

ANALYSIS & PDE

Volume 3

No. 2

2010



mathematical sciences publishers

Analysis & PDE

pjm.math.berkeley.edu/apde

EDITORS

EDITOR-IN-CHIEF

Maciej Zworski
University of California
Berkeley, USA

BOARD OF EDITORS

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

PRODUCTION

apde@mathscipub.org

Paulo Ney de Souza, Production Manager

Sheila Newbery, Production Editor

Silvio Levy, Senior Production Editor


See inside back cover or pjm.math.berkeley.edu/apde for submission instructions.

The subscription price for 2010 is US \$120/year for the electronic version, and \$180/year for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94720-3840, USA.

Analysis & PDE, at Mathematical Sciences Publishers, Department of Mathematics, 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 is managed by EditFLOW™ from Mathematical Sciences Publishers.

PUBLISHED BY

 **mathematical sciences publishers**
<http://www.mathscipub.org>

A NON-PROFIT CORPORATION

Typeset in L^AT_EX

Copyright ©2010 by Mathematical Sciences Publishers

POLYNOMIALS WITH NO ZEROS ON THE BIDISK

GREG KNESE

We prove a detailed sums of squares formula for two-variable polynomials with no zeros on the bidisk \mathbb{D}^2 , extending previous such formulas by Cole and Wermer and by Geronimo and Woerdeman. Our formula is related to the Christoffel–Darboux formula for orthogonal polynomials on the unit circle, but the extension to two variables involves issues of uniqueness in the formula and the study of ideals of two-variable orthogonal polynomials with respect to a positive Borel measure on the torus which may have infinite mass. We present applications to two-variable Fejér–Riesz factorizations, analytic extension theorems for a class of bordered curves called distinguished varieties, and Pick interpolation on the bidisk.

1. Introduction	109
2. An example	114
3. Sums of squares and uniqueness	114
4. Preliminaries	117
5. General properties of orthogonal polynomials on \mathbb{T}^2	121
6. OC measures	124
7. Bernstein–Szegő measures	130
8. Proof of the main theorem	135
9. Polynomials with unique decompositions	135
10. Application: Fejér–Riesz factorization	138
11. Application: distinguished varieties	143
12. Application: Agler’s Pick interpolation theorem	146
13. Questions	147
Notational index and conventions	148
Acknowledgements	148
References	148

1. Introduction

Let $q \in \mathbb{C}[z, w]$ be a polynomial of degree (n, m) (degree n in z and degree m in w). Suppose q has no zeros on the unit bidisk $\mathbb{D}^2 := \mathbb{D} \times \mathbb{D} \subset \mathbb{C}^2$. Then, q satisfies the following “sums of (Hermitian) squares” formula: there exist polynomials $A_j \in \mathbb{C}[z, w]$, for $j = 1, \dots, n$, and $B_k \in \mathbb{C}[z, w]$, for $k = 1, \dots, m$ such that

$$|q(z, w)|^2 - |\tilde{q}(z, w)|^2 = (1 - |z|^2) \sum_{j=1}^n |A_j(z, w)|^2 + (1 - |w|^2) \sum_{k=1}^m |B_k(z, w)|^2 \quad (1-1)$$

MSC2000: primary 42C05; secondary 47A57, 46C07, 42B05, 14M12.

Keywords: bidisk, Christoffel–Darboux, sums of squares, Fejér–Riesz, orthogonal polynomials, distinguished varieties, Pick interpolation, Andô’s inequality, Bernstein–Szegő measures, torus, stable polynomials.

where \tilde{q} is the “reflection” of q :

$$\tilde{q}(z, w) = z^n w^m \overline{q\left(\frac{1}{\bar{z}}, \frac{1}{\bar{w}}\right)}.$$

This was first proved in [Cole and Wermer 1999]. Here is an example.

Example 1.1. The polynomial $q(z, w) = 2 - z - w$ has degree $(1, 1)$ and no zeros on \mathbb{D}^2 . The reflection of q is $\tilde{q}(z, w) = 2zw - w - z$. The sum of squares decomposition for q is rather simple:

$$|2 - z - w|^2 - |2zw - w - z|^2 = (1 - |z|^2)2|1 - w|^2 + (1 - |w|^2)2|1 - z|^2.$$

There are several reasons why we deem the Cole–Wermer formula interesting. First, it can be used to give direct proofs of Andô’s inequality from operator theory (in [Cole and Wermer 1999]) and Agler’s Pick interpolation theorem for the bidisk (see Section 12 for this simple derivation). Second, (1-1) can be thought of as a two-variable version of the Christoffel–Darboux formula for orthogonal polynomials on the unit circle. The Christoffel–Darboux formula is fundamental in the theory of orthogonal polynomials on the unit circle [Simon 2005; 2008]. Third, the most obvious analogue of (1-1) in three or more variables is false as it would imply a three operator version of Andô’s inequality (something known to be false). Fourth, (1-1) can be used to prove a determinantal representation for a class of algebraic curves in \mathbb{C}^2 called distinguished varieties (as in [Knese 2009]).

One drawback to the Cole–Wermer formula is that the sums of squares decomposition is not unique.

Example 1.2.

$$\begin{aligned} & |3 - z - w|^2 - |3zw - z - w|^2 \\ &= (1 - |z|^2)3 \left| \frac{1 - \sqrt{5}}{2} + \frac{1 + \sqrt{5}}{2} w \right|^2 + (1 - |w|^2)3 \left| \frac{1 + \sqrt{5}}{2} + \frac{1 - \sqrt{5}}{2} z \right|^2 \\ &= (1 - |z|^2)3 \left| \frac{1 + \sqrt{5}}{2} + \frac{1 - \sqrt{5}}{2} w \right|^2 + (1 - |w|^2)3 \left| \frac{1 - \sqrt{5}}{2} + \frac{1 + \sqrt{5}}{2} z \right|^2 \end{aligned}$$

(Example 2.1 below is more interesting.) It turns out that we can make the Cole–Wermer sums of squares decomposition unique if we require more.

Here is an abridged version of our main theorem. We will fill in more details in Theorem 8.1. All new terminology in the theorem is explained immediately following its statement.

Theorem 1.3. *Let $q \in \mathbb{C}[z, w]$ be almost stable and $\deg q \leq (n, m)$. Then, there exist vector polynomials $\mathbf{E} \in \mathbb{C}^n[z, w]$ and $\mathbf{F} \in \mathbb{C}^m[z, w]$, $\deg \mathbf{E} \leq (n-1, m)$, $\deg \mathbf{F} \leq (n, m-1)$ such that*

- (1) \mathbf{E} is horizontally \mathbb{D} -stable;
- (2) \mathbf{F} is vertically \mathbb{E} -stable;
- (3) the following formula holds

$$|q(z, w)|^2 - |\tilde{q}(z, w)|^2 = (1 - |z|^2)|\mathbf{E}(z, w)|^2 + (1 - |w|^2)|\mathbf{F}(z, w)|^2; \quad (1-2)$$

- (4) $\mathbf{E} \in \mathbb{C}^n[z, w]$ and $\mathbf{F} \in \mathbb{C}^m[z, w]$ satisfying items (1) and (3) are unique up to unitary multiplication.

Definition 1.4. A polynomial $p \in \mathbb{C}[z, w]$ is *stable* if p has no zeros on $\overline{\mathbb{D}^2}$. A polynomial $p \in \mathbb{C}[z, w]$ is *almost stable* if p has no zeros on \mathbb{D}^2 and finitely many zeros on \mathbb{T}^2 .

For instance, $p(z, w) = 3 - z - w$ is stable; $q(z, w) = 2 - z - w$ is almost stable.

