|_{0,\mu_0+1,\gamma,T} + \langle \dot U|_{x_d=0}\rangle_{\mu_0+1,\gamma,T} \le C(K,\gamma)(|f|_{0,\mu_0+2,\gamma,T} + \langle g\rangle_{\mu_0+2,\gamma,T} + (|\dot U|_* + \langle \dot U\rangle_*)\kappa). \quad (2\text{-}44)$$

[26] If we were interested in a small time result for a given source term G, we would need to fix the constant K in terms of G and the parameters γ, T would be chosen accordingly.

By Remark 2.14 if κ is chosen small enough, we can absorb the last term on the right in (2-44) and obtain, with a new constant C,

$$|\dot{U}|_* + \langle \dot{U}|_{x_d=0}\rangle_* \le C(|f|_{0,\mu_0+2,\gamma,T} + \langle g\rangle_{\mu_0+2,\gamma,T}). \tag{2-45}$$

Substituting (2-45) in (2-43), we obtain:

Proposition 2.16 (tame estimate for the linearized system). *Let K and $\gamma = \gamma(K)$ be fixed as in Proposition 2.11 and suppose $|\varepsilon\partial_d U|_{C^{0,M_0-1}} + |U|_{C^{0,M_0}} \le K$ for $\varepsilon \in (0,1]$. Let $\mu_0 = [(d+1)/2]$ and $s \in [0, \tilde{\mu}]$, where $\tilde{\mu}$ is defined in (5-57). There exist positive constants $\kappa_0(\gamma, T)$, ε_0, and C such that if*

$$|U|_{0,\mu_0+2,\gamma,T} + \langle U|_{x_d=0}\rangle_{\mu_0+2,\gamma,T} \le \kappa_0,$$

solutions $\dot{U}$ of the linearized system (2-40) satisfy, for $0 < \varepsilon \le \varepsilon_0$,

$$\begin{aligned}|\dot{U}|_{E^s_{\gamma,T}} + \langle \dot{U}|_{x_d=0}\rangle_{s+1,\gamma,T}\\ \le C[|f|_{0,s+2,\gamma,T} + \langle g\rangle_{s+2,\gamma,T} + (|f|_{0,\mu_0+2,\gamma,T} + \langle g\rangle_{\mu_0+2,\gamma,T})(|U|_{0,s+2,\gamma,T} + \langle U|_{x_d=0}\rangle_{s+2,\gamma,T})].\end{aligned}$$

3. Profile equations

3A. ***The key subsystem in the 3 × 3 strictly hyperbolic case.*** To simplify the exposition, we first treat the case of a 3×3 strictly hyperbolic system and a boundary frequency β for which there is one single resonance in which two incoming modes interact to produce an outgoing mode. This case already contains the main difficulties and is exactly the one we emphasize in the example of Appendix B. We explain later the relatively small changes needed to treat the general case of systems satisfying the assumptions of Section 1A.

The leading profile is decomposed as

$$\mathcal{V}^0(x,\theta_1,\theta_2,\theta_3) = \sigma_1(x,\theta_1)r_1 + \sigma_3(x,\theta_3)r_3 \tag{3-1}$$

where ϕ_2 is the outgoing phase and the resonant triple $(n_1,n_2,n_3) \in \mathbb{Z}^3\setminus\{0\}$ satisfies

$$n_1\phi_1 = n_2\phi_2 + n_3\phi_3. \tag{3-2}$$

We can thus write

$$\mathcal{V}^0_{\mathrm{inc}} = \sigma_1(x,\theta_1)r_1 + \sigma_3(x,\theta_3)r_3, \quad \mathcal{V}^1_{\mathrm{out}} = \tau_2(x,\theta_2)r_2. \tag{3-3}$$

Furthermore, we have

$$\mathcal{V}^0_{\mathrm{inc}}|_{x_d=0,\theta_1=\theta_3=\theta_0} = a(x',\theta_0)e = a(x',\theta_0)(e_1+e_3), \tag{3-4}$$

so (recall that $e = e_1 + e_3$, where $e_i \in \operatorname{span}\{r_i\}$ spans $\ker B \cap \mathbb{E}^s(\beta)$)

$$\sigma_i(x',0,\theta_0)r_i = a(x',\theta_0)e_i, \quad i = 1,3, \tag{3-5}$$

which determines the trace of σ_i in terms of a.

Applying the operators E_i for $i = 1, 3$ to (1-42)(a) and for $i = 2$ to (1-42)(b) and using Corollary 1.27 for (1-42)(c), we obtain the following system for the unknowns $(\sigma_1, \tau_2, \sigma_3, a)$, where $\mathscr{A}(x', \theta_0)$ denotes the unique function with mean zero such that $\partial_{\theta_0}\mathscr{A} = a$:

$$\begin{aligned}
&X_{\phi_1}\sigma_1 + c_1\sigma_1 = 0,\\
&X_{\phi_3}\sigma_3 + c_3\sigma_3 = 0,\\
&X_{\phi_2}\tau_2 + c_0\tau_2 + c_2\int_0^{2\pi} \sigma_{1,n_1}\Big(x, \frac{n_2}{n_1}\theta_2 + \frac{n_3}{n_1}\theta_3\Big)\sigma_3(x, \theta_3)\, d\theta_3 = 0\\
&X_{\mathrm{Lop}}\mathscr{A} + c_4\mathscr{A} + c_5\tau_2|_{x_d=0} + c_6(a^2)^* = -b\cdot G^* \quad \text{on } b\Omega_T,
\end{aligned} \tag{3-6}$$

where the first three equations hold on Ω_T, and the constants c_i are readily computed real constants. Here $\sigma_{1,n_1}(x, \theta_1)$ is the image of the function σ_1 under the *preparation map*

$$\sigma_1(x, \theta_1) = \sum_{k\in\mathbb{Z}} f_k(x)e^{ik\theta_1} \to \sum_{k\in\mathbb{Z}} f_{kn_1}(x)e^{ikn_1\theta_1}, \tag{3-7}$$

a map designed so that the integral in (3-6) picks out resonances in the product of σ_1 and σ_3.[27]

Differentiating with respect to θ_0, we rewrite the last equation of (3-6) as

$$X_{\mathrm{Lop}}a + c_4a + c_5\partial_{\theta_0}\tau_2|_{x_d=0} + c_6\partial_{\theta_0}(a^2) = -b\cdot\partial_{\theta_0}G^* =: g \quad \text{on } b\Omega_T. \tag{3-8}$$

We now set $V := (\sigma_1(x, \theta_1), \sigma_3(x, \theta_3), \tau_2(x, \theta_2), a(x', \theta_0))$ and define the interior and boundary operators for the leading profile system:

$$\begin{aligned}
\mathscr{L}(V) &:= \begin{pmatrix} X_{\phi_1}\sigma_1 + c_1\sigma_1 \\ X_{\phi_3}\sigma_3 + c_3\sigma_3 \\ X_{\phi_2}\tau_2 + c_0\tau_2 + c_2\int_0^{2\pi}\sigma_{1,n_1}\big(x, \frac{n_2}{n_1}\theta_2 + \frac{n_3}{n_1}\theta_3\big)\sigma_3(x, \theta_3)\, d\theta_3 \end{pmatrix},\\
\mathscr{B}(V) &:= X_{\mathrm{Lop}}a + c_4a + c_5\partial_{\theta_0}\tau_2|_{x_d=0} + c_6\partial_{\theta_0}(a^2).
\end{aligned} \tag{3-9}$$

In this notation the profile subsystem becomes

$$\begin{aligned}
&\mathscr{L}(V) = 0 \quad \text{in } \Omega_T,\\
&\mathscr{B}(V) = g \quad \text{in } b\Omega_T,\\
&V = 0 \qquad \text{in } t \le 0,
\end{aligned} \tag{3-10}$$

where the additional relations (3-5) hold giving the traces of σ_1, σ_3 in terms of a. The following existence result for the key subsystem is proved in Section 5A using the tame estimates derived in Section 3B below.

Proposition 3.1. *Fix $T > 0$, let $\nu_0 := [(d+1)/2] + 1$, $\nu := 2\nu_0 + 4$, and $\tilde{\nu} := 2\nu - \nu_0$, and suppose $g \in H^{\tilde{\nu}-2}(b\Omega_T)$. Rewrite V as $V = (V', a)$. If $\langle g\rangle_\nu$ is small enough, there exists a solution V of the profile subsystem (3-10) with $V' \in H^{\nu-1}(\Omega_T)$, $(V'|_{x_d=0}, a) \in H^{\nu-1}(b\Omega_T)$.*

[27] Interaction integrals like the one in (3-6) are discussed further in [Coulombel et al. 2011, Proposition 2.13].

Remark 3.2. (1) Although the original problem is semilinear with a nonlinear zero-order boundary condition, the profile system (3-9) has a *quasilinear* first-order boundary operator and an interior operator that includes a nonlinear, nonlocal, integro-pseudodifferential operator given by the interaction integral. The nonlocality arises both from the $d\theta_3$-integration and from the pseudodifferential operator $\sigma_1 \to \sigma_{1,n_1}$.

(2) Attempts to solve the system (3-10) by a standard Picard iteration lead to a (fatal) loss of a derivative from one iterate to the next. The reason is that σ_1 and σ_3 have the regularity of a (incoming transport equation), and therefore τ_2 has the same regularity as a. However, the equation for a involves the derivative $\partial_{\theta_0}\tau_2|_{x_d=0}$ and this term induces the loss. Thus we shall use Nash–Moser iteration to prove Proposition 3.1.

3B. ***Tame estimates.*** With $V = (\sigma_1, \sigma_3, \tau_2, a)$ and $\dot V = (\dot\sigma_1, \dot\sigma_3, \dot\tau_2, \dot a)$, we compute the first derivatives of $\mathscr{L}$ and $\mathscr{B}$:

$$\text{(a)}\quad \mathscr{L}'(V)\dot V = \begin{pmatrix} X_{\phi_1}\dot\sigma_1 + c_1\dot\sigma_1 \\ X_{\phi_3}\dot\sigma_3 + c_3\dot\sigma_3 \\ X_{\phi_2}\dot\tau_2 + c_0\dot\tau_2 + c_2\int_0^{2\pi}\sigma_{1,n_1}\big(x, \frac{n_2}{n_1}\theta_2 + \frac{n_3}{n_1}\theta_3\big)\dot\sigma_3(x,\theta_3)\,d\theta_3 \\ \qquad + c_2\int_0^{2\pi}\sigma_{3,n_3}\big(x, -\frac{n_2}{n_3}\theta_2 + \frac{n_3}{n_1}\theta_1\big)\dot\sigma_1(x,\theta_1)\,d\theta_1 \end{pmatrix}, \tag{3-11}$$

$$\text{(b)}\quad \mathscr{B}'(V)\dot V = X_{\mathrm{Lop}}\dot a + c_4\dot a + c_5\partial_{\theta_0}\dot\tau_2|_{x_d=0} + 2c_6(a\partial_{\theta_0}\dot a + \dot a\partial_{\theta_0}a).$$

Here we have used the property

$$\int_0^{2\pi}\sigma_{3,n_3}\Big(x, -\frac{n_2}{n_3}\theta_2 + \frac{n_3}{n_1}\theta_1\Big)\dot\sigma_1(x,\theta_1)\,d\theta_1 = \int_0^{2\pi}\dot\sigma_{1,n_1}\Big(x, \frac{n_2}{n_1}\theta_2 + \frac{n_3}{n_1}\theta_3\Big)\sigma_3(x,\theta_3)\,d\theta_3, \tag{3-12}$$

which follows readily by looking at the Fourier series of the factors of the integrand. For the second derivatives we obtain

$$\mathscr{L}''(V)(\dot V^a, \dot V^b) = c_2\begin{pmatrix} 0 \\ 0 \\ \int_0^{2\pi}\dot\sigma^a_{1,n_1}\big(x, \frac{n_2}{n_1}\theta_2 + \frac{n_3}{n_1}\theta_3\big)\dot\sigma^b_3(x,\theta_3)\,d\theta_3 + \int_0^{2\pi}\dot\sigma^b_{1,n_1}\big(x, \frac{n_2}{n_1}\theta_2 + \frac{n_3}{n_1}\theta_3\big)\dot\sigma^a_3(x,\theta_3)\,d\theta_3 \end{pmatrix},$$
$$\mathscr{B}''(V)(\dot V^a, \dot V^b) = 2c_6(\dot a^a\partial_{\theta_0}\dot a^b + \dot a^b\partial_{\theta_0}\dot a^a). \tag{3-13}$$

Proposition 3.3 (tame estimates for second derivatives). (a) *Let ν_1 be the smallest integer greater than $(d+2)/2$ and let $s \ge 0$. We have*

$$|\mathscr{L}''(V)(\dot V^a, \dot V^b)|_{s,\gamma} \le C(|\dot V^a|_{s,\gamma}|\dot V^b|_{\nu_1} + |\dot V^b|_{s,\gamma}|\dot V^a|_{\nu_1}), \tag{3-14}$$

where C is independent of V, γ, and T.

(b) *Let ν_2 be the smallest integer greater than $(d+1)/2+1$ and let $s \ge 0$. We have*

$$\langle\mathscr{B}''(V)(\dot V^a, \dot V^b)\rangle_{s,\gamma} \le C(\langle\dot V^a\rangle_{s+1,\gamma}\langle\dot V^b\rangle_{\nu_2} + \langle\dot V^b\rangle_{s+1,\gamma}\langle\dot V^a\rangle_{\nu_2}), \tag{3-15}$$

where C is independent of V, γ, and T.

In (3-14) and (3-15), the constant C can be chosen independent of s in any fixed finite interval.

Proof. (a) Moser estimates imply

$$|\dot\sigma^a_{1,n_1}\Big(x,\frac{n_2}{n_1}\theta_2+\frac{n_3}{n_1}\theta_3\Big)\dot\sigma^b_3(x,\theta_3)|_{H^s_\gamma(x,\theta_2)}\le C(|\dot\sigma^a_{1,n_1}|_{s,\gamma}|\dot\sigma^b_3|_{L^\infty}+|\dot\sigma^a_{1,n_1}|_{L^\infty}|\dot\sigma^b_3|_{H^s_\gamma(x)}),\tag{3-16}$$

since $\dot\sigma^b_3$ is independent of θ_2. We have

$$\int_0^{2\pi}|\dot\sigma^b_3(x,\theta_3)|_{H^s_\gamma(x)}\,d\theta_3\le C|\dot\sigma^b_3|_{L^2(\theta_3,H^s_\gamma(x))}\le C|\dot\sigma^b_3|_{s,\gamma}.\tag{3-17}$$

The estimate (3-14) now follows by Sobolev embedding and the fact that

$$|\dot\sigma^a_{1,n_1}|_{s,\gamma}\le|\dot\sigma^a_1|_{s,\gamma}.\tag{3-18}$$

(b) Again Moser estimates imply

$$\langle\dot a^a\partial_{\theta_0}\dot a^b\rangle_{s,\gamma}\le C(\langle\dot a^a\rangle_{s,\gamma}\langle\partial_{\theta_0}\dot a^b\rangle_{L^\infty}+\langle\dot a^a\rangle_{L^\infty}\langle\partial_{\theta_0}\dot a^b\rangle_{s,\gamma}),\tag{3-19}$$

so the estimate (3-15) follows by Sobolev embedding. □

Next we derive tame energy estimates for the linearized problem

$$\begin{aligned}&\mathscr{L}'(V)\dot V=f\quad&&\text{in }\Omega_T,\\&\mathscr{B}'(V)\dot V=g\quad&&\text{on }b\Omega_T,\\&V=0&&\text{in }t<0,\end{aligned}\tag{3-20}$$

where f and g vanish in $t<0$. We begin with a simple proposition.

Proposition 3.4. (1) *If the phase ϕ_p is incoming, solutions of*

$$X_{\phi_p}\sigma_p+c_p\sigma_p=h\quad\text{in }\Omega_T,\quad\sigma_p=0\quad\text{in } t<0\tag{3-21}$$

satisfy, for γ large (depending on c_p),

$$\sqrt\gamma|\sigma_p|_{s,\gamma}\le C\Big(\langle\sigma_p\rangle_{s,\gamma}+\frac{|h|_{s,\gamma}}{\sqrt\gamma}\Big).\tag{3-22}$$

(2) *If the phase ϕ_p is outgoing, solutions of (3-21) satisfy, for γ large (depending on c_p),*

$$\sqrt\gamma|\sigma_p|_{s,\gamma}+\langle\sigma_p\rangle_{s,\gamma}\le C\frac{|h|_{s,\gamma}}{\sqrt\gamma}.\tag{3-23}$$

(3) *Solutions in ω_T of*

$$X_{\mathrm{Lop}}\dot a+c_4\dot a+2c_6(a\partial_{\theta_0}\dot a+\dot a\partial_{\theta_0}a)=g,\dot a=0\quad\text{in } t<0\tag{3-24}$$

satisfy, for C_K, $\gamma\ge\gamma_K$ (where $K=\langle a\rangle_{W^{1,\infty}}$),

$$\sqrt\gamma\langle\dot a\rangle_{s,\gamma}\le\frac{C_K}{\sqrt\gamma}\left(\langle g\rangle_{s,\gamma}+\langle a\rangle_{s+1,\gamma}\langle\dot a\rangle_{W^{1,\infty}}\right).\tag{3-25}$$

The second term on the right in (3-25) does not appear in the $s=0$ estimate.

Proof. (1) To prove (3-23) with $s = 0$, one considers the problem satisfied by $e^{-\gamma t}\sigma_p$, multiplies the equation by $e^{-\gamma t}\sigma_p$, integrates $dx\,d\theta_p$ on Ω_T, and performs obvious integrations by parts. One then applies the L^2 estimate to the problem satisfied by tangential derivatives $\gamma^{s-|\beta|}\partial^{\beta}_{x',\theta_p}\sigma_p$, $|\beta| \le s$. Normal derivatives are estimated using the equation and the tangential estimates. The proof of (3-23) is similar. We refer to [Benzoni-Gavage and Serre 2007] for a complete discussion of such estimates.

(2) The proof of (3-25) is similar, but in the higher derivative estimates one now has forcing terms that are commutators involving a. The commutators are linear combinations of terms of the form

$$\gamma^{s-|\beta|}(\partial^{\beta_1}_{x',\theta_0}a)(\partial^{\beta_2}_{x',\theta_0}\partial_{\theta_0}\dot a) \quad \text{where } |\beta_1|+|\beta_2| = |\beta|,\ |\beta_1| \ge 1, \tag{3-26}$$

or linear combinations of terms of the form

$$\gamma^{s-|\beta|}(\partial^{\beta_1}_{x',\theta_0}\dot a)(\partial^{\beta_2}_{x',\theta_0}\partial_{\theta_0}a) \quad \text{where } |\beta_1|+|\beta_2| = |\beta|,\ |\beta_2| \ge 1. \tag{3-27}$$

Applying Moser estimates to (3-26) after writing $\partial^{\beta_1}a = \partial^{\beta_1'}\partial a$, we obtain

$$\begin{aligned}\langle\gamma^{s-|\beta|}(\partial^{\beta_1}_{x',\theta_0}a)(\partial^{\beta_2}_{x',\theta_0}\partial_{\theta_0}\dot a)\rangle_{0,\gamma} &\le C(\langle\partial a\rangle_{L^\infty}\langle\partial_{\theta_0}\dot a\rangle_{m-1,\gamma} + \langle\partial a\rangle_{m-1,\gamma}\langle\partial_{\theta_0}\dot a\rangle_{L^\infty})\\ &\le C(\langle a\rangle_{W^{1,\infty}}\langle\dot a\rangle_{s,\gamma} + \langle a\rangle_{s,\gamma}\langle\dot a\rangle_{W^{1,\infty}}).\end{aligned} \tag{3-28}$$

The factor $C_K/\sqrt{\gamma}$ on the forcing term in the L^2 estimate allows the first term on the right to be absorbed by taking γ large.

The estimate of (3-27) is similar, but we do not split the ∂^{β_2} derivative, and after absorbing a term we are left with $(C_K/\sqrt{\gamma})\langle\dot a\rangle_{L^\infty}\langle a\rangle_{s+1,\gamma}$ on the right. □

We now use Proposition 3.4 to estimate solutions of the linearized problem (3-20) by treating the interaction integrals in (3-11)(a) and the term $c_5\partial_{\theta_0}\tau_2$ in (3-11)(b) as additional forcing terms. Setting

$$V_{\mathrm{inc},n} := (\sigma_{1,n_1}, \sigma_{3,n_3}), \quad V_{\mathrm{inc}} = (\sigma_1, \sigma_3),\ V_{\mathrm{out}} = \tau_2, \tag{3-29}$$

estimating interaction integrals as in (3-16) and (3-17), and using (3-18), we immediately obtain

$$\begin{aligned}\sqrt{\gamma}|\dot V_{\mathrm{out}}|_{s,\gamma} + \langle\dot V_{\mathrm{out}}\rangle_{s,\gamma} &\le \frac{C}{\sqrt{\gamma}}(|f|_{s,\gamma} + |V_{\mathrm{inc},n}|_{L^\infty}|\dot V_{\mathrm{inc}}|_{s,\gamma} + |V_{\mathrm{inc}}|_{s,\gamma}|\dot V_{\mathrm{inc}}|_{L^\infty}),\\ \sqrt{\gamma}|\partial_\theta\dot V_{\mathrm{out}}|_{s,\gamma} + \langle\partial_{\theta_0}\dot V_{\mathrm{out}}\rangle_{s,\gamma} &\le \frac{C}{\sqrt{\gamma}}(|\partial_\theta f|_{s,\gamma} + |\partial_\theta V_{\mathrm{inc},n}|_{L^\infty}|\dot V_{\mathrm{inc}}|_{s,\gamma} + |\partial_\theta V_{\mathrm{inc}}|_{m,\gamma}|\dot V_{\mathrm{inc}}|_{L^\infty}),\end{aligned} \tag{3-30}$$

and

$$\begin{aligned}\sqrt{\gamma}|\dot V_{\mathrm{inc}}|_{s,\gamma} &\le C\left(\langle\dot V_{\mathrm{inc}}\rangle_{s,\gamma} + \frac{|f|_{s,\gamma}}{\sqrt{\gamma}}\right),\\ \sqrt{\gamma}\langle\dot V_{\mathrm{inc}}\rangle_{s,\gamma} &\le \frac{C_K}{\sqrt{\gamma}}(\langle g\rangle_{s,\gamma} + \langle\partial_{\theta_0}\dot V_{\mathrm{out}}\rangle_{s,\gamma} + \langle V_{\mathrm{inc}}\rangle_{s+1,\gamma}\langle\dot V_{\mathrm{inc}}\rangle_{W^{1,\infty}}).\end{aligned} \tag{3-31}$$

This leads to the following "pretame" estimate.

Proposition 3.5. *Let $\mu_0 := [(d+1)/2]$, fix $K_1 > 0$, and suppose $|V_{\mathrm{inc}}|_{\mu_0+2} \le K_1$.*[28] *For $s \ge 0$ in any fixed finite interval, there exist constants $C(K_1)$, $\gamma(K_1)$ such that, for $\gamma \ge \gamma(K_1)$, solutions of the linearized problem (3-20) satisfy*

$$\sqrt{\gamma}|\dot V_{\mathrm{out}}, \partial_\theta \dot V_{\mathrm{out}}, \dot V_{\mathrm{inc}}|_{s,\gamma} + \langle \dot V_{\mathrm{out}}, \partial_{\theta_0} \dot V_{\mathrm{out}}\rangle_{s,\gamma} + \sqrt{\gamma}\langle \dot V_{\mathrm{inc}}\rangle_{s,\gamma} \le \frac{C(K_1)}{\sqrt{\gamma}}(|f|_{s+1,\gamma} + \langle g\rangle_{s,\gamma} + |V_{\mathrm{inc}}|_{s+1,\gamma}|\dot V_{\mathrm{inc}}|_{L^\infty} + \langle V_{\mathrm{inc}}\rangle_{s+1,\gamma}\langle \dot V_{\mathrm{inc}}\rangle_{W^{1,\infty}}). \quad (3\text{-}32)$$

Proof. We add the estimates (3-30) and (3-31) and absorb the terms

$$\frac{C_K}{\sqrt{\gamma}}(\langle \dot V_{\mathrm{inc}}\rangle_{s,\gamma} + \langle \partial_{\theta_0}\dot V_{\mathrm{out}}\rangle_{s,\gamma} + |V_{\mathrm{inc},n}, \partial_\theta V_{\mathrm{inc},n}|_{L^\infty}|\dot V_{\mathrm{inc}}|_{s,\gamma}) \quad (3\text{-}33)$$

by taking γ large, after observing that

$$|V_{\mathrm{inc},n}, \partial_\theta V_{\mathrm{inc},n}|_{L^\infty} \le C|V_{\mathrm{inc},n}|_{\mu_0+2} \le C|V_{\mathrm{inc}}|_{\mu_0+2} \quad (3\text{-}34)$$

and

$$K = \langle V_{\mathrm{inc}}\rangle_{W^{1,\infty}} \le C|V_{\mathrm{inc}}|_{\mu_0+2}. \qquad \square$$

We now set $\tilde\nu := 2\nu - \nu_0$ as in Proposition 3.1 and choose constants $C(K_1)$, $\gamma(K_1)$ as in Proposition 3.5 corresponding to the interval $s \in [0, \tilde\nu]$.[29] *In the remainder of Section 3B and also in Section 5A, γ is fixed as $\gamma = \gamma(K_1)$.*

To obtain a tame estimate, we need to remove the terms depending on $\dot V_{\mathrm{inc}}$ on the right side of (3-32). Let

$$K_2 = |V_{\mathrm{inc}}|_{\mu_0+2,\gamma} + \langle V_{\mathrm{inc}}\rangle_{\mu_0+2,\gamma}. \quad (3\text{-}35)$$

Applying (3-32) with $s = \mu_0 + 1$, we obtain

$$\sqrt{\gamma}|\dot V_{\mathrm{inc}}|_{\mu_0+1,\gamma} + \sqrt{\gamma}\langle \dot V_{\mathrm{inc}}\rangle_{\mu_0+1,\gamma} \le \frac{C(K_1)}{\sqrt{\gamma}}[|f|_{\mu_0+2,\gamma} + \langle g\rangle_{\mu_0+1,\gamma} + (|\dot V_{\mathrm{inc}}|_{L^\infty} + \langle \dot V_{\mathrm{inc}}\rangle_{W^{1,\infty}})K_2]. \quad (3\text{-}36)$$

By Sobolev embedding, if $K_2 = K_2(\gamma, T)$ is chosen small enough, we can absorb the last term on the right in (3-36) and obtain, with a new C,

$$|\dot V_{\mathrm{inc}}|_{L^\infty} + \langle \dot V_{\mathrm{inc}}\rangle_{W^{1,\infty}} \le C(|f|_{\mu_0+2,\gamma} + \langle g\rangle_{\mu_0+1,\gamma}). \quad (3\text{-}37)$$

For γ fixed as above, setting $|U|_{s,\gamma} = |U|_s$ now and substituting (3-37) in (3-32), we obtain the estimate in the following proposition.

Proposition 3.6 (tame estimate for the linearized system). *Let $\mu_0 = [(d+1)/2]$ and $s \in [0, \tilde\nu]$. There exists $\kappa_0 = \kappa_0(\gamma, T) > 0$ and a constant C depending on κ_0 such that if*

$$|V_{\mathrm{inc}}|_{\mu_0+2} + \langle V_{\mathrm{inc}}\rangle_{\mu_0+2} \le \kappa_0, \quad (3\text{-}38)$$

solutions of the linearized system (3-20) satisfy

$$|\dot V|_s + \langle \dot V\rangle_s \le C[|f|_{s+1} + \langle g\rangle_s + (|f|_{\mu_0+2} + \langle g\rangle_{\mu_0+1})(|V|_{s+1} + \langle V\rangle_{s+1})]. \quad (3\text{-}39)$$

[28] In this proposition $\mu_0 = [d/2]$ would work, but we make the above choice so as not to have to redefine μ_0 later.

[29] The choice of $\tilde\nu$ is explained in Section 5A.

Proof. We have proved the a priori estimate (3-39) for sufficiently smooth solutions of the linearized system. The existence of such solutions now follows by standard arguments, which we summarize here for completeness.

The unknown in the linearized system (3-20) is $(\dot\sigma_1, \dot\sigma_3, \dot\tau_2, \dot a)$. We can solve the linearized system by putting the terms that involve $\partial_{\theta_0}\dot\tau_2$ or $\partial_{\theta_0}\dot a$ on the right and replacing the operator ∂_{θ_0}, when it acts on those terms, by a finite difference operator $\partial^h_{\theta_0}$:

$$\begin{aligned}
&X_{\phi_1}\dot\sigma_1^h + c_1\dot\sigma_1^h = f_1,\\
&X_{\phi_3}\dot\sigma_3^h + c_3\dot\sigma_3^h = f_2,\\
&X_{\phi_2}\dot\tau_2^h + c_0\dot\tau_2^h = f_3 - c_2\textstyle\int_0^{2\pi}\sigma_{1,n_1}\big(x, \frac{n_2}{n_1}\theta_2 + \frac{n_3}{n_1}\theta_3\big)\dot\sigma_3^h(x,\theta_3)\,d\theta_3\\
&\qquad\qquad - c_2\textstyle\int_0^{2\pi}\sigma_{3,n_3}\big(x, -\frac{n_2}{n_3}\theta_2 + \frac{n_3}{n_1}\theta_1\big)\dot\sigma_1^h(x,\theta_1)\,d\theta_1,\\
&X_{\mathrm{Lop}}\dot a^h + c_4\dot a^h + 2c_6\dot a^h\partial_{\theta_0}a = g - c_5\partial^h_{\theta_0}\dot\tau_2^h - 2c_6 a\partial^h_{\theta_0}\dot a^h.
\end{aligned} \tag{3-40}$$

For fixed $h \in (0, 1]$ we can solve this system by Picard iteration, where n-th iterates appear on the right and $(n+1)$-st iterates appear on the left. All iterates are 0 in $t < 0$ and the iterates with index zero are all 0.

We then need an estimate that is uniform in h. This can be done by repeating the existing proof of tame estimates, using the operator $\partial^h_{\theta_0}$ in place of ∂_{θ_0}. This gives an estimate like (3-39):

$$|\dot V^h|_s + \langle \dot V^h\rangle_s \le C[|f|_{s+1} + \langle g\rangle_s + (|f|_{\mu_0+2} + \langle g\rangle_{\mu_0+1})(|V|_{s+1} + \langle V\rangle_{s+1})], \tag{3-41}$$

where $\dot V^h := (\dot\sigma_1^h, \dot\sigma_3^h, \dot\tau_2^h, \dot a^h)$ and C is uniform for $h \in (0, 1]$. Passing to a subsequence, we obtain the desired solution of the linearized system. □

Remark 3.7 (short time, given data). For a given $T > 0$, let K_1 and $\gamma = \gamma(K_1)$ be as in Proposition 3.5. As we saw above, to obtain a tame estimate, we need to take $|\mathcal{V}_{\mathrm{inc}}|_{\mu_0+2} + \langle\mathcal{V}_{\mathrm{inc}}\rangle_{\mu_0+2}$ small. In our formulation of Theorem 1.15, T is fixed ahead of time and we achieve (3-38) by taking G small in an appropriate norm on Ω_T. For a given G as in (1-2) vanishing in $t < 0$, another way to proceed is to shrink T; that is, to work on Ω_{T_1} where $0 < T_1 < T$ is chosen so that $\gamma_1 := 1/T_1 \ge \gamma(K_1)$ and so that

$$|V_{\mathrm{inc}}|_{H^{\mu_0+2}_{\gamma_1}(\Omega_{T_1})} + \langle V_{\mathrm{inc}}\rangle_{H^{\mu_0+2}_{\gamma_1}(\omega_{T_1})}$$

is small enough to absorb the terms involving $\dot V_{\mathrm{inc}}$ on the right in (3-36). One again obtains an estimate of the form (3-39), where now

$$|U|_s := |U|_{H^s_{\gamma_1}(\Omega_{T_1})}.$$

The iteration scheme described in Section 5A applies to this situation with no essential change as well.

3C. ***The key subsystem in the general case.*** Recall that $\{1, \dots, M\} = \mathbb{O} \cup \mathcal{I}$, where $\mathbb{O}$ and $\mathcal{I}$ contain the indices corresponding to outgoing and incoming phases. We further decompose $\mathbb{O} = \mathbb{O}_1 \cup \mathbb{O}_2$, where $\mathbb{O}_1$ consists of indices m such that ϕ_m is part of at least one triple of resonant phases with the property that the other two phases in that triple are incoming. For a given $m \in \mathbb{O}_1$ the phase ϕ_m might belong to more than one such triple.

Now instead of (3-3) we have

$$\mathcal{V}^0_{\rm inc} = \sum_{m\in\mathcal{J}} \sum_{k=1}^{\nu_{k_m}} \sigma_{m,k}(x,\theta_m) r_{m,k} \quad \text{and} \quad \mathcal{V}^1_{\rm out} = \sum_{m\in\mathbb{O}_1} \sum_{k=1}^{\nu_{k_m}} \tau_{m,k}(x,\theta_m) r_{m,k}, \tag{3-42}$$

since terms $\tau_{m,k}$ in the expansion of $\mathcal{V}^1_{\rm out}$ vanish if $m \in \mathbb{O}_2$ as a consequence of (1-36) and $\mathcal{V}^0 = \mathcal{V}^0_{\rm inc}$. Recalling that

$$e = \sum_{m\in\mathcal{J}} \sum_{k=1}^{\nu_{k_m}} e_{m,k}, \quad \text{where } e_{m,k} \in \operatorname{span}\{r_{m,k}\}, \tag{3-43}$$

we see that in place of (3-4) we now have

$$\mathcal{V}^0_{\rm inc}|_{x_d=0;\theta_m=\theta_0,m\in I} = a(x',\theta_0)e = \sum_{m\in\mathcal{J}} \sum_{k=1}^{\nu_{k_m}} a e_{m,k} = \sum_{m\in\mathcal{J}} \sum_{k=1}^{\nu_{k_m}} \sigma_{m,k}(x',0,\theta_0) r_{m,k}, \tag{3-44}$$

and thus

$$\sigma_{m,k}(x',0,\theta_0) r_{m,k} = a(x',\theta_0) e_{m,k} \quad \text{for } m \in \mathcal{J},\ k = 1,\dots,\nu_{k_m}. \tag{3-45}$$

Next we derive the formulas for $\mathcal{L}(V)$ and $\mathcal{B}(V)$ in the general case. The unknown is now

$$V = (\sigma_{m,k}, m \in \mathcal{J}, k = 1,\dots,\nu_{k_m}; \tau_{m,k}, m \in \mathbb{O}_1, k = 1,\dots,\nu_{k_m}; a). \tag{3-46}$$

Suppose $q \in \mathbb{O}_1$ and that (ϕ_p, ϕ_q, ϕ_s) is a resonant triple such that

$$n_p\phi_p = n_q\phi_q + n_s\phi_s \quad \text{where } p, s \in \mathcal{J} \text{ and } \gcd(n_p, n_q, n_s) = 1. \tag{3-47}$$

Applying the projectors $E_{m,k}$, $m \in \mathcal{J}$, $k = 1,\dots,\nu_{k_m}$ to (1-42)(a) and the projectors $E_{q,l}$, $q \in \mathbb{O}_1$, $l = 1,\dots,\nu_{k_q}$ to (1-42)(b), we obtain

$$\mathcal{L}(V) = \begin{pmatrix} X_{\phi_m}\sigma_{m,k} + c_{m,k}\sigma_{m,k}; \quad m \in \mathcal{J},\ k = 1,\dots,\nu_{k_m} \\ X_{\phi_q}\tau_{q,l} + c_{q,l}\tau_{q,l} + \sum_{k=1}^{\nu_{k_p}} \sum_{k'=1}^{\nu_{k_s}} d^{k,k'}_{q,l} \int_0^{2\pi} (\sigma_{p,k})_{n_p}\big(x, \tfrac{n_q}{n_p}\theta_q + \tfrac{n_s}{n_p}\theta_s\big)\sigma_{s,k'}(x,\theta_s)\mathrm{d}\theta_s + (\text{similar}); \\ q \in \mathbb{O}_1,\ l = 1,\dots,\nu_{k_q} \end{pmatrix}. \tag{3-48}$$

Here "similar" denotes a finite sum of families (i.e., sums over k and k') of integrals similar to the one explicitly given. One such family corresponds to each distinct resonant triple involving the outgoing phase ϕ_q and two incoming phases.[30] The values of the real constants $c_{m,k}$, $d^{k,k'}_{q,l}$ are not important for our analysis, but, for example, the $d^{k,k'}_{q,l}$ are given by[31]

$$d^{k,k'}_{q,l} = \frac{1}{2\pi} \ell_{q,l} \cdot [dD(0)(r_{p,k}, r_{s,k'}) + dD(0)(r_{s,k'}, r_{p,k})]. \tag{3-49}$$

By a computation similar to the one that produced (3-8), we obtain from (1-42)(c)

[30] We do not distinguish between (ϕ_p, ϕ_q, ϕ_s) and (ϕ_p, ϕ_s, ϕ_q). We do distinguish between (ϕ_p, ϕ_q, ϕ_s) and (ϕ_p, ϕ_q, ϕ_t).

[31] We have suppressed indices r, s on the $d^{k,k'}_{q,l}$.

$$\mathcal{B}(V) = X_{\mathrm{Lop}}a + f_1 a + \sum_{q\in\mathbb{O}_1}\sum_{l=1}^{\nu_{kq}} f_{q,l}\partial_{\theta_0}\tau_{q,l} + f_2\partial_{\theta_0}(a^2) \tag{3-50}$$

for some real constants f_1, f_2, $f_{q,l}$. For example, we have $f_2 = -b\cdot[\psi'(0)(e,e)]$. Thus the system (1-42) may be rewritten

$$\begin{aligned} &\mathcal{L}(V) = 0 && \text{in } \Omega_T,\\ &\mathcal{B}(V) = -b\cdot\partial_{\theta_0}G^* := g && \text{on } b\Omega_T,\\ &V = 0 && \text{in } t<0, \end{aligned} \tag{3-51}$$

where the relations (3-45) hold.

It is now a simple matter to write out the expressions for the first and second derivatives of $\mathcal{L}$ and $\mathcal{B}$. For example, just as the interaction integral in (3-9) gave rise to two integrals in the expression (3-11) for $\mathcal{L}'(V)$ in the 3×3 case, it is clear that each integral in (3-48) will give rise to two integrals in the new expression for $\mathcal{L}'(V)$. The tame estimates for second derivatives are proved exactly as before, and Proposition 3.3 holds verbatim in the general case. Proposition 3.4 is used exactly as before to prove estimates for the linearized system. With the unknown V as given in (3-46) and after defining V_{inc}, V_{out}, $\dot V_{\mathrm{inc}}$, $\dot V_{\mathrm{out}}$ in the obvious way, we see that the "pretame" estimate of Proposition 3.5 and the tame estimate of Proposition 3.6 hold verbatim in the general case. The iteration scheme of Section 5A depends only on the tame estimates. Thus it applies here without change and Proposition 3.1 holds verbatim in the general case.

Once the key subsystem is solved, we can easily complete the solution of the full profile system (1-35)–(1-36). The precise result for the full system will be proved in Theorem 5.11.

4. Error analysis

Here we carry out the error analysis sketched in Section 1E, beginning with the proof of Proposition 1.29.

Proof of Proposition 1.29. Step 1: *Noncharacteristic modes.* We write

$$F(x,\theta) = F_0(x) + \sum_{\alpha\notin\mathscr{C}} F_\alpha(x)e^{i\alpha\cdot\theta} + \sum_{m=1}^{M}\sum_{\alpha\in\mathscr{C}_m\setminus\{0\}} F_\alpha(x)e^{i\alpha\cdot\theta},$$

and recall that the sums are finite. Set

$$n_\alpha = \sum_{j=1}^{M}\alpha_j \quad\text{and}\quad \underline{\omega} = (\underline{\omega}_1,\dots,\underline{\omega}_M).$$

Since $EF = 0$, we first note that F_0 vanishes. For any α, we have

$$\left(F_\alpha(x)e^{i\alpha\cdot\theta}\right)|_{\theta\to(\theta_0,\xi_d)} = F_\alpha(x)e^{in_\alpha\theta_0 + i(\alpha\cdot\underline{\omega})\xi_d},$$

and when $\alpha\notin\mathscr{C}$, we look for $U_\alpha(x)$ such that

$$\mathcal{L}_0(\partial_{\theta_0},\partial_{\xi_d})U_\alpha(x)e^{in_\alpha\theta_0+i(\alpha\cdot\underline{\omega})\xi_d} = F_\alpha(x)e^{in_\alpha\theta_0+i(\alpha\cdot\underline{\omega})\xi_d}. \tag{4-1}$$

This holds if and only if

$$iL(n_\alpha\beta, \alpha\cdot\underline{\omega})U_\alpha = F_\alpha.$$

The matrix on the left is invertible, so we obtain a solution of (4-1) for $\alpha \notin \mathscr{C}$.

Step 2: *Characteristic modes.* When $\alpha \in \mathscr{C}_m \setminus \{0\}$, we have $\alpha\cdot\underline{\omega} = n_\alpha\underline{\omega}_m$, so

$$(F_\alpha(x)e^{i\alpha\cdot\theta})|_{\theta\to(\theta_0,\xi_d)} = F_\alpha(x)e^{in_\alpha(\theta_0+\underline{\omega}_m\xi_d)}.$$

We can write

$$\sum_{\alpha\in\mathscr{C}_m\setminus\{0\}} F_\alpha(x)e^{in_\alpha(\theta_0+\underline{\omega}_m\xi_d)} = \sum_{k\in\mathbb{Z}\setminus\{0\}} \mathscr{F}_{m,k}(x)e^{ik(\theta_0+\underline{\omega}_m\xi_d)},$$

where

$$\mathscr{F}_{m,k}(x) := \sum_{\{\alpha\in\mathscr{C}_m\setminus 0,\, n_\alpha=k\}} F_\alpha(x).$$

Since $E_mF = 0$, we have for each $k\in\mathbb{Z}\setminus\{0\}$ that $P_m\mathscr{F}_{m,k}(x) = 0$. So now we look for $U_{m,k}(x)$ such that

$$\mathscr{L}_0(\partial_{\theta_0}, \partial_{\xi_d})U_{m,k}(x)e^{ik(\theta_0+\underline{\omega}_m\xi_d)} = (I-P_m)\mathscr{F}_{m,k}e^{ik(\theta_0+\underline{\omega}_m\xi_d)}.$$

The latter relation holds if and only if

$$iL(k\beta, k\underline{\omega}_m)U_{m,k}(x) = ikL(d\phi_m)U_{m,k}(x) = (I-P_m)\mathscr{F}_{m,k}(x),$$

which is solvable even though $L(d\phi_m)$ is singular. Finally, we take

$$\mathscr{U}(x,\theta_0,\xi_d) = \sum_{\alpha\notin\mathscr{C}} U_\alpha(x)e^{in_\alpha\theta_0+i(\alpha\cdot\underline{\omega})\xi_d} + \sum_{m=1}^{M}\sum_{k\in\mathbb{Z}\setminus\{0\}} U_{m,k}(x)e^{ik(\theta_0+\underline{\omega}_m\xi_d)},$$

which solves (1-46) as claimed. □

The existence theorems for profiles and for the exact solution to the singular system, Theorems 5.11 and 5.13, respectively, are stated and proved in Section 5; we shall only use the statement of these theorems here. In order to formulate the main result of this section we must make some preliminary choices.

Choice of μ *and* $\tilde{\mu}$. The conditions on the boundary datum $G(x',\theta_0)$ are slightly different in Theorems 5.11 and 5.13. We need to choose μ, $\tilde{\mu}$, and $G(x',\theta_0)$ so that both theorems apply simultaneously. We also need μ large enough so that we can apply Proposition 2.9 in step (4-24) of the error analysis below. These conditions are met if we take[32]

$$\mu = \max(d+9, [(d+1)/2]+M_0+3) = [(d+1)/2]+M_0+3 \quad\text{and}\quad \tilde{\mu} = 2\mu - [(d+1)/2] \tag{4-2}$$

and choose $G\in H^{\tilde{\mu}}(b\Omega_T)$ such that $\langle G\rangle_{H^{\mu+2}(b\Omega_T)}$ is small enough. Applying Theorems 5.11 and 5.13, we now have, for $0<\varepsilon\le\varepsilon_0$, an exact solution $U_\varepsilon(x,\theta_0)\in E^{\mu-1}(\Omega_T)$ to the singular system (1-18) and profiles $\mathscr{V}^0(x,\theta)\in H^{\mu-1}(\Omega_T)$, $\mathscr{V}^1(x,\theta)\in H^{\mu-2}(\Omega_T)$ satisfying the equations (1-35) and (1-36).

[32]Recall that $M_0 = 3d+5$, $d\ge 2$.

Approximation. Fix $\delta > 0$. Using the Fourier series of $\mathcal{V}^0$ and $\mathcal{V}^1$, we choose trigonometric polynomials $\mathcal{V}^0_p(x,\theta)$ and $\mathcal{V}^1_p(x,\theta)$ such that

$$|\mathcal{V}^0 - \mathcal{V}^0_p|_{H^{\mu-1}(\Omega_T)} < \delta, \quad |\mathcal{V}^1 - \mathcal{V}^1_p|_{H^{\mu-2}(\Omega_T)} < \delta. \tag{4-3}$$

We can smooth the coefficients so that $\mathcal{V}^0_p$ and $\mathcal{V}^1_p$ lie in $H^\infty(\Omega_T)$ and so that (4-3) still holds. Having made these choices, we can now state the main result of this section, which yields the final convergence result of Theorem 1.15 as an immediate corollary.

Theorem 4.1. *We make the same assumptions as in Theorem 1.15 and let μ and $\tilde{\mu}$ be as just chosen. Consider the leading-order approximate solution to the singular semilinear system (1-18) given by*

$$\mathcal{U}^0_\varepsilon(x,\theta_0) := \mathcal{V}^0(x,\theta)|_{\theta\to(\theta_0,x_d/\varepsilon)}, \tag{4-4}$$

and let $U_\varepsilon(x,\theta_0) \in E^{\mu-1}(\Omega_T)$ be the exact solution to (1-18) just obtained. Then

$$\lim_{\varepsilon\to 0} |U_\varepsilon(x,\theta_0) - \mathcal{U}^0_\varepsilon(x,\theta_0)|_{E^{\mu-3}(\Omega_T)} = 0. \tag{4-5}$$

The following lemma, which is proved in [Coulombel et al. 2011, Lemmas 2.7 and 2.25] by a simple argument based on Fourier series, is an important tool in the proof.

Lemma 4.2 (relation between norms). *For $m \in \mathbb{N}$ suppose $f(x,\theta_j) \in H^{m+1}(\Omega_T)$, and set $f_\varepsilon(x,\theta_0) := f(x,\theta_0 + \underline{\omega}_j x_d/\varepsilon)$. Then*

$$|f_\varepsilon|_{E^m_T} \le C|f|_{H^{m+1}(\Omega_T)}. \tag{4-6}$$

Proof of Theorem 4.1. We shall fill in the sketch provided in Section 1E.

Step 1. First we use Proposition 1.29 to construct $\mathcal{U}^2_p(x,\theta_0,\xi_d)$ satisfying

$$\mathcal{L}_0(\partial_{\theta_0},\partial_{\xi_d})\mathcal{U}^2_p = [-(I-E)(L(\partial)\mathcal{V}^1_p + D(0)\mathcal{V}^1_p + dD(0)(\mathcal{V}^0_p,\mathcal{V}^0_p))]|_{\theta\to(\theta_0,\xi_d)}. \tag{4-7}$$

The function $\mathcal{U}^2_p$ is a trigonometric polynomial of the form (1-47) with H^∞ coefficients. We then define the corrected approximate solution

$$\mathcal{U}_\varepsilon(x,\theta_0) := (\mathcal{V}^0(x,\theta) + \varepsilon\mathcal{V}^1(x,\theta))|_{\theta\to(\theta_0,x_d/\varepsilon)} + \varepsilon^2\mathcal{U}^2_p\left(x,\theta_0,\frac{x_d}{\varepsilon}\right). \tag{4-8}$$

Since $\mathcal{V}^1 \in H^{\mu-2}(\Omega_T)$, Lemma 4.2 implies $\mathcal{U}_\varepsilon \in E^{\mu-3}(\Omega_T)$.

Step 2. Next we explain (1-48) and make precise the norms used on the right there. Using the identity (1-44), we compute

$$\begin{aligned}\mathbb{L}_\varepsilon(\mathcal{U}_\varepsilon) = \varepsilon[(\mathcal{L}_0(\partial_{\theta_0},\partial_{\xi_d})\mathcal{U}^2_p)|_{\xi_d=x_d/\varepsilon} + (L(\partial)\mathcal{V}^1 + D(0)\mathcal{V}^1 + dD(0)\mathcal{V}^0\mathcal{V}^0)|_{\theta\to(\theta_0,x_d/\varepsilon)}]\\ + \varepsilon^2(L(\partial)\mathcal{U}^2_p)|_{\xi_d=x_d/\varepsilon} + D(\varepsilon\mathcal{U}_\varepsilon)\mathcal{U}_\varepsilon - D(0)(\mathcal{V}^0 + \varepsilon\mathcal{V}^1) - \varepsilon dD(0)\mathcal{V}^0\mathcal{V}^0,\end{aligned} \tag{4-9}$$

where the second line represents an $O(\varepsilon^2)$ term (see below for a precise estimate). Here the profile equations (1-20)(a)–(b) imply that the terms of order ε^{-1} and ε^0 vanish. Using (4-7), we can rewrite the

coefficient of ε in (4-9) as

$$[L(\partial)(\mathcal{V}^1-\mathcal{V}^1_p)+D(0)(\mathcal{V}^1-\mathcal{V}^1_p)+dD(0)(\mathcal{V}^0\mathcal{V}^0-\mathcal{V}^0_p\mathcal{V}^0_p)]|_{\theta\to(\theta_0,x_d/\varepsilon)}$$
$$+[E(L(\partial)\mathcal{V}^1_p+D(0)\mathcal{V}^1_p+dD(0)\mathcal{V}^0_p\mathcal{V}^0_p)]|_{\theta\to(\theta_0,x_d/\varepsilon)}:=A+B. \quad (4\text{-}10)$$

Using (4-3), Lemma 4.2, and the fact that $E^s(\Omega_T)$ is a Banach algebra for $s\geq[(d+1)/2]$, we see that

$$|A|_{E^{\mu-4}(\Omega_T)}\leq K\delta. \quad (4\text{-}11)$$

To estimate B, let

$$F:=L(\partial)\mathcal{V}^1+D(0)\mathcal{V}^1+dD(0)\mathcal{V}^0\mathcal{V}^0 \quad \text{and} \quad F_p:=L(\partial)\mathcal{V}^1_p+D(0)\mathcal{V}^1_p+dD(0)\mathcal{V}^0_p\mathcal{V}^0_p. \quad (4\text{-}12)$$

The profile equation (1-36)(b) implies $EF=0$. Using continuity of the multiplication map (1-25), we see that (4-3) implies[33]

$$|F-F_p|_{H^{\mu-3;2}_T}\leq K\delta. \quad (4\text{-}13)$$

From the continuity of $E:H^{s;2}_T\to H^{s;1}_T$ and Lemma 4.2 we then obtain

$$|B|_{E^{\mu-4}(\Omega_T)}=|(EF_p)|_{\theta\to(\theta_0,x_d/\varepsilon)}|_{E^{\mu-4}(\Omega_T)}=|(E(F-F_p))|_{\theta\to(\theta_0,x_d/\varepsilon)}|_{E^{\mu-4}(\Omega_T)}\leq K\delta. \quad (4\text{-}14)$$

Step 3. The $O(\varepsilon^2)$ terms in (4-9) consist of

$$|\varepsilon^2(L(\partial)\mathcal{U}^2_p(x,\theta_0,\xi_d))|_{\xi_d\to x_d/\varepsilon}|_{E^{\mu-4}(\Omega_T)}\leq\varepsilon^2C(\delta), \quad (4\text{-}15)$$

as well as terms coming from the Taylor expansion of $D(\varepsilon\mathcal{U}_\varepsilon)\mathcal{U}_\varepsilon$ like $(\varepsilon^2dD(0)\mathcal{V}^0\mathcal{V}^1)|_{\theta\to(\theta_0,x_d/\varepsilon)}$, all of which satisfy an estimate like (4-15). Setting $R_\varepsilon(x,\theta_0):=\mathbb{L}_\varepsilon(\mathcal{U}_\varepsilon)$, we have shown

$$|R_\varepsilon|_{E^{\mu-4}(\Omega_T)}\leq\varepsilon(K\delta+C(\delta)\varepsilon). \quad (4\text{-}16)$$

Step 4. The boundary profile equations (1-22) and the fact that the traces of $\mathcal{V}^0$ and $\mathcal{V}^1$ lie in $H^{\mu-1}(b\Omega_T)$ and $H^{\mu-2}(b\Omega_T)$, respectively, imply

$$\langle r_\varepsilon(x',\theta_0)\rangle_{H^{\mu-2}(b\Omega_T)}\leq C(\delta)\varepsilon^2, \quad \text{where } r_\varepsilon:=\psi(\varepsilon\mathcal{U}_\varepsilon)\mathcal{U}_\varepsilon-\varepsilon G(x',\theta_0). \quad (4\text{-}17)$$

Indeed, these $O(\varepsilon^2)$ terms include

$$\langle\varepsilon^2B\mathcal{U}^2_p(x',0,\theta_0,0)\rangle_{H^{\mu-2}(b\Omega_T)}\leq C(\delta)\varepsilon^2, \quad (4\text{-}18)$$

and other terms satisfying the same estimate coming from the Taylor expansion of $\psi(\varepsilon\mathcal{U}_\varepsilon)\mathcal{U}_\varepsilon$.

Step 5. Next we consider the singular problem satisfied by the difference $W_\varepsilon:=U_\varepsilon-\mathcal{U}_\varepsilon$:

$$\begin{gathered}\partial_dW_\varepsilon+\mathbb{A}\Big(\partial_{x'}+\frac{\beta\partial_{\theta_0}}{\varepsilon}\Big)W_\varepsilon+D_2(\varepsilon U_\varepsilon,\varepsilon\mathcal{U}_\varepsilon)W_\varepsilon=-R_\varepsilon,\\ \psi_2(\varepsilon U_\varepsilon,\varepsilon\mathcal{U}_\varepsilon)W_\varepsilon=-r_\varepsilon \quad \text{on } x_d=0,\\ W_\varepsilon=0 \quad \text{in } t<0,\end{gathered} \quad (4\text{-}19)$$

[33] Here $H^{\mu-3;2}_T$ denotes the space defined in (1-24), but with the obvious restriction on the domain of t.

where

$$D_2(\varepsilon U_\varepsilon, \varepsilon {}^{\mathcal{Q}}u_\varepsilon)W_\varepsilon := D(\varepsilon U_\varepsilon)U_\varepsilon - D(\varepsilon {}^{\mathcal{Q}}u_\varepsilon){}^{\mathcal{Q}}u_\varepsilon$$
$$= D(\varepsilon U_\varepsilon)W_\varepsilon + \left(\int_0^1 dD(\varepsilon {}^{\mathcal{Q}}u_\varepsilon + s\varepsilon(U_\varepsilon - {}^{\mathcal{Q}}u_\varepsilon))\,ds\right)(W_\varepsilon, \varepsilon {}^{\mathcal{Q}}u_\varepsilon), \tag{4-20}$$

and $\psi_2(\varepsilon U_\varepsilon, \varepsilon {}^{\mathcal{Q}}u_\varepsilon)W_\varepsilon$ is defined similarly. Since $U_\varepsilon \in E^{\mu-1}(\Omega_T)$ and ${}^{\mathcal{Q}}u_\varepsilon \in E^{\mu-3}(\Omega_T)$, a short computation shows

$$\psi_2(\varepsilon U_\varepsilon, \varepsilon {}^{\mathcal{Q}}u_\varepsilon)W_\varepsilon = \psi(\varepsilon U_\varepsilon)W_\varepsilon + d\psi(\varepsilon U)(W_\varepsilon, \varepsilon {}^{\mathcal{Q}}u_\varepsilon) + O(C(\delta)\varepsilon^2) = \mathcal{B}(\varepsilon U, \varepsilon {}^{\mathcal{Q}}u)W_\varepsilon + O(C(\delta)\varepsilon^2), \tag{4-21}$$

where the error term is measured in $H^{\mu-3}(b\Omega_T)$ and $\mathcal{B}$ is defined in (1-9). Similarly,

$$D_2(\varepsilon U_\varepsilon, \varepsilon {}^{\mathcal{Q}}u_\varepsilon)W_\varepsilon = \mathcal{D}(\varepsilon U, \varepsilon {}^{\mathcal{Q}}u)W_\varepsilon + O(C(\delta)\varepsilon^2) \quad \text{in } E^{\mu-3}(\Omega_T). \tag{4-22}$$

Thus, using (4-16) and (4-17), we find

$$\begin{gathered}\partial_d W_\varepsilon + \mathbb{A}\left(\partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right)W_\varepsilon + \mathcal{D}(\varepsilon U_\varepsilon, \varepsilon {}^{\mathcal{Q}}u_\varepsilon)W_\varepsilon = \varepsilon(K\delta + C(\delta)\varepsilon) \quad \text{in } E^{\mu-4}(\Omega_T),\\ \mathcal{B}(\varepsilon U_\varepsilon, \varepsilon {}^{\mathcal{Q}}u_\varepsilon)W_\varepsilon|_{x_d=0} = O(C(\delta)\varepsilon^2) \quad \text{in } H^{\mu-3}(b\Omega_T),\\ W_\varepsilon = 0 \quad \text{in } t < 0.\end{gathered} \tag{4-23}$$

Applying the estimate of Proposition 2.9, we obtain

$$|W_\varepsilon|_{E^0(\Omega_T)} \le K\delta + C(\delta)\varepsilon, \tag{4-24}$$

which implies

$$|U_\varepsilon - {}^{\mathcal{Q}}u^0_\varepsilon|_{E^0(\Omega_T)} \le K\delta + C(\delta)\varepsilon. \tag{4-25}$$

Fixing first δ small and then letting $\varepsilon \to 0$, we have shown

$$\lim_{\varepsilon\to 0} |U_\varepsilon - {}^{\mathcal{Q}}u^0_\varepsilon|_{E^0(\Omega_T)} = 0. \tag{4-26}$$

The family $U_\varepsilon - {}^{\mathcal{Q}}u^0_\varepsilon$, $0 < \varepsilon \le \varepsilon_0$, is bounded in $E^{\mu-2}(\Omega_T)$, so, by interpolation, (4-26) implies

$$\lim_{\varepsilon\to 0} |U_\varepsilon - {}^{\mathcal{Q}}u^0_\varepsilon|_{E^{\mu-3}(\Omega_T)} = 0,$$

as required. □

5. Nash–Moser schemes

5A. ***Iteration scheme for profiles.*** A good reference for the Nash–Moser scheme is [Alinhac and Gérard 2007]. The method depends on having a family of smoothing operators with the following properties. For $T > 0$, $s \ge 0$, and $\gamma \ge 1$, we let

$$F^s_\gamma(\Omega_T) := \{u \in H^s_\gamma(\Omega_T),\ u = 0 \text{ for } t < 0\}. \tag{5-1}$$

Lemma 5.1 [Alinhac 1989, Section 4]. *There exists a family of operators* $S_\theta : F^0_\gamma(\Omega_T) \to \bigcap_{\beta\geq 0} F^\beta_\gamma(\Omega_T)$ *such that*

$$\begin{aligned} &\text{(a)} \quad |S_\theta u|_\beta \leq C\theta^{(\beta-\alpha)_+}|u|_\alpha && \text{for } \alpha,\beta\geq 0,\\ &\text{(b)} \quad |S_\theta u - u|_\beta \leq C\theta^{(\beta-\alpha)}|u|_\alpha && \text{for } 0\leq\beta\leq\alpha,\\ &\text{(c)} \quad \Big|\frac{d}{d\theta}S_\theta u\Big|_\beta \leq C\theta^{(\beta-\alpha-1)}|u|_\alpha && \text{for } \alpha,\beta\geq 0. \end{aligned} \tag{5-2}$$

The constants are uniform for α, β *in a bounded interval.*

There is another family of operators $\tilde S_\theta$ *acting on functions defined on the boundary and satisfying the above properties with respect to the norms* $\langle u\rangle_s$ *on* $b\Omega_T$.[34]

Description of the scheme. Our goal is to solve problem (3-10):

$$\begin{aligned} &\mathcal{L}(V) = 0 && \text{in } \Omega_T,\\ &\mathcal{B}(V) = g && \text{in } b\Omega_T,\\ &V = 0 && \text{in } t\leq 0. \end{aligned} \tag{5-3}$$

The scheme starts with $V_0 = 0$. Assume that V_k are already given for $k = 1,\dots,n$ and satisfy $V_k = 0$ for $t < 0$. We define

$$V_{n+1} = V_n + \dot V_n, \tag{5-4}$$

where the increment $\dot V_n$ is specified below. Given $\theta_0 \geq 1$, we set $\theta_n := (\theta_0^2 + n)^{1/2}$ and work with the smoothing operators S_{θ_n}. We write the decomposition

$$\mathcal{L}(V_{n+1}) - \mathcal{L}(V_n) = \mathcal{L}'(V_n)\dot V_n + e'_n = \mathcal{L}'(S_{\theta_n}V_n)\dot V_n + e'_n + e''_n, \tag{5-5}$$

where e'_n denotes the usual "quadratic error" of Newton's scheme and e''_n the "substitution error". Similarly,

$$\begin{aligned} \mathcal{B}((V_{n+1})|_{x_d=0}) - \mathcal{B}((V_n)|_{x_d=0}) &= \mathcal{B}'((V_n)|_{x_d=0}))(\dot V_n|_{x_d=0})) + e'_n\\ &= \mathcal{B}'((S_{\theta_n}V_n)|_{x_d=0})(\dot V_n|_{x_d=0}) + \tilde e'_n + \tilde e''_n. \end{aligned} \tag{5-6}$$

The increment $\dot V_n$ is computed by solving the linearized problem

$$\begin{aligned} &\mathcal{L}'(S_{\theta_n}V_n)\dot V_n = f_n,\\ &\mathcal{B}'((S_{\theta_n}V_n)|_{x_d=0})(\dot V_n|_{x_d=0}) = g_n,\\ &\dot V_n = 0 \quad \text{in } t<0, \end{aligned} \tag{5-7}$$

where f_n and g_n are computed as we now describe.

We set $e_n := e'_n + e''_n$ and $\tilde e_n := \tilde e'_n + \tilde e''_n$. Given

$$\begin{gathered} V_0 := 0, \quad f_0 := 0, \quad g_0 := \tilde S_{\theta_0} g, \quad E_0 := 0, \quad \tilde E_0 := 0,\\ V_1,\dots,V_n, \quad f_1,\dots,f_{n-1}, \quad g_1,\dots,g_{n-1}, \quad e_0,\dots,e_{n-1}, \quad \tilde e_0,\dots,\tilde e_{n-1}, \end{gathered} \tag{5-8}$$

[34] For u defined on Ω_T, we do not necessarily have equality of $(S_\theta u)|_{x_d=0}$ and $\tilde S_\theta(u|_{x_d=0})$.

we first compute for $n \geq 1$ the accumulated errors

$$E_n := \sum_{k=0}^{n-1} e_k, \qquad \tilde{E}_n := \sum_{k=0}^{n-1} \tilde{e}_k. \tag{5-9}$$

We then compute f_n and g_n from the equations

$$\sum_{k=0}^{n} f_k + S_{\theta_n} E_n = 0, \qquad \sum_{k=0}^{n} g_k + \tilde{S}_{\theta_n} \tilde{E}_n = \tilde{S}_{\theta_n} g, \tag{5-10}$$

solve (5-7) for $\dot{V}_n$, and finally compute V_{n+1} from (5-4).

Next e_n and $\tilde{e}_n$ can be computed from [35]

$$\begin{aligned} &\mathscr{L}(V_{n+1}) - \mathscr{L}(V_n) = f_n + e_n, \\ &\mathscr{B}((V_{n+1})|_{x_d=0}) - \mathscr{B}((V_n)|_{x_d=0}) = g_n + \tilde{e}_n. \end{aligned} \tag{5-11}$$

Thus the order of construction is

$$\cdots \to (e_{n-1}, \tilde{e}_{n-1}) \to (E_n, \tilde{E}_n) \to (f_n, g_n) \to \dot{V}_n \to V_{n+1} \to (e_n, \tilde{e}_n) \to \cdots. \tag{5-12}$$

Adding (5-11) from 0 to n and using (5-10) gives

$$\begin{aligned} &\mathscr{L}(V_{n+1}) = (I - S_{\theta_n}) E_n + e_n, \\ &\mathscr{B}((V_{n+1})|_{x_d=0}) - g = (\tilde{S}_{\theta_n} - I) g + (I - \tilde{S}_{\theta_n}) \tilde{E}_n + \tilde{e}_n. \end{aligned} \tag{5-13}$$

Since $S_{\theta_n} \to I$ and $\tilde{S}_{\theta_n} \to I$ as $n \to \infty$ and we expect $(e_n, \tilde{e}_n) \to 0$, we formally obtain a solution of (5-3) in the limit as $n \to \infty$.

Induction assumption. Let $\Delta_n := \theta_{n+1} - \theta_n$ and observe that

$$\frac{1}{3\theta_n} \leq \Delta_n = \sqrt{\theta_n^2 + 1} - \theta_n \leq \frac{1}{2\theta_n} \quad \text{for all } n \in \mathbb{N}. \tag{5-14}$$

With $\mu_0 = [(d+1)/2]$ as in Proposition 3.6, we now set $\nu_0 := \mu_0 + 1$ and fix a choice of integers $\nu_0 < \nu < \tilde{\nu}$, whose values are explained below:

$$\nu := 2\nu_0 + 4 \quad \text{and} \quad \tilde{\nu} := 2\nu - \nu_0. \tag{5-15}$$

Given $\delta > 0$ our induction assumption is this:

($\mathbf{H}_{n-1}$) For all $k = 0, \ldots, n-1$ and for all $s \in [0, \tilde{\nu}] \cap \mathbb{N}$,

$$|\dot{V}_k|_s + \langle \dot{V}_k \rangle_s \leq \delta \theta_k^{s-\nu-1} \Delta_k. \tag{5-16}$$

The main step in the proof of Theorem 5.11 is to show that, for correctly chosen parameters $\delta > 0$ (small) and $\theta_0 \geq 1$ (large) and for small enough g, ($\mathbf{H}_{n-1}$) implies ($\mathbf{H}_n$). At the end we will verify that ($\mathbf{H}_0$) holds for small enough g.

First we state some easy consequences of ($\mathbf{H}_{n-1}$).

[35] In the estimates of e_n and $\tilde{e}_n$, we instead use (5-20), (5-21) and (5-24).

Lemma 5.2. *If θ_0 is large enough, then, for $k = 0, \dots, n$ and all integers $s \in [0, \tilde{\nu}]$, we have*

$$|V_k|_s + \langle V_k \rangle_s \leq \begin{cases} C\delta\theta_k^{(s-\nu)_+}, & \nu \neq s, \\ C\delta \log \theta_k, & \nu = s. \end{cases} \tag{5-17}$$

Proof. This follows by writing $V_k = V_0 + \sum_{j=0}^{k-1} \dot{V}_j$ and using the triangle inequality and an elementary comparison between Riemann sums and integrals. □

Lemma 5.3. *If θ_0 is large enough, then, for $k = 0, \dots, n$ and all integers $s \in [0, \tilde{\nu} + 2]$, we have*

$$|S_{\theta_k} V_k|_s \leq \begin{cases} C\delta\theta_k^{(s-\nu)_+}, & \nu \neq s, \\ C\delta \log \theta_k, & \nu = s. \end{cases} \tag{5-18}$$

Moreover, for $k = 0, \dots, n$ and all integers $s \in [0, \tilde{\nu}]$, we have

$$|(I - S_{\theta_k})V_k|_s \leq \begin{cases} C\delta\theta_k^{s-\nu} \log \theta_k, & s \leq \nu, \\ C\delta\theta_k^{s-\nu}, & s > \nu. \end{cases} \tag{5-19}$$

Proof. This follows from Lemma 5.2 and the properties of the S_θ. For example, we have

$$\begin{aligned} &|(I - S_{\theta_k})V_k|_s \leq 2|V_k|_s \leq C\delta\theta^{s-\nu} && \text{for } s > \nu, \\ &|(I - S_{\theta_k})V_k|_s \leq C\theta^{s-\nu}|V_k|_\nu \leq C\delta\theta^{s-\nu} \log \theta_k && \text{for } s \leq \nu. \end{aligned}$$ □

Estimate of the quadratic errors. From (5-5) and (5-6) we have

$$e_k' = \mathscr{L}(V_{k+1}) - \mathscr{L}(V_k) - \mathscr{L}'(V_k)\dot{V}_k = \int_0^1 (1-\tau)\mathscr{L}''(V_k + \tau\dot{V}_k)(\dot{V}_k, \dot{V}_k)\, d\tau, \tag{5-20}$$

$$\tilde{e}_k' = \mathscr{B}(V_{k+1}) - \mathscr{B}(V_k) - \mathscr{B}'(V_k)\dot{V}_k = \int_0^1 (1-\tau)\mathscr{B}''(V_k + \tau\dot{V}_k)(\dot{V}_k, \dot{V}_k) d\tau, \tag{5-21}$$

where the arguments in (5-21) are evaluated at $x_d = 0$.

Lemma 5.4. (1) *For large enough θ_0 we have, for all $k = 0, \dots, n-1$ and all integers $s \in [0, \tilde{\nu}]$,*

$$|e_k'|_s \leq C\delta^2\theta_k^{L_1(s)-1}\Delta_k, \tag{5-22}$$

where $L_1(s) = s + \nu_0 - 2\nu - 2$.

(2) *For large enough θ_0 we have for all $k = 0, \dots, n-1$ and all integers $s \in [0, \tilde{\nu} - 1]$*

$$\langle \tilde{e}_k' \rangle_s \leq C\delta^2\theta_k^{L_2(s)-1}\Delta_k, \tag{5-23}$$

where $L_2(s) = s + \nu_0 - 2\nu - 1$.

Proof. Using (5-20), Proposition 3.3, and the fact that $\nu_0 > \nu_1$, we have

$$|e_k'|_s \leq C|\dot{V}_k|_s|\dot{V}_k|_{\nu_0}.$$

The estimate (5-22) then follows by applying assumption (5-16) and using $\Delta_k \sim 1/\theta_k$. The estimate (5-23) is proved similarly; the restriction $s \in [0, \tilde{\nu} - 1]$ reflects the loss of one derivative in (3-15). □

Estimate of the substitution errors. From (5-5) and (5-6) we have

$$\begin{aligned} &\text{(a)} \quad e_k'' = \int_0^1 \mathscr{L}''(S_{\theta_k} V_k + \tau(V_k - S_{\theta_k} V_k))(\dot{V}_k, (I - S_{\theta_k}) V_k)\, d\tau, \\ &\text{(b)} \quad \tilde{e}_k'' = \int_0^1 \mathscr{B}''(S_{\theta_k} V_k + \tau(V_k - S_{\theta_k} V_k))(\dot{V}_k, (I - S_{\theta_k}) V_k)\, d\tau, \end{aligned} \tag{5-24}$$

where in (5-24)(b) we have, for example, written $S_{\theta_k} V_k$ for $(S_{\theta_k} V_k)|_{x_d=0}$.

Lemma 5.5. (1) *For large enough θ_0 we have, for all $k = 0, \dots, n-1$ and all integers $s \in [0, \tilde{\nu}]$,*

$$|e_k''|_s \le C\delta^2 \theta_k^{L_3(s)-1} \Delta_k, \tag{5-25}$$

where $L_3(s) = s + \nu_0 - 2\nu + 1$.

(2) *For large enough θ_0 we have, for all $k = 0, \dots, n-1$ and all integers $s \in [0, \tilde{\nu} - 2]$,*

$$\langle \tilde{e}_k'' \rangle_s \le C\delta^2 \theta_k^{L_4(s)-1} \Delta_k, \tag{5-26}$$

where $L_4(s) = s + \nu_0 - 2\nu + 3$.

Proof. Using (5-24)(a) and Proposition 3.3, we obtain

$$|e_k''|_s \le C(|\dot{V}_k|_s |(I - S_{\theta_k}) V_k|_{\nu_0} + |(I - S_{\theta_k}) V_k|_s |\dot{V}_k|_{\nu_0}). \tag{5-27}$$

The estimate (5-25) now follows from ($\mathbf{H}_{n-1}$) and Lemma 5.3. The estimate (5-26) is proved the same way, after using the trace estimate

$$\langle (I - S_{\theta_k}) V_k \rangle_{s+1} \le C |(I - S_{\theta_k}) V_k|_{s+2}. \tag{5-28}$$

The restriction $s \in [0, \tilde{\nu} - 2]$ reflects the subscript $s + 2$ in (5-28). □

Estimate of $(E_n, \tilde{E}_n)$ and (f_n, g_n). Since $e_k = e_k' + e_k''$ and $\tilde{e}_k = \tilde{e}_k' + \tilde{e}_k''$, we have:

Lemma 5.6. *There exists θ_0 sufficiently large so that*

$$|E_n|_{\tilde{\nu}} \le C\delta^2 \theta_n^{L_3(\tilde{\nu})} \quad \text{and} \quad \langle \tilde{E}_n \rangle_{\tilde{\nu}-2} \le C\delta^2 \theta_n^{L_4(\tilde{\nu}-2)}. \tag{5-29}$$

Proof. Viewing $E_n = \sum_{k=0}^{n-1} e_k$ as a Riemann sum and using $L_3(\tilde{\nu}) > 0$,[36] we obtain the estimate of E_n from (5-22) and (5-25). Since $L_4(\tilde{\nu} - 2) > 0$, the estimate of $\tilde{E}_n$ is similar. □

From (5-10) we have

$$\begin{aligned} f_n &= -(S_{\theta_n} - S_{\theta_{n-1}}) E_{n-1} - S_{\theta_n} e_{n-1}, \\ g_n &= (\tilde{S}_{\theta_n} - \tilde{S}_{\theta_{n-1}}) g - (\tilde{S}_{\theta_n} - \tilde{S}_{\theta_{n-1}}) \tilde{E}_{n-1} - \tilde{S}_{\theta_n} \tilde{e}_{n-1}. \end{aligned} \tag{5-30}$$

Lemma 5.7. *There exists θ_0 sufficiently large so that, for $s \in [0, \tilde{\nu} + 1]$, we have*

$$\begin{aligned} &\text{(a)} \quad |f_n|_s \le C\delta^2 \theta_n^{L_3(s)-1} \Delta_n, \\ &\text{(b)} \quad \langle g_n \rangle_s \le C\delta^2 \theta_n^{L_4(s)-1} \Delta_n + C\theta_n^{s-\nu-1} \langle g \rangle_\nu \Delta_n. \end{aligned} \tag{5-31}$$

[36] This determines $\tilde{\nu}$ in (5-15).

Proof. Using (5-2)(c), (5-29), and $s - \tilde{\nu} + L_3(\tilde{\nu}) = L_3(s)$, we find

$$|(S_{\theta_n} - S_{\theta_{n-1}})E_{n-1}|_s \le C \int_{\theta_{n-1}}^{\theta_n} \theta^{s-\tilde{\nu}-1} |E_{n-1}|_{\tilde{\nu}} \, d\theta \le C\delta^2 \theta_{n-1}^{L_3(s)-1} \Delta_n. \tag{5-32}$$

From (5-22), (5-25), and the properties of S_θ, we readily obtain

$$|S_{\theta_n} e_{n-1}|_s \le C\delta^2 \theta_n^{L_3(s)-1} \Delta_n, \tag{5-33}$$

and this gives (5-31)(a).

The first term on the right in (5-31)(b) arises similarly. With

$$\langle (\tilde{S}_{\theta_n} - \tilde{S}_{\theta_{n-1}})g \rangle_s \le C \int_{\theta_{n-1}}^{\theta_n} \theta^{s-\nu-1} \langle g \rangle_\nu \, d\theta \le C\theta_n^{s-\nu-1} \langle g \rangle_\nu \Delta_n, \tag{5-34}$$

we obtain (5-31)(b). □

Induction step. We claim that, for $\delta > 0$ sufficiently small, the estimate for the linearized system (3-39) applies to (5-7) and gives for $s \in [0, \tilde{\nu}]$

$$|\dot{V}_n|_s + \langle \dot{V}_n \rangle_s \le C[|f_n|_{s+1} + \langle g_n \rangle_s + (|f_n|_{\nu_0+1} + \langle g_n \rangle_{\nu_0})(|S_{\theta_n} V_n|_{s+1} + \langle S_{\theta_n} V_n \rangle_{s+1})]. \tag{5-35}$$

Indeed, (5-18) and $\nu > \nu_0 + 2$ imply that, for $\delta > 0$ small enough, the requirement (3-38) holds.[37] For the terms involving f_n and g_n, except $\langle g_n \rangle_{\nu_0}$, we substitute directly into (5-35) the corresponding estimates from Lemma 5.7. For $\langle g_n \rangle_{\nu_0}$ we have

$$\langle g_n \rangle_{\nu_0} \le C(\delta^2 \theta_n^{L_4(\nu_0)-1} \Delta_n + \theta_n^{-\nu-2} \langle g \rangle_{\nu_0+\nu+1} \Delta_n), \tag{5-36}$$

where the last term arises from (5-34) with $s = \nu_0$ and ν replaced by $\nu + \nu_0 + 1$. We also use

$$\langle S_{\theta_n} V_n \rangle_{s+1} \le |S_{\theta_n} V_n|_{s+2} \le C\delta \theta_n^{(s+2-\nu)_+ + 1}, \tag{5-37}$$

and a similar estimate for $|S_{\theta_n} V_n|_{s+1}$, which follow directly from (5-18).

Since $L_4(s) > L_3(s+1)$, this gives for $s \in [0, \tilde{\nu}]$

$$\begin{aligned} &|\dot{V}_n|_s + \langle \dot{V}_n \rangle_s \\ &\quad \le C[\delta^2 \theta_n^{L_4(s)-1} \Delta_n + \theta_n^{s-\nu-1} \langle g \rangle_\nu \Delta_n + (\delta^2 \theta_n^{L_4(\nu_0)-1} \Delta_n + \theta_n^{-\nu-2} \langle g \rangle_{\nu_0+\nu+1} \Delta_n) \delta \theta_n^{(s+2-\nu)_+ + 1}]. \end{aligned} \tag{5-38}$$

For $s \in [0, \tilde{\nu}]$ the parameters ν_0 and ν (recall (5-15)) satisfy

$$\begin{aligned} &L_4(s) \le s - \nu, \\ &L_4(\nu_0) + (s + 2 - \nu)_+ + 1 \le s - \nu, \\ &(s + 2 - \nu)_+ < s. \end{aligned} \tag{5-39}$$

Thus we have proved ($\mathbf{H}_n$), which is the content of the following lemma.

[37] We use a trace estimate like (5-37) here as well.

Lemma 5.8 ($\mathbf{H}_n$). *If $\delta > 0$, $\langle g\rangle_\nu/\delta$ are sufficiently small, and θ_0 sufficiently large, we have*

$$|\dot V_n|_s + \langle \dot V_n\rangle_s \le \delta\theta_n^{s-\nu-1}\Delta_n \quad \text{for all integers } s\in[0,\tilde\nu]. \tag{5-40}$$

Still assuming ($\mathbf{H}_{n-1}$), we now show the following.

Lemma 5.9. *Suppose $n\ge 1$. If $\delta>0$ is sufficiently small and θ_0 sufficiently large, we have*

$$\begin{aligned} &\text{(a)}\quad |\mathscr{L}(V_n)|_s \le \delta\theta_n^{s-\nu-1} && \text{for all integers } s\in[0,\tilde\nu],\\ &\text{(b)}\quad \langle \mathscr{B}(V_n)-g\rangle_s \le \delta\theta_n^{s-\nu-1} && \text{for all integers } s\in[0,\tilde\nu-2]. \end{aligned} \tag{5-41}$$

Proof. From (5-13) we have

$$\begin{aligned} &\text{(a)}\quad |\mathscr{L}(V_n)|_s \le |(I-S_{\theta_{n-1}})E_{n-1}|_s + |e_{n-1}|_s,\\ &\text{(b)}\quad \langle B(V_n)-g\rangle_s \le \langle(\tilde S_{\theta_{n-1}}-I)g\rangle_s + \langle(I-\tilde S_{\theta_{n-1}})\tilde E_{n-1}\rangle_s + \langle\tilde e_{n-1}\rangle_s. \end{aligned} \tag{5-42}$$

Using (5-2) and the above estimates of E_{n-1} and e_{n-1}, we find

$$\begin{aligned} |(I-S_{\theta_{n-1}})E_{n-1}|_s &\le C\theta_n^{s-\tilde\nu}|E_{n-1}|_{\tilde\nu} \le C\delta^2\theta_n^{(s-\nu-1)+(\nu_0+2-\nu)},\\ |e_{n-1}|_s &\le C\delta^2\theta_n^{L_3(s)-1}\Delta_n, \end{aligned} \tag{5-43}$$

which imply (5-41)(a) since $\nu_0+2-\nu<0$ and $L_3(s)<s-\nu$.

The last two terms on the right in (5-42)(b) are estimated similarly. To finish, we use

$$\langle(\tilde S_{\theta_{n-1}}-I)g\rangle_s \le C\theta_n^{s-(\tilde\nu-2)}\langle g\rangle_{\tilde\nu-2} \quad \text{for } s\le\tilde\nu-2 \tag{5-44}$$

and observe that $s-\tilde\nu+2<s-\nu-1$. □

We now fix δ and θ_0 as above and check ($\mathbf{H}_0$).

Lemma 5.10. *If $\langle g\rangle_\nu$ is small enough, ($\mathbf{H}_0$) holds.*

Proof. Applying the estimate for the linearized system to

$$\begin{aligned} \mathscr{L}'(0)\dot V_0 &= 0,\\ \mathscr{B}'(0)\dot V_0 &= S_{\theta_0}g, \end{aligned} \tag{5-45}$$

we obtain for integer $s\in[0,\tilde\nu]$

$$|\dot V_0|_s + \langle\dot V_0\rangle_s \le C\langle S_{\theta_0}g\rangle_s \le C\begin{cases}\theta_0^{s-\nu}\langle g\rangle_\nu, & s\ge\nu,\\ \langle g\rangle_\nu, & s<\nu.\end{cases} \tag{5-46}$$

Thus, ($\mathbf{H}_0$) holds if $\langle g\rangle_\nu$ is small enough. □

Proof of Proposition 3.1. We have

$$V_n = V_{n-1} + \dot V_{n-1} = \sum_{k=0}^{n-1}\dot V_k. \tag{5-47}$$

Let $\nu' := \nu-1$. Since $\theta_k\sim\sqrt{k}$ we have by ($\mathbf{H}_n$)

$$\sum_{k=0}^{\infty}|\dot V_k|_{\nu'} + \sum_{k=0}^{\infty}\langle\dot V_k\rangle_{\nu'} \le \delta\sum_k\theta_k^{-2}\Delta_k \le C\sum_k k^{-3/2} < \infty. \tag{5-48}$$

Thus, for some V as described in Proposition 3.1, $V_k \to V$ in $H^{\nu'}(\Omega_T)$ and $V_k|_{x_d=0} \to V|_{x_d=0}$ in $H^{\nu'}(\Omega_T)$. This implies

$$\mathscr{L}(V_k) \to \mathscr{L}(V) \text{ in } H^{\nu'-1}(\Omega_T) \quad \text{and} \quad \mathscr{B}(V_k|_{x_d=0}) \to \mathscr{B}(V|_{x_d=0}) \text{ in } H^{\nu'-1}(b\Omega_T).$$

Applying Lemma 5.9 with $s = \nu' - 1$, we conclude that V is a solution of the profile system (3-10). □

Having solved the key subsystem we can now easily complete the solution of the full profile system (1-35)–(1-36) and obtain the following result.

Theorem 5.11. *Fix $T > 0$, let $\nu_0 = [(d+1)/2] + 1$, $\nu = 2\nu_0 + 4$, and $\tilde{\nu} = 2\nu - \nu_0$, and suppose $G \in H^{\tilde{\nu}-1}(\Omega_T)$. If $\langle G \rangle_{\nu+1}$ is small enough, there exist solutions*

$$\mathscr{V}^0 = \mathscr{V}^0_{\mathrm{inc}} \in H^{\nu-1}(\Omega_T), \quad \mathscr{V}^1 = \underline{\mathscr{V}}^1 + \mathscr{V}^1_{\mathrm{inc}} + \mathscr{V}^1_{\mathrm{out}} \in H^{\nu-2}(\Omega_T)$$

of the full profile system (1-35)–(1-36) satisfying[38]

$$\begin{aligned} &\mathscr{V}^0 = E\mathscr{V}^0 \in H^{\nu-1}(\Omega_T), \quad \mathscr{V}^0_{\mathrm{inc}}|_{x_d=0,\theta_j=\theta_0} \in H^{\nu-1}(b\Omega_T), \\ &\mathscr{V}^1_{\mathrm{out}} = E\mathscr{V}^1_{\mathrm{out}} \in H^{\nu-1}(\Omega_T), \quad (E\mathscr{V}^1_{\mathrm{out}})|_{x_d=0,\theta_j=\theta_0} \in H^{\nu-1}(b\Omega_T), \\ &\underline{\mathscr{V}}^1 \in H^{\nu-2}(\Omega_T), \quad (I-E)\mathscr{V}^1_{\mathrm{inc}} \in H^{\nu-2}(\Omega_T), \quad E\mathscr{V}^1_{\mathrm{inc}} \in H^{\nu-2}(\Omega_T). \end{aligned} \tag{5-49}$$

These statements remain true if ν is increased and if $\tilde{\nu} \geq 2\nu - \nu_0$.