Notation 1.5. We use \mathbb{T} to denote the unit circle $\partial\mathbb{D}$ and \mathbb{T}^2 is the two-dimensional torus, or just *torus*; $\mathbb{E} := \mathbb{C} \setminus \overline{\mathbb{D}}$ is the *exterior* disk. We use $\mathbb{C}^N[z]$ to denote the set of \mathbb{C}^N -valued polynomials in the variable z ; likewise, we use $\mathbb{C}^N[z, w]$ to denote the set of \mathbb{C}^N -valued polynomials in z and w . We define

$$\Lambda_N(z) := \begin{pmatrix} 1 \\ z \\ \vdots \\ z^{N-1} \end{pmatrix} \in \mathbb{C}^N[z]. \tag{1-3}$$

If $\mathbf{E}(z, w) = \sum_{j=0}^{n-1} \mathbf{E}_j(w)z^j \in \mathbb{C}^N[z, w]$ has degree less than n in z , we will frequently write \mathbf{E} in the matrix form

$$\mathbf{E}(z, w) = (\mathbf{E}_0(w), \mathbf{E}_1(w), \dots, \mathbf{E}_{n-1}(w))\Lambda_n(z) = E(w)\Lambda_n(z),$$

where $E(w) = (\mathbf{E}_0(w), \mathbf{E}_1(w), \dots, \mathbf{E}_{n-1}(w))$ is an $N \times n$ matrix valued polynomial in w .

Similarly, if $\mathbf{F} \in \mathbb{C}^M[z, w]$ has degree less than m in w , we may write

$$\mathbf{F}(z, w) = F(z)\Lambda_m(w),$$

where $F(z)$ is an $M \times m$ matrix polynomial in z .

Definition 1.6. Let $\Omega \subset \mathbb{C}$. Under the conventions above, we say \mathbf{E} is *horizontally Ω -stable* if $E(w)$ has full rank for all $w \in \Omega$; we say \mathbf{F} is *vertically Ω -stable* if $F(z)$ has full rank for all $z \in \Omega$.

Typically, Ω is one of following sets: \mathbb{D} , \mathbb{E} , $\mathbb{D} \cup \mathbb{E}$, or \mathbb{D} unioned with a subset of \mathbb{T} .

Let us explain the terminology. For fixed $w_0 \in \mathbb{D}$, call the set $\{(z, w_0) : z \in \mathbb{C}\}$ a horizontal line over \mathbb{D} . Supposing $N \leq n$, being *horizontally \mathbb{D} -stable* is equivalent to saying the image of

$$\mathbf{E} : \mathbb{C}^2 \rightarrow \mathbb{C}^N$$

when restricted to a horizontal line over \mathbb{D} sits in no linear subspace of dimension less than N . The reason is simple:

$$\mathbf{E}(z, w_0) = E(w_0)\Lambda_n(z),$$

and when $E(w_0)$ has full rank, the span of the right hand side as z varies over \mathbb{C} is \mathbb{C}^N . Being horizontally \mathbb{D} -stable is much stronger than saying \mathbf{E} is nonvanishing on $\mathbb{C} \times \mathbb{D}$. A similar interpretation holds for \mathbf{F} and “vertical” objects.

Notation 1.7. We let $|\cdot|$ denote the standard norm on \mathbb{C}^N (where the N will be understood from context) and therefore if $\mathbf{E} = (e_1, \dots, e_N)^t \in \mathbb{C}^N[z, w]$, then

$$|\mathbf{E}(z, w)|^2 = \sum_{j=1}^N |e_j(z, w)|^2$$

is evaluated pointwise (and does not represent any type of function space norm).

Definition 1.8. The degree of $p \in \mathbb{C}[z, w]$ will always refer to the *bidegree*. So,

$$\deg p = (n, m)$$

means p has degree n in z and m in w , while

$$\deg p \leq (n, m)$$

means p has at most degree n in z and at most m in w . The same notation applies to vector and matrix polynomials component-wise.

Frequent use will be made of the following notion of polynomial *reflection*.

Definition 1.9. If $p \in \mathbb{C}[z, w]$ is a polynomial of degree at most (j, k) we define the *reflection* (at the (j, k) degree) to be

$$\tilde{p}(z, w) := z^j w^k \overline{p(1/\bar{z}, 1/\bar{w})}.$$

Remark 1.10. In the case of a stable polynomial (no zeros on the *closed* bidisk $\overline{\mathbb{D}^2}$), the theorem is deducible from the work of Geronimo and Woerdeman [2004]. It is the goal of the present paper to extend the sums of squares decomposition with uniqueness to all polynomials with no zeros on the *open* bidisk \mathbb{D}^2 . Why are we concerned with such an extension?

First, it allows a direct, unified proof of the Cole–Wermer formula which does not make use of Andô’s inequality, Agler’s Pick interpolation theorem, or any of their close relatives (the original proof of Cole and Wermer relies heavily on these results). Our hope is that the uniqueness aspects could prove helpful in uniqueness issues of Pick interpolation on the bidisk.

Second, it allows us to improve a bounded analytic extension theorem (from [Knese 2009]) for the already alluded to curves called distinguished varieties. See Section 11.

Third, our method of proof may be of interest to some as we study orthogonal polynomials with respect to a positive Borel measure on \mathbb{T}^2 which may have infinite mass. Since such measures will not necessarily have finite moments, methods involving doubly Toeplitz matrices (as in [Geronimo and Woerdeman 2004]) are not directly available to us, and therefore our method of using reproducing kernels of subspaces of polynomials from [Knese 2008] is well adapted to the present situation. Our method of proof also allows us to improve a characterization of two-variable Fejér–Riesz factorizations from [Geronimo and Woerdeman 2004]. See Section 10.

Remark 1.11. The assumption that q is *almost stable* (i.e., has *finitely* many zeros on \mathbb{T}^2) is there to put us into the most interesting case and not to avoid a difficulty. Every polynomial q with no zeros on the bidisk can be factored into $q = q_1 q_2$ where q_1 has at most finitely many zeros on the two-torus and every factor of q_2 has infinitely many zeros on the two-torus. If q has a nontrivial factor of the type q_2 , then it can be factored out of the entire sums of squares formula. These polynomials with no zeros on the bidisk and infinitely many zeros on the two-torus can be studied separately; see [Knese 2009]. These notions will appear several places later on so we give the following definitions of *toral* and *atoral*.

Definition 1.12. A polynomial $p \in \mathbb{C}[z, w]$ is *toral* if every factor of p has infinitely many zeros on \mathbb{T}^2 .

Definition 1.13. A polynomial $p \in \mathbb{C}[z, w]$ is *atoral* if p has finitely many zeros on \mathbb{T}^2 .

These terms were introduced in [Agler et al. 2006] in a more natural way that makes sense for higher dimensions, but these definitions will suffice for our purposes.

Remark 1.14. The requirements on \mathbf{E} and \mathbf{F} in [Theorem 1.3](#) that make the decomposition unique are essential in proving our bounded analytic extension theorem for distinguished varieties. The requirements are also curiously asymmetric. In fact, the entire formula (1-2) can be “reflected”: replace (z, w) with $(1/\bar{z}, 1/\bar{w})$ and multiply through by $-|z^n w^m|^2$. The result will be a new sums of squares formula with \mathbf{E} and \mathbf{F} replaced with

$$\overleftarrow{\mathbf{E}}(z, w) = z^{n-1} w^m \overline{\mathbf{E}(1/\bar{z}, 1/\bar{w})} \quad \text{and} \quad \overleftarrow{\mathbf{F}}(z, w) = z^n w^{m-1} \overline{\mathbf{F}(1/\bar{z}, 1/\bar{w})},$$

respectively. These new choices will have the stability requirements reversed in [Theorem 1.3](#): \mathbf{E} will be horizontally \mathbb{E} -stable and \mathbf{F} will be vertically \mathbb{D} -stable. (Notice that in [Example 2.1](#), below, the two choices for the sums of squares decompositions are *not* simply obtained from one another by performing this reflection.)

These thoughts beg the following question. Which almost stable polynomials have a unique sums of squares decomposition?