Proof. After the subsystem (1-42) is solved, we know $\mathscr{V}^0 = \mathscr{V}^0_{\mathrm{inc}} = E\mathscr{V}^0_{\mathrm{inc}}$, $\mathscr{V}^1_{\mathrm{out}} = E\mathscr{V}^1_{\mathrm{out}}$, and these functions have the regularity described in Proposition 3.1. Taking the mean of equations (1-36)(b)–(d), using the fact that the mean of the quadratic term in (1-36)(b) lies in $H^{\nu-1}(\Omega_T)$, and applying the result of [Coulombel 2005] to the resulting weakly stable system, we conclude $\underline{\mathscr{V}}^1 \in H^{\nu-2}(\Omega_T)$. From (1-36)(a) we find

$$(I-E)\mathscr{V}^1 = (I-E)\mathscr{V}^1_{\mathrm{inc}} \in H^{\nu-2}(\Omega_T). \tag{5-50}$$

It remains to determine $E\mathscr{V}^1_{\mathrm{inc}}$. Since the solvability condition (1-41) holds, we can make a choice of $E\mathscr{V}^1_{\mathrm{inc}}|_{x_d=0,\theta_j=\theta_0} \in H^{\nu-2}(b\Omega_T)$ satisfying the boundary equation (1-40), whose right side is now known and lies in $H^{\nu-2}(b\Omega_T)$.[39] Finally, we determine the components of $E\mathscr{V}_{\mathrm{inc}}$ by solving the transport equations determined by (1-36)(b), the choice of initial data, and the initial condition (1-36)(d). Observe that the interaction integrals corresponding to the quadratic term in (1-36)(b) lie in $H^{\nu-1}(\Omega_T)$. □

5B. ***Iteration scheme for the exact solution.*** The Nash–Moser scheme for the exact solution will use the scale of spaces $E^s_{\gamma,T}$ on Ω_T and $H^s_{\gamma,T}$ on $b\Omega_T$. Since T was fixed at the start and γ was fixed in Section 2C, we now drop these subscripts in the notation for norms and function spaces. For $s \geq 0$ we let

$$\mathbb{F}^s := \{u(x,\theta_0) \in E^s,\, u = 0 \text{ for } t < 0\}. \tag{5-51}$$

Moreover, we shall now denote E^s norms simply by $|U|_s$ and H^s norms by $\langle U \rangle_s$.

[38] Here when we write $\mathscr{V}^0_{\mathrm{inc}} \in H^{\nu-1}(\Omega_T)$, for example, we mean that the individual components of $\mathscr{V}^0_{\mathrm{inc}}$ lie in that space.

[39] All terms on the right in (1-40) lie in $H^{\nu-1}(b\Omega_T)$, except the term involving $L(\partial)$. That term is actually more regular than $H^{\nu-2}(b\Omega_T)$, but we do not wish to introduce more refined spaces to capture this.

Lemma 5.12. *There exists a family of operators* $S_\theta : \mathbb{F}^0 \to \bigcap_{\beta\ge 0} \mathbb{F}^\beta$ *such that*

$$\begin{aligned} &\text{(a)} \quad |S_\theta u|_\beta \le C\theta^{(\beta-\alpha)_+}|u|_\alpha && \text{for } \alpha,\beta\ge 0,\\ &\text{(b)} \quad |S_\theta u - u|_\beta \le C\theta^{(\beta-\alpha)}|u|_\alpha && \text{for } 0\le\beta\le\alpha,\\ &\text{(c)} \quad \left|\frac{d}{d\theta}S_\theta u\right|_\beta \le C\theta^{(\beta-\alpha-1)}|u|_\alpha && \text{for } \alpha,\beta\ge 0. \end{aligned} \tag{5-52}$$

The constants are uniform for α, β *in a bounded interval.*

There is a family of operators $\tilde S_\theta$ *acting on functions defined on the boundary and satisfying the above properties with respect to the norms* $\langle u\rangle_s$ *on* $b\Omega_T$, *and we have*

$$(S_\theta u)|_{x_d=0} = \tilde S_\theta(u|_{x_d=0}). \tag{5-53}$$

Proof. Let $\tilde S_\theta$ be a standard family of smoothing operators, for example, as in [Alinhac 1989], acting in the (x',θ_0) variables on the scale of spaces H^s. For $U\in E^s$ simply treat x_d as a parameter and define

$$S_\theta U = \tilde S_\theta U(\,\cdot\,, x_d, \cdot\,). \tag{5-54}$$

The properties (5-52) then follow immediately from the corresponding properties of the operators $\tilde S_\theta$. □

To avoid excessive repetition, we use the notation and arguments of Section 5A as much as possible, and just point out where changes are needed. Thus, we now denote the solution to the semilinear singular problem (1-18) by V instead of U, and rewrite (1-18) as

$$\begin{aligned} &\mathcal{L}(V) = 0 && \text{on } \Omega_T,\\ &\mathcal{B}(V) = G && \text{on } b\Omega_T,\\ &V = 0 && \text{in } T<0, \end{aligned} \tag{5-55}$$

where

$$\begin{aligned} \mathcal{L}(V) &:= \frac{1}{\varepsilon}\left(\partial_d V + \mathbb{A}\left(\partial_{x'} + \frac{\beta\partial_{\theta_0}}{\varepsilon}\right)V + D(\varepsilon V)V\right),\\ \mathcal{B}(V) &:= \frac{1}{\varepsilon}(\psi(\varepsilon V)V). \end{aligned} \tag{5-56}$$

We now let[40]

$$\mu_0 := [(d+1)/2], \quad \mu_1 := \mu_0 + M_0, \quad \mu := \max(2\mu_0+3, \mu_1+1) = \mu_1+1, \quad \tilde\mu := 2\mu - \mu_0. \tag{5-57}$$

We now state the main result of this section.

Theorem 5.13. *Fix* $T>0$, *define* μ_0, μ_1, μ, *and* $\tilde\mu$ *as in* (5-57), *and suppose* $G\in H^{\tilde\mu}$. *There exists* $\varepsilon_0>0$ *such that if* $\langle G\rangle_{\mu+2}$ *is small enough, there exists a solution* V *of the system* (5-55) *on* Ω_T *for* $0<\varepsilon\le\varepsilon_0$ *with* $V\in E^{\mu-1}$, $V|_{x_d=0}\in H^\mu$. *Thus* $U_\varepsilon = V$ *is a solution of the singular system* (1-18) *on* Ω_T *for* $0<\varepsilon\le\varepsilon_0$. *These statements remain true if* μ *is increased and if* $\tilde\mu\ge 2\mu-\mu_0$.

[40] The parameter $\tilde\mu$ is determined so that $L_2(\tilde\mu)>0$ for $L_2(s)$ as in Lemma 5.18. The definition of μ is chosen so that $\mu_1<\mu$ and the conditions (5-76) hold.

The linearized singular problem (2-40) is now written

$$\begin{aligned} &\mathscr{L}'(V)\dot V = f \quad \text{on } \Omega_T,\\ &\mathscr{B}'(V)\dot V = g \quad \text{on } b\Omega_T,\\ &\dot V = 0 \qquad\quad \text{in } t<0. \end{aligned} \tag{5-58}$$

With this notation the description of the scheme in Section 5A starting at line (5-3) applies here word for word down to line (5-14).

Remark 5.14. (a) In order to apply the tame estimate of Proposition 2.16 to the linearized system (5-58), by Sobolev embedding (Remark 2.14), it suffices to have

$$|\varepsilon\partial_d V|_{\mu_1-1} + |V|_{\mu_1} \le K' \text{ for } \varepsilon\in(0,1] \quad \text{and} \quad |V|_{\mu_0+2}\le\kappa_0, \tag{5-59}$$

for some constant K' depending on K and κ_0 as in Proposition 2.16. In fact, we use the slightly weaker (because we use E^s norms on the right) estimate for $s\in[0,\tilde\mu]$:

$$|\dot V|_s + \langle \dot V\rangle_{s+1} \le C[|f|_{s+1} + \langle g\rangle_{s+2} + (|f|_{\mu_0+1} + \langle g\rangle_{\mu_0+2})(|U|_{s+1} + \langle U\rangle_{s+2})]. \tag{5-60}$$

(b) By Proposition 2.15, when $|V|_{\mu_0}\le K'$, the tame estimates for second derivatives now take the form

$$\begin{aligned} &\text{(a)}\quad |\mathscr{L}''(V)(\dot V^a,\dot V^b)|_s \le C(|\dot V^a|_s|\dot V^b|_{\mu_0} + |\dot V^b|_s|\dot V^a|_{\mu_0} + \varepsilon|V|_s|\dot V^a|_{\mu_0}|\dot V^b|_{\mu_0}),\\ &\text{(b)}\quad \langle\mathscr{B}''(V)(\dot V^a,\dot V^b)\rangle_s \le C(\langle\dot V^a\rangle_s\langle\dot V^b\rangle_{\mu_0} + \langle\dot V^b\rangle_s\langle\dot V^a\rangle_{\mu_0} + \varepsilon\langle V\rangle_s\langle\dot V^a\rangle_{\mu_0}\langle\dot V^b\rangle_{\mu_0}). \end{aligned} \tag{5-61}$$

With μ and $\tilde\mu$ redefined as in (5-57), for a given $\delta>0$, the induction hypothesis $(\mathbf{H}_{n-1})$ is now as follows.

$(\mathbf{H}_{n-1})$ For all $k=0,\dots,n-1$ and for all $s\in[0,\tilde\mu]\cap\mathbb{N}$

$$|\dot V_k|_s + \langle\dot V_k\rangle_{s+1} \le \delta\theta_k^{s-\mu-1}\Delta_k. \tag{5-62}$$

Lemmas 5.2 and 5.3 are now replaced, with no real change in the proofs, by the following two lemmas.

Lemma 5.15. *If θ_0 is large enough, then, for $k=0,\dots,n$ and all integers $s\in[0,\tilde\mu]$, we have*

$$|V_k|_s + \langle V_k\rangle_{s+1} \le \begin{cases} C\delta\theta_k^{(s-\mu)_+}, & \mu\ne s,\\ C\delta\log\theta_k, & \mu=s. \end{cases} \tag{5-63}$$

Lemma 5.16. *If θ_0 is large enough, then, for $k=0,\dots,n$ and all integers $s\in[0,\tilde\mu+2]$, we have*

$$|S_{\theta_k}V_k|_s + \langle S_{\theta_k}V_k\rangle_{s+1} \le \begin{cases} C\delta\theta_k^{(s-\mu)_+}, & \mu\ne s,\\ C\delta\log\theta_k, & \mu=s. \end{cases} \tag{5-64}$$

For $k=0,\dots,n$ and all integers $s\in[0,\tilde\mu]$, we have

$$|(I-S_{\theta_k})V_k|_s + \langle(I-S_{\theta_k})V_k\rangle_{s+1} \le \begin{cases} C\delta\theta_k^{s-\mu}\log\theta_k, & s\le\mu,\\ C\delta\theta_k^{s-\mu}, & s>\mu. \end{cases} \tag{5-65}$$

We have used (5-54) for the estimate on traces in Lemma 5.16. In place of Lemma 5.4 we now have:

Lemma 5.17. (1) *For large enough θ_0 and small enough δ we have, for all $k = 0, \dots, n-1$ and all integers $s \in [0, \tilde{\mu}]$,*

$$|e_k'|_s \le C\delta^2 \theta_k^{L_1(s)-1} \Delta_k, \tag{5-66}$$

where $L_1(s) := \max(s + \mu_0 - 2\mu - 2, (s-\mu)_+ + 2\mu_0 - 2\mu - 1)$.

(2) *For large enough θ_0 and small enough δ we have, for all $k = 0, \dots, n-1$ and all integers $s \in [0, \tilde{\mu}]$,*

$$\langle \tilde{e}_k' \rangle_{s+1} \le C\delta^2 \theta_k^{L_1(s)-1} \Delta_k. \tag{5-67}$$

Proof. Again we use (5-20) and (5-21). By Lemma 5.15 and ($\mathbf{H}_{n-1}$), we see that, for δ small enough, $|V_k + \tau \dot{V}_k|_{\mu_0} \le K'$, so we can apply the estimates (5-61). The new definition of $L_1(s)$ reflects the third term on the right in the estimates (5-61). □

In place of Lemma 5.5, the estimate of substitution errors, we now have:

Lemma 5.18. (1) *For large enough θ_0 and small enough δ we have, for all $k = 0, \dots, n-1$ and all integers $s \in [0, \tilde{\mu}]$,*

$$|e_k''|_s \le C\delta^2 \theta_k^{L_2(s)-1} \Delta_k, \tag{5-68}$$

where $L_2(s) := \max(s + \mu_0 - 2\mu + 1, (s-\mu)_+ + 2\mu_0 - 2\mu + 2)$.

(2) *For large enough θ_0 and small enough δ we have, for all $k = 0, \dots, n-1$ and all integers $s \in [0, \tilde{\mu}]$,*

$$\langle \tilde{e}_k'' \rangle_{s+1} \le C\delta^2 \theta_k^{L_2(s)-1} \Delta_k. \tag{5-69}$$

Proof. Again we use the formulas (5-24). By Lemma 5.3 we have $|S_{\theta_k} V_k + \tau(I - S_{\theta_k}) V_k|_{\mu_0} \le K'$ for δ small enough, so we can apply the estimates (5-61). When estimating the right sides of (5-61), we use, for example,

$$|(I - S_{\theta_k}) V_k|_s \le C\delta \theta_k^{s-\mu+1}. \qquad \square$$

In place of Lemma 5.6, the estimate of accumulated errors, we now have:

Lemma 5.19. *There exist θ_0 sufficiently large and δ_0 sufficiently small so that, for $0 < \delta \le \delta_0$,*

$$|E_n|_{\tilde{\mu}} \le C\delta^2 \theta_n^{L_2(\tilde{\mu})} \quad \text{and} \quad \langle \tilde{E}_n \rangle_{\tilde{\mu}+1} \le C\delta^2 \theta_n^{L_2(\tilde{\mu})}. \tag{5-70}$$

Proof. Since $\tilde{\mu} = 2\mu - \mu_0$, we have $L_2(\tilde{\mu}) > 0$, so the proof is the same as that of Lemma 5.6. □

The new version of Lemma 5.7, the estimate of f_n and g_n, is this:

Lemma 5.20. *There exist θ_0 sufficiently large and δ_0 sufficiently small so that, for $s \in [0, \tilde{\mu}+1]$, $0 < \delta \le \delta_0$, we have*

$$\begin{aligned} &\text{(a)} \quad |f_n|_s \le C\delta^2 \theta_n^{L_2(s)-1} \Delta_n, \\ &\text{(b)} \quad \langle g_n \rangle_{s+1} \le C\delta^2 \theta_n^{L_2(s)-1} \Delta_n + C\theta_n^{s-\mu-2} \langle G \rangle_{\mu+2} \Delta_n. \end{aligned} \tag{5-71}$$

Proof. Since $s - \tilde{\mu} + L_2(\tilde{\mu}) \le L_2(s)$, the proof of Lemma 5.7 can be repeated here. □

Induction step. For $\delta > 0$ sufficiently small, the estimate (5-60) for the linearized system applies to (5-7) and gives for $s \in [0, \tilde{\mu}]$

$$|\dot{V}_n|_s + \langle \dot{V}_n \rangle_{s+1} \le C[|f_n|_{s+1} + \langle g_n \rangle_{s+2} + (|f_n|_{\mu_0+1} + \langle g_n \rangle_{\mu_0+2})(|S_{\theta_n} V_n|_{s+1} + \langle S_{\theta_n} V_n \rangle_{s+2})]. \tag{5-72}$$

Indeed, (5-64) implies that for $\delta > 0$ small enough, $S_{\theta_n} V_n$ satisfies the requirement (5-59).[41] For the terms involving f_n and g_n, except $\langle g_n \rangle_{\mu_0+2}$, we substitute directly into (5-72) the corresponding estimates from Lemma 5.20. For $\langle g_n \rangle_{\mu_0+2}$, we have

$$\langle g_n \rangle_{\mu_0+2} \le C(\delta^2 \theta_n^{L_2(\mu_0+1)-1} \Delta_n + \theta_n^{-\mu-2} \langle G \rangle_{\mu_0+\mu+3} \Delta_n), \tag{5-73}$$

where the last term arises from an estimate like (5-34) with $s = \mu_0 + 2$ and μ replaced by $\mu + \mu_0 + 3$. We also use

$$\langle S_{\theta_n} V_n \rangle_{s+2} \le C\delta \theta_n^{(s+1-\mu)_+ + 1} \tag{5-74}$$

and a similar estimate for $|S_{\theta_n} V_n|_{s+1}$, which follow directly from (5-64).

Making these substitutions in (5-72) gives, for $s \in [0, \tilde{\mu}]$,

$$\begin{aligned} |\dot{V}_n|_s + \langle \dot{V}_n \rangle_{s+1} \le C[&\delta^2 \theta_n^{L_2(s+1)-1} \Delta_n + \theta_n^{s-\mu-1} \langle G \rangle_{\mu+2} \Delta_n \\ &+ (\delta^2 \theta_n^{L_2(\mu_0+1)-1} \Delta_n + \theta_n^{-\mu-2} \langle G \rangle_{\mu_0+\mu+3} \Delta_n) \delta \theta_n^{(s+1-\mu)_+ + 1}]. \end{aligned} \tag{5-75}$$

For $s \in [0, \tilde{\mu}]$ the parameters μ_0 and μ (recall (5-57)) satisfy

$$\begin{aligned} &\text{(a)} \quad L_2(s+1) \le s - \mu, \\ &\text{(b)} \quad L_2(\mu_0 + 1) + (s + 1 - \mu)_+ + 1 \le s - \mu, \\ &\text{(c)} \quad (s + 1 - \mu)_+ < s. \end{aligned} \tag{5-76}$$

Thus, we have proved $(\mathbf{H}_n)$, which is the content of the following lemma.

Lemma 5.21 $(\mathbf{H}_n)$. *If $\delta > 0$ and $\langle G \rangle_{\mu+2}/\delta$ are sufficiently small and θ_0 is sufficiently large, we have*

$$|\dot{V}_n|_s + \langle \dot{V}_n \rangle_{s+1} \le \delta \theta_n^{s-\mu-1} \Delta_n \quad \textit{for all integers } s \in [0, \tilde{\mu}]. \tag{5-77}$$

Still assuming $(\mathbf{H}_{n-1})$ we now show:

Lemma 5.22. *Suppose $n \ge 1$. If $\delta > 0$ is sufficiently small and θ_0 sufficiently large, we have*

$$\begin{aligned} &\text{(a)} \quad |\mathcal{L}(V_n)|_s \le \delta \theta_n^{s-\mu-1}, \\ &\text{(b)} \quad \langle \mathcal{B}(V_n) - G \rangle_{s+1} \le \delta \theta_n^{s-\mu-1}, \end{aligned} \tag{5-78}$$

for all integers $s \in [0, \tilde{\mu}]$.

Proof. From (5-13) we have

$$\begin{aligned} &\text{(a)} \quad |\mathcal{L}(V_n)|_s \le |(I - S_{\theta_{n-1}}) E_{n-1}|_s + |e_{n-1}|_s, \\ &\text{(b)} \quad \langle B(V_n) - G \rangle_{s+1} \le \langle (\tilde{S}_{\theta_{n-1}} - I) G \rangle_{s+1} + \langle (I - \tilde{S}_{\theta_{n-1}}) \tilde{E}_{n-1} \rangle_{s+1} + \langle \tilde{e}_{n-1} \rangle_{s+1}. \end{aligned} \tag{5-79}$$

Using (5-52) and the above estimates of E_{n-1} and e_{n-1}, we find

$$\begin{aligned} |(I - S_{\theta_{n-1}}) E_{n-1}|_s &\le C \theta_n^{s-\tilde{\mu}} |E_{n-1}|_{\tilde{\mu}} \le C\delta^2 \theta_n^{(s-\mu-1)+(\mu_0-\mu)}, \\ |e_{n-1}|_s &\le C\delta^2 \theta_n^{L_2(s)-1} \Delta_n, \end{aligned} \tag{5-80}$$

which imply (5-78)(a) since $\mu_0 - \mu < 0$ and $L_2(s) < s - \mu$.

[41] Here we use $\mu_1 < \mu$. Also, the term $|\varepsilon \partial_d (S_{\theta_n} V_n)|_{\mu_1 - 1}$ is estimated using (5-7).

The last two terms on the right in (5-79)(b) are estimated similarly. To finish, we use

$$\langle(\tilde{S}_{\theta_{n-1}} - I)G\rangle_{s+1} \le C\theta_n^{s+1-\tilde{\mu}}\langle G\rangle_{\tilde{\mu}} \quad \text{for } s \le \tilde{\mu} - 1, \tag{5-81}$$

and observe that $s - \tilde{\mu} + 1 < s - \mu - 1$. □

We now fix δ and θ_0 as above and check $(\mathbf{H}_0)$.

Lemma 5.23. *If $\langle G\rangle_{\mu+2}$ is small enough, $(\mathbf{H}_0)$ holds.*

Proof. Applying the estimate for the linearized system to

$$\begin{aligned} \mathscr{L}'(0)\dot{V}_0 &= 0, \\ \mathscr{B}'(0)\dot{V}_0 &= S_{\theta_0}G, \end{aligned} \tag{5-82}$$

we obtain, for integers $s \in [0, \tilde{\mu}]$,

$$|\dot{V}_0|_s + \langle\dot{V}_0\rangle_{s+1} \le C\langle S_{\theta_0}G\rangle_{s+2} \le C\begin{cases}\theta_0^{s-\mu}\langle G\rangle_{\mu+2}, & s \ge \mu, \\ \langle G\rangle_{\mu+2}, & s < \mu.\end{cases} \tag{5-83}$$

Thus $(\mathbf{H}_0)$ holds if $\langle G\rangle_{\mu+2}$ is small enough. □

Proof of Theorem 5.13. We have

$$V_n = V_{n-1} + \dot{V}_{n-1}\sum_{k=0}^{n-1}\dot{V}_k.$$

Let $\nu := \mu - 1$. Since $\theta_k \sim \sqrt{k}$ we have by $(\mathbf{H}_n)$

$$\sum_{k=0}^{\infty}|\dot{V}_k|_\nu + \sum_{k=0}^{\infty}\langle\dot{V}_k\rangle_{\nu+1} \le \delta\sum_k\theta_k^{-2}\Delta_k \le C\sum_k k^{-3/2} < \infty.$$

Thus, for some V as described in Theorem 5.13, $V_k \to V$ in E^ν and $V_k|_{x_d=0} \to V|_{x_d=0}$ in $H^{\nu+1}$ (in fact, uniformly for $0 < \varepsilon \le \varepsilon_0$). Lemma 5.22 applied with $s = \nu - 1$ now implies that V is a solution of the semilinear system (5-55). □

Appendix A: A calculus of singular pseudodifferential operators

Here we summarize the parts of the singular calculus constructed in [Coulombel et al. 2012] that are needed in this article.

Symbols. Our singular symbols are built from the following sets of classical symbols.

Definition A.1. Let $\mathbb{O} \subset \mathbb{R}^N$ be an open subset that contains the origin. For $m \in \mathbb{R}$ we let $\boldsymbol{S}^m(\mathbb{O})$ denote the class of all functions $\sigma : \mathbb{O}\times\mathbb{R}^d\times[1,\infty) \to \mathbb{C}^{N\times N}$, $N \ge 1$, such that σ is C^∞ on $\mathbb{O}\times\mathbb{R}^d$ and, for all compact sets $K \subset \mathbb{O}$,

$$\sup_{v\in K}\sup_{\xi'\in\mathbb{R}^d}\sup_{\gamma\ge1}(\gamma^2 + |\xi'|^2)^{-(m-|\nu|)/2}|\partial_v^\alpha\partial_{\xi'}^\nu\sigma(v,\xi',\gamma)| \le C_{\alpha,\nu,K}.$$

Let $\mathscr{C}_b^k(\mathbb{R}^d\times\mathbb{T})$, $k \in \mathbb{N}$, denote the space of continuous and bounded functions on $\mathbb{R}^d\times\mathbb{R}$ that are 2π-periodic in their last argument, and whose derivatives up to order k are continuous and bounded.

Definition A.2 (singular symbols). Let $m \in \mathbb{R}$, $n \in \mathbb{N}$, and fix $\beta \in \mathbb{R}^d \setminus 0$. We let S^m_n denote the family of functions $(a_{\varepsilon,\gamma})_{\varepsilon\in(0,1],\gamma\geq 1}$ that are constructed as follows:

$$\text{for all } (x', \theta_0, \xi', k) \in \mathbb{R}^d \times \mathbb{T} \times \mathbb{R}^d \times \mathbb{Z}, \quad a_{\varepsilon,\gamma}(x', \theta_0, \xi', k) := \sigma\left(\varepsilon V(x', \theta_0), \xi' + \frac{k\beta}{\varepsilon}, \gamma\right), \tag{A-1}$$

where $\sigma \in \boldsymbol{S}^m(\mathbb{O})$ and $V \in \mathscr{C}^n_b(\mathbb{R}^d \times \mathbb{T})$. Below and in the main text we often set

$$X := \xi' + \frac{k\beta}{\varepsilon}.$$

All results below extend to the case where in place of a function V that is independent of ε, the representation (A-1) is considered with a function V_ε that is indexed by ε, provided that we assume that all functions V_ε take values in a *fixed* convex compact subset K of $\mathbb{O}$ that contains the origin, and $(V_\varepsilon)_{\varepsilon\in(0,1]}$ is a bounded family of $\mathscr{C}^n_b(\mathbb{R}^d \times \mathbb{T})$. Such singular symbols with a function V_ε are exactly the kind of symbols that we manipulated in the construction of exact solutions to the singular system (1-18).

Singular pseudodifferential operators. To each symbol $a_{\varepsilon,\gamma}$ as in (A-1), we associate a singular pseudodifferential operator $\mathrm{Op}^{\varepsilon,\gamma}(a)$ whose action on Schwartz class functions $u \in \mathscr{S}(\mathbb{R}^d \times \mathbb{T} : \mathbb{C}^N)$ is defined by

$$\mathrm{Op}^{\varepsilon,\gamma}(a)u(x', \theta_0) := \frac{1}{(2\pi)^d} \sum_{k\in\mathbb{Z}} \int_{\mathbb{R}^d} e^{ix'\cdot\xi' + ik\theta_0} \sigma\left(\varepsilon V(x', \theta_0), \xi' + \frac{k\beta}{\varepsilon}, \gamma\right) \hat{u}(\xi', k)\, d\xi', \tag{A-2}$$

where $\hat{u}(\xi', k)$ denotes the Fourier transform at ξ' of the k-th Fourier coefficient of u with respect to θ_0. When $a_{\varepsilon,\gamma}$ is defined as in (A-1), below and in the main text of the article, we often write $\sigma(\varepsilon V(x, \theta_0), X, \gamma)$ in place of $a_{\varepsilon,\gamma}(x', \theta_0, \xi', k)$, and σ_D in place of $\mathrm{Op}^{\varepsilon,\gamma}(a)$. In particular, we let Λ_D denote the singular Fourier multiplier associated to the function

$$\Lambda(X, \gamma) := (\gamma^2 + |X|^2)^{1/2}.$$

When $V(x', x_d, \theta_0)$ depends also on a normal variable $x_d \geq 0$, we define the associated family of operators depending on the parameter x_d in the obvious way. The pseudodifferential calculus takes place only in the tangential directions (x', θ_0). To discuss mapping properties, we first define "singular" Sobolev spaces as follows.

Definition A.3. We let

$$H^{s,\varepsilon}(\mathbb{R}^d \times \mathbb{T}) := \left\{ u \in \mathscr{S}'(\mathbb{R}^d \times \mathbb{T}) : \sum_{k\in\mathbb{Z}} \int_{\mathbb{R}^d} (1 + |X|^2)^s |\hat{u}(\xi', k)|^2 \mathrm{d}\xi' < \infty \right\}.$$

This space is equipped with the family of norms[42]

$$|u|^2_{H^{s,\varepsilon},\gamma} := \frac{1}{(2\pi)^d} \sum_{k\in\mathbb{Z}} \int_{\mathbb{R}^d} (\gamma^2 + |X|^2)^s |\hat{u}(\xi', k)|^2\, d\xi'.$$

[42] In this appendix we use $|\cdot|$ instead of $\langle\cdot\rangle$ in the notation for norms on $\mathbb{R}^d \times \mathbb{T}$, but otherwise we retain notation from the main text.

Observe that, for s fixed, the space $H^{s,\varepsilon}$ depends on ε with no obvious inclusion if $\varepsilon_1 < \varepsilon_2$. However, for fixed $\varepsilon > 0$, the norms $|\cdot|_{H^{s,\varepsilon},\gamma_1}$ and $|\cdot|_{H^{s,\varepsilon},\gamma_2}$ are equivalent.

The next proposition describes some of the mapping properties of these operators. Detailed proofs can be found in [Coulombel et al. 2012]. The constant C is always independent of $\varepsilon \in (0, 1]$ and $\gamma \geq 1$, and we denote the $L^2(\mathbb{R}^d \times \mathbb{T})$ norm by $|\cdot|_0$ (which corresponds to $s = 0$ in Definition A.3).

Proposition A.4 (mapping properties). (a) *Suppose $\sigma(\varepsilon V(x, \theta_0), X, \gamma) \in S^m_n$, where $n \geq d+1$ and $m \leq 0$. Then $\sigma_D : L^2(\mathbb{R}^d \times \mathbb{T}) \to L^2(\mathbb{R}^d \times \mathbb{T})$ and*

$$|\sigma_D u|_0 \leq \frac{C}{\gamma^{|m|}} |u|_0.$$

(b) *Suppose $\sigma(\varepsilon V(x, \theta_0), X, \gamma) \in S^m_n$, where $n \geq d+1$ and $m > 0$. Then $\sigma_D : H^{m,\varepsilon}(\mathbb{R}^d \times \mathbb{T}) \to L^2(\mathbb{R}^d \times \mathbb{T})$ and*

$$|\sigma_D u|_0 \leq C |u|_{H^{m,\varepsilon},\gamma}.$$

(c) (*Smoothing property*) *Suppose $\sigma(\varepsilon V(x, \theta_0), X, \gamma) \in S^{-1}_n$, where $n \geq d + 2$. Then*

$$\sigma_D : L^2(\mathbb{R}^d \times \mathbb{T}) \to H^{1,\varepsilon}(\mathbb{R}^d \times \mathbb{T})$$

and

$$|\sigma_D u|_{H^{1,\varepsilon},\gamma} \leq C |u|_0.$$

(d) *Suppose $\sigma(\varepsilon V(x, \theta_0), X, \gamma) \in S^0_n$, where $n \geq d + 2$. Then $\sigma_D : H^{1,\varepsilon}(\mathbb{R}^d \times \mathbb{T}) \to H^{1,\varepsilon}(\mathbb{R}^d \times \mathbb{T})$ and*

$$|\sigma_D u|_{H^{1,\varepsilon},\gamma} \leq C |u|_{H^{1,\varepsilon},\gamma}.$$

Residual operators. We sometimes denote by $r_{0,D}$ an operator that maps $L^2(\mathbb{R}^d \times \mathbb{T}) \to L^2(\mathbb{R}^d \times \mathbb{T})$ and satisfies a uniform operator bound

$$|r_{0,D} u|_0 \leq C |u|_0,$$

even when $r_{0,D}$ is not necessarily defined by a symbol in some class S^0_n. Similarly, we sometimes let $r_{-1,D}$ denote an operator not necessarily associated to a symbol in S^{-1}_n such that

$$|r_{-1,D} u|_{H^{1,\varepsilon},\gamma} \leq C |u|_0. \tag{A-3}$$

For example, the composition $\sigma_{-1,D} \tau_{0,D} = r_{-1,D}$ of an operator of order -1 (case (c) in Proposition A.4) with an operator of order 0 (case (a) when $m = 0$) is of this latter type.

Remark A.5. Observe that a composition of the form $r_{0,D} r_{-1,D}$ is not necessarily an operator of type $r_{-1,D}$, a fact that is a source of difficulty in the proof of the main linear estimate, Proposition 2.2. This is the case, for example, if $r_{0,D}$ is the operator of multiplication by $V(x', \theta_0) \in \mathscr{C}^1_b(\mathbb{R}^d \times \mathbb{T})$. On the other hand we have

$$\varepsilon V(x', \theta_0) r_{-1,D} = r_{-1,D},$$

and, more generally, Proposition A.4(d) implies that if $\sigma \in S^0_n$, $n \geq d + 2$, we have

$$\sigma_D r_{-1,D} = r_{-1,D}.$$

Adjoints and products. In spite of the fact that singular symbols and their derivatives fail to decay in the classical way in $\langle \xi', k, \gamma \rangle$, it is possible to construct a crude calculus of singular pseudodifferential operators with useful formulas for adjoints and products, which, in particular, permit Gårding inequalities to be proved. This calculus was used repeatedly in the proof of the main linear estimate, Proposition 2.2. Detailed proofs can be found in [Coulombel et al. 2012].

In the next proposition, σ^* denotes the conjugate transpose of the $N \times N$ matrix valued symbol σ, while $(\sigma_D)^*$ denotes the adjoint operator for the L^2 scalar product.

Proposition A.6 (adjoints). (a) *Let $\sigma \in S^0_n$, where $n \geq 2d+3$. Then $(\sigma_D)^* - (\sigma^*)_D = r_{-1,D}$.*

(b) *Let $\sigma \in S^1_n$, where $n \geq 3d+4$. Then $(\sigma_D)^* - (\sigma^*)_D = r_{0,D}$.*

Proposition A.7 (products). (a) *Suppose σ and τ lie in S^0_n, where $n \geq 2d+3$. Then*

$$\sigma_D \tau_D - (\sigma\tau)_D = r_{-1,D}.$$

(b) *Suppose $\sigma \in S^1_n$, $\tau \in S^0_n$ or $\sigma \in S^0_n$, $\tau \in S^1_n$, where $n \geq 3d+4$. Then*

$$\sigma_D \tau_D - (\sigma\tau)_D = r_{0,D}.$$

(c) *Suppose $\sigma \in S^{-1}_n$, $\tau \in S^1_n$, where $n \geq 3d+4$. Then*

$$\sigma_D \tau_D - (\sigma\tau)_D = r_{-1,D}. \tag{A-4}$$

Remark A.8. Observe that when $\tau = \tau(X, \gamma)$ is independent of $\varepsilon V(x, \theta_0)$ and thus gives rise to a Fourier multiplier, the composition $\sigma_D \tau_D = (\sigma\tau)_D$ is exact, a fact that has been used several times in the proof of Proposition 2.2.

The equality (A-4) does not hold in general when $\sigma \in S^1_n$ and $\tau \in S^{-1}_n$, and this is one of the main difficulties in the proof of Proposition 2.4.

In the proof of Proposition 2.2 we use the following localized Gårding inequality for zero-order operators. As before, we write $\zeta = (\xi', \gamma)$.

Proposition A.9 (Gårding inequality). *Let $\sigma(v, \zeta) \in \boldsymbol{S}^0(\mathbb{O})$ and $\chi(v, \zeta) \in \boldsymbol{S}^0(\mathbb{O})$ be such that*

$$\operatorname{Re} \sigma(v, \zeta) \geq c > 0$$

on a conic neighborhood of $\operatorname{supp} \chi$. *Provided the corresponding singular symbols lie in S^0_n, $n \geq 2d+2$, we have*

$$\operatorname{Re}(\sigma_D \chi_D u, \chi_D u) \geq \frac{c}{2} |\chi_D u|_0^2 - \frac{C}{\gamma} |u|_0^2.$$

Extended calculus. In the proof of Corollary 2.3 we use a slight extension of the singular calculus. For given parameters $0 < \delta_1 < \delta_2 < 1$, we choose a cutoff $\chi^e(\xi', k\beta/\varepsilon, \gamma)$ such that

$$0 \leq \chi^e \leq 1, \quad \chi^e\left(\xi', \frac{k\beta}{\varepsilon}, \gamma\right) = 1 \text{ on } \left\{(\gamma^2 + |\xi'|^2)^{1/2} \leq \delta_1 \left|\frac{k\beta}{\varepsilon}\right|\right\}, \quad \operatorname{supp} \chi^e \subset \left\{(\gamma^2 + |\xi'|^2)^{1/2} \leq \delta_2 \left|\frac{k\beta}{\varepsilon}\right|\right\},$$

and define a corresponding Fourier multiplier χ_D in the extended calculus by the formula (A-2) with $\chi^e(\xi', k\beta/\varepsilon, \gamma)$ in place of $\sigma(\varepsilon V, X, \gamma)$. Composition laws involving such operators are proved in

[Coulombel et al. 2012], but here we need only the fact that part (a) of Proposition A.7 holds when either σ or τ is replaced by an extended cutoff χ^e.

Appendix B: An example derived from the Euler equations

In this appendix we explain in a particular example how one can derive a single nonlocal nonlinear equation that governs the evolution of the amplitude function a, which itself determines the leading profile $\mathcal{V}^0$; see Proposition 1.24. In the process, we provide explicit constructions of a number of the objects that appeared in our earlier discussion of approximate solutions.

As in [Coulombel and Guès 2010], we consider the linearized Euler equations in two space dimensions to which we add a nonlinear zero-order term (we slightly change notation compared with the introduction). More precisely, we consider the system

$$\begin{cases} \partial_t V^\varepsilon + A_1 \partial_{x_1} V^\varepsilon + A_2 \partial_{x_2} V^\varepsilon + \boldsymbol{D}(V^\varepsilon, V^\varepsilon) = 0, & (t, x_1, x_2) \in (-\infty, T] \times \mathbb{R}^2_+, \\ B V^\varepsilon|_{x_2=0} + \Psi(V^\varepsilon, V^\varepsilon)|_{x_2=0} = \varepsilon^2 G(t, x_1, \phi_0(t, x_1)/\varepsilon), & (t, x_1) \in (-\infty, T] \times \mathbb{R}, \\ V^\varepsilon|_{t<0} = 0, \end{cases} \tag{B-1}$$

where the 3×3 matrices A_1, A_2 are given by

$$A_1 := \begin{pmatrix} 0 & -v & 0 \\ -c^2/v & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_2 := \begin{pmatrix} u & 0 & -v \\ 0 & u & 0 \\ -c^2/v & 0 & u \end{pmatrix},$$

and the parameters v, u, c are chosen so that

$$v > 0, \quad 0 < u < c.$$

The latter assumption corresponds to the linearization of the Euler equations at a given specific volume v with corresponding sound speed c, and a *subsonic incoming* velocity $(0, u)$ (observe the difference with [Coulombel and Guès 2010]). We also assume that $\boldsymbol{D}$ in (B-1) is a symmetric bilinear operator from $\mathbb{R}^3 \times \mathbb{R}^3$ into $\mathbb{R}^3$, and that Ψ is a bilinear operator from $\mathbb{R}^3 \times \mathbb{R}^3$ into $\mathbb{R}^2$ (why we choose $\mathbb{R}^2$ is explained below).

For such parameters, the operator $\partial_t + A_1 \partial_{x_1} + A_2 \partial_{x_2}$ in (B-1) is strictly hyperbolic with three characteristic speeds:

$$\lambda_1(\xi_1, \xi_2) := u\xi_2 - c\sqrt{\xi_1^2 + \xi_2^2}, \quad \lambda_2(\xi_1, \xi_2) := u\xi_2, \quad \lambda_3(\xi_1, \xi_2) := u\xi_2 + c\sqrt{\xi_1^2 + \xi_2^2}.$$

There are two incoming characteristics and one outgoing characteristic, so B should be a 2×3 matrix of maximal rank. The choice of B is made precise below. Of course, the source term G in (B-1) is valued in $\mathbb{R}^2$. We assume moreover that G is 1-periodic and has mean zero with respect to its third variable θ_0. We choose a planar phase ϕ_0 for the oscillations of the boundary source term in (B-1):

$$\phi_0(t, x_1) := \underline{\tau} t + \underline{\eta} x_1, \quad (\underline{\tau}, \underline{\eta}) \neq (0, 0).$$

The hyperbolic region $\mathcal{H}$ can be explicitly computed and is given by

$$\mathcal{H} = \{(\tau, \eta) \in \mathbb{R} \times \mathbb{R} / |\tau| > \sqrt{c^2 - u^2}|\eta|\}.$$

For concreteness, we fix from now on parameters $(\underline{\tau}, \underline{\eta})$ such that $\underline{\eta} > 0$ and $\underline{\tau} = c\underline{\eta}$. In this way, we have $(\underline{\tau}, \underline{\eta}) \in \mathcal{H}$.

We determine the planar characteristic phases whose trace on $\{x_2 = 0\}$ equals ϕ_0. This amounts to finding the roots ω of the dispersion relation

$$\det[\underline{\tau} I + \underline{\eta} A_1 + \omega A_2] = 0.$$

We obtain three real roots that are given by

$$\underline{\omega}_1 := \frac{2M}{1 - M^2}\underline{\eta}, \quad \underline{\omega}_2 := 0, \quad \underline{\omega}_3 := -\frac{1}{M}\underline{\eta}, \quad M := \frac{u}{c} \in (0, 1).$$

The associated (real) phases are $\phi_i(t, x) := \phi_0(t, x_1) + \underline{\omega}_i x_2$, $i = 1, 2, 3$. The relations

$$\underline{\tau} + \lambda_1(\underline{\eta}, \underline{\omega}_1) = \underline{\tau} + \lambda_1(\underline{\eta}, \underline{\omega}_2) = \underline{\tau} + \lambda_2(\underline{\eta}, \underline{\omega}_3) = 0$$

yield the group velocity $\boldsymbol{v}_i$ associated with each phase ϕ_i:

$$\boldsymbol{v}_1 := \frac{1 - M^2}{1 + M^2}\begin{pmatrix} -c \\ -u \end{pmatrix}, \quad \boldsymbol{v}_2 := \begin{pmatrix} -c \\ u \end{pmatrix}, \quad \boldsymbol{v}_3 := \begin{pmatrix} 0 \\ u \end{pmatrix}.$$

Hence the phase ϕ_1 is outgoing while ϕ_2, ϕ_3 are incoming. With the notation of the introduction, we can also compute

$$r_1 := \begin{pmatrix} \dfrac{1 + M^2}{1 - M^2} v \\ c \\ \dfrac{2Mc}{1 - M^2} \end{pmatrix}, \qquad r_2 := \begin{pmatrix} v \\ c \\ 0 \end{pmatrix}, \qquad r_3 := \begin{pmatrix} 0 \\ c \\ u \end{pmatrix},$$

$$\ell_1 := \frac{1 - M^2}{2(1 + M^2)}\begin{pmatrix} 1/v \\ -1/c \\ 1/u \end{pmatrix}, \quad \ell_2 := \frac{1}{2}\begin{pmatrix} 1/v \\ 1/c \\ -1/u \end{pmatrix}, \quad \ell_3 := \frac{1}{1 + M^2}\begin{pmatrix} -1/v \\ 1/c \\ M/c \end{pmatrix},$$

from which one can obtain the expression of the projectors P_1, P_2, P_3 as well as the expression of the partial inverses R_1, R_2, R_3. The stable subspace at the frequency $(\underline{\tau}, \underline{\eta})$ is spanned by the vectors r_2, r_3. The matrix B in (B-1) is chosen as

$$B := \begin{pmatrix} 0 & v & 0 \\ u & 0 & v \end{pmatrix},$$

so that we can choose $e := r_2 - r_3$ as the vector that spans $\ker B \cap \mathbb{E}^s(\underline{\tau}, \underline{\eta})$. The reader can check that all our weak stability assumptions are satisfied with this particular choice of boundary conditions. (We skip the details, which are just slightly more complicated than those in [Coulombel and Guès 2010].) The one-dimensional space $B\mathbb{E}^s(\underline{\tau}, \underline{\eta})$ can be written as the orthogonal of the vector $b := (u, -c)^T$.

The leading profile $\mathcal{V}^0$ and the corrector $\mathcal{V}^1$ satisfy (see Proposition 1.24)

$$\mathcal{V}^0 = \mathcal{V}^0_{\mathrm{inc}} = \sigma_2(t,x,\theta_2)r_2 + \sigma_3(t,x,\theta_3)r_3, \quad \mathcal{V}^1_{\mathrm{out}} = \tau_1(t,x,\theta_1)r_1.$$

Moreover, we have

$$\mathcal{V}^0(t,x_1,0,\theta_0,\theta_0,\theta_0) = a(t,x_1,\theta_0)e = a(t,x_1,\theta_0)(r_2 - r_3),$$

where the scalar function a is 1-periodic with respect to θ_0 and has mean 0. The Fourier coefficients of a are denoted by a_k, $k \in \mathbb{Z}$, where a_0 equals 0 for all time t. Since the functions σ_2, σ_3 satisfy the transport equations[43]

$$\partial_t\sigma_2 + \boldsymbol{v}_2 \cdot \nabla_x\sigma_2 = \partial_t\sigma_3 + \boldsymbol{v}_3 \cdot \nabla_x\sigma_3 = 0,$$

and vanish for $t < 0$, we obtain the expressions

$$\sigma_2(t,x,\theta_2) = a\left(t - \frac{x_2}{u}, x_1 + \frac{x_2}{M}, \theta_2\right), \quad \sigma_3(t,x,\theta_3) = -a\left(t - \frac{x_2}{u}, x_1, \theta_3\right). \tag{B-2}$$

To compute $\mathcal{V}^1_{\mathrm{out}}$, we must solve

$$E_{\mathrm{out}}(L(\partial)\mathcal{V}^1_{\mathrm{out}} + A_2^{-1}\boldsymbol{D}(\mathcal{V}^0_{\mathrm{inc}}, \mathcal{V}^0_{\mathrm{inc}})) = 0 \quad (\text{here } E_{\mathrm{out}} = E_1), \tag{B-3}$$

and we thus need to determine the resonances between the phases. A simple calculation shows that there is a nontrivial $n \in \mathbb{Z}^3$ satisfying $n_1\phi_1 = n_2\phi_2 + n_3\phi_3$ if and only if M^2 is a rational number. We thus assume this to be the case from now on. The resonance between the phases reads

$$n_1 := q, \quad n_2 := p + q, \quad n_3 := -p, \quad \text{with } \frac{2M^2}{1-M^2} = \frac{p}{q},$$

and it is understood that p, q are both positive and have no common divisor (for instance $p = q = 1$ when M equals $1/\sqrt{3}$). Expanding the quadratic term $\boldsymbol{D}(\mathcal{V}^0_{\mathrm{inc}}, \mathcal{V}^0_{\mathrm{inc}})$ in Fourier series, and using the relation

$$\mathcal{C}_1 = \mathbb{Z}\begin{pmatrix}1\\0\\0\end{pmatrix} \cup \mathbb{Z}\begin{pmatrix}0\\n_2\\n_3\end{pmatrix},$$

we obtain (using the expressions (B-2))

$$E_1(A_2^{-1}\boldsymbol{D}(\mathcal{V}^0_{\mathrm{inc}}, \mathcal{V}^0_{\mathrm{inc}})) = -2\sum_{k\in\mathbb{Z}} a_{k(p+q)}\left(t - \frac{x_2}{u}, x_1 + \frac{x_2}{M}\right)a_{-kp}\left(t - \frac{x_2}{u}, x_1\right)e^{2i\pi kq\theta_1}P_1A_2^{-1}\boldsymbol{D}(r_2,r_3).$$

In terms of the interaction integral, we obtain the expression

$$E_1(A_2^{-1}\boldsymbol{D}(\mathcal{V}^0_{\mathrm{inc}}, \mathcal{V}^0_{\mathrm{inc}})) = -2\int_0^1 (a)_{n_2}\left(t - \frac{x_2}{u}, x_1 + \frac{x_2}{M}, \frac{n_1}{n_2}\theta_1 - \frac{n_3}{n_2}\theta_3\right)a\left(t - \frac{x_2}{u}, x_1, \theta_3\right)d\theta_3 P_1A_2^{-1}\boldsymbol{D}(r_2,r_3),$$

[43] Observe that there is no zero-order term in the transport equations because the zero-order term in (B-1) has only a quadratic part. This choice has been made for the sake of simplicity.

where $(a)_{n_2}$ still denotes the action of a under the preparation map that retains only Fourier coefficients that are multiples of n_2. Consequently, (B-3) reads

$$\left(\partial_t - \frac{1-M^2}{1+M^2}c\partial_{x_1} - \frac{1-M^2}{1+M^2}u\partial_{x_2}\right)\tau_1 = \boldsymbol{d}\int_0^1 (a)_{n_2}\left(t-\frac{x_2}{u}, x_1+\frac{x_2}{M}, \frac{n_1}{n_2}\theta_1 - \frac{n_3}{n_2}\theta_3\right) a\left(t-\frac{x_2}{u}, x_1, \theta_3\right) d\theta_3, \quad \text{(B-4)}$$

with

$$\boldsymbol{d} := -2u\frac{1-M^2}{1+M^2}\ell_1 \cdot A_2^{-1}\boldsymbol{D}(r_2, r_3).$$

The transport equation (B-4) is solved by integrating along the characteristics, and we obtain the expression

$$\tau_1(t, x_1, 0, \theta_1) = \boldsymbol{d}\int_0^t\int_0^1 (a)_{n_2}\left(\frac{2s-(1-M^2)t}{1+M^2}, x_1 + 2c\frac{1-M^2}{1+M^2}(t-s), \frac{n_1}{n_2}\theta_1 - \frac{n_3}{n_2}\theta_3\right) \times a\left(\frac{2s-(1-M^2)t}{1+M^2}, x_1 + c\frac{1-M^2}{1+M^2}(t-s), \theta_3\right) d\theta_3\, ds. \quad \text{(B-5)}$$

The Fourier series expansion of τ_1 reads

$$\tau_1(t, x_1, 0, \theta_1) = \boldsymbol{d}\sum_{k\in\mathbb{Z}}\int_0^t a_{k(p+q)}\left(\frac{2s-(1-M^2)t}{1+M^2}, x_1 + 2c\frac{1-M^2}{1+M^2}(t-s)\right) \times a_{-kp}\left(\frac{2s-(1-M^2)t}{1+M^2}, x_1 + c\frac{1-M^2}{1+M^2}(t-s)\right) ds\, e^{2i\pi kq\theta_1}.$$

The equation governing the amplitude a reads

$$b\cdot((a^2)^*\Psi(e,e) + \tau_1|_{x_2=0}Br_1 - BR(L(\partial)\mathcal{V}^0_{\text{inc}})|_{x_2=0}) = b\cdot G,$$

where functions are evaluated at $x_2 = 0$ and $\theta_1 = \theta_2 = \theta_3 = \theta_0$. Since we already have the expression of τ_1 in terms of a, the only task left is to compute the trace of the term $BR(L(\partial)\mathcal{V}^0_{\text{inc}})$. Recalling that $R_2r_2 = R_3r_3 = 0$, we have

$$BR(L(\partial)\mathcal{V}^0_{\text{inc}})|_{x_2=0} = (BR_2A_2^{-1}r_2 + BR_3A_2^{-1}r_3)\partial_t\mathfrak{a} + (BR_2A_2^{-1}A_1r_2 + BR_3A_2^{-1}A_1r_3)\partial_{x_1}\mathfrak{a},$$

with $\mathfrak{a}$ the unique primitive function of a with zero mean. Using the expressions of R_2, R_3 in terms of the projectors P_1, P_2, P_3, which themselves can be obtained from the vectors r_i, ℓ_i, we get

$$b\cdot(BR_2A_2^{-1}r_2 + BR_3A_2^{-1}r_3) = -\frac{uv(1+M^2)}{M^2\eta},$$
$$b\cdot(BR_2A_2^{-1}A_1r_2 + BR_3A_2^{-1}A_1r_3) = \frac{ucv(1+M^2)}{M^2\eta}.$$

The fact that both quantities are proportional to each other with a factor $-c$ comes from a general fact; see [Coulombel and Guès 2010, Lemma 5.1].

The function a should therefore satisfy the amplitude equation

$$\frac{uv(1+M^2)}{M^2\underline{\eta}}(\partial_t \mathfrak{a} - c\partial_{x_1}\mathfrak{a}) + b\cdot\Psi(e,e)(a^2)^* + b\cdot Br_1\tau_1|_{x_2=0} = b\cdot G,$$

or, equivalently,

$$\frac{uv(1+M^2)}{M^2\underline{\eta}}(\partial_t a - c\partial_{x_1}a) + b\cdot\Psi(e,e)\partial_{\theta_0}(a^2) + b\cdot Br_1\partial_{\theta_0}\tau_1|_{x_d=0} = b\cdot\partial_{\theta_0}G. \tag{B-6}$$

Let us define the two constants

$$\alpha_1 := \frac{M^2\underline{\eta}}{uv(1+M^2)}b\cdot\Psi(e,e), \quad \alpha_2 := \frac{4ucM^2\underline{\eta}}{1+M^2}\ell_1\cdot A_2^{-1}\boldsymbol{D}(r_2,r_3).$$

Then (B-6) reads

$$\partial_t a - c\partial_{x_1}a + \alpha_1\partial_{\theta_0}(a^2) + \alpha_2\partial_{\theta_0}\frac{\tau_1}{\boldsymbol{d}}|_{x_2=0} = \frac{M^2\underline{\eta}}{uv(1+M^2)}b\cdot\partial_{\theta_0}G,$$

where the derivative $\partial_{\theta_0}\tau_1/\boldsymbol{d}|_{x_2=0}$ is computed from the relation (B-5):

$$\partial_{\theta_0}\frac{\tau_1}{\boldsymbol{d}}\Big|_{x_2=0} = \frac{n_1}{n_2}\int_0^t\int_0^1 (\partial_{\theta_0}a)_{n_2}\Big(\frac{2s-(1-M^2)t}{1+M^2}, x_1+2c\frac{1-M^2}{1+M^2}(t-s), \frac{n_1}{n_2}\theta_0 - \frac{n_3}{n_2}\Theta\Big)$$
$$\times a\Big(\frac{2s-(1-M^2)t}{1+M^2}, x_1+c\frac{1-M^2}{1+M^2}(t-s), \Theta\Big)\,d\Theta\,ds.$$

In terms of the Fourier coefficients a_k, the latter equation is seen to be equivalent to the infinite system of transport equations

$$\partial_t a_k - c\partial_{x_1}a_k + 2i\pi k\alpha_1\sum_{k'\in\mathbb{Z}} a_{k'}a_{k-k'} = 2i\pi k\frac{M^2\underline{\eta}}{uv(1+M^2)}b\cdot G_k, \quad k\notin q\mathbb{Z},$$

and

$$\partial_t a_{kq} - c\partial_{x_1}a_{kq} + 2i\pi kq\alpha_1\sum_{k'\in\mathbb{Z}} a_{k'}a_{kq-k'} + 2i\pi kq\alpha_2$$
$$\times\int_0^t a_{k(p+q)}\Big(\frac{2s-(1-M^2)t}{1+M^2}, x_1+2c\frac{1-M^2}{1+M^2}(t-s)\Big)a_{-kp}\Big(\frac{2s-(1-M^2)t}{1+M^2}, x_1+c\frac{1-M^2}{1+M^2}(t-s)\Big)\,ds$$
$$= 2i\pi kq\frac{M^2\underline{\eta}}{uv(1+M^2)}b\cdot G_{kq}.$$

We recall that the coefficient a_0 vanishes.

In the special case $M = 1/\sqrt{3}$, the above system reduces to

$$\partial_t a_k - c\partial_{x_1}a_k + 2i\pi k\alpha_1\sum_{k'\in\mathbb{Z}} a_{k'}a_{k-k'} + 2i\pi k\alpha_2\int_0^t a_{2k}\Big(\frac{3s-t}{2}, x_1+c(t-s)\Big)a_{-k}\Big(\frac{3s-t}{2}, x_1+\frac{c}{2}(t-s)\Big)\,ds$$
$$= 2i\pi k\frac{\eta}{4uv}b\cdot G_k, \quad k\in\mathbb{Z},$$

with parameters α_1, α_2 computed from the nonlinearities $\boldsymbol{D}$, Ψ in (B-1):

$$\alpha_1 := \frac{\eta}{4uv} b \cdot \Psi(e,e), \quad \alpha_2 := uc\underline{\eta}\ell_1 \cdot A_2^{-1} \boldsymbol{D}(r_2, r_3).$$

References

[Alinhac 1989] S. Alinhac, "Existence d'ondes de raréfaction pour des systèmes quasi-linéaires hyperboliques multidimensionnels", *Comm. Partial Differential Equations* **14**:2 (1989), 173–230. MR 90h:35147b Zbl 0692.35063

[Alinhac and Gérard 2007] S. Alinhac and P. Gérard, *Pseudo-differential operators and the Nash–Moser theorem*, Graduate Studies in Mathematics **82**, Amer. Math. Soc., Providence, 2007. MR 2007m:35001 Zbl 1121.47033

[Artola and Majda 1987] M. Artola and A. J. Majda, "Nonlinear development of instabilities in supersonic vortex sheets, I: The basic kink modes", *Phys. D* **28**:3 (1987), 253–281. MR 88i:76025 Zbl 0632.76074

[Benzoni-Gavage and Serre 2007] S. Benzoni-Gavage and D. Serre, *Multidimensional hyperbolic partial differential equations: first-order systems and applications*, Oxford University Press, 2007. MR 2008k:35002 Zbl 1113.35001

[Benzoni-Gavage et al. 2002] S. Benzoni-Gavage, F. Rousset, D. Serre, and K. Zumbrun, "Generic types and transitions in hyperbolic initial-boundary-value problems", *Proc. Roy. Soc. Edinburgh Sect. A* **132**:5 (2002), 1073–1104. MR 2003j:35200 Zbl 1029.35165

[Chazarain and Piriou 1982] J. Chazarain and A. Piriou, *Introduction to the theory of linear partial differential equations*, Studies in Mathematics and its Applications **14**, North-Holland, Amsterdam, 1982. MR 83j:35001 Zbl 0487.35002

[Coulombel 2004] J.-F. Coulombel, "Weakly stable multidimensional shocks", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **21**:4 (2004), 401–443. MR 2005e:35157 Zbl 1072.35120

[Coulombel 2005] J.-F. Coulombel, "Well-posedness of hyperbolic initial boundary value problems", *J. Math. Pures Appl.* (9) **84**:6 (2005), 786–818. MR 2006h:35166 Zbl 1078.35066

[Coulombel and Guès 2010] J.-F. Coulombel and O. Guès, "Geometric optics expansions with amplification for hyperbolic boundary value problems: linear problems", *Ann. Inst. Fourier* (*Grenoble*) **60**:6 (2010), 2183–2233. MR 2012d:35209 Zbl 1218.35137

[Coulombel and Williams 2013] J.-F. Coulombel and M. Williams, "Amplification of pulses in nonlinear geometric optics", preprint, 2013. To appear in J. Hyp. Differential Equations. arXiv 1308.6686

[Coulombel et al. 2011] J.-F. Coulombel, O. Guès, and M. Williams, "Resonant leading order geometric optics expansions for quasilinear hyperbolic fixed and free boundary problems", *Comm. Partial Differential Equations* **36**:10 (2011), 1797–1859. MR 2012k:35315 Zbl 1241.35131

[Coulombel et al. 2012] J.-F. Coulombel, O. Guès, and M. Williams, "Singular pseudodifferential calculus for wavetrains and pulses", Preprint, 2012. To appear in *Bull. Soc. Math. France*. arXiv 1201.6202

[Joly et al. 1993] J.-L. Joly, G. Métivier, and J. Rauch, "Generic rigorous asymptotic expansions for weakly nonlinear multidimensional oscillatory waves", *Duke Math. J.* **70**:2 (1993), 373–404. MR 94c:35048 Zbl 0815.35066

[Joly et al. 1995] J.-L. Joly, G. Métivier, and J. Rauch, "Coherent and focusing multidimensional nonlinear geometric optics", *Ann. Sci. École Norm. Sup.* (4) **28**:1 (1995), 51–113. MR 95k:35035 Zbl 0836.35087

[Kreiss 1970] H.-O. Kreiss, "Initial boundary value problems for hyperbolic systems", *Comm. Pure Appl. Math.* **23** (1970), 277–298. MR 55 #10862 Zbl 0193.06902

[Lax 1957] P. D. Lax, "Asymptotic solutions of oscillatory initial value problems", *Duke Math. J.* **24** (1957), 627–646. MR 20 #4096 Zbl 0083.31801

[Lescarret 2007] V. Lescarret, "Wave transmission in dispersive media", *Math. Models Methods Appl. Sci.* **17**:4 (2007), 485–535. MR 2008c:78037 Zbl 1220.35170

[Majda and Artola 1988] A. J. Majda and M. Artola, "Nonlinear geometric optics for hyperbolic mixed problems", pp. 319–356 in *Analyse mathématique et applications*, Gauthier-Villars, Montrouge, 1988. MR 90c:35152 Zbl 0674.35057

[Majda and Rosales 1983] A. Majda and R. Rosales, "A theory for spontaneous Mach stem formation in reacting shock fronts, I: The basic perturbation analysis", *SIAM J. Appl. Math.* **43**:6 (1983), 1310–1334. MR 84i:76051 Zbl 0544.76135

[Majda and Rosales 1984] A. Majda and R. Rosales, "A theory for spontaneous Mach-stem formation in reacting shock fronts, II: Steady-wave bifurcations and the evidence for breakdown", *Stud. Appl. Math.* **71**:2 (1984), 117–148. MR 86b:35133

[Marcou 2010] A. Marcou, "Rigorous weakly nonlinear geometric optics for surface waves", *Asymptot. Anal.* **69**:3-4 (2010), 125–174. MR 2011m:35370 Zbl 1222.35118

[Métivier 2000] G. Métivier, "The block structure condition for symmetric hyperbolic systems", *Bull. London Math. Soc.* **32**:6 (2000), 689–702. MR 2001i:35198 Zbl 1073.35525

[Sablé-Tougeron 1988] M. Sablé-Tougeron, "Existence pour un problème de l'élastodynamique Neumann non linéaire en dimension 2", *Arch. Rational Mech. Anal.* **101**:3 (1988), 261–292. MR 89f:35191 Zbl 0652.73019

[Williams 1996] M. Williams, "Nonlinear geometric optics for hyperbolic boundary problems", *Comm. Partial Differential Equations* **21**:11-12 (1996), 1829–1895. MR 98d:35136 Zbl 0881.35068

[Williams 2000] M. Williams, "Boundary layers and glancing blow-up in nonlinear geometric optics", *Ann. Sci. École Norm. Sup.* (4) **33**:3 (2000), 383–432. MR 2001f:35392 Zbl 0962.35118

[Williams 2002] M. Williams, "Singular pseudodifferential operators, symmetrizers, and oscillatory multidimensional shocks", *J. Funct. Anal.* **191**:1 (2002), 132–209. MR 2003e:35334 Zbl 1028.35174

Received 28 Feb 2013. Accepted 29 Apr 2013.

JEAN-FRANCOIS COULOMBEL: jean-francois.coulombel@univ-nantes.fr
CNRS, Laboratoire de mathématiques Jean Leray (UMR CNRS 6629), Université de Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes, France

OLIVIER GUÈS: gues@cmi.univ-mrs.fr
Laboratoire d'Analyse, Topologie et Probabilités (UMR CNRS 6632), Université de Provence, Technopôle Château-Gombert, 39 rue F. Joliot Curie, 13453 Marseille 13, France

MARK WILLIAMS: williams@email.unc.edu
Mathematics Department, University of North Carolina, CB 3250, Phillips Hall, Chapel Hill, NC 27599, United States

ANALYSIS AND PDE
Vol. 7, No. 3, 2014
dx.doi.org/10.2140/apde.2014.7.627

THE 1-HARMONIC FLOW WITH VALUES IN A HYPEROCTANT OF THE N-SPHERE

LORENZO GIACOMELLI, JOSE M. MAZÓN AND SALVADOR MOLL

We prove the existence of solutions to the 1-harmonic flow — that is, the formal gradient flow of the total variation of a vector field with respect to the L^2-distance — from a domain of $\mathbb{R}^m$ into a hyperoctant of the N-dimensional unit sphere, $\mathbb{S}^{N-1}_+$, under homogeneous Neumann boundary conditions. In particular, we characterize the lower-order term appearing in the Euler–Lagrange formulation in terms of the "geodesic representative" of a BV-director field on its jump set. Such characterization relies on a lower semicontinuity argument which leads to a nontrivial and nonconvex minimization problem: to find a shortest path between two points on $\mathbb{S}^{N-1}_+$ with respect to a metric which penalizes the closeness to their geodesic midpoint.

1. Introduction 627
2. Preliminaries 631
3. Existence of solutions 639
4. A nonconvex variational problem 652
Acknowledgement 669
References 669

1. Introduction

Throughout the paper, $\Omega \subset \mathbb{R}^m$ is a bounded domain with Lipschitz continuous boundary $\partial\Omega$ and $\mathbb{S}^{N-1}$ is the unit sphere of $\mathbb{R}^N$. For a smooth map $\boldsymbol{u} : \Omega \to \mathbb{S}^{N-1}$ and $1 \le p < \infty$, the p-energy of $\boldsymbol{u}$ is given by

$$E_p(\boldsymbol{u}) = \int_\Omega |D\boldsymbol{u}|^p \, dx.$$

A critical point $\boldsymbol{u} \in C^1(\Omega; \mathbb{S}^{N-1})$ of the p-energy, a *p-harmonic map*, formally satisfies the Euler–Lagrange equation

$$-\operatorname{div}(|D\boldsymbol{u}|^{p-2} D\boldsymbol{u}) = |D\boldsymbol{u}|^p \boldsymbol{u}. \tag{1-1}$$

The term $|D\boldsymbol{u}|^p$ plays the role of a Lagrange multiplier corresponding to the pointwise constraint $|\boldsymbol{u}| = 1$.

Mazón and Moll have been partially supported by the Spanish MEC project MTM2012–31103.
MSC2010: primary 35K55, 49Q20, 53C22, 35K67, 35K92; secondary 49J45, 53C44, 58E20, 68U10.
Keywords: harmonic flows, total variation flow, nonlinear parabolic systems, lower semicontinuity and relaxation, nonconvex variational problems, geodesics, Riemannian manifolds with boundary, image processing.

One well-known method to obtain (distributional) solutions to (1-1), the so-called *heat-flow method*, introduced by J. Eells and J. H. Sampson [1964] for $p = 2$ in the general framework of Riemannian manifolds, consists in looking at long time limits of solutions to

$$\boldsymbol{u}_t = \operatorname{div}(|D\boldsymbol{u}|^{p-2} D\boldsymbol{u}) + \boldsymbol{u}|D\boldsymbol{u}|^p, \quad |\boldsymbol{u}| = 1. \tag{1-2}$$

Equation (1-2) is also a prototype for often quite complicated reaction-diffusion systems for the evolution of director fields which arise in various contexts — multigrain problems [Kobayashi et al. 2000], theory of liquid crystals [van der Hout 2001], ferromagnetism [DeSimone and Podio-Guidugli 1996], and image processing [Sapiro 2001]. For $p > 1$, (1-2) with various boundary conditions has been widely studied over the last decades; referenced discussions of the cases $p = 2$ and $p \in (1, \infty)$ may be found, for example, in [Bertsch et al. 2003; Bertsch et al. 2002; Chen 1989; Struwe 1992] and [Chen et al. 1994; Hungerbühler 2004; Misawa 2002], respectively.

Here we are interested in the case $p = 1$, for which (1-2) formally reads

$$\boldsymbol{u}_t = \operatorname{div}\left(\frac{D\boldsymbol{u}}{|D\boldsymbol{u}|}\right) + \boldsymbol{u}|D\boldsymbol{u}|, \quad \boldsymbol{u} \in \mathbb{S}^{N-1}. \tag{1-3}$$

More precisely, we focus on the homogeneous Neumann problem for (1-3) when the target space is a compact subset $\mathbb{A}$ of $\mathbb{S}^{N-1}$; that is,

$$\begin{cases} \boldsymbol{u}_t = \operatorname{div}\left(\dfrac{D\boldsymbol{u}}{|D\boldsymbol{u}|}\right) + \boldsymbol{u}|D\boldsymbol{u}|, \ \boldsymbol{u} \in \mathbb{A} \subseteq \mathbb{S}^{N-1}, & \text{in } Q_T = (0,T) \times \Omega, \\ \dfrac{D\boldsymbol{u}}{|D\boldsymbol{u}|}\nu = 0 & \text{on } S_T = (0,T) \times \partial\Omega, \\ \boldsymbol{u}(0,\cdot) = \boldsymbol{u}_0(\cdot), \ \boldsymbol{u}_0 \in \mathbb{A}, & \text{in } \Omega, \end{cases} \tag{1-4}$$

where ν denotes the outward unit normal to $\partial\Omega$. Problem (1-4) was proposed as a tool to denoise either two-dimensional image gradients and optical flows, in which case $N = 2$ and $\mathbb{A} = \mathbb{S}^1$ [Tang et al. 2000], or color images by smoothing the chromaticity data while preserving the contrast, in which case $N = 3$ and $\mathbb{A}$ is an octant of the sphere [Tang et al. 2001].

While the scalar and unconstrained version of (1-3), that is, the so-called total variation flow, is by now well understood after the pioneering paper [Andreu et al. 2001] (see the monograph [Andreu-Vaillo et al. 2004] and the references therein or [Bonforte and Figalli 2012] for an up-to-date reference list). An existence theory for (1-3) is still open in general. Special cases considered so far have dealt with piece-wise constant data [Giga and Kobayashi 2003; Giga et al. 2005; Giga et al. 2007], initial data with "small" energy [Giga et al. 2004], and rotationally symmetric solutions [Giga and Kuroda 2004; Dal Passo et al. 2008; Giacomelli and Moll 2010]. We refer to [Giacomelli et al. 2013a] for a detailed discussion of previous attempts to obtain a solution to (1-4) given in [Barrett et al. 2008; Feng 2010].

In dealing with (1-3), the most delicate issue is of course the interpretation of the bounded matrix $\boldsymbol{Z}$, which represents $D\boldsymbol{u}/|D\boldsymbol{u}|$, and of the measure $\boldsymbol{\mu}$, which represents $\boldsymbol{u}|D\boldsymbol{u}|$, the latter being the product between a measure and a possibly discontinuous function. Very recently, an interpretation of (1-3) has

been proposed in [Giacomelli et al. 2013a]: in summary,

$$\boldsymbol{u}_t(t) - \operatorname{div} \boldsymbol{Z}(t) \in \boldsymbol{u}_g\,|D\boldsymbol{u}|(t),\ \boldsymbol{u}(t) \in \mathbb{A} \quad \text{for a.e. } t \in [0,T] \tag{1-5}$$

in the sense of distributions, where $\boldsymbol{Z}(t)$ is a bounded matrix that represents $D\boldsymbol{u}(t)/|D\boldsymbol{u}(t)|$ (the precise meaning is given in Proposition 3.5) and $\boldsymbol{u}_g|D\boldsymbol{u}|(t)$ denotes a set of vector-valued measures which are oriented as $\boldsymbol{u}(t)^*$ (the precise representative of $\boldsymbol{u}(t)$) and have total variation density $|D\boldsymbol{u}(t)|$. For $N=2$, this interpretation has led to the existence and uniqueness of a solution to (1-4) when $\mathbb{A}$ is a semicircle [Giacomelli et al. 2013a, Theorems 4.1 and 5.1] together with the existence of a solution when $\mathbb{A}=\mathbb{S}^1$ and $\boldsymbol{u}_0 \in BV(\Omega;\mathbb{S}^1)$ has no jumps by an "angle" larger than π.

The aim of this paper is to prove an existence result, according to the same interpretation, for an arbitrary dimension of the target sphere. We consider (1-4) in the first hyperoctant of the N-sphere:

$$\mathbb{A} = \mathbb{S}^{N-1}_+ := \{(x_1,\dots,x_N) \in \mathbb{S}^{N-1} : x_i \geq 0 \text{ for } i=1,\dots,N\}$$

(a natural assumption in the context of image processing; see above). Note that in this case, for every pair $\boldsymbol{u}_-, \boldsymbol{u}_+ \in \mathbb{S}^{N-1}_+$ there exists a unique *geodesic midpoint*, $\boldsymbol{u}_g = (\boldsymbol{u}_+ + \boldsymbol{u}_-)/|\boldsymbol{u}_+ + \boldsymbol{u}_-|$ (see Definition 3.1). Hence we may define the *geodesic representative of* $\boldsymbol{u} \in BV(\Omega;\mathbb{S}^{N-1}_+)$, $\boldsymbol{u}_g := \boldsymbol{u}^*/|\boldsymbol{u}^*|$ (see Definition 3.2 and Remark 3.3) and the set of measures in (1-5) reduces to the singleton $\boldsymbol{u}(t)_g|D\boldsymbol{u}(t)|$.

The complete definition of a solution and the statement of the main result are given in Definition 3.4 and Theorem 3.6, respectively. We obtain a solution as the limit of a sequence of solutions to the following approximating problems (see Proposition 3.7 and Lemma 3.8):

$$\begin{cases} \boldsymbol{u}^\varepsilon_t = \operatorname{div} \boldsymbol{Z}^\varepsilon + \boldsymbol{\mu}^\varepsilon, \quad \boldsymbol{u}^\varepsilon \in \mathbb{S}^{N-1}_+ & \text{in } \Omega_T, \\ [\boldsymbol{Z}^\varepsilon, \nu] = 0 & \text{on } S_T, \\ \boldsymbol{u}^\varepsilon(0,\cdot) = \boldsymbol{u}^\varepsilon_0(\cdot) & \text{in } \Omega, \end{cases}$$

where

$$\boldsymbol{Z}^\varepsilon = \varepsilon^\alpha \nabla \boldsymbol{u}^\varepsilon + \frac{\nabla \boldsymbol{u}^\varepsilon}{\sqrt{|\nabla \boldsymbol{u}^\varepsilon|^2+\varepsilon^2}}, \quad \boldsymbol{\mu}^\varepsilon = \varepsilon^\alpha \boldsymbol{u}^\varepsilon |\nabla \boldsymbol{u}^\varepsilon|^2 + \boldsymbol{u}^\varepsilon \frac{|\nabla \boldsymbol{u}^\varepsilon|^2}{\sqrt{|\nabla \boldsymbol{u}^\varepsilon|^2+\varepsilon^2}}, \tag{1-6}$$

and the initial data suitably converge to a given $\boldsymbol{u}_0 \in BV(\Omega;\mathbb{S}^{N-1}_+)$ (see Lemma 3.9). The strategy we follow is completely different from that in [Giacomelli et al. 2013a], where the special structure of $\mathbb{S}^1$ was heavily used. Its core, neglecting any technicality and concentrating on the crucial issues, may be summarized as follows (see also [Giacomelli et al. 2013b] for a slightly more detailed discussion). By fairly standard compactness arguments, we obtain convergence of $\boldsymbol{u}^\varepsilon$, $\boldsymbol{Z}^\varepsilon$, and $\boldsymbol{\mu}^\varepsilon$ to $\boldsymbol{u}$, $\boldsymbol{Z}$, and $\boldsymbol{\mu}$, respectively (see Step 1 in the proof of Theorem 3.6). The functions $\boldsymbol{u}$ and $\boldsymbol{Z}$ can be seen to satisfy, for a.e. $t \in [0,T]$,

$$\boldsymbol{u}_t(t) - \operatorname{div} \boldsymbol{Z}(t) = \boldsymbol{\mu}(t) \quad \text{in } \mathcal{M}(\Omega;\mathbb{R}^N).$$

Then we show, by a relatively soft argument, which nevertheless requires quite a few preliminaries, that

$$\boldsymbol{\mu} = *(*(\boldsymbol{Z} \wedge \boldsymbol{u}) \wedge D\boldsymbol{u}) \quad \text{and} \quad |\boldsymbol{\mu}(t)| \leq |D\boldsymbol{u}(t)| \quad \text{as measures for a.e. } t \in [0,T] \tag{1-7}$$

(see Step 2 in the proof of Theorem 3.6). Hence, in order to identify $\boldsymbol{\mu}$ it suffices to show that

$$\boldsymbol{u}(t)_g \cdot \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \geq 1 \quad \text{for a.e. } t \in [0, T], \tag{1-8}$$

where $\boldsymbol{\mu}(t)/|D\boldsymbol{u}(t)|$ denotes the Radon–Nikodým derivative of $\boldsymbol{\mu}(t)$ with respect to $|D\boldsymbol{u}(t)|$. Indeed, (1-7) and simple vectorial identities then imply that

$$\boldsymbol{\mu}(t) = \boldsymbol{u}(t)_g |D\boldsymbol{u}(t)| \quad \text{for a.e. } t \in [0, T]$$

(see Step 6 in the proof of Theorem 3.6). In view of (1-6), the lower bound (1-8) for the diffuse part of $\boldsymbol{\mu}$ follows (see Step 4 in the proof of Theorem 3.6) from a suitable modification of a relaxation result [Alicandro et al. 2007], applied to each of the components of

$$\mathscr{F}(\boldsymbol{v}) := \int_\Omega \boldsymbol{v}(x) |\nabla \boldsymbol{v}(x)|\, dx$$

(see Section 2F). On the other hand, the same argument would lead to a suboptimal lower bound on $\boldsymbol{\mu}(t)$ over the jump set of $\boldsymbol{u}(t)$ (see Remark 3.10). Moreover, the results in [Alicandro et al. 2007] can not be directly applied to $\boldsymbol{u}(t)_g \cdot \boldsymbol{\mu}^\varepsilon(t)$, since $\boldsymbol{u}(t)_g$ is a discontinuous function (though a very special one). For these reasons, we revisit the blow-up argument in [Fonseca and Müller 1993] and the dimensional reduction argument in [Fonseca and Rybka 1992] to conclude that

$$\boldsymbol{u}(t)_g \cdot \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \geq \frac{1}{|\boldsymbol{u}(t)_+(x) - \boldsymbol{u}(t)_-(x)|} \inf_{\boldsymbol{\gamma} \in \widetilde{\Gamma}_N} \int_0^1 \boldsymbol{u}(t)_g(x) \cdot \boldsymbol{\gamma}(s) |\boldsymbol{\gamma}'(s)|\, ds \tag{1-9}$$

for a.e. t and $\mathscr{H}^{m-1}$-a.e. $x \in J_{\boldsymbol{u}(t)}$, where

$$\widetilde{\Gamma}_N := \{\boldsymbol{\gamma} \in W^{1,1}((0,1); \mathbb{S}^{N-1}_+) : \boldsymbol{\gamma}(0) = \boldsymbol{u}(t)_-(x),\ \boldsymbol{\gamma}(1) = \boldsymbol{u}(t)_+(x)\} \tag{1-10}$$

(see Step 5 in the proof of Theorem 3.6). The minimization problem which appears on the right-hand side of (1-9) is crucial in our argument. In Section 4 we argue that

$$\min_{\boldsymbol{\gamma} \in \Gamma} \int_0^1 \boldsymbol{u}_g \cdot \boldsymbol{\gamma}(s) |\boldsymbol{\gamma}'(s)|\, ds = |\boldsymbol{u}_+ - \boldsymbol{u}_-|, \quad \text{where } \Gamma = \{\boldsymbol{\gamma} \in W^{1,1}((0,1); \mathbb{S}^{N-1}_+) : \boldsymbol{\gamma}(0) = \boldsymbol{u}_-,\ \boldsymbol{\gamma}(1) = \boldsymbol{u}_+\} \tag{1-11}$$

(see Theorem 4.1). Together with (1-9), (1-11) yields the lower bound (1-8) on the jump set of $\boldsymbol{u}(t)$ too.

The minimization problem in (1-11) is equivalent to finding — and characterizing the length of — shortest paths between $\boldsymbol{u}_-$ and $\boldsymbol{u}_+$ in a Riemannian manifold with boundary whose metric penalizes the closeness to $\boldsymbol{u}_g$. In addition, the metric may degenerate at a point of the manifold: for instance, if $N = 3$, $\boldsymbol{u}_- = (0, 0, 1)$, and $\boldsymbol{u}_+ = (0, 1, 0)$, then $\boldsymbol{u}_g \cdot (1, 0, 0) = 0$. In these respects, the minimization problem has a geometrical interest of its own.

It turns out that the minimum in (1-11) is achieved by the standard geodesic on $\mathbb{S}^{N-1}_+$ connecting $\boldsymbol{u}_-$ and $\boldsymbol{u}_+$; see Lemma 4.2. Nevertheless, the analysis of (1-11) is highly nontrivial for two reasons. Firstly, one has to characterize the length of candidate shortest paths which may in principle intersect and/or de-touch from the boundary of the manifold. Secondly, the functional in (1-11) is genuinely nonconvex:

indeed, besides the aforementioned standard geodesic, it always possesses a second smooth critical point, which we show not to be a shortest path. In addition, in the extreme cases in which $\boldsymbol{u}_+$ and $\boldsymbol{u}_-$ are two distinct "vertices" of $\mathbb{S}^{N-1}_+$, the functional in (1-11) possesses a second shortest path which is not a critical point: it follows the boundary of $\mathbb{S}^{N-1}_+$ and passes through the point of degeneracy. For instance, if $N = 3$, $\boldsymbol{u}_- = (0, 0, 1)$, and $\boldsymbol{u}_+ = (0, 1, 0)$, then $\boldsymbol{u}_g = (0, 1, 1)/\sqrt{2}$ and the curve

$$\boldsymbol{\gamma}(s) = \begin{cases} (\sin(\pi s), 0, \cos(\pi s)) & \text{if } s \in [0, 1/2], \\ (\sin(\pi s), -\cos(\pi s), 0) & \text{if } s \in (1/2, 1] \end{cases}$$

is such that

$$\int_0^1 \boldsymbol{u}_g \cdot \boldsymbol{\gamma}(s)|\boldsymbol{\gamma}'(s)|\, ds = 2\int_0^{1/2} \frac{\cos(\pi s)}{\sqrt{2}}\pi\, ds = \sqrt{2} = |\boldsymbol{u}_+ - \boldsymbol{u}_-|.$$

Finally, we note that if the paths in Γ are allowed to take values in a set $\mathbb{A}$ which contains $\mathbb{S}^{N-1}_+$, then in general the standard geodesic is not a minimizer and (1-11) does not hold; an example is given in Remark 4.4.

The paper is organized as follows. In Section 2 we collect the definitions and results which we need concerning multivector fields, functions of bounded variations, a generalized Green's formula, tensor fields, and lower semicontinuity of integral functionals. In Section 3 we introduce the concept of and prove the existence of a solution to (1-4). Section 4 is devoted to the minimization problem in (1-11).

2. Preliminaries

In this section we introduce some notation and some preliminary results that we need in the sequel.

General notations. Throughout this paper $\mathscr{H}^{m-1}$ denotes the $(m-1)$-dimensional Hausdorff measure and $\mathscr{L}^m$ the m-dimensional Lebesgue measure. We denote by $\mathcal{M}(\Omega; \mathbb{R}^N)$ the space of $\mathbb{R}^N$-valued finite Radon measures on Ω; see [Ambrosio et al. 2000, Definition 1.40]. We recall that $\mathcal{M}(\Omega; \mathbb{R}^N)$ is the dual space of $C_0(\Omega; \mathbb{R}^N)$. Throughout, the subscript $_0$ denotes spaces of compactly supported functions. We denote $\mathcal{D}(\Omega; \mathbb{R}^N) := C_0^\infty(\Omega; \mathbb{R}^N)$. When $N = 1$, we often do not specify the target space (for example, $\mathcal{M}(\Omega) = \mathcal{M}(\Omega; \mathbb{R})$). Finally, if $\mathbb{A} \subset \mathbb{R}^N$ is compact and $\Upsilon(\Omega; \mathbb{R}^N)$ is a space of functions, we sometimes use the notation $\Upsilon(\Omega; \mathbb{A}) := \{\boldsymbol{u} \in \Upsilon(\Omega; \mathbb{R}^N) : \boldsymbol{u}(x) \in \mathbb{A} \text{ for } \mathscr{L}^m\text{-a.e. } x \in \Omega\}$.

2A. ***Multivectors.*** Here we recall some definitions and basic properties about multivectors that we need in our analysis. We refer to, for example, [Federer 1969, Chapter 1; Darling 1994, Chapter 1] for details.

The spaces $\Lambda_0(\mathbb{R}^N)$ and $\Lambda_1(\mathbb{R}^N)$ coincide with $\mathbb{R}$ and $\mathbb{R}^N$, respectively. For $2 \le k \le \mathbb{N}$, the k-th exterior power of $\mathbb{R}^N$, denoted by $\Lambda_k(\mathbb{R}^N)$, is a set spanned by elements of the form

$$\boldsymbol{u}_1 \wedge \cdots \wedge \boldsymbol{u}_k, \quad \boldsymbol{u}_i \in \mathbb{R}^N,\ i = 1, \ldots, k$$

(elements of this form are called "generators") and subject to the following rules:

(1) $(a\boldsymbol{v} + b\boldsymbol{w}) \wedge \boldsymbol{u}_2 \wedge \cdots \wedge \boldsymbol{u}_k = a(\boldsymbol{v} \wedge \boldsymbol{u}_2 \wedge \cdots \wedge \boldsymbol{u}_k) + b(\boldsymbol{w} \wedge \boldsymbol{u}_2 \wedge \cdots \wedge \boldsymbol{u}_k)$.

(2) $\boldsymbol{u}_1 \wedge \cdots \wedge \boldsymbol{u}_k$ changes sign if two entries are transposed.

(3) For any basis $\{\boldsymbol{e}_1, \dots \boldsymbol{e}_n\}$ of $\mathbb{R}^N$, $\{\boldsymbol{e}_\alpha := \boldsymbol{e}_{\alpha_1} \wedge \cdots \wedge \boldsymbol{e}_{\alpha_k} : \alpha \in I(k, N)\}$ is a basis for $\Lambda_k(\mathbb{R}^N)$.