Theorem 1.15. *Suppose $q \in \mathbb{C}[z, w]$ is almost stable with $\deg q \leq (n, m)$. The following are equivalent.*

- (1) *There exist unique nonnegative functions Γ_1, Γ_2 which can be written as the sum of the squared moduli of two-variable polynomials such that*

$$|q(z, w)|^2 - |\tilde{q}(z, w)|^2 = (1 - |z|^2)\Gamma_1(z, w) + (1 - |w|^2)\Gamma_2(z, w). \tag{1-4}$$

- (2) *There are no nonzero polynomials $f \in \mathbb{C}[z, w]$ with degree at most $(n-1, m-1)$ such that*

$$\frac{f}{q} \in L^2(\mathbb{T}^2).$$

- (3) *There exist vector polynomials $\mathbf{E} \in \mathbb{C}^n[z, w]$, $\deg \mathbf{E} \leq (n-1, m)$ and $\mathbf{F} \in \mathbb{C}^m[z, w]$, $\deg \mathbf{F} \leq (n, m-1)$ satisfying*

$$|q(z, w)|^2 - |\tilde{q}(z, w)|^2 = (1 - |z|^2)|\mathbf{E}(z, w)|^2 + (1 - |w|^2)|\mathbf{F}(z, w)|^2$$

that are symmetric in the sense that:

$$\mathbf{E}(z, w) = z^{n-1} w^m \overline{\mathbf{E}(1/\bar{z}, 1/\bar{w})} \quad \text{and} \quad \mathbf{F}(z, w) = z^n w^{m-1} \overline{\mathbf{F}(1/\bar{z}, 1/\bar{w})}.$$

Moreover, \mathbf{E} is horizontally $\mathbb{D} \cup \mathbb{E}$ -stable, and \mathbf{F} is vertically $\mathbb{D} \cup \mathbb{E}$ -stable.

The polynomial $q(z, w) = 2 - z - w$ from [Example 1.1](#) has a unique sums of squares decomposition, since the decomposition we gave satisfies (3), after multiplying by a suitable unimodular constant. [Item \(2\)](#) says that the polynomials with a unique decomposition must in some sense have as many zeros as possible on the torus. Because of this, polynomials with no zeros on the closed bidisk never have unique decompositions unless they are one variable polynomials.

Corollary 1.16. *If $q \in \mathbb{C}[z, w]$ is stable, then q has a unique sums of squares decomposition if and only if q is a function of only one variable (i.e., one of q 's partial derivatives vanishes identically).*

It would be interesting to have a parametrization of the polynomials in [Theorem 1.15](#). Both [Theorem 1.15](#) and [Corollary 1.16](#) are proved in [Section 9](#).

2. An example

Example 2.1. Let $f(z, w) = 2 - zw - z^2w$. One decomposition of f is

$$|2 - zw - z^2w|^2 - |2z^3w^2 - z^2w - zw|^2 = (1 - |z|^2)|\mathbf{E}(z, w)|^2 + (1 - |w|^2)|\mathbf{F}(z, w)|^2,$$

where

$$\mathbf{E}(z, w) = \sqrt{2} \begin{pmatrix} 1 - z^2w \\ w - zw^2 \\ zw - z^2w^2 \end{pmatrix} = \sqrt{2} \begin{pmatrix} 1 & 0 & -w \\ w & -w^2 & 0 \\ 0 & w & -w^2 \end{pmatrix} \begin{pmatrix} 1 \\ z \\ z^2 \end{pmatrix},$$

$$\mathbf{F}(z, w) = \sqrt{2} \begin{pmatrix} z - z^3w \\ 1 - zw \end{pmatrix} = \sqrt{2} \begin{pmatrix} z & -z^3 \\ 1 & -z \end{pmatrix} \begin{pmatrix} 1 \\ w \end{pmatrix}.$$

Alternatively, we could choose instead

$$\mathbf{E}(z, w) = \begin{pmatrix} \sqrt{2}(z - z^2w) \\ z - z^2 \\ 2 - zw - z^2w \end{pmatrix} = \begin{pmatrix} 0 & \sqrt{2} & -\sqrt{2}w \\ 0 & 1 & -1 \\ 2 & -w & -w \end{pmatrix} \begin{pmatrix} 1 \\ z \\ z^2 \end{pmatrix},$$

$$\mathbf{F}(z, w) = \begin{pmatrix} z + z^2 - 2z^3w \\ z^2 - z^3 \end{pmatrix} = \begin{pmatrix} z + z^2 & -2z^3 \\ z^2 - z^3 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ w \end{pmatrix}.$$

These two choices are not equivalent up to unitary multiplication (reflecting is no remedy either) as can be checked. The second choices of \mathbf{E} and \mathbf{F} fit the requirements of [Theorem 1.3](#), while the first choices do not.

3. Sums of squares and uniqueness

In this section we present several lemmas on sums of squares decompositions. [Lemma 3.4](#) proves uniqueness in [Theorem 1.3](#), namely, [item \(4\)](#). This section can easily be skipped and referred back to as necessary. It is included here because it does not require the more demanding notation of the rest of the paper.

The following theorem can be found in [\[D'Angelo 1993\]](#).

Theorem 3.1 (polarization for holomorphic functions). *Let Ω be a domain in \mathbb{C}^N and set*

$$\Omega^* = \{\bar{z} = (\bar{z}_1, \dots, \bar{z}_N) : z \in \Omega\}.$$

If $f : \Omega \times \Omega^ \rightarrow \mathbb{C}$ is a holomorphic function with the property that*

$$f(z, \bar{z}) = 0 \quad \text{for all } z \in \Omega$$

then

$$f(z, w) = 0 \quad \text{for all } (z, w) \in \Omega \times \Omega^*.$$

The following lemma holds equally well for multivariable polynomials, and may be well known to some readers. See [\[Cole and Wermer 1999, Appendix\]](#) for a proof.

Lemma 3.2. *Suppose $\Gamma(z)$ is a sum of squares of polynomials and let N be the rank of the matrix of coefficients of Γ . Then, there exists $\mathbf{A} \in \mathbb{C}^N[z]$ so that*

$$\Gamma(z) = |\mathbf{A}(z)|^2$$

and \mathbf{A} is minimal in the sense that

$$|\mathbf{A}(z)|^2 \equiv |\mathbf{B}(z)|^2, \quad \mathbf{B} \in \mathbb{C}^M[z]$$

implies $\mathbf{B}(z) = V\mathbf{A}(z)$ for some isometric $M \times N$ matrix V .

Lemma 3.3. *Let $\mathbf{E} \in \mathbb{C}^n[z, w]$, $\deg \mathbf{E} \leq (n-1, m)$. Suppose \mathbf{E} is horizontally \mathbb{D} -stable. Suppose further that $\mathbf{A} \in \mathbb{C}^N[z, w]$ satisfies*

$$|\mathbf{E}(z, w)|^2 = |\mathbf{A}(z, w)|^2 \quad \text{for } (z, w) \in \mathbb{C} \times \mathbb{T}.$$

Then, $n \leq N$, $\mathbf{A}(z, w)$ has degree at most $n-1$ in z and there exists an $N \times n$ matrix valued rational inner function $\Psi : \mathbb{D} \rightarrow \mathbb{C}^{N \times n}$, holomorphic on \mathbb{D} such that

$$\mathbf{A}(z, w) = \Psi(w)\mathbf{E}(z, w).$$

By $N \times n$ matrix valued inner function we mean that Ψ is isometry valued on the circle (or more appropriately, unitary valued in the case $n = N$).

Proof. We have assumed

$$|\mathbf{E}(z, w)|^2 = |\mathbf{A}(z, w)|^2,$$

for all $z \in \mathbb{C}$ but $w \in \mathbb{T}$. By the polarization theorem for holomorphic functions

$$\langle \mathbf{E}(z, w), \mathbf{E}(Z, w) \rangle = \langle \mathbf{A}(z, w), \mathbf{A}(Z, w) \rangle, \tag{3-1}$$

for all $z, Z \in \mathbb{C}$ and $w \in \mathbb{T}$. The left hand side has degree at most $n-1$ in z and this implies $\mathbf{A}(z, w)$ has degree at most $n-1$ in z as follows. If some component with the largest degree, say $A_1(z, w) = \sum_{j=0}^M a_j(w)z^j$, of $\mathbf{A}(z, w)$ has degree M larger than $n-1$, then

$$A_1(z, w)\overline{A_1(Z, w)} = |a_M(w)|^2 z^M \bar{Z}^M + \text{lower order terms.}$$

We necessarily have $a_M(w) \equiv 0$ on \mathbb{T} , which implies $a_M(w) \equiv 0$ for all $w \in \mathbb{C}$.

Therefore, we may write

$$\mathbf{A}(z, w) = A(w)\Lambda_n(z),$$

where $A(w)$ is an $N \times n$ matrix polynomial. Let us write

$$\mathbf{E}(z, w) = E(w)\Lambda_n(z), \quad E(w) \in \mathbb{C}^{n \times n}[w].$$

Saying \mathbf{E} is horizontally \mathbb{D} -stable means $E(w)$ is invertible for all $w \in \mathbb{D}$.

Rewriting (3-1) in matrix form we have

$$\Lambda_n(Z)^* E(w)^* E(w) \Lambda_n(z) = \Lambda_n(Z)^* A(w)^* A(w) \Lambda_n(z),$$

and since this holds for all $z, Z \in \mathbb{C}$

$$E(w)^* E(w) = A(w)^* A(w) \tag{3-2}$$

for all $w \in \mathbb{T}$ because $\Lambda_n(z)$ spans \mathbb{C}^n as z varies over any n points. Now define

$$\Psi(w) = A(w)E(w)^{-1}$$

for $w \in \mathbb{D}$, a rational matrix polynomial with no poles on the disk (since $E(w)$ is invertible in the disk). Equation (3-2) says that $\Psi(w)$ is isometric for $w \in \mathbb{T}$. In particular, $n \leq N$, any singularities of Ψ on the circle are removable (Ψ is rational and bounded on the circle), and by the maximum principle Ψ is contraction valued in the disk. By definition, $\mathbf{A}(z, w) = \Psi(w)\mathbf{E}(z, w)$ for all $z, w \in \mathbb{C}$. \square

Lemma 3.4 (uniqueness). *Let $\mathbf{E}, \tilde{\mathbf{E}} \in \mathbb{C}^n[z, w]$, $\deg \mathbf{E}, \tilde{\mathbf{E}} \leq (n-1, m)$. Suppose both \mathbf{E} and $\tilde{\mathbf{E}}$ are horizontally \mathbb{D} -stable.*

Suppose further that there are vector polynomials $\mathbf{F}, \tilde{\mathbf{F}} \in \mathbb{C}^m[z, w]$ such that

$$(1 - |z|^2)|\mathbf{E}(z, w)|^2 + (1 - |w|^2)|\mathbf{F}(z, w)|^2 = (1 - |z|^2)|\tilde{\mathbf{E}}(z, w)|^2 + (1 - |w|^2)|\tilde{\mathbf{F}}(z, w)|^2. \quad (3-3)$$

Then, there exists an $n \times n$ unitary U_1 and an $m \times m$ unitary U_2 such that

$$\mathbf{E}(z, w) = U_1 \tilde{\mathbf{E}}(z, w), \quad \mathbf{F}(z, w) = U_2 \tilde{\mathbf{F}}(z, w).$$

Proof. Setting $|w| = 1$ in (3-3) and canceling the factor $(1 - |z|^2)$ we have

$$|\mathbf{E}(z, w)|^2 = |\tilde{\mathbf{E}}(z, w)|^2 \quad \text{for } (z, w) \in \mathbb{C} \times \mathbb{T}.$$

Both \mathbf{E} and $\tilde{\mathbf{E}}$ satisfy the conditions of Lemma 3.3. Therefore, there exist $n \times n$ matrix valued rational inner functions $\Psi_1, \Psi_2 : \mathbb{D} \rightarrow \mathbb{C}^{n \times n}$ such that

$$\tilde{\mathbf{E}}(z, w) = \Psi_1(w)\mathbf{E}(z, w), \quad \mathbf{E}(z, w) = \Psi_2(w)\tilde{\mathbf{E}}(z, w).$$

This implies $\Psi_1(w)\Psi_2(w) = I$, and as Ψ_1, Ψ_2 are contractive valued, we must have Ψ_1 and Ψ_2 constant and equal to unitary matrices. Hence, there exists an $n \times n$ unitary matrix U_1 such that

$$\mathbf{E}(z, w) = U_1 \tilde{\mathbf{E}}(z, w),$$

which implies

$$|\mathbf{E}(z, w)|^2 = |\tilde{\mathbf{E}}(z, w)|^2 \quad \text{for all } (z, w) \in \mathbb{C}^2.$$

In turn, by (3-3) we have

$$|\mathbf{F}(z, w)|^2 = |\tilde{\mathbf{F}}(z, w)|^2 \quad \text{for all } (z, w) \in \mathbb{C}^2.$$

By Lemma 3.2, there exists an $m \times m$ unitary matrix U_2 such that

$$\mathbf{F}(z, w) = U_2 \tilde{\mathbf{F}}(z, w). \quad \square$$

We conclude this section with a lemma about the presence of zeros on the “undistinguished” portion of the boundary of \mathbb{D}^2 , namely $(\mathbb{D} \times \mathbb{T}) \cup (\mathbb{T} \times \mathbb{D})$.

Lemma 3.5. *Suppose $q \in \mathbb{C}[z, w]$ has no zeros on \mathbb{D}^2 . If $q(z_0, w_0) = 0$ for some $(z_0, w_0) \in \mathbb{T} \times \mathbb{D}$, then $q(z_0, w) = 0$ for all $w \in \mathbb{C}$; i.e., $(z - z_0)$ divides q . In particular, there can only be finitely many $z_0 \in \mathbb{T}$ such that $q(z_0, \cdot)$ has a zero in \mathbb{D} .*

Proof. There is no harm in assuming q is irreducible. Suppose $q(z_0, w)$ is not identically zero as a function of w . Then, we can apply the Weierstrass preparation theorem to q and write

$$q(z, w) = u(z, w)(z^k + a_1(w)z^{k-1} + \dots + a_k(w))$$

on some bidisk $D_1 \times D_2$ containing (z_0, w_0) where u is holomorphic and nonvanishing on $D_1 \times D_2$ and each a_j is holomorphic on D_2 . We also assume $D_2 \subset \mathbb{D}$. Furthermore, for $w \in D_2 \setminus \{w_0\}$, each $a_j(w)$ is a symmetric function of the k (necessarily) distinct roots (by irreducibility) $z_1(w), z_2(w), \dots, z_k(w) \in D_1$ of $q(\cdot, w)$ for $w \in D_2 \setminus \{w_0\}$. Note $a_k(w) = (-1)^k z_1(w) \dots z_k(w)$ for $w \neq w_0$ and $a_k(w_0) = (-z_0)^k$. Since q has no zeros in \mathbb{D}^2 , $|z_j(w)| \geq 1$ for all j and $w \in D_2$, and hence $|a_k(w)| \geq 1$ for all $w \in D_2$. Since $|a_k(w_0)| = 1$ the maximum principle implies a_k is a unimodular constant, which in turn implies the roots $z_1(w), \dots, z_k(w)$ are all unimodular valued. The roots must be constant and equal to z_0 ; that is, $q(z, w)$ can be divided by $z - z_0$. \square

4. Preliminaries

As in [Knese 2008], our approach will be to study two-variable orthogonal polynomials with respect to a positive Borel measure μ on the two-torus. The difference is that here we allow measures with infinite mass. In particular, we study ‘‘Bernstein–Szegő’’ measures on \mathbb{T}^2

$$\frac{1}{|q(z, w)|^2} d\sigma,$$

where $d\sigma$ is normalized Lebesgue measure on the torus:

$$d\sigma = d\sigma(z, w) = \frac{dz}{2\pi iz} \frac{dw}{2\pi iw}, \tag{4-1}$$

and $q \in \mathbb{C}[z, w]$ has finitely many zeros on \mathbb{T}^2 (and hence this measure can have infinite mass). On one hand, this causes a number of certain superficial (but still interesting) changes in the theory. For instance, we have to deal with the ideal $\mathbb{C}[z, w] \cap L^2(\mu)$ of polynomials in $L^2(\mu)$ as opposed to all of $\mathbb{C}[z, w]$ when studying orthogonal polynomials. (In particular, studying moment matrices will not be an option, because our measures may not have finite moments.) On the other hand, this change forces us to take greater care in certain situations. For instance, if $q \in \mathbb{C}[z, w]$ has no zeros on the bidisk and finitely many zeros on the two-torus, we *cannot* say (as we would in the case with no zeros on \mathbb{T}^2) that

$$\int_{\mathbb{T}^2} \frac{1}{q(z, w)} d\sigma(z, w) = \frac{1}{q(0, 0)}$$

since $1/q$ will not be integrable. Perhaps this integral could be understood in a principal value sense, however we confront this issue in our own way in Proposition 7.1.