Here we have used the standard notation for ordered multiindexes:

$$I(k, N) = \{\alpha = (\alpha_1, \dots, \alpha_k) : \alpha_i \in \boldsymbol{Z},\ 1 \le \alpha_1 < \cdots < \alpha_k \le N\}. \tag{2-1}$$

The elements of $\Lambda_k(\mathbb{R}^N)$ are called multivectors (or k-vectors), and $\Lambda_k(\mathbb{R}^N)$ is a vector space of dimension $\binom{N}{k}$. We will use the well-known equality [Darling 1994, Formula 1.68]

$$|\boldsymbol{a}|^2|\boldsymbol{b}|^2 = (\boldsymbol{a} \cdot \boldsymbol{b})^2 + (\boldsymbol{a} \wedge \boldsymbol{b})^2 \quad \text{for all } \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N. \tag{2-2}$$

Given $k, p \in \{0, \dots, N\}$ with $k + p \le N$, there exists a unique bilinear map $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \to \boldsymbol{\lambda} \wedge \boldsymbol{\mu}$ from $\Lambda_k(\mathbb{R}^N) \times \Lambda_p(\mathbb{R}^N)$ to $\Lambda_{k+p}(\mathbb{R}^N)$, whose effect on generators is

$$(\boldsymbol{u}_1 \wedge \boldsymbol{u}_2 \wedge \cdots \wedge \boldsymbol{u}_k) \wedge (\boldsymbol{v}_1 \wedge \boldsymbol{v}_2 \wedge \cdots \wedge \boldsymbol{v}_p) = \boldsymbol{u}_1 \wedge \boldsymbol{u}_2 \wedge \cdots \wedge \boldsymbol{u}_k \wedge \boldsymbol{v}_1 \wedge \boldsymbol{v}_2 \wedge \cdots \wedge \boldsymbol{v}_p.$$

This map satisfies

$$\boldsymbol{\lambda} \wedge \boldsymbol{\mu} = (-1)^{-kp}(\boldsymbol{\mu} \wedge \boldsymbol{\lambda}) \quad \text{for } \boldsymbol{\lambda} \in \Lambda_k(\mathbb{R}^N),\ \boldsymbol{\mu} \in \Lambda_p(\mathbb{R}^N). \tag{2-3}$$

The *Hodge-star operator* is an isomorphism from $\Lambda_k(\mathbb{R}^N)$ to $\Lambda_{N-k}(\mathbb{R}^N)$, defined on the basis as

$$*(\boldsymbol{e}_{\alpha_1} \wedge \cdots \wedge \boldsymbol{e}_{\alpha_k}) := \boldsymbol{e}_{\alpha_{k+1}} \wedge \cdots \wedge \boldsymbol{e}_{\alpha_N}, \tag{2-4}$$

where $\{\alpha_1, \dots, \alpha_N\}$ has positive signature. In particular, in what follows we will systematically identify $\Lambda_{N-1}(\mathbb{R}^N)$ with $\mathbb{R}^N$. We will use the following well-known formulas:

$$*(*\boldsymbol{\lambda}) = (-1)^{k(N-k)}\boldsymbol{\lambda} \quad \text{for all } \boldsymbol{\lambda} \in \Lambda_k(\mathbb{R}^N) \tag{2-5}$$

(see, for example, [Darling 1994, (1.64)]) and

$$\boldsymbol{a} \wedge *(\boldsymbol{b} \wedge \boldsymbol{c}) = (\boldsymbol{a} \cdot \boldsymbol{c}) * \boldsymbol{b} - (\boldsymbol{a} \cdot \boldsymbol{b}) * \boldsymbol{c} \quad \text{for all } \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^N \tag{2-6}$$

(see, for example, [Darling 1994, Table 1.2]). It follows from (2-3), (2-6), and (2-5) that

$$|\boldsymbol{b}|^2\boldsymbol{a} = (\boldsymbol{a} \cdot \boldsymbol{b})\boldsymbol{b} - *(*(\boldsymbol{a} \wedge \boldsymbol{b}) \wedge \boldsymbol{b}) \quad \text{for all } \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N. \tag{2-7}$$

Introducing the norm

$$|\boldsymbol{\lambda}|_k = \Big(\sum_{\alpha \in I(k,N)} |\lambda_\alpha|^2 \Big)^{1/2}, \quad \text{where } \boldsymbol{\lambda} = \sum_{\alpha \in I(k,N)} \lambda_\alpha \boldsymbol{e}_\alpha \tag{2-8}$$

and using (2-4), it is immediate to see that

$$|*\boldsymbol{\lambda}|_{N-k} = |\boldsymbol{\lambda}|_k \quad \text{for any } \boldsymbol{\lambda} \in \Lambda_k(\mathbb{R}^N). \tag{2-9}$$

Finally, we recall that, given $\boldsymbol{\lambda} \in \Lambda_k(\mathbb{R}^N)$ and $\boldsymbol{\eta} \in \Lambda_p(\mathbb{R}^N)$ such that one of them is a generator, we have

$$|\boldsymbol{\lambda} \wedge \boldsymbol{\eta}|_{k+p} \le |\boldsymbol{\lambda}|_k |\boldsymbol{\eta}|_p; \tag{2-10}$$

see [Federer 1969, p. 32].

2B. ***Vector-valued functions.*** Let $(X, \|\cdot\|)$ a Banach space with dual X' and let $U \subset \mathbb{R}^d$ be a bounded open set endowed with the Lebesgue measure $\mathscr{L}^d$. We denote by $\langle \cdot, \cdot \rangle$ the pairing between X and X'. A function $u : U \to X$ is called *simple* if there exist $x_1, \dots, x_n \in X$ and $U_1, \dots, U_n$ $\mathscr{L}^m$-measurable subsets of U such that $u = \sum_{i=1}^n x_i \chi_{U_i}$. The function u is called *strongly measurable* if there exists a sequence of simple functions $\{u_n\}$ such that $\|u_n(x) - u(x)\| \to 0$ as $n \to +\infty$ for almost all $x \in U$. If $1 \le p < \infty$, then $L^p(U; X)$ stands for the space of (equivalence classes of) strongly measurable functions $u : U \to X$ with

$$\|u\|_p := \left(\int_U \|u(x)\|^p \, dx \right)^{1/p} < \infty.$$

Endowed with this norm, $L^p(U; X)$ is a Banach space. For $p = \infty$, the symbol $L^\infty(U; X)$ stands for the space of (equivalence classes of) strongly measurable functions $u : U \to X$ such that

$$\|u\|_\infty := \operatorname{esssup}\{\|u(x)\| : x \in U\} < \infty.$$

If $U = (0, T)$, we write $L^p(0, T; X) = L^p((0, T); X)$. For $1 \le p < \infty$, $L^{p'}(0, T; X')$ $(1/p + 1/p' = 1)$ is isometric to a subspace of $(L^p(0, T; X))'$, with equality if and only if X' has the Radon–Nikodým property; see, for instance, [Diestel and Uhl 1977].

We consider the vector space $\mathscr{D}(U; X) := C_0^\infty(U; X)$, endowed with the topology for which a sequence $\varphi_n \to 0$ as $n \to +\infty$ if there exists $K \subset U$ compact such that $\operatorname{supp}(\varphi_n) \subset K$ for any $n \in \mathbb{N}$ and $D^\alpha \varphi_n \to 0$ uniformly on K as $n \to +\infty$ for all multiindexes α. We denote by $\mathscr{D}'(U; X)$ the space of distributions on U with values in X, that is, the set of all linear continuous maps $T : \mathscr{D}(U; X) \to \mathbb{R}$. As is well known, $L^p(U; X) \subset \mathscr{D}'(U; X)$ through the standard continuous injection. Given $T \in \mathscr{D}'(U; X)$, the distributional derivative of T is defined by

$$\langle D_i T, \varphi \rangle := -\langle T, \partial_i \varphi \rangle \quad \text{for any } \varphi \in \mathscr{D}(U; X) \text{ and any } i \in \{1, \dots, d\}. \tag{2-11}$$

General notations for matrices. If $\boldsymbol{A} = (a_i^\ell)$ is an $N \times m$ matrix, we write $a^\ell = (a_1^\ell, \dots a_m^\ell)$ for $1 \le \ell \le N$ and $\boldsymbol{a}_i = (a_i^1, \dots, a_i^N)$ for $1 \le i \le m$. If $\boldsymbol{B} = (b_i^\ell)$ is also an $N \times m$ matrix, we let

$$\boldsymbol{A} : \boldsymbol{B} = \sum_{\ell=1}^N \sum_{i=1}^m a_i^\ell b_i^\ell \quad \text{and} \quad |\boldsymbol{A}| = (\boldsymbol{A} : \boldsymbol{A})^{1/2} = \left(\sum_{\ell=1}^N \sum_{i=1}^m (a_i^\ell)^2 \right)^{1/2}.$$

Given $\boldsymbol{A} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_m) \in \mathbb{R}^{N \times m}$ and $\boldsymbol{b} \in \mathbb{R}^N$, we let

$$\boldsymbol{A} \wedge \boldsymbol{b} := (\boldsymbol{a}_1 \wedge \boldsymbol{b}, \dots, \boldsymbol{a}_m \wedge \boldsymbol{b}),$$
$$*(\boldsymbol{A} \wedge \boldsymbol{b}) := (*(\boldsymbol{a}_1 \wedge \boldsymbol{b}), \dots, *(\boldsymbol{a}_m \wedge \boldsymbol{b})).$$

2C. ***Functions of bounded variation.*** A vector-field $\boldsymbol{u} \in L^1(\Omega; \mathbb{R}^N)$ has bounded variation, and we write $\boldsymbol{u} \in \mathrm{BV}(\Omega; \mathbb{R}^N)$, if there is an $N \times m$ matrix $D\boldsymbol{u}$, whose components $D_i u^\ell$ are finite Radon measures, such that

$$\sum_{\ell=1}^{N}\int_\Omega u^\ell \operatorname{div}\varphi^\ell\,dx = -\sum_{\ell=1}^{N}\sum_{i=1}^{m}\int_\Omega \varphi_i^\ell\, dD_i u^\ell \quad \text{for all } \varphi\in (C_0^1(\Omega;\mathbb{R}^N))^m.$$

Its variation measure $|D\boldsymbol{u}|$ is a finite Radon measure defined on open sets $U\subseteq\Omega$ by

$$|D\boldsymbol{u}|(U) = \sup\left\{\sum_{\ell=1}^{N}\int_U u^\ell \operatorname{div}\varphi^\ell\,dx : \varphi\in (C_0^1(U;\mathbb{R}^N))^m,\ \|\varphi\|_\infty\le 1\right\}.$$

The matrix-valued Radon measure $D\boldsymbol{u}$ is decomposed into three mutually orthogonal measures (see [Ambrosio et al. 2000; Evans and Gariepy 1992; Ziemer 1989]):

$$D\boldsymbol{u} = \nabla\boldsymbol{u}\mathscr{L}^m + D^c\boldsymbol{u} + D^j\boldsymbol{u},$$

where $\nabla\boldsymbol{u}$ denotes the Radon–Nikodým derivative of $D\boldsymbol{u}$ with respect to $\mathscr{L}^m$. The *Cantor part* $D^c\boldsymbol{u}$ is supported on the set of *Lebesgue points of* $\boldsymbol{u}$, $\Omega\setminus S_{\boldsymbol{u}}$, that is, those points $x\in\Omega$ for which there exists $\tilde{\boldsymbol{u}}(x)\in\mathbb{R}^N$ such that

$$\lim_{\rho\downarrow 0}\frac{1}{\mathscr{L}^m(B_\rho(x))}\int_{B_\rho(x)}|\boldsymbol{u}(y)-\tilde{\boldsymbol{u}}(x)|\,dy = 0.$$

The *jump part* $D^j\boldsymbol{u}$ is supported on the set of *approximate jump points of* $\boldsymbol{u}$, $J_{\boldsymbol{u}}$, that is, those points $x\in\Omega$ for which there exist $\boldsymbol{u}_+(x)\neq\boldsymbol{u}_-(x)\in\mathbb{R}^N$ and $\nu_{\boldsymbol{u}}(x)\in\mathbb{S}^{m-1}$ such that

$$\lim_{\rho\downarrow 0}\frac{1}{\mathscr{L}^m(B_\rho^\pm(x,\nu_{\boldsymbol{u}}(x)))}\int_{B_\rho^\pm(x,\nu_{\boldsymbol{u}}(x))}|\boldsymbol{u}(y)-\boldsymbol{u}_\pm(x)|\,dy = 0,$$

where

$$B_\rho^\pm(x,\nu_{\boldsymbol{u}}(x)) = \{y\in B_\rho(x) : \langle y-x,\nu_{\boldsymbol{u}}(x)\rangle \gtrless 0\}.$$

The jump set $J_{\boldsymbol{u}}$ is a Borel subset of $S_{\boldsymbol{u}}$ that satisfies $\mathscr{H}^{m-1}(S_{\boldsymbol{u}}\setminus J_{\boldsymbol{u}})=0$. The *precise representative* $\boldsymbol{u}^*:\Omega\setminus(S_{\boldsymbol{u}}\setminus J_{\boldsymbol{u}})\to\mathbb{R}^N$ of $\boldsymbol{u}$ is defined to be equal to $\tilde{\boldsymbol{u}}$ on $\Omega\setminus S_{\boldsymbol{u}}$ and equal to $(\boldsymbol{u}_-+\boldsymbol{u}_+)/2$ on $J_{\boldsymbol{u}}$. In what follows, we identify $\boldsymbol{u}=\tilde{\boldsymbol{u}}=\boldsymbol{u}^*$ on $\Omega\setminus S_{\boldsymbol{u}}$.

2D. ***A generalized Green's formula.*** Let

$$X_{\mathcal{M}}(\Omega) = \{z\in L^\infty(\Omega;\mathbb{R}^m) : \operatorname{div} z\in\mathcal{M}(\Omega)\}$$

and

$$\mathcal{M}_{\mathscr{H}}(\Omega;\mathbb{R}^N) := \{\boldsymbol{\mu}\in\mathcal{M}(\Omega;\mathbb{R}^N) : |\boldsymbol{\mu}|(B)=0 \text{ for any Borel set } B\subset\Omega : \mathscr{H}^{m-1}(B)=0\}.$$

In [Anzellotti 1983, Theorem 1.2] (see also [Andreu-Vaillo et al. 2004; Chen and Frid 1999]), the weak trace on $\partial\Omega$ of the normal component of $z\in X_{\mathcal{M}}(\Omega)$ is defined. Namely, it is proved that there exists a linear operator $[\,\cdot\,,\nu]:X_{\mathcal{M}}(\Omega)\to L^\infty(\partial\Omega)$ such that $\|[z,\nu]\|_{L^\infty(\partial\Omega)}\le\|z\|_{L^\infty(\Omega)}$ for all $z\in X_{\mathcal{M}}(\Omega)$ and $[z,\nu]$ coincides with the pointwise trace of the normal component if z is smooth:

$$[z,\nu](x) = z(x)\cdot\nu(x) \quad \text{for all } x\in\partial\Omega \text{ if } z\in C^1(\bar\Omega,\mathbb{R}^m).$$

It follows from [Chen and Frid 1999, Proposition 3.1] or [Ambrosio et al. 2005, Proposition 3.4] that

$$\operatorname{div} z \in \mathcal{M}_{\mathcal{H}}(\Omega) \quad \text{for all } z \in X_{\mathcal{M}}(\Omega). \tag{2-12}$$

Therefore, given $z \in X_{\mathcal{M}}(\Omega)$ and $u \in BV(\Omega) \cap L^\infty(\Omega)$, the functional $(z, Du) \in \mathcal{D}'(\Omega)$ given by

$$\langle (z, Du), \varphi \rangle := -\int_\Omega u^* \varphi \, d(\operatorname{div} z) - \int_\Omega u z \nabla \varphi \, dx \tag{2-13}$$

is well defined, and the following holds (in [Caselles 2011], see Lemma 5.1, Theorem 5.3, and the discussion after Lemma 5.4):

Lemma 2.1. *Let $z \in X_{\mathcal{M}}(\Omega)$ and $u \in BV(\Omega) \cap L^\infty(\Omega)$. Then the functional $(z, Du) \in \mathcal{D}'(\Omega)$ defined by (2-13) is a Radon measure which is absolutely continuous with respect to $|Du|$. Furthermore,*

$$\int_\Omega u^* \, d(\operatorname{div} z) + (z, Du)(\Omega) = \int_{\partial\Omega} [z, \nu] u \, d\mathcal{H}^{m-1}$$

and

$$\operatorname{div}(zu) = u^* \operatorname{div} z + (z, Du) \quad \text{as measures.}$$

We will use the vector-valued version of Lemma 2.1. To this aim, we introduce the space

$$\boldsymbol{X}^N_{\mathcal{M}}(\Omega) = \{\boldsymbol{Z} = (z^1, \dots, z^N)^T : z^\ell \in X_{\mathcal{M}}(\Omega) \text{ for } \ell = 1, \dots, N\}.$$

Given $\boldsymbol{Z} \in \boldsymbol{X}^N_{\mathcal{M}}(\Omega)$ and $\boldsymbol{u} \in \mathrm{BV}(\Omega; \mathbb{R}^N) \cap L^\infty(\Omega; \mathbb{R}^N)$, we use the notation

$$\begin{aligned} \operatorname{div} \boldsymbol{Z} &:= (\operatorname{div} z^1, \dots, \operatorname{div} z^N), \\ [\boldsymbol{Z}, \nu] &:= ([z^1, \nu], \dots, [z^N, \nu]), \\ \boldsymbol{Z} : D\boldsymbol{u} &:= \sum_{\ell=1}^N (z^\ell, Du^\ell). \end{aligned}$$

Then, as an immediate consequence of (2-12) and Lemma 2.1, we have:

Corollary 2.2. *Let $\boldsymbol{Z} \in \boldsymbol{X}^N_{\mathcal{M}}(\Omega)$. Then*

$$\operatorname{div} \boldsymbol{Z} \in \mathcal{M}_{\mathcal{H}}(\Omega; \mathbb{R}^N).$$

Furthermore, for any $\boldsymbol{u} \in \mathrm{BV}(\Omega; \mathbb{R}^N) \cap L^\infty(\Omega; \mathbb{R}^N)$, $\boldsymbol{Z} : D\boldsymbol{u}$ is a Radon measure which is absolutely continuous with respect to $|D\boldsymbol{u}|$,

$$\int_\Omega \boldsymbol{u}^* \cdot \, d(\operatorname{div} \boldsymbol{Z}) + (\boldsymbol{Z} : D\boldsymbol{u})(\Omega) = \int_{\partial\Omega} [\boldsymbol{Z}, \nu] \cdot \boldsymbol{u} \, d\mathcal{H}^{m-1} \tag{2-14}$$

and

$$\operatorname{div}(\boldsymbol{Z}^T \boldsymbol{u}) = \boldsymbol{u}^* \cdot \operatorname{div} \boldsymbol{Z} + \boldsymbol{Z} : D\boldsymbol{u} \quad \text{as measures.} \tag{2-15}$$

2E. ***Multivector fields.*** Let $U \subset \mathbb{R}^d$. A multivector distribution in U is a linear continuous map $\boldsymbol{\lambda} \in \mathscr{D}'(U; \Lambda_k(\mathbb{R}^N))$ (see Section 2B). It may be expressed in terms of the basis (3) as

$$\boldsymbol{\lambda} = \sum_{\alpha \in I(k,N)} \lambda_\alpha \boldsymbol{e}_\alpha \quad \text{with } \lambda_\alpha \in \mathscr{D}'(U; \mathbb{R}^N) \text{ for any } \alpha \in I(k, N).$$

Then, according to (2-11),

$$D_i \boldsymbol{\lambda} = \sum_{\alpha \in I(k;N)} D_i \lambda_\alpha \boldsymbol{e}_\alpha \quad \text{for any } i \in \{1, \dots, d\}. \tag{2-16}$$

From (2-16), the following two identities are easily seen to hold for $k, p \in \mathbb{N}$ and $i \in \{1, \dots, d\}$:

$$D_i(\boldsymbol{\lambda} \wedge \boldsymbol{\eta}) = D_i \boldsymbol{\lambda} \wedge \boldsymbol{\eta} + \boldsymbol{\lambda} \wedge D_i \boldsymbol{\eta} \tag{2-17}$$

for any $\boldsymbol{\lambda} \in L^2(U; \Lambda_k(\mathbb{R}^N))$ such that $D_i \boldsymbol{\lambda} \in L^2(U; \Lambda_k(\mathbb{R}^N))$ and any $\boldsymbol{\eta} \in L^2(U; \Lambda_p(\mathbb{R}^N))$ such that $D_i \boldsymbol{\eta} \in L^2(U; \Lambda_p(\mathbb{R}^N))$;

$$*(D_i \boldsymbol{\lambda}) = D_i(*\boldsymbol{\lambda}) \quad \text{for any } \boldsymbol{\lambda} \in \mathscr{D}'(U; \Lambda_k(\mathbb{R}^N)). \tag{2-18}$$

For any $k \in \mathbb{N}$, $(\Lambda_k(\mathbb{R}^N))^m$ is a Banach space. We use the norm

$$\|\mathscr{A}\| := \left(\sum_{i=1}^m |\mathscr{A}_i|_k^2 \right)^{1/2} \quad \text{for } \mathscr{A} = (\mathscr{A}_1, \dots, \mathscr{A}_m)$$

with $|\cdot|_k$ given by (2-8).

We will now state and prove the analogue of Corollary 2.2 for a multivector field

$$\mathscr{A} = (\mathscr{A}_1, \dots, \mathscr{A}_m) \in L^\infty(\Omega; (\Lambda_k(\mathbb{R}^N))^m).$$

We define

$$\operatorname{div} \mathscr{A} := \sum_{i=1}^m D_i(\mathscr{A}_i). \tag{2-19}$$

Square-integrability of $\operatorname{div} \mathscr{A}$ suffices for our purposes. Hence, we introduce the space

$$\boldsymbol{X}_2(\Omega; \Lambda_{N-2}(\mathbb{R}^N)) := \{\mathscr{A} \in L^\infty(\Omega; (\Lambda_{N-2}(\mathbb{R}^N))^m) : \operatorname{div} \mathscr{A} \in L^2(\Omega; \Lambda_{N-2}(\mathbb{R}^N))\}.$$

The following holds:

Lemma 2.3. *Let $\mathscr{A} \in \boldsymbol{X}_2(\Omega; \Lambda_{N-2}(\mathbb{R}^N))$ and consider $\boldsymbol{u} \in \mathrm{BV}(\Omega; \mathbb{R}^N) \cap L^2(\Omega; \mathbb{R}^N)$. Then the functional $*(\mathscr{A} \wedge D\boldsymbol{u}) : \mathscr{D}(\Omega; \mathbb{R}^N) \to \mathbb{R}$ defined by*

$$\langle *(\mathscr{A} \wedge D\boldsymbol{u}), \Phi \rangle := -\int_\Omega *(\operatorname{div} \mathscr{A} \wedge \boldsymbol{u}) \cdot \Phi \, dx - \sum_{i=1}^m \int_\Omega *(\mathscr{A}_i \wedge \boldsymbol{u}) \cdot \partial_i \Phi \, dx \tag{2-20}$$

is an $\mathbb{R}^N$-valued Radon measure on Ω, absolutely continuous with respect to $|D\boldsymbol{u}|$, with

$$|*(\mathscr{A} \wedge D\boldsymbol{u})|(B) \le \|\mathscr{A}\|_\infty |D\boldsymbol{u}|(B) \quad \textit{for any Borel set } B \subseteq \Omega. \tag{2-21}$$

Furthermore,

$$\operatorname{div}(*(\mathscr{A}\wedge \boldsymbol{u})) = *(\mathscr{A}\wedge D\boldsymbol{u}) + *(\operatorname{div}\mathscr{A}\wedge \boldsymbol{u})\mathscr{L}^m \quad \textit{as measures}. \tag{2-22}$$

Proof. Since Ω has compact Lipschitz boundary, it follows from [Ambrosio et al. 2000, Theorem 3.21, Remark 3.22, and Corollary 3.80] that the sequence $\boldsymbol{u}_n := (T\boldsymbol{u}) \star \rho_n \in C^\infty(\bar{\Omega})$ (here T denotes an extension operator) is such that $\boldsymbol{u}_n \rightharpoonup \boldsymbol{u}$ in $\mathrm{BV}(\Omega;\mathbb{R}^N)$, $\int_\Omega |\nabla \boldsymbol{u}_n|\,dx \to |D\boldsymbol{u}|(\Omega)$, and $\boldsymbol{u}_n \to \boldsymbol{u}^*$ $\mathscr{H}^{m-1}$-a.e. in $\bar{\Omega}$. Furthermore, by construction and since $\boldsymbol{u}\in L^2(\Omega;\mathbb{R}^N)$, we have $\boldsymbol{u}_n\to\boldsymbol{u}$ in $L^2(\Omega;\mathbb{R}^N)$. Then

$$\langle *(\mathscr{A}\wedge D\boldsymbol{u}),\Phi\rangle \overset{(2\text{-}20)}{=} -\lim_{n\to\infty}\left(\int_\Omega *(\operatorname{div}\mathscr{A}\wedge \boldsymbol{u}_n)\cdot\Phi\,dx + \sum_{i=1}^m \int_\Omega *(\mathscr{A}_i\wedge \boldsymbol{u}_n)\cdot\partial_i\Phi\,dx\right).$$

Integrating by parts and using (2-18), we obtain

$$\begin{aligned}\langle *(\mathscr{A}\wedge D\boldsymbol{u}),\Phi\rangle &= -\lim_{n\to\infty}\left(\int_\Omega *(\operatorname{div}\mathscr{A}\wedge \boldsymbol{u}_n)\cdot\Phi\,dx - \sum_{i=1}^m\int_\Omega *(\partial_i(\mathscr{A}_i\wedge\boldsymbol{u}_n))\cdot\Phi\,dx\right)\\ &\overset{(2\text{-}17)}{=} \lim_{n\to\infty}\sum_{i=1}^m\int_\Omega *(\mathscr{A}_i\wedge\partial_i\boldsymbol{u}_n)\cdot\Phi\,dx.\end{aligned}$$

Therefore, applying the Hölder and Cauchy–Schwarz inequalities,

$$\begin{aligned}|\langle *(\mathscr{A}\wedge D\boldsymbol{u}),\Phi\rangle| &\overset{(2\text{-}9)}{\le} \|\Phi\|_\infty \lim_{n\to\infty}\sum_{i=1}^m\int_\Omega |\mathscr{A}_i\wedge\partial_i\boldsymbol{u}_n|_{N-1}\,dx\\ &\overset{(2\text{-}10)}{\le} \|\Phi\|_\infty \lim_{n\to\infty}\int_\Omega\sum_{i=1}^m |\partial_i\boldsymbol{u}_n||\mathscr{A}_i|_{N-2}\,dx\\ &\le \|\Phi\|_\infty \lim_{n\to\infty}\int_\Omega |\nabla\boldsymbol{u}_n|\left(\sum_{i=1}^m|\mathscr{A}_i|_{N-2}^2\right)^{1/2}dx\\ &\le \|\Phi\|_\infty\|\mathscr{A}\|_\infty\lim_{n\to\infty}\int_\Omega|\nabla\boldsymbol{u}_n|\,dx\\ &= \|\Phi\|_\infty\|\mathscr{A}\|_\infty|D\boldsymbol{u}|(\Omega).\end{aligned}$$

The arbitrariness of Φ completes the proof of (2-21). It follows from (2-20) and (2-19) that $\operatorname{div}(*(\mathscr{A}\wedge\boldsymbol{u}))$ is also an $\mathbb{R}^N$-valued Radon measure in Ω, and (2-22) follows from (2-19). □

2F. ***Lower semicontinuity of integral functionals over $W^{1,1}(\Omega;\mathbb{S}^{N-1}_+)$.*** Let $f:\Omega\times\mathbb{S}^{N-1}_+\to\mathbb{R}_+$ and consider the energy functional defined in $L^1(\Omega;\mathbb{S}^{N-1}_+)$ by

$$\mathscr{F}_f(\boldsymbol{v}) := \begin{cases}\displaystyle\int_\Omega f(x,\boldsymbol{v}(x))|\nabla\boldsymbol{v}(x)|\,dx & \text{if } \boldsymbol{v}\in W^{1,1}(\Omega;\mathbb{S}^{N-1}_+),\\ +\infty & \text{otherwise.}\end{cases}$$

The purpose of this section is to restate, to the extent we need in the present setting, a few lower semicontinuity results obtained in [Fonseca and Rybka 1992; Alicandro et al. 2007] (see also [Giaquinta

and Mucci 2006] for related results when the target space is a general manifold). We consider the following hypotheses for f:

(H1) f is continuous and nonnegative;

(H2) (uniform boundedness) a positive constant C_1 exists such that

$$|f(x,\boldsymbol{s})| \le C_1 \quad \text{for all } (x,\boldsymbol{s}) \in \Omega \times \mathbb{S}^{N-1}_+;$$

(H3) for every compact set $U \subset \Omega$, there exist a continuous function ω, with $\omega(0) = 0$, such that

$$|f(x,\boldsymbol{s}) - f(x',\boldsymbol{s}')| = \omega(|x-x'| + |\boldsymbol{s}-\boldsymbol{s}'|) \quad \text{for all } (x,\boldsymbol{s}),\,(x',\boldsymbol{s}') \in U \times \mathbb{S}^{N-1}_+.$$

For $\varsigma \in \mathbb{R}^m$ such that $|\varsigma| = 1$, we define $Q_\varsigma := R_\varsigma\big[-\frac{1}{2},\frac{1}{2}\big]^m$, where R_ς denotes a rotation such that $R_\varsigma e_m = \varsigma$. Given $\boldsymbol{a},\boldsymbol{b} \in \mathbb{S}^{N-1}_+$, we set

$$K_f(x,\boldsymbol{a},\boldsymbol{b},\varsigma) := \inf\left\{\int_{Q_\varsigma} f(x,\boldsymbol{v}(y))|\nabla \boldsymbol{v}(y)|\,dy : \boldsymbol{v} \in \mathcal{P}(\boldsymbol{a},\boldsymbol{b},\varsigma)\right\}, \tag{2-23}$$

where

$$\mathcal{P}(\boldsymbol{a},\boldsymbol{b},\varsigma) := \left\{\boldsymbol{v} \in W^{1,1}(Q_\varsigma;\mathbb{S}^{N-1}_+) : \boldsymbol{v}(x) = \boldsymbol{a} \text{ if } x\cdot\varsigma = -\tfrac{1}{2},\ \boldsymbol{v}(x) = \boldsymbol{b} \text{ if } x\cdot\varsigma = \tfrac{1}{2}\right\}. \tag{2-24}$$

Lemma 2.4. *Assume* (H1). *Then*

$$K_f(x,\boldsymbol{a},\boldsymbol{b},\varsigma) = \inf\left\{\int_0^1 f(x,\boldsymbol{\gamma}(t))|\dot{\boldsymbol{\gamma}}(t)|\,dt : \boldsymbol{\gamma} \in W^{1,1}((0,1);\mathbb{S}^{N-1}_+),\ \boldsymbol{\gamma}(0) = \boldsymbol{a},\ \boldsymbol{\gamma}(1) = \boldsymbol{b}\right\}. \tag{2-25}$$

The proof of Lemma 2.4 is identical to that of [Fonseca and Rybka 1992, Proposition 2.6], where the same result has been proved (under more general assumptions on the energy density) when the target space is $\mathbb{R}^N$ rather than $\mathbb{S}^{N-1}_+$. Therefore we omit it.

In order to obtain a lower bound on the lower semicontinuous envelope of $\mathcal{F}_f$, in particular of its jump part, one needs an approximation lemma which relates a generic sequence in $W^{1,1}(Q_\varsigma;\mathbb{S}^{N-1}_+)$, converging to a step function, to a nongeneric one in $\mathcal{P}(\boldsymbol{a},\boldsymbol{b},\varsigma)$:

Lemma 2.5. *Assume* (H1) *and* (H2).

Let $\boldsymbol{a},\boldsymbol{b} \in \mathbb{S}^{N-1}_+$ *and let* $\boldsymbol{v}_n \in W^{1,1}(Q_\varsigma;\mathbb{S}^{N-1}_+)$ *such that* $\boldsymbol{v}_n \to \boldsymbol{u}_0$ *in* $L^1(Q_\varsigma;\mathbb{S}^{N-1}_+)$, *where*

$$\boldsymbol{u}_0(x) := \begin{cases} \boldsymbol{b} & \text{if } \langle x,\varsigma\rangle \ge 0,\\ \boldsymbol{a} & \text{if } \langle x,\varsigma\rangle < 0.\end{cases}$$

Then a sequence $\boldsymbol{w}_n \in \mathcal{P}(\boldsymbol{a},\boldsymbol{b},\varsigma)$ *exists such that* $\boldsymbol{w}_n \to \boldsymbol{u}_0$ *in* $L^1(Q_\varsigma;\mathbb{S}^{N-1}_+)$ *and*

$$\liminf_{n\to\infty}\int_{Q_\varsigma} f(x,\boldsymbol{v}_n)|\nabla \boldsymbol{v}_n|\,dx \ge \limsup_{n\to\infty}\int_{Q_\varsigma} f(x,\boldsymbol{w}_n)|\nabla \boldsymbol{w}_n|\,dx.$$

Lemma 2.5 may be proved following line by line that of [Alicandro et al. 2007, Lemma 5.2], where the same result was proved (under more general assumptions on the energy density) when the target space is $\mathbb{S}^{N-1}$, and therefore we omit it. We just mention that the proof may in fact be simplified in the present

setting by using the standard projection onto $\mathbb{S}^{N-1}_+$ (see estimate (3-23) and Lemma 3.9 below for a related approximation result).

Let $\mathcal{G}_f$ be the functional defined in $\mathrm{BV}(\Omega; \mathbb{S}^{N-1}_+)$ by

$$\mathcal{G}_f(\boldsymbol{v}) := \int_\Omega f(x, \boldsymbol{v})|\nabla \boldsymbol{v}|\, dx + \int_{J(\boldsymbol{v})} K_f(x, \boldsymbol{v}_-, \boldsymbol{v}_+, \nu_{\boldsymbol{v}})\, d\mathcal{H}^{m-1} + \int_\Omega f(x, \boldsymbol{v})\, d|D^c \boldsymbol{v}|$$

(and $+\infty$ elsewhere). Under an additional coercivity assumption on f, and when the target space is $\mathbb{S}^{N-1}$, in [Alicandro et al. 2007, Proposition 5.1] it is proved that $\mathcal{G}_f$ coincides with the lower semicontinuous envelope of $\mathcal{F}_f$ with respect to the L^1-convergence. Of course, coercivity is crucial for the upper bound in that it guarantees that *any* sequence along which $\mathcal{G}_f$ is bounded has a convergent subsequence. However, it may be dropped when only a lower bound is needed, provided it is a priori known that a sequence has good convergence properties:

Proposition 2.6. *Let f satisfy (H1)–(H3) and let $\boldsymbol{v}_n \in W^{1,1}(\Omega; \mathbb{S}^{N-1}_+)$ such that $\boldsymbol{v}_n \rightharpoonup \boldsymbol{v} \in \mathrm{BV}(\Omega; \mathbb{S}^{N-1}_+)$ and $\boldsymbol{v}_n \to \boldsymbol{v}$ in $L^1(\Omega; \mathbb{S}^{N-1}_+)$. Then*

$$\mathcal{G}_f(\boldsymbol{v}) \le \liminf_{n\to\infty} \mathcal{F}_f(\boldsymbol{v}_n).$$

Given Lemma 2.5, the proof follows line by line that of [Alicandro et al. 2007, Proposition 5.1], and the difference between the target spaces ($\mathbb{S}^{N-1}$ versus $\mathbb{S}^{N-1}_+$) is harmless. Therefore we omit it.

3. Existence of solutions

In this section we introduce the notion of solutions to (1-4) and we prove their existence.

As is mentioned in Section 2C, on its jump set $J_{\boldsymbol{u}}$ a function $\boldsymbol{u} \in \mathrm{BV}(\Omega; \mathbb{R}^N)$ has a jump discontinuity between two distinct values, $\boldsymbol{u}_+$ and $\boldsymbol{u}_-$, and the value of the precise representative of $\boldsymbol{u}$ is given by $(\boldsymbol{u}_+ + \boldsymbol{u}_-)/2$. Note that $(\boldsymbol{u}_+ + \boldsymbol{u}_-)/2$ is the midpoint of the segment which connects $\boldsymbol{u}_+$ and $\boldsymbol{u}_-$. In this sense, $(\boldsymbol{u}_+ + \boldsymbol{u}_-)/2$ has natural counterparts in $\mathbb{S}^{N-1}$ endowed with the standard geodesic distance d_g on $\mathbb{S}^{N-1}$, the *geodesic midpoints*:

Definition 3.1. Let $\mathbb{A}$ be a geodesically convex subset of $\mathbb{S}^{N-1}$ and let $\boldsymbol{u}_-, \boldsymbol{u}_+ \in \mathbb{A}$. A point $\boldsymbol{u}_g \in \mathbb{A}$ is called a *geodesic midpoint* on $\mathbb{A}$ between $\boldsymbol{u}_-$ and $\boldsymbol{u}_+$ if

(i) $\boldsymbol{u}_g$ belongs to a greatest circle of $\mathbb{S}^{N-1}$ passing through $\boldsymbol{u}_-$ and $\boldsymbol{u}_+$, and

(ii) $d_g(\boldsymbol{u}_g, \boldsymbol{u}_-) = d_g(\boldsymbol{u}_g, \boldsymbol{u}_+)$.

In particular, when $\mathbb{A} = \mathbb{S}^{N-1}_+$, geodesic midpoints are uniquely determined:

$$\boldsymbol{u}_g = \frac{\boldsymbol{u}_- + \boldsymbol{u}_+}{|\boldsymbol{u}_- + \boldsymbol{u}_+|} \quad \text{for all } \boldsymbol{u}_-, \boldsymbol{u}_+ \in \mathbb{S}^{N-1}_+.$$

Thus we can introduce the notion of *geodesic representative of* $\boldsymbol{u} \in \mathrm{BV}(\Omega; \mathbb{S}^{N-1}_+)$:

Definition 3.2. Let $\boldsymbol{u} \in \mathrm{BV}(\Omega; \mathbb{S}^{N-1}_+)$. The *geodesic representative* $\boldsymbol{u}_g : \Omega \backslash (S_{\boldsymbol{u}} \backslash J_{\boldsymbol{u}}) \to \mathbb{S}^{N-1}_+$ of $\boldsymbol{u}$ is defined by

$$\boldsymbol{u}_g = \begin{cases} \boldsymbol{u}^* & \text{on } \Omega \backslash S_{\boldsymbol{u}}, \\ \boldsymbol{u}^*/|\boldsymbol{u}^*| & \text{on } J_{\boldsymbol{u}}. \end{cases}$$

Note that $\boldsymbol{u}_g \in \mathrm{BV}(\Omega; \mathbb{S}^{N-1}_+)$ since $\boldsymbol{u}_+$ and $\boldsymbol{u}_-$ are $\mathscr{H}^{m-1}$-measurable on $J_{\boldsymbol{u}}$; see [Ambrosio et al. 2000, Proposition 3.69]. Hence the following Radon measures are well defined:

$$|\boldsymbol{u}^*||D\boldsymbol{u}| := |\nabla \boldsymbol{u}|\mathscr{L}^m + |D^c \boldsymbol{u}| + |\boldsymbol{u}^*||\boldsymbol{u}_+ - \boldsymbol{u}_-|\mathscr{H}^{m-1} \llcorner_{J_{\boldsymbol{u}}}, \tag{3-1}$$

$$\boldsymbol{u}_g|D\boldsymbol{u}| := \boldsymbol{u}(|\nabla \boldsymbol{u}|\mathscr{L}^m + |D^c \boldsymbol{u}|) + \boldsymbol{u}_g|\boldsymbol{u}_+ - \boldsymbol{u}_-|\mathscr{H}^{m-1} \llcorner_{J_{\boldsymbol{u}}}. \tag{3-2}$$

Moreover, $\boldsymbol{u}_g|D\boldsymbol{u}| \in \mathscr{M}_{\mathscr{H}}(\Omega; \mathbb{R}^N)$ (see Section 2D).

Remark 3.3. As shown in the proof of Lemma 3.9, the projections onto $\mathbb{S}^{N-1}_+$ of the mollifications of $\boldsymbol{u}$ point-wise converge to $\boldsymbol{u}_g$ in Ω. In this sense, the geodesic representative $\boldsymbol{u}_g$ is a natural representative for BV-vector fields with values into $\mathbb{S}^{N-1}_+$.

We are now ready to introduce the concept of solution for (1-4).

Definition 3.4. Let $\mathbb{A} = \mathbb{S}^{N-1}_+$, $T > 0$, and $\boldsymbol{u}_0 \in \mathrm{BV}(\Omega; \mathbb{S}^{N-1}_+)$. A function

$$\boldsymbol{u} \in L^\infty(0, T; \mathrm{BV}(\Omega; \mathbb{R}^N)) \cap C(0, T; L^1(\Omega; \mathbb{R}^N)), \quad \boldsymbol{u}_t \in L^2(0, T; L^2(\Omega; \mathbb{R}^N))$$

is a solution to (1-4) in Q_T if $\boldsymbol{u}(0) = \boldsymbol{u}_0$, $\boldsymbol{u} \in \mathbb{S}^{N-1}_+$ a.e. in Q_T, and there exists a matrix-valued function $\boldsymbol{Z} \in L^\infty(Q_T, \mathbb{R}^{N\times m})$, with $\|\boldsymbol{Z}\|_\infty \le 1$ and $\boldsymbol{Z}(t) \in \boldsymbol{X}^N_{\mathscr{M}}(\Omega)$ for almost all $t \in (0, T)$, such that

$$\boldsymbol{u}_t(t) - \operatorname{div} \boldsymbol{Z}(t) = \boldsymbol{u}(t)_g|D\boldsymbol{u}(t)| \qquad \text{as measures for a.e. } t \in [0, T], \tag{3-3}$$

$$\boldsymbol{u}_t(t) \wedge \boldsymbol{u}(t) = \operatorname{div}(\boldsymbol{Z}(t) \wedge \boldsymbol{u}(t)) \qquad \text{in } L^2(\Omega; \Lambda_2(\mathbb{R}^N)) \text{ for a.e. } t \in [0, T], \tag{3-4}$$

$$\boldsymbol{Z}^T \boldsymbol{u} = 0 \qquad \text{a.e. in } Q_T, \tag{3-5}$$

and

$$[\boldsymbol{Z}(t), \nu] = 0 \quad \mathscr{H}^{m-1}\text{-a.e. on } \partial\Omega \text{ for a.e. } t \in [0, T]. \tag{3-6}$$

The next observation clarifies the concept of solution given in Definition 3.4.

Proposition 3.5. *Let $\boldsymbol{u}$ be a solution of (1-4) in the sense of Definition 3.4. Then*

$$\boldsymbol{Z}(t) : D\boldsymbol{u}(t) = |\boldsymbol{u}(t)^*||D\boldsymbol{u}(t)| \quad \textit{as measures for a.e. } t \in (0, T). \tag{3-7}$$

Proof. We take any $\varphi \in \mathscr{D}(\Omega)$. Then

$$\begin{aligned} \int_\Omega \varphi \, d(\boldsymbol{Z}(t) : D\boldsymbol{u}(t)) &\overset{(2\text{-}15)}{=} -\int_\Omega \varphi \boldsymbol{u}(t)^* \cdot d(\operatorname{div} \boldsymbol{Z}(t)) - \int_\Omega (\boldsymbol{Z}(t)^T \boldsymbol{u}(t)) \cdot \nabla\varphi \, dx \\ &\underset{(3\text{-}3)}{\overset{(3\text{-}5)}{=}} \int_\Omega \varphi \boldsymbol{u}^*(t) \cdot d(\boldsymbol{u}_t(t) + \boldsymbol{u}(t)_g|D\boldsymbol{u}(t)|) \\ &= \int_\Omega \varphi \boldsymbol{u}(t)^* \cdot d(\boldsymbol{u}(t)_g|D\boldsymbol{u}(t)|), \end{aligned}$$

where in the last line we have used the facts that $|\boldsymbol{u}(t)| = 1$, $\boldsymbol{u}_t \in L^2(Q_T; \mathbb{R}^N)$ and the fact that $\boldsymbol{u}(t)_g|D\boldsymbol{u}(t)| \in \mathcal{M}_{\mathcal{H}}(\Omega; \mathbb{R}^N)$ a.e. $t \in (0, T)$. Finally, by (3-1) we get

$$\int_\Omega \varphi\, d(\boldsymbol{Z}(t) : D\boldsymbol{u}(t)) = \int_\Omega \varphi\, d(|\nabla \boldsymbol{u}(t)|\mathcal{L}^m + |D^c(\boldsymbol{u}(t))|) + \int_{J_{\boldsymbol{u}(t)}} \varphi |\boldsymbol{u}(t)^*|\,|\boldsymbol{u}(t)_+ - \boldsymbol{u}(t)_-|\, d\mathcal{H}^{m-1}$$
$$= \int_\Omega \varphi\, d(|\boldsymbol{u}(t)^*|\,|D\boldsymbol{u}(t)|). \qquad \square$$

Our main result is the following existence theorem.

Theorem 3.6. *For any $T > 0$ and any $\boldsymbol{u}_0 \in \mathrm{BV}(\Omega; \mathbb{S}^{N-1}_+)$, there exists a solution $\boldsymbol{u}$ to (1-4) in the sense of Definition 3.4.*

To prove Theorem 3.6 we need to recall or establish several results. The first one follows as a particular case from [Barrett et al. 2008, Theorem 4.1, (4.24), and (4.25)] (with $\lambda = \boldsymbol{g} = 0$ and $p = 2$).

Proposition 3.7. *Let $\varepsilon > 0$, $T > 0$, and $\alpha > 0$. If $\boldsymbol{u}_0^\varepsilon \in W^{1,2}(\Omega; \mathbb{S}^{N-1})$, then there exists*

$$\boldsymbol{u}^\varepsilon \in L^\infty(0, T; W^{1,2}(\Omega; \mathbb{R}^N)) \cap W^{1,2}(0, T; L^2(\Omega; \mathbb{R}^N))$$

such that $\boldsymbol{u}^\varepsilon(0, \cdot) = \boldsymbol{u}_0^\varepsilon$,

$$|\boldsymbol{u}^\varepsilon| = 1 \quad \text{a.e. in } Q_T, \tag{3-8}$$

and $\boldsymbol{u}^\varepsilon$ is a weak solution to

$$\begin{cases} \boldsymbol{u}_t^\varepsilon = \operatorname{div} \boldsymbol{Z}^\varepsilon + \boldsymbol{\mu}^\varepsilon & \text{in } Q_T, \\ [\boldsymbol{Z}^\varepsilon, \nu] = 0 & \text{in } S_T, \end{cases} \tag{3-9}$$

where

$$\boldsymbol{Z}^\varepsilon = \varepsilon^\alpha \nabla \boldsymbol{u}^\varepsilon + \frac{\nabla \boldsymbol{u}^\varepsilon}{\sqrt{|\nabla \boldsymbol{u}^\varepsilon|^2 + \varepsilon^2}} \quad \text{and} \quad \boldsymbol{\mu}^\varepsilon = \varepsilon^\alpha \boldsymbol{u}^\varepsilon |\nabla \boldsymbol{u}^\varepsilon|^2 + \boldsymbol{u}^\varepsilon \frac{|\nabla \boldsymbol{u}^\varepsilon|^2}{\sqrt{|\nabla \boldsymbol{u}^\varepsilon|^2 + \varepsilon^2}} \tag{3-10}$$

in the sense that

$$\int_0^T \int_\Omega (\boldsymbol{u}_t^\varepsilon \cdot \boldsymbol{v} + \boldsymbol{Z}^\varepsilon : \nabla \boldsymbol{v} - \boldsymbol{\mu}^\varepsilon \cdot \boldsymbol{v})\, dx\, dt = 0 \quad \text{for all } \boldsymbol{v} \in C^1(\overline{Q}_T; \mathbb{R}^N). \tag{3-11}$$

Furthermore,

$$(\boldsymbol{Z}^\varepsilon)^T \boldsymbol{u}^\varepsilon = 0 \quad \text{a.e. in } Q_T, \tag{3-12}$$
$$\boldsymbol{u}_t^\varepsilon \cdot \boldsymbol{u}^\varepsilon = 0 \quad \text{a.e. in } Q_T, \tag{3-13}$$
$$\boldsymbol{u}_t^\varepsilon \wedge \boldsymbol{u}^\varepsilon = \operatorname{div}(\boldsymbol{Z}^\varepsilon \wedge \boldsymbol{u}^\varepsilon), \tag{3-14}$$

and

$$J_\alpha^\varepsilon(\boldsymbol{u}^\varepsilon(t)) + \int_0^t \int_\Omega |\boldsymbol{u}_t^\varepsilon|^2\, dx\, ds \le J_\alpha^\varepsilon(\boldsymbol{u}_0) \quad \text{for a.e. } t \in [0, T], \tag{3-15}$$

where the energy functional J_α^ε is defined as

$$J_\alpha^\varepsilon(\boldsymbol{v}) := \varepsilon^\alpha \int_\Omega |\nabla \boldsymbol{v}(x)|^2\, dx + \int_\Omega \sqrt{|\nabla \boldsymbol{v}(x)|^2 + \varepsilon^2}\, dx, \quad \boldsymbol{v} \in W^{1,2}(\Omega; \mathbb{R}^N),$$

and a positive ε-independent constant C exists such that

$$\|\operatorname{div} \boldsymbol{Z}^\varepsilon\|_{L^2(0,T;L^1(\Omega;\mathbb{R}^N))} \le C, \tag{3-16}$$

$$\|\operatorname{div}(\boldsymbol{Z}^\varepsilon \wedge \boldsymbol{u}^\varepsilon)\|_{L^2(0,T;L^2(\Omega;\Lambda_2(\mathbb{R}^N)))} \le C, \tag{3-17}$$

$$\varepsilon^{\alpha/2}\|\nabla \boldsymbol{u}^\varepsilon(t)\|_{L^\infty(0,T;L^2(\Omega;\mathbb{R}^{N\times m}))} \le C. \tag{3-18}$$

We next show that if $\boldsymbol{u}_0^\varepsilon$ takes values in the first hyperoctant, then $\boldsymbol{u}^\varepsilon$ does too:

Lemma 3.8. *If $\boldsymbol{u}_0^\varepsilon \in W^{1,2}(\Omega;\mathbb{S}_+^{N-1})$, then the weak solution to Problem (3-9) given by Proposition 3.7 verifies $\boldsymbol{u}^\varepsilon \in \mathbb{S}_+^{N-1}$ a.e. in Q_T.*

Proof. Let $(s)^- = \max\{0,-s\}$ and let $(\boldsymbol{u}^\varepsilon)^- = ((u^{\varepsilon,1})^-,\dots,(u^{\varepsilon,N})^-)$. Pick a sequence of smooth functions $\boldsymbol{v}_n$ such that $\boldsymbol{v}_n \to (\boldsymbol{u}^\varepsilon)^-$ in $L^2(0,T;W^{1,2}(\Omega)) \cap W^{1,2}(0,T;L^2(\Omega))$ as $n\to+\infty$. Choosing $\boldsymbol{v}=\boldsymbol{v}_n$ in (3-11) and passing to the limit as $n\to+\infty$, we obtain on the one hand

$$\int_0^T\int_\Omega (\boldsymbol{u}^\varepsilon)^- \cdot \boldsymbol{u}_t^\varepsilon \,dx\,dt = \int_0^T\int_\Omega \left(\varepsilon^\alpha + \frac{1}{\sqrt{\varepsilon^2+|\nabla \boldsymbol{u}^\varepsilon|^2}}\right)|\nabla(\boldsymbol{u}^\varepsilon)^-|^2(1-|(\boldsymbol{u}^\varepsilon)^-|^2)\,dx\,dt \ge 0. \tag{3-19}$$

On the other hand, since $\boldsymbol{u}^\varepsilon \in W^{1,2}(0,T;L^2(\Omega;\mathbb{R}^N))$,

$$0 \le \int_0^T\int_\Omega (\boldsymbol{u}^\varepsilon)^- \cdot \boldsymbol{u}_t^\varepsilon \,dx\,dt = \int_\Omega (|(\boldsymbol{u}_0)^-|^2 - |(\boldsymbol{u}^\varepsilon(T))^-|^2)\,dx = -\int_\Omega |(\boldsymbol{u}^\varepsilon(T))^-|^2\,dx, \tag{3-20}$$

hence the negative part of each component remains 0 for all times. □

Provided α is large enough, any function in $\mathrm{BV}(\Omega;\mathbb{S}_+^{N-1})$ can be approximated in $W^{1,2}(\Omega;\mathbb{S}_+^{N-1})$ in such a way that the initial energy is controlled.

Lemma 3.9. *Given $\boldsymbol{u}_0 \in \mathrm{BV}(\Omega;\mathbb{S}_+^{N-1})$ and $\alpha > m$, there exist $\boldsymbol{u}_0^\varepsilon \in W^{1,2}(\Omega;\mathbb{S}_+^{N-1})$ such that*

(i) *$\boldsymbol{u}_0^\varepsilon \to \boldsymbol{u}_0$ in $L^p(\Omega;\mathbb{R}^N)$ for all $p<\infty$ as $\varepsilon\to 0$,*

(ii) *$\boldsymbol{u}_0^\varepsilon \to (\boldsymbol{u}_0)_g$ $\mathscr{H}^{m-1}$-a.e. in Ω as $\varepsilon\to 0$,*

(iii) *$J_\alpha^\varepsilon(\boldsymbol{u}_0^\varepsilon) \to L < +\infty$ as $\varepsilon\to 0$.*

Proof. We will construct $\boldsymbol{u}_0^\varepsilon$ as the projection onto $\mathbb{S}_+^{N-1}$ of the convolution of a suitable extension $T\boldsymbol{u}_0$ of $\boldsymbol{u}_0$ with a standard mollifier. In order to do this, we proceed as in [Ambrosio et al. 2000, Proposition 3.21], to which we refer for further details; see also [Brezis 2011, Theorem 9.7].

Since $\overline{\Omega}$ is compact, there exists a finite collection $\{R_i\}_{i\in I}$ of open rectangles whose union B contains $\overline{\Omega}$, which satisfies the following property: for any $i\in I$, either

(a) $R_i \subset \Omega$ or

(b) $\partial\Omega \cap R_i$ is the graph of a Lipschitz function defined on one face L_i of R_i and the closure of $\partial\Omega\cap R_i$ intersects neither $\bar{L}_i$ nor the closure of the face opposite to L_i.

Let $\Omega_i = \Omega\cap R_i$. In case (b), up to a translation, a rotation, and a homothety, we have $R_i = L_i\times(-1,1)$ with Ω_i on the upper side of R_i (that is, $\Omega_i = \{x=(y,z): z>\phi_i(y)\}$). A vertical deformation $\varphi: R_i \to R_i$

exists such that $\varphi(\Omega_i) = R_i^+ = L_i \times (0,1)$ and both φ and its inverse are Lipschitz. Given $u \in \mathrm{BV}(\Omega)$, the operator $T_i : R_i \to \mathbb{R}$ is defined as the identity in case (a) and as

$$T_i(u) = T_i'(u \circ \varphi^{-1}) \circ \varphi, \quad \text{where } T_i'(u)(y,z) = u(y,|z|),$$

in case (b). Note that $|\boldsymbol{u}_0| = 1$ a.e. in Ω, the maps φ and its inverse are Lipschitz, and T_i' does not change the value of u. Hence

$$U_i := \{x \in R_i : |(T_i(u_0^1), \dots, T_i(u_0^N))| \neq 1\} \quad \text{has zero measure.}$$

Let $\{\eta_i\}_{i\in I}$ be a partition of unity relative to $\{R_i\}_{i\in I}$, that is, $\operatorname{supp}(\eta_i) \subset R_i$, $0 \le \eta_i \le 1$ for any $i \in I$ and there exists $r > 0$ such that $\sum_{i\in I} \eta_i \equiv 1$ in a neighborhood of $\overline{\Omega}$ containing $\Omega \oplus B_r$. We now define

$$T\boldsymbol{u}_0 : B = \bigcup_{i\in I} R_i \to \mathbb{R}^N, \quad T\boldsymbol{u}_0 := \Big(\sum_{i\in I} T_i(u_0^1)\eta_i, \dots, \sum_{i\in I} T_i(u_0^N)\eta_i\Big).$$

It is readily checked that $T \in \mathrm{BV}(\Omega \oplus B_r; \mathbb{R}^N)$. Now let $k > 0$ be the cardinality of I and $U = \bigcup_{i\in I} U_i$ (a set of zero measure). We observe that

$$|T\boldsymbol{u}_0(x)| \ge \frac{1}{k} \quad \text{for all } x \in \Omega \oplus B_r \setminus U. \tag{3-21}$$

Indeed, for each $x \in (\Omega \oplus B_r) \setminus U$, there exists $i(x) \in I$ such that $\eta_{i(x)}(x) \ge 1/k$: since each component of $\boldsymbol{u}_0$ is nonnegative and $x \notin U_{i(x)}$,

$$|T\boldsymbol{u}_0(x)|^2 \ge \frac{1}{k^2}\big((T_{i(x)}(u_0^1))^2 + \cdots + (T_{i(x)}(u_0^N))^2\big) = \frac{1}{k^2}.$$

Given $\varepsilon < r$, let $\rho_\varepsilon(x) := \varepsilon^{-m}\rho(x/\varepsilon)$ be a standard mollifier. As is well known (see, for example, [Ambrosio et al. 2000, Remark 3.22]) $T\boldsymbol{u}_0 \star \rho_\varepsilon$ converges to $T\boldsymbol{u}_0$ strictly in $\mathrm{BV}(\Omega;\mathbb{R}^N)$ and strongly in $L^1(\Omega;\mathbb{R}^N)$. Since $\|T\boldsymbol{u}_0 \star \rho_\varepsilon\|_\infty \le 1$, the last convergence upgrades to

$$T\boldsymbol{u}_0 \star \rho_\varepsilon \to T\boldsymbol{u}_0 \quad \text{in } L^p(\Omega;\mathbb{R}^N) \quad \text{for all } 1 \le p < \infty. \tag{3-22}$$

By (3-21) and since $(T(\boldsymbol{u}_0))^\ell \ge 0$ for $\ell = 1, \dots, N$, a direct computation shows that

$$|T\boldsymbol{u}_0 \star \rho_\varepsilon(x)| \ge \frac{1}{k\sqrt{N}} \quad \text{for all } x \in \Omega. \tag{3-23}$$

In addition, it follows from [Ambrosio et al. 2000, Corollary 3.80] that $T\boldsymbol{u}_0 \star \rho_\varepsilon \to (T\boldsymbol{u}_0)^* = \boldsymbol{u}_0^*$ pointwise in $\Omega \setminus (S_{\boldsymbol{u}_0} \setminus J_{\boldsymbol{u}_0})$. Together with (3-23), this implies that

$$\boldsymbol{u}_0^\varepsilon := \frac{T\boldsymbol{u}_0 \star \rho_\varepsilon}{|T\boldsymbol{u}_0 \star \rho_\varepsilon|} \to (\boldsymbol{u}_0)_g \quad \mathscr{H}^{m-1}\text{-a.e. in } \Omega. \tag{3-24}$$

Furthermore, (3-23) and (3-22) easily imply that $\boldsymbol{u}_0^\varepsilon \to \boldsymbol{u}_0$ in $L^p(\Omega;\mathbb{R}^N)$ for all $1 \le p < \infty$. Finally, applying the chain rule and (3-23), [Ambrosio et al. 2000, Proposition 3.2], and [Ambrosio et al. 2000, Theorem 2.2(b)] (in this order), we see that

$$\int_\Omega |\nabla \boldsymbol{u}_0^\varepsilon|\,dx \le C\int_\Omega |\nabla(T\boldsymbol{u}_0 \star \rho_\varepsilon)|\,dx = C\int_\Omega |(DT\boldsymbol{u}_0) \star \rho_\varepsilon|\,dx \le C|DT\boldsymbol{u}_0|(\Omega \oplus B_\varepsilon). \tag{3-25}$$

Similarly,

$$\int_\Omega |\nabla \boldsymbol{u}_0^\varepsilon|^2\,dx \le C\int_\Omega |(DT\boldsymbol{u}_0)\star\rho_\varepsilon|^2\,dx \le C\|(DT\boldsymbol{u}_0)\star\rho_\varepsilon\|_\infty \int_\Omega |(DT\boldsymbol{u}_0)\star\rho_\varepsilon|\,dx,$$

and, using the definition of ρ_ε, we conclude that

$$\varepsilon^\alpha \int_\Omega |\nabla \boldsymbol{u}_0^\varepsilon|^2\,dx \le C\varepsilon^{\alpha-m}(|DT\boldsymbol{u}_0|(\Omega\oplus B_\varepsilon))^2. \tag{3-26}$$

Inequalities (3-25) and (3-26), together with (3-24), complete the proof. □

Proof of Theorem 3.6. We proceed in steps. In the first step, we use the previous lemmas, together with standard compactness arguments, to identify a triplet $(\boldsymbol{u}, \boldsymbol{Z}, \boldsymbol{\mu})$. In the second step we identify $\boldsymbol{\mu}$ in terms of $\boldsymbol{u}$ and $\boldsymbol{Z}$, which automatically yields an upper bound on $|\boldsymbol{\mu}|$. In the third step, collecting the information of the previous two steps, we note that $\boldsymbol{u}$ satisfies all the properties in Definition 3.4 except for

$$\boldsymbol{\mu}(t) = \boldsymbol{u}(t)_g|D\boldsymbol{u}(t)| \quad \text{as measures for a.e. } t\in[0,T], \tag{3-27}$$

to which the rest of the proof is devoted. In the fourth step we use the lower semicontinuity results in Section 2F to prove a lower bound on $\boldsymbol{\mu}(t)$ over the diffuse support of $|D\boldsymbol{u}(t)|$. In the fifth step, we revise the blow-up argument given in [Fonseca and Müller 1993; Fonseca and Rybka 1992] to obtain a lower bound on $\boldsymbol{\mu}(t)$ over $J_{\boldsymbol{u}(t)}$. Finally, in the sixth step we complete the proof.

Step 1: *Passage to the limit.* Let $\boldsymbol{u}_0^\varepsilon$ and $\boldsymbol{u}^\varepsilon$ be as given by Lemma 3.9 and Proposition 3.7, respectively. By Lemma 3.8, $\boldsymbol{u}^\varepsilon \in \mathbb{S}_+^{N-1}$ a.e. in Q_T. By (3-8), Lemma 3.9(iii), and (3-15), a positive constant C (independent of ε) exists such that

$$\sup_{t\in(0,T)} \|\boldsymbol{u}^\varepsilon\|_{W^{1,1}(\Omega)} \le C, \tag{3-28}$$

$$\|\boldsymbol{u}_t^\varepsilon\|_{L^2(0,T;L^2(\Omega;\mathbb{R}^N))} \le C. \tag{3-29}$$

We recall that $\mathrm{BV}(\Omega;\mathbb{R}^N)$ is compactly embedded in $L^1(\Omega;\mathbb{R}^N)$ [Ambrosio et al. 2000, Theorem 3.23]. Hence the Aubin–Simon compactness criterion [Simon 1987, Corollary 8.4], together with (3-28) and (3-29), implies that

$$\boldsymbol{u}^\varepsilon \to \boldsymbol{u} \quad \text{in } C(0,T;L^1(\Omega;\mathbb{R}^N)) \text{ and a.e. in } Q_T \tag{3-30}$$

for a subsequence. By the lower semicontinuity of the total variation [Ambrosio et al. 2000, Remark 3.5], (3-30) and (3-15) imply that

$$\boldsymbol{u} \in L^\infty(0,T;\mathrm{BV}(\Omega;\mathbb{R}^N)). \tag{3-31}$$

From (3-30) and Lemma 3.9(i), we have

$$\boldsymbol{u}(0) = \boldsymbol{u}_0 \tag{3-32}$$

and, using also (3-8),

$$|\boldsymbol{u}| = 1 \quad \text{a.e. in } Q_T. \tag{3-33}$$

By a standard interpolation argument, the boundedness of $\boldsymbol{u}^\varepsilon$ in $L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^N))$ and (3-30) imply that

$$\boldsymbol{u}^\varepsilon \to \boldsymbol{u} \quad \text{in } L^p(0,T;L^q(\Omega;\mathbb{R}^N)) \text{ for all } p,q\in[1,\infty) \text{ and a.e. in } Q_T. \tag{3-34}$$

Moreover, it follows from (3-29) that

$$\boldsymbol{u}_t^\varepsilon \rightharpoonup \boldsymbol{u}_t \quad \text{in } L^2(0,T;L^2(\Omega;\mathbb{R}^N)). \tag{3-35}$$

By (3-15), Lemma 3.9(iii), and (3-18), a subsequence exists such that

$$\varepsilon^\alpha \nabla \boldsymbol{u}^\varepsilon \rightharpoonup 0 \quad \text{in } L^2(0,T;L^2(\Omega;\mathbb{R}^{N\times m})), \tag{3-36}$$

$$\frac{\nabla \boldsymbol{u}^\varepsilon}{\sqrt{|\nabla \boldsymbol{u}^\varepsilon|^2+\varepsilon^2}} \overset{*}{\rightharpoonup} \boldsymbol{Z} \quad \text{in } L^\infty(Q_T;\mathbb{R}^{N\times m}). \tag{3-37}$$

Recalling the definition (3-10) of $\boldsymbol{Z}^\varepsilon$, by (3-36) and (3-37) we obtain that

$$\boldsymbol{Z}^\varepsilon \rightharpoonup \boldsymbol{Z} \quad \text{in } L^2(0,T;L^2(\Omega;\mathbb{R}^{N\times m})), \tag{3-38}$$

and from (3-37) we also obtain that

$$\|\boldsymbol{Z}\|_{L^\infty(Q_T)} \le 1. \tag{3-39}$$

Since $\{\boldsymbol{\mu}^\varepsilon\}$ is bounded in $L^\infty(0,T;L^1(\Omega;\mathbb{R}^N))$ and

$$L^\infty(0,T;L^1(\Omega;\mathbb{R}^N)) \subset L^\infty(0,T;\mathcal{M}(\Omega;\mathbb{R}^N)) \subset (L^1(0,T;C_0(\Omega;\mathbb{R}^N)))'$$

(see Section 2B), we have

$$\boldsymbol{\mu}^\varepsilon \overset{*}{\rightharpoonup} \boldsymbol{\mu} \quad \text{in } (L^1(0,T;C_0(\Omega;\mathbb{R}^N)))'. \tag{3-40}$$

Analogously, by (3-16),

$$\operatorname{div} \boldsymbol{Z}^\varepsilon \overset{*}{\rightharpoonup} \operatorname{div} \boldsymbol{Z} \quad \text{in } (L^2(0,T;C_0(\Omega;\mathbb{R}^N)))'. \tag{3-41}$$

Passing to the limit as $\varepsilon\to 0$ in $(3\text{-}9)_1$ (using (3-35), (3-41), and (3-40)), we obtain

$$\boldsymbol{u}_t - \operatorname{div} \boldsymbol{Z} = \boldsymbol{\mu} \quad \text{in } (L^2(0,T;C_0(\Omega;\mathbb{R}^N)))'. \tag{3-42}$$

Passing to the limit as $\varepsilon\to 0$ in (3-12) (using (3-38) and (3-34)), in (3-13) (using (3-35) and (3-34)), and in (3-14) (using (3-35), (3-38), and (3-34)), we get that

$$\boldsymbol{Z}^T\boldsymbol{u} = 0 \qquad \text{a.e. in } Q_T, \tag{3-43}$$

$$\boldsymbol{u}_t \cdot \boldsymbol{u} = 0 \qquad \text{a.e. in } Q_T, \tag{3-44}$$

$$\boldsymbol{u}_t(t)\wedge\boldsymbol{u}(t) = \operatorname{div}(\boldsymbol{Z}(t)\wedge\boldsymbol{u}(t)) \quad \text{in } L^2(\Omega;\Lambda_2(\mathbb{R}^N)) \text{ for a.e. } t\in[0,T]. \tag{3-45}$$

Step 2: *The intermediate identification of* $\boldsymbol{\mu}$ *and its upper bound.* We claim that

$$\boldsymbol{\mu} = *(*(\boldsymbol{Z}\wedge\boldsymbol{u})\wedge D\boldsymbol{u}) \in L^\infty(0,T;\mathcal{M}(\Omega;\mathbb{R}^N)) \tag{3-46}$$

with

$$|\boldsymbol{\mu}(t)| \le |D\boldsymbol{u}(t)| \quad \text{as measures for a.e. } t \in [0,T]. \tag{3-47}$$

Let

$$\mathscr{A} = (\mathscr{A}_1, \dots, \mathscr{A}_m) := *(\boldsymbol{Z} \wedge \boldsymbol{u}) \in L^\infty(Q_T; (\Lambda_{N-2}(\mathbb{R}^N))^m). \tag{3-48}$$

We have

$$*(\boldsymbol{u}_t \wedge \boldsymbol{u}) \overset{(3\text{-}45)}{=} *(\operatorname{div}(\boldsymbol{Z} \wedge \boldsymbol{u})) \overset{(2\text{-}18)}{=} \operatorname{div}(*(\boldsymbol{Z} \wedge \boldsymbol{u})) = \operatorname{div} \mathscr{A}, \tag{3-49}$$

hence $\mathscr{A}(t) \in \boldsymbol{X}_2(\Omega; \Lambda_{N-2}(\mathbb{R}^N))$ for a.e. t. Therefore, by Lemma 2.3, $*(\mathscr{A}(t) \wedge D\boldsymbol{u}(t)) \in \mathcal{M}(\Omega; \mathbb{R}^N)$ for almost every t with

$$|*(\mathscr{A}(t) \wedge D\boldsymbol{u}(t))| \overset{(2\text{-}21)}{\le} \|\mathscr{A}(t)\|_\infty |D\boldsymbol{u}(t)| \overset{(2\text{-}9)}{=} \|\boldsymbol{Z}(t) \wedge \boldsymbol{u}(t)\|_\infty |D\boldsymbol{u}(t)| \overset{(2\text{-}10),(3\text{-}33),(3\text{-}39)}{\le} |D\boldsymbol{u}(t)|, \tag{3-50}$$

and in addition

$$*(\mathscr{A}(t) \wedge D\boldsymbol{u}(t)) \overset{(2\text{-}22)}{=} - * (\operatorname{div} \mathscr{A}(t) \wedge \boldsymbol{u}(t)) \mathscr{L}^m + \operatorname{div}(*(\mathscr{A}(t) \wedge \boldsymbol{u}(t))). \tag{3-51}$$

It follows from (3-50) and (3-31) that

$$*(\mathscr{A} \wedge D\boldsymbol{u}) \in L^\infty(Q_T; \mathcal{M}(\Omega; \mathbb{R}^N)). \tag{3-52}$$

Using (3-51), we see that

$$\begin{aligned} \boldsymbol{u}_t &\overset{(3\text{-}33)}{=} |\boldsymbol{u}|^2 \boldsymbol{u}_t \overset{(2\text{-}7)}{=} (\boldsymbol{u}_t \cdot \boldsymbol{u})\boldsymbol{u} - *(*(\boldsymbol{u}_t \wedge \boldsymbol{u}) \wedge \boldsymbol{u}) \\ &\underset{(3\text{-}49)}{\overset{(3\text{-}44)}{=}} -*(\operatorname{div} \mathscr{A} \wedge \boldsymbol{u}) \overset{(3\text{-}51)}{=} *(\mathscr{A} \wedge D\boldsymbol{u}) - \operatorname{div}((*(\mathscr{A} \wedge \boldsymbol{u}))). \end{aligned} \tag{3-53}$$

On the other hand,

$$\begin{aligned} - * (\mathscr{A} \wedge \boldsymbol{u}) &\overset{(3\text{-}48)}{=} -(*(*(\boldsymbol{Z} \wedge \boldsymbol{u}) \wedge \boldsymbol{u})) \\ &= -(*(*(\boldsymbol{z}_1 \wedge \boldsymbol{u}) \wedge \boldsymbol{u}), \dots, *(*(\boldsymbol{z}_m \wedge \boldsymbol{u}) \wedge \boldsymbol{u})) \\ &\overset{(2\text{-}7)}{=} (|\boldsymbol{u}|^2 \boldsymbol{z}_1 - (\boldsymbol{u} \cdot \boldsymbol{z}_1)\boldsymbol{u}, \dots, |\boldsymbol{u}|^2 \boldsymbol{z}_m - (\boldsymbol{u} \cdot \boldsymbol{z}_m)\boldsymbol{u}) \\ &= \boldsymbol{Z} - (\boldsymbol{Z}^T \boldsymbol{u})\boldsymbol{u} \overset{(3\text{-}43)}{=} \boldsymbol{Z}. \end{aligned} \tag{3-54}$$

Combining (3-53) and (3-54), we obtain

$$\boldsymbol{u}_t = *(\mathscr{A} \wedge D\boldsymbol{u}) + \operatorname{div} \boldsymbol{Z},$$

which together with (3-42), (3-48), and (3-52), implies (3-46). Finally, (3-47) follows immediately from (3-46) and (3-50).

Step 3: *Intermediate summary.* It follows from (3-42), (3-35), and (3-46) that $\operatorname{div} \boldsymbol{Z} \in L^2(0, T; \mathcal{M}(\Omega; \mathbb{R}^N))$. Hence (3-42) upgrades to

$$\boldsymbol{u}_t(t) - \operatorname{div} \boldsymbol{Z}(t) = \boldsymbol{\mu}(t) \quad \text{as measures for a.e. } t \in [0,T]. \tag{3-55}$$

In particular,

$$\boldsymbol{Z}(t) \in \boldsymbol{X}^N_{\mathcal{M}}(\Omega) \quad \text{for a.e. } t \in [0,T]. \tag{3-56}$$

Thus the weak trace $[\boldsymbol{Z}(t), \nu]$ on $\partial\Omega$ of the normal component of $\boldsymbol{Z}(t)$ is well defined, and for all smooth $\boldsymbol{w}$ we have

$$\int_0^T \int_{\partial\Omega} [\boldsymbol{Z}, \nu] \cdot \boldsymbol{w}\, d\mathscr{H}^{m-1}\, dt \overset{(2\text{-}14)}{=} \int_0^T \left(\int_\Omega \boldsymbol{w} \cdot d(\operatorname{div} \boldsymbol{Z}) + \boldsymbol{Z} : \nabla \boldsymbol{w}\, dx \right) dt$$
$$\overset{\substack{(3\text{-}38)\\(3\text{-}41)}}{=} \lim_{\varepsilon\to 0} \left(\int_0^T \int_\Omega (\boldsymbol{w} \cdot \operatorname{div} \boldsymbol{Z}^\varepsilon + \boldsymbol{Z}^\varepsilon : \nabla \boldsymbol{w})\, dx\, dt \right) \overset{(3\text{-}9)_2}{=} 0.$$

Hence

$$[\boldsymbol{Z}(t), \nu] = 0 \quad \mathscr{H}^{m-1}\text{-a.e. on } \partial\Omega \text{ for a.e. } t \in [0,T]. \tag{3-57}$$

Collecting (3-31), (3-30), (3-35), (3-32), (3-33), (3-39), (3-56), (3-43), (3-45), and (3-57), we see that all the properties of $\boldsymbol{u}$ stated in Definition 3.4 are satisfied except for (3-3). In view of (3-55), in order to prove (3-3), it remains to show (3-27).

Step 4: *The lower bound on $\boldsymbol{\mu}$ over the diffuse support of $|D\boldsymbol{u}|$.* In view of (3-47), $\boldsymbol{\mu}(t)$ can be decomposed as

$$\boldsymbol{\mu}(t) = \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|}(|\nabla \boldsymbol{u}(t)|\mathscr{L}^m + |D^c(\boldsymbol{u}(t))|) + \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|}|\boldsymbol{u}(t)_+ - \boldsymbol{u}(t)_-|\mathscr{H}^{m-1} \llcorner J_{\boldsymbol{u}(t)}, \tag{3-58}$$

where $\boldsymbol{\mu}(t)/|D\boldsymbol{u}(t)| \in (L^1(\Omega; |D\boldsymbol{u}(t)|))^N$ denotes the Radon–Nikodým derivative of $\boldsymbol{\mu}(t)$ with respect to $|D\boldsymbol{u}(t)|$. We claim that

$$\boldsymbol{u}(t) \cdot \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \geq 1 \quad (|\nabla \boldsymbol{u}(t)|\mathscr{L}^m + |D^c(\boldsymbol{u}(t))|)\text{-a.e. in } \Omega. \tag{3-59}$$

We first notice that

$$\mu^{\varepsilon,\ell} \overset{(3\text{-}10)}{\geq} u^{\varepsilon,\ell}(\sqrt{\varepsilon^2 + |\nabla \boldsymbol{u}^\varepsilon|^2} - \varepsilon) \geq u^{\varepsilon,\ell}(|\nabla \boldsymbol{u}^\varepsilon| - \varepsilon), \quad \ell = 1, \dots N. \tag{3-60}$$

For any $\varphi \in C(\overline{\Omega}; [0,\infty))$, $0 \leq \psi \in L^1((0,T))$, and $\ell \in \{1, \dots, N\}$, we have

$$\int_0^T \psi(t) \left(\int_\Omega \varphi\, d\mu^\ell(t) \right) dt \overset{(3\text{-}40)}{=} \lim_{\varepsilon\to 0} \int_0^T \psi(t) \left(\int_\Omega \varphi \mu^{\varepsilon,\ell}(t)\, dx \right) dt$$
$$\overset{(3\text{-}60)}{\geq} \liminf_{\varepsilon\to 0} \int_0^T \psi(t) \left(\int_\Omega \varphi u^{\varepsilon,\ell}(t) |\nabla \boldsymbol{u}^\varepsilon(t)|\, dx \right) dt.$$

We claim that, for a.e. $t \in (0,T)$,

$$\boldsymbol{u}^\varepsilon(t) \rightharpoonup \boldsymbol{u}(t) \quad \text{in } \mathrm{BV}(\Omega; \mathbb{R}^N) \text{ as } \varepsilon \to 0. \tag{3-61}$$

Indeed, in view of (3-28), for a.e. t we have $\|\boldsymbol{u}^\varepsilon(t)\|_{W^{1,1}(\Omega)} < \infty$. Take any such t and assume for a contradiction that (3-61) does not hold, that is, that $\boldsymbol{u}^\varepsilon(t) \not\rightharpoonup \boldsymbol{u}(t)$ for a subsequence. By (3-28), a further subsequence would exist such that $\boldsymbol{u}^\varepsilon(t) \rightharpoonup \tilde{\boldsymbol{u}}$ for some $\tilde{\boldsymbol{u}} \in \mathrm{BV}(\Omega; \mathbb{R}^N)$. On the other hand, because of (3-30), $\boldsymbol{u}^\varepsilon(t) \to \boldsymbol{u}(t)$ in $L^1(\Omega; \mathbb{R}^N)$: hence $\tilde{\boldsymbol{u}} = \boldsymbol{u}(t)$, a contradiction.

In view of (3-61) and (3-30), we may apply Proposition 2.6 to the right-hand side of (3-60) with $f = f_{\varphi,\ell} : \Omega \times \mathbb{R}^N \to [0,\infty)$ defined by $f_{\varphi,\ell}(x,\boldsymbol{s}) := \varphi(x)s^\ell|\xi|$. This implies that

$$\int_0^T \psi(t)\left(\int_\Omega \varphi\, d\mu^\ell(t)\right)dt \geq \int_0^T \psi(t)\left(\int_\Omega \varphi u^\ell(t)(|\nabla \boldsymbol{u}(t)|\,dx + d|D^c\boldsymbol{u}(t)|) + \int_{J_{\boldsymbol{u}}(t)} \varphi K_t^\ell\, d\mathcal{H}^{m-1}\right)dt,$$

where

$$K_t^\ell = \inf\left\{\int_0^1 \gamma^\ell(\tau)|\dot{\boldsymbol{\gamma}}(\tau)|\,d\tau : \boldsymbol{\gamma} \in W^{1,1}((0,1);\mathbb{S}_+^{N-1}),\ \boldsymbol{\gamma}(0) = \boldsymbol{u}(t)_-, \boldsymbol{\gamma}(1) = \boldsymbol{u}(t)_+\right\}. \tag{3-62}$$

By the arbitrariness of ψ, we conclude that

$$\int_\Omega \varphi\, d\mu^\ell(t) \geq \int_\Omega \varphi u^\ell(t)(|\nabla \boldsymbol{u}(t)|\,dx + d|D^c\boldsymbol{u}(t)|) + \int_{J_{\boldsymbol{u}}(t)} \varphi K_t^\ell\, d\mathcal{H}^{m-1} \quad \text{for all } \varphi \in C(\overline{\Omega}) \tag{3-63}$$

for a.e. $t \in [0,T]$ and for all $\ell \in \{1,\dots,N\}$. Recalling (3-58), (3-63) yields

$$\frac{\mu^\ell(t)}{|D\boldsymbol{u}(t)|} \geq u^\ell(t) \quad (|\nabla \boldsymbol{u}(t)|\mathcal{L}^m + |D^c(\boldsymbol{u}(t))|)\text{-a.e. in } \Omega$$

for a.e. $t \in [0,T]$ and all $\ell = 1,\dots,N$. Now, recalling that $|\boldsymbol{u}(t)| = 1$ a.e. in Ω, we obtain the inequality (3-59) at once.

Remark 3.10. On the jump set $J_{\boldsymbol{u}(t)}$, the above argument would yield

$$|\boldsymbol{u}_+(t) - \boldsymbol{u}_-(t)|\boldsymbol{u}(t)_g \cdot \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \geq \boldsymbol{u}(t)_g \cdot (K_t^1,\dots,K_t^N) \quad \mathcal{H}^{m-1}\text{-a.e. on } J_{\boldsymbol{u}}(t).$$

Unfortunately, by (3-62) and obvious properties of the infimum,

$$\begin{aligned}&\boldsymbol{u}(t)_g \cdot (K_t^1,\dots,K_t^N)\\ &\quad \leq \inf\left\{\int_0^1 \boldsymbol{u}(t)_g \cdot \boldsymbol{\gamma}(\tau)|\dot{\boldsymbol{\gamma}}(\tau)|\,d\tau : \boldsymbol{\gamma} \in W^{1,1}((0,1);\mathbb{S}_+^{N-1}), \boldsymbol{\gamma}(0) = \boldsymbol{u}(t)_-, \boldsymbol{\gamma}(1) = \boldsymbol{u}(t)_+\right\},\end{aligned} \tag{3-64}$$

whilst, as we shall see, it is the right-hand side of (3-64) which yields the sharp lower bound on the jump part (cf. (3-70)–(3-73) below). On the other hand, we can not use the results in Proposition 2.6 directly on $\boldsymbol{u}^* \cdot \boldsymbol{\mu}^\varepsilon$, since $\boldsymbol{u}^*$ is a discontinuous function (though a very special one). This motivates the discussion that follows.

Step 5: *The lower bound on* $\boldsymbol{\mu}$ *over* $J_{\boldsymbol{u}(t)}$. We claim that

$$\boldsymbol{u}(t)_g \cdot \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \geq 1 \quad \mathcal{H}^{m-1}\text{-a.e. on } J_{\boldsymbol{u}(t)}. \tag{3-65}$$

It follows from (3-8) and (3-28) that, for a.e. $t \in [0,T]$, there exists a subsequence ε_k such that

$$\boldsymbol{u}^{\varepsilon_k}(t)|\nabla \boldsymbol{u}^{\varepsilon_k}(t)| \overset{*}{\rightharpoonup} \tilde{\boldsymbol{\mu}}(t) \quad \text{in } \mathcal{M}(\Omega;\mathbb{R}^N). \tag{3-66}$$

Then (3-60) and the fact that $u^{\varepsilon,\ell} \geq 0$ imply that

$$\mu^\ell(t) \geq \tilde{\mu}^\ell(t) \geq 0 \quad \text{as measures for a.e. } t \in [0,T], \quad \ell \in \{1,\dots,N\}. \tag{3-67}$$

Hereafter, we argue for a fixed t and we do not specify dependence on t for notational convenience. Using the Radon–Nikodým theorem [Ambrosio et al. 2000, Theorem 1.28], we decompose $\tilde{\boldsymbol{\mu}}$ into four mutually orthogonal measures:

$$\tilde{\boldsymbol{\mu}} = \frac{\tilde{\boldsymbol{\mu}}}{|D\boldsymbol{u}|}|\nabla \boldsymbol{u}|\mathscr{L}^N + \frac{\tilde{\boldsymbol{\mu}}}{|D\boldsymbol{u}|}|D^c\boldsymbol{u}| + \frac{\tilde{\boldsymbol{\mu}}}{|D\boldsymbol{u}|}|\boldsymbol{u}_+ - \boldsymbol{u}_-|\mathscr{H}^{m-1} \llcorner J_{\boldsymbol{u}} + (\tilde{\boldsymbol{\mu}})^o$$

with $(\tilde{\boldsymbol{\mu}})^o \perp |D\boldsymbol{u}|$. It follows from (3-67) and (3-58) that

$$\boldsymbol{u}_g \cdot \frac{\boldsymbol{\mu}}{|D\boldsymbol{u}|} \geq \boldsymbol{u}_g \cdot \frac{\tilde{\boldsymbol{\mu}}}{|D\boldsymbol{u}|} \quad \mathscr{H}^{m-1}\text{-a.e. on } J_{\boldsymbol{u}}.$$

Therefore, (3-65) is proved once we have shown that

$$\boldsymbol{u}_g \cdot \frac{\tilde{\boldsymbol{\mu}}}{|D\boldsymbol{u}|} \geq 1 \quad \mathscr{H}^{m-1}\text{-a.e. on } J_{\boldsymbol{u}}. \tag{3-68}$$

To prove (3-68) we apply the same blow-up argument as in [Fonseca and Müller 1993, Section 3].

From the Besicovitch differentiation theorem [Ambrosio et al. 2000, Theorem 2.22], for $\mathscr{H}^{m-1}$-a.e. $x_0 \in J_{\boldsymbol{u}}$, we have

$$\frac{\tilde{\boldsymbol{\mu}}}{|D\boldsymbol{u}|}(x_0) = \lim_{\delta\to 0} \frac{\tilde{\boldsymbol{\mu}}(x_0 + \delta Q_{\nu_{\boldsymbol{u}}(x_0)})}{|\boldsymbol{u}_+ - \boldsymbol{u}_-|\mathscr{H}^{m-1}(J_{\boldsymbol{u}} \cap (x_0 + \delta Q_{\nu_{\boldsymbol{u}}(x_0)}))},$$

where Q_ς is defined in Section 2F. On the other hand, by [Fonseca and Müller 1993, Lemma 2.6], for $\mathscr{H}^{m-1}$ a.e. $x_0 \in J_{\boldsymbol{u}}$, we also have

$$\lim_{\delta\to 0} \frac{1}{\delta^{m-1}} \int_{(x_0+\delta Q_{\nu_{\boldsymbol{u}}(x_0)})\cap J_{\boldsymbol{u}}} |\boldsymbol{u}_+(x) - \boldsymbol{u}_-(x)|\, d\mathscr{H}^{m-1} = |\boldsymbol{u}_+(x_0) - \boldsymbol{u}_-(x_0)|.$$

Therefore, letting

$$M = |\boldsymbol{u}_+(x_0) - \boldsymbol{u}_-(x_0)|$$

for notational convenience, we obtain that

$$M\frac{\tilde{\boldsymbol{\mu}}}{|D\boldsymbol{u}|}(x_0) = \lim_{\delta\to 0} \frac{1}{\delta^{m-1}} \int_{x_0+\delta Q_{\nu_{\boldsymbol{u}}(x_0)}} d\tilde{\boldsymbol{\mu}}.$$

Then, for any $\ell \in \{1, \dots, N\}$, since the function $\chi_{x_0+\delta Q_{\nu_{\boldsymbol{u}}(x_0)}}$ is upper semicontinuous with compact support in Ω if δ is sufficiently small, we have

$$\begin{aligned} M\frac{\tilde{\mu}^\ell}{|D\boldsymbol{u}|}(x_0) &= \lim_{\delta\to 0} \frac{1}{\delta^{m-1}} \int_{x_0+\delta Q_{\nu_{\boldsymbol{u}}(x_0)}} d\tilde{\mu}^\ell \\ &\overset{(3\text{-}66)}{\geq} \limsup_{\delta\to 0} \limsup_{k\to\infty} \frac{1}{\delta^{m-1}} \int_{x_0+\delta Q_{\nu_{\boldsymbol{u}}(x_0)}} u^{\varepsilon_k,\ell} |\nabla \boldsymbol{u}^{\varepsilon_k}|\, dx \\ &= \limsup_{\delta\to 0} \limsup_{k\to\infty} \int_{Q_{\nu_{\boldsymbol{u}}(x_0)}} v^\ell_{\delta,k}(y) |\nabla \boldsymbol{v}_{\delta,k}(y)|\, dy, \end{aligned} \tag{3-69}$$

where

$$\boldsymbol{v}_{\delta,k}(y) := \boldsymbol{u}^{\varepsilon_k}(x_0 + \delta y).$$

We now observe that $\boldsymbol{v}_{\delta,k} \in W^{1,1}(Q_{\nu_{\boldsymbol{u}(x_0)}}; \mathbb{R}^N)$ and (see [Fonseca and Müller 1993, formula (3.2)])

$$\lim_{\delta\to 0}\lim_{k\to\infty} \|\boldsymbol{v}_{\delta,k} - \boldsymbol{w}_0\|_{L^1(Q_{\nu_{\boldsymbol{u}(x_0)}};\mathbb{R}^N)} = 0,$$

where

$$\boldsymbol{w}_0(y) := \begin{cases} cc\boldsymbol{u}_+(x_0) & \text{if } y\cdot \nu_{\boldsymbol{u}(x_0)} > 0,\\ \boldsymbol{u}_-(x_0) & \text{if } y\cdot \nu_{\boldsymbol{u}(x_0)} < 0.\end{cases}$$

Then, by a diagonalization argument, we may extract a subsequence $\boldsymbol{v}_k$ converging to $\boldsymbol{w}_0$ in $L^1(Q_{\nu_{\boldsymbol{u}(x_0)}}; \mathbb{R}^N)$. It follows from (3-69) that

$$M\frac{\tilde{\mu}^\ell}{|D\boldsymbol{u}|}(x_0) \geq \lim_{k\to\infty}\int_{Q_{\nu_{\boldsymbol{u}(x_0)}}} v_k^\ell(y)|\nabla \boldsymbol{v}_k(y)|\,dy.$$

Since $(u^\ell)^* \geq 0$ for all $\ell \in \{1,\dots,N\}$, this implies that

$$M\left(\boldsymbol{u}_g\cdot\frac{\tilde{\boldsymbol{\mu}}}{|D\boldsymbol{u}|}\right)(x_0) \geq \lim_{k\to\infty}\int_{Q_{\nu_{\boldsymbol{u}(x_0)}}} \boldsymbol{u}_g(x_0)\cdot\boldsymbol{v}_k(y)|\nabla \boldsymbol{v}_k(y)|\,dy.$$

The function $f(x,\boldsymbol{s}) = f(\boldsymbol{s}) = \boldsymbol{u}_g(x_0)\cdot\boldsymbol{s}$ is continuous, nonnegative, and bounded. Then, applying Lemma 2.5, we obtain a new sequence

$$\boldsymbol{w}_k \in \mathcal{P}(\boldsymbol{u}_+(x_0), \boldsymbol{u}_-(x_0), \nu_{\boldsymbol{u}(x_0)})$$

(with $\mathcal{P}$ given by (2-24)) converging to $\boldsymbol{w}_0$ in $L^1(Q_{\nu_{\boldsymbol{u}(x_0)}}; \mathbb{R}^N)$ and such that

$$M\left(\boldsymbol{u}_g\cdot\frac{\tilde{\boldsymbol{\mu}}}{|D\boldsymbol{u}|}\right)(x_0) \geq \limsup_{k\to\infty}\int_{Q_{\nu_{\boldsymbol{u}(x_0)}}} \boldsymbol{u}_g(x_0)\cdot\boldsymbol{w}_k(y)|\nabla \boldsymbol{w}_k(y)|\,dy.$$

We may now apply Lemma 2.4. It follows from (2-23) and (2-25) that

$$M\left(\boldsymbol{u}_g\cdot\frac{\tilde{\boldsymbol{\mu}}}{|D\boldsymbol{u}|}\right)(x_0) \geq \inf_{\boldsymbol{\gamma}\in\widetilde{\Gamma}_N(\boldsymbol{u}_+(x_0),\boldsymbol{u}_-(x_0))} J_N[\boldsymbol{u}_+(x_0),\boldsymbol{u}_-(x_0)](\boldsymbol{\gamma}), \tag{3-70}$$

where

$$J_N[\boldsymbol{v}_0,\boldsymbol{v}_1](\boldsymbol{\gamma}) := \int_0^1 \boldsymbol{v}_g\cdot\boldsymbol{\gamma}(t)|\dot{\boldsymbol{\gamma}}(t)|\,dt, \quad \boldsymbol{v}_g := \frac{\boldsymbol{v}_0+\boldsymbol{v}_1}{|\boldsymbol{v}_0+\boldsymbol{v}_1|}, \tag{3-71}$$

and

$$\widetilde{\Gamma}_N(\boldsymbol{v}_0,\boldsymbol{v}_1) := \{\boldsymbol{\gamma}\in W^{1,1}((0,1);\mathbb{S}^{N-1}_+) : \boldsymbol{\gamma}(0) = \boldsymbol{v}_0,\ \boldsymbol{\gamma}(1) = \boldsymbol{v}_1\}. \tag{3-72}$$

In view of (3-70), (3-68) and therefore (3-65) follows from

$$\inf_{\boldsymbol{\gamma}\in\widetilde{\Gamma}_N(\boldsymbol{u}_+(x_0),\boldsymbol{u}_-(x_0))} J_N[\boldsymbol{u}_+(x_0),\boldsymbol{u}_-(x_0)](\boldsymbol{\gamma}) \geq M = |\boldsymbol{u}_+(x_0)-\boldsymbol{u}_-(x_0)|. \tag{3-73}$$

This last inequality will be proved in Theorem 4.1, to which the next section is devoted.

Step 6: *Conclusion*. Recalling (3-58), the upper bound on $|\boldsymbol{\mu}|$ given by (3-47) immediately implies that

$$\left|\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|}\right| \leq 1 \quad |D\boldsymbol{u}(t)|\text{-a.e. in } \Omega \tag{3-74}$$

for a.e. $t \in [0, T]$. In particular, recalling (3-33),

$$\begin{aligned}
\boldsymbol{u}(t) \cdot \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} &\leq 1 \quad (|\nabla \boldsymbol{u}(t)|\mathscr{L}^m + |D^c(\boldsymbol{u}(t))|)\text{-a.e. in } \Omega,\\
\boldsymbol{u}(t)_g \cdot \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} &\leq 1 \quad \mathscr{H}^{m-1}\text{-a.e. on } J_{\boldsymbol{u}(t)}
\end{aligned}$$

for a.e. $t \in [0, T]$. Combining these inequalities with the lower bounds in (3-59) and (3-65), we obtain

$$\boldsymbol{u}(t) \cdot \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} = 1 \quad (|\nabla \boldsymbol{u}(t)|\mathscr{L}^m + |D^c(\boldsymbol{u}(t))|)\text{-a.e. in } \Omega, \tag{3-75}$$

$$\boldsymbol{u}(t)_g \cdot \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} = 1 \quad \mathscr{H}^{m-1}\text{-a.e. on } J_{\boldsymbol{u}(t)} \tag{3-76}$$

for a.e. $t \in [0, T]$. We are now ready to complete the proof. By (2-2), we have

$$\left|\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \wedge \boldsymbol{u}(t)\right|^2 = \left|\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|}\right|^2 - \left|\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \cdot \boldsymbol{u}(t)\right|^2 \quad (|\nabla \boldsymbol{u}(t)|\mathscr{L}^m + |D^c(\boldsymbol{u}(t))|)\text{-a.e. in } \Omega \tag{3-77}$$

and

$$\left|\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \wedge \boldsymbol{u}(t)_g\right|^2 = \left|\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|}\right|^2 |\boldsymbol{u}(t)_g|^2 - \left|\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \cdot \boldsymbol{u}(t)_g\right|^2 \quad \mathscr{H}^{m-1}\text{-a.e. on } J_{\boldsymbol{u}(t)}. \tag{3-78}$$

Now, from (3-74),(3-75), and (3-77), we get

$$\left|\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \wedge \boldsymbol{u}(t)\right|^2 = 0 \quad (|\nabla \boldsymbol{u}(t)|\mathscr{L}^m + |D^c(\boldsymbol{u}(t))|)\text{-a.e. in } \Omega,$$

and from (3-74),(3-76), and (3-78), we get

$$\left|\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} \wedge \boldsymbol{u}(t)_g\right|^2 = 0 \quad \mathscr{H}^{m-1}\text{-a.e. on } J_{\boldsymbol{u}(t)}.$$

Hence the wedge products on the left-hand side are zero.

Therefore, applying (2-7) and using once more the equalities in (3-75) and (3-76), we conclude that

$$\begin{aligned}
\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} &= |\boldsymbol{u}(t)|^2 \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} = \boldsymbol{u}(t) \qquad (|\nabla \boldsymbol{u}(t)|\mathscr{L}^m + |D^c(\boldsymbol{u}(t))|)\text{-a.e. in } \Omega,\\
\frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} &= |\boldsymbol{u}(t)_g|^2 \frac{\boldsymbol{\mu}(t)}{|D\boldsymbol{u}(t)|} = \boldsymbol{u}(t)_g \quad \mathscr{H}^{m-1}\text{-a.e. on} J_{\boldsymbol{u}(t)}.
\end{aligned}$$

Plugging these expressions into (3-58), we obtain (3-27), and the proof is complete. □

4. A nonconvex variational problem

In this section we study the minimization of a nonconvex functional, and, as a result, we prove the inequality (3-73).

Theorem 4.1. *Let $\boldsymbol{v}_0, \boldsymbol{v}_1 \in \mathbb{S}^{N-1}$ and let $J_N[\boldsymbol{v}_0, \boldsymbol{v}_1](\boldsymbol{\gamma})$ and $\widetilde{\Gamma}_N(\boldsymbol{v}_0, \boldsymbol{v}_1)$ be given by (3-71) and (3-72). If $\boldsymbol{v}_0 \cdot \boldsymbol{v}_1 \geq 0$, then*

$$\min_{\boldsymbol{\gamma} \in \widetilde{\Gamma}_N(\boldsymbol{v}_0, \boldsymbol{v}_1)} J_N[\boldsymbol{v}_0, \boldsymbol{v}_1](\boldsymbol{\gamma}) = |\boldsymbol{v}_1 - \boldsymbol{v}_0|.$$

Of course, it suffices to consider $\boldsymbol{v}_0 \neq \boldsymbol{v}_1$. Up to a rotation, we may assume without loss of generality that

$$\boldsymbol{v}_g = \frac{\boldsymbol{v}_0 + \boldsymbol{v}_1}{|\boldsymbol{v}_0 + \boldsymbol{v}_1|} = \boldsymbol{e}_N \quad \text{and} \quad \boldsymbol{v}_0, \boldsymbol{v}_1 \in \operatorname{span}\{\boldsymbol{e}_{N-1}, \boldsymbol{e}_N\}.$$

Since $\boldsymbol{v}_g$ is the geodesic midpoint and $\boldsymbol{v}_0 \cdot \boldsymbol{v}_1 \geq 0$, there exists $\theta_0 \in (0, \pi/4]$ such that

$$\boldsymbol{v}_0 = (0, \dots, 0, \sin\theta_0, \cos\theta_0) \quad \text{and} \quad \boldsymbol{v}_1 = (0, \dots, 0, -\sin\theta_0, \cos\theta_0).$$

Then

$$|\boldsymbol{v}_1 - \boldsymbol{v}_0| = 2\sin\theta_0. \tag{4-1}$$

A curve which attains the equality in (4-3) is easily obtained: it is just the geodesic with respect to the standard metric of $\mathbb{S}^{N-1}$.

Lemma 4.2. *Let $\boldsymbol{\gamma}_{\min}(t) = \big(0, \dots, 0, \sin((1-2t)\theta_0), \cos((1-2t)\theta_0)\big)$. Then $J(\boldsymbol{\gamma}_{\min}) = 2\sin\theta_0$.*

After the above-mentioned rotation, $\mathbb{S}^{N-1}_+$ is transformed into a geodesic simplex $\mathcal{T}$ in $\mathbb{S}^{N-1}$. We consider a larger set of curves: let $\mathcal{P}_N(\boldsymbol{v}_0, \boldsymbol{v}_1)$ be given by

$$\mathcal{P}_N(\boldsymbol{v}_0, \boldsymbol{v}_1) = \{\boldsymbol{v} \in \mathbb{S}^{N-1} : \boldsymbol{v} \cdot \boldsymbol{v}_0 \geq 0,\ \boldsymbol{v} \cdot \boldsymbol{v}_1 \geq 0\}$$

and let

$$\Gamma_N(\boldsymbol{v}_0, \boldsymbol{v}_1) = \{\boldsymbol{\gamma} \in W^{1,1}((0,1); \mathcal{P}_N(\boldsymbol{v}_0, \boldsymbol{v}_1)) : \boldsymbol{\gamma}(0) = \boldsymbol{v}_0,\ \boldsymbol{\gamma}(1) = \boldsymbol{v}_1\}.$$

Then

$$J_N[\boldsymbol{v}_0, \boldsymbol{v}_1](\boldsymbol{\gamma}) = \int_0^1 \gamma^N(t)|\dot{\boldsymbol{\gamma}}(t)|\,dt \quad \text{for } \boldsymbol{\gamma} = (\gamma^1, \dots, \gamma^N) \in \Gamma_N(\boldsymbol{v}_0, \boldsymbol{v}_1)\,.$$

Hence, recalling (4-1) and Lemma 4.2, it suffices to prove that

$$\inf_{\boldsymbol{\gamma} \in \Gamma_N(\boldsymbol{v}_0, \boldsymbol{v}_1)} J_N[\boldsymbol{v}_0, \boldsymbol{v}_1](\boldsymbol{\gamma}) \geq 2\sin\theta_0. \tag{4-2}$$

We now show that the problem in $\mathbb{S}^{N-1}$ may be reduced to the same problem in $\mathbb{S}^2$. Let

$$\bar{\boldsymbol{v}}_i = (0, (-1)^i \sin\theta_0, \cos\theta_0), \quad i = 0, 1,$$

denote the projection of $\boldsymbol{v}_i$ onto the three-dimensional subspace $\operatorname{span}\{\boldsymbol{e}_{N-2}, \boldsymbol{e}_{N-1}, \boldsymbol{e}_N\}$.

Lemma 4.3. *Let $N \geq 4$. Then*

$$\inf_{\boldsymbol{\gamma} \in \Gamma_N(\boldsymbol{v}_0, \boldsymbol{v}_1)} J_N[\boldsymbol{v}_0, \boldsymbol{v}_1](\boldsymbol{\gamma}) \geq \inf_{\boldsymbol{\gamma} \in \Gamma_3(\bar{\boldsymbol{v}}_0, \bar{\boldsymbol{v}}_1)} J_3[\bar{\boldsymbol{v}}_0, \bar{\boldsymbol{v}}_1](\boldsymbol{\gamma}).$$

Proof. For any $\boldsymbol{\gamma} \in \Gamma_N(\boldsymbol{v}_0, \boldsymbol{v}_1)$, consider the curve

$$\tilde{\boldsymbol{\gamma}} = \big(0, \dots, 0, \sqrt{(\gamma^1)^2 + \cdots + (\gamma^{N-2})^2}, \gamma^{N-1}, \gamma^N\big).$$

Clearly $\tilde{\boldsymbol{\gamma}} \in W^{1,1}((0,1), \mathbb{S}^{N-1})$. Since $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ belong to $\operatorname{span}\{\boldsymbol{e}_{N-1}, \boldsymbol{e}_N\}$ and the projections of $\tilde{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$ onto $\operatorname{span}\{\boldsymbol{e}_{N-1}, \boldsymbol{e}_N\}$ coincide, $\tilde{\boldsymbol{\gamma}} \in W^{1,1}((0,1); \mathscr{P}_N(\boldsymbol{v}_0, \boldsymbol{v}_1))$ and the end-point conditions are satisfied. Therefore $\tilde{\boldsymbol{\gamma}} \in \Gamma_N(\boldsymbol{v}_0, \boldsymbol{v}_1)$. In addition, letting

$$\delta = (\gamma^1, \dots, \gamma^{N-2}),$$

we may apply the chain rule [Ambrosio and Dal Maso 1990, Corollary 3.2]: since $\delta \in W^{1,1}((0,1); \mathbb{R}^{N-2})$ and $f(x) = |x|$ is a Lipschitz function with $f(0) = 0$, we have $|\delta| = f \circ \delta \in W^{1,1}((0,1); \mathbb{R})$, for almost every $t \in (0,1)$ the restriction of f to the affine space

$$T_t^\delta := \{y \in \mathbb{R}^{N-2} : y = \delta(t) + \eta\dot{\delta}(t) \text{ for some } \eta \in \mathbb{R}\}$$

is differentiable at $\delta(t)$, and finally

$$\frac{d}{dt}|\delta| = \nabla(f|_{T_t^\delta})(\delta(t)) \cdot \dot{\delta}(t) \quad \text{for a.e. } t \in (0,1).$$

Since the Lipschitz constant of f is 1, we get that $\big|(d/dt)|\delta|\big| \le |(d/dt)\delta|$. Hence $|(d/dt)\tilde{\boldsymbol{\gamma}}| \le |(d/dt)\boldsymbol{\gamma}|$, which implies that $J_N[\boldsymbol{v}_0, \boldsymbol{v}_1](\tilde{\boldsymbol{\gamma}}) \le J_N[\boldsymbol{v}_0, \boldsymbol{v}_1](\boldsymbol{\gamma})$, since $\tilde{\gamma}^N = \gamma^N$. Arguing as above, we also see that

$$\bar{\boldsymbol{\gamma}} = (\sqrt{(\gamma^1)^2 + \cdots + (\gamma^{N-2})^2}, \gamma^{N-1}, \gamma^N)$$

belongs to $\Gamma_3(\bar{\boldsymbol{v}}_0, \bar{\boldsymbol{v}}_1)$. Since $J_N[\boldsymbol{v}_0, \boldsymbol{v}_1](\tilde{\boldsymbol{\gamma}}) = J_3[\bar{\boldsymbol{v}}_0, \bar{\boldsymbol{v}}_1](\bar{\boldsymbol{\gamma}})$, the proof is complete. □

Hereafter we let

$$\boldsymbol{v}_i := \bar{\boldsymbol{v}}_i, \quad J := J_3[\boldsymbol{v}_0, \boldsymbol{v}_1], \quad \mathscr{P} := \mathscr{P}_3(\boldsymbol{v}_0, \boldsymbol{v}_1), \quad \Gamma := \Gamma_3(\boldsymbol{v}_0, \boldsymbol{v}_1).$$

Because of (4-2) and Lemma 4.3, it suffices to prove that

$$\inf_{\boldsymbol{\gamma} \in \Gamma} J(\boldsymbol{\gamma}) \ge 2 \sin\theta_0. \tag{4-3}$$

Proving (4-3) is far from trivial, both since the functional is genuinely nonconvex (see Lemmas 4.9 and 4.10) and since the curves are constrained to an octant of the sphere. However, it is exactly for this reason that the lower bound holds:

Remark 4.4. In the extremal case $\theta_0 = \pi/4$, there are exactly two paths $\boldsymbol{\gamma}$ such that $J(\boldsymbol{\gamma}) = 2\sin\theta_0$: the one given in Lemma 4.2, and the one which coincides with $\partial\mathscr{P}$ (see Section 1 or Lemma 4.14 with $\varphi_0 = 0$ and $\varphi_1 = \pi/2$). If the constraint is removed, the lower bound (4-3) does not hold any more: for instance, the curve

$$\boldsymbol{\gamma}(t) := \begin{cases} (0, \sin\theta, \cos\theta),\ \theta = \theta_0 + 3t(\pi/2 - \theta_0) \in (\theta_0, \pi/2) & \text{if } 0 \le t \le \frac{1}{3}, \\ (\sin\varphi, \cos\varphi, 0),\ \varphi = 3\pi(t - \frac{1}{3}) \in (0, \pi) & \text{if } \frac{1}{3} < t \le \frac{2}{3}, \\ (0, -\sin\theta, \cos\theta),\ \theta = \pi/2 + 3(t - \frac{2}{3})(\theta_0 - \pi/2) \in (\theta_0, \pi/2) & \text{if } \frac{2}{3} < t \le 1 \end{cases}$$

is such that

$$J(\boldsymbol{\gamma}) = 2\int_0^{1/3} \cos\theta\, |\dot\theta|\, dt = 2\int_{\theta_0}^{\pi/2} \cos\theta\, d\theta = 2(1-\sin\theta_0),$$

hence $J(\boldsymbol{\gamma}) = 2(1-\sin\theta_0) < 2\sin\theta_0$ if $\theta_0 > \pi/6$.

We will often use spherical coordinates centered at $(0,0,1)$:

$$X(\varphi,\theta) := (\sin\varphi\sin\theta, \cos\varphi\sin\theta, \cos\theta). \tag{4-4}$$

In this case $\boldsymbol{v}_0 = X(0,\theta_0)$, $\boldsymbol{v}_1 = X(\pi,\theta_0)$, the functional reads

$$J(\boldsymbol{\gamma}) = \int_0^1 \cos\theta(t)\sqrt{(\dot\theta(t))^2 + (\dot\varphi(t))^2\sin^2\theta(t)}\, dt, \quad \text{where } \boldsymbol{\gamma}(t) = X(\varphi(t),\theta(t)), \tag{4-5}$$

and the constraint $\boldsymbol{\gamma}(t) \in \mathcal{P}$ is equivalent to

$$\theta(t) \in [0,\pi/2], \quad \theta(t) \le \arctan\frac{1}{\tan\theta_0\,|\cos\varphi(t)|} =: \theta^*(\varphi(t)). \tag{4-6}$$

It is convenient to cut-off from $\mathcal{P}$ a neighborhood of $z=0$: in this way, the new constraint has a smooth boundary and the density of J does not degenerate. Thus, let $\theta_\varepsilon^* \in C^\infty(\mathbb{R})$ be such that

$$\begin{aligned} &\theta_\varepsilon^* \text{ is } \pi\text{-periodic, even w.r.t. } \pi/2, \text{ increasing in } (0,\pi/2),\\ &\theta_\varepsilon^*(\varphi) = \theta^*(\varphi) \text{ if } |\pi/2-\varphi| \ge \varepsilon,\ \theta_\varepsilon^* < \pi/2, \text{ and } |(\theta_\varepsilon^*)'| \le C \end{aligned} \tag{4-7}$$

for some ε-independent positive constant C. Note that here and after primes denote differentiation with respect to φ, and that the latter property of θ_ε^* may be fulfilled since θ^* is Lipschitz-continuous. Now let

$$\begin{aligned} \mathcal{P}_\varepsilon &:= \{X(\varphi,\theta) : \varphi \in [0,2\pi],\ 0 \le \theta \le \theta_\varepsilon^*(\varphi)\},\\ \Gamma_\varepsilon &:= \{\boldsymbol{\gamma} \in W^{1,1}((0,1);\mathcal{P}_\varepsilon) : \boldsymbol{\gamma}(0) = \boldsymbol{v}_0,\ \boldsymbol{\gamma}(1) = \boldsymbol{v}_1\}. \end{aligned}$$

In what follows, $\omega(\varepsilon)$ denotes a generic positive universal function which goes to zero as $\varepsilon \to 0$. The next lemma shows that we may equivalently work on $\mathcal{P}_\varepsilon$:

Lemma 4.5. *Assume that*

$$\inf_{\boldsymbol{\gamma}\in\Gamma_\varepsilon} J(\boldsymbol{\gamma}) \ge 2\sin\theta_0 - \omega(\varepsilon). \tag{4-8}$$

Then (4-3) *holds true, and therefore so does Theorem 4.1.*

Proof. Given $\boldsymbol{\gamma} \in \Gamma$, we replace the parts of $\boldsymbol{\gamma}$ which enter into $\mathcal{P}\setminus\mathcal{P}_\varepsilon$ by arcs of $\partial\mathcal{P}_\varepsilon$. More precisely, let

$$I_\varepsilon = \{t \in (0,1) : \boldsymbol{\gamma}(t) \in \mathcal{P}\setminus\mathcal{P}_\varepsilon\}.$$

Since the spherical coordinates (4-4) are a bijection away from the north pole $(0,0,1)$, in I_ε we may define $\varphi(t)$ and $\theta(t)$ through $\boldsymbol{\gamma}(t) =: X(\varphi(t),\theta(t))$. Then we let

$$\boldsymbol{\gamma}_\varepsilon(t) = \begin{cases} \boldsymbol{\gamma}(t) & \text{if } t \notin I_\varepsilon,\\ (\varphi(t), \theta_\varepsilon^*(\varphi(t))) & \text{if } t \in I_\varepsilon. \end{cases}$$

It follows from (4-7) that

$$\left|\frac{\pi}{2} - \theta_\varepsilon^*(\varphi(t))\right| = \omega(\varepsilon). \tag{4-9}$$

We may now estimate

$$J(\boldsymbol{\gamma}) - J(\boldsymbol{\gamma}_\varepsilon) \overset{(4\text{-}5)}{\geq} -\int_{I_\varepsilon} \cos\theta_\varepsilon^*(\varphi)|\dot\varphi|\sqrt{(\theta_\varepsilon^{*'})^2 + \sin^2\theta_\varepsilon^*(\varphi)}\,dt \underset{(4\text{-}9)}{\overset{(4\text{-}7)}{\geq}} -\omega(\varepsilon)\int_0^1 |\dot\varphi|\,dt.$$

Therefore

$$J(\boldsymbol{\gamma}) \overset{(4\text{-}8)}{\geq} 2\sin\theta_0 - \omega(\varepsilon)\left(1 + \int_0^1 |\dot\varphi|\,dt\right).$$

Passing to the limit as $\varepsilon \to 0$, the arbitrariness of $\boldsymbol{\gamma} \in \Gamma$ yields (4-3). □

The rest of the section will be concerned with the proof of (4-8). Let

$$\Gamma_\varepsilon(\boldsymbol{w}_0, \boldsymbol{w}_1) := \{\boldsymbol{\gamma} \in W^{1,1}((0,1); \mathcal{P}_\varepsilon) : \boldsymbol{\gamma}(0) = \boldsymbol{w}_0,\ \boldsymbol{\gamma}(1) = \boldsymbol{w}_1\} \quad \text{for } \boldsymbol{w}_0, \boldsymbol{w}_1 \in \mathcal{P}_\varepsilon.$$

Lemma 4.6. *For any $\boldsymbol{w}_0, \boldsymbol{w}_1 \in \mathcal{P}_\varepsilon$, there exists a minimizer $\boldsymbol{\gamma}$ of J in $\Gamma_\varepsilon(\boldsymbol{w}_0, \boldsymbol{w}_1)$. Furthermore $\boldsymbol{\gamma}$ lies in $W^{1,\infty}((0,1); \mathbb{R}^3)$, satisifes $\gamma^3|\dot{\boldsymbol{\gamma}}| = J(\boldsymbol{\gamma})$ a.e. in $[0,1]$, and is also a minimizer of*

$$E(\boldsymbol{\gamma}) := \int_0^1 (\gamma^3(t))^2|\dot{\boldsymbol{\gamma}}(t)|^2\,dt$$

among all $\boldsymbol{\gamma} \in \Gamma_\varepsilon(\boldsymbol{w}_0, \boldsymbol{w}_1) \cap H^1((0,1); \mathbb{R}^3)$.

Though we could appeal to general results on geodesics for Riemannian manifolds with boundary (see [Alexander et al. 1993] and the references therein), we prefer to give a self-contained proof.

Proof. We preliminarily observe that

$$\text{for all } \boldsymbol{\gamma} \in \Gamma_\varepsilon(\boldsymbol{w}_0, \boldsymbol{w}_1), \text{ there exists } \tilde{\boldsymbol{\gamma}} \in \Gamma_\varepsilon(\boldsymbol{w}_0, \boldsymbol{w}_1) \cap W^{1,\infty}((0,1); \mathbb{R}^3)$$
$$\text{such that } \tilde\gamma^3(t)|\dot{\tilde{\boldsymbol{\gamma}}}(t)| = L := J(\boldsymbol{\gamma}) \text{ for a.e. } t \in [0,1]. \tag{4-10}$$

To see this, let

$$s(t) = \frac{1}{L}\int_0^t \gamma^3(\tau)|\dot{\boldsymbol{\gamma}}(\tau)|\,d\tau. \tag{4-11}$$

Obviously $s \in W^{1,1}([0,1]; [0,1])$, s is nondecreasing, and $s(t_1) = s(t_2)$ if and only if $\boldsymbol{\gamma}(t) = \boldsymbol{\gamma}(t_1)$ in $[t_1, t_2]$. Therefore, for any $\sigma \in [0,1]$, either there exists a unique $t(\sigma)$ such that $s(t(\sigma)) = \sigma$, or there exists an interval I_σ such that $s(t) = \sigma$ for all $t \in I_\sigma$, and in this case we let, for example, $t(\sigma) = \inf I_\sigma$, so that again $s(t(\sigma)) = \sigma$. Now let $\tilde{\boldsymbol{\gamma}}(\sigma) := \boldsymbol{\gamma}(t(\sigma))$. By construction,

$$\boldsymbol{\gamma}(t) = \tilde{\boldsymbol{\gamma}}(s(t)) \quad \text{for all } t \in [0,1]. \tag{4-12}$$

Note that $\tilde{\boldsymbol{\gamma}} \in W^{1,\infty}((0,1);\mathcal{P}_\varepsilon)$. Indeed,

$$\begin{aligned}|\tilde{\boldsymbol{\gamma}}(\sigma_1)-\tilde{\boldsymbol{\gamma}}(\sigma_2)| &= |\boldsymbol{\gamma}(t(\sigma_1))-\boldsymbol{\gamma}(t(\sigma_2))| \le \int_{t(\sigma_1)}^{t(\sigma_2)} |\dot{\boldsymbol{\gamma}}(\tau)|\,d\tau \\ &\overset{(4\text{-}11)}{\le} \frac{L}{\inf_{\tau\in[t(\sigma_1),t(\sigma_2)]}\gamma^3(\tau)}|s(t(\sigma_1))-s(t(\sigma_2))| \\ &\overset{(4\text{-}7)}{\le} \frac{L}{\omega(\varepsilon)}|\sigma_1-\sigma_2|. \end{aligned} \tag{4-13}$$

Hence, it follows from (4-12) and the chain rule formula given in [Ambrosio et al. 2000, Theorem 3.101] that

$$\dot{\boldsymbol{\gamma}}(t) = \frac{d\tilde{\boldsymbol{\gamma}}}{ds}(s(t))\dot{s}(t) \quad \text{in } L^1((0,1)). \tag{4-14}$$

Therefore,

$$L = \int_0^1 \gamma^3(t)\,|\dot{\boldsymbol{\gamma}}(t)|\,dt \overset{(4\text{-}14)}{\underset{(4\text{-}12)}{=}} \int_0^1 \tilde{\gamma}^3(s(t))\left|\frac{d\tilde{\boldsymbol{\gamma}}}{ds}(s(t))\right|\dot{s}(t)\,dt = \int_0^1 \tilde{\gamma}^3(s)\left|\frac{d\tilde{\boldsymbol{\gamma}}}{ds}(s)\right|ds. \tag{4-15}$$

On the other hand, given $s\in[0,1]$ and $\varepsilon>0$, let $s_1, s_2 \in [0,1]$ with $|s_i - s| < \varepsilon$. Then

$$\tilde{\gamma}^3(s)|\tilde{\boldsymbol{\gamma}}(s_1)-\tilde{\boldsymbol{\gamma}}(s_2)| \overset{(4\text{-}13)}{\le} \frac{\tilde{\gamma}^3(s)}{\inf_{\tau\in[t(s_1),t(s_2)]}\gamma^3(\tau)}L|s_2-s_1|. \tag{4-16}$$

If $\tau\in[t(s_1),t(s_2)]$, then, by the monotonicity of s and since $s(\tau(s))=s$, we have $s(\tau)\in[s_1,s_2]$. Hence

$$\inf_{\tau\in[t(s_1),t(s_2)]}\gamma^3(\tau) \overset{(4\text{-}12)}{=} \inf_{\tau\in[t(s_1),t(s_2)]}\tilde{\gamma}^3(s(\tau)) \ge \inf_{s\in[s_1,s_2]}\tilde{\gamma}^3(s). \tag{4-17}$$

Combining (4-16) and (4-17) and passing to the limit as $\varepsilon\to 0$, we obtain

$$\tilde{\gamma}^3(s)\left|\frac{d\tilde{\boldsymbol{\gamma}}}{ds}(s)\right| \le L \quad \text{for a.e. } s\in[0,1],$$

which together with (4-15) concludes the proof of the claim (4-10).

We consider the functional E defined on $G_\varepsilon(\boldsymbol{w}_0,\boldsymbol{w}_1) := \Gamma_\varepsilon(\boldsymbol{w}_0,\boldsymbol{w}_1)\cap H^1((0,1);\mathbb{R}^3)$. By the Cauchy–Schwarz inequality,

$$(J(\boldsymbol{\gamma}))^2 \le E(\boldsymbol{\gamma}) \quad \text{for all } \boldsymbol{\gamma}\in G_\varepsilon(\boldsymbol{w}_0,\boldsymbol{w}_1). \tag{4-18}$$

Hence $\inf E(\boldsymbol{\gamma}) \ge \inf(J(\boldsymbol{\gamma}))^2$. On the other hand, let $\boldsymbol{\gamma}_n$ be a minimizing sequence for J, and let $\tilde{\boldsymbol{\gamma}}_n$ be as given by (4-10): then $E(\tilde{\boldsymbol{\gamma}}_n) = (J(\boldsymbol{\gamma}_n))^2$, which means that $\inf E \le \inf J^2$. Therefore,

$$\inf_{\boldsymbol{\gamma}\in G_\varepsilon(\boldsymbol{w}_0,\boldsymbol{w}_1)} E(\boldsymbol{\gamma}) = \inf_{\boldsymbol{\gamma}\in\Gamma_\varepsilon(\boldsymbol{w}_0,\boldsymbol{w}_1)} (J(\gamma))^2.$$

The inf on the left-hand side is attained. Indeed, let $\boldsymbol{\gamma}_n$ be a minimizing sequence. By the coercivity of E ensured by the definition of $\mathcal{P}_\varepsilon$, a subsequence (not relabeled) exists such that $\boldsymbol{\gamma}_n\to\boldsymbol{\gamma}$ weakly in $H^1((0,1);\mathcal{P}_\varepsilon)$ and in $C([0,1];\mathcal{P}_\varepsilon)$. Therefore $E(\boldsymbol{\gamma}) \le \liminf_{n\to+\infty} E(\boldsymbol{\gamma}_n)$.

Let $\boldsymbol{\gamma}_0$ be a minimizer of E, and let $\tilde{\boldsymbol{\gamma}}_0$ be as given by (4-10). Then

$$E(\boldsymbol{\gamma}_0) \overset{(4\text{-}18)}{\geq} (J(\boldsymbol{\gamma}_0))^2 = (J(\tilde{\boldsymbol{\gamma}}_0))^2 = E(\tilde{\boldsymbol{\gamma}}_0),$$

that is, $\tilde{\boldsymbol{\gamma}}_0$ is also a minimizer of E, and

$$(J(\boldsymbol{\gamma}))^2 = (J(\tilde{\boldsymbol{\gamma}}))^2 = E(\tilde{\boldsymbol{\gamma}}) \geq E(\tilde{\boldsymbol{\gamma}}_0) = (J(\tilde{\boldsymbol{\gamma}}_0))^2 \quad \text{for all } \boldsymbol{\gamma} \in \Gamma_\varepsilon(\boldsymbol{w}_0, \boldsymbol{w}_1),$$

hence $\tilde{\boldsymbol{\gamma}}_0$ (or $\boldsymbol{\gamma}_0$) is a minimizer of J. Therefore J has a minimizer too. □

The rest of the section is concerned with estimating the length of a minimizer of J in Γ_ε as given by Lemma 4.6, a *shortest path* in what follows. Our first observation concerns those shortest paths which pass through the north pole:

Lemma 4.7. *If a shortest path $\boldsymbol{\gamma}$ passes through $(0, 0, 1)$, then $J(\boldsymbol{\gamma}) \geq 2 \sin\theta_0$.*

Proof. Let t_0 and t_1 be the first, respectively the last, time in which $\boldsymbol{\gamma} = (0, 0, 1)$. Then, using the spherical coordinates (4-4),

$$\begin{aligned} J(\boldsymbol{\gamma}) &\geq \int_0^{t_0} \cos\theta \sqrt{(\dot\theta)^2 + (\dot\varphi)^2 \sin^2\theta}\, dt + \int_{t_1}^1 \cos\theta \sqrt{(\dot\theta)^2 + (\dot\varphi)^2 \sin^2\theta}\, dt \\ &\geq \int_0^{t_0} \cos\theta |\dot\theta|\, dt + \int_{t_1}^1 \cos\theta |\dot\theta|\, dt = \int_0^{t_0} \left|\frac{d}{dt} \sin\theta\right| dt + \int_{t_1}^1 \left|\frac{d}{dt} \sin\theta\right| dt, \end{aligned}$$

and the lemma follows, since $\theta(t_0) = \theta(t_1) = 0$ and $\theta(0) = \theta(1) = \theta_0$. □

We may therefore restrict our attention to shortest paths not passing through the north pole. There, the spherical coordinates (4-4) are a diffeomorphism. In fact, we may also restrict our attention to those paths for which φ is nondecreasing and which are symmetric with respect to $\varphi = \pi/2$. In what follows, we shall call them *symmetric* shortest paths.

Lemma 4.8. *Let $\boldsymbol{\gamma} = X(\varphi, \theta)$ be a shortest path not passing through $(0, 0, 1)$. Then $\varphi \in [0, \pi]$ and φ is nondecreasing. Moreover, there exists a shortest path $\tilde{\boldsymbol{\gamma}} = X(\tilde\varphi, \tilde\theta)$ not passing through $(0, 0, 1)$ such that $\tilde\varphi$ is symmetric with respect to $\pi/2$:*

$$\{(\tilde\varphi(t), \tilde\theta(t)) : t \in [0, 1]\} = \{(\pi - \tilde\varphi(t), \tilde\theta(t)) : t \in [0, 1]\}.$$

Proof. Without loss of generality, $\varphi(0) = 0$ and $\varphi(1) = (2k + 1)\pi$ with $k \geq 0$. It is straightforward to see that $\max\{\varphi, 0\}$ and $\min\{\varphi, \pi\}$ both decrease the value of J, hence $k = 0$. Analogously, if $t_0 < t_1 < t_2$ are such that $\varphi(t_1) < \varphi(t_2) = \varphi(t_0)$, then replacing φ with $\varphi(t_0)$ in (t_0, t_2) decreases the value of J. Therefore, φ is nondecreasing along a shortest path.

In order to construct $\tilde{\boldsymbol{\gamma}}$, we claim that

$$J_1 := \int_0^{t_*} \cos\theta \sqrt{(\dot\theta)^2 + (\dot\varphi)^2 \sin^2\theta}\, dt = \int_{t_*}^1 \cos\theta \sqrt{(\dot\theta)^2 + (\dot\varphi)^2 \sin^2\theta}\, dt =: J_2 \tag{4-19}$$

for any $t^* \in (0,1)$ such that $\varphi(t_*) = \pi/2$. Suppose by contradiction that (4-19) does not hold. Then, without loss of generality, we can suppose $J_1 < J_2$. We define $\tilde{\boldsymbol{\gamma}}(t) = X(\tilde{\varphi}(t), \tilde{\theta}(t))$, where

$$\tilde{\theta}(t) = \begin{cases} \theta(t) & \text{if } t \le t_*, \\ \theta\left(\dfrac{t_*(1-t)}{1-t_*}\right) & \text{if } t > t_*, \end{cases} \quad \text{and} \quad \tilde{\varphi}(t) = \begin{cases} \varphi(t) & \text{if } t \le t_*, \\ \pi - \varphi\left(\dfrac{t_*(1-t)}{1-t_*}\right) & \text{if } t > t_*. \end{cases} \tag{4-20}$$

Then, by letting $\hat{t} = t_*(1-t)/(1-t_*)$ and using the 1-homogeneity of the integrands with respect to t, we see that

$$\begin{aligned} J(\tilde{\boldsymbol{\gamma}}) &= J_1 + \frac{t_*}{1-t_*} \int_{t_*}^1 \cos\theta(\hat{t}) \sqrt{(\dot\theta(\hat{t}))^2 + (\dot\varphi(\hat{t}))^2 \sin^2\theta(\hat{t})}\, dt \\ &= J_1 + \int_0^{t_*} \cos\theta(\hat{t}) \sqrt{(\dot\theta(\hat{t}))^2 + (\dot\varphi(\hat{t}))^2 \sin^2\theta(\hat{t})}\, d\hat{t} = 2J_1 \\ &< J_1 + J_2 = J(\boldsymbol{\gamma}), \end{aligned} \tag{4-21}$$

a contradiction, since $\boldsymbol{\gamma}$ is a shortest path. Therefore (4-19) holds. Then, defining $\tilde{\boldsymbol{\gamma}}(t) = X(\tilde{\varphi}(t), \tilde{\theta}(t))$ as in (4-20), it follows from (4-21) that $J(\tilde{\boldsymbol{\gamma}}) = 2J_1 = J_1 + J_2 = J(\boldsymbol{\gamma})$, hence $\tilde{\boldsymbol{\gamma}}$ is also a shortest path. □

We now characterize arcs of shortest paths contained in $\mathring{\mathcal{P}}_\varepsilon$.

Lemma 4.9. *Let $\boldsymbol{\gamma}$ be a shortest path not passing through $(0,0,1)$ and let (t_0, t_1) be an interval in which $\boldsymbol{\gamma}|_{(t_0,t_1)} \subset \mathring{\mathcal{P}}_\varepsilon$. Then*

$$\frac{\cos\theta(t)\sin^2\theta(t)\dot\varphi(t)}{\sqrt{(\dot\theta(t))^2 + (\dot\varphi(t))^2 \sin^2\theta(t)}} = K \quad \text{for all } t \in (t_0, t_1) \tag{4-22}$$

for some $K \in [0, 1/2]$. If $K = 0$, then φ is constant. If $K > 0$, then φ is strictly increasing, the function

$$(\varphi(t_0), \varphi(t_1)) =: I \ni \varphi \mapsto \theta(t(\varphi)) \tag{4-23}$$

is a smooth solution of

$$\theta'' \sin\theta\cos\theta = (\theta')^2(\cos^2\theta + \cos(2\theta)) + \cos(2\theta)\sin^2\theta \tag{4-24}$$

with

$$\sin^2\theta\,(\cos^2\theta\sin^2\theta - K^2) = K^2(\theta')^2, \tag{4-25}$$

and

$$J(\boldsymbol{\gamma}\,\chi_{(t_0,t_1)}) = \int_{\varphi(t_0)}^{\varphi(t_1)} \cos\theta \sqrt{(\theta')^2 + \sin^2\theta}\, d\varphi. \tag{4-26}$$

Proof. Up to a linear reparametrization, $\boldsymbol{\gamma}$ is also a minimizer of J in $\Gamma_\varepsilon(\boldsymbol{\gamma}(t_0), \boldsymbol{\gamma}(t_1))$. Hence, by Lemma 4.6, it is also a minimizer of E in $\Gamma_\varepsilon(\boldsymbol{\gamma}(t_0), \boldsymbol{\gamma}(t_1)) \cap H^1((0,1);\mathbb{R}^3)$. Since it does not touch the north pole, we can write $\boldsymbol{\gamma} = X(\varphi, \theta)$ with φ and θ Lipschitz, and

$$E(\boldsymbol{\gamma}\,\chi_{(t_0,t_1)}) = \int_{t_0}^{t_1} \cos^2\theta\,(\dot\theta^2 + \sin^2\theta\dot\varphi^2)\, dt.$$

Taking the first variation with respect to φ, we obtain $\sin^2\theta\cos^2\theta\dot\varphi = H$, and, recalling that $\gamma^3|\dot{\boldsymbol{\gamma}}|$ is constant, (4-22) follows. Since $\sin\theta > 0$ ($\boldsymbol{\gamma}$ does not cross the north pole), $\cos\theta > 0$ ($\boldsymbol{\gamma} \in P_\varepsilon$), and

φ is nondecreasing (by Lemma 4.8), we see that $K \geq 0$. If $K = 0$, then φ is constant. If $K > 0$, then $\dot{\varphi} > 0$ in (t_0, t_1) and we may use φ as independent variable: letting θ be as in (4-23), we have $\theta' = d\theta/d\varphi = \dot{\theta}/\dot{\varphi} \in L^\infty((t_0, t_1))$ (because of (4-22)). Then (4-26) follows at once from (4-5) and the definition of K may be rewritten as

$$\frac{\cos\theta \sin^2\theta}{\sqrt{(\theta')^2 + \sin^2\theta}} = K, \tag{4-27}$$

which is equivalent to (4-25). From (4-25) one sees immediately that $K \leq 1/2$. Differentiating (4-27), we obtain (4-24) in the sense of distributions, and a bootstrap argument starting from $\theta \in W^{1,\infty}((t_0, t_1))$ yields smoothness. □

If $\boldsymbol{\gamma} = X(\varphi, \theta(\varphi)) : (t_0, t_1) \to \mathbb{S}^2$ is a curve which does not pass through $(0, 0, 1)$ and such that $\varphi \in I := (\varphi(t_0), \varphi(t_1))$ is strictly increasing, then, following (4-26), we hereafter write (with a slight abuse of notation)

$$J(\boldsymbol{\gamma}\chi_{(t_0,t_1)}) = J_I(\theta) := \int_I \cos\theta \sqrt{(\theta')^2 + \sin^2\theta}\, d\varphi, \quad J(\theta) := J_{(0,\pi)}(\theta).$$

In view of Lemma 4.9, it is convenient to state a few properties of the solutions to (4-24), some of which are visualized in Figure 1.

Lemma 4.10. *Let θ be any solution of (4-24) such that $\theta \in (0, \pi/2)$ at some point of its domain. Then*:

(a) *θ is globally defined, periodic, and $\theta \in (0, \pi/2)$;*

(b) *within a period, θ has a unique local (and therefore global) maximum, $\theta_M \geq \pi/4$, and a unique local (and therefore global) minimum, $\theta_m = \pi/2 - \theta_M$, and it is symmetric with respect to its maximum (minimum) point;*

(c) *the period P is larger than π;*

(d) *the length of each interval in which $\theta \leq \pi/4$ is at least $\pi/\sqrt{2}$;*

(e) *θ' has a unique local (and therefore global) maximum and a unique local (and therefore global) minimum.*

Proof. (a) and (b) easily follow from (4-25) rewritten as

$$(\theta')^2 = \frac{1}{K^2}\sin^2\theta\,(\sin^2\theta\cos^2\theta - K^2) =: f_K(\theta), \quad K \in [0, 1/2] \tag{4-28}$$

and plotted in the phase space (see Figure 1). We just observe explicitly that, since $\theta' = 0$ at the extremal values of θ, we can characterize K from (4-25) as

$$K = \cos\theta_m \sin\theta_m = \cos\theta_M \sin\theta_M, \tag{4-29}$$

which explains why $\theta_M = \pi/2 - \theta_m$. Also (e) follows immediately from (4-28), since after differentiation we see that

$$2\theta'' = f_K'(\theta),$$

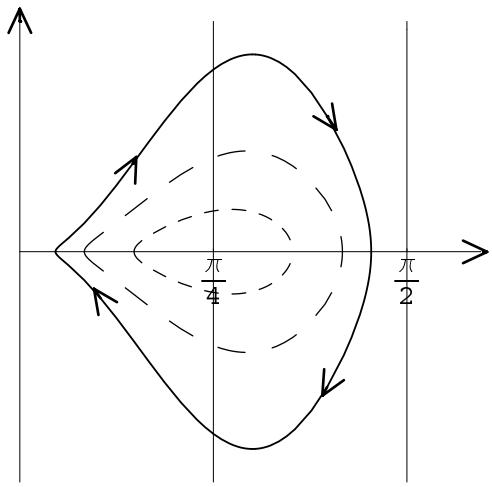


Figure 1. The phase plane (θ, θ').

whence the arrows in Figure 1.

To prove (c), we let φ_m and φ_M be a point of minimum and of maximum, respectively, chosen such that no local extremum exists in between. Then, in view of (b),

$$\frac{P}{2} = \int_{\varphi_m}^{\varphi_M} d\varphi \overset{(4\text{-}25)}{=} \int_{\varphi_m}^{\varphi_M} \frac{K\theta'}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\, d\varphi = \int_{\theta_m}^{\theta_M} \frac{K}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\, d\theta.$$

We now observe that

$$K \overset{(4\text{-}29)}{=} \cos\theta_M \sin\theta_M = \cos\theta_M \cos\theta_m \geq \cos\theta_M \cos\theta \quad \text{for all } \theta \in (\theta_m, \theta_M).$$

Therefore,

$$\frac{P}{2} \geq \cos\theta_M \int_{\theta_m}^{\theta_M} \frac{\cos\theta}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\, d\theta,$$

whose primitives may be computed explicitly:

$$\cos\theta_M \int \frac{\cos\theta}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\, d\theta = \frac{1}{2\sin\theta_M} \arcsin \frac{\sin^2\theta - 2\sin^2\theta_M\cos^2\theta_M}{\sin^2\theta\,|1 - 2\cos^2\theta_M|}.$$

Hence

$$\frac{P}{2} \geq \frac{1}{2\sin\theta_M}\left(\frac{\pi}{2} + \frac{\pi}{2}\right) = \frac{\pi}{2\sin\theta_M} > \frac{\pi}{2},$$

which proves (c).

To prove (d), let φ_m be a minimum point and let φ_* be the closest point to φ_m such that $\varphi_m \leq \varphi_*$ and $\theta(\varphi_*) = \pi/4$. By (b), the length of the interval within a period where $\theta \leq \pi/4$ is exactly

$$2\int_{\varphi_m}^{\varphi_*} d\varphi \overset{(4\text{-}25)}{=} 2\int_{\theta_m}^{\pi/4} \frac{K}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\, d\theta.$$

Since $\cos\theta \geq 1/\sqrt{2}$ if $\theta \in [0, \pi/4]$,

$$2\int_{\varphi_m}^{\varphi_*} d\varphi \geq \sqrt{2}\int_{\theta_m}^{\pi/4} \frac{K}{\sin\theta\cos\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\,d\theta.$$

The primitives of the right-hand side may be computed explicitly (via the substitution $z = \sin^2(2\theta)$):

$$\int \frac{K}{\sin\theta\cos\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\,d\theta = -\arctan\frac{2K\cos(2\theta)}{\sqrt{\sin^2(2\theta) - 4K^2}} + C. \tag{4-30}$$

After a substitution we get (d). □

Lemma 4.11. *Let $\boldsymbol{\gamma}$ be a symmetric shortest path not passing through $(0, 0, 1)$ and let $\boldsymbol{\gamma} = X(\varphi, \theta)$. If $t_1 \in [0, 1]$ is such that $\theta(t_1) < \pi/6$ and $\varphi(t_1) < \pi/2$, then $\theta(t) < \pi/6$ as long as $\varphi(t) < \pi - \varphi(t_1)$.*

Proof. Let $w = \sin\theta$ and let $t_2 > t_1$ be the first time in which $\varphi(t_2) = \pi - \varphi(t_1)$. We have

$$J(\boldsymbol{\gamma}\chi_{(t_1,t_2)}) = \int_{t_1}^{t_2} \sqrt{(\dot{w}(t))^2 + (\dot{\varphi}(t))^2 w(t)^2(1 - w(t)^2)}\,dt.$$

By assumption, $w(t_1) < \frac{1}{2}$. If there is an interval $\tilde{I} \subset [t_1, t_2]$ where $w(t) > \frac{1}{2}$, then a symmetrization of w with respect to $\frac{1}{2}$ would strictly decrease the value of J, since

$$(1 - w)^2(1 - (1 - w)^2) - w^2(1 - w^2) = 2w(1 - w)(1 - 2w) < 0 \quad \text{if } w \in (1/2, 1).$$

This contradicts that $\boldsymbol{\gamma}$ is a shortest path and thus proves the lemma. □

We are now ready to exclude shortest paths which are contained in $\mathring{\mathcal{P}}_\varepsilon$:

Lemma 4.12. *There is no symmetric shortest path $\boldsymbol{\gamma}$ not passing through $(0, 0, 1)$ such that $\boldsymbol{\gamma}((0, 1)) \subset \mathring{\mathcal{P}}_\varepsilon$.*

Proof. Assume for a contradiction that such a $\boldsymbol{\gamma}$ exists. We will argue that $J(\boldsymbol{\gamma}) > 2\sin\theta_0$, in contradiction with Lemma 4.2 (note that $\boldsymbol{\gamma}_{\min} \subseteq \mathcal{P}_\varepsilon$ for all ε).

Since φ has to travel from 0 to π, it can not be constant in $[0, 1]$. Then, it follows from Lemma 4.9 that $\boldsymbol{\gamma}(t) = X(\varphi(t), \theta(t))$, where $\varphi \mapsto \theta(t(\varphi))$ is a smooth solution of (4-24) such that $\theta(0) = \theta(\pi) = \theta_0$. Since $\boldsymbol{\gamma}$ is symmetric, we have $\theta'(\pi/2) = 0$. Because of (b) and (c) in Lemma 4.10, θ is monotone in $(0, \pi/2)$. Hence, letting $\theta_1 = \theta(\pi/2)$, we have

$$K \overset{(4\text{-}25)}{=} K(\theta_1) = \cos\theta_1\sin\theta_1 \quad \text{and} \quad \cos^2\theta\sin^2\theta \geq K^2 \quad \text{for all } t \in (0, 1). \tag{4-31}$$

We claim that $\theta_1 > \pi/4$. If not, it follows from (4-31) that $\theta_1 \leq \theta_0$. Hence θ is nonincreasing in $(0, \pi/2)$, and

$$\begin{aligned}\frac{\pi}{2} \overset{(4\text{-}25)}{=} \int_0^{\pi/2} \frac{K\theta'}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\,d\varphi &= \int_{\theta_1}^{\theta_0} \frac{K}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\,d\theta \\ &\leq \cos(\theta_1)\int_{\theta_1}^{\pi/4} \frac{K}{\sin\theta\cos\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\,d\theta. \end{aligned} \tag{4-32}$$

By (4-30), we would have $\pi \leq \pi\cos(\theta_1)$, a contradiction. Hence $\theta_1 > \pi/4$.

We note the obvious bound

$$\tfrac{1}{2}J(\theta) \geq \int_0^{\pi/2} \cos\theta \sin\theta \, d\varphi \overset{(4\text{-}31)}{\geq} \frac{\pi}{2} \sin\theta_1 \cos\theta_1.$$

Hence we are done if

$$\frac{\pi}{2} \sin\theta_1 \cos\theta_1 > \sin\theta_0,$$

that is, if

$$\theta_0 < \arcsin\Big(\frac{\pi}{2} \sin\theta_1 \cos\theta_1\Big) = \arcsin\Big(\frac{\pi}{4} \sin(2\theta_1)\Big). \tag{4-33}$$

We claim that (4-33) does hold. If not, recalling Lemma 4.11, we would have

$$\theta_0 \geq \max\Big\{\arcsin\Big(\frac{\pi}{4} \sin(2\theta_1)\Big), \frac{\pi}{6}\Big\} =: f(\theta_1).$$

Then, arguing as in (4-32), we write

$$\begin{aligned}
\frac{\pi}{2} &= \int_{\theta_0}^{\theta_1} \frac{K}{\sin\theta \sqrt{\cos^2\theta \sin^2\theta - K^2}} \, d\theta \\
&< \cos\frac{\pi}{6} \int_{f(\theta_1)}^{\pi/4} \frac{K}{\sin\theta\cos\theta \sqrt{\cos^2\theta \sin^2\theta - K^2}} \, d\theta + \cos\frac{\pi}{4} \int_{\pi/4}^{\theta_1} \frac{K}{\sin\theta\cos\theta \sqrt{\cos^2\theta \sin^2\theta - K^2}} \, d\theta \\
&= \frac{\sqrt{3}}{2} \arctan \frac{\sin\theta_1 \cos\theta_1 \cos(2f(\theta_1))}{\sqrt{\sin^2(2f(\theta_1)) - 4\sin^2\theta_1 \cos^2\theta_1}} + \frac{\sqrt{2}}{2}\frac{\pi}{2} =: F(\theta_1).
\end{aligned}$$

It is now a calculus exercise to check that F is increasing in $(\pi/4, \bar\theta) := (0, \frac{1}{2}(\pi - \arcsin(2/\pi)))$ and decreasing in $(\bar\theta, \pi/2)$: therefore F has a global maximum at $\bar\theta$, with $F(\bar\theta) < \pi/2$. Since this is impossible, (4-33) holds and the proof is complete. □

The rest of the section is concerned with estimating the length of candidate symmetric shortest paths which intersect $\partial\mathcal{P}_\varepsilon$ (and do not pass through the north pole). We firstly infer some properties of those candidate shortest paths which reach $\partial\mathcal{P}_\varepsilon$.

Lemma 4.13. *Let $\theta_0 < \pi/4$, let ε be sufficiently small, and let $\boldsymbol{\gamma} = X(\varphi, \theta)$ be a symmetric shortest path not intersecting the north pole. If $t_1 > 0$ exists such that $\boldsymbol{\gamma}(t_1) \in \partial\mathcal{P}_\varepsilon$ and $\boldsymbol{\gamma}(t) \in \mathring{\mathcal{P}}_\varepsilon$ in $[0, t_1)$, then*:

(i) $\theta(t) \geq \pi/6$ *for all* $t \in [0, t_1)$;

(ii) θ *is increasing in* $[0, t_1)$;

(iii) $\varphi(t_1) \leq \pi/2 - \varepsilon$.

Proof. (i) follows immediately from Lemma 4.11.

To prove (ii), we note that by Lemma 4.9, (4-22) holds in $[0, t_1)$. Let $\varphi_1 = \varphi(t_1)$. By symmetry, $\varphi_1 \leq \pi/2$. If $K = 0$, we would have $\varphi(t) = \varphi_1$ in $(0, t_1)$: since $\boldsymbol{\gamma}$ does not reach the north pole, this means that $\varphi_1 = 0$ and θ is increasing from θ_0 up to $\theta(t_1) = \pi/2 - \theta_0$. If instead $K > 0$, then (4-23) holds in $(0, \varphi_1)$. We will prove that $\theta' \geq 0$ in $(0, \varphi_1)$, which implies (ii). Assume by contradiction that $\theta' < 0$ somewhere in $(0, \varphi_1)$. Then, by Lemma 4.10(b), there exists $\varphi_2 \in (0, \varphi_1)$ such that $\theta(\varphi_2) = \theta_m$. By

Lemma 4.10(d) and since $\theta(\varphi_1) > \pi/4$, we have $\varphi_1 \geq \varphi_1 - \varphi_2 \geq \pi/(2\sqrt{2})$. Then, since θ_ε^* is increasing in $(0, \pi/2)$ and provided ε is sufficiently small,

$$\theta_1 := \theta(\varphi_1) = \theta_\varepsilon^*(\varphi_1) \geq \theta_\varepsilon^*\left(\frac{\pi}{2\sqrt{2}}\right) = \theta^*\left(\frac{\pi}{2\sqrt{2}}\right) \overset{(4\text{-}6)}{\geq} \arctan\frac{1}{\cos(\pi/(2\sqrt{2}))} > \frac{\pi}{3}.$$

By (4-25), this implies that $\sin\theta_m \cos\theta_m \leq \sin\theta_1 \cos\theta_1 < \sqrt{3}/4$, that is, $\theta_m < \pi/6$, which is impossible in view of Lemma 4.11.

To prove (iii), assume for a contradiction that $\varphi(t_1) \in (\pi/2 - \varepsilon, \pi/2]$. We have

$$\frac{\pi}{2} - \varepsilon \leq \int_0^{\varphi(t_1)} d\varphi \overset{(4\text{-}25)}{=} \int_{\theta_0}^{\theta(\varphi(t_1))} \frac{K}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\,d\theta \leq \int_{\theta_0}^{\theta_M} \frac{K}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\,d\theta,$$

where in the last inequality we have used (ii). Splitting the right-hand side and applying (i), we then obtain

$$\begin{aligned}\frac{\pi}{2} - \varepsilon \;&\leq\; \int_{\pi/6}^{\pi/4} \frac{K}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\,d\theta + \frac{\sqrt{2}}{2}\int_{\pi/4}^{\theta_M} \frac{K}{\sin\theta\cos\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\,d\theta \\ &\overset{(4\text{-}30)}{=} \int_{\pi/6}^{\pi/4} \frac{K}{\sin\theta\sqrt{\cos^2\theta\sin^2\theta - K^2}}\,d\theta + \frac{\pi\sqrt{2}}{4}. \end{aligned} \tag{4-34}$$

Furthermore, again by (ii), we have

$$K = \sin\theta_M\cos\theta_M \leq \sin\theta(\varphi(t_1))\cos\theta(\varphi(t_1)) \leq \sin\theta^*(\varphi - \varepsilon)\cos\theta^*(\varphi - \varepsilon) \to 0 \quad \text{as } \varepsilon \to 0.$$

Therefore the integral on the right-hand side of (4-34) vanishes as $\varepsilon \to 0$, yielding a contradiction for ε sufficiently small. □

Lemma 4.14. *Let $\varphi \in I = (\varphi_0, \varphi_1) \subseteq [0, \pi/2 - \varepsilon]$. Then*

$$J_I(\theta_\varepsilon^*) = \left[\frac{\sin\theta_0\sin\varphi}{\sqrt{1 + \tan^2\theta_0\cos^2\varphi}}\right]_{\varphi=\varphi_0}^{\varphi=\varphi_1}. \tag{4-35}$$

Proof. Since $\theta_\varepsilon^* = \theta^*$ for $\varphi \leq \pi/2 - \varepsilon$, a straightforward computation shows that

$$\cos\theta_\varepsilon^*\sqrt{(\theta_\varepsilon^{*'})^2 + \sin^2\theta_\varepsilon^*} = \frac{4\sin\theta_0\cos\varphi}{\sqrt{1 + \tan^2\theta_0\cos^2\varphi}\,(3 + \cos 2\theta_0 + 2\cos 2\varphi\sin^2\theta_0)}.$$

An integration of this expression yields (4-35). □

We now show that if the graph of a solution to (4-24) emanates from $\partial\mathcal{P} \cap \partial\mathcal{P}_\varepsilon$ into $\mathring{\mathcal{P}}_\varepsilon$, then it does not return to $\partial\mathcal{P} \cap \partial\mathcal{P}_\varepsilon$.

Lemma 4.15. *Let $\varphi_1 \in [0, \pi/2 - \varepsilon)$ and let θ be a solution of (4-24) such that $\theta(\varphi_1) = \theta^*(\varphi_1)$ and $\theta'(\varphi_1) \leq \theta^{*'}(\varphi_1)$. Then $X(\varphi, \theta(\varphi)) \subset \mathring{\mathcal{P}}_\varepsilon$ for all $\varphi \in (\varphi_1, \pi/2 - \varepsilon]$.*

Proof. We let $\theta_1 = \theta(\varphi_1)$ and we distinguish two cases.

Case 1: $\theta'(\varphi_1) \le 0$. If $\theta_0 = \pi/4$, $\varphi_1 = 0$, and $\theta'(\varphi_1) = 0$, then $\theta \equiv \pi/4$ and the lemma is trivially true. Else Lemma 4.10 implies that θ decreases until either $\varphi = \pi$ or it reaches its minimum. In the former case the lemma is proved. In the latter, part (d) of Lemma 4.10 implies that $\theta < \theta^*(\varphi_1)$ at least until $\varphi = \varphi_1 + \pi/\sqrt{2} > \pi/2$.

Case 2: $\theta'(\varphi_1) > 0$. It is convenient to set

$$v(\varphi) = \log\tan(\tfrac{1}{2}\theta(\varphi)).$$

Lengthy but straightforward computations show that

$$v'' = \frac{\cosh(2v) - 3}{\sinh(2v)}(1 + (v')^2).$$

We now observe that

$$\begin{aligned}
\cosh(2v) < 3 &\iff \tfrac{1}{2}\log(3 - 2\sqrt{2}) < v < \tfrac{1}{2}\log(3 + 2\sqrt{2}) \\
&\iff \log\tan(\pi/8) < \log\tan(\theta/2) < \log\tan(3\pi/8) \\
&\impliedby \theta \in (\pi/4, \pi/2), \\
\sinh(2v) > 0 &\iff v > 0 \impliedby \theta \in (\pi/8, \pi/2).
\end{aligned}$$

Hence $v'' < 0$ if $\theta > \pi/4$. On the other hand, as long as $\varphi \le \pi/2 - \varepsilon$, we have

$$\theta < \theta^*_\varepsilon = \theta^* \iff v < v^*(\varphi) := \log\tan\left(\frac{1}{2}\arctan\frac{1}{\tan\theta_0\,|\cos\varphi|}\right)$$

with

$$v^{*''} = \frac{\sin\theta_0\cos\varphi}{(\sin^2\theta_0\cos^2\varphi + \cos^2\theta_0)^{3/2}} > 0.$$

Hence $(v - v^*)'' < 0$ as long as $\theta > \pi/4$ and $\varphi \le \pi/2 - \varepsilon$. Since $v = v^*$ and $v' \le v^{*'}$ at $\varphi = \varphi_1$, we have $v < v^*$ as long as either $\varphi = \pi/2 - \varepsilon$ or $\theta = \pi/4$. In the former case the proof is complete. In the latter case, part (d) of Lemma 4.10 implies that θ will then remain below $\pi/4$ at least in an interval of length $\pi/\sqrt{2} > \pi/2$, and the proof is complete. □

We now estimate J over a candidate symmetric shortest path which de-touches from $\partial\mathcal{P} \cap \partial\mathcal{P}_\varepsilon$ and reaches $\varphi = \pi/2 - \varepsilon$:

Lemma 4.16. *Let $\boldsymbol{\gamma}$ be a symmetric shortest path not passing through* $(0, 0, 1)$ *and let $\boldsymbol{\gamma} = X(\varphi, \theta)$. If $t_1 \ge 0$ exists such that $\varphi_1 = \varphi(t_1) \in [0, \pi/2 - \varepsilon)$, $\theta(t_1) = \theta^*(\varphi_1)$, and $\boldsymbol{\gamma} \not\subset \partial\mathcal{P}$ in a right-neighborhood of t_1, then*

$$J_{(\varphi_1, \pi/2-\varepsilon)}(\theta) > J_{(\varphi_1, \pi/2)}(\theta^*) - \frac{\varepsilon}{2}.$$

Proof. By assumption, for all $\sigma > 0$, there exists $t_\sigma \in (t_1, t_1 + \sigma)$ such that $\boldsymbol{\gamma}(t_\sigma) \in \mathring{\mathcal{P}}_\varepsilon$. By continuity, there exists $\tilde{t}_\sigma \in [t_1, t_\sigma)$ such that $\boldsymbol{\gamma}(\tilde{t}_\sigma) \in \partial\mathcal{P}_\varepsilon$ and $\boldsymbol{\gamma}(t) \in \mathring{\mathcal{P}}_\varepsilon$ for all $t \in (\tilde{t}_\sigma, t_\sigma]$. Then we may apply Lemma 4.9 in $(\tilde{t}_\sigma, t_\sigma]$.

If $K = 0$, then φ is constant, and since the curve is on $\partial\mathcal{P}_\varepsilon$ at $t = \tilde{t}_\sigma$, θ must decrease. Hence $\boldsymbol{\gamma}$ remains smooth down to $\theta = 0$, the north pole. Therefore this case is excluded.

Then $K > 0$, φ is strictly increasing, and $\theta(\varphi)$ solves (4-24) in $(\tilde{t}_\sigma, t_\sigma)$. By Lemma 4.15, we in fact have $X(\varphi, \theta(\varphi)) \subset \mathring{\mathcal{P}}_\varepsilon$ as long as $\varphi \le \pi/2 - \varepsilon$, which in particular implies that $\tilde{t}_\sigma = t_1$ and that θ solves (4-24) as long as $\varphi \le \pi/2 - \varepsilon$. We let

$$\theta_1 = \theta(\varphi_1) = \theta^*(\varphi_1)$$

and we distinguish two cases.

Case 1: $\theta'(\varphi_1) \le 0$. Lemma 4.10 and the symmetry of the path imply that θ does not increase until $\pi/2$ and $\theta(\pi/2) = \theta_m > 0$. If $\theta_1 = \theta_0 = \pi/4$ and $\varphi_1 = 0$, then $\boldsymbol{\gamma}((0,1)) \subset \mathring{\mathcal{P}}_\varepsilon$, a case which has already been ruled out in Lemma 4.12. Hence $\theta_1 > \pi/4$.

We claim that

$$\min_{\varphi\in[0,\pi/2]} \sin\theta\cos\theta = \sin\theta_m\cos\theta_m. \tag{4-36}$$

By (4-25),

$$\min_{\varphi\in[\varphi_1,\pi/2]} \sin\theta\cos\theta = \sin\theta_m\cos\theta_m. \tag{4-37}$$

In particular,

$$\sin(\theta_1)\cos(\theta_1) \ge \sin\theta_m\cos\theta_m. \tag{4-38}$$

On the other hand, by Lemma 4.13(ii), $\theta \in [\theta_0, \theta_1]$ for $\varphi \in [0, \varphi_1]$. Hence

$$\min_{\varphi\in[0,\varphi_1]} \sin\theta\cos\theta = \min\{\sin\theta_0\cos\theta_0, \sin\theta_1\cos\theta_1\}. \tag{4-39}$$

Since θ^* is increasing,

$$\sin(\theta_1)\cos(\theta_1) \le \sin(\theta^*(0))\cos(\theta^*(0)) = \sin\Big(\frac{\pi}{2} - \theta_0\Big)\cos\Big(\frac{\pi}{2} - \theta_0\Big) = \sin(\theta_0)\cos(\theta_0),$$

therefore (4-39) reads

$$\min_{\varphi\in[0,\varphi_1]} \sin\theta\cos\theta = \sin\theta_1\cos\theta_1 \overset{(4\text{-}38)}{\ge} \sin\theta_m\cos\theta_m \tag{4-40}$$

and (4-36) follows from (4-37) and (4-40).

We denote by $\tilde{\varphi}_1 \in (\varphi_1, \pi/2)$ the unique point such that $\theta(\tilde{\varphi}) = \theta_0$ (recall that $\theta_1 > \pi/4$, $\theta(\pi/2) = \theta_m < \pi/4$ and θ is decreasing in $(\varphi_1, \pi/2)$), and we define (see Figure 2)

$$\tilde{\theta}(\varphi) := \begin{cases} \theta(\varphi + \tilde{\varphi}_1) & \text{if } 0 \le \varphi \le \dfrac{\pi}{2} - \tilde{\varphi}_1, \\ \theta_m & \text{if } \dfrac{\pi}{2} - \tilde{\varphi}_1 \le \varphi \le \dfrac{\pi}{2}. \end{cases}$$

We have

$$J_{(0,\tilde{\varphi}_1)}(\theta) > \sin\theta_m\cos\theta_m\tilde{\varphi}_1 = J_{(\pi/2-\tilde{\varphi}_1,\pi/2)}(\tilde{\theta}) \quad \text{and} \quad J_{(0,\pi/2-\tilde{\varphi}_1)}(\tilde{\theta}) = J_{(\tilde{\varphi}_1,\pi/2)}(\theta).$$

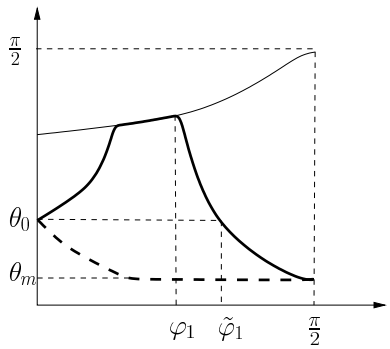


Figure 2. Case 1 in the proof of Lemma 4.16. The path $(\varphi, \theta(\varphi))$ (continuous) is beaten by its competitor $(\varphi, \tilde{\theta}(\varphi))$ (dashed).

Therefore $\boldsymbol{\gamma}$ is not a shortest path and this case is excluded.

Case 2: $\theta'(\varphi_1) > 0$. Lemma 4.10 and the symmetry of the path imply that θ increases until $\pi/2 - \varepsilon$. We now estimate its length in $I_\varepsilon = (\varphi_1, \pi/2 - \varepsilon)$. By the assumption of Case 2, and since $\boldsymbol{\gamma} \in \mathcal{P}$ in I_ε,

$$0 < \theta'(\varphi_1) \leq (\theta^*)'(\varphi_1). \tag{4-41}$$

By (4-25),

$$\sin^2\theta\,(\cos^2\theta \sin^2\theta - K^2) = K^2(\theta')^2 \quad \text{for some } K \in (0, 1/2). \tag{4-42}$$

Evaluating this expression at φ_1, we have

$$K = \frac{\sin^2\theta^*(\varphi_1)\cos\theta^*(\varphi_1)}{\sqrt{(\theta'(\varphi_1))^2 + \sin^2\theta^*(\varphi_1)}} \overset{(4\text{-}41)}{\geq} \frac{\sin^2\theta^*(\varphi_1)\cos\theta^*(\varphi_1)}{\sqrt{(\theta^{*'}(\varphi_1))^2 + \sin^2\theta^*(\varphi_1)}}.$$

Of course, we have

$$J_{I_\varepsilon}(\theta) \geq \int_{I_\varepsilon} \cos\theta \sin\theta\, d\varphi \overset{(4\text{-}42)}{\geq} |I_\varepsilon| K.$$

This chain of inequalities implies that

$$J_{I_\varepsilon}(\theta) > |I_\varepsilon| \frac{\sin^2\theta^*(\varphi_1)\cos\theta^*(\varphi_1)}{\sqrt{(\theta^{*'}(\varphi_1))^2 + \sin^2\theta^*(\varphi_1)}} = \frac{|I_\varepsilon| \sin\theta_0 \cos\varphi_1}{\sqrt{1 + \tan^2\theta_0 \cos^2\varphi_1}}$$

(the latter equality follows from an explicit computation). On the other hand, by Lemma 4.14, the curve which just stays on the obstacle, $\boldsymbol{\gamma}^* = X(\varphi, \theta^*(\varphi))$, $\varphi \in (\varphi_1, \pi/2)$, is such that

$$J_{(\varphi_1, \pi/2)}(\theta^*) = \sin\theta_0 \left(1 - \frac{\sin\varphi_1}{\sqrt{1 + \tan^2\theta_0 \cos^2\varphi_1}}\right).$$

Hence

$$(J_{I_\varepsilon}(\theta) - J_{(\varphi_1,\pi/2)}(\theta^*))\frac{\sqrt{1+\tan^2\theta_0\cos^2\varphi_1}}{\sin\theta_0}$$
$$> |I_\varepsilon|\cos\varphi_1 + \sin\varphi_1 - \sqrt{1+\tan^2\theta_0\cos^2\varphi_1}$$
$$> \left(\frac{\pi}{2} - \varepsilon - \varphi_1\right)\cos\varphi_1 + \sin\varphi_1 - \sqrt{1+\cos^2\varphi_1} =: F(\varphi_1) - \varepsilon\cos\varphi_1.$$

Another calculus exercise shows that F is decreasing $[0, \pi/2]$: since $F(\pi/2) = 0$, F is positive. Therefore

$$J_{I_\varepsilon}(\theta) > J_{(\varphi_1,\pi/2)}(\theta^*) - \varepsilon\frac{\sin\theta_0\cos\varphi_1}{\sqrt{1+\tan^2\theta_0\cos^2\varphi_1}} > J_{(\varphi_1,\pi/2)}(\theta^*) - \tfrac{1}{2}\varepsilon. \qquad \square$$

Next we characterize the candidate shortest paths joining $X(0, \theta^*(0))$ with another point on $\partial\mathcal{P} \cap \partial\mathcal{P}_\varepsilon$ which is on the same side with respect to $\pi/2$.

Lemma 4.17. *Let $0 < \bar\varphi \le \pi/2 - \varepsilon$. The shortest path which connects $X(0, \pi/2 - \theta_0)$ and $X(\bar\varphi, \theta^*(\bar\varphi))$ is (a smooth reparametrization of) $\boldsymbol{\gamma}^* = X(\varphi, \theta^*(\varphi))$, $\varphi \in [0, \bar\varphi]$.*

Proof. Let $I = (0, \bar\varphi)$. We recall by Lemma 4.14 that

$$J_I(\theta^*) = \frac{\sin\theta_0\sin\bar\varphi}{\sqrt{1+\tan^2\theta_0\cos^2\bar\varphi}} < \sin\theta_0.$$

First note that $\boldsymbol{\gamma}$ does not reach the north pole. For if it did, at a time $\bar t \in (0, 1)$, we would have

$$J(\boldsymbol{\gamma}) \ge \int_0^{\bar t} \cos\theta|\dot\theta|\,dt \ge \sin(\theta(0)) = \sin\left(\frac{\pi}{2} - \theta_0\right) = \cos\theta_0 > \sin\theta_0 > J_I(\theta^*),$$

which is impossible.

Therefore we may use the spherical coordinates (4-4), and arguing as in the proof of Lemma 4.8 we see that φ in nondecreasing.

Assume for a contradiction that $\boldsymbol{\gamma}$ does not coincide (up to a smooth reparametrization) with $\boldsymbol{\gamma}^*$. Then $t_1 > 0$ and a right-neighborhood $\tilde I$ of t_1 exist such that $\varphi_1 := \varphi(t_1) < \bar\varphi$ and $\boldsymbol{\gamma}(\tilde I) \not\subset \partial\mathcal{P}$. Arguing as in the first lines of the proof of Lemma 4.16, one finds that there is $t_2 > t_1$ such that $\boldsymbol{\gamma}(t) \in \mathring{\mathcal{P}}$ for all $t \in (t_1, t_2)$. Then, arguing as in the proof of Lemma 4.9, one finds that (4-22) holds, and $K \ge 0$ since φ is nondecreasing. If $K > 0$, then $\theta(\varphi)$ would solve (4-24) in (t_1, t_2); but in view of Lemma 4.15, such a solution will not rehit the constraint until $\varphi = \pi/2 - \varepsilon$, hence $K > 0$ can not occur. If $K = 0$, then $\varphi \equiv \varphi_1$, and since we are on $\partial\mathcal{P}$ at time t_1, θ must move inwards. Hence $\boldsymbol{\gamma}$ remains smooth up to $\theta = 0$, the north pole, a contradiction. $\square$

Proof of Theorem 4.1. First of all, we note that Lemma 4.14 implies that $J(\theta^*) = 2\sin\theta_0$. Hence, in view of Lemma 4.5, it suffices to show that

$$\inf_{\boldsymbol{\gamma}\in\Gamma_\varepsilon} J(\boldsymbol{\gamma}) \ge J(\theta^*) - \omega(\varepsilon) = 2\sin\theta_0 - \omega(\varepsilon), \tag{4-43}$$

where ω is a universal function which vanishes as $\varepsilon \to 0$. By Lemma 4.6, the inf on the left-hand side of (4-43) is attained. Let $\boldsymbol{\gamma}$ be one such shortest path. If $\boldsymbol{\gamma}$ passes through $(0, 0, 1)$, then (4-43) follows from

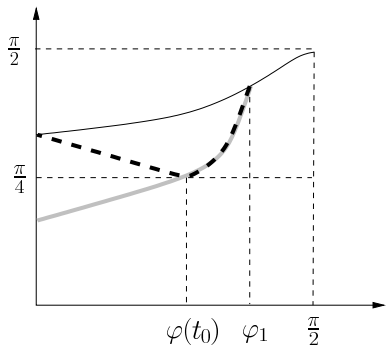


Figure 3. Case 1 in the proof of Theorem 4.1. The path $(\varphi(t), \theta(t))$ (gray) is beaten by its competitor $(\varphi(t), \tilde{\theta}(t))$ (dashed).

Lemma 4.7. If not, we let $\boldsymbol{\gamma} = X(\varphi, \theta)$ and, by Lemma 4.8, we assume without loss of generality that $\boldsymbol{\gamma}$ is symmetric. For simplicity, we distinguish between $\theta_0 < \pi/4$ and $\theta_0 = \pi/4$.

Case 1: $\theta_0 < \pi/4$. We already know from Lemma 4.12 that $\boldsymbol{\gamma}$ has to intersect $\partial\mathcal{P}_\varepsilon$. Let t_0 and t_1 be, respectively, the first time in which $\theta(t) = \pi/4$ and the first time in which $\boldsymbol{\gamma}$ intersects $\partial\mathcal{P}_\varepsilon$:

$$t_1 := \sup\{t > 0 : \boldsymbol{\gamma} \in \mathring{\mathcal{P}}_\varepsilon \text{ in } [0, t)\} \quad \text{and} \quad \varphi_1 = \varphi(t_1).$$

Provided ε is sufficiently small, by Lemma 4.13(ii), θ is increasing in $(0, t_1)$. Hence the curve

$$\tilde{\boldsymbol{\gamma}}(t) := X(\varphi(t), \tilde{\theta}(t)), \quad \tilde{\theta}(t) := \begin{cases} \dfrac{\pi}{2} - \theta(t) & t \in [0, t_0], \\ \theta(t) & t \in [t_0, t_1] \end{cases}$$

is contained in $\mathcal{P}_\varepsilon$ (see Figure 3). We claim that

$$J(\tilde{\boldsymbol{\gamma}}\chi_{(0,t_1)}) < J(\boldsymbol{\gamma}\chi_{(0,t_1)}), \tag{4-44}$$

which is equivalent to

$$J(\tilde{\boldsymbol{\gamma}}\chi_{(0,t_0)}) < J(\boldsymbol{\gamma}\chi_{(0,t_0)}). \tag{4-45}$$

By Lemma 4.9, $\boldsymbol{\gamma}$ satisfies (4-22) in $(0, t_0)$. If $K = 0$, then $\varphi \equiv 0$ and (4-45) follows from the expression (4-5) of J:

$$\cos\left(\frac{\pi}{2} - \theta\right) = \sin(\theta) < \cos\theta \quad \text{if } \theta \le \pi/4.$$

Otherwise, by Lemma 4.9 $\varphi \mapsto \theta(\varphi)$ solves (4-24) in $(0, \varphi_0)$, where $\varphi_0 = \varphi(t_0)$. Then it follows by Lemma 4.13(iii) that $\varphi_0 < \pi/2 - \varepsilon$, and we may use the equivalent expression (4-26) for J: since

$$\cos^2\tilde{\theta}((\tilde{\theta}')^2 + \sin^2\tilde{\theta}) = \sin^2\theta((\theta')^2 + \cos^2\theta) < \cos^2\theta((\theta')^2 + \sin^2\theta) \quad \text{in } (0, \varphi_0),$$

(4-45) follows.

Since $\tilde{\boldsymbol{\gamma}}$ is a path connecting $X(0, \pi/2 - \theta_0)$ to $X(\varphi_1, \theta^*(\varphi_1))$, Lemma 4.17 implies that $J(\tilde{\boldsymbol{\gamma}}\chi_{(0,t_1)}) \ge J_{(0,\varphi_1)}(\theta^*)$. Together with (4-44), we obtain

$$J_{(0,\varphi_1)}(\theta^*) < J(\boldsymbol{\gamma}\chi_{(0,t_1)}). \tag{4-46}$$

Now let $t_2 \geq t_1$ be defined by

$$t_2 := \max\{t \geq t_1 : \boldsymbol{\gamma} \in \partial \mathcal{P}_\varepsilon \text{ in } [t_1, t]\} \quad \text{and} \quad \varphi_2 = \varphi(t_2).$$

The estimate in (t_1, t_2) is trivial since $\boldsymbol{\gamma}$ coincides with $\boldsymbol{\gamma}^* := X(\varphi, \theta^*)$:

$$J(\boldsymbol{\gamma}\chi_{(t_1,t_2)}) = J_{(\varphi_1,\varphi_2)}(\theta^*).$$

On $(\varphi_2, \pi/2 - \varepsilon)$, Lemma 4.16 implies that

$$J_{(\varphi_2,\pi/2-\varepsilon)}(\theta) > J_{(\varphi_2,\pi/2)}(\theta^*) - \frac{\varepsilon}{2} \quad \text{if } \varphi_2 < \frac{\pi}{2} - \varepsilon. \tag{4-47}$$

Finally, we just observe that

$$J_{(\pi/2-\varepsilon,\pi/2)}(\theta^*) \overset{(4\text{-}7)}{\leq} \omega(\varepsilon). \tag{4-48}$$

Collecting (4-46)–(4-48) and recalling the symmetry of $\boldsymbol{\gamma}$, we obtain (4-43).