Let us begin to provide some details. We shall make the following standing assumptions:

- μ is a positive Borel measure on \mathbb{T}^2 ;
- the ideal

$$\mathcal{F}_\mu := L^2(\mu) \cap \mathbb{C}[z, w] \tag{4-2}$$

is nontrivial, where elements of $\mathbb{C}[z, w]$ here are thought of as measurable functions on \mathbb{T}^2 ;

- the support of μ is not contained in the zero set of a nonzero polynomial, thus ensuring that $\|q\|_{L^2(\mu)} \neq 0$ if $q \neq 0$.

The inner product on $L^2(\mu)$ will be denoted by

$$\langle f, g \rangle_\mu = \int_{\mathbb{T}^2} f \bar{g} d\mu. \tag{4-3}$$

We shall make use of the machinery of reproducing kernel Hilbert spaces.

Notation 4.1. Given a finite-dimensional subspace $V \subset L^2(\mu) \cap \mathbb{C}[z, w]$, we shall use KV to denote the reproducing kernel of V . Namely, for each $(Z, W) \in \mathbb{C}^2$, $KV_{(Z, W)}$ is the unique element of V satisfying

$$f(Z, W) = \langle f, KV_{(Z, W)} \rangle_\mu \quad \text{for all } f \in V$$

and we define $KV : \mathbb{C}^2 \times \mathbb{C}^2 \rightarrow \mathbb{C}$ by

$$KV((z, w), (Z, W)) := KV_{(Z, W)}(z, w).$$

It is not hard to show KV is conjugate symmetric:

$$KV((z, w), (Z, W)) = \overline{KV((Z, W), (z, w))},$$

and if $\{e_1, \dots, e_N\}$ is an orthonormal basis of V , then

$$KV((z, w), (Z, W)) = \sum_{j=1}^N e_j(z, w) \overline{e_j(Z, W)}.$$

Given $q \in \mathbb{C}[z, w]$ we use

$$\hat{q}(j, k) \tag{4-4}$$

to denote the coefficient of $z^j w^k$ in the Fourier series of q .

Remark 4.2. Throughout, we fix positive integers n and m . The notations below depend on this.

We use the following notations as in [Knese 2008] which define subspaces of polynomials based on what frequencies may appear in their Fourier series (or in other language, we define subspaces based on the *carrier* of the polynomials). The symbols should be thought of a lying in the grid \mathbb{Z}^2 with the lower left corners representing the origin. A blackened section denotes excluded Fourier support. The box with the lower left corner missing \square denotes the polynomials of degree at most (n, m) which vanish at $(0, 0)$, while the box with the upper right corner missing \blacksquare denotes the polynomials of degree at most (n, m) with no (n, m) Fourier coefficient.

Notation 4.3.

- $\square := \{q \in \mathbb{C}[z, w] : \deg(q) \leq (n, m)\}$
- $\blacksquare := \{q \in \mathbb{C}[z, w] : \deg(q) \leq (n-1, m)\}$
- $\blacksquare := \{q \in \mathbb{C}[z, w] : \deg(q) \leq (n, m-1)\}$
- $\blacksquare := \{q \in \mathbb{C}[z, w] : \deg(q) \leq (n-1, m-1)\}$
- $\square := \{q \in \square : q(0, 0) = 0\}$
- $\square := \{q \in \square : \hat{q}(n, m) = 0\}$

For any of the above subspaces (and similar variations) we shall use a subscript μ to denote the intersection with $L^2(\mu)$. Namely,

$$\begin{aligned}\square_\mu &:= \square \cap L^2(\mu) \\ \blacksquare_\mu &:= \blacksquare \cap L^2(\mu) \\ \blacksquare_\mu &:= \blacksquare \cap L^2(\mu), \dots\end{aligned}$$

We use the following notations for shifts and certain orthogonal complements using the inner product on $L^2(\mu)$.

Notation 4.4.

$$\begin{aligned}w\square_\mu &:= \{wp : p \in \square_\mu\} & z\square_\mu &:= \{zp : p \in \square_\mu\} \\ \boxplus_\mu &:= \square_\mu \ominus \square_\mu & \boxminus_\mu &:= \square_\mu \ominus (w\square_\mu) \\ \boxtimes_\mu &:= \square_\mu \ominus \square_\mu & \boxtimes_\mu &:= \square_\mu \ominus (z\square_\mu) \\ \boxdot_\mu &:= \square_\mu \ominus \square_\mu & \boxdot_\mu &:= \square_\mu \ominus (w\square_\mu) \\ \boxplus_\mu &:= \square_\mu \ominus \square_\mu & \boxplus_\mu &:= \square_\mu \ominus \square_\mu\end{aligned}$$

For instance, \boxplus_μ denotes all $p \in \mathbb{C}[z, w] \cap L^2(\mu)$ of degree at most $(n-1, m)$ which are orthogonal to the polynomials in $\mathbb{C}[z, w] \cap L^2(\mu)$ of degree at most $(n-1, m-1)$.

A discussion of the notation. A more traditional notation for the subspaces above might work as follows:

$$\begin{aligned}\mathcal{P}_{n,m} &:= \{q \in \mathbb{C}[z, w] : \deg(q) \leq (n, m)\} \\ \mathcal{P}_{n,m-1} &:= \{q \in \mathbb{C}[z, w] : \deg(q) \leq (n, m-1)\} \\ \mathcal{P}_{n-1,m-1} &:= \{q \in \mathbb{C}[z, w] : \deg(q) \leq (n-1, m-1)\} \\ \mathcal{P}_{n,m}^{(n,m)} &:= \{q \in \mathbb{C}[z, w] : \deg(q) \leq (n, m), \hat{q}(n, m) = 0\}\end{aligned}$$

In which case one could write out orthogonal complements in detail as in:

$$\mathcal{P}_{n,m} \ominus \mathcal{P}_{n,m-1}.$$

To illustrate how cumbersome this becomes let us compare this more traditional notation with the box notation above. In the rest of this paper it will be important to decompose $(\mathcal{P}_{n,m})_\mu$ (or \square_μ) in a variety of ways. With more traditional notation we have:

$$\begin{aligned}\mathcal{P}_{n,m} &= (\mathcal{P}_{n,m} \ominus \mathcal{P}_{n,m}^{(n,m)}) \oplus \mathcal{P}_{n,m}^{(n,m)} \\ &= (\mathcal{P}_{n,m} \ominus \mathcal{P}_{n,m}^{(n,m)}) \oplus (\mathcal{P}_{n,m}^{(n,m)} \ominus \mathcal{P}_{n,m-1}) \oplus \mathcal{P}_{n,m-1}.\end{aligned}$$

All orthogonal sums and complements are taken with respect to $L^2(\mu)$. With our notation we have:

$$\begin{aligned}\square_\mu &= \underbrace{(\square_\mu \ominus \square_\mu)}_{\boxplus_\mu} \oplus \underbrace{\square_\mu}_{\boxtimes_\mu} \\ &= \boxplus_\mu \oplus \underbrace{(\square_\mu \ominus \square_\mu)}_{\boxtimes_\mu} \oplus \square_\mu \\ &= \boxplus_\mu \oplus \boxtimes_\mu \oplus \square_\mu\end{aligned}$$