Case 2: $\theta_0 = \pi/4$. This case is simpler. We let

$$t_2 = \max\{t \geq 0 : \boldsymbol{\gamma} \in \partial \mathcal{P}_\varepsilon \text{ in } [0, t]\} \geq 0 \quad \text{and} \quad \varphi_2 = \varphi(t_2),$$

and we argue exactly as above to obtain $J(\boldsymbol{\gamma}\chi_{(0,t_2)}) = J_{(0,\varphi_2)}(\theta^*)$ and (4-47)–(4-48). □

Acknowledgement

We thank Manuel Ritoré for fruitful discussions on the geometrical problem in Section 4.

References

[Alexander et al. 1993] S. B. Alexander, I. D. Berg, and R. L. Bishop, "Cut loci, minimizers, and wavefronts in Riemannian manifolds with boundary", *Michigan Math. J.* **40**:2 (1993), 229–237. MR 94c:53053 Zbl 0817.53023

[Alicandro et al. 2007] R. Alicandro, A. Corbo Esposito, and C. Leone, "Relaxation in BV of integral functionals defined on Sobolev functions with values in the unit sphere", *J. Convex Anal.* **14**:1 (2007), 69–98. MR 2008b:49007 Zbl 1138.49017

[Ambrosio and Dal Maso 1990] L. Ambrosio and G. Dal Maso, "A general chain rule for distributional derivatives", *Proc. Amer. Math. Soc.* **108**:3 (1990), 691–702. MR 90j:26019 Zbl 0685.49027

[Ambrosio et al. 2000] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of bounded variation and free discontinuity problems*, Clarendon Press, New York, 2000. MR 2003a:49002 Zbl 0957.49001

[Ambrosio et al. 2005] L. Ambrosio, G. Crippa, and S. Maniglia, "Traces and fine properties of a BD class of vector fields and applications", *Ann. Fac. Sci. Toulouse Math.* (6) **14**:4 (2005), 527–561. MR 2007b:35040 Zbl 1091.35007

[Andreu et al. 2001] F. Andreu, C. Ballester, V. Caselles, and J. M. Mazón, "Minimizing total variation flow", *Differential Integral Equations* **14**:3 (2001), 321–360. MR 2002e:35109 Zbl 1020.35037

[Andreu-Vaillo et al. 2004] F. Andreu-Vaillo, V. Caselles, and J. M. Mazón, *Parabolic quasilinear equations minimizing linear growth functionals*, Progress in Mathematics **223**, Birkhäuser, Basel, 2004. MR 2005c:35002 Zbl 1053.35002

[Anzellotti 1983] G. Anzellotti, "Pairings between measures and bounded functions and compensated compactness", *Ann. Mat. Pura Appl.* (4) **135** (1983), 293–318. MR 85m:46042 Zbl 0572.46023

[Barrett et al. 2008] J. W. Barrett, X. Feng, and A. Prohl, "On p-harmonic map heat flows for $1 \leq p < \infty$ and their finite element approximations", *SIAM J. Math. Anal.* **40**:4 (2008), 1471–1498. MR 2009m:35202 Zbl 1182.35154

[Bertsch et al. 2002] M. Bertsch, R. Dal Passo, and R. van der Hout, "Nonuniqueness for the heat flow of harmonic maps on the disk", *Arch. Ration. Mech. Anal.* **161**:2 (2002), 93–112. MR 2003a:35093 Zbl 1006.35050

[Bertsch et al. 2003] M. Bertsch, R. Dal Passo, and A. Pisante, "Point singularities and nonuniqueness for the heat flow for harmonic maps", *Comm. Partial Differential Equations* **28**:5-6 (2003), 1135–1160. MR 2004c:53099 Zbl 1029.58008

[Bonforte and Figalli 2012] M. Bonforte and A. Figalli, "Total variation flow and sign fast diffusion in one dimension", *J. Differential Equations* **252**:8 (2012), 4455–4480. MR 2881044 Zbl 1242.35049

[Brezis 2011] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Springer, New York, 2011. MR 2012a:35002 Zbl 1220.46002

[Caselles 2011] V. Caselles, "On the entropy conditions for some flux limited diffusion equations", *J. Differential Equations* **250**:8 (2011), 3311–3348. MR 2012d:35179 Zbl 1231.35101

[Chen 1989] Y. M. Chen, "The weak solutions to the evolution problems of harmonic maps", *Math. Z.* **201**:1 (1989), 69–74. MR 90i:58030 Zbl 0685.58015

[Chen and Frid 1999] G.-Q. Chen and H. Frid, "Divergence-measure fields and hyperbolic conservation laws", *Arch. Ration. Mech. Anal.* **147**:2 (1999), 89–118. MR 2000d:35136 Zbl 0942.35111

[Chen et al. 1994] Y. M. Chen, M. C. Hong, and N. Hungerbühler, "Heat flow of p-harmonic maps with values into spheres", *Math. Z.* **215**:1 (1994), 25–35. MR 94k:58145 Zbl 0793.53049

[Dal Passo et al. 2008] R. Dal Passo, L. Giacomelli, and S. Moll, "Rotationally symmetric 1-harmonic maps from D^2 to S^2", *Calc. Var. Partial Differential Equations* **32**:4 (2008), 533–554. MR 2009f:58024 Zbl 1147.58015

[Darling 1994] R. W. R. Darling, *Differential forms and connections*, Cambridge University Press, 1994. MR 95j:53038 Zbl 0822.53001

[DeSimone and Podio-Guidugli 1996] A. DeSimone and P. Podio-Guidugli, "On the continuum theory of deformable ferromagnetic solids", *Arch. Rational Mech. Anal.* **136**:3 (1996), 201–233. MR 98a:73051 Zbl 1002.74521

[Diestel and Uhl 1977] J. Diestel and J. J. Uhl, Jr., *Vector measures*, Mathematical Surveys **15**, Amer. Math. Soc., Providence, RI, 1977. MR 56 #12216 Zbl 0369.46039

[Eells and Sampson 1964] J. Eells, Jr. and J. H. Sampson, "Harmonic mappings of Riemannian manifolds", *Amer. J. Math.* **86** (1964), 109–160. MR 29 #1603 Zbl 0122.40102

[Evans and Gariepy 1992] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, CRC Press, Boca Raton, FL, 1992. MR 93f:28001 Zbl 0804.28001

[Federer 1969] H. Federer, *Geometric measure theory*, Die Grundlehren der mathematischen Wissenschaften **153**, Springer, New York, 1969. MR 41 #1976 Zbl 0176.00801

[Feng 2010] X. Feng, "Divergence-L^q and divergence-measure tensor fields and gradient flows for linear growth functionals of maps into the unit sphere", *Calc. Var. Partial Differential Equations* **37**:1-2 (2010), 111–139. MR 2011a:35281 Zbl 1184.35171

[Fonseca and Müller 1993] I. Fonseca and S. Müller, "Relaxation of quasiconvex functionals in $\mathrm{BV}(\Omega, \mathbb{R}^p)$ for integrands $f(x, u, \nabla u)$", *Arch. Rational Mech. Anal.* **123**:1 (1993), 1–49. MR 94h:49023 Zbl 0788.49039

[Fonseca and Rybka 1992] I. Fonseca and P. Rybka, "Relaxation of multiple integrals in the space $\mathrm{BV}(\Omega, \mathbb{R}^p)$", *Proc. Roy. Soc. Edinburgh Sect. A* **121**:3-4 (1992), 321–348. MR 94g:49032 Zbl 0794.49012

[Giacomelli and Moll 2010] L. Giacomelli and S. Moll, "Rotationally symmetric 1-harmonic flows from D^2 to S^2: local well-posedness and finite time blowup", *SIAM J. Math. Anal.* **42**:6 (2010), 2791–2817. MR 2011k:35084 Zbl 1230.35046

[Giacomelli et al. 2013a] L. Giacomelli, J. M. Mazón, and S. Moll, "The 1-harmonic flow with values into $\mathbb{S}^1$", *SIAM J. Math. Anal.* **45**:3 (2013), 1723–1740. MR 3063148 Zbl 06200948

[Giacomelli et al. 2013b] L. Giacomelli, J. M. Mazón, and S. Moll, "Solutions to the 1-harmonic flow with values into a hyper-octant of the N-sphere", *Appl. Math. Lett.* **26**:11 (2013), 1061–1064. MR 3089565

[Giaquinta and Mucci 2006] M. Giaquinta and D. Mucci, "The BV-energy of maps into a manifold: Relaxation and density results", *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (5) **5**:4 (2006), 483–548. MR 2008k:49090 Zbl 1150.49020

[Giga and Kobayashi 2003] Y. Giga and R. Kobayashi, "On constrained equations with singular diffusivity", *Methods Appl. Anal.* **10**:2 (2003), 253–277. MR 2005e:58030 Zbl 1058.58006

[Giga and Kuroda 2004] Y. Giga and H. Kuroda, "On breakdown of solutions of a constrained gradient system of total variation", *Bol. Soc. Parana. Mat.* (3) **22**:1 (2004), 9–20. MR 2005f:35149 Zbl 1064.35084

[Giga et al. 2004] Y. Giga, Y. Kashima, and N. Yamazaki, "Local solvability of a constrained gradient system of total variation", *Abstr. Appl. Anal.* 8 (2004), 651–682. MR 2005k:35195 Zbl 1068.35054

[Giga et al. 2005] Y. Giga, H. Kuroda, and N. Yamazaki, "An existence result for a discretized constrained gradient system of total variation flow in color image processing", *Interdiscip. Inform. Sci.* **11**:2 (2005), 199–204. MR 2006h:65161

[Giga et al. 2007] Y. Giga, H. Kuroda, and N. Yamazaki, "Global solvability of constrained singular diffusion equation associated with essential variation", pp. 209–218 in *Free boundary problems*, edited by I. N. Figueiredo et al., Internat. Ser. Numer. Math. **154**, Birkhäuser, Basel, 2007. MR 2007m:35141 Zbl 1119.35028

[van der Hout 2001] R. van der Hout, "Flow alignment in nematic liquid crystals in flows with cylindrical symmetry", *Differential Integral Equations* **14**:2 (2001), 189–211. MR 2001m:76007 Zbl 1021.35028

[Hungerbühler 2004] N. Hungerbühler, "Heat flow into spheres for a class of energies", pp. 45–65 in *Variational problems in Riemannian geometry*, edited by P. Baird et al., Progr. Nonlinear Differential Equations Appl. **59**, Birkhäuser, Basel, 2004. MR 2005d:53106 Zbl 1063.58010

[Kobayashi et al. 2000] R. Kobayashi, J. A. Warren, and W. C. Carter, "A continuum model of grain boundaries", *Phys. D* **140**:1-2 (2000), 141–150. MR 2000m:74071 Zbl 0956.35123

[Misawa 2002] M. Misawa, "On the p-harmonic flow into spheres in the singular case", *Nonlinear Anal.* **50**:4, Ser. A: Theory Methods (2002), 485–494. MR 2003f:53119 Zbl 1163.53339

[Sapiro 2001] G. Sapiro, *Geometric partial differential equations and image analysis*, Cambridge University Press, 2001. MR 2002a:68142 Zbl 0968.35001

[Simon 1987] J. Simon, "Compact sets in the space $L^p(0, T; B)$", *Ann. Mat. Pura Appl.* (4) **146** (1987), 65–96. MR 89c:46055 Zbl 0629.46031

[Struwe 1992] M. Struwe, "The evolution of harmonic maps: Existence, partial regularity, and singularities", pp. 485–491 in *Nonlinear diffusion equations and their equilibrium states, 3* (Gregynog, 1989), edited by N. G. Lloyd et al., Progr. Nonlinear Differential Equations Appl. **7**, Birkhäuser, Boston, MA, 1992. MR 1167858 Zbl 0769.58016

[Tang et al. 2000] B. Tang, G. Sapiro, and V. Caselles, "Diffusion of general data on non-flat manifolds via harmonic maps theory: The direction diffusion case", *Int. J. Comput. Vis.* **36**:2 (2000), 149–161.

[Tang et al. 2001] B. Tang, G. Sapiro, and V. Caselles, "Color image enhancement via chromaticity diffusion", *IEEE Trans. Image Process.* **10**:5 (2001), 701–707. Zbl 1037.68792

[Ziemer 1989] W. P. Ziemer, *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*, Graduate Texts in Mathematics **120**, Springer, New York, 1989. MR 91e:46046 Zbl 0692.46022

Received 18 Apr 2013. Accepted 27 Nov 2013.

LORENZO GIACOMELLI: lorenzo.giacomelli@sbai.uniroma1.it
SBAI Department, Sapienza University of Rome, Via Scarpa, 16, I-00161 Roma, Italy

JOSE M. MAZÓN: mazon@uv.es
Departament d'Anàlisi Matemàtica, Universitat de València, Dr. Moliner, 50, 46100 Burjassot, Spain

SALVADOR MOLL: j.salvador.moll@uv.es
Departament d'Anàlisi Matemàtica, Universitat de València, Dr. Moliner, 50, 46100 Burjassot, Spain

ANALYSIS AND PDE
Vol. 7, No. 3, 2014
dx.doi.org/10.2140/apde.2014.7.673

msp

DECOMPOSITION RANK OF $\mathcal{Z}$-STABLE C*-ALGEBRAS

AARON TIKUISIS AND WILHELM WINTER

We show that C*-algebras of the form $C(X) \otimes \mathcal{Z}$, where X is compact and Hausdorff and $\mathcal{Z}$ denotes the Jiang–Su algebra, have decomposition rank at most 2. This amounts to a dimension reduction result for C*-bundles with sufficiently regular fibres. It establishes an important case of a conjecture on the fine structure of nuclear C*-algebras of Toms and Winter, even in a nonsimple setting, and gives evidence that the topological dimension of noncommutative spaces is governed by fibres rather than base spaces.

1. Introduction

The structure and classification theory of nuclear C*-algebras has seen rapid progress in recent years, largely spurred by the subtle interplay between certain topological and algebraic regularity properties, such as finite topological dimension, tensorial absorption of suitable strongly self-absorbing C*-algebras, and order completeness of homological invariants; see [Elliott and Toms 2008] for an overview. In the simple and unital case, these relations were formalized by A. Toms and W. Winter as follows.

Conjecture 1.1. For a separable, simple, unital, nonelementary, stably finite and nuclear C*-algebra A, the following are equivalent:

(i) A has finite decomposition rank: in symbols, $\operatorname{dr} A < \infty$.

(ii) A is $\mathcal{Z}$-stable: $A \cong A \otimes \mathcal{Z}$.

(iii) A has strict comparison of positive elements.

Here, decomposition rank is a notion of noncommutative topological dimension introduced in [Kirchberg and Winter 2004], $\mathcal{Z}$ denotes the Jiang–Su algebra introduced in [Jiang and Su 1999], and strict comparison essentially means that positive elements may be compared in terms of tracial values of their support projections; compare [Rørdam 2006]. If one drops the finiteness assumption on A, one should replace (i) by

(i′) A has finite nuclear dimension, $\dim_{\mathrm{nuc}} A < \infty$,

where nuclear dimension [Winter and Zacharias 2010] is a variation of the decomposition rank that can have finite values also for infinite C*-algebras.

The conjecture still makes sense in the nonsimple situation, provided one asks A to have no elementary

Both authors were supported by DFG (SFB 878). Winter was also supported by EPSRC (grant numbers EP/G014019/1 and EP/I019227/1).
MSC2010: 46L35, 46L85.
Keywords: nuclear C*-algebras, decomposition rank, nuclear dimension, Jiang–Su algebra, classification, $C(X)$-algebras.

subquotients (this is a minimal requirement for $\mathcal{Z}$-stability); one also has to be slightly more careful about the definition of comparison in this case.

Nuclearity in this context manifests itself most prominently via approximation properties with particularly nice completely positive maps [Christensen et al. 2012; Hirshberg et al. 2012a].

Conjecture 1.1 has a number of important consequences for the structure of nuclear C*-algebras and it has turned out to be pivotal for many recent classification results, especially in view of the examples given in [Villadsen 1999; Rørdam 2003; Toms 2008]. Moreover, it highlights the striking analogy between the classification program for nuclear C*-algebras (see [Elliott 1995]) and Connes' [1976] celebrated classification of injective II_1 factors.

Implications (i), (i′) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii) of Conjecture 1.1 are by now known to hold in full generality [Rørdam 2004; Winter 2010; 2012]; (iii) $\Longrightarrow$ (ii) has been established under certain additional structural hypotheses [Matui and Sato 2012; Winter 2012], all of which, in particular, guarantee sufficient divisibility properties.

Arguably, it is (ii) $\Longrightarrow$ (i) which remains the least well understood of these implications. While there are promising partial results available [Winter and Zacharias 2010; Lin 2011b; Winter 2012], all of these factorize through classification theorems of some sort. This in turn makes it hard to explicitly identify the origin of finite dimensionality.[1]

In the simple purely infinite (hence $\mathcal{O}_\infty$-stable, hence $\mathcal{Z}$-stable [Kirchberg and Rørdam 2002; Kirchberg 2006]) case, one has to use Kirchberg–Phillips classification [Kirchberg 1995; Kirchberg and Phillips 2000] as well as a range result providing models to exhaust the invariant [Rørdam 2002] and then again Kirchberg–Phillips classification to show that these models have finite nuclear dimension [Winter and Zacharias 2010].[2]

In the simple stably finite case, at this point only approximately homogeneous (AH) algebras or approximately subhomogeneous (ASH) algebras for which projections separate traces are covered [Lin 2011a; Lin and Niu 2008; Winter 2004; 2007]. (This approach also includes crossed products associated to uniquely ergodic minimal dynamical systems [Toms and Winter 2009; 2013].) While both of these classes after stabilizing with $\mathcal{Z}$ can by now be shown directly to consist of TAI and TAF algebras [Lin 2011b], again finite topological dimension will only follow from classification results [Elliott et al. 2007; Winter 2010; Lin 2011a; Toms 2011] and after comparing to models which exhaust the invariant [Elliott 1996; Villadsen 1998]; see also [Rørdam 2002] for an overview. (Note that certain crossed products are shown directly to have finite nuclear dimension, or even finite decomposition rank [Hirshberg et al. 2012b; Szabo 2013]; however, $\mathcal{Z}$-stability is not assumed in these cases.)

[1] After this article appeared on the arXiv, Matui and Sato [2013] posted a very nice paper in which they prove finite decomposition rank for separable, simple, unital, nuclear, and $\mathcal{Z}$-stable C*-algebras provided these are quasidiagonal and have a unique tracial state. While this result is restricted to the simple and monotracial case (conditions we do not need at all), it only uses quasidiagonality as additional structural hypothesis (and this is of course much more general than our local homogeneity); see also [Sato et al. 2014]. Matui and Sato's approach heavily relies on deep results of Connes and of Haagerup and, in a sense, is almost perpendicular to ours; we believe that the two methods nicely complement each other.

[2] [Matui and Sato 2013] also contains a proof of finite nuclear dimension for simple purely infinite C*-algebras, which does not rely on classification; see also [Barlak et al. 2014].

Once again, the classification procedure does not make it entirely transparent where the finite topological dimension comes from, but at least Elliott–Gong–Li classification of simple AH algebras (of very slow dimension growth — later shown to be equivalent to slow dimension growth and to $\mathcal{Z}$-stability [Winter 2012]) heavily relies on Gong's deep dimension reduction theorem [2002]. Gong gives an essentially explicit way of replacing a given AH limit decomposition with one of low topological dimension. However, this method is technically very involved and requires both simplicity and the given inductive limit decomposition. It does not fully explain to what extent the two are necessary; in particular, it is in principle conceivable that a decomposition similar to that of Gong exists for algebras of the form $C(X) \otimes \mathcal{Q}$, where $\mathcal{Q}$ is the universal UHF algebra.[3]

In this article we show how finite topological dimension indeed arises for algebras of this type; in fact, we are able to cover algebras of the form $C(X) \otimes \mathcal{Z}$, and hence also locally homogeneous $\mathcal{Z}$-stable C*-algebras (not necessarily simple, or with a prescribed inductive limit structure). We hope our argument will shed new light on the conceptual reasons why finite topological dimension should arise in the presence of sufficient C*-algebraic regularity. Our method is based on approximately embedding the cone over the Cuntz algebra $\mathcal{O}_2$ into tracially small subalgebras of the algebra in question; these play a similar role as the small corners used in the definition of TAF algebras [Lin 2004] or the small hereditary subalgebras in property SI [Matui and Sato 2012]. We mention that we only obtain (a strong version of) finite decomposition rank, whereas Gong's reduction theorem yields an inductive limit decomposition; however, for many purposes, finite decomposition rank is sufficient; see [Winter 2010; Toms and Winter 2013].

In [Kirchberg and Rørdam 2005], algebras of the form $C(X) \otimes \mathcal{O}_2$ were shown to be approximated by algebras of the form $C(\Gamma) \otimes \mathcal{O}_2$ with Γ one-dimensional. Since $\mathcal{O}_2$ is by now known to have finite nuclear dimension [Winter and Zacharias 2010], this may be regarded as strong evidence that the topological dimension of a C*-bundle depends on the noncommutative size of the fibres more than the size of the base space. (A somewhat similar phenomenon was already observed for stable rank by Rieffel [1983].)

It is remarkable that [Kirchberg and Rørdam 2005] does not rely on a classification result in any way. It does, however, mix commutativity (of the structure algebra) and pure infiniteness (of the fibres).

It is not clear from [Kirchberg and Rørdam 2005] whether such a dimension type reduction also occurs in the setting of stably finite fibres. In the present article we show that it does, by developing a method to transport [Kirchberg and Rørdam 2005] to the situation where the fibres are UHF algebras (to pass to the case where each fibre is $\mathcal{Z}$ then requires a certain amount of additional machinery — at least if one wants to increase the dimension by no more than one). The crucial concept to link purely infinite and stably finite fibres is quasidiagonality of the cone over $\mathcal{O}_2$, discovered by Voiculescu [1993] and Kirchberg [1991]. In many ways it is most interesting just to know that the $\mathcal{Z}$-stable C*-algebras in our main result have finite decomposition rank, and the very small bound that we are able to derive is secondary. Certain technicalities can be circumvented, using [Carrión 2011, Lemma 3.1] in order to prove just finite decomposition rank, as we describe in Remarks 4.8 and 4.9. We are indebted to one of the referees for suggesting this shortcut.

[3] Added in proof: It turns out that this is not, in fact, conceivable; see [Tikuisis 2014].

One should mention that the fact that the fibres are specific strongly self-absorbing algebras in both [Kirchberg and Rørdam 2005] and in our result plays an important but in some sense secondary role: In [Kirchberg and Rørdam 2005] (combined with [Winter and Zacharias 2010]) one can replace $\mathcal{O}_2$ with $\mathcal{O}_\infty$, or in fact with any UCT Kirchberg algebra, and still arrive at finite nuclear dimension. More generally, our result yields the respective statement if the fibres have finite nuclear dimension and are $\mathcal{Z}$-stable, for example, in the simple, nuclear, classifiable case.

While at the current stage we only cover the case of highly homogeneous bundles, it will be an important task to handle bundles with non-Hausdorff spectrum, for example, $B \otimes \mathcal{Z}$ with B subhomogeneous, in order to also cover transformation group C*-algebras. This will be pursued in subsequent work by combining our technical Lemma 4.7, with the methods of [Winter 2004]; in preparation, we have stated Lemma 4.7 in a form slightly more general than necessary for the current main result Theorem 4.1. One of the referees has raised the question of whether (local) triviality of $C(X) \otimes \mathcal{Z}$ is needed to show that it has finite decomposition rank, particularly in light of the interesting examples of $C(X)$-algebras appearing in [Dadarlat 2009b; Hirshberg et al. 2007]; in response, we have added Section 5, in which we show that our result easily extends to nontrivial bundles such as these examples.

We would like to take this opportunity to thank both referees for their careful proofreading and inspiring comments.

We remind the reader that the Jiang–Su algebra $\mathcal{Z}$ is an inductive limit of so-called dimension-drop C*-algebras

$$\mathcal{Z}_{p_0,p_1} := \{f \in C([0,1], M_{p_0} \otimes M_{p_1}) : f(0) \in M_{p_0} \otimes \mathbb{C} \cdot 1_{p_1} \text{ and } f(1) \in \mathbb{C} \cdot 1_{p_0} \otimes M_{p_1}\}, \tag{1-1}$$

where $p_0, p_1 \in \mathbb{N}$ are coprime, and it can be defined as the unique simple, monotracial limit of such algebras. It has also been realized as an inductive limit of generalized dimension-drop algebras, which are defined as in (1-1), but with p_0, p_1 taken to be coprime supernatural numbers (so that M_{p_i} denotes a UHF algebra) [Rørdam and Winter 2010, Theorem 3.4]. The connecting maps in this inductive limit have the crucial feature of being trace-collapsing.

2. Decomposition rank of homomorphisms

In this section we introduce the notions of decomposition rank and nuclear dimension of *-homomorphisms, building naturally on the respective notions for C*-algebras, just as nuclearity for *-homomorphisms arises from the completely positive approximation property for C*-algebras. We first recall from [Winter 2003] the notion of completely positive contractive (c.p.c.) order zero maps.

Definition 2.1. Let A, B be C*-algebras and let $\phi : A \to B$ be a c.p.c. map. We say that ϕ has order zero if it preserves orthogonality in the sense that if $a, b \in A_+$ satisfy $ab = 0$, then $\phi(a)\phi(b) = 0$.

Definition 2.2. Let $\alpha : A \to B$ be a *-homomorphism of C*-algebras. We say that α has decomposition rank at most n, and write $\mathrm{dr}(\alpha) \leq n$, if, for any finite subset $\mathcal{F} \subset A$ and any $\epsilon > 0$, there exists a finite dimensional C*-algebra F and c.p.c. maps

$$\psi : A \to F \quad \text{and} \quad \phi : F \to B$$

such that ϕ is $(n+1)$-colourable, in the sense that we can write

$$F = F^{(0)} \oplus \cdots \oplus F^{(n)}$$

and $\phi|_{F^{(i)}}$ has order zero for all i, and such that $\phi\psi$ is point-norm close to α, in the sense that, for $a \in \mathcal{F}$,

$$\|\alpha(a) - \phi\psi(a)\| < \epsilon.$$

We may define nuclear dimension of α similarly (and write $\dim_{\mathrm{nuc}}(\alpha) \leq n$), where instead of requiring that ϕ is contractive, we only ask that $\phi|_{F^{(i)}}$ is contractive for each i.

Remark 2.3. The decomposition rank (respectively nuclear dimension) of a C*-algebra, as defined in [Kirchberg and Winter 2004, Definition 3.1] (respectively [Winter and Zacharias 2010, Definition 2.1]) is just the decomposition rank (respectively nuclear dimension) of the identity map.

The following generalizes some permanence properties for decomposition rank and nuclear dimension of C*-algebras. Proofs are omitted, as they are essentially the same as those found in [Kirchberg and Winter 2004; Winter 2003; Winter and Zacharias 2010].

Proposition 2.4. *Let A, B be C*-algebras and let $\alpha : A \to B$ be a *-homomorphism.*

(i) *Suppose that A is locally approximated by a family of C*-subalgebras $(A_\lambda)_\Lambda$, in the sense that, for every finite subset $\mathcal{F} \subset A$ and every tolerance $\epsilon > 0$, there exists λ such that $\mathcal{F} \subset_\epsilon A_\lambda$. Then*

$$\mathrm{dr}(\alpha) \leq \sup_\Lambda \mathrm{dr}(\alpha|_{A_\lambda}) \quad \textit{and} \quad \dim_{\mathrm{nuc}}(\alpha) \leq \sup_\Lambda \dim_{\mathrm{nuc}}(\alpha|_{A_\lambda}).$$

(ii) *If $C \subset A$ is a hereditary C*-subalgebra, then*

$$\mathrm{dr}(\alpha_C) \leq \mathrm{dr}(\alpha) \quad \textit{and} \quad \dim_{\mathrm{nuc}}(\alpha_C) \leq \dim_{\mathrm{nuc}}(\alpha),$$

where $\alpha_C := \alpha|_C : C \to \mathrm{her}(\alpha(C))$.

When computing the decomposition rank (or nuclear dimension), it is often convenient to replace the codomain by its sequence algebra, defined to be

$$A_\infty := \Big(\prod_{\mathbb{N}} A\Big) \Big/ \Big(\bigoplus_{\mathbb{N}} A\Big).$$

We shall denote by

$$\pi_\infty : \prod_{\mathbb{N}} A \to A_\infty$$

the quotient map, and by $\iota_\infty : A \to A_\infty$ the canonical embedding as constant sequences.

Proposition 2.5. *Let $\alpha : A \to B$ be a *-homomorphism.*

Then

$$\mathrm{dr}(\alpha) = \mathrm{dr}(\iota_\infty \circ \alpha) \quad \textit{and} \quad \dim_{\mathrm{nuc}}(\alpha) = \dim_{\mathrm{nuc}}(\iota_\infty \circ \alpha).$$

Proof. Straightforward, using stability of the relations defining c.p.c. order zero maps on finite dimensional domains [Kirchberg and Winter 2004]. □

Proposition 2.6. *Let $\mathcal{D}$ be a strongly self-absorbing* C*-*algebra* (as defined in [Toms and Winter 2007]), *and let A be a $\mathcal{D}$-stable* C*-*algebra.*

Then

$$\mathrm{dr}(A) = \mathrm{dr}(\mathrm{id}_A \otimes 1_{\mathcal{D}}) \quad \textit{and} \quad \dim_{\mathrm{nuc}}(A) = \dim_{\mathrm{nuc}}(\mathrm{id}_A \otimes 1_{\mathcal{D}}).$$

Proof. This follows easily from the fact that id_A has approximate factorizations of the form

$$\mathcal{D} \xrightarrow{\mathrm{id}_A \otimes 1_{\mathcal{D}}} \mathcal{D} \otimes \mathcal{D} \xrightarrow{\phi} A \otimes \mathcal{D},$$

where ϕ is a *-isomorphism. □

3. $C(X)$-algebras and decomposition rank

For a locally compact Hausdorff space X, a $C_0(X)$-algebra is a C*-algebra A equipped with a nondegenerate *-homomorphism $C_0(X) \to Z\mathcal{M}(A)$, called the structure map [Kasparov 1988, Definition 1.5]. Here $\mathcal{M}(A)$ refers to the multiplier algebra of A and $Z\mathcal{M}(A)$ to its centre; note that if A is unital, so is the structure map. In this section, we study the decomposition rank of such structure maps. Proposition 3.2 below is reminiscent of [Winter 2003, Proposition 2.19], which shows that the completely positive rank of $C(X)$ equals the covering dimension of X.

Definition 3.1. Let A be a $C_0(X)$-algebra and let $a \in A$. Define the support of a to be the smallest closed set $F \subset X$ such that $ag = 0$ whenever $g \in C_0(X \setminus F) \subset C_0(X)$. (This is easily seen to be well defined.)

We note the following property of order zero maps, which was obtained in the proof of [Kirchberg and Winter 2004, Proposition 5.1] (sixth line from the bottom of p. 79): if $\phi : A \to B$ is an order zero map and A is a unital C*-algebra, then

$$\|\phi(x)\| = \|\phi(1_A)\|\,\|x\| \quad \text{for any } x \in A. \tag{3-1}$$

Proposition 3.2. *Let X be a compact Hausdorff space, and let A be a unital $C(X)$-algebra with structure map $\iota : C(X) \to Z(A)$.*

The following are equivalent:

(i) $\mathrm{dr}(\iota) \le n$.

(ii) $\dim_{\mathrm{nuc}}(\iota) \le n$.

(iii) *The definition of* $\mathrm{dr}(\iota) \le n$ *holds with the additional requirements that F is abelian and ψ is a unital* *-*homomorphism.*

(iv) *For any finite open cover $\mathfrak{U}$ of X, any $\epsilon > 0$, and any $b \in C(X)_+$, there exists an $(n+1)$-colourable ϵ-approximate finite partition of b; that is, positive elements $b_j^{(i)} \in A$ for $i = 0, \dots, n$, $j = 1, \dots, r$, such that*

  (a) *for each i, the elements $b_1^{(i)}, \dots, b_r^{(i)}$ are pairwise orthogonal,*

  (b) *for each i, j, the support of $b_j^{(i)}$ is contained in some open set in the given cover $\mathfrak{U}$, and*

  (c) $\left\| \sum_{i,j} b_j^{(i)} - \iota(b) \right\| \le \epsilon$.

Proof. (iii) $\Rightarrow$ (i) $\Rightarrow$ (ii) is obvious.

(ii) $\Rightarrow$ (iv). Let us first assume $b = 1$. Let $\mathcal{F}$ be a finite partition of unity such that, for each $f \in \mathcal{F}$, there exists $U_f \in \mathcal{U}$ such that supp $f \subset U_f$. Set

$$\eta := \frac{\epsilon}{2|\mathcal{F}|(n+1)}. \tag{3-2}$$

Use $\dim_{\mathrm{nuc}}(\iota) \leq n$ to obtain

$$C(X) \xrightarrow{\psi} F^{(0)} \oplus \cdots \oplus F^{(n)} \xrightarrow{\phi} A$$

such that ψ is c.p.c., $\phi|_{F^{(i)}}$ is c.p.c. and order zero for all $i = 0, \ldots, n$, $\phi(\psi(f)) =_\eta f$ for $f \in \mathcal{F}$, and $\phi(\psi(1)) =_{\epsilon/2} 1$. Set

$$F^{(i)} := \bigoplus_{j=1}^{r_i} M_{m(i,j)}.$$

(By throwing in some zero summands if necessary, we may as well assume all the r_i to be equal.)

For each $i = 0, \ldots, n$ and $j = 1, \ldots, r_i$, we set

$$a_j^{(i)} := \left(\phi(\psi(1_{C(X)})1_{M_{m(i,j)}}) - \frac{\epsilon}{2(n+1)}\right)_+.$$

For each i, since $\phi|_{F^{(i)}}$ is order zero, $a_1^{(i)}, \ldots, a_{r_i}^{(i)}$ are orthogonal. We estimate

$$1 =_{\epsilon/2} \phi(\psi(1)) = \sum_{i=0}^{n} \sum_{j=1}^{r_i} \phi(\psi(1)1_{M_{m(i,j)}}) =_{\frac{(n+1)\epsilon}{2(n+1)}} \sum_{i,j} a_j^{(i)},$$

where the last approximation is obtained using the fact that the inner summands are orthogonal.

Lastly, we must verify that each $a_j^{(i)}$ has support contained in an open set from the cover $\mathcal{U}$. Fix i and j. Let $f_{i,j} \in \mathcal{F}$ maximize $f \mapsto \|\psi(f)1_{M_{m(i,j)}}\|$. We shall show that the support of $a_j^{(i)}$ is contained in the support of $f_{i,j}$ by showing that $a_j^{(i)}|_K = 0$, where

$$K := \{x \in X : f_{i,j} = 0\}.$$

Since $1 = \sum_{f \in \mathcal{F}} f$, we must have

$$\|\psi(f_{i,j})1_{M_{m(i,j)}}\| \geq \frac{1}{|\mathcal{F}|}\|\psi(1)1_{M_{m(i,j)}}\|. \tag{3-3}$$

Noting that

$$f_{i,j} =_\eta \phi(\psi(f_{i,j})) \geq \phi(\psi(f_{i,j})1_{M_{m(i,j)}}),$$

we must have

$$\|\phi(\psi(f_{i,j})1_{M_{m(i,j)}})|_K\| \leq \eta. \tag{3-4}$$

We get

$$\begin{aligned}\|\phi(\psi(1)1_{M_{m(i,j)}})|_K\| &\overset{(3\text{-}1)}{=} \|\phi(1_{M_{m(i,j)}})|_K\|\,\|\psi(1)1_{M_{m(i,j)}}\| \\ &\overset{(3\text{-}3)}{\leq} \|\phi(1_{M_{m(i,j)}})|_K\|\,|\mathcal{F}|\,\|\psi(f_{i,j})1_{M_{m(i,j)}}\| \\ &\overset{(3\text{-}1)}{=} \|\phi(\psi(f_{i,j})1_{M_{m(i,j)}})|_K\| \\ &\underset{(3\text{-}4)}{\overset{(3\text{-}2)}{\leq}} \frac{\epsilon}{2(n+1)};\end{aligned}$$

therefore, $a_j^{(i)}|_K = 0$, as required.

If b is not the unit, we may still assume that $\|b\| \leq 1$ and use the argument above to obtain an $(n+1)$-colourable approximate partition of unity $(a_j^{(i)})$ subordinate to $\mathcal{U}$. Then simply set $b_j^{(i)} = ba_j^{(i)}$.

(iv) $\Rightarrow$ (iii). It will suffice to prove the condition in (iii) assuming that $\mathcal{F}$ consists of self-adjoint contractions.

Take an open cover $\mathcal{U}$ of X along with points $x_U \in U$ for every $U \in \mathcal{U}$ such that, for any $f \in \mathcal{F}$, $U \in \mathcal{U}$, and $x \in U$,

$$|f(x) - f(x_U)| < \frac{\epsilon}{2}. \tag{3-5}$$

Use (iv) with $b = 1$ to find an $(n+1)$-colourable $\epsilon/2$-approximate partition of unity

$$(a_j^{(i)})_{i=0,\ldots,n;\, j=1,\ldots,r}$$

subordinate to $\mathcal{U}$. By a standard rescaling argument, we may assume that $\sum a_j^{(i)} \leq 1$. For each i, j, let $U(i,j) \in \mathcal{U}$ be such that $\operatorname{supp} a_j^{(i)} \subset U(i,j)$.

Define $\psi : C(X) \to (\mathbb{C}^r)^n$ by

$$\psi(f) = (f(x_{U(i,j)}))_{i=0,\ldots,n;\, j=1,\ldots,r}$$

and define $\phi : (\mathbb{C}^r)^n \to C(X, A)$ by

$$\phi(\lambda_{i,j})_{i=0,\ldots,n;\, j=1,\ldots,r} = \sum_{i,j} \lambda_{i,j} \cdot a_j^{(i)}.$$

Clearly, ψ is a *-homomorphism, while ϕ is c.p.c. and its restriction to each copy of $\mathbb{C}^r$ is order zero.

To verify that $\phi \circ \psi$ approximates θ in the appropriate sense, fix $f \in \mathcal{F}$ and $x \in X$. We shall show that $\|\phi\psi(f)(x) - f(x)\| < \epsilon$ (in the fibre $A(x)$). Let

$$S = \{(i,j) \in \{0,\ldots,n\} \times \{1,\ldots,r\} : x \in U(i,j)\},$$

so that

$$\phi(\psi(f))(x) = \sum_{(i,j)\in S} f(x_{U(i,j)}) \cdot a_j^{(i)}(x) \quad \text{and} \quad 1 =_{\epsilon/2} \sum_{(i,j)\in S} a_j^{(i)}(x).$$

By (3-5),

$$(f(x)-\epsilon/2)\cdot\sum_{(i,j)\in S}a_j^{(i)}(x)\le\sum_{(i,j)\in S}f(x_{U(i,j)})\cdot a_j^{(i)}(x)\le(f(x)+\epsilon/2)\cdot\sum_{(i,j)\in S}a_j^{(i)}(x).$$

It follows that

$$\phi(\psi(f))=\sum_{(i,j)\in S}f(x_{U(i,j)})\cdot a_j^{(i)}=_{\epsilon/2}f(x)\cdot\sum_{(i,j)\in S}a_j^{(i)}=_{\epsilon/2}f(x),$$

as required. □

Proposition 3.3. *Let X be a locally compact metrizable space with finite covering dimension, and let A be a $C_0(X)$-algebra all of whose fibres are isomorphic to $\mathbb{O}_2$. Let $U\subset X$ be an open subset such that $\overline{U}$ is compact.*

Then $C_0(U)A\cong C_0(U,\mathbb{O}_2)$ as $C_0(U)$-algebras.

Proof. [Dadarlat 2009a, Theorem 1.1] says that $A|_{\overline{U}}\cong C(\overline{U},\mathbb{O}_2)$, as $C(\overline{U})$-algebras. Viewing $C_0(U)A$ as an ideal of $A|_{\overline{U}}$, the result follows. □

4. Decomposition rank of $C_0(X,\mathcal{Z})$

In this section, we prove our main result.

Theorem 4.1. *Let A be a* C*-*algebra which is locally approximated by hereditary subalgebras of* C*-*algebras of the form $C(X,\mathcal{K})$, with X compact Hausdorff.*

Then

$$\mathrm{dr}(A\otimes\mathcal{Z})\le 2.$$

In particular, any $\mathcal{Z}$-stable AH C*-*algebra has decomposition rank at most* 2.

In our proof, we will make use of the huge amount of space provided by the noncommutative fibres in two ways. First, we exhaust the identity on X by pairwise orthogonal functions up to a tracially small hereditary subalgebra. This will be designed to host an algebra of the form $C_0(Z)\otimes\mathbb{O}_2$, which is possible by quasidiagonality of the cone over $\mathbb{O}_2$. The first factor embedding of $C_0(Z)$ into the latter can be approximated by 2-colourable maps as shown by Kirchberg and Rørdam (see below). Together with the initial set of functions, we obtain a 3-colourable, hence 2-dimensional, approximation of the first factor embedding of $C(X)$ into $C(X)\otimes\mathcal{Z}$.

We will first carry out this construction with a UHF algebra in place of $\mathcal{Z}$; a slight modification will then allow us to pass to certain $C([0,1])$-algebras with UHF fibres, which immediately yields the general case.

In fact, if one is only concerned with showing that $A\otimes\mathcal{Z}$ has finite decomposition rank, our argument can be significantly shortened; using [Carrión 2011, Lemma 3.1], it suffices to show that $A\otimes U$ has finite decomposition rank, when U is an infinite dimensional, self-absorbing UHF algebra. Remarks 4.8

and 4.9 describe how one can easily modify (and skip some long technicalities in) the arguments below in order to efficiently prove that $A \otimes \mathcal{Z}$ has finite decomposition rank.

As noted above, a result of Kirchberg and Rørdam [2005, Proposition 3.7] on 1-dimensional approximations in the case of $\mathbb{O}_2$-fibred bundles is a crucial ingredient; this in turn relies on the fact that the unitary group of $C(S^1, \mathbb{O}_2)$ is connected [Cuntz 1981]. We note the following direct consequence which is more adapted to our needs.

Theorem 4.2. *For any locally compact Hausdorff space X, the decomposition rank of the first factor embedding $C_0(X) \to C_0(X, \mathbb{O}_2)$ is at most one.*

Proof. We begin with the case that X is compact and metrizable. By [Kirchberg and Rørdam 2005, Proposition 3.7], there exists a *-subalgebra $A \subset C(X, \mathbb{O}_2)$ which contains $C(X) \otimes 1_{\mathbb{O}_2}$ and is isomorphic to $C(Y)$ where Y is compact metrizable with covering dimension at most one. Therefore, the decomposition rank of the first factor embedding $C(X) \to C(X) \otimes \mathbb{O}_2$ is at most the decomposition rank of the inclusion $C(X) \otimes 1_{\mathbb{O}_2} \subset A$, which in turn is at most $\operatorname{dr} A \leq 1$.

For X compact but not metrizable, $C(X)$ is locally approximated by finitely generated unital subalgebras, which are of the form $C(Y)$ where Y is compact and metrizable. Therefore, by Proposition 2.4(i), the claim holds in this case too.

For the case that X is not compact, we let $\widetilde{X}$ denote the one-point compactification of X. Then $C_0(X, \mathbb{O}_2)$ is the hereditary subalgebra of $C(\widetilde{X}, \mathbb{O}_2)$ generated by $C_0(X)$, and therefore the result follows from Proposition 2.4(ii). □

Remark 4.3. The preceding result also implies that $\dim_{\mathrm{nuc}}(A \otimes \mathbb{O}_2) \leq 3$ for A as in Theorem 4.1 — this can be seen using Proposition 2.6, [Winter and Zacharias 2010, Theorem 7.4], and the analogue of [Winter and Zacharias 2010, Proposition 2.3(ii)].

In what follows, D_n denotes the diagonal subalgebra of M_n.

Lemma 4.4. *Let $I_1, \ldots, I_n \subset (0, 1)$ be nonempty closed intervals and let $a_{1/2} \in C_0((0, 1), D_n)_+$ be a function of norm 1 such that, for $t \in I_s$, the s-th diagonal entry of $a_{1/2}(t)$ is 1.*

Then there exist $a_0, a_1, e_0, e_{1/2}, e_1 \in C([0, 1], D_n)_+$ such that

(i) *e_0 and e_1 are orthogonal,*

(ii) *$a_0 + a_{1/2} + a_1 = e_0 + e_{1/2} + e_1 = 1$,*

(iii) *for $i = 0, 1$, we have $a_i(i) = 1_n$,*

(iv) *e_0, e_1 act like a unit on a_0, a_1, respectively, and*

(v) *$a_{1/2}$ acts like a unit on $e_{1/2}$.*

Proof. Since $D_n \cong \mathbb{C}^n$, it suffices to work in one coordinate at a time — that is, to assume that $n = 1$. Then define

$$a_0(x) := \begin{cases} 1 - a_{1/2}(x) & \text{if } x \text{ is to the left of } I_1, \\ 0 & \text{otherwise,} \end{cases}$$

$$a_1(x) := \begin{cases} 1 - a_{1/2}(x) & \text{if } x \text{ is to the right of } I_1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that since $a_{1/2} \equiv 1$ on I_1, these are continuous. Now, we may find continuous orthogonal functions e_0, e_1 such that e_0 is 1 to the left of I_1, and e_1 is 1 to the right of I_1. Finally, set $e_{1/2} := 1 - (e_0 + e_1)$. Then (i), (ii), and (iii) clearly hold by construction. (iv) holds since each a_i is nonzero only on one side of I_1, and the corresponding e_i is identically 1 on that side. Likewise, (v) holds since $e_{1/2}$ is nonzero only on I_1, where $a_{1/2}$ is identically 1. □

We mention the following well-known fact explicitly for convenience. Here $\otimes$ denotes the minimal tensor product.

Proposition 4.5. *Let A_1, A_2, B_1, and B_2 be C*-algebras, and suppose that $\phi^{(i)} : A_i \to (B_i)_\infty$ is a *-homomorphism for $i = 1, 2$ with a c.p. lift $(\phi_k^{(i)})_\mathbb{N} : A_i \to \prod_\mathbb{N} B_i$.*

Then

$$\phi_1 \otimes \phi_2 = \pi_\infty \circ (\phi_k^{(1)} \otimes \phi_k^{(2)})_\mathbb{N} : A_1 \otimes A_2 \to (B_1 \otimes B_2)_\infty$$

*is a *-homomorphism.*

Lemma 4.6. *Let A be an infinite dimensional UHF algebra.*

Then there exist positive orthogonal contractions

$$a_0, a_1 \in C([0, 1], A)_\infty,$$

*a *-homomorphism*

$$\psi : C_0(Z, \mathbb{O}_2) \to C_0((0, 1), A)_\infty,$$

where $Z = (0, 1]^2$, and a positive element $c \in C_c(Z, \mathbb{C} \cdot 1_{\mathbb{O}_2})$ such that $\psi(c)$ commutes with a_0, a_1,

$$a_0 + a_1 + \psi(c) = 1, \tag{4-1}$$

and $a_0(0) = a_1(1) = 1$. In addition, there exist positive contractions $e_0, e_{1/2}, e_1 \in C([0, 1], A)_\infty$ such that

(i) *e_0, e_1 are orthogonal,*

(ii) *$e_0 + e_{1/2} + e_1 = 1$,*

(iii) *$\psi(c)$ acts like a unit on $e_{1/2}$,*

(iv) *e_i acts like a unit on a_i for $i = 0, 1$, and*

(v) *$e_0, e_{1/2}, e_1, a_0, a_1, \psi(c)$ all commute.*

Proof. Let $A = M_{n_1 n_2 \cdots}$ where $n_1, n_2, \ldots$ are a sequence of natural numbers ≥ 2. Since the cone over $\mathbb{O}_2$ is quasidiagonal (see [Voiculescu 1991] and [Kirchberg 1993, Theorem 5.1]) there exists a sequence of c.p.c. maps

$$\phi_k : C_0((0, 1], \mathbb{O}_2) \to M_{n_1 \cdots n_k}$$

which are approximately multiplicative and approximately isometric, meaning that

$$\|\phi_k(a)\phi_k(b) - \phi_k(ab)\| \to 0 \quad \text{and} \quad \|\phi_k(a)\| \to \|a\| \quad \text{as } k \to \infty$$

for all $a, b \in C_0((0,1], \mathbb{O}_2)$. Fix a positive element

$$d \in C_c((0,1], \mathbb{C} \cdot 1_{\mathbb{O}_2})$$

of norm 1.

For each k, let λ_k denote the greatest eigenvalue of $\phi_k(d)$. Note that

$$\lambda_k = \|\phi_k(d)\| \to 1 \quad \text{as } k \to \infty.$$

Fix k for a moment and let $l = n_1 \cdots n_k$. Let

$$I_1, \dots, I_l$$

be nonempty disjoint closed intervals in $(0,1)$. Let

$$u_1, \dots, u_l \in M_l$$

be unitaries such that, for each s, $u_s \phi_k(d) u_s^*$ is a diagonal matrix whose s-th diagonal entry is λ_k. Let

$$h_1, \dots, h_l \in C_0((0,1))$$

be positive normalized functions with disjoint support, such that $h_s|_{I_s} \equiv 1$ for each s. Set $Z := (0,1]^2$ and define

$$\psi_k : C_0(Z, \mathbb{O}_2) \cong C_0((0,1]) \otimes C_0((0,1], \mathbb{O}_2) \to C([0,1]) \otimes M_l \cong C([0,1], M_l) \subset C([0,1], A)$$

by

$$\psi_k(f \otimes b) = \sum_{s=1}^{l} f(h_s) \otimes u_s \phi_k(b) u_s^*.$$

Let $f \in C_c((0,1])$ be a function satisfying $f(1) = 1$ and set

$$c = f \otimes d \in C_c(Z, \mathbb{C} \cdot 1_{\mathbb{O}_2}).$$

By construction, $\psi_k(c) \in C((0,1), D_l)_+$, and for $t \in I_s$, the s-th diagonal entry is λ_k. Let

$$c_k' \in C([0,1], D_l)_+$$

be of norm 1, such that

$$\|c_k' - \psi_k(c)\| = |1 - \lambda_k|$$

and for $t \in I_s$, the s-th diagonal entry is 1. Feeding

$$a_{1/2} := c_k'$$

to Lemma 4.4, let

$$a_{0,k}, a_{1,k}, e_{0,k}, e_{1/2,k}, e_{1,k} \in C([0,1], D_l)_+$$

be the output, satisfying (i)–(v) of Lemma 4.4.

Having found these for each k, set

$$\psi := \pi_\infty \circ (\psi_1, \psi_2, \dots) : C_0(Z, \mathbb{O}_2) \to C([0,1], A)_\infty.$$

Set

$$a_i := \pi_\infty(a_{i,1}, a_{i,2}, \dots)$$

for $i = 0, 1$ and

$$e_i := \pi_\infty(e_{i,1}, e_{i,2}, \dots)$$

for $i = 0, \frac{1}{2}, 1$.

Since all unitaries in M_l (and in particular, all the u_s) are in the same path component, ψ_k is unitarily equivalent to $\alpha \otimes \phi_k$, where

$$\alpha : C_0((0,1]) \to C([0,1])$$

is the *-homomorphism given by

$$f \mapsto f(h_1 + \cdots + h_l).$$

From this observation and Proposition 4.5, it follows that ψ is a *-homomorphism.

Notice further that

$$\psi(c) = \pi_\infty(c'_1, c'_2, \dots),$$

and therefore, drawing on the finite stage results, we see that

$$a_0 + a_1 + \psi(c) = 1$$

and that (i)–(v) hold. □

Lemma 4.7. *Let $p, q > 1$ be natural numbers. Let $X = [0,1]^m$ for some m and let $\epsilon > 0$.*

Then there exist positive orthogonal elements

$$h_0, \dots, h_k \in C(X, \mathcal{Z})_\infty,$$

*a *-homomorphism*

$$\phi : C_0(Z, \mathbb{O}_2) \to C(X, \mathcal{Z})_\infty$$

for some locally compact, metrizable, finite dimensional space Z, and a positive element $c \in C_c(Z, \mathbb{C}\cdot 1_{\mathbb{O}_2})$ such that $\phi(c)$ commutes with $h_0, \dots, h_k$,

$$h_0 + \cdots + h_k + \phi(c) = 1,$$

and the support of h_i has diameter at most ϵ for $i = 0, \dots, k$ with respect to the ℓ^∞ metric on $[0,1]^m$.

In addition, there exist positive contractions $e_0, e_{1/2}, e_1 \in C(X, \mathcal{Z})_\infty$ such that

(i) *e_0, e_1 are orthogonal,*

(ii) *$e_0 + e_{1/2} + e_1 = 1$,*

(iii) *e_j is identically 1 on $\{j\} \times [0,1]^{m-1}$, for $j = 0, 1$,*

(iv) *$\phi(c)$ acts like a unit on $e_{1/2}$,*

(v) *$e_0 + e_1$ acts like a unit on h_i for all $i = 0, \dots, k$, and*

(vi) *$e_0, e_{1/2}, e_1, h_0, \dots, h_k, \phi(c)$ all commute.*

Remark 4.8. This lemma of course holds with a self-absorbing UHF algebra in place of $\mathcal{Z}$ (since such a UHF algebra contains $\mathcal{Z}$). But, in fact, this variation is shown in Steps 1 and 2 of the proof below, and, as we will see in Remark 4.9, this variation is sufficient to prove that $A \otimes \mathcal{Z}$ as in Theorem 4.1 has finite decomposition rank. A reader only interested in showing finite decomposition rank may therefore skip the third step of the proof below.

Proof. This will be proven in three steps. In Step 1, we will prove the statement of the proposition with $\mathcal{Z}$ replaced by a UHF algebra of infinite type and with $m = 1$. In Step 2, we will still replace A by a UHF algebra of infinite type, but we will allow any $m \in \mathbb{N}$. Step 3 will be the proof of the proposition.

Step 1. Let A be a UHF algebra of infinite type. Let

$$a_0, a_1, \psi, c, e_0', e_{1/2}', e_1', Z$$

be as in Lemma 4.6, with e_i' in place of e_i. Note that each a_i has a positive normalized lift

$$(a_{i,j})_{j=1}^{\infty} \in \prod_{\mathbb{N}} C([0,1], A)$$

such that $a_{i,j}(t) = \delta_{i,t} 1$ for all $i, t = 0, 1$ and all j; likewise, each e_i', $i = 0, \frac{1}{2}, 1$, has a positive normalized lift

$$(e_{i,j}')_{j=1}^{\infty} \in \prod_{\mathbb{N}} C([0,1], A)$$

such that, for $i = 0, 1$, $e_{i,j}'(i) = 1$.

Let $k \geq 2/\epsilon$ be a natural number. For $i = 0, \ldots, k$, $j \in \mathbb{N}$, and $t \in [0,1]$, set

$$h_{i,j}(t) := \begin{cases} 0 & \text{if } t \leq \dfrac{i-1}{k} \text{ or } t \geq \dfrac{i+1}{k}, \\ a_{1,j}(kt - (i-1)) & \text{if } t \in \left[\dfrac{i-1}{k}, \dfrac{i}{k}\right], \\ a_{0,j}(kt - i) & \text{if } t \in \left[\dfrac{i}{k}, \dfrac{i+1}{k}\right]. \end{cases} \tag{4-2}$$

Note that the endpoint conditions on $a_{i,j}$ make $h_{i,j}$ well defined and continuous on $[0,1]$. Likewise, set

$$e_{i,j}(t) := \begin{cases} e_{i,j}'(0) & \text{if } t = 0, \\ e_{i,j}'(1) & \text{if } t \geq \dfrac{1}{k}, \\ e_{i,j}'(kt) & \text{if } t \in \left[0, \dfrac{1}{k}\right]. \end{cases} \tag{4-3}$$

Set

$$h_i := \pi_\infty(h_{i,1}, h_{i,2}, \ldots), e_i := \pi_\infty(e_{i,1}, e_{i,2}, \ldots) \in C([0,1], A)_\infty$$

for $i = 0, 1$. Choose a c.p.c. lift for ψ, that is, c.p.c. maps

$$\psi_j : C_0(Z, \mathbb{O}_2) \to C_0((0,1), A) \subset C([0,1], A)$$

such that

$$\psi := \pi_\infty \circ (\psi_1, \psi_2, \ldots).$$

Define

$$\phi_j : C_0(Z, \mathbb{O}_2) \to C([0,1], A)$$

by

$$\phi_j(a)(t) = \psi_j(a)(kt - i), \tag{4-4}$$

if $i \in \mathbb{N}$ is such that $t \in [i/k, (i+1)/k]$. Note that this is well defined since the image of ψ_j is contained in $C_0((0,1), A)$. Use $(\phi_j)_{j=1}^{\infty}$ to define

$$\phi = \pi_\infty \circ (\phi_1, \phi_2, \dots) : C_0(Z, \mathbb{O}_2) \to C([0,1], A)_\infty.$$

Then ϕ is a *-homomorphism.

Let us first show that $h_0 + \cdots + h_k + \phi(c) = 1$, and then that (i)–(vi) hold. For $t \in [0,1]$, let i be such that $t \in [i/k, (i+1)/k]$. Then, by (4-2), we have, for all i,

$$h_i(t) = a_0(kt - i), \quad h_{i+1}(t) = a_1(kt - i), \quad h_j(t) = 0$$

for $j \neq i, i+1$. Thus

$$(h_0 + \cdots + h_k + \phi(c))(t) \overset{(4\text{-}4)}{=} a_0(kt-i) + a_1(kt-i) + \psi(c)(kt-i) \overset{(4\text{-}1)}{=} 1.$$

Properties (i) and (ii) hold by Lemma 4.6(i) and (ii), and since, for each $t \in [0,1]$, there exists s such that $e_j(t) = e'_j(s)$ for $j = 0, \frac{1}{2}, 1$ (by (4-3)). Property (iii) holds since $e_i(i) = e'_i(i)$ (by (4-3)) and since $a_i(i) = 1$.

(iv): $e_{1/2}$ is supported on $[0, 1/k]$, so it suffices to show that

$$(\phi(c)e_{1/2})(t) = e_{1/2}(t)$$

for $t \in [0, 1/k]$. But, for such t,

$$(\phi(c)e_{1/2})(t) \underset{(4\text{-}4)}{\overset{(4\text{-}3)}{=}} \psi(c)(kt)e'_{1/2}(kt) \overset{\text{Lemma 4.6(iii)}}{=} e'_{1/2}(kt) \overset{(4\text{-}3)}{=} e_{1/2}(t).$$

(v): By a similar computation, this time using Lemma 4.6(iv), we see that $e_0 a_0 = a_0$, while $e_1 a_i = a_i$ for $i = 1, \dots, k$.

(vi) is clear from (4-2), (4-3), (4-4), and Lemma 4.6(v).

Finally, also, for each i, the support of h_i is contained in $\left[\frac{i-1}{k}, \frac{i+1}{k}\right]$, which has diameter at most ϵ.

Step 2. From Step 1, let

$$g_0, \dots, g_{k'} \in C([0,1], A)_\infty$$

be orthogonal positive contractions,

$$\psi : C_0(Y, \mathbb{O}_2) \to C([0,1], A)_\infty$$

a *-homomorphism for some locally compact, metrizable, finite dimensional space Y, and $d \in C_c(Y, \mathbb{C}\cdot 1_{\mathbb{O}_2})$ a positive contraction such that $\psi(d)$ commutes with $g_0, \dots, g_{k'}$,

$$g_0 + \cdots + g_{k'} + \psi(d) = 1$$

and the support of g_i has diameter at most ϵ for $i = 0, \dots, k'$; furthermore, let

$$e'_0, e'_{1/2}, e'_1 \in C([0,1], A)_\infty$$

be such that

(i′) e'_0, e'_1 are orthogonal,

(ii′) $e'_0 + e'_{1/2} + e'_1 = 1$,

(iii′) e'_j is identically 1 on $\{j\} \times [0,1]^{m-1}$, for $j = 0, 1$,

(iv′) $\psi(d)$ acts like a unit on $e'_{1/2}$,

(v′) $e'_0 + e'_1$ acts like a unit on g_i for all $i = 0, \dots, k'$, and

(vi′) $e'_0, e'_{1/2}, e'_1, g_0, \dots, g_{k'}, \psi(d)$ all commute.

For $i = (i_1, \dots, i_m) \in \{0, \dots, k'\}^m$, set

$$h_i := g_{i_1} \otimes \cdots \otimes g_{i_m} \in (C([0,1], A)^{\otimes m})_\infty,$$

where we have used the canonical inclusion

$$(C([0,1], A)_\infty)^{\otimes m} \to (C([0,1], A)^{\otimes m})_\infty;$$

compare Proposition 4.5.

Then $\{h_i\}$ is a set of pairwise orthogonal positive contractions, and each one has support with diameter at most ϵ (recall that we are using the ℓ^∞ metric on $[0,1]^m$). Proposition 4.5 gives us a *-homomorphism

$$\phi' := (\psi^\sim)^{\otimes m} : C := (C_0(Y, \mathbb{O}_2)^\sim)^{\otimes m} \to (C([0,1], M_n\infty)^{\otimes m})_\infty.$$

Set

$$c := 1 - (1-d)^{\otimes m} \in C.$$

We can easily see that $\phi'(c)$ commutes with each h_i; a simple computation shows that

$$\sum_i h_i + \phi'(c) = 1.$$

Setting

$$e_i := e'_i \otimes 1^{\otimes(m-1)} \quad \text{for } i = 0, \tfrac{1}{2}, 1,$$

it is easy to see that (i), (ii), (iii), (v), and (vi) hold (with ϕ' in place of ϕ). To see that (iv) holds, we compute

$$\begin{aligned} \phi'(c)e_{1/2} &= (1 - (1-\psi(d))^{\otimes m})(e'_{1/2} \otimes 1^{\otimes(m-1)}) \\ &= \phi'(c) - (e'_{1/2} - \psi(d)e'_{1/2}) \otimes (1-\psi(d))^{\otimes(m-1)} \\ &\overset{(\text{iv}')}{=} \phi'(c). \end{aligned}$$

We may set

$$k := (k'+1)^m - 1$$

and relabel the h_i as $h_0, \dots, h_k$.

All that remains is to modify ϕ' to make it a map whose domain is $C_0(Z, \mathbb{O}_2)$ for some Z. Set

$$Z' := (Y \amalg \{\infty\})^{\times m}.$$

Then C may be identified with a certain $C(Z')$-subalgebra of $C(Z', \mathbb{O}_2^{\otimes m})$. All of the fibres of C are isomorphic to $\mathbb{O}_2$ except for the fibre at $(\infty, \ldots, \infty)$, which is $\mathbb{C}$. One can easily verify that the element c is in $C_0(U, \mathbb{C} \cdot 1_{\mathbb{O}_2^{\otimes m}})$ where U is some open subset of Z' whose closure does not contain $(\infty, \ldots, \infty)$.

Let Z be an open subset of Z' such that $\overline{U} \subset Z$ and whose closure does not contain $(\infty, \ldots, \infty)$; in particular, $\overline{Z}$ is a compact subset of $Z' \backslash \{(\infty, \ldots, \infty)\}$. By Proposition 3.3, $C_0(Z)C \cong C_0(Z, \mathbb{O}_2)$ as $C_0(Z)$-algebras. With this identification, we have $c \in C_c(Z, \mathbb{C} \cdot 1_{\mathbb{O}_2})$ (since c is in the image of the structure map, which is fixed by the isomorphism $C_0(Z)C \cong C_0(Z, \mathbb{O}_2)$), and we may define

$$\phi := \phi'|_{C_0(Z)C} : C_0(Z, \mathbb{O}_2) \to C(X, A)_\infty.$$

Step 3. Let p_0, p_1 be coprime natural numbers. Since $\mathscr{Z}_{p_0^\infty, p_1^\infty}$ (as defined in [Rørdam and Winter 2010, Section 2]) embeds unitally into $\mathscr{Z}$ [Rørdam 2004, Proposition 2.2], it suffices to do this part with $\mathscr{Z}_{p_0^\infty, p_1^\infty}$ in place of $\mathscr{Z}$.

From Step 2, for $i = 0, 1$, we may find

$$h_0^{(i)}, \ldots, h_k^{(i)} \in C(X, M_{p_i^\infty})_\infty,$$

a *-homomorphism

$$\phi_i : C_0(Z_i, \mathbb{O}_2) \to C(X, M_{p_i^\infty})_\infty$$

for some locally compact, metrizable, finite dimensional space Z_i, and a positive element

$$c_i \in C_c(Z_i, \mathbb{C} \cdot 1_{\mathbb{O}_2})$$

such that ϕ_i commutes with $h_0^{(i)}, \ldots, h_k^{(i)}$,

$$h_0^{(i)} + \cdots + h_k^{(i)} + \phi_i(c_i) = 1,$$

and the support of $h_j^{(i)}$ has diameter at most ϵ for $j = 1, \ldots, k$. We may also find $e_l^{(i)}$ for $l = 0, \frac{1}{2}, 1$ satisfying (i)–(vi).

From Lemma 4.6, let

$$a_0, a_1, e_0', e_{1/2}', e_1' \in C\left(\left[\tfrac{1}{3}, \tfrac{2}{3}\right], A\right)_\infty$$

be positive orthogonal contractions, let

$$\psi : C_0(Y, \mathbb{O}_2) \to C_0\left(\left(\tfrac{1}{3}, \tfrac{2}{3}\right), M_{(p_0 p_1)^\infty}\right)_\infty$$

be a *-homomorphism for some locally compact, metrizable, finite dimensional space Y, and let

$$d \in C_c(Z, \mathbb{C} \cdot 1_{\mathbb{O}_2})$$

be positive such that $\psi(d)$ commutes with a_0, a_1,

$$a_0 + a_1 + \psi(d) = 1,$$

$a_0\left(\frac{1}{3}\right) = a_1\left(\frac{2}{3}\right) = 1$, and such that (i)–(v) of Lemma 4.6 hold. We continuously extend a_0, a_1, e'_0, $e'_{1/2}$, e'_1 to $[0,1]$ by allowing them to be constant on $\left[0, \frac{1}{3}\right]$ and on $\left[\frac{2}{3}, 1\right]$.

Upon choosing an isomorphism

$$M_{(p_0p_1)^\infty} \otimes M_{p_0^\infty} \otimes M_{p_1^\infty} \cong M_{p_0^\infty} \otimes M_{p_1^\infty}$$

and using the diagonal restriction $C(X, M_{p_0^\infty})_\infty \otimes C(X, M_{p_1^\infty})_\infty \to C(X, M_{p_0^\infty} \otimes M_{p_1^\infty})_\infty$, we obtain a *-homomorphism

$$\begin{aligned}\rho : C([0,1], M_{(p_0p_1)^\infty})_\infty \otimes C(X, M_{p_0^\infty})_\infty \otimes C(X, M_{p_1^\infty})_\infty &\to C([0,1]\times X, M_{(p_0p_1)^\infty} \otimes M_{p_0^\infty} \otimes M_{p_1^\infty})_\infty \\ &\cong C([0,1]\times X, M_{p_0^\infty} \otimes M_{p_1^\infty})_\infty,\end{aligned}$$

and define

$$\hat{h}_{0,j} := \rho\big(a_0 \otimes h_j^{(0)} \otimes 1_{C(X, M_{p_1^\infty})_\infty}\big) \quad \text{and} \quad \hat{h}_{1,j} := \rho\big(a_1 \otimes 1_{C(X, M_{p_0^\infty})_\infty} \otimes h_j^{(1)}\big)$$

for $j = 1, \ldots, k$. Note that a_i has a lift

$$(a_{i,k})_{k=1}^\infty \in \textstyle\prod_{\mathbb{N}} C([0,1], M_{(p_0p_1)^\infty})$$

such that $a_{i,k}(t) \in \mathbb{C} \cdot 1$ for $t = 0, 1$, and, consequently,

$$\hat{h}_{i,j} \in C(X, \mathscr{Z}_{p_0^\infty, p_1^\infty})_\infty.$$

Define a *-homomorphism

$$\phi : C([0,1]) \otimes C_0(Y, \mathbb{O}_2)^\sim \otimes C_0(Z_0, \mathbb{O}_2)^\sim \otimes C_0(Z_1, \mathbb{O}_2)^\sim \to C([0,1] \times X, M_{p_0^\infty} \otimes M_{p_1^\infty})_\infty$$

by

$$\phi := \rho \circ \big(\mathrm{id}_{C([0,1])} \otimes (\psi^\sim) \otimes (\phi_0^\sim) \otimes (\phi_1^\sim)\big).$$

Let

$$Y' := \{y \in Y : d(y) > 0\} \quad \text{and} \quad Z_i' := \{z \in Z_i : c_i(z) \neq 0\},$$

and, using these, set

$$\begin{aligned}C := \mathrm{C}^*\big(&C_0[0,1) \otimes 1_{C_0(Y,\mathbb{O}_2)^\sim} \otimes C_0(Z_0', \mathbb{O}_2) \otimes 1_{C_0(Z_1,\mathbb{O}_2)^\sim}, \\ &C_0(0,1] \otimes 1_{C_0(Y,\mathbb{O}_2)^\sim} \otimes 1_{C_0(Z_0,\mathbb{O}_2)^\sim} \otimes C_0(Z_1', \mathbb{O}_2), \\ &1_{C([0,1])} \otimes C_0(Y', \mathbb{O}_2) \otimes 1_{C_0(Z_0,\mathbb{O}_2)^\sim} \otimes 1_{C_0(Z_1,\mathbb{O}_2)^\sim}\big).\end{aligned}$$

Using Proposition 3.3 as in Step 2, C is a subalgebra of some $C_0(Z)$-algebra

$$D \subset C[0,1] \otimes C_0(Y, \mathbb{O}_2)^\sim \otimes C_0(Z_0, \mathbb{O}_2) \otimes C_0(Z_1, \mathbb{O}_2)$$

for some open subset Z of

$$[0,1] \times (Y' \cup \{\infty\}) \times (Z_0' \cup \infty) \times (Z_1' \cup \infty),$$

and D is isomorphic, as a $C_0(Z)$-algebra, to $C_0(Z, \mathbb{O}_2)$, via an isomorphism taking C into $C_c(Z, \mathbb{O}_2)$. One easily sees that $\phi(C) \subset C(X, \mathcal{Z}_{p_0,p_1})$.

Let $f_0 \in C_0[0,1)_+$ be identically 1 on $\left[0, \frac{2}{3}\right]$, and let $f_1 \in C_0(0,1]_+$ be identically 1 on $\left[\frac{1}{3}, 1\right]$. Set

$$\hat{c} := f_0 \otimes 1 \otimes c_0 \otimes 1 + f_1 \otimes 1 \otimes 1 \otimes c_1 + 1 \otimes d \otimes 1 \otimes 1 \in C.$$

Identifying D with $C_0(Z, \mathbb{O}_2)$, we see that $\hat{c} \in C_c(Z, \mathbb{C} \cdot 1_{\mathbb{O}_2})$. It is straightforward to check that $\phi(\hat{c})$ commutes with $\hat{h}_{i,j}$ for all i, j, and we may easily compute

$$\phi(\hat{c}) + \sum_{i,j} \hat{h}_{i,j} \geq 1.$$

Let $g \in C_0(0, \infty]$ be the function $g(t) = \max\{t, 1\}$ and set

$$c := g(\hat{c}).$$

Then, by commutativity, it follows that

$$\phi(c) + \sum_{i,j} \hat{h}_{i,j} \geq 1. \tag{4-5}$$

Let $g_0, g_{1/2}, g_1 \in C(X)_+$ be a partition of unity such that g_j is identically 1 on $\{j\} \times [0,1]^{m-1}$ for $j = 0, 1$, g_0 is supported on $\left[0, \frac{1}{3}\right] \times [0,1]^{m-1}$, and g_1 is supported on $\left[\frac{2}{3}, 1\right] \times [0,1]^{m-1}$. Let us define

$$e_j := \rho(e'_0 \otimes e_j^{(0)} \otimes 1 + e'_1 \otimes 1 \otimes e_j^{(1)}) + g_j \rho(e'_{1/2} \otimes 1 \otimes 1) \tag{4-6}$$

for $j = 0, \frac{1}{2}, 1$. It is clear by their definitions that $e_0, e_{1/2}, e_1, \hat{h}_0, \ldots, \hat{h}_k, \phi(c)$ all commute.

Let us now check that $(e_0 + e_1)\hat{h}_{i,j} = \hat{h}_{i,j}$. Certainly

$$\begin{aligned}
(e_0 + e_1)\hat{h}_{0,j} &\overset{(4\text{-}6)}{=} \big(\rho(e'_0 \otimes (e_0^{(0)} + e_1^{(0)}) \otimes 1 + e'_1 \otimes 1 \otimes (e_0^{(1)} \otimes e_1^{(1)})) + (g_0 + g_1)\rho(e'_{1/2} \otimes 1 \otimes 1)\big)\rho(a_0 \otimes h_j^{(0)} \otimes 1) \\
&\underset{4.6(\text{ii,iv})}{\overset{\text{Lemma}}{=}} \rho(a_0 \otimes ((e_0^{(0)} + e_1^{(0)})h_j^{(0)}) \otimes 1) \\
&\overset{\text{Step 2(v)}}{=} \rho(a_0 \otimes h_j^{(0)} \otimes 1) = \hat{h}_{0,j},
\end{aligned}$$

and likewise, $(e_0 + e_1)\hat{h}_{1,j} = \hat{h}_{1,j}$ as required.

Since all terms in (4-5) commute, it is easy to see that for any $\epsilon > 0$, there exist orthogonal elements $\tilde{h}_{i,j} \leq \hat{h}_{i,j}$ which commute with $e_0, e_{1/2}, e_1$ and $\phi(c)$, such that

$$\phi(c) + \sum_{i,j} \tilde{h}_{i,j} =_\epsilon 1.$$

Then, by a diagonal sequence argument, it follows that there exist orthogonal elements $h_{i,j}$ with supports contained in those of $\hat{h}_{i,j}$, commuting with $e_0, e_{1/2}, e_1$ and $\phi(c)$, and such that

$$\phi(c) + \sum_{i,j} h_{i,j} = 1 \quad \text{and} \quad (e_0 + e_1)h_{i,j} = h_{i,j}.$$

Hence (v) holds.

Now let us verify (i)–(iv).

(i) holds using the following orthogonalities:

$$\begin{aligned} e_0^{(i)} &\perp e_1^{(i)}, \qquad i = 0, 1, \\ g_0 &\perp g_1, \\ e_0' &\perp e_1', \\ \rho(1 \otimes e_j^{(0)} \otimes 1) &\perp g_{1-j}, \quad j = 0, 1, \\ \rho(1 \otimes 1 \otimes e_j^{(1)}) &\perp g_{1-j}, \quad j = 0, 1. \end{aligned}$$

(ii): We compute

$$\begin{aligned} \epsilon_0 + e_{1/2} + e_1 &\overset{(4\text{-}6)}{=} \rho\big(e_0' \otimes (e_0^{(0)} + e_{1/2}^{(0)} + e_1^{(0)}) \otimes 1 + e_1' \otimes 1 \otimes (e_0^{(1)} + e_{1/2}^{(1)} + e_1^{(1)})\big) \\ &\qquad + (g_0 + g_{1/2} + g_1)\rho(e_{1/2}' \otimes 1 \otimes 1) \\ &\overset{\text{Step 2(ii)}}{=} \rho((e_0' + e_1' +) \otimes 1 \otimes 1) \overset{\text{Lemma 4.6(ii)}}{=} 1. \end{aligned}$$

(iii): For $x \in \{j\} \times [0,1]^{m-1}$,

$$e_j(x) \overset{\text{Step 2(iii)}}{=} e_0' + e_1' + g_j(x) e_{1/2}' \overset{\text{Lemma 4.6(ii)}}{=} 1.$$

(iv) follows from the fact that $\phi(\hat{c}) e_{1/2} = e_{1/2}\phi(\hat{c}) \geq e_{1/2}$, by considering irreducible representations of $\mathrm{C}^*(\phi(\hat{c}), e_{1/2})$. □

Proof of Theorem 4.1. By Proposition 2.4(i) and [Kirchberg and Winter 2004, Proposition 3.8], it suffices to verify the theorem for C*-algebras A of the form $C(X, \mathcal{H})$, where X is compact and Hausdorff. By [ibid., (3.5)], it suffices to prove it for $A = C(X)$. Again by Proposition 2.4(i), it suffices to assume that $C(X)$ is finitely generated. Finally, when $C(X)$ is finitely generated, it is a quotient of $C([0,1]^m)$ for some m, and so, by [ibid., (3.3)], the result reduces to showing that $\dim_{\mathrm{nuc}} C(X, \mathcal{Z}) \leq 2$ for $X = [0,1]^m$. By Proposition 2.6, we must show that the first factor embedding $C(X, \mathcal{Z}) \to C(X, \mathcal{Z}) \otimes \mathcal{Z}$ has decomposition rank at most 2.

We will do this in two steps. In Step 1, we will use Lemma 4.7 to show that the first factor embedding $\iota_0 : C(X) \to C(X) \otimes \mathcal{Z}$ has decomposition rank at most 2. In Step 2, we will use Step 1, with X replaced by $X \times [0,1]$, to prove the theorem.

Step 1. Due to Proposition 2.5, it suffices to replace ι_0 by its composition with the inclusion $C(X) \otimes \mathcal{Z} \subset (C(X) \otimes \mathcal{Z})_\infty$, that is, ι_0 is now

$$C(X) \cong C(X) \otimes 1_{\mathcal{Z}} \subset C(X) \otimes \mathcal{Z} \subset (C(X) \otimes \mathcal{Z})_\infty.$$

To show that $\dim_{\mathrm{nuc}} \iota_0 \leq 2$, we verify condition (iv) of Proposition 3.2. Let $\mathcal{U}$ be an open cover of X and let $\epsilon > 0$. By the Lebesgue covering lemma, we may possibly reduce ϵ so that $\mathcal{U}$ is refined by the set of all open sets of diameter at most ϵ. Then, it suffices to assume that $\mathcal{U}$ is in fact the set of all open sets of diameter at most ϵ.

Let $h_0, \dots, h_k, \phi, c$ be as in Lemma 4.7. By Theorem 4.2 and condition (iv) of Proposition 3.2, we may find

$$b_j^{(i)} \in C_0(X \times Z, \mathbb{O}_2) \cong C(X) \otimes C_0(Z) \otimes \mathbb{O}_2$$

for $i = 0, 1$, $j = 0, \dots, r$ such that

(i) for each $i = 0, 1$, the elements $b_0^{(i)}, \dots, b_r^{(i)}$ are pairwise orthogonal,

(ii) for each i, j, the support of $b_j^{(i)}$ is contained in $U \times Z$ for some $U \in \mathcal{U}$, and

(iii) $\left\| \sum_{i,j} b_j^{(i)} - 1_{C(X)} \otimes c \right\| < \epsilon$ (note that $c \in C_0(Z) \otimes 1_{\mathbb{O}_2}$).

Define

$$\hat{\phi} : C_0(X \times Z, \mathbb{O}_2) \cong C(X) \otimes C_0(Z, \mathbb{O}_2) \to C(X, \mathcal{Z})_\infty$$

by $\hat{\phi}(f \otimes a) = f\phi(a)$. This is a *-homomorphism. For $i = 0, 1$ and $j = 0, \dots, r$, set

$$a_j^{(i)} := \hat{\phi}(b_j^{(i)}),$$

and, for $j = 0, 1 \dots, k$, set

$$a_j^{(2)} := h_j.$$

Since $\hat{\phi}$ is a homomorphism, $a_0^{(i)}, \dots, a_r^{(i)}$ are pairwise orthogonal for $i = 0, 1$. Also, by the definition of $\hat{\phi}$ and the choice of $b_j^{(i)}$, the support of each $a_j^{(i)}$ is contained in some set in $\mathcal{U}$, for $i = 0, 1$. Since the supports of the h_j have diameter at most ϵ, the respective statement holds for the $a_j^{(2)}$ as well. Finally,

$$\sum_{i,j} a_j^{(i)} = \hat{\phi}\Big(\sum_{i=0,1} \sum_{j=0}^{k} b_j^{(i)} \Big) + \sum_{j=0}^{k} h_j =_\epsilon \hat{\phi}(1 \otimes c) + \sum_{j=0}^{k} h_j = \phi(c) + \sum_{j=0}^{k} h_j = 1,$$

as required.

Step 2. Since $\mathcal{Z}$ is an inductive limit of algebras of the form $\mathcal{Z}_{p,q}$ (for $p, q \in \mathbb{N}$), by Proposition 2.4(i), it suffices to show that the decomposition rank of the first factor embedding

$$\iota := \mathrm{id}_{C(X, \mathcal{Z}_{p,q})} \otimes 1_{\mathcal{Z}} : C(X, \mathcal{Z}_{p,q}) \to C(X, \mathcal{Z}_{p,q}) \otimes \mathcal{Z} \tag{4-7}$$

is at most 2. The proof will combine Step 1 with the idea of Proposition 3.2(iv) $\Rightarrow$ (iii).

For $t \in [0, 1]$, we let $\mathrm{ev}_t : \mathcal{Z}_{p,q} \to M_p \otimes M_q$ denote the point-evaluation at t, while we also let

$$\overline{\mathrm{ev}}_0 : \mathcal{Z}_{p,q} \to M_p \quad \text{and} \quad \overline{\mathrm{ev}}_1 : \mathcal{Z}_{p,q} \to M_q$$

denote the irreducible representations which satisfy

$$\mathrm{ev}_0(\cdot) = \overline{\mathrm{ev}}_0(\cdot) \otimes 1_{M_q} \quad \text{and} \quad \mathrm{ev}_1(\cdot) = 1_{M_p} \otimes \overline{\mathrm{ev}}_1(\cdot).$$

Let $\mathcal{F} \subset C(X, \mathcal{Z}_{p,q})$ be the finite set to approximate, and let $\epsilon > 0$ be the tolerance. Let us assume that $\mathcal{F}$ consists of contractions. Let $\mathcal{U}$ be an open cover of $X \times [0, 1]$, such that, for all $f \in \mathcal{F}$ and $U \in \mathcal{U}$, if

$(x,t),(x',t')\in U$, then

$$\|\mathrm{ev}_t(f(x))-\mathrm{ev}_{t'}(f(x'))\|<\epsilon/2.$$

Let us also assume that no $U\in\mathcal{U}$ intersects both $X\times\{0\}$ and $X\times\{1\}$.

Using Step 1 (with $X\times[0,1]$ in place of X) and Proposition 3.2(iv), we may find a 3-colourable $\epsilon/2$-approximate partition of unity

$$(a_j^{(i)})_{i=0,1,2;j=0,\dots,r}\subset C(X\times[0,1])\otimes\mathscr{Z}$$

subordinate to $\mathcal{U}$, and such that

$$\sum a_j^{(i)}\leq 1.$$

Upon replacing $\mathcal{U}$ by a subcover, we may clearly assume that $\mathcal{U}$ is of the form $(U_j^{(i)})_{i=0,1,2;j=0,\dots,r}$, with the support of each $a_j^{(i)}$ being contained in $U_j^{(i)}$.

For each i,j, we shall choose a matrix algebra $F_j^{(i)}$ and produce maps

$$C(X,\mathscr{Z}_{p,q})\xrightarrow{\psi_j^{(i)}}F_j^{(i)}\xrightarrow{\phi_j^{(i)}}C(X,\mathscr{Z}_{p,q})\otimes\mathscr{Z}.$$

We distinguish three cases, depending on properties of the set $U_j^{(i)}\in\mathcal{U}$. In every case, we arrange that

$$\phi_j^{(i)}\psi_j^{(i)}(f)=a_j^{(i)}\otimes\mathrm{ev}_{t_j^{(i)}}(f(x_j^{(i)})),$$

where $(x_j^{(i)},t_j^{(i)})$ is a point from $U_j^{(i)}$, and we make sense of the right-hand side by using the canonical identification of $C(X,\mathscr{Z}_{p,q})\otimes\mathscr{Z}$ with a subalgebra of

$$C(X\times[0,1])\otimes\mathscr{Z}\otimes M_p\otimes M_q$$

(determined by boundary conditions at $X\times\{0\}$ and at $X\times\{1\}$).

Case 1. If $U_j^{(i)}\cap(X\times\{0\})\neq\varnothing$, let $(x_j^{(i)},t_j^{(i)}=0)$ be a point in this intersection. We set $F_j^{(i)}:=M_p$ and define

$$\psi_j^{(i)}(f):=\overline{\mathrm{ev}}_0(f(x_j^{(i)})),$$
$$\phi_j^{(i)}(T):=a_j^{(i)}\otimes T\otimes 1_{M_q}.$$

By assumption, $U_j^{(i)}\cap(X\times\{1\})=\varnothing$, so, for all $x\in X$,

$$\mathrm{ev}_1(\phi_j^{(i)}(T)(x))=0,$$

and therefore, the range of $\phi_j^{(i)}$ lies in $C(X,\mathscr{Z}_{p,q})\otimes\mathscr{Z}$.

Case 2. If $U_j^{(i)}\cap(X\times\{1\})\neq\varnothing$, as in Case 1, let $(x_j^{(i)},t_j^{(i)}=1)$ be a point in this intersection. We set $F_j^{(i)}:=M_q$ and define

$$\psi_j^{(i)}(f):=\overline{\mathrm{ev}}_1(f(x_j^{(i)}))\quad\text{and}\quad\phi_j^{(i)}(T):=a_j^{(i)}\otimes 1_{M_p}\otimes T.$$

Case 3. If $U_j^{(i)} \cap (X \times \{0\}) = \varnothing$ and $U_j^{(i)} \cap (X \times \{1\}) = \varnothing$, then let $(x_j^{(i)}, t_j^{(i)})$ be any point in $U_j^{(i)}$. We set $F_j^{(i)} := M_p \otimes M_q$ and define

$$\psi_j^{(i)}(f) := \mathrm{ev}_{t_j^{(i)}}(f(x_j^{(i)})) \quad \text{and} \quad \phi_j^{(i)}(T) := a_j^{(i)} \otimes T.$$

We now set $F := \bigoplus_{i,j} F_j^{(i)}$ and use $(\psi_j^{(i)})$ and $(\phi_j^{(i)})$ to define

$$C(X, \mathscr{Z}_{p,q}) \xrightarrow{\psi} F \xrightarrow{\phi} C(X, \mathscr{Z}_{p,q}) \otimes \mathscr{Z}.$$

We have that ψ is c.p.c. since all of its components are. Each $\phi_j^{(i)}$ is c.p. and order zero. For each i, j_1, j_2, if $j_1 \neq j_2$, the images of $\phi_{j_1}^{(i)}$ and $\phi_{j_2}^{(i)}$ are orthogonal. Thus, for each i,

$$\phi|_{\bigoplus_j F_j^{(i)}}$$

is order zero. Also, $\phi(1) = \sum a_j^{(i)} \leq 1$, so that ϕ is contractive.

Finally, let $f \in \mathscr{F}$ and let us check that $\phi\psi(f) =_\epsilon f$. As in the proof of Proposition 3.2(iv) $\Rightarrow$ (iii), we have for each i, j that if $x \in U_j^{(i)}$, then

$$\mathrm{ev}_{t_j^{(i)}}(f(x_j^{(i)})) =_{\epsilon/2} \mathrm{ev}_t(f(x)),$$

and therefore,

$$\mathrm{ev}_t(f(x)) - \frac{\epsilon}{2} \cdot 1_{M_p \otimes M_q} \leq \mathrm{ev}_{t_j^{(i)}}(f(x_j^{(i)})) \leq \mathrm{ev}_t(f(x)) + \frac{\epsilon}{2} \cdot 1_{M_p \otimes M_q}.$$

Since $a_j^{(i)}$ commutes with f, this gives

$$\begin{aligned} a_j^{(i)}(x,t)\Big(\mathrm{ev}_t(f(x)) - \frac{\epsilon}{2} \cdot 1_{M_p \otimes M_q}\Big) &\leq a_j^{(i)}(x,t)\mathrm{ev}_{t_j^{(i)}}(f(x_j^{(i)})) \\ &\leq a_j^{(i)}(x,t)\Big(\mathrm{ev}_t(f(x)) + \frac{\epsilon}{2} \cdot 1_{M_p \otimes M_q}\Big). \end{aligned}$$

Moreover, since $a_j^{(i)}$ vanishes outside of $U_j^{(i)}$, these inequalities continue to hold for all $x \in X$ and all $t \in [0,1]$.

Summing over i, j, we find that

$$\begin{aligned} \sum_{i,j} a_j^{(i)}(x,t)\Big(\mathrm{ev}_t(f(x)) - \frac{\epsilon}{2} \cdot 1_{M_p \otimes M_q}\Big) &\leq \sum_{i,j} a_j^{(i)}(x,t)\mathrm{ev}_{t_j^{(i)}}(f(x_j^{(i)})) \\ &\leq \sum_{i,j} a_j^{(i)}(x,t)\Big(\mathrm{ev}_t(f(x)) + \frac{\epsilon}{2} \cdot 1_{M_p \otimes M_q}\Big), \end{aligned}$$

and therefore

$$\mathrm{ev}_t(f(x)) =_{\epsilon/2} \sum_{i,j} a_j^{(i)}(x,t)\mathrm{ev}_t(f(x)) =_{\epsilon/2} \sum_{i,j} a_j^{(i)}(x,t)\mathrm{ev}_{t_j^{(i)}}(f(x_j^{(i)})) = \mathrm{ev}_t(\phi\psi(f)(x)).$$

Since this holds for all $x \in X, t \in [0,1]$, this means that $\|f - \phi\psi(f)\| < \epsilon$, as required. □

Remark 4.9. Here we describe how one can give a shorter proof that $A \otimes \mathcal{Z}$ has decomposition rank at most 5, for A as in Theorem 4.1. Since $A \otimes \mathcal{Z}$ is an inductive limit of $A \otimes \mathcal{Z}_{p^\infty,q^\infty}$, it suffices to show that the latter has decomposition rank at most 5. This algebra is a $C([0,1])$-algebra whose fibres are all of the form $A \otimes U$, where U is a self-absorbing UHF algebra. Hence, by [Carrión 2011, Lemma 3.1], $A \otimes \mathcal{Z}_{p^\infty,q^\infty}$ has decomposition rank at most $5 = (\dim[0,1]+1)(2+1)-1$ if we show that $A \otimes U$ has decomposition rank at most 2 for every infinite dimensional, self-absorbing UHF algebra.

As in the first paragraph of the proof above, it suffices to show that the first-factor embedding $C(X,U) \to C(X,U) \otimes U$ has decomposition rank at most 2, when $X = [0,1]^m$. Since U is a limit of finite dimensional C*-algebras, by Proposition 2.4(i), the decomposition rank of this first-factor embedding agrees with the decomposition rank of the first-factor embedding $\iota_0 : C(X) \to C(X) \otimes U$. Then following Step 1 of the above proof verbatim, except with U in place of $\mathcal{Z}$, shows that this ι_0 has decomposition rank at most 2; moreover, we only need to use the variation of Lemma 4.7 where $\mathcal{Z}$ is replaced by U, and, as explained in Remark 4.8, the proof of that lemma can be simplified in that case.

5. $\mathcal{Z}$-stable $C(X)$-algebras

The proof of [Carrión 2011, Lemma 3.1] actually shows the following.

Lemma 5.1. *Let X be a compact metric space, let A be a $C(X)$-algebra, and let B be a unital* C*-*algebra. Denote by $\iota_{C(X)} : C(X) \to C(X) \otimes B$ and $\iota_A : A \to A \otimes B$ the first-factor embeddings. Then*

$$\operatorname{dr} \iota_A \leq (\operatorname{dr} \iota_{C(X)} + 1)\big(\max_{x \in X} \operatorname{dr} A(x) + 1\big) - 1 \tag{5-1}$$

and

$$\dim_{\mathrm{nuc}} \iota_A \leq (\dim_{\mathrm{nuc}} \iota_{C(X)} + 1)\big(\max_{x \in X} \dim_{\mathrm{nuc}} A(x) + 1\big) - 1. \tag{5-2}$$

Proof. Although this is essentially the same as the proof of [Carrión 2011, Lemma 3.1], we provide a detailed proof of (5-1) for the reader's convenience.

Set $k := \max_{x \in X} \operatorname{dr} A(x)$ and $l := \operatorname{dr} \iota_{C(X)}$. Let $\mathcal{F} \subset A$ be a finite subset and let $\epsilon > 0$. Without loss of generality, $\mathcal{F}$ consists of self-adjoint contractions. As shown in the proof of [Carrión 2011, Lemma 3.1], there exists an open cover $\mathcal{U}$ of X such that, for each $U \in \mathcal{U}$, there exists a finite dimensional C*-algebra F_U and c.p.c. maps $\psi_U : A \to F_U$, $\phi_U : F_U \to A$ such that ϕ_U is $(k+1)$-colourable and $\phi_U \psi_U(a)(x) =_{\epsilon/2} a(x)$ for all $a \in \mathcal{F}$ and all $x \in U$. By Proposition 3.2(iv), let $(b_j^{(i)})_{j=1,\dots,r;\, i=0,\dots,l} \subset C(X) \otimes B$ be an $(l+1)$-colourable, $\epsilon/2$-approximate partition of 1, subordinate to $\mathcal{U}$, and, by a rescaling argument, we may assume $b_j^{(i)} \leq 1$ for each i, j. Hence, for each i, j, we may pick some $U_j^{(i)} \in \mathcal{U}$ containing the support of $b_j^{(i)}$. Define

$$\psi := \bigoplus_{i,j} \psi_{U_j^{(i)}} : A \to \bigoplus_{i,j} F_{U_j^{(i)}}$$

and $\phi : \bigoplus_{i,j} F_{U_j^{(i)}} \to A \otimes B$ by

$$\phi((a_j^{(i)})) := \sum_{i,j} \phi_{U_j^{(i)}}(a_j^{(i)}) \otimes b_j^{(i)}.$$

One readily verifies that ϕ is $(k+1)(l+1)$-colourable, and, as in the proof of Proposition 3.2(iv) $\Rightarrow$ (iii), that $\phi\psi(a) =_\epsilon a \otimes 1_B$ for all $a \in \mathcal{F}$. □

Corollary 5.2. *If A is a $\mathcal{Z}$-stable $C(X)$-algebra whose fibres have decomposition rank (respectively nuclear dimension) bounded by M, the decomposition rank (respectively nuclear dimension) of A is at most $3(M+1)-1$.*

Proof. We shall apply Lemma 5.1 with $\mathcal{Z}$ in place of B. Using the notation of Lemma 5.1, Theorem 4.1 tells us that $\dim_{\mathrm{nuc}} \iota_{C(X)} \leq \mathrm{dr}\, \iota_{C(X)} \leq 2$. Thus, if the fibres of A have decomposition rank at most M, then, by Lemma 5.1, $\mathrm{dr}\, \iota_A \leq (2+1)(M+1)-1$. □

This shows in particular that the $C(X)$-algebra in [Hirshberg et al. 2007, Example 4.7] (which is $\mathcal{Z}$-stable by [Dadarlat and Toms 2009]) has decomposition rank at most 2, and that the $C(X)$-algebra E in [Dadarlat 2009b, Section 3] (which is $\mathcal{Z}$-stable since it is an extension of patently $\mathcal{Z}$-stable C*-algebras) has nuclear dimension at most 5. On the other hand, the $C(X)$-algebra in [Hirshberg et al. 2007, Example 4.8] is not $\mathcal{Z}$-stable, and it is shown in [Robert and Tikuisis 2013, Section 7.4] that it does not have finite nuclear dimension.

Another immediate application is the following strengthening of Theorem 4.1. See [Dadarlat and Pennig 2013] for a discussion of $C(X)$-algebras with fibres $\mathcal{D} \otimes \mathcal{K}$, where $\mathcal{D}$ is either $\mathcal{Z}$ or an infinite dimensional UHF algebra.

Corollary 5.3. *If A is a $\mathcal{Z}$-stable $C(X)$-algebra whose fibres are all AF algebras tensored by $\mathcal{Z}$, then* $\mathrm{dr}\, A \leq 2$.

Proof. It suffices to show that $B := A \otimes \mathcal{Z}_{p^\infty,q^\infty}$ has decomposition rank at most 2. Note that B is a $\mathcal{Z}$-stable $C(X \times [0,1])$-algebra with AF fibres. Therefore, by Corollary 5.2, $\mathrm{dr}\, B$ is at most $3(0+1)-1=2$, as required. □

References

[Barlak et al. 2014] S. Barlak, D. Enders, H. Matui, G. Szabó, and W. Winter, "The Rokhlin property vs. Rokhlin dimension 1 on $\mathbb{O}_2$", preprint, 2014. arXiv 1312.6289v2

[Carrión 2011] J. R. Carrión, "Classification of a class of crossed product C^*-algebras associated with residually finite groups", *J. Funct. Anal.* **260**:9 (2011), 2815–2825. MR 2012e:46146 Zbl 1220.46042

[Christensen et al. 2012] E. Christensen, A. M. Sinclair, R. R. Smith, S. A. White, and W. Winter, "Perturbations of nuclear C^*-algebras", *Acta Math.* **208**:1 (2012), 93–150. MR 2910797 Zbl 1252.46047

[Connes 1976] A. Connes, "Classification of injective factors: cases II_1, II_∞, III_λ, $\lambda \neq 1$", *Ann. of Math.* (2) **104**:1 (1976), 73–115. MR 56 #12908 Zbl 0343.46042

[Cuntz 1981] J. Cuntz, "K-theory for certain C^*-algebras", *Ann. of Math.* (2) **113**:1 (1981), 181–197. MR 84c:46058 Zbl 0475.46051

[Dadarlat 2009a] M. Dadarlat, "Continuous fields of C^*-algebras over finite dimensional spaces", *Adv. Math.* **222**:5 (2009), 1850–1881. MR 2010j:46102 Zbl 1190.46040

[Dadarlat 2009b] M. Dadarlat, "Fiberwise KK-equivalence of continuous fields of C^*-algebras", *J. K-Theory* **3**:2 (2009), 205–219. MR 2010j:46122 Zbl 1173.46050

[Dadarlat and Pennig 2013] M. Dadarlat and U. Pennig, "A Dixmier–Douady theory for strongly self-absorbing C*-algebras", preprint, 2013. To appear in *J. Reine Angew. Math.* arXiv 1302.4468

[Dadarlat and Toms 2009] M. Dadarlat and A. S. Toms, "$\mathcal{Z}$-stability and infinite tensor powers of C^*-algebras", *Adv. Math.* **220**:2 (2009), 341–366. MR 2010c:46132 Zbl 1160.46039

[Elliott 1995] G. A. Elliott, "The classification problem for amenable C^*-algebras", pp. 922–932 in *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), vol. 2, edited by S. Chatterji, Birkhäuser, Basel, 1995. MR 97g:46072 Zbl 0946.46050

[Elliott 1996] G. A. Elliott, "An invariant for simple C^*-algebras", pp. 61–90 in *Canadian Mathematical Society, 1945–1995*, vol. 3, edited by J. B. Carrell and R. Murty, Canadian Math. Soc., Ottawa, ON, 1996. MR 2000b:46095 Zbl 1206.46046

[Elliott and Toms 2008] G. A. Elliott and A. S. Toms, "Regularity properties in the classification program for separable amenable C^*-algebras", *Bull. Amer. Math. Soc.* (*N.S.*) **45**:2 (2008), 229–245. MR 2009k:46111 Zbl 1151.46048

[Elliott et al. 2007] G. A. Elliott, G. Gong, and L. Li, "On the classification of simple inductive limit C^*-algebras, II: The isomorphism theorem", *Invent. Math.* **168**:2 (2007), 249–320. MR 2010g:46102 Zbl 1129.46051

[Gong 2002] G. Gong, "On the classification of simple inductive limit C^*-algebras, I: The reduction theorem", *Doc. Math.* **7** (2002), 255–461. MR 2007h:46069 Zbl 1024.46018

[Hirshberg et al. 2007] I. Hirshberg, M. Rørdam, and W. Winter, "$\mathcal{C}_0(X)$-algebras, stability and strongly self-absorbing C^*-algebras", *Math. Ann.* **339**:3 (2007), 695–732. MR 2008j:46040 Zbl 1128.46020

[Hirshberg et al. 2012a] I. Hirshberg, E. Kirchberg, and S. White, "Decomposable approximations of nuclear C^*-algebras", *Adv. Math.* **230**:3 (2012), 1029–1039. MR 2921170 Zbl 1256.46019

[Hirshberg et al. 2012b] I. Hirshberg, W. Winter, and J. Zacharias, "Rokhlin dimension and C*-dynamics", preprint, 2012. arXiv 1209.1618

[Jiang and Su 1999] X. Jiang and H. Su, "On a simple unital projectionless C^*-algebra", *Amer. J. Math.* **121**:2 (1999), 359–413. MR 2000a:46104 Zbl 0923.46069

[Kasparov 1988] G. G. Kasparov, "Equivariant KK-theory and the Novikov conjecture", *Invent. Math.* **91**:1 (1988), 147–201. MR 88j:58123 Zbl 0647.46053

[Kirchberg 1993] E. Kirchberg, "On nonsemisplit extensions, tensor products and exactness of group C^*-algebras", *Invent. Math.* **112**:3 (1993), 449–489. MR 94d:46058 Zbl 0803.46071

[Kirchberg 1995] E. Kirchberg, "Exact C*-algebras, tensor products, and the classification of purely infinite algebras", pp. 943–954 in *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), vol. 2, edited by S. Chatterji, Birkhäuser, Basel, 1995. MR 97g:46074 Zbl 0897.46057

[Kirchberg 2006] E. Kirchberg, "Central sequences in C^*-algebras and strongly purely infinite algebras", pp. 175–231 in *Operator Algebras: The Abel Symposium 2004*, edited by O. Brattelli et al., Abel Symp. **1**, Springer, Berlin, 2006. MR 2009c:46075 Zbl 1118.46054

[Kirchberg and Phillips 2000] E. Kirchberg and N. C. Phillips, "Embedding of exact C^*-algebras in the Cuntz algebra $\mathbb{O}_2$", *J. Reine Angew. Math.* **525** (2000), 17–53. MR 2001d:46086a Zbl 0973.46047

[Kirchberg and Rørdam 2002] E. Kirchberg and M. Rørdam, "Infinite non-simple C^*-algebras: absorbing the Cuntz algebras $\mathbb{O}_\infty$", *Adv. Math.* **167**:2 (2002), 195–264. MR 2003k:46080 Zbl 1030.46075

[Kirchberg and Rørdam 2005] E. Kirchberg and M. Rørdam, "Purely infinite C^*-algebras: ideal-preserving zero homotopies", *Geom. Funct. Anal.* **15**:2 (2005), 377–415. MR 2006d:46070 Zbl 1092.46044

[Kirchberg and Winter 2004] E. Kirchberg and W. Winter, "Covering dimension and quasidiagonality", *Internat. J. Math.* **15**:1 (2004), 63–85. MR 2005a:46148 Zbl 1065.53057

[Lin 2004] H. Lin, "Classification of simple C^*-algebras of tracial topological rank zero", *Duke Math. J.* **125**:1 (2004), 91–119. MR 2005i:46064 Zbl 1068.46032

[Lin 2011a] H. Lin, "Asymptotic unitary equivalence and classification of simple amenable C^*-algebras", *Invent. Math.* **183**:2 (2011), 385–450. MR 2012c:46157 Zbl 1255.46031

[Lin 2011b] H. Lin, "On locally AH algebras", preprint, 2011. To appear in *Mem. Amer. Math. Soc.* arXiv 1104.0445

[Lin and Niu 2008] H. Lin and Z. Niu, "Lifting KK-elements, asymptotic unitary equivalence and classification of simple C^*-algebras", *Adv. Math.* **219**:5 (2008), 1729–1769. MR 2009g:46118 Zbl 1162.46033

[Matui and Sato 2012] H. Matui and Y. Sato, "Strict comparison and $\mathcal{Z}$-absorption of nuclear C^*-algebras", *Acta Math.* **209**:1 (2012), 179–196. MR 2979512

[Matui and Sato 2013] H. Matui and Y. Sato, "Decomposition rank of UHF-absorbing C*-algebras", preprint, 2013. To appear in *Duke Math. J.* arXiv 1303.4371

[Rieffel 1983] M. A. Rieffel, "Dimension and stable rank in the K-theory of C^*-algebras", *Proc. London Math. Soc.* (3) **46**:2 (1983), 301–333. MR 84g:46085 Zbl 0533.46046

[Robert and Tikuisis 2013] L. Robert and A. Tikuisis, "Nuclear dimension and $\mathcal{Z}$-stability of non-simple C*-algebras", preprint, 2013. arXiv 1308.2941

[Rørdam 2002] M. Rørdam, "Classification of nuclear, simple C^*-algebras", pp. 1–145 in *Classification of nuclear C^*-algebras. Entropy in operator algebras*, Encyclopaedia Math. Sci. **126**, Springer, Berlin, 2002. MR 2003i:46060 Zbl 1016.46037

[Rørdam 2003] M. Rørdam, "A simple C^*-algebra with a finite and an infinite projection", *Acta Math.* **191**:1 (2003), 109–142. MR 2005m:46096 Zbl 1072.46036

[Rørdam 2004] M. Rørdam, "The stable and the real rank of $\mathcal{Z}$-absorbing C^*-algebras", *Internat. J. Math.* **15**:10 (2004), 1065–1084. MR 2005k:46164 Zbl 1077.46054

[Rørdam 2006] M. Rørdam, "Structure and classification of C^*-algebras", pp. 1581–1598 in *International Congress of Mathematicians*, vol. II, edited by M. Sanz-Solé et al., Eur. Math. Soc., Zürich, 2006. MR 2008e:46070 Zbl 1104.46033

[Rørdam and Winter 2010] M. Rørdam and W. Winter, "The Jiang–Su algebra revisited", *J. Reine Angew. Math.* **642** (2010), 129–155. MR 2011i:46074 Zbl 1209.46031

[Sato et al. 2014] Y. Sato, S. White, and W. Winter, "Nuclear dimension and $\mathcal{Z}$-stability", preprint, 2014. arXiv 1403.0747

[Szabo 2013] G. Szabo, "The Rokhlin dimension of topological $\mathbb{Z}^m$-actions", preprint, 2013. arXiv 1308.5418

[Tikuisis 2014] A. Tikuisis, "High-dimensional $\mathcal{Z}$-stable AH algebras", preprint, 2014. arXiv 1406.0883

[Toms 2008] A. S. Toms, "On the classification problem for nuclear C^*-algebras", *Ann. of Math.* (2) **167**:3 (2008), 1029–1044. MR 2009g:46119 Zbl 1181.46047

[Toms 2011] A. Toms, "K-theoretic rigidity and slow dimension growth", *Invent. Math.* **183**:2 (2011), 225–244. MR 2012h:46093 Zbl 1237.19009

[Toms and Winter 2007] A. S. Toms and W. Winter, "Strongly self-absorbing C^*-algebras", *Trans. Amer. Math. Soc.* **359**:8 (2007), 3999–4029. MR 2008c:46086 Zbl 1120.46046

[Toms and Winter 2009] A. S. Toms and W. Winter, "Minimal dynamics and the classification of C^*-algebras", *Proc. Natl. Acad. Sci. USA* **106**:40 (2009), 16942–16943. MR 2011d:46139 Zbl 1203.46046

[Toms and Winter 2013] A. S. Toms and W. Winter, "Minimal dynamics and K-theoretic rigidity: Elliott's conjecture", *Geom. Funct. Anal.* **23**:1 (2013), 467–481. MR 3037905 Zbl 06183911

[Villadsen 1998] J. Villadsen, "The range of the Elliott invariant of the simple AH-algebras with slow dimension growth", *K-Theory* **15**:1 (1998), 1–12. MR 99m:46143 Zbl 0916.19003

[Villadsen 1999] J. Villadsen, "On the stable rank of simple C^*-algebras", *J. Amer. Math. Soc.* **12**:4 (1999), 1091–1102. MR 2000f:46075 Zbl 0937.46052

[Voiculescu 1991] D. Voiculescu, "A note on quasi-diagonal C^*-algebras and homotopy", *Duke Math. J.* **62**:2 (1991), 267–271. MR 92c:46062 Zbl 0833.46055

[Winter 2003] W. Winter, "Covering dimension for nuclear C^*-algebras", *J. Funct. Anal.* **199**:2 (2003), 535–556. MR 2004c:46134 Zbl 1026.46049

[Winter 2004] W. Winter, "Decomposition rank of subhomogeneous C^*-algebras", *Proc. London Math. Soc.* (3) **89**:2 (2004), 427–456. MR 2005d:46121 Zbl 1081.46049

[Winter 2007] W. Winter, "Localizing the Elliott conjecture at strongly self-absorbing C*-algebras", preprint, 2007. To appear in *J. Reine Angew. Math.* arXiv 0708.0283

[Winter 2010] W. Winter, "Decomposition rank and $\mathcal{Z}$-stability", *Invent. Math.* **179**:2 (2010), 229–301. MR 2011a:46092 Zbl 1194.46104

[Winter 2012] W. Winter, "Nuclear dimension and $\mathcal{Z}$-stability of pure C^*-algebras", *Invent. Math.* **187**:2 (2012), 259–342. MR 2885621 Zbl 06010393

[Winter and Zacharias 2010] W. Winter and J. Zacharias, "The nuclear dimension of C^*-algebras", *Adv. Math.* **224**:2 (2010), 461–498. MR 2011e:46095 Zbl 1201.46056

Received 30 Apr 2013. Revised 5 Sep 2013. Accepted 4 Oct 2013.

AARON TIKUISIS: a.tikuisis@abdn.ac.uk
Institute of Mathematics, University of Aberdeen, Fraser Noble Building, Aberdeen, AB24 3UE, United Kingdom

WILHELM WINTER: wwinter@uni-muenster.de
Mathematisches Institut, Universität Münster, Einsteinstraße 62, D-48149 Münster, Germany

ANALYSIS AND PDE
Vol. 7, No. 3, 2014
dx.doi.org/10.2140/apde.2014.7.701

SCATTERING FOR A MASSLESS CRITICAL NONLINEAR WAVE EQUATION IN TWO SPACE DIMENSIONS

MARTIN SACK

We prove scattering for a massless wave equation which is critical in two space dimensions. Our method combines conformal inversion with decay estimates from Struwe's previous work on global existence of a similar equation.

1. Introduction

We study the asymptotic behavior of solutions to the nonlinear wave equation

$$u_{tt} - \Delta u + u(e^{u^2} - 1 - u^2) = 0 \quad \text{on } \mathbb{R} \times \mathbb{R}^2, \tag{1}$$

with compactly supported initial data

$$(u, u_t)|_{t=0} = (u_0, u_1) \in C_c^\infty(\mathbb{R}^2) \times C_c^\infty(\mathbb{R}^2). \tag{2}$$

Their initial energy is given by

$$E_0 = \frac{1}{2} \int_{\mathbb{R}^2} \left(u_1^2 + |\nabla u_0|^2 + e^{u_0^2} - 1 - u_0^2 - \tfrac{1}{2}u_0^4\right) dx. \tag{3}$$

Interest in this equation arises because it lies at the boundary of what one considers an energy-critical equation. For the defocusing nonlinear wave equation with power nonlinearity in dimension $d \geq 3$,

$$u_{tt} - \Delta u + |u|^{p-2}u = 0 \quad \text{on } \mathbb{R} \times \mathbb{R}^d,$$

this border is marked by the Sobolev-critical power $p^* = 2d/(d-2)$. In the subcritical case $p < p^*$ as well as in the critical case $p = p^*$ well-posedness in the energy space is known to hold. However, little is known for the supercritical case $p > p^*$. In two space dimensions the embedding $H^1(\mathbb{R}^2) \subset L^p(\mathbb{R}^2)$ for $p < \infty$ renders every power nonlinearity subcritical. However, $H^1(\mathbb{R}^2) \not\subseteq L^\infty(\mathbb{R}^2)$. Instead, we have the Trudinger–Moser inequality

$$\sup_{\substack{u \in H_0^1(\Omega) \\ \|\nabla u\|_{L^2(\mathbb{R}^2)} \leq 1}} \int_\Omega e^{\alpha u^2} dx \begin{cases} \leq C|\Omega| & \text{if } \alpha \leq 4\pi, \\ = \infty & \text{if } \alpha > 4\pi, \end{cases} \tag{4}$$

This work was supported by SNF project 200021_140467 / 1.
MSC2010: primary 35L71; secondary 35B40.
Keywords: nonlinear wave equation, energy critical, scattering theory.

for a smooth bounded domain $\Omega \subset \mathbb{R}^2$. Since

$$\sup_{\substack{u\in H_0^1(\Omega)\\ \|\nabla u\|^2_{L^2(\mathbb{R}^2)}=1}} \int_\Omega e^{\alpha u^2}\,dx = \sup_{\substack{u\in H_0^1(\Omega)\\ \|\nabla u\|^2_{L^2(\mathbb{R}^2)}=\alpha}} \int_\Omega e^{u^2}\,dx,$$

it seems that well-posedness, for instance of the initial value problem for the equation

$$u_{tt} - \Delta u + u e^{u^2} = 0 \quad \text{on } \mathbb{R}\times\mathbb{R}^2, \tag{5}$$

may depend on the size of the initial energy

$$E := \frac{1}{2}\int_{\mathbb{R}^2} (u_1^2 + |\nabla u_0|^2 + e^{u_0^2} - 1)\,dx,$$

(or, in the case of (1), on the size of E_0).

For small data, global well-posedness for (5) was shown in [Nakamura and Ozawa 1999]. Ibrahim, Majdoub, and Masmoudi [Ibrahim et al. 2006] proved global existence for data with energy $E \le 2\pi$, which they define to be (sub)critical. Due to the dispersive nature of (5), they also expected u to decay in time and to scatter towards a solution of the linear Klein–Gordon equation

$$u_{tt} - \Delta u + u = 0. \tag{6}$$

Indeed, together with Nakanishi [Ibrahim et al. 2009], they established scattering for the modified equation

$$u_{tt} - \Delta u + u(e^{u^2} - u^2) = 0 \quad \text{on } \mathbb{R}\times\mathbb{R}^2, \tag{7}$$

as long as

$$E_1 = \frac{1}{2}\int_{\mathbb{R}^2} \left(u_1^2 + |\nabla u_0|^2 + e^{u_0^2} - 1 - \tfrac{1}{2}u_0^4\right) dx \le 2\pi,$$

leaving open the corresponding questions in the *supercritical* regime when $E > 2\pi$ or $E_1 > 2\pi$. Note that we reserve the notation E for the context of (5), while E_0 and E_1 refer to equations (1) and (7), respectively.

Surprisingly, Struwe [2013] was able to establish global existence for (5) for arbitrary smooth initial data using only energy estimates.

Here we show that for scattering, too, there is no restriction on the energy, at least when we consider the massless wave equation (1) for radially symmetric initial data. As a consequence of the next result, we consider (1), (5), and (7) to be critical problems only, regardless of the size of the initial energy.

Theorem 1.1. *For any solution u to the Cauchy problem (1), (2) with smooth compactly supported radial data (u_0, u_1), $u_0(x) = u_0(|x|)$, $u_1(x) = u_1(|x|)$, there exists $(v_0, v_1) \in \dot{H}^1(\mathbb{R}^2)\times L^2(\mathbb{R}^2)$ such that*

$$\|(u(t) - v(t), \partial_t u(t) - \partial_t v(t))\|_{\dot{H}^1(\mathbb{R}^2)\times L^2(\mathbb{R}^2)} \to 0 \quad \text{as } t\to\infty, \tag{8}$$

where v is the solution to the linear wave equation

$$v_{tt} - \Delta v = 0 \tag{9}$$

with Cauchy data $(v, v_t)|_{t=0} = (v_0, v_1)$.

We assume smooth data. We remark, however, that to our knowledge Struwe's result has not been extended to data in energy space.

To prepare for the proof of Theorem 1.1, we rewrite (1) abstractly as

$$u_{tt} - \Delta u + N = 0, \tag{10}$$

with the nonlinearity

$$N(u) := (e^{u^2} - 1 - u^2)u.$$

The solution to (10) is given by the Duhamel formula

$$u(t) = \partial_t R(t) * u_0 + R(t) * u_1 + \int_0^t R(t-s) * N(u(s))\, ds \tag{11}$$

with R the fundamental solution to (9). In Fourier space it reads

$$\mathcal{F}(R(t))(\xi) = \frac{\sin(|\xi|t)}{|\xi|}.$$

From the Duhamel formula (11), we read off how the initial data are propagated. We define

$$v_0 := \mathcal{F}^{-1}\left(\hat{u}_0 - \int_0^\infty \frac{\sin(|\xi|s)}{|\xi|}\widehat{N}(s)\, ds\right), \quad v_1 := \mathcal{F}^{-1}\left(\hat{u}_1 + \int_0^\infty \cos(|\xi|s)\widehat{N}(s)\, ds\right)$$

as initial data for the linear wave equation, which we understand in the trace sense by energy control (compare [Lions and Magenes 1970]). We call v the solution to the corresponding Cauchy problem. Using the Duhamel formula (11), one calculates

$$\|u(t) - v(t)\|_{\dot{H}^1(\mathbb{R}^2)} = \left\| \int_t^\infty \frac{\sin(|\xi|(t-s))}{|\xi|}\widehat{N}(s)\, ds \right\|_{\dot{H}^1(\mathbb{R}^2)}, \tag{12}$$

and a corresponding expression for the time derivative. To prove scattering, we need to establish convergence of the integrals defining the initial data (v_0, v_1) in the norm $\dot{H}^1 \times L^2$. In the following lemma we reduce this problem to a bound on the nonlinearity N.

Lemma 1.2. *If*

$$\|N\|_{L^1([0,\infty);L^2(\mathbb{R}^2))} < \infty,$$

the integrals

$$\int_0^\infty \frac{\sin(|\xi|s)}{|\xi|}\widehat{N}(s)\, ds, \quad \int_0^\infty \cos(|\xi|s)\widehat{N}(s)\, ds$$

converge in $\dot{H}^1 \times L^2$.

The lemma follows from the equivalences

$$\|u\|_{\dot{H}^1} \simeq \big\||\xi|\hat{u}\big\|_{L^2}, \quad \|v\|_{L^2} \simeq \|\hat{v}\|_{L^2}.$$

Thus, once $N \in L^1_t L^2_x$ is established, the assertion of Theorem 1.1 follows from (12).

In the case of the nonlinear Klein–Gordon equation, we find similar representation formulæ and analogous results with the fundamental solution replaced by

$$\mathcal{F}(R(t))(\xi) = \frac{\sin(\langle\xi\rangle t)}{\langle\xi\rangle},$$

where $\langle\xi\rangle = \sqrt{1+|\xi|^2}$. Then scattering takes place in the norm $H^1 \times L^2$.

This discussion highlights the significance of leaving out the cubic term in (1). Informally, to ensure that $N(u) = u(e^{u^2}-1)$ lies in $L_t^1 L_x^2$ we need to control $\|u\|_{L_t^3 L_x^6}$. However, $L_t^3 L_x^6$ is not an admissible Strichartz norm in two space dimensions. In this respect, we agree with [Ibrahim et al. 2009]. In the course of our argument we will encounter further reasons that justify omission of the cubic term.

Moreover, for large data we restrict our result to the massless equation (1). The reason is that the method of conformal inversion that we employ in Section 3 to control the nonlinearity in this case only seems to work for the massless equation. It is not clear whether a similar control can also be achieved when working in the original coordinates. However, even then, the contribution to the energy from the mass term might spoil the validity of an estimate like Lemma 3.1.

Our work is organized as follows. In Section 2 we derive estimates for the nonlinear term. As a by-product we obtain a scattering result for the massive equation (7) for small data, where we only use standard $L_t^p L_x^q$ Strichartz estimates, instead of the more elaborate estimates for Besov spaces used in [Nakamura and Ozawa 1999; Ibrahim et al. 2009].

In Section 3 we prove Theorem 1.1 for large radially symmetric data. In a first step, by applying the method of conformal inversion as in [Grillakis 1990] and adapting the decay estimates from [Struwe 2013], we find a hyperboloid contained inside the support of the solution u such that $\|N\|_{L_t^1 L_x^2}$ is bounded inside the hyperboloid. For this part of the argument, we need not assume the initial data to be radial. In the final step, we use the radial symmetry of the data to bound $\|N\|_{L_t^1 L_x^2}$ in the complement of the hyperboloid.

2. Scattering for small data

For small data, scattering for (7) was first shown in [Nakamura and Ozawa 1999]. In [Ibrahim et al. 2009], the authors extend the result to include initial data with energy $E_1 \le 2\pi$. Both these works rely on Besov space techniques. In this section, we show scattering for small data via a more direct approach. We assume $u_0, u_1 \in C_c^\infty(\mathbb{R}^2)$ with E_1 bounded by an absolute constant ε_0 to be determined later.

The modulus of the nonlinearity $|N| = (e^{u^2}-u^2-1)|u|$ behaves like $|u|^5$ for small values of $|u|$. For large values of $|u|$ the exponential dominates. More precisely, we have the pointwise estimate

$$|N| = |(e^{u^2}-u^2-1)u| = |u|^3 \sum_{k=1}^{\infty} \frac{u^{2k}}{(k+1)!} \le |u|^3(e^{u^2}-1) \le \begin{cases} |u|^{40/9}(e^{u^2}-1) & \text{if } |u| \ge 1, \\ e|u|^5 & \text{if } 0 \le |u| < 1. \end{cases} \tag{13}$$

By Hölder's inequality,

$$\|u^{40/9}(e^{u^2}-1)\|_{L_t^1 L_x^2} \le \|u\|^{40/9}_{L_t^{40/9} L_x^{20}} \|e^{u^2}-1\|_{L_t^\infty L_x^{18/5}}.$$

To control the norm of the exponential term we roughly estimate

$$(e^{u^2}-1)^{\frac{18}{5}} \le e^{\frac{18}{5}u^2}-1 \le e^{4\pi u^2}-1.$$

Then we can use a version of the Trudinger–Moser inequality [Ruf 2005]:

$$\sup_{\|u\|_{L^2}+\|\nabla u\|_{L^2}\le 1} \int_\Omega (e^{4\pi u^2}-1)\,dx \le C_{\mathrm{TM}} \tag{14}$$

with a constant C_{TM} independent of the region $\Omega \subset \mathbb{R}^2$. Because of the finite speed of propagation, the support of u stays bounded locally uniformly in time. Since the energy is nonincreasing in time, if $\varepsilon_0 \le \frac{1}{2}$, the condition $\|u\|_{L^2}+\|\nabla u\|_{L^2}\le 1$ is satisfied for all times. Therefore we may combine (14) with (13) to obtain

$$\|N\|_{L_t^1([0,T);L_x^2(\mathbb{R}^2))} \le C_{\mathrm{TM}}\|u\|^{40/9}_{L_t^{40/9}([0,T);L_x^{20}(\mathbb{R}^2))} + e\|u\|^5_{L_t^5([0,T);L_x^{10}(\mathbb{R}^2))}. \tag{15}$$

We have chosen the power $\frac{40}{9}$ for convenience. However, we are not free in our choice, as we want to estimate u in $L_t^qL_x^r$ with Strichartz estimates. Wave admissibility [Keel and Tao 1998] demands that

$$\frac{1}{q}+\frac{1}{2r} \le \frac{1}{4},$$

so we need $q \ge 4$. By Strichartz estimates (as in [Nakanishi and Schlag 2011, Corollary 2.41, Lemma 2.43]),

$$\begin{aligned} f(T) &:= \|u\|_{L_t^{40/9}([0,T);L_x^{20}(\mathbb{R}^2))} + \|u\|_{L_t^5([0,T);L_x^{10}(\mathbb{R}^2))} \\ &\le C_S\big(\|(u_0,u_1)\|_{H^1(\mathbb{R}^2)\times L^2(\mathbb{R}^2)} + \|N\|_{L_t^1([0,T);L_x^2(\mathbb{R}^2))}\big) \end{aligned} \tag{16}$$

with a constant C_S that does not depend on the initial data. Then, by (15) and (16), we have

$$f(T) \le C_S\big(\|(u_0,u_1)\|_{H^1(\mathbb{R}^2)\times L^2(\mathbb{R}^2)} + C_{\mathrm{TM}} f(T)^{40/9} + e f(T)^5\big).$$

The function $f(T)$ is continuous and nondecreasing with $f(0)=0$. Therefore there exists a time $T_0>0$ such that $f(T)\le 1$ for $0\le T<T_0$ and

$$f(T) \le C_S\big(\|(u_0,u_1)\|_{H^1(\mathbb{R}^2)\times L^2(\mathbb{R}^2)} + (e+C_{\mathrm{TM}}) f(T)^{40/9}\big) \tag{17}$$

for all times $T\in[0,T_0)$. Let $A=\min\{1,A_0\}$, where A_0 satisfies

$$C_S(e+C_{\mathrm{TM}})(2A_0)^{40/9} = \tfrac{1}{2}A_0.$$

Suppose $\|(u_0,u_1)\|_{H^1(\mathbb{R}^2)\times L^2(\mathbb{R}^2)} < \varepsilon_0$, where

$$C_S\varepsilon_0 = \tfrac{1}{2}A.$$

Then relation (17) implies $f(T)\le A$ as long as $f(T)\le 2A$. Hence, by continuity, $f(T_0)\le A$. By the definition of A and continuity again, T_0 can be arbitrarily extended and the bound $f(T)\le A$ holds for all times. By (15) we have

$$\|N\|_{L_t^1([0,\infty);L_x^2(\mathbb{R}^2))} \le C_{\mathrm{TM}}A^{40/9} + eA^5 < \infty.$$

Therefore u scatters for $\|(u_0,u_1)\|_{H^1(\mathbb{R}^2)\times L^2(\mathbb{R}^2)} < \varepsilon_0$, and in particular for $E_1<\varepsilon_0$.

3. Scattering for large data

Conformal inversion. Suppose we are given initial data at time $a > 0$. We assume they are compactly supported inside a ball of radius $a/2$. Because of the finite speed of propagation, the solution is confined within the forward light cone emanating from the origin at time $a/2$:

$$\operatorname{supp} u(t,\cdot) \subset B_{t-a/2}(0), \quad t \geq a.$$

We perform a conformal inversion

$$\Phi : (t,x,u) \mapsto (T,X,U),$$

as in [Grillakis 1990]; that is, we define

$$T := \frac{t}{t^2-r^2}, \quad X := \frac{x}{t^2-r^2}, \quad U := \Omega^{-\frac{1}{2}} u$$

with the weight

$$\Omega := \frac{1}{t^2-r^2} = T^2 - R^2,$$

where $r = |x|$, $R = |X|$. Conformal inversion leaves the structure of the d'Alembert operator invariant [Godin 1994, Lemma 4.2] and

$$(\partial_T^2 - \Delta_X) U = \Omega^{-\frac{5}{2}} (\partial_t^2 - \Delta_x) u.$$

In fact, conformal inversion can be regarded as a Kelvin transform of Minkowski space $(\mathbb{R}^{1,2}, \eta)$ with metric $\eta_{\mu\nu} = \operatorname{diag}(+1,-1,-1)$. This can be seen by writing the transformation as

$$G : x^\lambda \mapsto x^\lambda (x^\mu x^\nu \eta_{\mu\nu})^{-1} = x^\lambda \langle x, x \rangle_\eta^{-1}.$$

One then calculates the differential,

$$dG_x(y) = \frac{d}{dt}\Big|_{t=0} G(x+ty) = \frac{d}{dt}\Big|_{t=0} \left(\frac{x+ty}{\langle x,x\rangle_\eta + 2t\langle x,y\rangle_\eta + t^2\langle y,y\rangle_\eta} \right) = \frac{y}{\langle x,x\rangle_\eta} - \frac{2x\langle x,y\rangle_\eta}{\langle x,x\rangle_\eta^2},$$

so that $\langle (dG_x)y, (dG_x)y \rangle_\eta = \langle x,x\rangle_\eta^{-2} \langle y,y\rangle_\eta$ and the differential is a conformal transformation with respect to the metric η.

In the new variables T, X, (1) becomes

$$\partial_T^2 U - \Delta U + \Omega^{-2} U (e^{\Omega U^2} - 1 - \Omega U^2) = 0. \tag{18}$$

Note that we changed the direction of time. The transformed function U has support inside the set

$$\operatorname{supp} U = \left\{ (T,X) : T + R \leq \frac{2}{a} \text{ and } \frac{T}{T^2-R^2} \geq a \text{ and additionally } R \leq T \right\}.$$

For the following arguments we fix a. This is not a restriction. In fact, for any initial data with compact support, we may shift the initial time such that the support of the initial data at the starting time is contained inside our fixed cone. We choose $a = 1$. This leads to $\Omega \leq 1$ for $T \leq 1$.

Energy-flux relation in conformal coordinates. For the remainder of the argument we closely follow [Struwe 2013]. We multiply (18) with U_T. Then we obtain

$$\partial_T e - \operatorname{div} m = TP \tag{19}$$

with the scaled energy density

$$e := \tfrac{1}{2}\big(U_T^2 + |\nabla U|^2 + \Omega^{-3}\big(e^{\Omega U^2} - 1 - \Omega U^2 - \tfrac{1}{2}\Omega^2 U^4\big)\big),$$

the momentum density

$$m := U_T \nabla U,$$

and the remainder

$$P := \Omega^{-4}\big(\Omega U^2(e^{\Omega U^2} - 1 - \Omega U^2) - 3\big(e^{\Omega U^2} - 1 - \Omega U^2 - \tfrac{1}{2}\Omega^2 U^4\big)\big) = U^8 \sum_{k=0}^{\infty} \frac{(\Omega U^2)^k}{(k+4)!}(k+1) \geq 0.$$

The power series expansion of P shows that the right-hand side of (19) is positive. Therefore the scaled energy is nonincreasing as we approach the origin. Note that removing the mass term is crucial at this point. Without doing so, we are left with an additional term $-2\Omega^{-2}U^2$ in P that spoils the definite sign of the remainder. Furthermore, the same observation holds for the u^3-term in the original equation.

For $T_0 < 1$, we integrate (19) over the forward light cone $\{R \leq T\}$ where we truncate by the initial data surface and the support of U, that is, we integrate over

$$K := \{(T, X) \in \operatorname{supp} U,\ T_0 \leq T,\ |X| = R \leq T\}.$$

Its boundary ∂K has four components. The first one is the initial data surface. It contributes the energy E_a on the initial data surface. The second is the boundary of the support of U inside $\{R < T\}$. Its contribution vanishes. The third boundary is the mantle of the light cone,

$$M_{T_0}^1 := \{(T, X) : T_0 \leq T \leq 1,\ |X| = R = T\}.$$

We write

$$V(Y) := U(|Y|, Y)$$

for the restriction of U to the mantle. We call the quantity

$$\int_{M_{T_1}^{T_2}} \tfrac{1}{2}\big(|\nabla V|^2 + \Omega^{-3}\big(e^{\Omega V^2} - 1 - \Omega V^2 - \tfrac{1}{2}\Omega^2 V^4\big)\big)\, dY$$

the *flux* of U through the mantle $M_{T_1}^{T_2}$. The last boundary yields the energy in new coordinates:

$$E(T_0) := \int_{B_{T_0}(0)} e\, dX.$$

Putting everything together, we find

$$E(T_0) + \frac{1}{\sqrt{2}} \operatorname{Flux}(M_{T_0}^1) = E_a - \int_K PT\, dX\, dT.$$

In particular, we have the energy inequality

$$E(T_0) + \frac{1}{\sqrt{2}} \operatorname{Flux}(M^1_{T_0}) \le E_a.$$

Therefore the limit $\lim_{T\to 0} E(T, B_T(0))$ exists and the flux decays:

$$\operatorname{Flux}(M^T_0) := \sup_{0<S<T} \operatorname{Flux}(M^T_S) \to 0, \quad T \to 0. \tag{20}$$

Moreover, the remainder term PT is bounded by the initial energy

$$\int_K PT\, dX\, dT \le E_a. \tag{21}$$

Pointwise estimates for the average on the mantle. We derive pointwise estimates for the spherical averages

$$\overline{V} = \overline{V}(T) = \frac{1}{2\pi} \int_0^{2\pi} V(e^{i\phi} T)\, d\phi \tag{22}$$

of V, the trace of U on $M^{T_0}_0$. By Hölder's inequality, for any $0 < T \le T_1$,

$$\begin{aligned} |\overline{V}(T)| &\le |\overline{V}(T_1)| + \int_T^{T_1} |\overline{V}'(S)|\, dS \le |\overline{V}(T_1)| + \left(\int_T^{T_1} |\overline{\nabla V}|^2 S\, dS \cdot \int_T^{T_1} \frac{dS}{S} \right)^{\frac{1}{2}} \\ &\le |\overline{V}(T_1)| + \pi^{-\frac{1}{2}} \operatorname{Flux}^{\frac{1}{2}}(M^{T_1}_T) \log^{\frac{1}{2}} \frac{T_1}{T}. \end{aligned}$$

Flux decays towards the origin by (20). So there exists a time $T_0 \le 1$ such that, for smaller times $0 < T \le T_0$, we have

$$\operatorname{Flux}^{\frac{1}{2}}(M^{T_0}_T) \le \operatorname{Flux}^{\frac{1}{2}}(M^{T_0}_0) \le \tfrac{1}{8}.$$

With this explicit bound on the flux, we can fix a second time T_1, $0 < T_1 \le T_0$, such that $8|\overline{V}(T_0)| \le \log^{1/2}(1/T)$ for $0 < T \le T_1$. By $T_0 \le 1$ we have $\log(T_0/T) \le \log(1/T)$. Therefore,

$$4|\overline{V}(T)| \le \log^{\frac{1}{2}} \frac{1}{T} \quad \text{for all } 0 < T \le T_1. \tag{23}$$

Decay of energy. We introduce polar coordinates R, ϕ. The energy law (19) becomes

$$\partial_T(Re) - \partial_R(Rm) = \frac{1}{R} \partial_\phi(U_T U_\phi) + RTP, \tag{24}$$

where now

$$\begin{aligned} e &:= \tfrac{1}{2}\big(U_T^2 + U_R^2 + R^{-2} U_\phi^2 + \Omega^3 \big(e^{\Omega U^2} - 1 - \Omega U^2 - \tfrac{1}{2}\Omega^2 U^4\big)\big), \\ m &:= U_T U_R. \end{aligned}$$

We multiply (18) with $X \cdot \nabla U$. Then

$$\partial_T(X\cdot m)-\operatorname{div}\Big(X\cdot\nabla U\nabla U-\frac{X}{2}\big(|\nabla U|^2-U_T^2+\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac12\Omega^2U^4\big)\big)\Big)$$
$$+U_T^2-\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac12\Omega^2U^4\big)=-R^2P.$$

In polar coordinates,

$$\partial_T(R^2m)-\tfrac12\partial_R\big(R^2\big(U_T^2+U_R^2-R^{-2}U_\phi^2-\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac12\Omega^2U^4\big)\big)\big)$$
$$+R\big(U_T^2-\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac12\Omega^2U^4\big)\big)=\partial_\phi(U_RU_\phi)-R^3P. \quad (25)$$

Multiplying (18) with $(U-\overline{V})$, we obtain

$$\partial_T(U_T(U-\overline{V}))-\operatorname{div}(\nabla U(U-\overline{V}))+|\nabla U|^2-U_T^2+U_T\overline{V}_T+\Omega^{-2}U(U-\overline{V})(e^{\Omega U^2}-1-\Omega U^2)=0.$$

Or, again in polar coordinates,

$$\partial_T(RU_T(U-\overline{V}))-\partial_R(RU_R(U-\overline{V}))+R\big(|\nabla U|^2-U_T^2+U_T\overline{V}_T+\Omega^{-2}U(U-\overline{V})(e^{\Omega U^2}-1-\Omega U^2)\big)$$
$$=\frac{1}{R}\partial_\phi((U-\overline{V})U_\phi). \quad (26)$$

We rescale the energy identity (24) with R/T. Then

$$\partial_T\Big(\frac{R^2}{T}e\Big)-\partial_R\Big(\frac{R^2}{T}m\Big)+\frac{R^2}{T^2}e+\frac{R}{T}m=\partial_\phi\Big(\frac{1}{T}U_TU_\phi\Big)+R^2P. \quad (27)$$

We divide both (25) and (26) by T. Then

$$\partial_T\Big(\frac{R^2}{T}m\Big)-\tfrac12\partial_R\Big(\frac{R^2}{T}\big(U_T^2+U_R^2-R^{-2}U_\phi^2-\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac12\Omega^2U^4\big)\big)\Big)$$
$$+\frac{R^2}{T^2}m+\frac{R}{T}\big(U_T^2-\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac12\Omega^2U^4\big)\big)=\partial_\phi\Big(\frac{1}{T}U_RU_\phi\Big)-\frac{R^3}{T}P. \quad (28)$$

and

$$\partial_T\Big(\frac{R}{T}U_T(U-\overline{V})\Big)-\partial_R\Big(\frac{R}{T}U_R(U-\overline{V})\Big)$$
$$+\frac{R}{T}\Big(|\nabla U|^2-U_T^2+U_T\overline{V}_T+U_T\frac{U-\overline{V}}{T}+\Omega^{-2}U(U-\overline{V})(e^{\Omega U^2}-1-\Omega U^2)\Big)$$
$$=\partial_T\Big(\frac{R}{T}\Big(U_T(U-\overline{V})+\frac{(U-\overline{V})^2}{2T}\Big)\Big)-\partial_R\Big(\frac{R}{T}U_R(U-\overline{V})\Big)$$
$$+\frac{R}{T}\Big(|\nabla U|^2-U_T^2+U_T\overline{V}_T+V_T\frac{U-\overline{V}}{T}+\frac{(U-\overline{V})^2}{T^2}+\Omega^{-2}U(U-\overline{V})(e^{\Omega U^2}-1-\Omega U^2)\Big)$$
$$=\partial_\phi\Big(\frac{U-\overline{V}}{RT}U_\phi\Big). \quad (29)$$

Adding (27) and (28) with one half of (29) yields

$$\begin{aligned}
&\partial_T\left(\frac{R^2}{T}\left(e+m+\tfrac{1}{2}U_T\frac{U-\overline{V}}{R}+\frac{(U-\overline{V})^2}{4TR}\right)\right)\\
&\quad-\partial_R\left(\frac{R^2}{T}\left(e+m-R^{-2}U_\phi^2-\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac{1}{2}\Omega^2U^4\big)+U_R\frac{U-\overline{V}}{2R}\right)\right)\\
&\qquad+\frac{R}{T}\left(\left(1+\frac{R}{T}\right)(e+m)+\tfrac{1}{2}U_T\overline{V}_T+\overline{V}_T\frac{U-\overline{V}}{2T}+\frac{(U-\overline{V})^2}{2T^2}\right)\\
&=\partial_\phi\left(\frac{1}{T}\left(U_R+U_T+\frac{U-\overline{V}}{2R}\right)U_\phi\right)\\
&\quad+\frac{R}{T}\big(\tfrac{3}{2}\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac{1}{2}\Omega^2U^4\big)-\tfrac{1}{2}\Omega^{-2}U(U-\overline{V})(e^{\Omega U^2}-1-\Omega U^2)\big)+R^2\left(1-\frac{R}{T}\right)P. \quad (30)
\end{aligned}$$

Lemma 3.1. *For any time T_2 with $0<T_2<T_1$, we have*

$$\int_{K^{T_2}}\left(\left(1\pm\frac{R}{T}\right)(e\pm m)+\frac{(U-\overline{V})^2}{2T^2}\right)\frac{dX\,dT}{T}\le C(1+E_a+T_2^2E_a^3),$$

where K^{T_2} is the truncated light cone

$$K^{T_2}:=\{(T,X):T\le T_2,\ |X|\le T\}.$$

Proof. Fix $T_2<T_1$. We integrate (30) over the truncated cone K^{T_2}. Then

$$I_+=\int_{K^{T_2}}\left(\left(1+\frac{R}{T}\right)(e+m)+\frac{(U-\overline{V})^2}{2T^2}\right)\frac{dX\,dT}{T}\le \mathrm{II}+\mathrm{IV}+\mathrm{V},$$

where we label the terms I_+, II, IV, and V as in the proof of [Struwe 2013, Lemma 3.1]. As shown there, by Poincaré's inequality, we obtain

$$\mathrm{II}\le E_a,\quad \mathrm{IV}\le C\ \mathrm{Flux}(M_0^{T_2})\le CE_a.$$

The first two terms of our error term

$$\begin{aligned}
\mathrm{V}=\int_{K^{T_2}}\Bigg(&-\tfrac{1}{2}U_T\overline{V}_T-\overline{V}_T\frac{U-\overline{V}}{2T}+RT\left(1-\frac{R}{T}\right)P\\
&+\tfrac{3}{2}\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac{1}{2}\Omega^2U^4\big)-\tfrac{1}{2}U(U-\overline{V})\Omega^{-2}(e^{\Omega U^2}-1-\Omega U^2)\Bigg)\frac{dX\,dT}{T}
\end{aligned}$$

are the same as in Struwe's work and so, for any $\delta>0$, we have

$$\left|\int_{K^{T_2}}\left(U_T\overline{V}_T+\overline{V}_T\frac{U-\overline{V}}{T}\right)\frac{dX\,dT}{T}\right|\le C\delta\int_0^{T_2}\int_{B_{T/2}(0)}|\nabla U|^2\frac{dX\,dT}{T}+C\delta I_++C\delta^{-1}\,\mathrm{Flux}(M_0^{T_2}).$$

By (21),

$$\int_{K^{T_2}}R\left(1-\frac{R}{T}\right)P\,dX\,dT\le\int_{K^{T_2}}TP\,dX\,dT\le E_a.$$

For the remaining terms we add and subtract in the spherical averages as defined in (22):

$$\begin{aligned}
&\tfrac{3}{2}\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac{1}{2}\Omega^2U^4\big)-\tfrac{1}{2}U(U-\overline{V})\Omega^{-2}\big(e^{\Omega U^2}-1-\Omega U^2\big)\\
&\quad=\tfrac{3}{2}\Omega^{-3}\big(e^{\Omega U^2}-1-\Omega U^2-\tfrac{1}{2}\Omega^2U^4-\big(e^{\Omega\overline{V}^2}-1-\Omega\overline{V}^2-\tfrac{1}{2}\Omega^2\overline{V}^4\big)\big)\\
&\qquad\qquad-\tfrac{1}{2}U(U-\overline{V})\Omega^{-2}\big(e^{\Omega U^2}-1-\Omega U^2\big)+\tfrac{3}{2}\Omega^{-3}\big(e^{\Omega\overline{V}^2}-1-\Omega\overline{V}^2-\tfrac{1}{2}\Omega^2\overline{V}^4\big)\\
&\quad=f(U,\overline{V})+\tfrac{3}{2}\Omega^{-3}\big(e^{\Omega\overline{V}^2}-1-\Omega\overline{V}^2-\tfrac{1}{2}\Omega^2\overline{V}^4\big).
\end{aligned}$$

We can compensate for the second term with the pointwise bound from (23):

$$\begin{aligned}
\tfrac{3}{2}\Omega^{-3}(e^{\Omega\overline{V}^2}-1-\Omega\overline{V}^2-\tfrac{1}{2}\Omega^2\overline{V}^4)&=\tfrac{3}{2}\sum_{k=3}^{\infty}\frac{\Omega^{k-3}\overline{V}^{2k}}{k!}\\
&=\tfrac{3}{2}\overline{V}^6\sum_{k=0}^{\infty}\frac{(\Omega\overline{V}^2)^k}{(k+3)!}\le\tfrac{3}{2}\overline{V}^6e^{\Omega\overline{V}^2}\\
&\le C\log^3\Big(\frac{1}{T}\Big)\frac{1}{T^{\frac{1}{16}\Omega}}\le C\log^3\Big(\frac{1}{T}\Big)\frac{1}{T},
\end{aligned}$$

where we used $\Omega\le 1$. Then

$$\int_{K^{T_2}}\log^3\Big(\frac{1}{T}\Big)\frac{1}{T}\frac{dX\,dT}{T}\le C\int_0^T\log^3\Big(\frac{1}{T}\Big)\,dT\le C<\infty.$$

In the following, we analyze the nonlinear function f as above by comparing $U(T,X)$ with $\overline{V}(T)$ pointwise in X for a fixed time slice. Recalling that

$$f(U,\overline{V})=\frac{3}{2}\sum_{k=3}^{\infty}\frac{\Omega^{k-3}(U^{2k}-\overline{V}^{2k})}{k!}-\tfrac{1}{2}U(U-\overline{V})\sum_{k=2}^{\infty}\frac{\Omega^{k-2}U^{2k}}{k!},$$

we observe that $f(-U,-\overline{V})=f(U,\overline{V})$. Furthermore, if U and $\overline{V}$ have opposite sign, say $U\ge 0$, $\overline{V}\le 0$, then $U(U-\overline{V})\ge U^2$. Comparing coefficients, we see that the second power series dominates the first and f is negative. Therefore, we only need to analyze the case $U,\overline{V}>0$. We distinguish three subcases.

(i) If $U\le\overline{V}$, then

$$f(U,\overline{V})\le\tfrac{1}{2}\overline{V}^2\Omega^{-2}(e^{\Omega\overline{V}^2}-1-\Omega\overline{V}^2)\le\tfrac{1}{2}\overline{V}^6e^{\Omega\overline{V}^2},$$

which we estimate with the bound on $|\overline{V}|$ as above.

(ii) If $\overline{V}<U\le 4\overline{V}$, then

$$f(U,\overline{V})\le\tfrac{3}{2}\Omega^{-3}\big(e^{16\Omega\overline{V}^2}-1-16\Omega\overline{V}^2-\tfrac{1}{2}(16\Omega)^2\overline{V}^4\big)\le\tfrac{3}{2}16^3\overline{V}^6e^{16\Omega\overline{V}^2}\le C\log^3\Big(\frac{1}{T}\Big)\frac{1}{T},$$

where the factor 4 in (23) together with $\Omega\le 1$ ensures that the power in $1/T$ stays smaller than 1.

(iii) For the remaining case $U>4\overline{V}$, we write $\overline{V}=\alpha U$, that is, $\alpha<\frac{1}{4}$. Then we analyze the power series

$$f(U,\overline{V})=\tfrac{1}{4}(U^6-\overline{V}^6)+\frac{3}{2}\sum_{k=4}^{\infty}\frac{\Omega^{k-3}(U^{2k}-\overline{V}^{2k})}{k!}-\tfrac{1}{2}U(U-\overline{V})\sum_{k=2}^{\infty}\frac{\Omega^{k-2}U^{2k}}{k!}.$$

For the leading term, we use $\alpha < \frac{1}{4}$ to compare with $(U-\overline{V})^6$,

$$U^6 - \overline{V}^6 = U^6(1-\alpha^6) \le CU^6(1-\alpha)^6 = C(U-\overline{V})^6,$$

pointwise. Then, by the Poincaré–Sobolev inequality, on each time slice,

$$\int_{B_T(0)} \frac{(U-\overline{V})^6}{T}\,dX \le \frac{C}{T}\left(\int_{B_T(0)} |\nabla U|^{\frac{3}{2}}\,dX\right)^6 \le CT\left(\int_{B_T(0)} |\nabla U|^2\,dX\right)^3 \le CTE_a^3.$$

Integration in time yields a term bounded by $T_2^2 E_a^3$. The remaining power series is negative, as

$$\begin{aligned}
\frac{3}{2}\sum_{k=4}^{\infty} \frac{\Omega^{k-3}(U^{2k}-\overline{V}^{2k})}{k!} - \tfrac{1}{2}U(U-\overline{V})\sum_{k=2}^{\infty}\frac{\Omega^{k-2}U^{2k}}{k!} \\
&= \tfrac{3}{2}U^6\sum_{k=1}^{\infty}\frac{(\Omega U^2)^k(1-\alpha^{2(k+3)})}{(k+3)!} - \tfrac{1}{4}(1-\alpha)U^6\sum_{k=0}^{\infty}\frac{(\Omega U^2)^k}{(k+2)!} \\
&= U^6\left(-\tfrac{1}{2}(1-\alpha) + \sum_{k=1}^{\infty}\frac{(\Omega U^2)^k}{(k+3)!}\left(\tfrac{1}{2}(3(1-\alpha^{2(k+3)}) - (1-\alpha)(k+3))\right)\right) \\
&\le 0.
\end{aligned}$$

Note that this calculation further motivates the exclusion of u^3 in the original equation.

Combining, we arrive at the estimate

$$\mathrm{V} \le C(1 + E_a + T_2^2 E_a^3 + \delta I_+) + C\delta \int_0^{T_2}\int_{B_{T/2}(0)} |\nabla U|^2 \frac{dX\,dT}{T} + C\delta^{-1}\,\mathrm{Flux}(M_0^{T_2}).$$

By the energy inequality, $\mathrm{Flux}(M_0^{T_2}) \le E_a$. Therefore,

$$I_+ \le C(1 + E_a + T_2^2 E_a^3 + \delta I_+ + \delta^{-1} E_a) + C\delta \int_0^{T_2}\int_{B_{T/2}(0)} |\nabla U|^2 \frac{dX\,dT}{T},$$

and, in the same fashion,

$$\begin{aligned}
I_- &= \int_{K^{T_2}} \left(\left(1-\frac{R}{T}\right)(e-m) + \frac{(U-\overline{V})^2}{2T^2}\right)\frac{dX\,dT}{T} \\
&\le C(1 + E_a + T_2^2 E_a^3 + \delta I_+ + \delta^{-1} E_a) + C\delta \int_0^{T_2}\int_{B_{T/2}(0)} |\nabla U|^2 \frac{dX\,dT}{T}.
\end{aligned}$$

We have $|\nabla U|^2 \le 2e = (e+m) + (e-m)$, and hence

$$\int_0^{T_2}\int_{B_{T/2}(0)} |\nabla U|^2 \frac{dX\,dT}{T} \le I_+ + 2I_-.$$

Choosing $\delta > 0$ sufficiently small, we conclude that

$$I_+ + I_- \le C(1 + E_a + T_2^2 E_a^3). \qquad \square$$

Bound inside a hyperboloid. Recall that T_1 was fixed to bound $|\overline{V}(T)|$ as in (23), which in turn was crucial for smallness in Lemma 3.1.

For any $\varepsilon > 0$, we fix a time $0 < T_\varepsilon < T_1$ such that

$$\operatorname{Flux}(u, M^{T_\varepsilon}) + \int_{K^{T_\varepsilon}} \left(\left(1 \pm \frac{R}{T}\right)(e \pm m) + \frac{(U - \overline{V})^2}{T^2} \right) \frac{dX\, dT}{T} < \varepsilon.$$

In the same fashion as in [Struwe 2013, Lemma 4.3], we obtain:

Lemma 3.2. *There exists $\varepsilon > 0$ and a constant $C < \infty$ such that, for any $0 < T \le 4^{-1} T_\varepsilon$, we have*

$$\int_{K^T} e^{4U^2}\, dX\, dT \le C\, T.$$

The region $\Phi^{-1}(K^T)$ is a hyperboloid. Its asymptote is the cone $\{r = t - 1/(2T)\}$.

In the following we fix $T \le 4^{-1} T_\epsilon$. Let $t_0 = 1/T$, the smallest time inside the hyperboloid. Furthermore, we denote $D = \Phi^{-1}(K^T)$.

Using Lemma 3.2, we obtain decay of the nonlinearity in $L_t^2 L_x^2$ locally in time.

Lemma 3.3. *Let $t_2 \ge t_1 \ge t_0$. Then*

$$\int_{D \cap \{t_1 \le t \le t_2\}} |N(u)|^2\, dx\, dt \le C t_1^{-2}.$$

Proof. Inside $D_{t_1}^{t_2} = D \cap \{t_1 \le t \le t_2\}$ we have $t + r \ge t$ and $t - r \ge 1/(2T)$. Therefore, $\Omega \le C/t_1$ with a constant C that is uniform over $D_{t_1}^{t_2}$. Then we calculate

$$\begin{aligned} \int_{D_{t_1}^{t_2}} |u(e^{u^2} - 1 - u^2)|^2\, dx\, dt &= \int_{\Phi(D_{t_1}^{t_2})} \Omega U^2 (e^{\Omega U^2} - 1 - \Omega U^2)^2 \Omega^{-3}\, dX\, dT \\ &\le \int_{\Phi(D_{t_1}^{t_2})} \tfrac{1}{4} \Omega^2 U^{10} e^{2\Omega U^2}\, dX\, dT \le \frac{C}{t_1^2} \int_{\Phi(D_{t_1}^{t_2})} e^{3U^2}\, dX\, dT \le C \frac{T}{t_1^2}. \quad \square \end{aligned}$$

We conclude:

Lemma 3.4. *Inside D the nonlinearity is bounded in $L_t^1 L_x^2$, that is,*

$$\int_{t_0}^{\infty} \left(\int_{D \cap (\{t\} \times \mathbb{R}^2)} |N|^2\, dx \right)^{\frac{1}{2}} dt < \infty.$$

Proof. Divide $[t_0, \infty)$ into intervals $I_n = [t_0 2^n, t_0 2^{n+1})$. Then, by Hölder's inequality and Lemma 3.3,

$$\begin{aligned} \int_{t_0}^{\infty} \left(\int_{D \cap (\{t\} \times \mathbb{R}^2)} |N|^2\, dx \right)^{\frac{1}{2}} dt &= \sum_{n=0}^{\infty} \int_{I_n} \left(\int_{D \cap (\{t\} \times \mathbb{R}^2)} |N|^2\, dx \right)^{\frac{1}{2}} dt \\ &\le \sum_{n=0}^{\infty} (t_0 2^n)^{\frac{1}{2}} \left(\int_{I_n} \int_{D \cap (\{t\} \times \mathbb{R}^2)} |N|^2\, dx\, dt \right)^{\frac{1}{2}} \\ &\le \sum_{n=0}^{\infty} C t_0^{-\frac{1}{2}} 2^{-n/2} < \infty. \end{aligned}$$

$\square$

The case of radial data. In the previous section we have obtained control of the nonlinearity inside a hyperboloid $\Phi^{-1}(K^T)$, where $T \le 4^{-1}T_\epsilon$. Let $t_0 = 1/T$, the smallest time in the hyperboloid. Now fix T and choose $d > 1/(2T)$. Let

$$A_{t_1} = \{(t,x) : t \ge t_1,\ t-d \le |x| \le t\}.$$

Then there exists a time $t_1 \ge t_0$ such that

$$\{(t,x) : t \ge t_1,\ |x| \le t\} \subset (\Phi^{-1}(K^T) \cap \{(t,x) : t \ge t_1\}) \cup A_{t_1},$$

that is, the thinned cone A_{t_1} covers the region where we have not yet obtained control over the nonlinearity.

In the following, we will restrict ourselves to the case of radial solutions. We will show that we can bound the nonlinearity inside A_{t_1} in $L^1_t L^2_x$.

In the case of radially symmetric data, we employ the following bound. Let $t > t_1$ fixed, $t-d \le r \le t$. Recall that u is compactly supported within $B_t(0)$. Then

$$\begin{aligned} |u(t,r)| &\le \int_r^t |\partial_s u(t,s)|\,ds \le \int_{t-d}^t |\partial_s u(t,s)|\,ds \\ &\le \left(\int_{t-d}^t |\partial_s u(t,s)|^2 s\,ds\right)^{\frac12} \left(\int_{t-d}^t \frac{1}{s}\,ds\right)^{\frac12} \le CE^{\frac12}\left(\log\frac{t}{t-d}\right)^{\frac12}. \end{aligned}$$

Therefore there exists $t_2 \ge t_1$ such that $|u(t,r)| \le \dfrac{C}{t^{1/2}}$ for all $t \ge t_2$, with a constant C independent of $t \ge t_2$.

Lemma 3.5. *Let t_2 be as above. Then N is bounded in $L^1_t L^2_x$ inside A_{t_2}.*

Proof. Again we estimate

$$|N(u)| = |u|(e^{u^2} - 1 - u^2) \le \tfrac12 |u|^5 e^{u^2}$$

pointwise. Then

$$\int_{B_t(0)\setminus B_{t-d}(0)} u^{10} e^{2u^2}\,dx \le Ct \cdot t^{-5} = Ct^{-4}.$$

Therefore,

$$\int_{t_2}^\infty \left(\int_{B_t(0)\setminus B_{t-d}(0)} u^{10} e^{2u^2}\,dx\right)^{\frac12} dt \le C\int_{t_2}^\infty t^{-2}\,dt < \infty. \qquad \square$$

Combining Lemmas 3.3 and 3.5, we obtain $\|N\|_{L^1([t_2,\infty);L^2(\mathbb{R}^2))} < \infty$. Using Lemma 1.2, we conclude the proof of Theorem 1.1

References

[Godin 1994] P. Godin, "Global sound waves for quasilinear second order wave equations", *Math. Ann.* **298**:3 (1994), 497–531. MR 95f:35156 Zbl 0790.35071

[Grillakis 1990] M. G. Grillakis, "Regularity and asymptotic behaviour of the wave equation with a critical nonlinearity", *Ann. of Math.* (2) **132**:3 (1990), 485–509. MR 92c:35080 Zbl 0736.35067

[Ibrahim et al. 2006] S. Ibrahim, M. Majdoub, and N. Masmoudi, "Global solutions for a semilinear, two-dimensional Klein–Gordon equation with exponential-type nonlinearity", *Comm. Pure Appl. Math.* **59**:11 (2006), 1639–1658. MR 2007h:35229 Zbl 1117.35049

[Ibrahim et al. 2009] S. Ibrahim, M. Majdoub, N. Masmoudi, and K. Nakanishi, "Scattering for the two-dimensional energy-critical wave equation", *Duke Math. J.* **150**:2 (2009), 287–329. MR 2010k:35313 Zbl 1206.35175

[Keel and Tao 1998] M. Keel and T. Tao, "Endpoint Strichartz estimates", *Amer. J. Math.* **120**:5 (1998), 955–980. MR 2000d:35018 Zbl 0922.35028

[Lions and Magenes 1970] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, vol. 3, Travaux et Recherches Mathématiques **20**, Dunod, Paris, 1970. Translated as *Non-homogeneous boundary value problems and applications*, vol. 3, Grundlehren der Mathematischen Wissenschaften **183**, Springer, New York, 1973. MR 45 #975 Zbl 0197.06701

[Nakamura and Ozawa 1999] M. Nakamura and T. Ozawa, "Global solutions in the critical Sobolev space for the wave equations with nonlinearity of exponential growth", *Math. Z.* **231**:3 (1999), 479–487. MR 2001b:35216 Zbl 0931.35107

[Nakanishi and Schlag 2011] K. Nakanishi and W. Schlag, *Invariant manifolds and dispersive Hamiltonian evolution equations*, European Mathematical Society, Zürich, 2011. MR 2012m:37120 Zbl 1235.37002

[Ruf 2005] B. Ruf, "A sharp Trudinger–Moser type inequality for unbounded domains in $\mathbb{R}^2$", *J. Funct. Anal.* **219**:2 (2005), 340–367. MR 2005k:46082 Zbl 1119.46033

[Struwe 2013] M. Struwe, "The critical nonlinear wave equation in two space dimensions", *J. Eur. Math. Soc.* **15**:5 (2013), 1805–1823. MR 3082244 Zbl 1282.35245

Received 4 Jun 2013. Revised 14 Feb 2014. Accepted 11 Apr 2014.

MARTIN SACK: sackm@phys.ethz.ch
Department of Mathematics, ETH Zürich, CH-8092 Zürich, Switzerland

ANALYSIS AND PDE
Vol. 7, No. 3, 2014
dx.doi.org/10.2140/apde.2014.7.717

LARGE-TIME BLOWUP FOR A PERTURBATION OF THE CUBIC SZEGŐ EQUATION

HAIYAN XU

We consider the following Hamiltonian equation on a special manifold of rational functions:

$$i\,\partial_t u = \Pi(|u|^2 u) + \alpha(u|1), \quad \alpha \in \mathbb{R},$$

where Π denotes the Szegő projector on the Hardy space of the circle $\mathbb{S}^1$. The equation with $\alpha = 0$ was first introduced by Gérard and Grellier as a toy model for totally nondispersive evolution equations. We establish the following properties for this equation. For $\alpha < 0$, any compact subset of initial data leads to a relatively compact subset of trajectories. For $\alpha > 0$, there exist trajectories on which high Sobolev norms exponentially grow in time.

1. Introduction

The study on the long time behavior of solutions of Schrödinger type Hamiltonian equations is a central issue in the theory of dispersive nonlinear partial differential equations. For instance, Colliander, Keel, Staffilani, Takaoka, and Tao [Colliander et al. 2010] studied the cubic defocusing nonlinear Schrödinger equation,

$$i\,\partial_t u + \triangle u = \pm|u|^2 u, \quad (t, x) \in \mathbb{R} \times \mathbb{T}^2. \tag{1-1}$$

In that paper, they constructed solutions with small H^s norm at the initial moment, which present a large Sobolev H^s norm at a sufficiently long time T. Guardia and Kaloshin [2012] improved this result by refining the estimates on the time T. Zaher Hani [2014] studied a version of the nonlinear Schrödinger equation obtained by canceling the least resonant part, and showed the existence of unbounded trajectories in high Sobolev norms. Hani, Pausader, Tzvetkov, and Visciglia [Hani et al. 2013] studied the nonlinear Schrödinger equation (1-1) on the spatial domain $\mathbb{R} \times \mathbb{T}^d$, and obtained global solutions to the defocusing and focusing problems (for any $d \geq 2$) with infinitely growing high Sobolev norms H^s.

Gérard and Grellier [2012a] achieved a related result by considering the following degenerate half wave equation on the one-dimensional torus:

$$i\,\partial_t u - |D|u = |u|^2 u. \tag{1-2}$$

This work was supported by grants from Région Île-de-France.
MSC2010: 37J35, 47B35, 35B44.
Keywords: Szegő equation, integrable Hamiltonian systems, Lax pair, large-time blowup.

They found solutions with small Sobolev norms at initial time which become much larger as time grows. More precisely, there exist sequences of solutions u^n and t^n such that $\|u_0^n\|_{H^r} \to 0$ for any r, but

$$\|u^n(t^n)\|_{H^s} \sim \|u_0^n\|_{H^s} \left(\log \frac{1}{\|u_0^n\|_{H^s}} \right)^{2s-1}, \quad s > 1.$$

This result is a consequence of studies on the so-called *cubic Szegő equation*, introduced by Gérard and Grellier [2010; 2012b] as a model of nondispersive dynamics:

$$i\,\partial_t u = \Pi(|u|^2 u). \tag{1-3}$$

The above equation turns out to be the resonant part of the half wave equation (1-2). The operator Π, called the Szegő operator, is defined as a projector onto the nonnegative frequencies. If $u \in \mathcal{D}'(\mathbb{S}^1)$ is a distribution on the circle $\mathbb{S}^1 = \{z \in \mathbb{C} : |z| = 1\}$, then

$$\Pi(u) = \Pi\left(\sum_{k \in \mathbb{Z}} \hat{u}(k) e^{ik\theta} \right) = \sum_{k \geq 0} \hat{u}(k) e^{ik\theta}. \tag{1-4}$$

Notice that, on the Hilbert space $L^2(\mathbb{S}^1)$ endowed with the inner product

$$(u \,|\, v) = \frac{1}{2\pi} \int_{-\pi}^{\pi} u(e^{ix}) \overline{v(e^{ix})}\, dx, \tag{1-5}$$

Π is the orthogonal projector on the subspace $L^2_+(\mathbb{S}^1)$ defined by the conditions

$$\hat{u}(k) = 0 \quad \text{for all } k < 0.$$

Gérard and Grellier [2010; 2012b] studied the Szegő equation on the space

$$H^{1/2}_+(\mathbb{S}^1) := H^{1/2}(\mathbb{S}^1) \cap L^2_+(\mathbb{S}^1)$$

and displayed two Lax pair structures for this completely integrable system. Moreover, they established an explicit formula of every solution with rational initial data [Gérard and Grellier 2013] and illustrated the large-time behavior of Sobolev norms of the solutions; for instance:

Theorem 1.1 [Gérard and Grellier 2010]. *Every solution u of (1-3) on*

$$\widetilde{\mathcal{M}}(1) := \left\{ u = \frac{a + bz}{1 - pz} : 0 \neq a \in \mathbb{C},\ b \in \mathbb{C},\ p \in \mathbb{C},\ |p| < 1,\ a + bp \neq 0 \right\}$$

satisfies

$$\sup_{t \in \mathbb{R}} \|u(t)\|_{H^s} < \infty \quad \textit{for all } s > \tfrac{1}{2}.$$

However, there exists a family of Cauchy data u_0^ε in $\widetilde{\mathcal{M}}(1)$ which converges in $\widetilde{\mathcal{M}}(1)$ for the $C^\infty(\mathbb{S}^1)$ topology as $\varepsilon \to 0$, and $K > 0$ such that the corresponding solutions of (1-3) u^ε satisfy the following condition, for all $\varepsilon > 0$:

$$\textit{for some } t^\varepsilon > 0,\ \|u^\varepsilon(t^\varepsilon)\|_{H^s} \geq K (t^\varepsilon)^{2s-1} \textit{ as } t^\varepsilon \to \infty \quad \textit{for all } s > \tfrac{1}{2}.$$

Another result on this Szegő equation was obtained by Pocovnicu [2011b; 2011a], who studied this equation by replacing the circle $\mathbb{S}^1$ with the real line and got a polynomial growth of high Sobolev norms [Pocovnicu 2011a, Corollary 4], which says that there exists a solution u of the Szegő equation and a constant $C > 0$ such that $\|u(t)\|_{H^s} \geq C|t|^{2s-1}$ for sufficiently large $|t|$.

The aim of this article is to study the properties of global solutions for the following Hamiltonian equation on $L^2_+(\mathbb{S}^1)$, which is the cubic Szegő equation with a linear perturbation:

$$\begin{cases} i\,\partial_t u = \Pi(|u|^2 u) + \alpha(u\,|\,1), \quad \alpha \in \mathbb{R}, \\ u(0,x) = u_0(x). \end{cases} \tag{1-6}$$

In view of (1-5),

$$(u\,|\,1) = \frac{1}{2\pi}\int_{-\pi}^{\pi} u(e^{ix})\,dx$$

is the average of u on $\mathbb{S}^1$.

Equation (1-6), called the α-Szegő equation, inherits three formal conservation laws:

$$\begin{aligned} &\text{mass:} && Q(u) := \int_{\mathbb{S}^1} |u|^2 \frac{d\theta}{2\pi} = \|u\|_{L^2}^2, \\ &\text{momentum:} && M(u) := (Du\,|\,u), \quad D := -i\,\partial_\theta = z\,\partial_z, \\ &\text{energy:} && E_\alpha(u) := \frac{1}{4}\int_{\mathbb{S}^1} |u|^4 \frac{d\theta}{2\pi} + \frac{1}{2}\alpha|(u\,|\,1)|^2. \end{aligned}$$

Slight modifications of the proof of the well-posedness result in [Gérard and Grellier 2010] lead to the result that the α-Szegő equation is globally well posed in $H^s_+(\mathbb{S}^1) = H^s(\mathbb{S}^1) \cap L^2_+(\mathbb{S}^1)$ for $s \geq \frac{1}{2}$:

Theorem 1.2. *Given $u_0 \in H^{1/2}_+(\mathbb{S}^1)$, there exists a unique global solution $u \in C(\mathbb{R}; H^{1/2}_+)$ of (1-6) with u_0 as the initial condition. Moreover, if $u_0 \in H^s_+(\mathbb{S}^1)$ for some $s > \frac{1}{2}$, then $u \in C^\infty(\mathbb{R}; H^s_+)$. Furthermore, if $u_0 \in H^s_+(\mathbb{S}^1)$ with $s > 1$, the Wiener norm of u is bounded uniformly in time*:

$$\sup_{t\in\mathbb{R}} \|u(t)\|_W := \sup_{t\in\mathbb{R}} \sum_{k=0}^{\infty} |\widehat{u(t)}(k)| \leq C_s \|u_0\|_{H^s}. \tag{1-7}$$

Now we present our main results. In our case with a perturbation term, it turns out that if $\alpha < 0$, the Sobolev norm stays bounded uniformly in time, while if $\alpha > 0$, it may grow exponentially fast:

Theorem 1.3. *Let $u_0 = b_0 + c_0 z/(1 - p_0 z)$, $c_0 \neq 0$, $|p_0| < 1$.*

For $\alpha < 0$, the Sobolev norm of the solution will stay bounded:

$$\|u(t)\|_{H^s} \leq C, \quad C \text{ does not depend on time } t,\ s \geq 0. \tag{1-8}$$

For $\alpha > 0$, the solution u of the α-Szegő equation (1-6) has a Sobolev norm growing exponentially in time:

$$\|u(t)\|_{H^s} \simeq e^{C_{\alpha,s}|t|}, \quad s > \tfrac{1}{2},\ C_{\alpha,s} > 0,\ |t| \to \infty \tag{1-9}$$

if and only if

$$E_\alpha = \tfrac{1}{4}Q^2 + \tfrac{1}{2}\alpha Q. \tag{1-10}$$

Remark 1.4. (1) Together with the results in [Gérard and Grellier 2010; 2012b], we now have a complete picture for the high Sobolev norm of the solutions to the α-Szegő equation. For $\alpha < 0$, it stays bounded (uniformly on time). For $\alpha > 0$, it turns out to have an exponential growth for some initial data satisfying the condition in Theorem 1.3. Finally, for $\alpha = 0$, the trajectories of the Szegő equation with rational initial data are quasiperiodic with instability of the H^s norm as in Theorem 1.1.

(2) Our result is in strong contrast with Bourgain's [1996] and Staffilani's [1997] results for the dispersive equations, which say that the dispersive equations admit polynomial upper bounds on Sobolev norms. Here, we give an example of exponential growth of Sobolev norms for a nondispersive model.

(3) The solutions to the α-Szegő equation admit an exponential upper bound of the Sobolev norms. Assuming $s > 1$, it is easy to solve (1-6) locally in time. More precisely, one has to solve the integral equation

$$u(t) = u_0 - i \int_0^t \big(\Pi(|u|^2 u) + \alpha(u|1)\big)\, dt'.$$

Thus

$$\|u(t)\|_{H^s} \le \|u_0\|_{H^s} + c \int_0^t (1 + \|u(t')\|_W^2)\, \|u(t')\|_{H^s}\, dt',$$

since, by Theorem 1.2, the Wiener norm is uniformly bounded. Then, by Gronwall's inequality, we have

$$\|u(t)\|_{H^s} \le \|u_0\|_{H^s} e^{ct}.$$

This shows that (1-9) is the worst that can happen.

This paper is organized as follows. In Section 2, we prove that there exists a Lax pair for the α-Szegő equation based on Hankel operators. Then we define the manifolds $\mathcal{L}(k) := \{u : \operatorname{rk} K_u = k,\ k \in \mathbb{Z}^+\}$ with the shifted Hankel operator K_u. These manifolds are proved to be invariant by the flow and can be represented as sets of rational functions. In this paper we will just consider the solutions $u \in \mathcal{L}(1)$. We plan to address the other cases in a forthcoming work. In Section 3, we prove the large-time blowup result and the boundedness of the Wiener norm to show that our result is optimal. Furthermore, we provide an example which describes the energy cascade. Finally, we present some perspectives in Section 4.