It becomes necessary to take this even further:

$$\begin{aligned} \square_\mu &= \boxplus_\mu \oplus \boxminus_\mu \oplus \underbrace{\square_\mu}_{(\boxplus_\mu \ominus \boxminus_\mu) \oplus \boxplus_\mu} \\ &= \boxplus_\mu \oplus \boxminus_\mu \oplus \underbrace{(\boxplus_\mu \ominus \boxminus_\mu)}_{\boxplus_\mu} \oplus \boxplus_\mu \\ &= \boxplus_\mu \oplus \boxminus_\mu \oplus \boxplus_\mu \oplus \boxplus_\mu \end{aligned}$$

Another way of decomposing \square_μ is as

$$\square_\mu = \boxplus_\mu \oplus \boxminus_\mu \oplus \boxplus_\mu \oplus \boxplus_\mu$$

All of these decompositions translate into formulas for reproducing kernels since the reproducing kernel of a direct sum is the sum of the reproducing kernels [Knese 2008, Section 3]. Therefore,

$$K\square_\mu = K\boxplus_\mu + K\boxminus_\mu + K\boxplus_\mu + K\boxplus_\mu, \tag{4-5}$$

$$K\square_\mu = K\boxplus_\mu + K\boxminus_\mu + K\boxplus_\mu + K\boxplus_\mu. \tag{4-6}$$

□

The two subspaces $\boxplus_\mu, \boxminus_\mu$ are either one-dimensional or trivial and are important enough to warrant special names:

$$\text{Max}_\mu := \boxplus_\mu = \square_\mu \ominus \square_\mu = \{p \in \mathbb{C}[z, w] \cap L^2(\mu) : p \in \square_\mu, p \perp \square_\mu\}, \tag{4-7}$$

$$\text{Min}_\mu := \boxminus_\mu = \square_\mu \ominus \square_\mu = \{p \in \mathbb{C}[z, w] \cap L^2(\mu) : p \in \square_\mu, p \perp \square_\mu\}. \tag{4-8}$$

We choose these names because $p \in \text{Max}_\mu$ maximizes the quantity

$$\frac{|f(0, 0)|}{\|f\|_{L^2_\mu}}$$

among all $f \in \mathbb{C}[z, w] \cap L^2(\mu)$ of degree at most (n, m) . This follows from the fact that $p \in \text{Max}_\mu$ if and only if p is orthogonal to all $f \in \square_\mu$ vanishing at $(0, 0)$. Elements of Min_μ maximize the value of

$$\frac{|\hat{f}(n, m)|}{\|f\|_{L^2_\mu}}$$

among $f \in \square_\mu$.

We continue [Example 1.1](#) to make all these definitions concrete.

Example 4.5. Let $q(z, w) = 2 - z - w$. Let

$$d\mu = \frac{1}{|2 - z - w|^2} d\sigma(z, w) = \frac{1}{(2\pi i)^2 |2 - z - w|^2} \frac{dz}{z} \frac{dw}{w}.$$

It turns out that $\mathcal{F}_\mu = L^2(\mu) \cap \mathbb{C}[z, w]$ equals the maximal ideal $(z - 1, w - 1) \subset \mathbb{C}[z, w]$. Indeed, a double application of Cauchy’s formula shows

$$1 \notin L^2(\mu) \text{ and } z - 1, w - 1 \in L^2(\mu).$$

Also,

$$2 - z - w \perp w(z - 1), z(w - 1), \quad 2zw - z - w \perp (z - 1), (w - 1).$$

If we set $n = 1$ and $m = 1$, then

$$\begin{aligned} \square_\mu &= \{0\}, & \text{since } 1 \notin L^2(\mu), \\ \square_\mu &= (w - 1)\mathbb{C}, \\ \square_\mu &= (z - 1)\mathbb{C}, \\ \square_\mu &= \text{span}\{z - 1, w - 1, z + w - 2zw\}, \\ \boxplus_\mu &= (2 - z - w)\mathbb{C}, \\ \boxplus_\mu &= (2zw - z - w)\mathbb{C}. \end{aligned}$$

Since \square_μ is trivial,

$$\boxplus_\mu = \square_\mu \ominus \square_\mu = \square_\mu.$$

In general, $\boxplus_\mu \neq \square_\mu$, but the singularity of μ forces certain subspaces to degenerate.

5. General properties of orthogonal polynomials on \mathbb{T}^2

This section is about orthogonal polynomials on \mathbb{T}^2 with respect to a (not necessarily finite) positive Borel measure on \mathbb{T}^2 . We use reproducing kernels to study entire subspaces of polynomials all at once, so the ‘‘orthogonal polynomials’’ are in some sense disguised.

The heart of the following two propositions should be familiar to those who know something about orthogonal polynomials on the unit circle. Namely, if ρ is a probability measure on \mathbb{T} , and if $q \in \mathbb{C}[z]$, $\deg q \leq n$, then in $L^2(\rho)$

$$q \perp z, z^2, \dots, z^n \implies q \text{ is stable.}$$

In two variables, consider the *subspace* of polynomials $\boxplus_\mu = \square_\mu \ominus \square_\mu$; that is, all

$$p \in \mathbb{C}[z, w] \cap L^2(\mu), \deg p \leq (n, m - 1)$$

satisfying

$$p \perp \text{span}\{z^j w^k : 1 \leq j \leq n, 0 \leq k \leq m - 1\} \cap L^2(\mu).$$

The conclusion of the first proposition below is that $p(z, w)$ has no factors of the form $(z - z_0)$ with $z_0 \in \mathbb{D}$, and the second proposition says that a vector consisting of an orthonormal basis for \boxplus_μ is vertically \mathbb{D} -stable. Both of these notions are generalizations of one variable stability.

Another way to generalize orthogonal polynomials from one to two variables is to consider $p \in \mathbb{C}[z, w] \cap L^2(\mu)$, $\deg p \leq (n, m)$ satisfying

$$p \perp \text{span}\{z^j w^k : 0 \leq j \leq n, 0 \leq k \leq m, (j, k) \neq (0, 0)\} \cap L^2(\mu),$$

namely, $p \in \text{Max}_\mu = \boxplus_\mu$. This situation is much more subtle and is the topic of [Section 6](#).

Definition 5.1. We say an element p of $\mathbb{C}[z, w]$ is a *divisor of the ideal* \mathcal{F}_μ if whenever $pq \in \mathcal{F}_\mu$, then $q \in \mathcal{F}_\mu$.

Polynomials with no zeros on \mathbb{T}^2 are always divisors of \mathcal{F}_μ .

Proposition 5.2.

- (1) (a) If p is a nonzero element of \mathbb{A}_μ or \mathbb{B}_μ , then p is not divisible by a polynomial of the form $L(z, w) = z - z_0$ for $z_0 \in \mathbb{D}$.
- (b) If p is a nonzero element of \mathbb{A}_μ or \mathbb{B}_μ then p is not divisible by any $L(z, w) = z - z_0$ when $z_0 \in \mathbb{C} \setminus \overline{\mathbb{D}}$.
- (c) In addition, if $z_0 \in \mathbb{T}$, and $L(z, w) = z - z_0$ happens to be a divisor in \mathcal{F}_μ , then nonzero elements of \mathbb{A}_μ , \mathbb{B}_μ , \mathbb{C}_μ , \mathbb{D}_μ cannot have L as a factor.
- (2) (a) If p is a nonzero element of \mathbb{E}_μ or \mathbb{F}_μ , then p cannot have a factor of the form $J(z, w) = w - w_0$ when $w_0 \in \mathbb{D}$.
- (b) If p is a nonzero element of \mathbb{E}_μ or \mathbb{F}_μ , then p cannot have a factor of the form $J(z, w) = w - w_0$ when $w_0 \in \mathbb{C} \setminus \overline{\mathbb{D}}$.
- (c) In addition, if $w_0 \in \mathbb{T}$, and $J(z, w) = w - w_0$ happens to be a divisor in \mathcal{F}_μ , then nonzero elements of \mathbb{E}_μ , \mathbb{F}_μ , \mathbb{G}_μ , \mathbb{H}_μ cannot have J as a factor.