2. The Lax pair structure

For $u \in E \subset \mathcal{D}'(\mathbb{S}^1)$, we define E_+ by canceling the negative Fourier modes of u:

$$E_+ = \{u \in E : \text{for all } k < 0, \hat{u}(k) = 0\}.$$

In particular, L^2_+ is the Hardy space of L^2 functions which extend to the unit disc $D = \{z \in \mathbb{C} : |z| < 1\}$ as holomorphic functions

$$u(z) = \sum_{k \ge 0} \hat{u}(k) z^k, \quad \sum_{k \ge 0} |\hat{u}(k)|^2 < \infty.$$

An element of L^2_+ can therefore be seen either as a square integrable function $u = u(e^{i\theta})$ on the circle with only nonnegative Fourier modes, or a holomorphic function $u = u(z)$ on the unit disc with square summable Taylor coefficients.

Using the Szegő projector defined as (1-4), we first introduce two important classes of operators on $L^2_+(\mathbb{S}^1)$, namely, the Hankel and Toeplitz operators.

Given $u \in H^{1/2}_+(\mathbb{S}^1)$, a Hankel operator $H_u : L^2_+ \to L^2_+$ is defined by

$$H_u(h) = \Pi(u\bar{h}).$$

Notice that H_u is $\mathbb{C}$-antilinear and symmetric with respect to the real scalar product $\mathrm{Re}(u|v)$. In fact, it satisfies

$$(H_u(h_1)\,|\,h_2) = (H_u(h_2)\,|\,h_1).$$

Moreover, H_u is a Hilbert-Schmidt operator with

$$\mathrm{Tr}(H_u^2) = \sum_{n=0}^{\infty}(n+1)|\hat{u}(n)|^2.$$

Given $b \in L^\infty(\mathbb{S}^1)$, a Toeplitz operator $T_b : L^2_+ \to L^2_+$ is defined by

$$T_b(h) = \Pi(bh).$$

T_b is $\mathbb{C}$-linear, bounded, and self-adjoint if and only if b is real valued.

The cubic Szegő equation was proved to admit two Lax pairs as follows:

Theorem 2.1 [Gérard and Grellier 2010, Theorem 3.1]. *Let $u \in C(\mathbb{R}, H^s(\mathbb{S}^1))$ for some $s > \frac{1}{2}$. The cubic Szegő equation*

$$i\,\partial_t u = \Pi(|u|^2 u) \tag{2-1}$$

has two Lax pairs (H_u, B_u) and (K_u, C_u), namely, if u solves (2-1), then

$$\frac{dH_u}{dt} = [B_u, H_u], \quad \frac{dK_u}{dt} = [C_u, K_u], \tag{2-2}$$

where

$$B_u = \frac{i}{2}H_u^2 - iT_{|u|^2}, \quad K_u := T_z^* H_u, \quad C_u = \frac{i}{2}K_u^2 - iT_{|u|^2}.$$

Corollary 2.2. *The perturbed Szegő equation (1-6) with $\alpha \neq 0$ still has one Lax pair (K_u, C_u).*

Proof of Corollary 2.2. We need an identity from [Gérard and Grellier 2013, Lemma 1]:

$$H_{\Pi(|u|^2u)} = T_{|u|^2}H_u + H_u T_{|u|^2} - H_u^3. \tag{2-3}$$

Using (1-6) and (2-3),

$$\frac{dH_u}{dt} = H_{-i\,\Pi(|u|^2u) - i\alpha(u|1)} = -i\,(T_{|u|^2}H_u + H_u T_{|u|^2} - H_u^3) - i\alpha(u\,|\,1)H_1.$$

Using the antilinearity of H_u, we deduce that

$$\frac{dH_u}{dt} = [B_u, H_u] - i\alpha(u\,|\,1)H_1, \tag{2-4}$$

which means that (H_u, B_u) is no longer a Lax pair. Fortunately, we have $T_z^* H_1 = 0$, which leads to the identity

$$\frac{dK_u}{dt} = [C_u, K_u]. \qquad \square$$

An important consequence of this Lax pair structure is the existence of finite dimensional submanifolds of $L^2_+(\mathbb{S}^1)$, which are invariant by the flow of (1-6). To describe these manifolds, Gérard and Grellier [2010, Appendix 4] proved a Kronecker-type theorem to the effect that the Hankel operator H_u is of finite rank k if and only if u is a rational function of the complex variable z with no poles in the unit disc and of the form $u(z) = A(z)/B(z)$ with $A \in \mathbb{C}_{k-1}[z]$, $B \in \mathbb{C}_k[z]$, $B(0) = 1$, $\deg A = k-1$ or $\deg B = k$, A and B having no common factors, and $B(z) \neq 0$ if $|z| \leq 1$. In fact, we can prove a similar theorem for our case.

Definition 2.3. Letting k be a positive integer, we define

$$\mathcal{L}(k) := \{u \in H^{1/2}_+(\mathbb{S}^1) : \operatorname{rk} K_u = k\}. \tag{2-5}$$

Due to the Lax pair structure, the manifolds $\mathcal{L}(k)$ are invariant by the flow.

Theorem 2.4. *The elements of $\mathcal{L}(k)$ are the rational functions $u = \dfrac{A(z)}{B(z)}$, where*

$$A, B \in \mathbb{C}_k[z], \quad A \wedge B = 1, \quad \deg A = k \text{ or } \deg B = k, \quad B^{-1}(\{0\}) \cap \overline{D} = \varnothing. \tag{2-6}$$

Here $A \wedge B = 1$ means A and B have no common factors.

Proof. Gérard and Grellier [2010, Appendix 4] proved that

$$\begin{aligned}\mathcal{M}(k+1) &= \{u : \operatorname{rk} H_u = k+1\} \\ &= \left\{ u(z) = \frac{A(z)}{B(z)} : A \in \mathbb{C}_k[z], B \in \mathbb{C}_{k+1}[z], B(0) = 1, \right. \\ &\qquad \left. \deg A = k \text{ or } \deg B = k+1,\ A \wedge B = 1,\ B^{-1}(0) \cap \overline{D} = \varnothing \right\}.\end{aligned}$$

For $u \in \mathcal{M}(k+1)$ we have $\dim \operatorname{Im} H_u = k+1$. Then $u, T_z^* u, \ldots, (T_z^*)^{k+1} u$ are linearly dependent, that is, there exist C_ℓ, not all zero, such that $\sum_{\ell=0}^{k+1} C_\ell (T_z^*)^\ell u = 0$. We get

$$\sum_{\ell=0}^{k+1} C_\ell \hat{u}(\ell + n) = 0 \quad \text{for all } n \geq 0.$$

This is a recurrence equation for the sequence $\hat{u}$, and can be solved by using linear algebra. Define

$$P(X) = \sum_{\ell=0}^{k+1} C_\ell X^\ell = C \prod_{p \in \mathcal{P}} (X - p)^{m_p},$$

where $\mathcal{P} = \{p \in \mathbb{C} : P(p) = 0\}$ and m_p is the multiplicity of p. Then $(\hat{u}(n))_{n\geq 0}$ is a linear combination of the sequences

$$n^\ell p^{n-\ell},\ p \neq 0,\ 0 \leq \ell \leq m_p - 1 \quad \text{and} \quad \delta_{nm},\ p = 0,\ 0 \leq m \leq m_0 - 1.$$

Recall that

$$u(z) = \sum_{n\geq 0} \hat{u}(n) z^n \quad \text{for } |z| < 1.$$

Thus u is a linear combination of terms $\dfrac{1}{(1-pz)^{\ell+1}}$ with $0 < |p| < 1$ and $0 \leq \ell \leq m_p - 1$, and terms z^ℓ for $0 \leq \ell \leq m_0 - 1$.

Consequently, $u(z) = A(z)/B(z)$ with

$$\begin{aligned} &\deg A \leq k,\ \deg B = k+1 && \text{if } 0 \notin \mathcal{P},\\ &\deg A = k,\ \deg B \leq k && \text{if } 0 \in \mathcal{P}. \end{aligned}$$

But $0 \in \mathcal{P}$ is equivalent to $1 \in \operatorname{Im} H_u$, or again to $\ker K_u \cap \operatorname{Im} H_u \neq \{0\}$, since $K_u = T_z^* H_u$ and $\operatorname{rk} H_u - 1 \leq \operatorname{rk} K_u \leq \operatorname{rk} H_u$. For $u \in \mathcal{L}(k)$ we have $\operatorname{rk} K_u = k$. Thus $u = A(z)/B(z)$ with

$$\begin{aligned} &\deg A \leq k-1,\ \deg B = k && \text{if } \operatorname{rk} H_u = \operatorname{rk} K_u = k,\\ &\deg A = k, \quad \deg B \leq k && \text{if } \operatorname{rk} H_u = \operatorname{rk} K_u + 1 = k+1. \end{aligned}$$

The proof of the converse is similar. It follows that $\mathcal{L}(k) = \{u : \operatorname{rk} K_u = k+1\}$ contains precisely the quotients $u = A/B$, with A and B as in (2-6). □

3. Proof of the main theorem

We will now prove that the α-Szegő equation (1-6) has a large-time blowup as in Theorem 1.3. We also give an example to describe this phenomenon in terms of energy transfer to high frequencies. We start by proving the boundedness of the Wiener norm as in Theorem 1.2.

Proposition 3.1. *Assume $u_0 \in H^s_+(\mathbb{S}^1)$ with $s > 1$ and let u be the corresponding unique solution of (1-6). Then*

$$\|u(t)\|_W \leq C_s \|u_0\|_{H^s} \quad \textit{for all } t \in \mathbb{R}.$$

Proof. By Peller's theorem [2003], the regularity of u ensures that H_u is trace class and the trace norm of H_u is equivalent to the $B^1_{1,1}$ norm of u. Recall the definition of $B^s_{p,q}(\mathbb{S}^1)$.

Let $\chi \in C^\infty(\mathbb{R}^+)$ satisfy $\chi|_{t<1}(t) = 1$, $\chi|_{t>2}(t) = 0$, $0 \leq \chi \leq 1$. Set ψ as $\psi_0(t) = 1 - \chi(t)$, $\psi_j(t) = \chi(2^{-j+1}t) - \chi(2^{-j}t)$. Define the operator Δ_j for $f \in \mathcal{D}'(\mathbb{S}^1)$ as

$$\Delta_j f = \sum_{k\in\mathbb{Z}} \psi_j(k) \hat{f}(k) e^{ik\theta}.$$

Then the Besov space is defined as

$$B^s_{p,q}(\mathbb{S}^1) := \{u \in \mathcal{D}'(\mathbb{S}^1) : 2^{js} \|\Delta_j f\|_{L^p} \in l^q_j,\ 1 \leq p, q \leq +\infty,\ 0 \leq j \leq +\infty\},$$

with norm

$$\|u\|_{B^s_{p,q}(\mathbb{S}^1)} = \left(\sum_{j=0}^{+\infty} (2^{js}\|\Delta_j f\|_{L^p})^q\right)^{1/q}.$$

Observe that there exist $C, C_s > 0$ such that

$$\begin{aligned}\|u\|_{B^1_{1,1}} &= \sum_{j=0}^{+\infty} 2^j \|\Delta_j u\|_{L^1} \le C \sum_{j=0}^{+\infty} 2^j \|\Delta_j u\|_{L^2} \\ &\le C\left(\sum_{j=0}^{+\infty} 2^{2js}\|\Delta_j u\|_{L^2}^2\right)^{1/2}\left(\sum_{j=0}^{+\infty} 2^{2j(1-s)}\right)^{1/2} \le C_s\|u\|_{H^s} \quad \text{for all } s>1. \end{aligned} \tag{3-1}$$

So, for $u \in H^s$ with $s > 1$, H_u is trace class, and

$$\operatorname{Tr}(|H_u|) \le C_s\|u\|_{H^s}.$$

Since $K_u = T_z^* H_u$, we have $K_u^2 = H_u^2 - (\cdot|u)u$, and so $\operatorname{Tr}(|K_u|) \le \operatorname{Tr}(|H_u|)$. Due to the Lax pair structure, we conclude that $K_{u(t)}$ is isospectral to K_{u_0}, an in particular $\operatorname{Tr}(|K_{u(t)}|) = \operatorname{Tr}(|K_{u_0}|)$. Therefore

$$\operatorname{Tr}(|K_{u(t)}|) \le C_s\|u_0\|_{H^s}.$$

Since $\|u\|_W = |\hat{u}(0)| + \sum_{n\ge1} |\hat{u}(n)|$ and $|\hat{u}(0)| \le \|u\|_{L^2}$, we just need to show that

$$\sum_{n\ge1} |\hat{u}(n)| \le C \operatorname{Tr}(|K_u|).$$

Let $\{e_n\}$ be an orthonormal basis of L^2_+. Then, for any bounded operator B,

$$\sum_n |(K_u e_n \,|\, Be_n)| \le \operatorname{Tr}(|K_u|)\|B\|.$$

Then we see that $\sum_{n\ge1} |\hat{u}(2n)| + \sum_{n\ge1} |\hat{u}(2n+1)| \le \operatorname{Tr}(|K_u|)$ by taking $B = T_z$ and $B = \mathrm{Id}$. This completes the proof. □

Remark 3.2. In fact, to prove the global well-posedness, it is natural to use the Brezis–Gallouët type estimate from [Gérard and Grellier 2010, Appendix 2]: for $s > \frac{1}{2}$,

$$\|u\|_W \le C_s\|u\|_{H^{1/2}}\left[\log\left(1 + \frac{\|u\|_{H^s}}{\|u\|_{H^{1/2}}}\right)\right]^{\frac{1}{2}}.$$

This leads to a growth doubly exponential on time for the Sobolev norm of u. Fortunately, by the estimate in Proposition 3.1, we know the H^s norm of the solutions will admit an exponential on time upper bound for $s > 1$ (see Remark 1.4).

Now, let us start the large-time blowup theorem.

Theorem 3.3. *For $\alpha > 0$, we consider the solution of the Szegő equation (1-6) with initial data $u_0 \in \mathcal{L}(1)$.*

(1) *If the trajectory issued from u_0 is not relatively compact in $\mathcal{L}(1)$, then*

$$\left| b + \frac{\bar{p}c}{1-|p|^2} \right| = \sqrt{\alpha}, \tag{3-2}$$

or, equivalently,

$$E_\alpha = \tfrac{1}{4}Q^2 + \tfrac{1}{2}\alpha Q. \tag{3-3}$$

(2) *If (3-2) holds, then*

$$\|u(t)\|_{H^s} \simeq e^{C_{\alpha,s}|t|}, \quad s > \tfrac{1}{2},\ C_{\alpha,s} > 0,\ |t| \to \infty. \tag{3-4}$$

Thus the equality (3-3), which is invariant by the flow, is a necessary and sufficient condition to cause large-time blowup.

Proof. First, since the trajectory of the solution is not relatively compact in $\mathcal{L}(1)$, the level set $L(u_0) := \{u \in \mathcal{L}(1) : Q(u) = Q(u_0),\ M(u) = M(u_0),\ E_\alpha(u) = E_\alpha(u_0)\}$ is not compact in $\mathcal{L}(1)$.

We rewrite $u \in \mathcal{L}(1)$ as

$$u = b + \frac{cz}{1-pz}.$$

Then the conservation laws under the coordinates b, p, c are given as

$$Q = \|u\|_{L^2}^2 = \frac{|c|^2}{1-|p|^2} + |b|^2,$$

$$M = (Du\,|\,u) = \frac{|c|^2}{(1-|p|^2)^2},$$

$$E_\alpha = \tfrac{1}{4}\|u\|_{L^4}^4 + \tfrac{1}{2}\alpha|(u|1)|^2 = \frac{1}{4}\left[|b|^4 + \frac{4|b|^2|c|^2}{1-|p|^2} + \frac{|c|^4(1+|p|^2)}{(1-|p|^2)^3} + \frac{4|c|^2\,\mathrm{Re}(bp\bar{c})}{(1-|p|^2)^2}\right] + \tfrac{1}{2}\alpha|b|^2.$$

Now, $u \in \mathcal{L}(1)$ stays in a compact of $\mathcal{L}(1)$ if and only if $|b| \le C$, $1/C \le |c| \le C$, and $|p| \le k < 1$ with some constant C and k. Otherwise, due to the formulas of mass Q and momentum M, there exist $t_n \to \infty$ such that $|c(t_n)|$ and $1-|p(t_n)|^2$ tend to 0 at the same order. Using the formula of Q and E_α, we have

$$|b(t_n)|^2 \to Q, \quad \tfrac{1}{4}|b(t_n)|^4 + \tfrac{1}{2}\alpha|b(t_n)|^2 \to E_\alpha.$$

Since the limit should be unique,

$$E_\alpha = \tfrac{1}{4}Q^2 + \tfrac{1}{2}\alpha Q.$$

Using the formula of mass and energy, (3-3) can be rewritten under coordinates of b, p, c as

$$|b|^2 + \frac{|c|^2|p|^2}{(1-|p|^2)^2} + 2\,\mathrm{Re}\,\frac{bp\bar{c}}{1-|p|^2} = \alpha.$$

Simplifying the left hand side, we get

$$\left| b + \frac{\bar{p}c}{1-|p|^2} \right| = \sqrt{\alpha}.$$

Now we turn to proving that (3-2) is sufficient to cause the exponential growth of Sobolev norms. Writing, as before,

$$u(t) = b(t) + \frac{c(t)z}{1-p(t)z},$$

the terms $\partial_t u$, $\Pi(|u|^2u)$, $(u|1)$ can be represented as linear combinations of 1, $\frac{z}{1-pz}$, $\frac{z^2}{(1-pz)^2}$:

$$\begin{cases} \partial_t u = \partial_t b + \partial_t c\dfrac{z}{1-pz} + \partial_t p\dfrac{z^2}{(1-pz)^2}, \\ \Pi(|u|^2u) = |b|^2 b + \dfrac{2b|c|^2}{1-|p|^2} + \dfrac{|c|^2 c\bar{p}}{1-|p|^2} \\ \qquad\qquad + \left[2|b|^2 c + \dfrac{2b|c|^2 p}{1-|p|^2} + \dfrac{1+|p|^2}{1-|p|^2}|c|^2 c\right]\dfrac{z}{1-pz} + \left[c^2\bar{b} + \dfrac{|c|^2 cp}{1-|p|^2}\right]\dfrac{z^2}{(1-pz)^2}, \\ (u\,|\,1) = b. \end{cases}$$

Then (1-6) reads

$$\begin{cases} i\,\partial_t b = |b|^2 b + \dfrac{2b|c|^2}{1-|p|^2} + \dfrac{|c|^2 c\bar{p}}{(1-|p|^2)^2} + \alpha b, \\ i\,\partial_t c = 2|b|^2 c + \dfrac{2b|c|^2 p}{1-|p|^2} + \dfrac{|c|^2 c}{(1-|p|^2)^2}, \\ i\,\partial_t p = c\bar{b} + \dfrac{|c|^2 p}{1-|p|^2}. \end{cases} \tag{3-5}$$

Using the second equation of (3-5), we obtain

$$\frac{d|c|}{dt} = \frac{2|c|}{1-|p|^2}\,\mathrm{Im}(bp\bar{c}). \tag{3-6}$$

This equality together with (3-2) gives us

$$\begin{aligned} \left(\frac{d|c|}{|c|dt}\right)^2 &= \frac{4(\mathrm{Im}(bp\bar{c}))^2}{(1-|p|^2)^2} = \frac{4|bp\bar{c}|^2}{(1-|p|^2)^2} - \frac{4(\mathrm{Re}(bp\bar{c}))^2}{(1-|p|^2)^2} \\ &= \frac{4|bp\bar{c}|^2}{(1-|p|^2)^2} - \left[\alpha - |b|^2 - \frac{|c|^2|p|^2}{(1-|p|^2)^2}\right]^2 = \frac{4|bp\bar{c}|^2}{(1-|p|^2)^2} - \left[\alpha - |b|^2 - \frac{|c|^2}{(1-|p|^2)^2} + \frac{|c|^2}{1-|p|^2}\right]^2 \\ &= \frac{4|bp\bar{c}|^2}{(1-|p|^2)^2} - \left[\alpha - Q - M + 2\frac{|c|^2}{1-|p|^2}\right]^2 \\ &= \frac{4|bp\bar{c}|^2}{(1-|p|^2)^2} - \frac{4|c|^4}{(1-|p|^2)^2} - \frac{4|c|^2}{1-|p|^2}\left[\alpha - |b|^2 - \frac{|c|^2}{(1-|p|^2)^2} - \frac{|c|^2}{1-|p|^2}\right] - (\alpha - Q - M)^2 \\ &= \frac{4|b|^2|c|^2}{(1-|p|^2)^2} + \frac{4|c|^4}{(1-|p|^2)^3} - \alpha\frac{4|c|^2}{1-|p|^2} - (\alpha - Q - M)^2 \\ &= 4\left(|b|^2 + \frac{|c|^2}{1-|p|^2}\right)\frac{|c|^2}{(1-|p|^2)^2} - \alpha\frac{4|c|^2}{1-|p|^2} - (\alpha - Q - M)^2 \\ &= 4QM - 4\alpha\sqrt{M}|c| - (\alpha - Q - M)^2. \end{aligned}$$

Thus

$$\left(\frac{d\log|c|}{dt}\right)^2 = -4\alpha\sqrt{M}|c| + 4QM - (\alpha - M - Q)^2.$$

Since $0 \le |c| \le 1$, it follows that $c_{\alpha,M,Q} \le \left(\frac{d\log|c|}{dt}\right)^2 \le C_{\alpha,M,Q}$, which leads to exponential decay in time for $|c|$:

$$|c|(t) \simeq |c(0)|e^{-C|t|}$$

with the positive constant C depending on α and M, Q.

Notice that $\hat{u}(k,t) = cp^{k-1}$ for $k \ge 1$. Using Fourier expansion, we obtain, as $|p|$ approaches 1,

$$\|u\|_{H^s}^2 \simeq \frac{|c|^2}{(1-|p|^2)^{2s+1}}.$$

Since $M(u) = |c|^2/(1-|p|^2)^2 =$ constant, we get $\|u\|_{H^s}^2 \simeq |c|^{-(2s-1)} \simeq e^{C(2s-1)|t|}$, which has an exponential growth as $s > \frac{1}{2}$. This completes the proof. □

Corollary 3.4. *We do not have the growth of H^s norms for small data in $\mathcal{L}(1)$. In other words, if $\|u(0)\|_{H_+^{1/2}} \ll \sqrt{\alpha}$, the higher Sobolev norm will never grow to infinity.*

Proof. $\|u(0)\|_{H_+^{1/2}} \ll \sqrt{\alpha}$. Then

$$\left|b + \frac{c\bar{p}}{1-|p|^2}\right| \le \sqrt{Q} + \sqrt{M} \lesssim \|u(0)\|_{H_+^{1/2}} \ll \sqrt{\alpha}.$$

According to the necessary and sufficient condition (3-2), there is no norm explosion. □

Remark 3.5. Consider a family of Cauchy data given by

$$u_0^\varepsilon = z + \varepsilon, \quad \varepsilon \in \mathbb{C} \text{ and } \varepsilon \ne \sqrt{\alpha}.$$

For the case $\alpha = 0$, Gérard and Grellier got the following instability of H^s norms:

$$\|u^\varepsilon(t^\varepsilon)\|_{H^s} \simeq \varepsilon^{-(2s-1)}, \quad s > \tfrac{1}{2}.$$

However, we do not have such an instability result for $\alpha > 0$. In fact, using Theorem 3.3, we know there exists a constant $C = C(\alpha)$ such that

$$\sup_{\varepsilon \ne \sqrt{\alpha}} \sup_{t\in\mathbb{R}} \|u^\varepsilon(t)\|_{H^s} < C.$$

Now we give an example to display the energy cascade in Theorem 3.3.

Theorem 3.6. *Given $\alpha > 0$,*

$$\begin{cases} i\partial_t u = \Pi(|u|^2 u) + \alpha(u|1), \\ u|_{t=0} = z + \sqrt{\alpha}, \qquad\qquad z \in \mathbb{S}^1. \end{cases} \tag{3-7}$$

For all $s > \frac{1}{2}$, the above equation is globally well posed in H^s and the solution satisfies

$$\|u(t)\|_{H^s} \simeq e^{(2s-1)\sqrt{\alpha}t}, \quad t \to \infty.$$

Proof. Since $u_0 = z + \sqrt{\alpha}$, the conserved quantities are $Q = 1+\alpha$, $M = 1$, $E_\alpha = \frac{1}{4}(1+\alpha)(1+3\alpha)$. Thus $u_0 \in \mathscr{L}(1)$. So, by the proof of Theorem 3.3,

$$\left(\frac{d}{dt}|c|\right)^2 = 4\alpha|c|^2(1-|c|).$$

Together with the initial condition $|c|(0) = 1$, we get, for $t > 0$ (same strategy for $t < 0$),

$$\frac{d}{dt}|c| = -2\sqrt{\alpha}|c|\sqrt{1-|c|}, \tag{3-8}$$

and then

$$|c|(t) = \frac{4e^{2\sqrt{\alpha}t}}{(1+e^{2\sqrt{\alpha}t})^2}.$$

By (3-2), we get $\mathrm{Re}(bp\bar{c}) = |c|^2 - |c|$, and, by (3-6) and (3-8), we have $\mathrm{Im}(bp\bar{c}) = -\sqrt{\alpha}|c|\sqrt{1-|c|}$, so

$$bp\bar{c} = \mathrm{Re}(bp\bar{c}) + i\,\mathrm{Im}(bp\bar{c}) = |c|^2 - |c| - i\sqrt{\alpha}|c|\sqrt{1-|c|}.$$

The second equation of (3-5) can be simplified as follows:

$$\begin{cases} i\,\partial_t c = (1+2\alpha - 2i\sqrt{\alpha}\sqrt{1-|c|})c, \\ c(0) = 1. \end{cases}$$

Thus

$$c(t) = \frac{4e^{2\sqrt{\alpha}t}}{(1+e^{2\sqrt{\alpha}t})^2} e^{-i(1+2\alpha)t}. \tag{3-9}$$

Now we turn to calculating b and p. In fact, we only need to calculate their angles. Let us denote

$$b = |b|e^{i\theta(t)} = \sqrt{1+\alpha-|c|}e^{i\theta(t)}, \quad p = |p|e^{i\sigma(t)} = \sqrt{1-|c|}e^{i\sigma(t)}.$$

Then, using the differential equation on p, we get

$$\partial_t\sigma|p| = |c||p| + \mathrm{Re}(c\bar{b}e^{-i\sigma}) = |c||p| + \mathrm{Re}\left(\frac{c\bar{b}\,\bar{p}}{|p|}\right) = |c||p| + \frac{1}{|p|}(|c|^2 - |c|) = 0,$$

which means

$$\sigma(t) = \sigma(0).$$

Since

$$\begin{aligned} bp = \frac{c(bp\bar{c})}{|c|^2} &= (|c| - 1 - i\sqrt{\alpha}\sqrt{1-|c|})e^{-i(1+2\alpha)t} \\ &= \sqrt{(1+\alpha-|c|)(1-|c|)}\left(-\frac{\sqrt{1-|c|}}{\sqrt{1+\alpha-|c|}} - i\frac{\sqrt{\alpha}}{\sqrt{1+\alpha-|c|}}\right)e^{-i(1+2\alpha)t}, \\ e^{i(\theta+\sigma)} &= \left(-\frac{\sqrt{1-|c|}}{\sqrt{1+\alpha-|c|}} - i\frac{\sqrt{\alpha}}{\sqrt{1+\alpha-|c|}}\right)e^{-i(1+2\alpha)t}, \end{aligned}$$

and $e^{i\theta(0)} = 1$, we get

$$e^{i\sigma(t)} = e^{i\sigma(0)} = e^{i(\sigma(0)+\theta(0))} = -i.$$

Then

$$e^{i\theta(t)} = \left(-i\frac{\sqrt{1-|c|}}{\sqrt{1+\alpha-|c|}} + \frac{\sqrt{\alpha}}{\sqrt{1+\alpha-|c|}}\right)e^{-i(1+2\alpha)t}.$$

Finally, we have

$$\begin{aligned} p(t) &= -i\sqrt{1-|c|} = -i\frac{e^{2\sqrt{\alpha}t}-1}{e^{2\sqrt{\alpha}t}+1}, \\ b(t) &= \left(\sqrt{\alpha} - i\frac{e^{2\sqrt{\alpha}t}-1}{e^{2\sqrt{\alpha}t}+1}\right)e^{-i(1+2\alpha)t}. \end{aligned} \tag{3-10}$$

Now we get the explicit formula for the solution $u(t) = b(t) + c(t)z/(1-p(t)z)$:

$$\begin{cases} b(t) = \left(\sqrt{\alpha} - i\dfrac{e^{2\sqrt{\alpha}t}-1}{e^{2\sqrt{\alpha}t}+1}\right)e^{-i(1+2\alpha)t}, \\ c(t) = \dfrac{4e^{2\sqrt{\alpha}t}}{(1+e^{2\sqrt{\alpha}t})^2}e^{-i(1+2\alpha)t}, \\ p(t) = -i\dfrac{e^{2\sqrt{\alpha}t}-1}{e^{2\sqrt{\alpha}t}+1}. \end{cases} \tag{3-11}$$

In this case, $M(u) = |c|^2/(1-|p|^2)^2 = 1$ and we get, for $t \to +\infty$,

$$\|u(t)\|_{H^s}^2 \simeq |c|^{-(2s-1)} \simeq Ce^{2(2s-1)\sqrt{\alpha}t}. \qquad \square$$

Remark 3.7. One can illustrate this instability of Sobolev norms from the viewpoint of transfer of energy to high frequencies. The Fourier coefficients for $u = b + cz/(1-pz)$ are

$$\hat{u}(k) = c(t)p(t)^{k-1} \quad \text{for all } k \geq 1.$$

Then

$$M(u) = 1 = \sum_{k\geq 1}|k||\hat{u}(k)|^2 = \sum_{k\geq 1}|k||c(t)|^2|p(t)|^{2(k-1)}.$$

With (3-11), we have

$$\sum_{k\geq 1}\left|\frac{1-e^{-2\sqrt{\alpha}t}}{1+e^{-2\sqrt{\alpha}t}}\right|^{2k}\frac{16|k|}{|(1+e^{-2\sqrt{\alpha}t})(1-e^{-2\sqrt{\alpha}t})|^2} = 1.$$

As $t \to \infty$, we get

$$\sum_{k\geq 1} 4|k|e^{-2\sqrt{\alpha}t}\exp\left(-4|k|e^{-2\sqrt{\alpha}t}\right) \sim \tfrac{1}{4},$$

so the main part of the summation is on the ks satisfying

$$|k| \sim e^{2\sqrt{\alpha}t}.$$

So as time increases, the main part of the energy concentrates on the Fourier modes as large as $e^{2\sqrt{\alpha}t}$.

On the other hand, from the viewpoint of the space variable, we find that as time grows to infinity, the energy will concentrate on one point. In fact, rewriting $z = e^{ix}$, we get

$$\left| u(t,x) - \sqrt{\alpha} - i\,\frac{1-e^{-2\sqrt{\alpha}t}}{1+e^{-2\sqrt{\alpha}t}} \right| = \frac{|c(t)|}{|1-p(t)z|} = \frac{1-|p(t)|^2}{|1-p(t)z|} \sim \frac{1-|p(t)|}{|1-p(t)z|} \sim \frac{1}{\sqrt{2(e^{4\sqrt{\alpha}t}-1)(1-\sin x)+4}},$$

which tends to 0 as $t \to \infty$ if and only if $x \neq \pi/2$. Therefore, as time tends to infinity, the value of $|u|$ will concentrate on the point $i \in \mathbb{S}^1$.

This example shows that the radius of analyticity of the solution of (1-6) may decay exponentially. This shows the optimality of the result in [Gérard et al. 2013].

Now, let us turn to the case $\alpha < 0$.

Theorem 3.8. *In the case $\alpha < 0$, for any given initial data $u_0 \in \mathscr{L}(1)$, let $u = (az+b)/(1-pz)$ be the corresponding solution of (1-6). Then there exists a constant $C = C(\alpha)$ such that, for all t,*

$$\|u(t)\|_{H^s} < C, \quad s \geq \tfrac{1}{2},$$

where the constant $C > 0$ is uniform for u_0 in a compact subset of $\mathscr{L}(1)$.

Proof. Assume for a contradiction that $u(t_n)$ leaves any compact subset of $\mathscr{L}(1)$. Then Theorem 3.3 leads to (3-3), or equivalently to the equality

$$\|u_0\|_{L^2}^4 - \|u_0\|_{L^4}^4 = 2\alpha(|(u_0\,|\,1)|^2 - \|u_0\|_{L^2}^2).$$

Via the Cauchy–Schwarz inequality and $\alpha < 0$, we get

$$\|u_0\|_{L^2} = \|u\|_{L^4} \quad \text{and} \quad |(u_0\,|\,1)| = \|u_0\|_{L^2}.$$

Then u_0 is a constant, which contradicts the fact that $u_0 \in \mathscr{L}(1)$. □

4. Further studies and open problems

In this paper, we just considered the data on the (complex) three-dimensional manifold

$$\mathscr{L}(1) := \{u : \operatorname{rk} K_u = 1\}.$$

It is of course natural to consider the higher-dimensional case, which will probably be much more complicated. Since we also have enough conservation laws for the case $\operatorname{rk} K_u = 2$, we have a conjecture that the system stays completely integrable for $\operatorname{rk} K_u \geq 2$. It would be interesting to know how the results of this paper extend to this bigger phase space. In particular, do small data generate large-time blowup of high Sobolev norms?

References

[Bourgain 1996] J. Bourgain, "On the growth in time of higher Sobolev norms of smooth solutions of Hamiltonian PDE", *Internat. Math. Res. Notices* **1996**:6 (1996), 277–304. MR 97k:35016 Zbl 0934.35166

[Colliander et al. 2010] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, "Transfer of energy to high frequencies in the cubic defocusing nonlinear Schrödinger equation", *Invent. Math.* **181**:1 (2010), 39–113. MR 2011f:35320 Zbl 1197.35265

[Gérard and Grellier 2010] P. Gérard and S. Grellier, "The cubic Szegő equation", *Ann. Sci. Éc. Norm. Supér.* (4) **43**:5 (2010), 761–810. MR 2012b:37188 Zbl 1228.35225

[Gérard and Grellier 2012a] P. Gérard and S. Grellier, "Effective integrable dynamics for a certain nonlinear wave equation", *Anal. PDE* **5**:5 (2012), 1139–1155. MR 3022852 Zbl 1268.35013

[Gérard and Grellier 2012b] P. Gérard and S. Grellier, "Invariant tori for the cubic Szegő equation", *Invent. Math.* **187**:3 (2012), 707–754. MR 2944951 Zbl 1252.35026

[Gérard and Grellier 2013] P. Gérard and S. Grellier, "An explicit formula for the cubic Szegő equation", preprint, 2013. To appear in *Trans. Amer. Math. Soc.* arXiv 1304.2619

[Gérard et al. 2013] P. Gérard, Y. Guo, and E. S. Titi, "On the radius of analyticity of solutions to the cubic Szegő equation", preprint, 2013. To appear in *Ann. Inst. H. Poincaré Anal. Non Linéaire*. arXiv 1303.6148v2

[Guardia and Kaloshin 2012] M. Guardia and V. Kaloshin, "Growth of Sobolev norms in the cubic defocusing nonlinear Schrödinger equation", preprint, 2012. arXiv 1205.5188

[Hani 2014] Z. Hani, "Long-time instability and unbounded Sobolev orbits for some periodic nonlinear Schrödinger equations", *Arch. Ration. Mech. Anal.* **211**:3 (2014), 929–964. MR 3158811

[Hani et al. 2013] Z. Hani, B. Pausader, N. Tzvetkov, and N. Visciglia, "Modified scattering for the cubic Schrödinger equation on product spaces and applications", preprint, 2013. arXiv 1311.2275

[Peller 2003] V. V. Peller, *Hankel operators and their applications*, Springer, New York, 2003. MR 2004e:47040 Zbl 1030.47002

[Pocovnicu 2011a] O. Pocovnicu, "Explicit formula for the solution of the Szegő equation on the real line and applications", *Discrete Contin. Dyn. Syst.* **31**:3 (2011), 607–649. MR 2012h:35330 Zbl 1235.35263

[Pocovnicu 2011b] O. Pocovnicu, "Traveling waves for the cubic Szegő equation on the real line", *Anal. PDE* **4**:3 (2011), 379–404. MR 2012k:35521 Zbl 1270.35172

[Staffilani 1997] G. Staffilani, "On the growth of high Sobolev norms of solutions for KdV and Schrödinger equations", *Duke Math. J.* **86**:1 (1997), 109–142. MR 98b:35192 Zbl 0874.35114

Received 19 Jul 2013. Accepted 28 Apr 2014.

HAIYAN XU: haiyan.xu@math.u-psud.fr
Laboratoire de Mathématique d'Orsay, Université Paris-Sud (XI), 91405 Paris Orsay, France

ANALYSIS AND PDE
Vol. 7, No. 3, 2014
dx.doi.org/10.2140/apde.2014.7.733

A GEOMETRIC TANGENTIAL APPROACH TO SHARP REGULARITY FOR DEGENERATE EVOLUTION EQUATIONS

EDUARDO V. TEIXEIRA AND JOSÉ MIGUEL URBANO

That the weak solutions of degenerate parabolic PDEs modelled on the inhomogeneous p-Laplace equation

$$u_t - \operatorname{div}(|\nabla u|^{p-2}\nabla u) = f \in L^{q,r}, \quad p > 2$$

are $C^{0,\alpha}$, for some $\alpha \in (0, 1)$, has been known for almost 30 years. What was hitherto missing from the literature was a precise and sharp knowledge of the Hölder exponent α in terms of p, q, r and the space dimension n. We show in this paper that

$$\alpha = \frac{(pq - n)r - pq}{q[(p-1)r - (p-2)]}$$

using a method based on the notion of geometric tangential equations and the intrinsic scaling of the p-parabolic operator. The proofs are flexible enough to be of use in a number of other nonlinear evolution problems.

1. Introduction

The understanding of the local behaviour of solutions to singular and degenerate parabolic equations has witnessed an impressive progress in the last three decades. At the heart of most developments lies a single unifying idea, namely that regularity results have to be interpreted in an intrinsic geometric configuration, a sort of signature to each particular PDE. The pioneering work of DiBenedetto [1993] was the starting point to a theory that has, in many aspects, reached its maturity (see [DiBenedetto et al. 2012] and [Urbano 2008] for recent accounts).

A central aspect in this endeavour has always been the Hölder continuity of bounded weak solutions, which ultimately follows from Harnack-type inequalities. Although powerful, this approach only provides qualitative estimates that depend solely on the structure of the equations and thus hold in a very general setting. The quest for precise, quantitative derivations of the Hölder exponent has hitherto eluded the community, the only exception being the two-dimensional result in [Iwaniec and Manfredi 1989] concerning p-harmonic functions. This type of quantitative information, apart from its own intrinsic value, plays an important role in the analysis of a number of qualitative issues for parabolic PDEs, such as blow-up analysis, Liouville type results, free boundary problems, and so forth.

MSC2010: 35K55, 35K65, 35B65.
Keywords: degenerate parabolic equations, sharp Hölder regularity, tangential equations, intrinsic scaling.

The main goal of this paper is to fill this gap, bringing the theory to a new level of understanding. We show that weak solutions of degenerate p-parabolic equations whose prototype is

$$u_t - \operatorname{div}(|\nabla u|^{p-2}\nabla u) = f \in L^{q,r}, \quad p \geq 2, \tag{1}$$

are locally of class $C^{0,\alpha}$ in space, with

$$\alpha := \frac{(pq-n)r - pq}{q[(p-1)r - (p-2)]}$$

a precise and sharp expression for the Hölder exponent in terms of p, the integrability of the source and the space dimension n. We also show that u is of class $C^{0,\alpha/\theta}$ in time, where θ is the α-interpolation between 2 and p. What makes the parabolic case more delicate to analyse is the inhomogeneity in the equation, the fact that it scales differently with respect to space and time. It is worth stressing that the integrability in time (respectively, in space) of the source affects the regularity in space (respectively, in time) of the solution.

To highlight the extent to which our result is sharp, we project it into the state of the art of the theory. For the linear case $p = 2$, we obtain

$$\alpha = 1 - \left(\frac{2}{r} + \frac{n}{q} - 1\right),$$

which is the optimal Hölder exponent for the nonhomogeneous heat equation, and is in accordance with estimates obtained by energy considerations. When $p \to \infty$, we have $\alpha \to 1^-$, which gives an indication of the expected locally Lipschitz regularity for the case of the parabolic infinity-Laplacian. When the source f is independent of time, or else bounded in time, that is $r = \infty$, we obtain

$$\alpha = \frac{pq-n}{q(p-1)} = \frac{p}{p-1} \cdot \frac{q - n/p}{q},$$

which is exactly the optimal exponent obtained in [Teixeira 2013] for the elliptic case. It might also be interesting to compare our optimal result with the estimates from [Misawa 2013, Section 4], and also with the continuity estimates on p-parabolic obstacle problems from [Kuusi et al. 2014].

Within the general theory of p-parabolic equations, our result reveals a surprising feature. From the applied point of view, it is relevant to know what is the effect on the diffusion properties of the model as we dim the exponent p. Naïve physical interpretations could indicate that the higher the value of p, the less efficient should the diffusion properties of the p-parabolic operator turn out to be, *i.e.*, one should expect a less efficient smoothness effect of the operator. For instance, this is verified in the sharp regularity estimate for p-harmonic functions in 2D [Iwaniec and Manfredi 1989]. On the contrary, our estimate implies that for p-parabolic inhomogeneous equations, the Hölder regularity theory improves as p increases. In fact, a direct computation shows

$$\operatorname{sign}(\partial_p \alpha(p, n, q, r)) = \operatorname{sign}(q(2-r) + nr) = +1,$$

in view of standard assumptions on the integrability exponents of the source term.

Although regularity estimates for degenerate evolution equations have been successfully obtained in great generality (see [Kinnunen and Lewis 2000; Acerbi and Mingione 2007]), explicit expressions for the Hölder exponent of continuity for weak solutions have only been known in the linear setting. For nonlinear equations, the classical tools from harmonic analysis, such as singular integrals, are precluded from being used and an entirely new approach is needed. The new estimates we obtain are striking in their simplicity but perhaps the most relevant contribution we offer is the technique employed. We develop a method based on the notion of geometric tangential equations, which explores the intrinsic scaling of the p-parabolic operator and the integrability of the forcing term. By means of appropriate scaled iterative arguments, we show that at each inhomogeneous equation there is a universal tangential space formed by $C^{0,1}$ in space and $C^{0,1/2}$ in time functions. The method then imports such regularity back to the original equation, properly corrected through the scaling used to access the tangential space. The method is new to the field and robust enough to be adapted to other evolutionary problems, as well as to a number of other issues in the theory.

2. Preliminary tools

Let $U \subset \mathbb{R}^n$ be open and bounded, and $T > 0$. We consider the space-time domain $U_T = U \times (0, T)$. We work with the prototype inhomogeneous equation

$$u_t - \operatorname{div}(|\nabla u|^{p-2}\nabla u) = f \quad \text{in } U_T, \tag{2}$$

with a source term $f \in L^{q,r}(U_T) \equiv L^r(0, T; L^q(U))$ satisfying

$$\frac{1}{r} + \frac{n}{pq} < 1 \tag{3}$$

and

$$\frac{2}{r} + \frac{n}{q} > 1. \tag{4}$$

The first assumption is the standard minimal integrability condition that guarantees the existence of bounded weak solutions, while (4) defines the borderline setting for optimal Hölder type estimates. For instance, when $r = \infty$, conditions (3) and (4) enforce

$$\frac{n}{p} < q < n,$$

which corresponds to the known range of integrability required in the elliptic theory for local $C^{0,\alpha}$ estimates to be available.

We start with the definition of weak solution to (2).

Definition 2.1. We say a function

$$u \in C_{\mathrm{loc}}(0, T; L^2_{\mathrm{loc}}(U)) \cap L^p_{\mathrm{loc}}(0, T; W^{1,p}_{\mathrm{loc}}(U))$$

is a weak solution to (2) if, for every compact $K \subset U$ and every subinterval $[t_1, t_2] \subset (0, T]$, there holds

$$\int_K u\varphi\,dx\Big|_{t_1}^{t_2} + \int_{t_1}^{t_2}\int_K \left\{-u\varphi_t + |\nabla u|^{p-2}\nabla u\cdot\nabla\varphi\right\} dx\,dt = \int_{t_1}^{t_2}\int_K f\varphi\,dx\,dt,$$

for all $\varphi \in H^1_{\mathrm{loc}}(0, T; L^2(K)) \cap L^p_{\mathrm{loc}}(0, T; W^{1,p}_0(K))$.

The following alternative definition makes use of the Steklov average of a function $v \in L^1(U_T)$, defined for $0 < h < T$ by

$$v_h := \begin{cases} \dfrac{1}{h}\displaystyle\int_t^{t+h} v(\cdot\,,\tau)\,d\tau & \text{if } t \in (0, T-h], \\ 0 & \text{if } t \in (T-h, T], \end{cases}$$

and circumvents the difficulties related to the low regularity in time. In fact, these difficulties are more of a technical nature since the time derivative u_t is shown in [Lindqvist 2008] to be an element of a certain Lebesgue space.

Definition 2.2. We say a function

$$u \in C_{\mathrm{loc}}(0, T; L^2_{\mathrm{loc}}(U)) \cap L^p_{\mathrm{loc}}(0, T; W^{1,p}_{\mathrm{loc}}(U))$$

is a weak solution to (2) if, for every compact $K \subset U$ and every $0 < t < T - h$, there holds

$$\int_{K\times\{t\}} \left\{(u_h)_t\varphi + (|\nabla u|^{p-2}\nabla u)_h\cdot\nabla\varphi\right\} dx = \int_{K\times\{t\}} f_h\varphi\,dx, \tag{5}$$

for all $\varphi \in W^{1,p}_0(K)$.

One key ingredient in our analysis is the following Caccioppoli-type energy estimate enjoyed by weak solutions of (2).

Lemma 2.3 (Caccioppoli estimate). *Let u be a weak solution to (2). Given $K \times [t_1, t_2] \subset U \times (0, T]$, there exists a constant C, depending only on n, p, $K \times [t_1, t_2]$ and $\|f\|_{L^{q,r}}$, such that*

$$\sup_{t_1<t<t_2}\int_K u^2\xi^p\,dx + \int_{t_1}^{t_2}\int_K |\nabla u|^p\xi^p\,dx\,dt \le \int_{t_1}^{t_2}\int_K |u|^p(\xi^p + |\nabla\xi|^p)\,dx\,dt + \int_{t_1}^{t_2}\int_K u^2\xi^{p-1}|\xi_t|\,dx\,dt + \|f\|_{q,r} \tag{6}$$

for every $\xi \in \mathcal{C}^\infty_0(K \times (t_1, t_2))$ such that $\xi \in [0, 1]$.

Proof. Choose $\varphi = u_h\xi^p$ as a test function in (5) and perform the usual combination of integrating in time, passing to the limit in $h \to 0$ and applying Young's inequality to derive the estimate. □

We finally recall that, if v is a function belonging to $L^p(Q)$, its averaged norm is

$$\|v\|_{p,\mathrm{avg},Q} := \left(\fint_Q |v|^p\,dx\,dt\right)^{1/p} = |Q|^{-1/p}\|v\|_{p,Q},$$

where, as usual, the integral average is defined by

$$\fint_A \psi = \frac{1}{|A|}\int_A \psi.$$

3. Sharp Hölder estimate

We start by fixing universal constants, that depend only on the data. The intrinsic exponent to (2), with $f \in L^{q,r}$, is

$$\alpha := \frac{(pq-n)r - pq}{q[(p-1)r-(p-2)]} = \frac{p\Big(1-\frac{1}{r}-\frac{n}{pq}\Big)}{\Big(\frac{2}{r}+\frac{n}{q}-1\Big)+p\Big(1-\frac{1}{r}-\frac{n}{pq}\Big)}, \tag{7}$$

which, in view of (3) and (4), satisfies $0<\alpha<1$. Next, let

$$\theta := \alpha + p - (p-1)\alpha = p-(p-2)\alpha = \alpha 2 + (1-\alpha)p. \tag{8}$$

Clearly $2<\theta<p$, since $0<\alpha<1$. For such θ, we define the intrinsic θ-parabolic cylinder

$$G_\tau := (-\tau^\theta, 0)\times B_\tau(0), \quad \tau>0.$$

We first establish a key compactness result that states that if the source term f has a small norm in $L^{q,r}$, then a solution u to (2) is close to a p-caloric function in an inner subdomain. It is worth comparing such a result with the A-caloric approximation lemma obtained in [Duzaar and Mingione 2005, Lemma 4.1].

Lemma 3.1 (approximation to p-caloric functions). *For every $\delta>0$, there exists $0<\epsilon\ll 1$, such that if $\|f\|_{L^{q,r}(G_1)}\le\epsilon$ and u is a local weak solution of (2) in G_1, with $\|u\|_{p,\mathrm{avg},G_1}\le 1$, then there exists a ϕ that is p-caloric in $G_{1/2}$ in the sense that*

$$\phi_t - \operatorname{div}(|\nabla\phi|^{p-2}\nabla\phi) = 0 \quad \text{in } G_{1/2}, \tag{9}$$

and moreover satisfies

$$\|u-\phi\|_{p,\mathrm{avg},G_{1/2}} \le \delta. \tag{10}$$

Proof. Suppose, for the sake of contradiction, that the thesis of the lemma fails. That is, assume, for some $\delta_0>0$, that there exists a sequence

$$(u^j)_j \in C_{\mathrm{loc}}(-1,0;L^2_{\mathrm{loc}}(B_1))\cap L^p_{\mathrm{loc}}(-1,0;W^{1,p}_{\mathrm{loc}}(B_1))$$

and a sequence $(f^j)_j \in L^{q,r}(G_1)$ such that

$$u^j_t - \operatorname{div}(|\nabla u^j|^{p-2}\nabla u^j) = f^j \quad \text{in } G_1, \tag{11}$$

$$\|u^j\|_{p,\mathrm{avg},G_1} \le 1, \tag{12}$$

$$\|f^j\|_{L^{q,r}(G_1)} \le 1/j, \tag{13}$$

but still, for any j and any p-caloric function ϕ in $G_{1/2}$,

$$\|u^j-\phi\|_{p,\mathrm{avg},G_{1/2}} > \delta_0. \tag{14}$$

Fix a cutoff function $\xi \in C_0^\infty(G_1)$, such that $\xi \in [0,1]$, $\xi \equiv 1$ in $G_{1/2}$ and $\xi \equiv 0$ near $\partial_p G_1$. From the Caccioppoli estimate, using the notation

$$V(I \times U) = L^\infty(I; L^2(U)) \cap L^p(I; W^{1,p}(U)),$$

we obtain

$$\begin{aligned}
\|u^j\|_{V(G_{1/2})} &\le \sup_{-1<t<0} \int_{B_1} (u^j)^2 \xi^p \, dx + \int_{-1}^0 \int_{B_1} |\nabla u^j|^p \xi^p \, dx \, dt \\
&\le \int_{-1}^0 \int_{B_1} \left\{ |u^j|^p (\xi^p + |\nabla \xi|^p) + (u^j)^2 \xi^{p-1} |\xi_t| \right\} dx \, dt + \|f^j\|_{L^{q,r}(G_1)} \\
&\le c \, \|u^j\|^p_{p,\mathrm{avg},G_1} + c' \|u^j\|^2_{2,\mathrm{avg},G_1} + \frac{1}{j} \\
&\le c.
\end{aligned}$$

A control of the time derivative, along the lines of [Lindqvist 2008] (see also [Acerbi et al. 2004]), gives

$$\|u_t^j\|_{L^{s,1}(G_{1/2})} \le c,$$

with $s = \min\{q, \, p/(p-1)\} < p$. We now use a classical compactness result (see [Simon 1987, Corollary 4]), with

$$W^{1,p} \hookrightarrow L^p \subset L^s,$$

to conclude that

$$u^j \to \psi,$$

strongly in $L^p(G_{1/2})$, in addition to the weak convergence in $V(G_{1/2})$.

Passing to the limit in (11), we find that

$$\psi_t - \operatorname{div}(|\nabla \psi|^{p-2} \nabla \psi) = 0 \quad \text{in } \ G_{1/2},$$

which contradicts (14), for $j \gg 1$. The proof is complete. □

Next, by means of geometric iteration, we shall establish the optimal Hölder continuity for solutions to the heterogeneous p-parabolic equation (2). Our approach explores the approximation by p-caloric functions, given by Lemma 3.1, and the fact that p-caloric functions are *universally* Lipschitz continuous in space and $C^{0,1/2}$ in time. The following is the crucial first iterative step.

Lemma 3.2. *Let $0 < \alpha < 1$ be fixed. There exists $\epsilon > 0$ and $0 < \lambda \ll 1/2$, depending only on p, n and α, such that if $\|f\|_{L^{q,r}(G_1)} \le \epsilon$ and u is a local weak solution of (2) in G_1, with $\|u\|_{p,\mathrm{avg},G_1} \le 1$, then there exists a universally bounded constant c_0 such that*

$$\|u - c_0\|_{p,\mathrm{avg},G_\lambda} \le \lambda^\alpha. \tag{15}$$

Proof. Take $0 < \delta < 1$, to be chosen later, and apply Lemma 3.1 to obtain $0 < \epsilon \ll 1$ and a p-caloric function ϕ in $G_{1/2}$, such that

$$\|u - \phi\|_{p,\mathrm{avg},G_{1/2}} \leq \delta.$$

Observe that

$$\|\phi\|_{p,\mathrm{avg},G_{1/2}} \leq \|u - \phi\|_{p,\mathrm{avg},G_{1/2}} + \|u\|_{p,\mathrm{avg},G_1} \leq \delta + 2^{(\theta+n)/p} \leq C. \tag{16}$$

Since ϕ is p-caloric, it follows from standard theory that ϕ is universally $C^{0,1/2}_{\mathrm{loc}}$ in time and $C^{0,1}_{\mathrm{loc}}$ in space. That is, for $\lambda \ll 1$, to be chosen soon, we have

$$\sup_{(x,t)\in G_\lambda} |\phi(x,t) - \phi(0,0)| \leq C\,\lambda,$$

for $C > 1$ universal. In fact, for $(x,t) \in G_\lambda$,

$$\begin{aligned} |\phi(x,t) - \phi(0,0)| &\leq |\phi(x,t) - \phi(0,t)| + |\phi(0,t) - \phi(0,0)| \\ &\leq C' |x - 0| + C'' |t - 0|^{1/2} \\ &\leq C'\lambda + C''\lambda^{\theta/2} \leq C\,\lambda, \end{aligned}$$

since $\theta > 2$. We can therefore estimate

$$\begin{aligned} \|u(x,t) - \phi(0,0)\|_{p,\mathrm{avg},G_\lambda} &\leq \|u(x,t) - \phi(x,t)\|_{p,\mathrm{avg},G_\lambda} + \|\phi(x,t) - \phi(0,0)\|_{p,\mathrm{avg},G_\lambda} \\ &\leq \left(\frac{1}{2\lambda}\right)^{\frac{\theta+n}{p}} \delta + C\,\lambda. \end{aligned} \tag{17}$$

Note that we will choose $\lambda \ll 1/2$ and thus

$$G_\lambda = (-\lambda^\theta, 0) \times B_\lambda \subset (-(1/2)^\theta, 0) \times B_{1/2} = G_{1/2}.$$

We put $c_0 := \phi(0,0)$, observing that, due to (16) and the fact that ϕ is p-caloric, c_0 is universally bounded. The next step is to fix the constants. We choose $\lambda \ll \frac{1}{2}$ so small that

$$C\,\lambda \leq \tfrac{1}{2}\lambda^\alpha,$$

and then we define

$$\delta = \tfrac{1}{2}\lambda^\alpha (2\lambda)^{(\theta+n)/p},$$

thus fixing, via Lemma 3.1, also $\epsilon > 0$. The lemma now follows from estimate (17) with the indicated choices. □

Our next step involves iterating Lemma 3.2 in the appropriate geometric scaling.

Theorem 3.3. *Under the conditions of the previous lemma, there exists a convergent sequence of real numbers* $\{c_k\}_{k\geq 1}$, *with*

$$|c_k - c_{k+1}| \leq c(n,p)(\lambda^\alpha)^k, \tag{18}$$

such that

$$\|u - c_k\|_{p,\mathrm{avg},G_{\lambda^k}} \leq (\lambda^k)^\alpha. \tag{19}$$

Proof. The proof is by induction on $k \in \mathbb{N}$. For $k = 1$, (19) holds due to Lemma 3.2, with $c_1 = c_0$. Suppose the conclusion holds for k and let's show it also holds for $k+1$. We start by defining the function $v\colon G_1 \to \mathbb{R}$ by

$$v(x,t) = \frac{u(\lambda^k x, \lambda^{k\theta} t) - c_k}{\lambda^{\alpha k}}. \tag{20}$$

We compute

$$v_t(x,t) = \lambda^{k\theta - \alpha k} u_t(\lambda^k x, \lambda^{k\theta} t)$$

and

$$\operatorname{div}(|\nabla v(x,t)|^{p-2}\nabla v(x,t)) = \lambda^{pk-(p-1)\alpha k} \operatorname{div}(|\nabla u(\lambda^k x, \lambda^{k\theta} t)|^{p-2}\nabla u(\lambda^k x, \lambda^{k\theta} t))$$

to conclude, recalling (8), that

$$v_t - \operatorname{div}(|\nabla v|^{p-2}\nabla v) = \lambda^{pk-(p-1)\alpha k} f(\lambda^k x, \lambda^{k\theta} t) = \tilde{f}(x,t).$$

We now compute

$$\begin{aligned}
\|\tilde{f}\|^r_{L^{q,r}(G_1)} &= \int_{-1}^0 \left(\int_{B_1} |\tilde{f}(x,t)|^q \, dx \right)^{r/q} dt \\
&= \int_{-1}^0 \left(\int_{B_1} \lambda^{(pk-(p-1)\alpha k)q} |f(\lambda^k x, \lambda^{k\theta} t)|^q \, dx \right)^{r/q} dt \\
&= \int_{-1}^0 \left(\int_{B_{\lambda^k}} \lambda^{(pk-(p-1)\alpha k)q - kn} |f(x, \lambda^{k\theta} t)|^q \, dx \right)^{r/q} dt \\
&= \lambda^{((pk-(p-1)\alpha k)q - kn)\frac{r}{q}} \int_{-1}^0 \left(\int_{B_{\lambda^k}} |f(x, \lambda^{k\theta} t)|^q \, dx \right)^{r/q} dt \\
&= \lambda^{((pk-(p-1)\alpha k)q - kn)\frac{r}{q} - k\theta} \int_{-\lambda^{k\theta}}^0 \left(\int_{B_{\lambda^k}} |f(x,t)|^q \, dx \right)^{r/q} dt.
\end{aligned} \tag{21}$$

Due to the crucial and sharp choice (7) of α, we have, recalling again (8),

$$((pk - (p-1)\alpha k)q - kn)\frac{r}{q} - k\theta = 0.$$

We go back to (21) to conclude

$$\|\tilde{f}\|_{L^{q,r}(G_1)} = \|f\|_{L^{q,r}((-\lambda^{k\theta},0)\times B_{\lambda^k})} \le \|f\|_{L^{q,r}(G_1)} \le \epsilon,$$

which entitles v to Lemma 3.2 (note that $\|v\|_{p,\operatorname{avg},G_1} \le 1$, due to the induction hypothesis).

It then follows that there exists a constant $\tilde{c_0}$, with $|\tilde{c_0}| \le c(n,p)$, such that

$$\|v - \tilde{c_0}\|_{p,\operatorname{avg},G_\lambda} \le \lambda^\alpha,$$

which is the same as

$$\|u - c_{k+1}\|_{p,\operatorname{avg},G_{\lambda^{k+1}}} \le \lambda^{\alpha(k+1)},$$

for $c_{k+1} := c_k + \tilde{c_0}\lambda^{\alpha k}$; the induction is complete. We readily observe that

$$|c_{k+1} - c_k| \le c(n, p)(\lambda^\alpha)^k,$$

thus obtaining also (18). □

Theorem 3.4. *A locally bounded weak solution of (2), with $f \in L^{q,r}$, satisfying (3)–(4), is locally Hölder continuous in the space variables, with exponent*

$$\alpha = \frac{(pq - n)r - pq}{q[(p-1)r - (p-2)]}$$

and locally Hölder continuous in time with exponent α/θ. In addition, there exists a constant C, that depends only on $p, n, \|f\|_{q,r}$ and $\|u\|_{p,\mathrm{avg},G_1}$, such that

$$\|u\|_{C^{0;\alpha,\alpha/\theta}(G_{1/2})} \le C.$$

Proof. We start by observing (see also [Araújo et al. 2013, Section 7]) that the smallness regime required in the assumptions of Theorem 3.3 is not restrictive since we can fall into that framework by scaling and contraction. Indeed, given a solution u, let

$$v(x, t) = \varrho u\big(\varrho^a x, \varrho^{(p-2)+ap} t\big)$$

(ϱ, a to be fixed), which is a solution of (2) with

$$\tilde{f}(x, t) = \varrho^{(p-1)+ap} f\big(\varrho^a x, \varrho^{(p-2)+ap} t\big).$$

We choose $a > 0$ such that

$$a < \frac{2}{n+p} \qquad \text{and} \qquad [(p-1)+ap]r - a(n+p) - (p-2) > 0,$$

which is always possible (observe that the second condition holds for $a = 0$ and use its continuity with respect to a), and then $0 < \varrho < 1$ such that

$$\|v\|^p_{p,\mathrm{avg},G_1} \le \varrho^{2-a(n+p)} \|u\|^p_{p,\mathrm{avg},G_1} \le 1$$

and

$$\|\tilde{f}\|^r_{L^{q,r}(G_1)} = \varrho^{[(p-1)+ap]r - a(n+p) - (p-2)} \|f\|^r_{L^{q,r}(G_1)} \le \epsilon^r.$$

Due to (18), the sequence $\{c_k\}_{k\ge1}$ is convergent and we put

$$\bar{c} := \lim_{k\to\infty} c_k.$$

It follows from (19) that, for arbitrary $0 < r < 1/2$,

$$\fint_{G_r} |u - \bar{c}|^p \, dx\, dt \le C r^{p\alpha}.$$

Standard covering arguments, a remark in [Teixeira 2013, Lemma 3.2] and the characterisation of Hölder continuity of Campanato–Da Prato give the local $C^{0;\alpha,\alpha/\theta}$-continuity and thus the result. □

4. Generalisations and beyond

The ideas and methods employed in this paper only explore the degenerate p-structure of the operator. The underlying heuristics is to interpret the homogeneous problem as the geometric tangential equation of its inhomogeneous counterpart, for small perturbations $f \in L^{r,q}$, $\|f\|_{r,q} \ll 1$. The proofs adapt to more general degenerate parabolic equations

$$u_t - \operatorname{div} \mathcal{A}(x, t, Du) = f \in L^{r,q} \tag{22}$$

satisfying the usual structure assumptions for $p \geq 2$.

We briefly comment on the modifications required. Lemma 2.3 is based on pure energy considerations, thus the very same proof works in the general case. Lemma 3.1 can be carried out universally in the structural class of operators, provided integrability bounds for the time-derivative are available (cf. [Acerbi et al. 2004, Section 7], where a more general version of the result in [Lindqvist 2008] on this issue is proved). As for Lemma 3.2, the very same proof works since solutions to the general homogeneous equation are also Lipschitz in space and $C^{0;1/2}$ in time. The only modification occurs when we iterate Lemma 3.2. The rescaled function v defined in (20) now solves the equation

$$v_t - \operatorname{div} \mathcal{A}_k(x, t, Dv) = \lambda^{pk-(p-1)\alpha k} f(\lambda^k x, \lambda^{k\theta} t),$$

where

$$\mathcal{A}_k(x, t, \xi) := (\lambda^{-\alpha k})^{1-p} \mathcal{A}(\lambda^k x, \lambda^{\theta k} t, \lambda^{-\alpha k} \xi)$$

belongs to the same structural class of $\mathcal{A}$. In particular, v is entitled to the conditions of Lemma 3.2 and the proof then follows exactly as before.

We would like to conclude by explaining how the idea of finding geometrical tangential equations can be employed to derive analytical tools for p-parabolic operators, continuously on p. For instance, one can access regularity estimates for degenerate parabolic equations by interpreting the heat operator as the tangential equation obtained when we differentiate the family of p-parabolic operators with respect to the exponent p, near $p = 2$.

It is possible to obtain a universal compactness device. Let $Q_\tau := I_\tau \times B_\tau = (-\tau, \tau) \times B_\tau$. We fix $M_0 \gg 2$ and work within the range $p \in [2, M_0]$.

Lemma 4.1 (uniform in p compactness). *Given $\delta > 0$, there exists $\epsilon > 0$, depending only on n, M_0 and δ, such that if $q \in [2, M_0]$, u is a q-caloric function in Q_1, with $|u| \leq 1$, and $|q - p| < \epsilon$, then we can find a p-caloric function w in $Q_{1/2}$, with $|w| \leq 1$, such that*

$$\sup_{Q_{1/2}} |w - u| \leq \delta. \tag{23}$$

Proof. Suppose, for the sake of contradiction, that the thesis of the lemma does not hold true. This means that for a certain $\delta_0 > 0$, there exist sequences $(q_j)_j$, $(u_j)_j$ and $(p_j)_j$, with

$$\begin{cases} q_j \in [2, M_0], \\ (u_j)_t - \operatorname{div}(|\nabla u_j|^{q_j-2}\nabla u_j) = 0 \quad \text{in } Q_1, \\ |u_j| \le 1, \\ |p_j - q_j| \le \dfrac{1}{j}, \end{cases} \tag{24}$$

but such that, for every p_j-caloric function w in $Q_{1/2}$,

$$\sup_{Q_{1/2}} |u_j - w| > \delta_0. \tag{25}$$

By compactness, we have, up to subsequences,

$$q_j \to q_\infty \in [2, M_0] \tag{26}$$

and, from the last assertion in (24), also $p_j \to q_\infty$. As in the proof of Lemma 3.1, up to a subsequence, $u_j \to u_\infty$ in the appropriate space. Since $q_j \to q_\infty$, by stability (see [Kinnunen and Parviainen 2010]), we can pass to the limit in the equation satisfied by the u_j to conclude that u_∞ is q_∞-caloric in $Q_{2/3}$.

We now solve, for each p_j, the boundary value problem

$$\begin{cases} (w_j)_t - \operatorname{div}(|\nabla w_j|^{p_j-2}\nabla w_j) = 0 & \text{in } Q_{2/3}, \\ w_j = u_\infty & \text{on } \partial Q_{2/3}, \end{cases} \tag{27}$$

and pass to the limit in j, concluding that also $w_j \to u_\infty$ uniformly in $Q_{1/2}$.

Finally, choosing j sufficiently large, we obtain

$$|u_j - w_j| \le |u_j - u_\infty| + |w_j - u_\infty| \le \frac{\delta_0}{2} + \frac{\delta_0}{2} = \delta_0 \quad \text{in } Q_{1/2},$$

which is a contradiction to (25). □

Heuristically, Lemma 4.1 implies the continuity of the underlying regularity theory for p-parabolic operators with respect to p. In particular, improved sharp Hölder estimates can be derived by these methods for problems governed by p-parabolic operators, near the heat equation, *i.e.*, for p close to 2. We leave the development of these heuristics for a future work.

Acknowledgements

This work was developed in the framework of the Brazilian Program "Ciência sem Fronteiras". The authors thank Giuseppe Mingione and Rico Zacher for interesting conversations on the topic of the paper, and acknowledge the warm hospitality of Universidade Federal do Ceará and Universidade de Coimbra.

The research of Teixeira was partially supported by CNPq–Brazil. The research of Urbano was supported by projects UTAustin/MAT/0035/2008, PTDC/MAT/098060/2008, UTA-CMU/MAT/0007/2009 and PTDC/MAT-CAL/0749/2012, by FCT grant SFRH/BSAB/1273/2012, and by CMUC, funded by the European Regional Development Fund through the program COMPETE and by FCT, under the project PEst-C/MAT/UI0324/2011.

References

[Acerbi and Mingione 2007] E. Acerbi and G. Mingione, "Gradient estimates for a class of parabolic systems", *Duke Math. J.* **136**:2 (2007), 285–320. MR 2007k:35211 Zbl 1113.35105

[Acerbi et al. 2004] E. Acerbi, G. Mingione, and G. A. Seregin, "Regularity results for parabolic systems related to a class of non-Newtonian fluids", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **21**:1 (2004), 25–60. MR 2005a:35118 Zbl 1052.76004

[Araújo et al. 2013] D. Araújo, G. Ricarte, and E. V. Teixeira, "Optimal gradient continuity for degenerate elliptic equations", preprint, 2013. Accepted for publication in *Calc. Var. Partial Differential Equations* under the title "Geometric gradient estimates for solutions to degenerate elliptic equation". arXiv 1206.4089

[DiBenedetto 1993] E. DiBenedetto, *Degenerate parabolic equations*, Springer, New York, 1993. MR 94h:35130 Zbl 0794.35090

[DiBenedetto et al. 2012] E. DiBenedetto, U. Gianazza, and V. Vespri, *Harnack's inequality for degenerate and singular parabolic equations*, Springer, New York, 2012. MR 2865434 Zbl 1237.35004

[Duzaar and Mingione 2005] F. Duzaar and G. Mingione, "Second order parabolic systems, optimal regularity, and singular sets of solutions", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **22**:6 (2005), 705–751. MR 2008h:35139 Zbl 1099.35042

[Iwaniec and Manfredi 1989] T. Iwaniec and J. J. Manfredi, "Regularity of p-harmonic functions on the plane", *Rev. Mat. Iberoamericana* **5**:1-2 (1989), 1–19. MR 91i:35071 Zbl 0805.31003

[Kinnunen and Lewis 2000] J. Kinnunen and J. L. Lewis, "Higher integrability for parabolic systems of p-Laplacian type", *Duke Math. J.* **102**:2 (2000), 253–271. MR 2001b:35152 Zbl 0994.35036

[Kinnunen and Parviainen 2010] J. Kinnunen and M. Parviainen, "Stability for degenerate parabolic equations", *Adv. Calc. Var.* **3**:1 (2010), 29–48. MR 2011b:35222 Zbl 1185.35123

[Kuusi et al. 2014] T. Kuusi, G. Mingione, and K. Nyström, "Sharp regularity for evolutionary obstacle problems, interpolative geometries and removable sets", *J. Math. Pures Appl.* (9) **101**:2 (2014), 119–151. MR 3158698 Zbl 06251653

[Lindqvist 2008] P. Lindqvist, "On the time derivative in a quasilinear equation", *Skr. K. Nor. Vidensk. Selsk.* **2008**:2 (2008), 1–7. MR 2010d:35207 Zbl 1165.35402

[Misawa 2013] M. Misawa, "A Hölder estimate for nonlinear parabolic systems of p-Laplacian type", *J. Differential Equations* **254**:2 (2013), 847–878. MR 2990053 Zbl 1273.35173

[Simon 1987] J. Simon, "Compact sets in the space $L^p(0, T; B)$", *Ann. Mat. Pura Appl.* (4) **146** (1987), 65–96. MR 89c:46055 Zbl 0629.46031

[Teixeira 2013] E. V. Teixeira, "Sharp regularity for general Poisson equations with borderline sources", *J. Math. Pures Appl.* (9) **99**:2 (2013), 150–164. MR 3007841 Zbl 1263.35071

[Urbano 2008] J. M. Urbano, *The method of intrinsic scaling: a systematic approach to regularity for degenerate and singular PDEs*, Lecture Notes in Math. **1930**, Springer, Berlin, 2008. MR 2011a:35245 Zbl 1158.35003

Received 12 Aug 2013. Revised 20 Jan 2014. Accepted 17 Feb 2014.

EDUARDO V. TEIXEIRA: teixeira@mat.ufc.br
Department of Mathematics, Universidade Federal do Ceará, Campus of Pici, Bloco 914, 60455–760 Fortaleza-CE, Brazil

JOSÉ MIGUEL URBANO: jmurb@mat.uc.pt
CMUC, Department of Mathematics, University of Coimbra, 3001–501 Coimbra, Portugal

ANALYSIS AND PDE
Vol. 7, No. 3, 2014
dx.doi.org/10.2140/apde.2014.7.745

THE THEORY OF HAHN-MEROMORPHIC FUNCTIONS, A HOLOMORPHIC FREDHOLM THEOREM, AND ITS APPLICATIONS

JÖRN MÜLLER AND ALEXANDER STROHMAIER

We introduce a class of functions near zero on the logarithmic cover of the complex plane that have convergent expansions into generalized power series. The construction covers cases where noninteger powers of z and also terms containing $\log z$ can appear. We show that, under natural assumptions, some important theorems from complex analysis carry over to this class of functions. In particular, it is possible to define a field of functions that generalize meromorphic functions, and one can formulate an analytic Fredholm theorem in this class. We show that this modified analytic Fredholm theorem can be applied in spectral theory to prove convergent expansions of the resolvent for Bessel type operators and Laplace–Beltrami operators for manifolds that are Euclidean at infinity. These results are important in scattering theory, as they are the key step in establishing analyticity of the scattering matrix and the existence of generalized eigenfunctions at points in the spectrum.

1. Introduction

Asymptotic expansions of the form

$$f(z) \sim \sum_{k,m} a_{k,m} z^{\alpha_k} (-\log z)^{\beta_m} \quad \text{as } z \to 0,$$

with nonintegers α_k and β_m, defined for functions f in some sector centered at 0 in the complex plane, appear frequently in mathematics and mathematical physics. Classical examples are solutions for differential equations (for example, in Frobenius' method) and expansions of algebraic functions at singularities. It has been shown that low energy resolvent expansions in scattering problems are of this form; see, for example, [Jensen and Kato 1979; Jensen and Nenciu 2001] for Schrödinger operators in $\mathbb{R}^n$, [Murata 1982] for operators with constant leading coefficients in $\mathbb{R}^n$, and [Guillarmou and Hassell 2009] for the Laplace operator on a general manifold with a conical end. The resolvent expansion for $|\lambda| \to \infty$ of cone degenerate differential operators leads to similar asymptotics; see, for example, [Gil et al. 2011]. In many of these examples, the expansions can be shown to be convergent under more restrictive assumptions on the structure at infinity of the underlying geometry.

The algebraic theory of generalized power series is well developed and can be found in the literature under the name Hahn series or Maltsev–Neumann series (see, for example, [Hahn 1907; Passman 1977, Chapter 13; Ribenboim 1992]). In this paper we are concerned with the analytic theory of such generalized

Both authors were supported by the *SFB 647: Space–Time–Matter. Analytic and Geometric Structures*.
MSC2010: 47A56, 58J50.
Keywords: Hahn series, holomorphic Fredholm theorem, scattering theory.

power series. Namely, we will define a ring of functions, the Hahn-holomorphic functions, that have convergent expansions into generalized power series, and we will show that this ring is actually a division ring. We show that the quotient field, the field of Hahn-meromorphic functions, has a nice description in terms of Hahn series, and we generalize the notions of Hahn-holomorphic and Hahn-meromorphic functions to the operator valued case. The theory turns out to be very close to the case of analytic function theory. In particular, one of our main theorems states that an analog of the analytic Fredholm theorem holds in the class of Hahn-holomorphic functions.

The holomorphic Fredholm theorem plays an important role in geometric scattering theory as a tool to prove the existence of a meromorphic continuation of resolvent kernels of elliptic differential operators such as the Laplace operator. The extension is typically from the resolvent set across the continuous spectrum to a branched covering of the complex plane. As soon as such a meromorphic continuation of the resolvent kernel is established, resonances can be defined as poles of its continuation, generalized eigenfunctions may be defined as meromorphic functions of a suitably chosen spectral parameter, and an analytic continuation of the scattering matrix may be constructed. This in many situations leads to a rich mathematical structure that results in functional equations for the scattering matrix and Maass–Selberg relations for the generalized eigenfunctions (see, for example, [Müller 1987] for the case of manifolds with cusps of rank one, [Melrose 1993; Guillopé 1989; Müller and Strohmaier 2010] for manifolds with cylindrical ends, and [Müller 2011] for manifolds with fibered cusps). In particular, the analytic continuation of Eisenstein series may be regarded as a special case of this more general construction.

Often, as for example in the case of $\mathbb{R}^{2n+1}$ on asymptotically hyperbolic manifolds [Mazzeo and Melrose 1987; Guillarmou 2005], geometrically finite hyperbolic manifolds [Guillarmou and Mazzeo 2012], and on globally symmetric spaces of odd rank [Mazzeo and Vasy 2005; Strohmaier 2005], the branch points of the covering of the complex plane are algebraic and can be resolved by a change of variables. In this way, one can make sense of the statement that the resolvent is meromorphic at the branch point. In other examples, as in $\mathbb{R}^{2n}$ on symmetric spaces of even rank [Mazzeo and Vasy 2005; Strohmaier 2005] and on manifolds with generalized cusps [Hunsicker et al. 2014], the branch point is logarithmic, and this statement loses its meaning. The analytic Fredholm theorem can then only be applied away from the branching points. Our philosophy is that, at such branching points, it still makes sense to say when functions are Hahn-holomorphic, that is, have a convergent expansion into more general power series possibly containing log terms. Our Hahn analytic Fredholm theorem therefore allows us to analyze the resolvents at nonalgebraic branching points. Our theorem implies, for example, that the Hahn-meromorphic properties of the resolvent of the Laplace operator on a Riemannian manifold are stable under perturbations of the topology and the metric that are supported in compact regions. The theory can be developed further to establish Hahn-analyticity of the scattering matrix and of the generalized eigenfunctions in this context, but we decided to focus on the theoretical properties of Hahn-meromorphic functions first and keep the presentation self-contained. The applications in scattering theory will be developed elsewhere.

The article is organized as follows. Section 2 deals with the definition and the general theory of Hahn-holomorphic functions and some of their basic properties. In Section 3 we define Hahn-meromorphic

functions, and in Section 4 we prove our generalization of the meromorphic Fredholm theorem in the framework of Hahn-holomorphic functions. Sections 5 and 6 deal with two important examples of convergent Hahn series: those that can be expanded into real powers of z and those that have such expansions with additional $\log z$ terms. The theory has a nice application: convergent resolvent expansions for Bessel type operators and Laplace–Beltrami operators on manifolds that are Euclidean at infinity can be shown to be simple consequences of the Hahn-holomorphic Fredholm theorem. These examples are treated in detail in Section 7; the main results here are Theorems 7.6 and 7.9.

We would like to thank the anonymous referee for suggestions leading to considerable simplifications in some of the arguments in Section 7.

2. Hahn-holomorphic functions

Let $(\Gamma, +)$ be a linearly ordered abelian group and let $(G, \cdot)$ be a group. Suppose $e : \Gamma \to G$, $\gamma \mapsto e_\gamma$ is a group homomorphism; in particular

$$e_0 = \mathbf{1} \in G, \quad e_{\gamma_1+\gamma_2} = e_{\gamma_1} \cdot e_{\gamma_2} \quad \text{for all } \gamma_1, \gamma_2 \in \Gamma.$$

The following definition and proposition are due to H. Hahn [1907].

Definition 2.1. Let $\mathcal{R}$ be a ring. A formal series

$$\mathfrak{h} = \sum_{\gamma\in\Gamma} a_\gamma e_\gamma, \quad a_\gamma \in \mathcal{R}$$

is called a *Hahn series* if the support of $\mathfrak{h}$,

$$\operatorname{supp}(\mathfrak{h}) := \{g \in \Gamma \mid a_g \neq 0 \in \mathcal{R}\},$$

is a well-ordered subset of Γ. The set of Hahn series will be denoted by $\mathcal{R}[\![e_\Gamma]\!]$.

Proposition 2.2. *The set of Hahn series $\mathcal{R}[\![e_\Gamma]\!]$ is a ring with multiplication*

$$\left(\sum_{\alpha\in\Gamma} a_\alpha e_\alpha\right)\left(\sum_{\beta\in\Gamma} b_\beta e_\beta\right) = \sum_{\gamma\in\Gamma} c_\gamma e_\gamma, \quad c_\gamma := \sum_{\substack{(\alpha,\beta)\in\Gamma\times\Gamma \\ \alpha+\beta=\gamma}} a_\alpha b_\beta \tag{1}$$

and addition

$$\sum_{\alpha\in\Gamma} a_\alpha e_\alpha + \sum_{\beta\in\Gamma} b_\beta e_\beta = \sum_{\gamma\in\Gamma} (a_\gamma + b_\gamma) e_\gamma.$$

If $\mathcal{R}$ is a field, so is $\mathcal{R}[\![e_\Gamma]\!]$.

It is well known that if the support of $\mathfrak{h}$ is contained in $\Gamma^+ = \{\gamma \mid \gamma > 0\}$, then $\mathbf{1} - \mathfrak{h}$ is invertible in $\mathcal{R}[\![e_\Gamma]\!]$ and its inverse is given by the Neumann series

$$(\mathbf{1} - \mathfrak{h})^{-1} = \sum_{k=0}^{\infty} \mathfrak{h}^k.$$

This is due to the fact that, for any well-ordered subset W of Γ^+, the semigroup generated by W is also well ordered; see, for example, [Passman 1977, Lemma 2.10]. Here convergence of a sequence

$(\mathfrak{p}_n) \subset \mathcal{R}[\![e_\Gamma]\!]$ to $\mathfrak{p} \in \mathcal{R}[\![e_\Gamma]\!]$ is understood in the sense that, for every element $\alpha \in \Gamma$, there exists an $N > 0$ such that, for all $n > N$, the coefficients of e_α in $\mathfrak{p}$ and $\mathfrak{p}_n$ are equal.

In the following, let $\mathcal{Z}$ be the logarithmic covering surface of the complex plane without the origin. We will use polar coordinates (r, φ) as global coordinates to identify $\mathcal{Z}$ as a set with $\mathbb{R}_+ \times \mathbb{R}$. Adding a single point $\{0\}$ to $\mathcal{Z}$, we obtain a set $\mathcal{Z}_0$ and a projection map $\pi : \mathcal{Z}_0 \to \mathbb{C}$ by extending the covering map $\mathcal{Z} \to \mathbb{C}\backslash\{0\}$ sending $0 \in \mathcal{Z}_0$ to $0 \in \mathbb{C}$. We endow $\mathcal{Z}$ with the covering topology and $\mathcal{Z}_0$ with the topology generated by the open sets in $\mathcal{Z}$ together with the open discs $D_\varepsilon := \{0\} \cup \{(r, \varphi) \mid 0 \le r < \varepsilon\}$. This means a sequence $((r_n, \varphi_n))_n$ converges to zero if and only if $r_n \to 0$. The covering map is continuous with respect to this topology. For a point $z \in \mathcal{Z}_0$, we denote by $|z|$ its r-coordinate and by $\arg z$ its φ coordinate. We will think of the positive real axis as embedded in $\mathcal{Z}$ as the subset $\{z \mid \arg z = 0\}$. In the following, $Y \subset \mathcal{Z}$ will always denote an open subset containing an open interval $(0, \delta)$ for some $\delta > 0$ and such that $0 \notin Y$. The set Y_0 will denote $Y \cup \{0\}$. In the applications we have in mind, the set Y is typically of the form $D_\delta^{[\sigma]}\backslash\{0\}$, where $D_\delta^{[\sigma]} = \{z \in \mathcal{Z}_0 \mid 0 \le |z| < \delta,\ |\varphi| < \sigma\}$. For the discussion and the general theorems, it is not necessary to restrict ourselves to this case.

In the remainder of this article we assume that $G := (\mathrm{Hol}(Y \cap D_\varepsilon), \cdot)^\times$ is a set of nonvanishing holomorphic functions and that the group homomorphism e satisfies the condition

$$\text{for all } \gamma > 0, \quad e_\gamma \text{ is bounded on } Y \text{ and } \lim_{z \to 0} |e_\gamma(z)| = 0. \tag{E1}$$

Definition 2.3. Suppose that $\mathcal{R}$ is a vector space with norm $\|\cdot\|$. A Hahn series $\mathfrak{f} = \sum_{\alpha \in \Gamma} a_\alpha e_\alpha$ is called *normally convergent* in $Y \cap D_\varepsilon$ if its support is countable and

$$\sum_{\alpha \in \Gamma} \|a_\alpha\| \, \|e_\alpha\|_{Y,\varepsilon} < \infty,$$

where $\|e_\alpha\|_{Y,\varepsilon} := \sup_{z \in Y \cap D_\varepsilon} |e_\alpha(z)|$.

Since a normally convergent series converges absolutely and uniformly, the value of the function

$$f(z) = \sum_{\alpha \in \Gamma} a_\alpha e_\alpha(z), \quad z \in Y \cap D_\varepsilon,$$

does not depend on the order of summation and f is holomorphic in $z \neq 0$.

Definition 2.4. Let $S \subset \Gamma_0^+ = \Gamma^+ \cup \{0\}$ be a subset of the nonnegative group elements.

- The family $\{e_\alpha\}_{\alpha \in S}$ is called *weakly monotonic* if there exists an $r_S > 0$ such that, for every $x \in (0, r_S)$, there is a *radius* $\rho(x)$ with $0 < \rho(x) \le x$ and with the property

$$\alpha \in S \implies \|e_\alpha\|_{Y,\rho(x)} \le |e_\alpha(x)|.$$

- The set S is called *admissible for* e (or simply *admissible*) if $\{e_\alpha\}_{\alpha \in S}$ is weakly monotonic, and if, for every $B \subset S$, the family

$$\{e_{\alpha - \min B}\}_{\substack{\alpha \in S \\ \alpha > \min B}}$$

is also weakly monotonic.