Proof. We prove item (1a). Let $p \in \mathbb{A}_\mu$ and suppose $p = gL$ for some $g \in \mathbb{A}$ where $L(z, w) = z - z_0$ with $|z_0| < 1$. Since L has no zeros on \mathbb{T}^2 , $g = p/L \in L^2(\mu)$. Then, $z_0g(z, w) = zg(z, w) - p(z, w)$ and

$$|z_0|^2 \|g\|_{L^2(\mu)}^2 = \|-p + zg\|_{L^2(\mu)}^2 = \|p\|_{L^2(\mu)}^2 + \|zg\|_{L^2(\mu)}^2 = \|p\|_{L^2(\mu)}^2 + \|g\|_{L^2(\mu)}^2.$$

since $p \perp_\mu zg$. Rearranging we arrive at

$$\|p\|_{L^2(\mu)}^2 = (|z_0|^2 - 1)\|g\|_{L^2(\mu)}^2 < 0,$$

a contradiction. The proofs of the other statements are variations on the above idea. \square

Curiously, slightly more complicated factors can be ruled out by a similar argument. For instance, if $|a| < 1$, then $P(z, w) = z^2 - aw^3$ cannot be a factor of any polynomial in \mathbb{A}_μ . If $|a| = 1$ and P is a divisor of \mathcal{F}_μ then the same conclusion holds.

The next proposition shows that *horizontal \mathbb{D} -stability* occurs naturally (recall [Definition 1.6](#)).

Proposition 5.3. *Let $\{e_1, \dots, e_N\} \subset \mathbb{C}[z, w]$ be an orthonormal basis for \mathbb{E}_μ which we write vectorially as $\mathbf{E}(z, w) = (e_1(z, w), \dots, e_N(z, w))^t$. Then, $N \leq n$ and \mathbf{E} is horizontally $\mathbb{D} \cup X$ -stable, where $X \subset \mathbb{T}$ is the set of $w_0 \in \mathbb{T}$ such that $L(z, w) = w - w_0$ is a divisor of \mathcal{F}_μ .*

The same results hold for \mathbb{B}_μ with the roles of z and w switched.

Proof. First, we claim $\dim \mathbb{E}_\mu := N \leq n$. Given $n + 1$ polynomials in \mathbb{E}_μ , some linear combination of them will be a multiple of w (since the degree in z is at most $n - 1$); such a combination would be orthogonal to itself (by definition of \mathbb{E}_μ) and therefore zero; and hence any $n + 1$ polynomials in \mathbb{E}_μ are dependent. So, $\dim \mathbb{E}_\mu \leq n$.

Write

$$\mathbf{E}(z, w) = E(w)\Lambda_n(z),$$

where $E(w)$ is an $(N \times n)$ -matrix valued polynomial in w of degree at most m . We must prove \mathbf{E} is horizontally $\mathbb{D} \cup X$ -stable which means $E(w)$ has rank N for all $w \in \mathbb{D} \cup X$.

So, suppose $E(w_0)$ has rank less than N at some point $w_0 \in \mathbb{C}$. Since $E(w_0)$ is $N \times n$ and $N \leq n$ there must be a nonzero vector $\mathbf{v} \in \mathbb{C}^N$ such that $\mathbf{v}^t E(w_0) = \mathbf{0}^t$; that is, the following (necessarily nonzero) polynomial

$$q(z, w) = \mathbf{v}^t E(w) \Lambda_n(z) = \mathbf{v}^t \mathbf{E}(z, w)$$

is in \square_μ and vanishes on the set $\{w = w_0\}$. By the previous proposition this can only happen if $w_0 \notin \mathbb{D}$ and when $w_0 \in \mathbb{T}$, $w - w_0$ cannot be a divisor of \mathcal{F}_μ . So, $E(w_0)$ has full rank N everywhere in \mathbb{D} and at all points $w_0 \in \mathbb{T}$ for which $w - w_0$ is a divisor of \mathcal{F}_μ ; that is, \mathbf{E} is horizontally $\mathbb{D} \cup X$ -stable. \square

Continuing our previous aside, we can also say that $\mathbf{E} \in \mathbb{C}^N[z, w]$ as above when restricted to the variety $\{z^2 - aw^3 = 0\}$ (here $|a| < 1$) does not sit inside any proper subspace of \mathbb{C}^N .

Remark 5.4. The main ideas of the previous two propositions appeared in the appendix of [Knese 2009] in a less detailed form.

The following is an analogue of the one variable Christoffel–Darboux formula.

Proposition 5.5 (Christoffel–Darboux type formulas). *Suppressing $((z, w), (z, w))$ in front of each kernel we have*

$$K_{\square_\mu} - K_{\square_\mu} = (1 - |z|^2)K_{\square_\mu} \quad \text{and} \quad K_{\square_\mu} - K_{\square_\mu} = (1 - |w|^2)K_{\square_\mu}.$$

Proof. Let us decompose \square_μ , the subspace of polynomials $p \in \mathbb{C}[z, w] \cap L^2(\mu)$, $\deg p \leq (n, m-1)$, in two ways:

$$\begin{aligned} \square_\mu &= (\square_\mu \ominus \square_\mu) \oplus \square_\mu = \square_\mu \oplus z\square_\mu, \\ \square_\mu &= (\square_\mu \ominus \square_\mu) \oplus \square_\mu = \square_\mu \oplus \square_\mu. \end{aligned}$$

The reproducing kernel of a direct sum is the sum of the reproducing kernels [Knese 2008, Section 3], and so

$$\begin{aligned} K_{\square_\mu} + \underbrace{K(z\square_\mu)} &= K_{\square_\mu} + K_{\square_\mu}, \\ K_{\square_\mu} + z\bar{Z}K_{\square_\mu} &= K_{\square_\mu} + K_{\square_\mu}, \end{aligned}$$

since shifting a subspace by z “shifts” the reproducing kernel by the factor $z\bar{Z}$. Here we have suppressed the argument $((z, w), (Z, W))$ in front of every reproducing kernel. After rearranging we get the first equation of the proposition:

$$K_{\square_\mu} - K_{\square_\mu} = K_{\square_\mu} - z\bar{Z}K_{\square_\mu} = (1 - z\bar{Z})K_{\square_\mu}.$$

The proof of the second equation is similar. \square

Definition 5.6. A polynomial $p \in \mathbb{C}[z, w]$ is \mathbb{T}^2 -symmetric if it equals a unimodular constant μ times its reflection:

$$p(z, w) = \mu \tilde{p}(z, w) = \mu z^j w^k \overline{p(1/\bar{z}, 1/\bar{w})};$$

here p has degree exactly (j, k) .

Proposition 5.7. *Let P be the greatest common divisor of \square_μ . Then, every factor of P is \mathbb{T}^2 -symmetric and the zero set of every factor of P intersects \mathbb{T}^2 .*

Proof. The greatest common divisor P is necessarily \mathbb{T}^2 -symmetric (basically since the set \square_μ is). Let q be an irreducible factor of P and let j be the highest power such that q^j divides P . Suppose q is not a multiple of \bar{q} . Then $q^j \bar{q}^j$ divides P . Let p be an element of \square_μ divisible by the maximal number of factors of q ; that is, q^k divides p and no nonzero element of \square_μ is divisible by q^{k+1} . Since \bar{q}^j divides p we may write $p = q^k \bar{q}^j g$ for some $g \in \mathbb{C}[z, w]$. Since $|q| = |\bar{q}|$ on \mathbb{T}^2 , it follows that p being in $L^2(\mu)$ implies $q^{k+j} g \in L^2(\mu)$. In particular, $q^{k+j} g \in \square_\mu$ contradicting the maximality property of p and k . Hence, q must be \mathbb{T}^2 -symmetric.

The zero set of every factor q of P must intersect \mathbb{T}^2 since otherwise $qg \in L^2(\mu)$ implies $g \in L^2(\mu)$ for any $g \in \mathbb{C}[z, w]$. \square

Question 5.8. Is P toral? That is, does the zero set of every factor of P intersect \mathbb{T}^2 on an infinite set?

This question is made more difficult by the fact that there exist irreducible, atoral, \mathbb{T}^2 -symmetric polynomials:

$$p(z, w) = (3z + 1)w^2 - (z + 3)(3z + 1)w + z(z + 3)$$

is such a polynomial taken from [Agler et al. 2008].