Definition 2.5 (Hahn-holomorphic functions). Suppose that $\mathcal{R}$ is a Banach algebra. A continuous function $h : Y_0 \to \mathcal{R}$ which is holomorphic in Y is called (Y, Γ)-Hahn-holomorphic (or simply *Hahn-holomorphic*) if there is a Hahn series

$$\mathfrak{h} = \sum_{\gamma\in\Gamma} a_\gamma e_\gamma, \quad a_\gamma \in \mathcal{R},$$

with countable admissible support, converging normally on $Y \cap D_\delta$ for some $\delta > 0$, and such that

$$h(z) = \sum_{\gamma\in\Gamma} a_\gamma e_\gamma(z), \quad z \in Y \cap D_\delta.$$

We will denote the Hahn series of a Hahn-holomorphic function h by the corresponding "fraktur" letter $\mathfrak{h}$. Note that (E1) together with uniform convergence imply that $\operatorname{supp}\mathfrak{h} \subset \Gamma_0^+$ and $h(0) = a_0$. Of course any normally convergent Hahn series with admissible support gives rise to a Hahn-holomorphic function.

Here is a direct consequence of the support of Hahn-holomorphic functions being admissible:

Lemma 2.6. *Let*

$$h(z) = \sum_{\gamma\in\Gamma} a_\gamma e_\gamma(z), \quad z \in Y \cap D_{2r},$$

be Hahn-holomorphic with $\mathfrak{m} = \min\operatorname{supp}(\mathfrak{h})$. *Then*

$$e_{-\mathfrak{m}}(z)h(z) = \sum_{\gamma\geq\mathfrak{m}} a_\gamma e_{\gamma-\mathfrak{m}}(z)$$

is Hahn-holomorphic.

Proof. Let ρ_1 be the radius for $\{e_\gamma\}$ such that, for all $\gamma \in \operatorname{supp}(\mathfrak{h})$,

$$\|e_\gamma\|_{\rho_1(r)} \leq |e_\gamma(r)|,$$

and similarly let ρ_2 be the radius for $\{e_{\gamma-\mathfrak{m}}\}$. For $\rho(r) = \min\{\rho_1(r), \rho_2(r)\}$,

$$\|e_\mathfrak{m}\|_{\rho(r)} \sum_{\gamma\in\Gamma} \|a_\gamma\| \, \|e_{\gamma-\mathfrak{m}}\|_{\rho(r)} \leq |e_\mathfrak{m}(r)| \sum_{\gamma\in\Gamma} \|a_\gamma\| \, |e_{\gamma-\mathfrak{m}}(r)| = \sum_{\gamma\in\Gamma} \|a_\gamma\| \, |e_\gamma(r)| < \infty.$$

Thus $\sum_{\gamma\in\Gamma} a_\gamma e_{\gamma-\mathfrak{m}}$ converges normally on $D_{\rho(r)}$. □

Proposition 2.7. *Let $f : Y \to \mathcal{R}$ be a Hahn-holomorphic function represented by a Hahn series $\mathfrak{f}$ on $Y \cap D_\delta$. Suppose the zeros of f accumulate in $Y \cup \{0\}$. Then $f \equiv 0$ and $\mathfrak{f} = 0$. In particular, the Hahn series of a Hahn-holomorphic function is completely determined by the germ of the function at zero.*

Proof. If the zero set of f has accumulation points in Y, the statement follows from the fact that f is holomorphic in this set. It remains to show that if $f \neq 0$, then 0 can not be an accumulation point of the zero set of f. Let $\mathfrak{f}$ be a Hahn series that represents the function on $Y \cap D_\varepsilon$. Let $f \neq 0$; then $\mathfrak{f} \neq 0$. Let $\mathfrak{m} = \min\operatorname{supp}\mathfrak{f}$. If there is no other element in the support of $\mathfrak{f}$, then $f(z) = a_\mathfrak{m} e_\mathfrak{m}(z)$ and the statement follows from the fact that $e_\mathfrak{m}$ has no zeros in Y. Otherwise, let $\mathfrak{m}_1$ be the smallest element in $\operatorname{supp}\mathfrak{f}$ which is larger than $\mathfrak{m}$. Then

$$f(z) = \sum_\alpha a_\alpha e_\alpha(z) = e_\mathfrak{m}(z)\left(a_\mathfrak{m} + e_{\mathfrak{m}_1-\mathfrak{m}}(z) \sum_{\alpha\geq\mathfrak{m}_1} a_\alpha e_{\alpha-\mathfrak{m}_1}(z)\right) = e_\mathfrak{m}(z)(a_\mathfrak{m} + h(z))$$

with a Hahn-holomorphic function $h(z)$ such that $h(0) = 0$. Since h is continuous and $e_{\mathfrak{m}}(z) \neq 0$, this shows that $f(z) \neq 0$ in a neighborhood of 0. □

Now suppose that Y, Γ and the family of functions $(e_\gamma)_{\gamma\in\Gamma}$ are fixed and satisfy (E1).

We want to show that the space of Hahn-holomorphic functions at 0 with values in a Banach algebra $\mathcal{R}$ is a ring. To that end we need the following.

Lemma 2.8. *Let $A_1, A_2 \subset \Gamma^+$ be admissible sets. Then the sets $A_1 \cup A_2$, $A_1 + A_2$, and $n \cdot A_1 := A_1 + \cdots + A_1$ (n times), $\bigcup_{n=0}^{\infty} n \cdot A_1$ are admissible.*

Proof. First we show that $A_1 \cup A_2$, $A_1 + A_2$ and $n \cdot A_1$ are weakly monotonic. Let ρ_i, $i = 1, 2$, be the radius for A_i and $\rho(x) = \min\{\rho_1(x), \rho_2(x)\}$. Then ρ is a radius for $A_1 \cup A_2$ and for $A_1 + A_2$ as well, because, for $\alpha_i \in A_i$,

$$\begin{aligned}\|e_{\alpha_1+\alpha_2}\|_{\rho(r)} &\leq \|e_{\alpha_1}\|_{\rho(r)}\|e_{\alpha_2}\|_{\rho(r)} \leq \|e_{\alpha_1}\|_{\rho_1(r)}\|e_{\alpha_2}\|_{\rho_2(r)} \\ &\leq |e_{\alpha_1}(r)|\,|e_{\alpha_2}(r)| = |e_{\alpha_1+\alpha_2}(r)|.\end{aligned}$$

The same argument shows that ρ_1 is a radius for $n \cdot A_1$.

Now let $B \subset A := A_1 + A_2$. Then $B = B_1 + B_2$ for some $B_i \subset A_i$, $i = 1, 2$, and $\min B = \min B_1 + \min B_2$. Let $\alpha \in A$ with $\alpha = \alpha_1 + \alpha_2$, $\alpha_i \in A_i$. Let $\rho_i(r)$ be the radius for $\{e_{\alpha_i - \min B_i}\}$ and $\rho = \min\{\rho_1, \rho_2\}$. The estimate

$$\|e_{\alpha-\min B}\|_{\rho(r)} = \|e_{\alpha_1-\min B_1+\alpha_2-\min B_2}\|_{\rho(r)} \leq \|e_{\alpha_1-\min B_1}\|_{\rho_1(r)}\|e_{\alpha_2-\min B_2}\|_{\rho_2(r)}$$

shows that $A_1 + A_2$ is admissible. The other statements are proven similarly. □

Let $f(z) = \sum_\alpha a_\alpha e_\alpha$ and $g(z) = \sum_\beta b_\beta e_\beta$ be Hahn-holomorphic functions on Y_f and Y_g, respectively. First it is easy to see that $f + g$ is Hahn-holomorphic on $Y = Y_f \cap Y_g$. Since $\mathfrak{f}$ and $\mathfrak{g}$ are Hahn series with support contained in Γ_0^+, we also have $\operatorname{supp}(\mathfrak{f} \cdot \mathfrak{g}) \subset \Gamma_0^+$ for the multiplication defined in (1). From Lemma 2.8 we obtain that the support of $\mathfrak{f} \cdot \mathfrak{g}$ is admissible. We claim that $h(z) = f(z) \cdot g(z)$ is represented by the product of Hahn series $\mathfrak{h} = \mathfrak{f} \cdot \mathfrak{g}$ on $Y_f \cap Y_g$. Because f and g are normally convergent,

$$\sum_\gamma \Big\| \sum_{\alpha+\beta=\gamma} a_\alpha b_\beta \Big\| \|e_\gamma\| \leq \sum_\gamma \Big(\sum_{\alpha+\beta=\gamma} \|a_\alpha\|\,\|b_\beta\| \Big) \|e_\gamma\| \leq \Big(\sum_\alpha \|a_\alpha\|\,\|e_\alpha\| \Big) \Big(\sum_\beta \|b_\beta\|\,\|e_\beta\| \Big)$$

so that the series $\mathfrak{f} \cdot \mathfrak{g}$ is normally convergent in $Y_f \cap Y_g$. Thus the series $\mathfrak{f} \cdot \mathfrak{g}$ defines a Hahn-holomorphic function on Y with values in $\mathcal{R}$, and this function equals $h(z)$.

Altogether we have this:

Proposition 2.9. *Let $\mathcal{R}$ be a Banach algebra. The Hahn-holomorphic functions with values in $\mathcal{R}$ on Y form a ring under usual addition and multiplication, and the map $\psi_{\mathcal{R}} : f \mapsto \mathfrak{f}$ is a ring isomorphism onto its image in $\mathcal{R}[\![e_\gamma]\!]$.*

Corollary 2.10. *The ring of Hahn-holomorphic functions on Y with values in an integral domain $\mathcal{R}$ is an integral domain.*

Proof. By looking at the coefficient c_γ with $\gamma = \min \operatorname{supp} \mathfrak{f}$ in (1), we observe that $\mathcal{R}[\![e_\Gamma]\!]$ is an integral domain if $\mathcal{R}$ is an integral domain. Because $\psi_{\mathcal{R}}$ is an isomorphism, the Hahn-holomorphic functions must be an integral domain. □

Theorem 2.11. *Let $\mathcal{R}$ be a Banach algebra and suppose $f : Y_0 \to \mathcal{R}$ is Hahn-holomorphic and $f(z)$ is invertible for all $z \in Y_0$. Then $f(z)^{-1}$ is also Hahn-holomorphic on Y_0.*

Proof. Since $1/f$ is holomorphic in Y, we only have to show that there is a Hahn series for $f(z)^{-1}$ that converges normally on some $Y_0 \cap D_\varepsilon$. Since $f(z)^{-1} = f(0)^{-1}(f(z)f(0)^{-1})^{-1}$, we can assume without loss of generality that $f(0) = \mathrm{Id}$. Thus we can write $f(z) = \mathrm{Id} - h(z)$, where $\mathfrak{m} := \min \operatorname{supp}(\mathfrak{h}) > 0$. By assumption, the series $\mathfrak{h} := \sum_{\alpha\in\Gamma} a_\alpha e_\alpha$ defining $h(z)$ converges normally on the set $Y_0 \cap D_{\delta_0}$ for some $\delta_0 > 0$. The function $\tilde{h}$ defined by

$$\tilde{h}(t) = \sum_{\alpha\in\Gamma} \|a_\alpha\| \, \|e_\alpha\|_{Y_0,t} \le \|e_\mathfrak{m}\|_{Y_0,t} \sum_{\alpha\ge\mathfrak{m}} \|a_\alpha\| \, \|e_{\alpha-\mathfrak{m}}\|_{Y_0,t}$$

converges to 0 for $t \to 0$ due to (E1) and Lemma 2.6. Therefore we can choose $\delta > 0$ so small that $\tilde{h} := \tilde{h}(\delta) < 1/2$. Because $|h(z)| \le \tilde{h}$ for $z \in Y_0 \cap D_\delta$, the geometric series

$$f(z)^{-1} = \sum_{n=0}^{\infty} h(z)^n$$

then converges normally on $Y_0 \cap D_\delta$. But we also know that $\mathfrak{f}$ is invertible:

$$\mathfrak{f}^{-1} = \sum_{n=0}^{\infty} \mathfrak{h}^n =: \sum_{\alpha\in\mathcal{S}} b_\alpha e_\alpha \quad \text{with } \operatorname{supp}(\mathfrak{f}^{-1}) \subset \mathcal{S} := \bigcup_{n\ge0} \operatorname{supp}(\mathfrak{h}^n).$$

From Lemma 2.8 we obtain that $\mathcal{S}$ is admissible. It remains to show that $\sum_{\alpha\in\mathcal{S}} b_\alpha e_\alpha(z)$ is normally convergent on $Y_0 \cap D_\delta$ and represents $f(z)^{-1}$. We have the implication

$$\sum_{n=0}^{N} \mathfrak{h}^n = \sum_{\alpha\in\mathcal{S}} c_\alpha(N) e_\alpha \quad \Longrightarrow \quad \sum_{\alpha\in\mathcal{S}} \|c_\alpha(N)\| \, \|e_\alpha\| \le \sum_{n=0}^{N} \tilde{h}^n \quad \text{in } Y_0 \cap D_\delta,$$

as a simple consequence of the triangle inequality. For every fixed finite set $A \subset \mathcal{S}$, there exists an $N_A > 0$ such that, for all $N \ge N_A$,

$$\mathfrak{f}^{-1} - \sum_{n=0}^{N} \mathfrak{h}^n = \sum_{\alpha\in\mathcal{S}\setminus A} (b_\alpha - c_\alpha(N)) e_\alpha$$

has support away from A. In particular, $c_\alpha(N) = b_\alpha$ for $\alpha \in A$ and $N \ge N_A$. Therefore, for $N > N_A$,

$$\sum_{\alpha\in A} \|b_\alpha\| \, \|e_\alpha\| \le \sum_{\alpha\in\mathcal{S}} \|c_\alpha(N)\| \, \|e_\alpha\| \le \sum_{n=0}^{N} \tilde{h}^n < \frac{1}{1-\tilde{h}},$$

and this proves convergence, since this bound is independent of A. In particular, $\sum_{\alpha\in\mathcal{S}} b_\alpha e_\alpha(z)$ converges absolutely in $\mathcal{R}$, hence it converges and the value does not depend on the order of summation. After

reordering,

$$\sum_{\alpha\in\mathcal{S}} b_\alpha e_\alpha(z) = \sum_{n=0}^{\infty} h(z)^n = f(z)^{-1}. \qquad \square$$

Because of Lemma 2.6, every complex valued Hahn-holomorphic f that is not identically 0 can be inverted away from its zeros. Let $\mathfrak{m} := \min \operatorname{supp}(\mathfrak{f}) \geq 0$. Then

$$f^{-1}(z) = a_{\mathfrak{m}}^{-1} e_{-\mathfrak{m}}(z) \sum_{n=0}^{\infty} (1 - a_{\mathfrak{m}}^{-1} e_{-\mathfrak{m}}(z) f(z))^n.$$

Theorem 2.12. *Suppose that $f : Y_0 \to \mathbb{C}$ is a Hahn-holomorphic function with Hahn series $\mathfrak{f}$. Suppose that U is an open neighborhood of $f(0)$ and $h : U \to \mathbb{C}$ is holomorphic. Then $h \circ f$ is Hahn-holomorphic on its domain.*

Proof. Since holomorphicity away from zero is obvious, it is enough to show that $h \circ f$ has a normally convergent expansion into a Hahn series. Replacing $f(z)$ by $f(z) - f(0)$ and $h(z)$ by $h(z - f(0))$, we can assume without loss of generality that $f(0) = 0$ and thus $\operatorname{supp}(\mathfrak{f}) \subset \Gamma^+$. Since h is holomorphic near $f(0)$, it has a uniformly and absolutely convergent expansion

$$h(z) = \sum_{k=0}^{\infty} a_k (z - f(0))^k.$$

Thus

$$h \circ f(z) = \sum_{k=0}^{\infty} a_k (f(z))^k.$$

Note that $\sum_{k=0}^{\infty} a_k \mathfrak{f}^k$ is a Hahn series. A similar argument as in the proof of Theorem 2.11 shows that this Hahn series is normally convergent and represents $h \circ f(z)$. $\square$

3. Hahn-meromorphic functions

Definition 3.1. A meromorphic function $h : Y \to \mathbb{C}$ is called *Hahn-meromorphic* if h is represented by a Hahn series $\mathfrak{h}$ in $Y \cap D_\varepsilon$ for some $\varepsilon > 0$ and there exist Hahn-holomorphic functions $f, g \not\equiv 0$ on $Y_0 \cap D_\varepsilon$ such that $\mathfrak{h} \cdot \mathfrak{g} = \mathfrak{f}$.

In this sense, a Hahn-meromorphic function can be written as a quotient $h = f/g$ of Hahn-holomorphic functions in a neighborhood of 0.

Remark 3.2. Since $\mathbb{C}$-valued Hahn-holomorphic functions form an integral domain, Hahn-meromorphic functions form a field. More generally, let $\mathcal{R}$ be a (commutative) integral domain. From Corollary 2.10 we know that Hahn-holomorphic functions with coefficients in $\mathcal{R}$ are a commutative integral domain whose quotient field is defined. Furthermore, the map $f \mapsto \mathfrak{f}$ induces an injective morphism from the quotient field of Hahn-holomorphic functions to the quotient field $\mathcal{R}((e_\Gamma))$ of Hahn series $\mathcal{R}[\![e_\Gamma]\!]$. Note that $\mathcal{R}((e_\Gamma)) = \mathcal{R}[\![e_\Gamma]\!]$ if $\mathcal{R}$ is a field.

An important difference with usual meromorphic functions is that Hahn-meromorphic functions may have infinitely many negative exponents. For example, the function

$$f(x) = \sum_{n=1}^{\infty} \frac{1}{n^2} z^{1-1/n}$$

is Hahn-holomorphic and therefore

$$\sum_{n=1}^{\infty} \frac{1}{n^2} z^{-1/n-1} = \frac{f(z)}{z^2}$$

is Hahn-meromorphic.

It follows from our analysis for Hahn-holomorphic functions that every $\mathbb{C}$-valued Hahn-meromorphic function h can be written as

$$h(z) = e_{\min \operatorname{supp} \mathfrak{h}}(z) f(z),$$

where f is Hahn-holomorphic. Moreover, if $h \neq 0$, then $f(0) \neq 0$. In particular, this implies that Hahn-meromorphic functions bounded on $(0, \delta)$ are Hahn-holomorphic in some neighborhood of 0.

We can also define Hahn-meromorphic functions with values in a Banach algebra.

Definition 3.3. Let $\mathcal{R}$ be a Banach algebra. A function $h : Y \to \mathcal{R}$ is called *Hahn-meromorphic* if it is meromorphic on Y and there exists a $\delta > 0$ and a nonzero Hahn-holomorphic function f on $Y_0 \cap D_\delta$ such that $f(z)h(z)$ is a Hahn-holomorphic function on $Y_0 \cap D_\delta$ with values in $\mathcal{R}$.

Remark 3.4. Let $R > 0$ and $\sigma > 0$. If there exists one nonzero Hahn-holomorphic function on $Y \cap D_R^{[\sigma]}$ that vanishes with positive order at 0, then one can use the Weierstrass product theorem together with Theorem 2.12 to show that the set of complex valued Hahn-meromorphic functions on $Y \cap D_R^{[\sigma]}$ can be identified with the quotient field of the division ring of Hahn-holomorphic functions on $Y \cap D_R^{[\sigma]}$.

4. A Hahn-holomorphic Fredholm theorem

Let $\mathcal{H}$ be a complex Hilbert space and denote by $\mathcal{K}(\mathcal{H})$ the space of compact operators on $\mathcal{H}$.

Theorem 4.1. *Suppose $Y_0 \subset \mathfrak{L}$ is connected and let $f : Y \to \mathcal{K}(\mathcal{H})$ be either Hahn-holomorphic or Hahn-meromorphic such that all coefficients of e_γ with $\gamma < 0$ and all Laurent coefficients in the principal part away from the point $z = 0$ have range in a common finite-dimensional subspace $\mathcal{H}_0 \subset \mathcal{H}$.*

Then either $(\mathrm{Id} - f(z)) \in \mathcal{B}(\mathcal{H})$ is invertible nowhere in Y_0 or its inverse $(\mathrm{Id} - f(z))^{-1}$ exists everywhere except at a discrete set of points in Y_0 and defines a Hahn-meromorphic function. Moreover, in the Hahn series of $(\mathrm{Id} - f(z))^{-1}$, the coefficients of e_γ with $\gamma < 0$ are finite-rank operators, and the coefficients in the principal part of its Laurent expansion away from $z = 0$ are finite-rank operators too.

Proof. The proof generalizes that of [Reed and Simon 1980, Theorem VI.14]. The assumptions imply that there exists a Hahn-meromorphic function $B(z)$ with range in $\mathcal{H}_0$, a finite-rank operator A, and a $\delta > 0$ such that $f(z) - A - B(z)$ is Hahn-holomorphic and $\| f(z) - A - B(z) \| < 1$ for all $z \in U^{[\sigma]} := D_\delta^{[\sigma]} \cap Y$. Thus $(\mathrm{Id} - f(z) + A + B(z))^{-1}$ exists and is Hahn-holomorphic by Theorem 2.11. Consequently, $g(z) = (A + B(z))(\mathrm{Id} - f(z) + A + B(z))^{-1}$ is a Hahn-meromorphic function on $U^{[\sigma]}$ with values in the

Banach space $\mathcal{B}(\mathcal{H}, V)$, where V is the finite-dimensional subspace of $\mathcal{H}$ spanned by $\mathcal{H}_0$ and $\mathrm{rg}(A)$. It is easy to see that

$$(\mathrm{Id} - f(z))^{-1} = (\mathrm{Id} - f(z) + A + B(z))^{-1}(\mathrm{Id} - g(z))^{-1}, \tag{2}$$

where equality means here that the left hand side exists if and only if the right hand side exists. Now let P be the orthogonal projection onto V and let $G(z)$ be the endomorphisms of V defined by restricting $g(z)$ to V, that is, $G(z) = g(z) \circ P$. Invertibility of $\mathrm{Id} - g(z)$ in $\mathcal{B}(\mathcal{H})$ is equivalent to invertibility of

$$P(\mathrm{Id} - g(z))P : V \to V,$$

and this is equivalent to $\det(\mathrm{Id}_V - G(z)) \neq 0$. Moreover, a straightforward computation shows

$$(\mathrm{Id} - g(z))^{-1} = (P(\mathrm{Id} - g(z))P)^{-1}(P + g(z)(\mathrm{Id} - P)) + (\mathrm{Id} - P). \tag{3}$$

Now note that $G(z)$ is a Hahn-holomorphic family of endomorphisms of V. In particular, $\det(\mathrm{Id} - G(z))$ is a Hahn-meromorphic $\mathbb{C}$-valued function. As such, it is meromorphic in $U^{[\sigma]} \setminus \{0\}$, and together with Proposition 2.7, this shows that the set

$$S = \{z \in U^{[\sigma]} \mid \det(\mathrm{Id} - G(z)) = 0\}$$

is either discrete in $U^{[\sigma]}$ or $S = U^{[\sigma]}$. If $\det(\mathrm{Id} - G(z)) \neq 0$, then, after a choice of basis of V, the inverse $(\mathrm{Id} - G(z))^{-1}$ can be computed with Cramer's rule, showing that, with respect to this basis,

$$\det(\mathrm{Id} - G(z))(\mathrm{Id} - G(z))^{-1} \in \mathrm{Mat}(\dim V, \mathbb{C}[\![e_\Gamma]\!])$$

is represented by a matrix with Hahn-meromorphic entries. After the identification

$$\mathrm{Mat}(\dim V, \mathbb{C}[\![e_\Gamma]\!]) = \mathrm{Mat}(\dim V, \mathbb{C})[\![e_\Gamma]\!],$$

we see that the function $(\mathrm{Id} - G(z))^{-1}$ is Hahn-meromorphic with coefficients in $\mathrm{End}(V)$ if there is only a single point in $U^{[\sigma]}$ for which it exists. Consequently, due to (2) and (3), $(\mathrm{Id} - f(z))^{-1}$ is Hahn-meromorphic with all coefficients of $e_\gamma(z)$ with $\gamma < 0$ being of finite rank if there is only a single point in $U^{[\sigma]}$ for which $\mathrm{Id} - f(z)$ is invertible. So far we have proved the statement in $U^{[\sigma]}$. By the usual analytic Fredholm theorem, invertibility of $\mathrm{Id} - f(z)$ at a single point in Y implies that the inverse exists as a meromorphic function on Y. Conversely, we have seen that invertibility of $\mathrm{Id} - f(z)$ at a single point in $U^{[\sigma]}$ implies that $(\mathrm{Id} - f(z))^{-1}$ exists as a Hahn-meromorphic function on $U^{[\sigma]}$. By the usual meromorphic Fredholm theorem, it then exists as a Hahn-meromorphic function on Y. □

5. z-Hahn-holomorphic functions

The prominent class of Hahn-holomorphic functions is defined by convergent power series with noninteger powers.

Let $\Gamma \subset \mathbb{R}$ be a subgroup with order inherited from the standard ordering of $\mathbb{R}$. As the group G we will take the group generated by the set of functions

$$e_\alpha(z) := z^\alpha, \quad \alpha \in \Gamma,\ z \in D_r^{[\sigma]} \setminus \{0\}.$$

In this definition we choose the principal branch of the logarithm with $|\operatorname{Im}\log z| < \pi$ for $z \in \mathbb{C} \setminus (-\infty, 0]$, and, as usual, we set $\log(re^{i\varphi}) = \log r + i\varphi$, $|\varphi| < \sigma$, and $z^\alpha := e^{\alpha \log z}$.

A z-Hahn-holomorphic function f with values $\mathbb{C}$ then is a holomorphic function on $D_r^{[\sigma]} \setminus \{0\}$ such that the generalized power series

$$f(z) = \sum_\gamma a_\gamma z^\gamma, \quad a_\gamma \in \mathbb{C},$$

is normally convergent in $Y \cap D_\delta^{[\sigma]}$ for some $\delta > 0$.

Note that every well-ordered subset of $W \subset \Gamma^+$ is admissible for e, because for every $\alpha \in W$,

$$|z^\alpha| = |z|^\alpha \le |z|^{\min W}, \quad z \in D_{1/2}^{[\sigma]}. \tag{4}$$

Example 5.1. If $\Gamma = \mathbb{Z}$ and $e_k(z) = z^k$, the set of Hahn series corresponds to the formal power series and the set of z-Hahn-holomorphic functions can be identified with the set of functions that are holomorphic on the disc of radius $\delta > 0$ centered at the origin.

Example 5.2. The series

$$z^\pi \sum_{k=0}^{\infty} \frac{z^{2k}}{(2k)!}$$

converges normally on D_r for any $r > 0$ and defines a z-Hahn-holomorphic function for $\Gamma = \pi\mathbb{Z} + 2\mathbb{Z}$.

Example 5.3. Puiseux series and Levi-Civita series, as defined in, for example, [Ribenboim 1992], are special cases of Hahn series with certain $\Gamma \subset \mathbb{Q}$. When they are normally convergent, they define z-Hahn-holomorphic functions.

In the following, let $D_R = D_R^{[\infty]} \setminus \{0\}$ be the pointed disk of radius R in the logarithmic covering of the complex plane. The next result is in analogy with complex analysis, where series expansions converge normally on the maximal disc embedded in the domain of holomorphicity.

Theorem 5.4. *Let $\mathcal{R}$ be a Banach algebra, and suppose f is z-Hahn-holomorphic. Suppose further that f is bounded on $D_{\tilde{R}R+\varepsilon}$ for some ε, $\tilde{R} > 0$, and let*

$$f(z) = \sum_{\alpha \in \operatorname{supp} f} a_\alpha z^\alpha$$

be its expansion (which we do not assume converges normally on $D_{\tilde{R}}$).

Then, for all R with $0 < R < \tilde{R}$,

$$\sum_{\alpha \in \operatorname{supp} f} \|a_\alpha\| R^\alpha \le \sup_{|z|=R} \|f(z)\| \sum_{\alpha \in \operatorname{supp} f} (R/\tilde{R})^\alpha.$$

In particular, if $\sum_\alpha (R/\tilde{R})^\alpha < \infty$, the Hahn series converges normally on D_R.

Proof. As a Hahn-holomorphic function, f converges normally on $D_{2\delta}$ for some $\delta > 0$ and is holomorphic in $D_{\tilde{R}}$. Let $\Lambda_{R,L}$ be the averaging operator

$$\Lambda_{R,L}(f) = \frac{1}{2\pi i L} \int_{S_R^{(L)}} \frac{f(z)}{z}\, dz,$$

where $S_R^{(L)}(t) = Re^{i\pi t}$, $t \in (-L, L]$, is the L-fold cover of the circle with radius R. Certainly

$$\|\Lambda_{R,L}(f)\| \leq \sup_{|z|=R} \|f(z)\|.$$

Since f is holomorphic for $0 < |z| < R$, we have

$$\frac{1}{2\pi i L} \int_{S_R^{(L)}} \frac{f(z)}{z}\, dz = \frac{1}{2\pi i L} \int_{S_\delta^{(L)}} \frac{f(z)}{z}\, dz + O(L^{-1}).$$

This shows that

$$\Lambda_R(f) := \lim_{L\to\infty} \Lambda_{R,L}(f) = \lim_{L\to\infty} \sum_{\alpha\in \operatorname{supp} f} \frac{a_\alpha}{2\pi i L} \int_{S_\delta^{(L)}} z^{\alpha-1}\, dz = a_0.$$

Suppose (I_k) is a family of finite subsets of $\operatorname{supp} f$ such that

$$I_1 \subset I_2 \subset \cdots \quad \text{and} \quad \bigcup_k I_k = \operatorname{supp} f.$$

For $z \in D_{\tilde{R}}$, let $g_k(z) = \sum_{\alpha\in I_k} \lambda_\alpha z^{-\alpha} R^\alpha$, where $\lambda_\alpha \in \mathcal{R}^* := \mathcal{B}(\mathcal{R}, \mathbb{R})$ are chosen such that

$$\|\lambda_\alpha\| = 1, \quad \lambda_\alpha(a_\alpha) = \|a_\alpha\|.$$

Such λ_α exist by the Hahn–Banach theorem.

Then g_k is holomorphic in $D_{\tilde{R}}$ and $\|g_k(z)\| \leq \sum_{\alpha\in I_k} |z|^{-\alpha} R^\alpha$. Moreover,

$$\langle g_k, f\rangle(z) = \sum_{\substack{\alpha\in I_k \\ \gamma=-\alpha+\beta}} \lambda_\alpha(a_\beta) R^\alpha z^\gamma,$$

and the constant term of this function is

$$\sum_{\alpha\in I_k} \|a_\alpha\| R^\alpha = \Lambda_\delta(\langle g_k, f\rangle) = \Lambda_{\tilde{R}}(\langle g_k, f\rangle).$$

Therefore

$$\sum_{\alpha\in I_k} \|a_\alpha\| R^\alpha \leq \sup_{|z|=R} |\langle g_k, f\rangle(z)| \leq \sup_{|z|=R} \|g_k(z)\| \|f(z)\| \leq \sup_{|z|=R} \|f(z)\| \sum_{\alpha\in\operatorname{supp} f} (R/\tilde{R})^\alpha,$$

and the theorem follows by letting $k \to \infty$. □

Theorem 5.5. *Let $R > 0$, and assume $f_k \colon D_R \to V$ is a sequence of bounded z-Hahn-holomorphic functions that converge uniformly to a bounded function $f \colon D_R \to V$. Suppose that there exist constants $C > 0$, $\hat{\varepsilon} > 0$ such that, for each $k \in \mathbb{N}$,*

$$\sum_{\alpha\in\operatorname{supp} f_k} \hat{\varepsilon}^\alpha < C.$$

Suppose furthermore that there exists $I \subset \mathbb{R}$ such that $\operatorname{supp} f_k \to I$ *in the following sense. For each compact subset $K \Subset \mathbb{R}$, there exists $N > 0$ such that* $\operatorname{supp} f_k \cap K = I \cap K$ *for all $k \geq N$.*

Then f is Hahn-holomorphic on D_R with $\operatorname{supp} f \subset I$.

Proof. First, I is well ordered because $\operatorname{supp} f_k \to I$. Let $f_k(z) = \sum_{\alpha\in\operatorname{supp} f_k} a_\alpha^{(k)} z^\alpha$ be the expansion of f_k.

Let $\varepsilon > 0$. Then there exists $N_1 > 0$ such that $\|f_\ell(z) - f_k(z)\| < \varepsilon$ for all $k, \ell > N_1$ and all $z \in D_R$. Given a finite subset $\tilde I \subset I$, we can choose $N > N_1$ such that $\tilde I \cap \operatorname{supp} f_k = \tilde I$ for all $k > N$. Theorem 5.4 then shows that, for all $k, \ell > N$ and $\tilde R < \hat\varepsilon R$,

$$\sum_{\alpha\in\tilde I} \|a_\alpha^{(\ell)} - a_\alpha^{(k)}\| \cdot \tilde R^\alpha \;\le\; \sup_{|z|=\tilde R} \|f_\ell(z) - f_k(z)\| \sum_{\alpha\in\operatorname{supp} f_k \cup \operatorname{supp} f_\ell} \hat\varepsilon^\alpha < 2\,C\,\varepsilon.$$

It follows that $(a_\alpha^{(k)})_k$ is a Cauchy sequence for each α. Let $a_\alpha := \lim_{k\to\infty} a_\alpha^{(k)}$. Given a finite subset $\tilde I \subset I$ and $\varepsilon > 0$, we can find N such that $\|a_\alpha^{(k)} - a_\alpha\| < \varepsilon$ for all $k > N$, $\alpha \in \tilde I$. Then, for $|z| < \tilde R < \hat\varepsilon$,

$$\left\| \sum_{\alpha\in\tilde I} a_\alpha^{(k)} z^\alpha - \sum_{\alpha\in\tilde I} a_\alpha z^\alpha \right\| < \sum_{\alpha\in\tilde I} \|a_\alpha^{(k)} - a_\alpha\| \tilde R^\alpha < C\,\varepsilon.$$

This shows that $\sum_{\alpha\in I} a_\alpha z^\alpha$ is a Hahn series for f. By the uniform convergence of (f_k), f is analytic in $D_R^{[\sigma]} \setminus \{0\}$. Its Hahn series converges normally on $D_{\tilde R}$ because

$$\begin{aligned}\sum_{\alpha\in\tilde I} \|a_\alpha\| \tilde R^\alpha &\le \sum_{\alpha\in\tilde I} \|a_\alpha^{(\ell)}\| \tilde R^\alpha + \sum_{\alpha\in\tilde I} \|a_\alpha^{(k)} - a_\alpha\| \tilde R^\alpha + \sum_{\alpha} \|a_\alpha^{(k)} - a_\alpha^{(\ell)}\| \tilde R^\alpha \\ &\le \sum_{\alpha\in\operatorname{supp} f_\ell} \|a_\alpha^{(\ell)}\| \tilde R^\alpha + C\varepsilon + 2C\varepsilon < \infty\end{aligned}$$

for all finite $\tilde I \subset I$, ℓ sufficiently large, and $k \gg \ell$ depending on $\tilde I$. □

6. $z \log z$-Hahn-holomorphic functions

In the following, let $\mathbb{R}^2$ be equipped with the lexicographical order and let $\Gamma \subset \mathbb{R}^2$ be a subgroup with order inherited from that of $\mathbb{R}^2$. Let $Y = D_{1/2}^{[\sigma]}$ for fixed $\sigma > 0$. The group G will be generated by

$$e_{(\alpha,\beta)}(z) := z^\alpha (-\log z)^{-\beta}, \quad (\alpha, \beta) \in \Gamma,\ |z| < 1.$$

With the inclusion $\mathbb{R} \times \{0\} \subset \mathbb{R}^2$, this comprises the power functions z^α from Section 5. Note that

$$\lim_{z\to 0} e_{(\alpha,\beta)}(z) = 0 \iff \alpha > 0 \vee (\alpha = 0 \wedge \beta > 0),$$

which is equivalent to $(\alpha, \beta) > (0, 0)$ in the lexicographical ordering of $\mathbb{R}^2$. The monotonicity (4) of power functions z^α has to be replaced by the following "weak monotonicity" property.

Lemma 6.1. *Let $\mathcal{S} \subset \Gamma^+ = \{\gamma \in \Gamma \mid \gamma > 0\}$ be a set such that there exists an $N \in \mathbb{N}_0$ with*

$$-\beta \le N\alpha \quad \textit{for all } (\alpha, \beta) \in \mathcal{S}. \tag{$*$}$$

(a) *There exists $r_N < 1$ such that, for $(\alpha, \beta) \in \mathcal{S}$ and $|\theta| < \sigma$, the function*

$$r \mapsto |re^{i\theta}|^\alpha |\log(re^{i\theta})|^{-\beta}$$

is monotonically increasing on $[0, r_N)$.

(b) *Given x with $0 < x < r_N$, there exists $\rho_N(x) \le x$ such that, for all z with $0 \le |z| \le \rho_N(x)$ and $|\arg z| < \sigma$, we have*

$$(\alpha, \beta) \in \mathcal{S} \implies |e_{(\alpha,\beta)}(z)| \le e_{(\alpha,\beta)}(x)$$

The proof is elementary and will be omitted.

It is not difficult to see that if $\mathcal{S}$ satisfies (∗), a similar inequality holds for the set $(\mathcal{S} - A) \cap \Gamma^+$, where $A \subset \mathcal{S}$ and the constant N depends on A. Thus a set $\mathcal{S}$ with (∗) is admissible for e.

Now the assumptions from Section 2 are all satisfied and we can consider Hahn-holomorphic and meromorphic functions: A *$z \log z$-Hahn-holomorphic function* with values in a Banach algebra $\mathcal{R}$ is defined by a normally convergent series

$$f(z) = \sum_{(\alpha,\beta)\in\Gamma} a_{(\alpha,\beta)} z^\alpha (-\log z)^{-\beta}, \quad a_{(\alpha,\beta)} \in \mathcal{R},\ z \in D_{1/2}^{[\sigma]},$$

such that $\operatorname{supp}(f)$ is contained in a set $\mathcal{S} \cup \{(0,0)\}$ with $\mathcal{S}$ as in Lemma 6.1.

Note that the property (∗) is invariant under addition and multiplication of Hahn-holomorphic functions, so that $z \log z$-Hahn-holomorphic functions indeed are a ring, and all results from Section 2 apply.

Example 6.2. The series

$$\sum_{n=0}^{\infty} z^n (-\log z)^n = (1 + z \log z)^{-1}$$

is a Hahn series in $\Gamma = \mathbb{Z} \times \mathbb{Z}$ with support $\{(n, -n) \mid n \in \mathbb{N}_0\}$. It converges normally on the set $\left\{z \in \mathcal{X} \;\middle|\; |z \log z| < \frac{1}{2}\right\}$ and therefore defines a $z \log z$-Hahn-holomorphic function on $D_r^{[\sigma]}$ for any $\sigma > 0$ and sufficiently small $r = r(\sigma)$.

Example 6.3. The formal series

$$\sum_{n=0}^{\infty} \frac{1}{n!} z(-\log z)^n$$

is *not* a Hahn series for $\Gamma = \mathbb{Z} \times \mathbb{Z}$, because the support

$$\{(1, -n) \mid n \in \mathbb{N}_0\}$$

is not a well-ordered subset of Γ.

Example 6.4. The logarithm $\log z = \dfrac{z \log z}{z}$ is Hahn-meromorphic for $\Gamma \subset \mathbb{Z} \times \mathbb{Z}$.

Example 6.5. The series

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \frac{1}{m^2} z^n (-\log z)^{2n-1+1/m}$$

defines a $z \log z$-Hahn-holomorphic function in a neighborhood $D_\varepsilon^{[\sigma]}$ of 0 for any $\sigma > 0$ and for small enough $\varepsilon = \varepsilon(\sigma)$ with $\Gamma = \mathbb{Z} \times \mathbb{Q}$. Its support is

$$\{(n, 1 - 2n - 1/m) \mid n, m \in \mathbb{N}\}.$$

7. Applications: Hahn-meromorphic continuation of resolvent kernels

7A. Suppose that $\nu \geq 0$. The differential operator B_ν associated to the Bessel differential equation in its Liouville normal form,

$$B_\nu := -\frac{\partial^2}{\partial x^2} + \frac{\nu^2 - \frac{1}{4}}{x^2}\mathrm{Id}, \tag{5}$$

is a nonnegative symmetric operator on the space $C_c^\infty((0,\infty))$ equipped with the inner product inherited from $L^2((0,\infty), dx)$. We will denote the Friedrichs extension of B_ν by the same symbol B_ν.

The kernel $r_\lambda^{(\nu)}$ of the resolvent $(B_\nu - \lambda^2)^{-1}$ can be constructed directly out of the fundamental system of the corresponding Sturm–Liouville equation and this results in (see, for example, [Brüning and Seeley 1987, p. 371])

$$r_\lambda^{(\nu)}(x,y) = \frac{\mathrm{i}\pi}{2}\sqrt{xy}\cdot J_\nu(\lambda \min(x,y)) H_\nu^{(1)}(\lambda \max(x,y)), \quad 0 < x, y < \infty, \tag{6}$$

where $H_\nu^{(1)}$ is the Hankel function of order ν of the first kind and J_ν is the Bessel function.

The proof of the following lemma uses the well-known expansion of Bessel and Hankel functions, and will be given at the end of this section.

Lemma 7.1. *For every $\nu > 0$ and $(x,y) \in (0,\infty)\times(0,\infty)$, the kernel $\lambda \mapsto r_\lambda^{(\nu)}(x,y)$ defines a $z\log z$-Hahn-holomorphic function.*

(a) *For $\nu \in \mathbb{R}_+ \setminus \mathbb{N}_0$, we have*

$$r_\lambda^{(\nu)}(x,y) = \lambda^{2\nu} f_1^{(\nu)}(x,y)(\lambda) + f_2^{(\nu)}(x,y)(\lambda),$$

where the maps $\lambda \mapsto f_j^{(\nu)}(x,y)(\lambda)$ are even and entire. In particular, $r_\lambda^{(\nu)}(x,y)$ is z-Hahn-holomorphic with support contained in $2\mathbb{Z} + 2\nu\mathbb{Z}$.

Let $a_{j;2k}^{(\nu)}(x,y)$ be the coefficient of λ^{2k} in the Taylor series expansion of $f_j^{(\nu)}(x,y)$.

There is a constant C_1 such that, for $0 \leq x \leq y$,

$$|a_{1;2k}^{(\nu)}(x,y)| \leq R^{-2k} C_1(\nu)(xy)^{\nu+1/2} e^{R(x+y)}, \quad R > 0.$$

For $c > 0$ and every $r_0 > 0$, there is a constant C_2 such that, for all $y \geq x \geq c$,

$$|a_{2;2k}^{(\nu)}(x,y)| \leq R^{-2k}\frac{C_2(\nu, r_0)}{\sqrt{R}}\sqrt{x}(x/y)^\nu e^{R(x+y)}, \quad R \geq r_0.$$

(b) *The kernel $\lambda \mapsto r_\lambda^{(\nu)}(x,y)$ is a $z\log z$-Hahn-holomorphic function with support contained in $2\mathbb{Z}\times\mathbb{Z}$ if $\nu = n \in \mathbb{N}$:*

$$r_\lambda^{(n)}(x,y) = \log(\lambda) g_1^{(n)}(x,y)(\lambda) + g_2^{(n)}(x,y)(\lambda),$$

where the maps $\lambda \mapsto g_j^{(\nu)}(x,y)(\lambda)$ are even and entire.

The coefficients $b_{j;2k}^{(n)}(x,y)$ in its Hahn-series expansion can be estimated by

$$|b^{(n)}_{1;2k}(x,y)| \le R^{-2k}\sqrt{xy}\frac{(R/2)^{2n}}{(n!)^2}e^{R(x+y)}, \qquad R>0,$$

$$|b^{(n)}_{2;2k}(x,y)| \le R^{-2k}e^{R(x+y)}\left(\hat{c}_1 x^{n+1}y^{n+1/2}\frac{(R/2)^{2n}}{n!\,(n-1)!}+c_2\right), \quad R>0.$$

Remark 7.2. For $\nu=0$, the expansion (11) below gives

$$r^{(0)}_\lambda(x,y) = -\sqrt{xy}\log\frac{\lambda y}{2} + \underline{h}(x,y)(\lambda)$$

with a Hahn-holomorphic function $\underline{h}(x,y)$. In particular, $\lambda\mapsto r^{(0)}_\lambda(x,y)$ is $z\log z$-Hahn-meromorphic.

For $c>0$, let $\chi_c:[0,\infty)\to\mathbb{R}_+$ be a smooth cutoff function with

$$\chi_c(x)=\begin{cases}0 & \text{if } x\le c,\\ 1 & \text{if } x\ge 2c.\end{cases}$$

Multiplication by this function defines a bounded operator on $L^2((0,\infty))$. The "restricted resolvent" $\chi_c(B_\nu-\lambda^2)^{-1}\chi_c$ then is the bounded operator on $L^2((0,\infty))$ with integral kernel

$$(\chi_c\circ r^{(\nu)}_\lambda)(x,y) := \chi_c(x)\cdot r^{(\nu)}_\lambda(x,y)\cdot\chi_c(y).$$

Proposition 7.3. *Let $I=(0,\infty)$, $\nu>0$, and $c>0$. For any $\kappa>0$ and $\sigma>0$, the restricted resolvent $\chi_c(B_\nu-\lambda^2)^{-1}\chi_c$ extends, as a function of λ, to a $z\log z$-Hahn-holomorphic function on some neighborhood $D^{[\sigma]}_r$ of 0 with values in the compact operators*

$$\mathcal{K}\big(L^2(I,e^{\kappa x}\,dx),L^2(I,e^{-\kappa x}\,dx)\big).$$

Proof. First let $\nu\notin\mathbb{N}_0$. In Lemma 7.1(a), let $r_0=R=\kappa/3$. Using

$$\int_c^\infty\int_c^\infty \min(x,y)\left(\frac{\min(x,y)}{\max(x,y)}\right)^{2\nu}e^{(2R-\kappa)(x+y)}\,dx\,dy \le C(\kappa) \tag{7a}$$

and

$$\int_c^\infty\int_c^\infty (xy)^{2\nu+1}e^{(2R-\kappa)(x+y)}\,dx\,dy \le \left(\frac{\Gamma(2+2\nu)}{(\kappa/3)^{2+2\nu}}\right)^2, \tag{7b}$$

it is easy to see that the coefficients $a^{(\nu)}_{j;2k}(x,y)$ of the Hahn series expansion of $r^{(\nu)}$ satisfy

$$|\chi_c\circ a^{(\nu)}_{j;2k}|\in L^2(I\times I, e^{-\kappa(x+y)}\,dx\otimes dy), \quad j=1,2.$$

Therefore the kernels $\{\chi_c\circ a^{(\nu)}_{j;2k}\}_k$ define Hilbert–Schmidt operators

$$A^{(\nu)}_{j;2k}: L^2(I,e^{\kappa x}\,dx)\to L^2(I,e^{-\kappa x}\,dx)=:\mathcal{H}_\kappa$$

with norm bounded from above by

$$\|A^{(\nu)}_{j;2k}\| \le \|\chi_c\circ a^{(\nu)}_{j;2k}\|_{\mathcal{H}_\kappa\times\mathcal{H}_\kappa} \le R^{-2k}C(\nu,\kappa), \tag{8}$$

where $C(\nu,\kappa)$ can be obtained from (10a), (7a), (10b), (7b). But then the series

$$\lambda^{2\nu}\sum_{k=0}^{\infty}\|A^{(\nu)}_{1;2k}\|\,|\lambda|^{2k}+\sum_{k=0}^{\infty}\|A^{(\nu)}_{2;2k}\|\,|\lambda|^{2k}$$

converges normally in some neighborhood $U\subset D_r^{[\sigma]}$ of 0 and the kernel $r^{(\nu)}_\lambda$ defines a z-Hahn-holomorphic family of Hilbert–Schmidt operators in

$$\mathcal{H}\big(L^2(I,e^{\kappa x}\,dx),L^2(I,e^{-\kappa x}\,dx)\big).$$

For integral $\nu=n\in\mathbb{N}$, we can argue similarly, using Lemma 7.1(b). □

Proof of Lemma 7.1. First let $\nu\notin\mathbb{N}_0$. Recall that $J_\nu(z)=\left(\frac{z}{2}\right)^\nu h_\nu(z)$, where

$$h_\nu(z)=\sum_{k=0}^{\infty}a_k^{(\nu)}z^{2k}\quad\text{with } a_k^{(\nu)}=\frac{(-1)^k}{4^k\,k!\Gamma(k+\nu+1)}.$$

The function h_ν is entire. We have

$$H^{(1)}_\nu(z)=\frac{\mathrm{i}}{\sin\nu\pi}(J_\nu(z)e^{-\mathrm{i}\nu\pi}-J_{-\nu}(z)),\quad H^{(1)}_n(z)=\lim_{\nu\to n}H^{(1)}_\nu(z),\quad n\in\mathbb{Z}.$$

(a) Let $x\le y$. Then

$$-\frac{2\mathrm{i}}{\pi}r^{(\nu)}_\lambda(x,y)=\sqrt{xy}J_\nu(\lambda x)H^{(1)}_\nu(\lambda y)=\lambda^{2\nu}f^{(\nu)}_1(x,y)(\lambda)+f^{(\nu)}_2(x,y)(\lambda)$$

with even, analytic functions in λ:

$$f^{(\nu)}_1(x,y)(\lambda)=\frac{\mathrm{i}e^{-\mathrm{i}\nu\pi}}{4^\nu\sin\nu\pi}(xy)^{\nu+1/2}h_\nu(\lambda x)h_\nu(\lambda y)$$
$$f^{(\nu)}_2(x,y)(\lambda)=\frac{-\mathrm{i}}{\sin\nu\pi}\sqrt{xy}\left(\frac{x}{y}\right)^\nu h_\nu(\lambda x)h_{-\nu}(\lambda y)$$

Due to Cauchy's integral formula,

$$|a^{(\nu)}_{j,2k}(x,y)|\le R^{-2k}\sup_{|\lambda|=R}|f^{(\nu)}_j(x,y)|,\quad j=1,2.$$

We know from, for example, [Olver and Maximon 2011, (10.14.4)] that, for $\nu\ge0$,

$$|h_\nu(z)|\le\frac{e^{|\operatorname{Im}z|}}{\Gamma(\nu+1)}.\tag{9}$$

Using $J_\nu=\frac12(H^{(1)}_\nu+H^{(2)}_\nu)$, $|H^{(*)}_{-\nu}|=|H^{(*)}_\nu|$, and that $h_{-\nu}$ is a holomorphic and even function,

$$(R/2)^\nu\sup_{|z|=R}|h_{-\nu}(z)|=\sup_{\substack{|z|=R\\ \operatorname{Re}z>0}}|J_{-\nu}(z)|\le\sup_{\substack{|z|=R\\ \operatorname{Re}z>0}}\tfrac12(|H^{(1)}_\nu(z)|+|H^{(2)}_\nu(z)|).$$

But from [Olver and Maximon 2011, (10.17.13)], for $-\pi/2<\arg z<\pi/2$,

$$|H^{(1;2)}_\nu(z)|\le\sqrt{\frac{2}{\pi|z|}}e^{\mp\operatorname{Im}z}(1+\tau_\nu(|z|)e^{\tau_\nu(|z|)}),\quad \tau_\nu(s):=\frac{\pi}{2}\left|\nu^2-\tfrac14\right|\cdot s^{-1}.$$

Thus there exists a constant $C_1 > 0$ with

$$\sup_{\substack{|\lambda|=R\\ \operatorname{Re}\lambda>0}} |H_\nu^{(1;2)}(\lambda y)| \le C_1 \frac{e^{Ry}}{\sqrt{Ry}}(1+\tau(R)e^{\tau(R)}), \quad y \ge c, \quad \tau(R) := \frac{\pi}{2}\frac{\left|\nu^2-\frac{1}{4}\right|}{Rc}.$$

This shows that, for every $r_0 > 0$, there is a constant C such that, for every $R \ge r_0$,

$$|a_{2;2k}^{(\nu)}(x,y)| \le R^{-2k-\nu}\frac{C_\nu}{\sqrt{R}}\sqrt{x}(x/y)^\nu e^{R(x+y)}, \quad C_\nu := \frac{C\cdot 2^\nu(1+\tau(r_0)e^{\tau(r_0)})}{|\sin(\nu\pi)|\Gamma(\nu+1)}. \tag{10a}$$

Also from (9),

$$|a_{1;2k}^{(\nu)}(x,y)| \le R^{-2k}\frac{(xy)^{\nu+1/2}e^{R(x+y)}}{4^\nu|\sin\nu\pi|\,\Gamma(\nu+1)^2}. \tag{10b}$$

(b) Let $\nu = n \in \mathbb{N}$ and $x \le y$. Then, from $H_n^{(1)} = J_n + \mathrm{i}Y_n$ and [Olver and Maximon 2011, (10.8.1)],

$$\frac{-2\mathrm{i}}{\pi\sqrt{xy}}r_\lambda^{(n)}(x,y) = \frac{2\mathrm{i}}{\pi}\Big(\log\lambda+\log\frac{y}{2}\Big)J_n(\lambda x)J_n(\lambda y)+J_n(\lambda x)J_n(\lambda y)-\frac{\mathrm{i}}{\pi}h_n(\lambda x)\sum_{k=0}^{n-1}\frac{(n-k-1)!}{k!}\Big(\frac{\lambda y}{2}\Big)^{2k}$$
$$-J_n(\lambda x)\Big(\frac{\lambda y}{2}\Big)^n\frac{1}{\pi}\sum_{k=0}^{\infty}\frac{\psi(k+1)+\psi(n+k+1)}{k!\,(n+k)!}(-1)^k(\lambda y/2)^{2k} \tag{11}$$

with $\psi(x) = \Gamma'(x)/\Gamma(x)$. The only logarithmic terms in the Hahn-series expansion of $r^{(n)}(x,y)$ are $e_{(2k,-1)}(\lambda)$, $k \ge n$. Because of (9), the coefficient of $e_{(2k,-1)}$ is bounded by

$$R^{-2k}\sqrt{xy}\frac{(R/2)^{2n}}{(n!)^2}e^{R(x+y)}, \quad R > 0.$$

From Stirling's inequalities for Γ, we obtain, for $0 \le k \to \infty$,

$$1 \ge \sqrt{k+1}\frac{(2k)!}{4^k(k!)^2} \sim \frac{1}{\sqrt{\pi}}; \tag{12}$$

hence

$$\frac{(n-k-1)!\,(2k)!}{4^k k!n!} = \frac{1}{(n-k)\binom{n}{k}}\cdot\frac{(2k)!}{4^k(k!)^2} \le \frac{1}{n}, \quad 0 \le k < n.$$

Because $|z|^{2k} \le (2k)!\,e^{|z|}$ and because of (9), the norm of the sum in the second line of (11) can be bounded by $(e^{|\operatorname{Im}\lambda|x}/\pi)e^{|\lambda y|}$.

For the last line in (11), we first note that the polygamma function is monotonically increasing and $\psi(k) \lesssim \log k$, $k > 0$, and estimate as above

$$\sum_{k=0}^{\infty}\left|\frac{\psi(k+1)+\psi(n+k+1)}{k!\,(n+k)!4^k}\right||\lambda y|^{2k} \le e^{|\lambda y|}\sum_{k=0}^{\infty}\frac{1}{\sqrt{k+1}}\frac{2\log(n+k+1)}{(k+n)\cdot(k+n-1)\cdots(k+1)}$$
$$\le 2e^{|\lambda y|}\sum_{k=0}^{\infty}\frac{1}{\sqrt{k+1}(k+n)^{2/3}}\frac{\log(n+k+1)}{(k+n)^{1/3}}\frac{1}{(n-1)!} \le \frac{c_1\,e^{|\lambda y|}}{(n-1)!},$$

where the constant c_1 can be obtained from $\zeta\left(\frac{7}{6}\right)$ and $x^{-1/3}\log(x+1) < \frac{3}{2}$ for $x > 0$.

Altogether, this shows that the coefficient of λ^{2k} in $r_\lambda^{(\nu)}(x, y)$ is bounded by

$$C R^{-2k}\sqrt{xy}e^{R(x+y)}\left(\left(\log\frac{y}{2}+1\right)\frac{(xy)^n(R/2)^{2n}}{(n!)^2}+\frac{1}{\pi}+c_1(xy)^n\frac{(R/2)^{2n}}{n!\,(n-1)!}\right), \quad R>0,$$

and, for $y \geq x \geq c$, this is smaller than

$$R^{-2k}e^{R(x+y)}\left(\hat{c}_1 x^{n+1}y^{n+1/2}\frac{(R/2)^{2n}}{n!\,(n-1)!}+c_2\right), \quad R>0. \qquad \square$$

7B. ***The resolvent of the Laplace operator on cones.*** Let $Z = (0, \infty) \times M$ be equipped with the cone metric $g^Z = dx^2 + x^2 g^M$, where (M, g^M) is a compact n-dimensional Riemannian manifold (without boundary); we will call Z a *cone*. We consider the Friedrichs extension Δ of the Laplace operator on compactly supported functions $C_0^\infty(Z)$ to $L^2(Z, g^Z)$. Under the isometry

$$\Psi : L^2(Z, dx^2 + g^M) \to L^2(Z, g^Z), \quad f(x, p) \mapsto x^{-n/2} f(x, p),$$

this Laplacian becomes

$$\widetilde{\Delta} := \Psi^{-1} \circ \Delta \circ \Psi = -\frac{\partial^2}{\partial x^2} + \frac{1}{x^2}\left(\Delta_M + \frac{n}{2}\left(\frac{n}{2}-1\right)\right).$$

Let $\{\mu_k\}$ be the eigenvalues of the Laplace operator Δ_M on $L^2(M)$, and define $\nu_k := \nu(\mu_k)$ as the positive solution of $\nu_k^2 - \frac{1}{4} = (n/2)(n/2-1) + \mu_k$. Let V be the set of these solutions and let $\{\phi_\nu\}_{\nu\in V}$ be the corresponding orthonormal Hilbert space basis of $L^2(M)$ consisting of eigenfunctions of Δ_M such that $\Delta_M\phi_{\nu(\mu)} = \mu\phi_{\nu(\mu)}$.

For a smooth function $f(x, p) = \sum_{\nu\in V} f_\nu(x)\phi_\nu(p) \in L^2(Z)$, we obtain

$$\Psi^{-1}(\Delta - \lambda^2)\Psi f(x, p) = \sum_{\nu\in V}((B_\nu - \lambda^2) f_\nu)(x)\phi_\nu(p),$$

where B_ν is the Bessel operator defined in (5).

Let $\lambda \in \mathbb{C}$ with $\operatorname{Im}\lambda > 0$; in particular, λ^2 lies in the resolvent set of Δ. Then the integral kernel of the resolvent $(\widetilde{\Delta} - \lambda^2)^{-1}$ is given by

$$K((x, p), (y, q), \lambda) = \sum_{\nu\in V} r^{(\nu)}(x, y)(\lambda)\phi_\nu(q) \otimes \phi_\nu(p), \tag{13}$$

where $r^{(\nu)}$ is defined in (6). Recall from Lemma 7.1 that $r^{(\nu)}$ is a $z\log z$-Hahn-holomorphic function,

$$r^{(\nu)}(x, y)(\lambda) = \sum_{\gamma\in\tilde{S}_\nu\subset\mathbb{R}^2} a_\gamma^{(\nu)}(x, y)e_\gamma(\lambda), \quad e_{(\alpha,\beta)}(\lambda) := \lambda^\alpha(-\log\lambda)^{-\beta},$$

where $\tilde{S}_\nu$ is the Hahn series support of $r^{(\nu)}$. In this expansion, logarithmic terms occur only for $\nu \in \mathbb{N}$.

Take $\mathcal{G} \subset \mathbb{R}^2$ to be the group generated by $\bigcup_\nu S_\nu$ for $S_\nu := \bigcup_{x,y\in(0,\infty)} \tilde{S}_\nu(x, y)$. Then it is clear that the

resolvent kernel is a Hahn series with support in $\mathcal{G} \subset \mathbb{R}^2$:

$$\begin{aligned} K((x,p),(y,q),\lambda) &:= \sum_{\nu\in V} r^{(\nu)}(x,y)(\lambda)\phi_\nu(q)\otimes\phi_\nu(p) = \sum_{\nu\in V}\sum_{\gamma\in S_\nu} a_\gamma^{(\nu)}(x,y)e_\gamma(\lambda)\phi_\nu(q)\otimes\phi_\nu(p) \\ &= \sum_{\gamma\in\mathcal{G}} e_\gamma(\lambda)\Bigg(\sum_{\nu\in V:\gamma\in S_\nu} a_\gamma^{(\nu)}(x,y)\phi_\nu(q)\otimes\phi_\nu(p)\Bigg) \\ &=: \sum_{\gamma\in\mathcal{G}} \tilde{a}_\gamma((x,p),(y,q))e_\gamma(\lambda), \end{aligned} \tag{14}$$

where we have set $\tilde{a}_\gamma = 0$ if $\gamma \notin \bigcup_{\nu\in V} S_\nu$.

To show normal convergence of the operator-valued series defined by (14), we will make the additional assumption that each $\nu \in V$ either is an integer, or is not "too close" to an integer in the following sense.

Definition 7.4. For $\kappa > 0$, a family of orders $V \subset \mathbb{R}_{\geq 0}$ is called *κ-suitable* if the set

$$\left\{ \frac{1}{(2\kappa)^\nu \sin(\nu\pi)\Gamma(\nu+1)} \;\middle|\; \nu \in V \setminus \mathbb{N} \right\} \tag{15}$$

is bounded.

Example 7.5. For $M = S^n$, $n \geq 1$, the n-sphere equipped with the standard metric, it is well known (see, for example, [Shubin 2001, Section 22]) that the eigenvalues of the Laplace operator on functions are $\mu_k = k(k+n-1)$, $k \in \mathbb{N}_0$ with multiplicity

$$m_k := \binom{n+k}{n} - \binom{n+k-2}{n}.$$

Then $\nu_k := \nu(\mu_k) := (n-1)/2 + k$ is (half-)integral for odd (even) n and $V = (\nu_0, \nu_1, \ldots, \nu_1, \nu_2, \ldots)$, where each ν_j appears m_j times. For $n \geq 2$, all $\nu(\mu_k)$ are positive.

In Section 7A we defined a smooth cutoff function χ, which can be extended to a bounded operator χ on $L^2((0,\infty)\times M, dx)$ by setting

$$\chi(f)(x,p) = \chi(x)f(x,p) \quad \text{for } f \in C_0^\infty(Z),\ x \in (0,\infty),\ p \in M,$$

and taking the closure.

Here is the main result of this section:

Theorem 7.6. *Let $c, \sigma, \kappa > 0$ and assume that the family $V = \{\nu\}$ of orders is κ-suitable. Then the restricted resolvent $\chi_c(\widetilde{\Delta} - \lambda^2)^{-1}\chi_c$ extends, as a function of λ, to a $z\log z$-Hahn-meromorphic function on some $D_r^{[\sigma]}$ with values in*

$$\mathcal{H}\big(L^2(Z, e^{\kappa x^2}\,dx\otimes \mathrm{vol}_M), L^2(Z, e^{-\kappa x^2}\,dx\otimes\mathrm{vol}_M)\big), \tag{16}$$

where the only term $\lambda^\alpha(-\log\lambda)^{-\beta}$ in its Hahn-series expansion with $(\alpha,\beta) < 0$ that possibly has a nonzero coefficient is the one with $(\alpha,\beta) = (0,-1)$, and its coefficient has finite rank. If V does not contain $\nu = 0$, then, in a (possibly smaller) neighborhood of zero, this function is Hahn-holomorphic.

Proof. Let A_γ be the operator on $L^2(Z)$ defined by the "restricted kernel"

$$(\chi_c \circ \tilde{a}_\gamma)((x,q),(y,q)) := \chi_c(x)\tilde{a}_\gamma((x,p),(y,q))\chi_c(y)$$

with $\tilde{a}$ from (14), so that $\sum_{\gamma\in\mathcal{G}} A_\gamma e_\gamma(\lambda)$ is the Hahn series of the restricted resolvent.

As in (8), in the proof of Proposition 7.3, we can estimate

$$\|\chi_c \circ \tilde{a}_\gamma^{(\nu)}\| \le R^{-k(\gamma)} C(\nu,\kappa);$$

now instead of (7b) we choose $R < c\kappa/4$ and use

$$\int_c^\infty x^{2\nu+1} e^{2Rx-\kappa x^2}\,dx \le \int_c^\infty x^{2\nu+1} e^{-(\kappa/2)x^2}\,dx \le \frac{\Gamma(\nu+1)}{2(\kappa/2)^{\nu+1}}. \tag{17}$$

Because the family V is κ-suitable, the constants $C(\nu,\kappa)$ are bounded in ν. Thus the kernel A_γ defines a Hilbert–Schmidt operator

$$A_\gamma : L^2(Z, e^{\kappa x^2}\,dx\otimes \mathrm{vol}_M) \to L^2(Z, e^{-\kappa x^2}\,dx\otimes \mathrm{vol}_M)$$

between weighted L^2-spaces, with norm bounded by

$$\|A_\gamma\| \le \|\chi_c \circ \tilde{a}_\gamma\| \le \sup_{\nu:\gamma\in\operatorname{supp} r_\nu} \|\chi_c \circ a_\gamma^{(\nu)}\| \le C$$

for all $\gamma \in \bigcup_\nu S_\nu$.

Therefore the Hahn series $\sum_{\gamma\in\mathcal{G}} A_\gamma e_\gamma(\lambda)$ is normally convergent in $D_\delta^{[\sigma]}$ for some $\delta > 0$, provided that

$$\sum_{\gamma\in\mathcal{S}} \|e_\gamma\|_\delta < \infty.$$

Due to Lemma 7.1, the support $\mathcal{S}$ is given by

$$\mathcal{S} = \bigcup_\nu \operatorname{supp} r_\nu = \mathcal{S}_\mathrm{r} \cup \mathcal{S}_\mathrm{i} \subset \mathbb{R}_+ \times (-\mathbb{N}_0), \tag{18}$$

where $\mathcal{S}_\mathrm{i}$ and $\mathcal{S}_\mathrm{r}$ correspond to integer and noninteger real coefficients ν. Furthermore, elements in $\mathcal{S}_\mathrm{i}$ are of the form $(2sn+2\ell, -s)$ with $\ell\in\mathbb{N}_0$, $s\in\{0,1\}$, and those in $\mathcal{S}_\mathrm{i}$ have the form $(2s\nu+2\ell, 0)$, $\ell\in\mathbb{N}_0$, $s\in\{0,1\}$ for ν noninteger.

For $0<|\lambda|<\delta<1$ and $\nu\in V\setminus\mathbb{N}_0$,

$$\sum_{\gamma\in\mathcal{S}_\mathrm{r}} |e_\gamma(\lambda)| \le \sum_{\ell\in\mathbb{N}_0} |\lambda^{2\ell}| + \sum_\nu \sum_{\ell\in\mathbb{N}_0} |\lambda^{2\ell+2\nu}| \le \frac{1}{1-\delta^2}\left(1+\sum_\nu |\lambda^{2\nu}|\right).$$

Now, from Weyl's formula, we obtain that there exists an $R>0$ such that $\sum_{\nu\in V} R^\nu < \infty$. This shows that, for $|\lambda| < \min(\delta, \sqrt{R})$, the partial series $\sum_{\gamma\in\mathcal{S}_\mathrm{r}} |e_\gamma(\lambda)|$ converges absolutely.

Finally, for $\nu = n \in \mathbb{N}$, we use $|\log\lambda||\lambda|^{2k} \le C_\sigma |\lambda|^{2k-1}$ to estimate $\sum_{\gamma\in\mathcal{S}_\mathrm{i}} |e_\gamma(\lambda)|$ by the geometric series.

Note that the only term that gives rise to a nonzero coefficient of e_γ with $\gamma<0$ is the order $\nu=0$. □

Remark 7.7. From the proof of Theorem 7.6, it is clear that a similar statement holds if the weights in (16) are replaced by $e^{\pm\kappa x^{1+\varepsilon}}$ for any $\varepsilon > 0$.

For the Laplace operator on differential forms $L^2(S^n, \Lambda^* T^* S^n)$, the eigenvalues μ are integers (cf. [Ikeda and Taniguchi 1978, Theorem 4.2]), and the corresponding $\nu = \nu(\mu_k, p)$ are square roots of integers. In this case we have:

Lemma 7.8. *Any family $V = (\sqrt{q_i}\,)_i$ with $q_i \in \mathbb{N}_0$ is κ-suitable for every $\kappa > 0$.*

Proof. First we show for $n \in \mathbb{N}_0$ and q with $n^2 < q < (n+1)^2$ that

$$\min\{\sqrt{q} - n, (n+1) - \sqrt{q}\} > \frac{1}{2(n+1)}.$$

We then use $|\sin x\pi| > 2|x|$ for $0 < |x| < \frac{1}{2}$ to prove that, for $\nu = \sqrt{q}$,

$$\frac{1}{(\nu+1)|\sin \nu\pi|} < 1, \quad \frac{1}{\nu|\sin \nu\pi|} < \frac{3}{2}.$$

Together with Stirling's formula for the asymptotics of $\Gamma(\nu)$, this shows the boundedness of (15). □

Therefore Theorem 7.6 has a straightforward extension to differential forms. A similar statement can be proven for the Laplacian acting on differential forms on $(0, \infty) \times P^n(\mathbb{C})$, where $P^n(\mathbb{C})$ is equipped with the Fubini–Study metric. The eigenvalues for the Laplace operator on sections of $\Lambda^p T^* P^n(\mathbb{C})$ have been computed in [Ikeda and Taniguchi 1978, Theorem 5.2], they are integers.

7C. ***The resolvent of the Laplace operator on compact perturbations of $\mathbb{R}^n$ or conic spaces.*** Set $Z = (0, \infty) \times M$ and let (Z, g^Z) be a cone as defined in the previous section. Let X be a Riemannian manifold that is isometric to Z away from a compact set. This means that, for some $a > 0$, we can identify X with $X = X_a \cup_{M_a} Z_a$, where $Z_a = [a, \infty) \times M$, $M_a = \{a\} \times M$, and X_a is a compact Riemannian manifold with boundary M_a.

In this section we denote by Δ_0 the self-adjoint operator on the cone that is obtained from the Friedrichs extension of the Laplace operator on $C_0^\infty(Z)$. Let Δ be the Laplace operator acting on compactly supported functions on X, and let L be a formally self-adjoint first order differential operator that is compactly supported in X_a for some $a > 0$. Then, of course $P := \Delta + L$ is of Laplace type and therefore essentially self-adjoint on compactly supported smooth functions. We will denote its self-adjoint extension by the same symbol P whenever there is no danger of confusion. It follows from standard results in perturbation theory that the essential spectrum of P equals the essential spectrum of the Laplace operator on the cone, namely, $[0, \infty)$. Moreover, it is well known that the distributional kernel of the resolvent $(P - \lambda^2)^{-1}$ has a meromorphic continuation across the spectrum away from the point $\lambda = 0$. Now Theorem 4.1 allows us to refine this statement and show that the resolvent kernel is Hahn-meromorphic at $\lambda = 0$ if this is true for the (restricted) kernel of $(\Delta_0 - \lambda^2)^{-1}$. The precise statement is formulated in the following theorem.

Theorem 7.9. *Let $a > 0, \kappa > 0$ and suppose that, for some $\sigma > 0$, the restricted resolvent $\chi_a(\Delta_0 - \lambda^2)^{-1}\chi_a$ extends, as a function of λ, to a $z \log z$-Hahn-meromorphic function on $D_r^{[\sigma]}$ with values in*

$$\mathcal{H}\big(L^2(Z, e^{\kappa x^2}\, dx \otimes \mathrm{vol}_M), L^2(Z, e^{-\kappa x^2}\, dx \otimes \mathrm{vol}_M)\big)$$

for a group $\Gamma \subset \mathbb{Z} \times \mathbb{Z}$, such that the range of the coefficients of e_γ with $\gamma < 0$ of its Hahn series are finite-rank operators with range contained in a fixed finite-dimensional subspace. Let $\widetilde{\Gamma}$ be the subgroup of $\mathbb{Z} \times \mathbb{Z}$ generated by Γ and $2\mathbb{Z} \times \{0\}$. Then $(P - \lambda^2)^{-1}$ has an extension, as a function of λ, to a $z \log z$ Hahn-meromorphic function on $D_r^{[\sigma]}$ for the group $\widetilde{\Gamma}$ with values in

$$\mathcal{H}\big(L^2(X, w(x)\mathrm{vol}_X), L^2(X, w(x)^{-1}\mathrm{vol}_X)\big),$$

where $w(x)$ is any positive function on X such that $w(x) = e^{\kappa x^2}$ on Z_a. Moreover, in the Hahn series expansion of this extension, the coefficients of e_γ with $\gamma < 0$ are finite-rank operators.

Proof. The proof is identical to the standard proof that the meromorphic properties of the resolvent do not change under compactly supported topological or metric perturbations. The only difference is that we apply our Hahn-meromorphic Fredholm theorem. For the sake of completeness, we give the full argument here. By assumption, we can choose $b > a > 0$ such that the operators Δ_0 and P agree on $C_0^\infty(Z_a)$. Suppose $\psi_1, \psi_2, \phi_1, \phi_2$ are smooth functions on X such that $\operatorname{supp}\phi_1 \subset X_b$ and $\operatorname{supp}\psi_1 \subset X_a$ and such that

$$\psi_1 + \psi_2 = 1, \quad \psi_1\phi_1 + \psi_2\phi_2 = 1, \quad \operatorname{dist}(\operatorname{supp} d\phi_i, \operatorname{supp}\psi_i) > 0.$$

Now denote by P_0 the self-adjoint operator obtained from P by imposing Dirichlet boundary conditions at M_b. Since P is an elliptic operator and the boundary conditions are elliptic, P_0 has compact resolvent and therefore $Q_1(\Lambda) := (P_0 - \lambda^2)^{-1}$ is a meromorphic function with values in $\mathcal{B}(L^2(X_b))$ and the residues of its poles are finite-rank operators. Let us denote by $Q_2(\lambda)$ the Hahn-meromorphic extension of $(\Delta_0 - \lambda^2)^{-1}$ that exists by assumption. Then

$$Q(\lambda) := \phi_1 Q_1(\lambda)\psi_1 + \phi_2 Q_2(\lambda)\psi_2$$

is a Hahn-meromorphic family with values in

$$\mathcal{H}\big(L^2(X, w(x)\mathrm{vol}_X), L^2(X, w(x)^{-1}\mathrm{vol}_X)\big)$$

with respect to the group $\widetilde{\Gamma}$ and the coefficients of e_γ with $\gamma < 0$ of its Hahn series are finite-rank operators with range contained in a fixed finite-dimensional subspace. By construction, for $\lambda \in D_r^{[\sigma]}$, $\operatorname{Im}\lambda > 0$,

$$Q(\lambda)(P - \lambda^2) = \mathrm{Id} + K(\lambda)$$

with

$$K(\lambda) = K_1(\lambda) + K_2(\lambda), \quad K_i(\lambda) := \phi_i Q_i(\lambda)(\Delta\psi_i - 2\nabla_{\operatorname{grad}\psi_i}).$$

Since the integral kernels of Q_i are smooth off the diagonal, the operator $K(\lambda)$ is smoothing. Moreover, its integral kernel has compact support in the second variable.

Given the previous remarks, since $Q_1(\lambda)$ is meromorphic and $Q_2(\lambda)$ is Hahn-meromorphic, $K(\lambda)$ is a Hahn-meromorphic family with values in $\mathcal{H}\big(L^2(X, w(x)^{-1}\mathrm{vol}_X)\big)$ for the group $\widetilde{\Gamma}$, and the coefficients of the e_γ in its Hahn series, for $\gamma < 0$, are finite-rank operators with range contained in a fixed finite-dimensional subspace. Furthermore, for $\lambda = \mathrm{i}r$ purely imaginary, one derives $\|K_i(\mathrm{i}r)\| \leq c/r$ for $r > 1$. Therefore, for a sufficiently large r, the operator $\mathrm{Id} + K(\mathrm{i}r)$ is invertible. By the meromorphic Fredholm

theory and Theorem 4.1, $(\mathrm{Id}+K(\lambda))^{-1}$ is a family of operators in $\mathcal{H}\big(L^2(X, w(x)^{-1}\mathrm{vol}_X)\big)$ which is meromorphic away from zero with finite-rank negative Laurent coefficients at its nonzero poles and finite-rank coefficients of e_γ with $\gamma < 0$. It is Hahn-meromorphic at zero for the group $\widetilde{\Gamma}$. Hence we have

$$(\mathrm{Id}+K(\lambda))^{-1}Q(\lambda)(P-\lambda^2) = \mathrm{Id},$$

and $(\mathrm{Id}+K(\lambda))^{-1}Q(\lambda)$ extends the resolvent of P to a Hahn-meromorphic function with the desired properties, as claimed. □

Combining Theorems 7.6 and 7.9, we obtain:

Corollary 7.10. *Let M be a Riemannian manifold that is isometric to $\mathbb{R}^n\setminus B_R$, $n \geq 2$, outside a compact set for some sufficiently large $R > 0$. Let P be a compactly supported perturbation of the Laplace operator in the sense of Theorem 7.9, and let $w(x)$ be as in that theorem. Then the resolvent $\lambda \mapsto (P-\lambda^2)^{-1}$, as a map*

$$\{\mathrm{Im}\,\lambda > 0\} \to \mathcal{H}\big(L^2(X, w(x)\mathrm{vol}_X), L^2(X, w(x)^{-1}\mathrm{vol}_X)\big),$$

has a continuation to a function in λ that is $z\log z$-Hahn-meromorphic for the group $\mathbb{Z}\times\mathbb{Z}$.

When n is odd, from Example 7.5 we conclude that $\Gamma = \mathbb{Z}\times\{0\}$. In this case, Theorem 7.9 and its corollary are well known and follow from the usual meromorphic Fredholm theorem. Similar convergent expansions in the case of two-dimensional potential scattering with suitable decay at infinity were obtained in [Bollé et al. 1988]. For example, in [Bollé et al. 1988, Theorem 3.3], it was shown by more direct methods that the transition operator $T(k)$ in $L^2(\mathbb{R}^2)$ has a convergent expansion in powers of k and $\log k$.

Remark 7.11. Set $Z = [1,\infty)\times N$. Let X be a Riemannian manifold with an end isometric to

$$\big(Z, dx^2 + x^{-2a}g^N\big), \quad a > 0,$$

for some closed Riemannian manifold (N, g^N). The spectral theory of the Laplace operator on X is examined in detail in [Hunsicker et al. 2014]. There the authors show that the spectral decomposition of the Laplace operator on differential forms on Z can also be described with the Weber transform. The same arguments as in Section 7A together with the proof of Theorem 7.9 then implies that the resolvent of the Laplace operator on X is $z\log z$-Hahn-meromorphic, provided that the eigenvalues of the Laplace operator on N lead to suitable ν.

Remark 7.12. Our method may also be applied to noncompactly supported perturbations of the Laplace operator on $\mathbb{R}^n$, such as for example potential perturbations that have a suitable decay rate at infinity. This is in line with the well-known result in the odd-dimensional case that uniform exponential decay of the potential at infinity guarantees the existence of an analytic continuation of the resolvent into a neighborhood of the spectrum.

References

[Bollé et al. 1988] D. Bollé, F. Gesztesy, and C. Danneels, "Threshold scattering in two dimensions", *Ann. Inst. H. Poincaré Phys. Théor.* **48**:2 (1988), 175–204. MR 89k:81184 Zbl 0696.35040

[Brüning and Seeley 1987] J. Brüning and R. Seeley, "The resolvent expansion for second order regular singular operators", *J. Funct. Anal.* **73**:2 (1987), 369–429. MR 88g:35151 Zbl 0625.47040

[Gil et al. 2011] J. B. Gil, T. Krainer, and G. A. Mendoza, "Dynamics on Grassmannians and resolvents of cone operators", *Anal. PDE* **4**:1 (2011), 115–148. MR 2012d:58040 Zbl 1228.58015

[Guillarmou 2005] C. Guillarmou, "Meromorphic properties of the resolvent on asymptotically hyperbolic manifolds", *Duke Math. J.* **129**:1 (2005), 1–37. MR 2006k:58051 Zbl 1099.58011

[Guillarmou and Hassell 2009] C. Guillarmou and A. Hassell, "Resolvent at low energy and Riesz transform for Schrödinger operators on asymptotically conic manifolds, II", *Ann. Inst. Fourier* (*Grenoble*) **59**:4 (2009), 1553–1610. MR 2011d:58073 Zbl 1175.58011

[Guillarmou and Mazzeo 2012] C. Guillarmou and R. Mazzeo, "Resolvent of the Laplacian on geometrically finite hyperbolic manifolds", *Invent. Math.* **187**:1 (2012), 99–144. MR 2874936 Zbl 1252.58015

[Guillopé 1989] L. Guillopé, "Théorie spectrale de quelques variétés à bouts", *Ann. Sci. École Norm. Sup.* (4) **22**:1 (1989), 137–160. MR 90g:58136 Zbl 0682.58049

[Hahn 1907] H. Hahn, "Über die nicht-archimedischen Größensysteme", *Sitzungsber. Akad. Wiss. Wien Math. Naturwiss.* **116** (1907), 601–655. Reprinted as 445–499 in his *Collected works*, vol. I, edited by L. Schmetterer and K. Sigmund, Springer, Vienna, 1995. MR 96j:01046 Zbl 0859.01030

[Hunsicker et al. 2014] E. Hunsicker, N. Roidos, and A. Strohmaier, "Scattering theory of the p-form Laplacian on manifolds with generalized cusps", *J. Spectr. Theory* **4**:1 (2014), 177–209. MR 3181390

[Ikeda and Taniguchi 1978] A. Ikeda and Y. Taniguchi, "Spectra and eigenforms of the Laplacian on S^n and $P^n(\mathbf{C})$", *Osaka J. Math.* **15**:3 (1978), 515–546. MR 80b:53037 Zbl 0392.53033

[Jensen and Kato 1979] A. Jensen and T. Kato, "Spectral properties of Schrödinger operators and time-decay of the wave functions", *Duke Math. J.* **46**:3 (1979), 583–611. MR 81b:35079 Zbl 0448.35080

[Jensen and Nenciu 2001] A. Jensen and G. Nenciu, "A unified approach to resolvent expansions at thresholds", *Rev. Math. Phys.* **13**:6 (2001), 717–754. MR 2002e:81031 Zbl 1029.81067

[Mazzeo and Melrose 1987] R. R. Mazzeo and R. B. Melrose, "Meromorphic extension of the resolvent on complete spaces with asymptotically constant negative curvature", *J. Funct. Anal.* **75**:2 (1987), 260–310. MR 89c:58133 Zbl 0636.58034

[Mazzeo and Vasy 2005] R. Mazzeo and A. Vasy, "Analytic continuation of the resolvent of the Laplacian on symmetric spaces of noncompact type", *J. Funct. Anal.* **228**:2 (2005), 311–368. MR 2006m:58047 Zbl 1082.58029

[Melrose 1993] R. B. Melrose, *The Atiyah–Patodi–Singer index theorem*, Research Notes in Mathematics **4**, A K Peters, Wellesley, MA, 1993. MR 96g:58180 Zbl 0796.58050

[Müller 1987] W. Müller, *Manifolds with cusps of rank one: Spectral theory and L^2-index theorem*, Lecture Notes in Mathematics **1244**, Springer, Berlin, 1987. MR 89g:58196 Zbl 0632.58001

[Müller 2011] J. Müller, "A Hodge-type theorem for manifolds with fibered cusp metrics", *Geom. Funct. Anal.* **21**:2 (2011), 443–482. MR 2012f:58070 Zbl 1223.58004

[Müller and Strohmaier 2010] W. Müller and A. Strohmaier, "Scattering at low energies on manifolds with cylindrical ends and stable systoles", *Geom. Funct. Anal.* **20**:3 (2010), 741–778. MR 2011h:58050 Zbl 1207.58025

[Murata 1982] M. Murata, "Asymptotic expansions in time for solutions of Schrödinger-type equations", *J. Funct. Anal.* **49**:1 (1982), 10–56. MR 85d:35019 Zbl 0499.35019

[Olver and Maximon 2011] F. W. J. Olver and L. C. Maximon, "Bessel functions", Digital Library of Mathematical Functions, 2011, Available at http://dlmf.nist.gov/10.

[Passman 1977] D. S. Passman, *The algebraic structure of group rings*, Wiley-Interscience, New York, 1977. MR 81d:16001 Zbl 0368.16003

[Reed and Simon 1980] M. Reed and B. Simon, *Methods of modern mathematical physics, I: Functional analysis*, 2nd ed., Academic Press, New York, 1980. MR 85e:46002 Zbl 0459.46001

[Ribenboim 1992] P. Ribenboim, "Fields: Algebraically closed and others", *Manuscripta Math.* **75**:2 (1992), 115–150. MR 93f:13014 Zbl 0767.12001

[Shubin 2001] M. A. Shubin, *Pseudodifferential operators and spectral theory*, 2nd ed., Springer, Berlin, 2001. MR 2002d:47073 Zbl 0980.35180

[Strohmaier 2005] A. Strohmaier, "Analytic continuation of resolvent kernels on noncompact symmetric spaces", *Math. Z.* **250**:2 (2005), 411–425. MR 2006g:58060 Zbl 1081.58027

Received 3 Sep 2013. Revised 22 Nov 2013. Accepted 22 Dec 2013.

JÖRN MÜLLER: jmueller@math.hu-berlin.de
Institut für Mathematik, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany

ALEXANDER STROHMAIER: a.strohmaier@lboro.ac.uk
Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire, LE11 3TU, United Kingdom

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 7 No. 3 2014

Prescription du spectre de Steklov dans une classe conforme 529
Pierre Jammes

Semilinear geometric optics with boundary amplification 551
Jean-Francois Coulombel, Olivier Guès and Mark Williams

The 1-harmonic flow with values in a hyperoctant of the N-sphere 627
Lorenzo Giacomelli, Jose M. Mazón and Salvador Moll

Decomposition rank of $\mathcal{Z}$-stable C^*-algebras 673
Aaron Tikuisis and Wilhelm Winter

Scattering for a massless critical nonlinear wave equation in two space dimensions 701
Martin Sack

Large-time blowup for a perturbation of the cubic Szegő equation 717
Haiyan Xu

A geometric tangential approach to sharp regularity for degenerate evolution equations 733
Eduardo V. Teixeira and José Miguel Urbano

The theory of Hahn-meromorphic functions, a holomorphic Fredholm theorem, and its applications 745
Jörn Müller and Alexander Strohmaier