6. OC measures

The following theorem should be thought of as an attempt to prove a two-variable Christoffel–Darboux formula for general positive Borel measures which fails. The expression ϵ below measures how much it fails.

Theorem 6.1. *Let μ be a positive Borel measure on \mathbb{T}^2 for which $\mathbb{C}[z, w] \cap L^2(\mu) \neq \{0\}$ and for which $\text{Max}_\mu = \square_\mu$ is one-dimensional. Let*

$$\epsilon := (K^{\square_\mu} - K^{\bar{\square}_\mu}) - (K^{\square_\mu} - K^{\bar{\square}_\mu}).$$

If q is any unit norm polynomial in Max_μ , then writing

$$q\bar{q} = q(z, w)\overline{q(Z, W)}$$

and omitting the expressions $((z, w), (Z, W))$, we get:

$$\begin{aligned} q\bar{q} - \bar{q}q &= (1 - z\bar{Z})(1 - w\bar{W})K^{\square_\mu} \\ &\quad + (1 - z\bar{Z})K^{\bar{\square}_\mu} + (1 - w\bar{W})K^{\bar{\square}_\mu} + \epsilon \\ &= (1 - z\bar{Z})K^{\bar{\square}_\mu} + (1 - w\bar{W})K^{\bar{\square}_\mu} + \epsilon \\ &= (1 - z\bar{Z})K^{\bar{\square}_\mu} + (1 - w\bar{W})K^{\bar{\square}_\mu} + \epsilon. \end{aligned}$$

The proof of this theorem is identical to the proof of Theorem 4.5 in [Knese 2008], which is for probability measures. We already have many of the details in place so it seems worthwhile to include the proof.

Proof. By Equation (4-5),

$$\begin{aligned} K \square_{\mu} &= K \boxplus_{\mu} + K \boxminus_{\mu} + K \boxtimes_{\mu} + K \square_{\mu} \\ &= K \boxplus_{\mu} + K \boxminus_{\mu} + K \boxtimes_{\mu} + K \square_{\mu} + (K \boxminus_{\mu} - K \boxplus_{\mu}), \end{aligned} \quad (6-1)$$

and, by Equation (4-6),

$$\begin{aligned} K \square_{\mu} &= K \boxplus_{\mu} + K \boxminus_{\mu} + K \boxtimes_{\mu} + K \square_{\mu} \\ &= K \boxplus_{\mu} + K \boxminus_{\mu} + K \boxtimes_{\mu} + K \square_{\mu} + (K \boxplus_{\mu} - K \boxminus_{\mu}) \\ &= K \boxplus_{\mu} + z \bar{Z} K \boxminus_{\mu} + w \bar{W} K \boxtimes_{\mu} + z \bar{Z} w \bar{W} K \square_{\mu} + (K \boxplus_{\mu} - K \boxminus_{\mu}). \end{aligned}$$

Using the formulas in Proposition 5.5 to eliminate $K \boxplus_{\mu}$ and $K \boxminus_{\mu}$, we get:

$$\begin{aligned} K \square_{\mu} &= K \boxplus_{\mu} + z \bar{Z} (K \boxminus_{\mu} + (1 - w \bar{W}) K \square_{\mu}) + w \bar{W} (K \boxtimes_{\mu} + (1 - z \bar{Z}) K \square_{\mu}) \\ &\quad + z \bar{Z} w \bar{W} K \square_{\mu} + (K \boxplus_{\mu} - K \boxminus_{\mu}) \\ &= K \boxplus_{\mu} + z \bar{Z} K \boxminus_{\mu} + w \bar{W} K \boxtimes_{\mu} \\ &\quad + (z \bar{Z} + w \bar{W} - z \bar{Z} w \bar{W}) K \square_{\mu} + (K \boxplus_{\mu} - K \boxminus_{\mu}). \end{aligned}$$

Combined with Equation (6-1) above we have

$$K \boxplus_{\mu} - K \boxminus_{\mu} = (1 - z \bar{Z}) K \boxminus_{\mu} + (1 - w \bar{W}) K \boxtimes_{\mu} + (1 - z \bar{Z})(1 - w \bar{W}) K \square_{\mu} + \epsilon.$$

Note that since $\text{Max}_{\mu} = \boxplus_{\mu}$ is one-dimensional, $q\bar{q}$ is its reproducing kernel. Likewise, $\text{Min}_{\mu} = \boxminus_{\mu}$ is the reflection of Max_{μ} and therefore has reproducing kernel $\overleftarrow{q\bar{q}}$. This proves the first formula of the theorem.

The remaining formulas follow from Proposition 5.5 by eliminating either $K \boxplus_{\mu}$ or $K \boxminus_{\mu}$. See [Knese 2008] for more details. \square

The ϵ in Theorem 6.1 is identically zero for measures of the following type, as we explain below.

Definition 6.2. We will call the measure μ an *OC measure* if it satisfies this *orthogonality condition*:

$$\boxplus_{\mu} = \boxminus_{\mu}. \quad (\text{OC})$$

These measures are so fundamental to the rest of the paper that they warrant extra discussion. Note that being an OC measure is only a constraint on how μ behaves with respect to polynomials of degree at most (n, m) . When μ is a finite measure, being an OC measure is a condition on the moments of μ , as is explained in [Knese 2008, Appendix].

Discussion of OC measures. Recall $\mathcal{F}_{\mu} = \mathbb{C}[z, w] \cap L^2(\mu)$. Here are four ways to interpret OC measures:

- Every $p \in \mathcal{F}_{\mu}$ of degree at most (n, m) with $\hat{p}(n, m) = 0$ which is orthogonal to polynomials in \mathcal{F}_{μ} of degree at most $(n, m-1)$ automatically satisfies

$$\hat{p}(n, k) = 0 \text{ for } k = 0, 1, \dots, m-1.$$

In symbols:

$$(p \in \square_{\mu} \text{ and } p \perp \square_{\mu}) \implies p \in \square_{\mu}.$$

- Every $p \in \mathcal{F}_\mu$ of degree at most $(n-1, m)$ which is orthogonal to all polynomials in \mathcal{F}_μ of degree at most $(n-1, m-1)$ is automatically orthogonal to all polynomials in \mathcal{F}_μ of degree at most $(n, m-1)$. In symbols:

$$(p \in \square_\mu \text{ and } p \perp \square_\mu) \implies p \perp \square_\mu.$$

- An OC measure satisfies a certain inclusion-exclusion principle:

$$0 = K\square_\mu - K\boxplus_\mu - K\boxminus_\mu + K\boxtimes_\mu. \tag{6-2}$$

To see this, consider the decompositions

$$\begin{aligned} K\square_\mu &= K\boxplus_\mu + K\boxminus_\mu + K\boxtimes_\mu, \\ K\boxplus_\mu &= K\boxtimes_\mu + K\boxminus_\mu, \\ K\boxminus_\mu &= K\boxtimes_\mu + K\boxplus_\mu. \end{aligned}$$

When μ is an OC measure, $\boxplus_\mu = \boxminus_\mu$. This yields [Equation \(6-2\)](#).

The symmetry in [Equation \(6-2\)](#) also proves that

$$\boxplus_\mu = \boxminus_\mu \quad \text{if and only if} \quad \boxtimes_\mu = \boxminus_\mu.$$

- An OC measure behaves like a *Bernstein–Szegő measure*

$$\frac{1}{|q(z, w)|^2} d\sigma(z, w);$$

here $q \in \mathbb{C}[z, w]$ has no zeros on \mathbb{D}^2 . [Section 7](#) is devoted to this fact and its converse: Bernstein–Szegő measures are OC measures! See [Corollary 7.6](#) and [Theorem 7.4](#). □

Additionally, if $\boxplus_\mu = \boxminus_\mu$ holds, then we have

$$\boxplus_\mu = \boxminus_\mu$$

by reflecting these subspaces (polynomial reflection is an antiunitary and so preserves orthogonality relations).

Therefore, if μ is an OC measure then the ϵ in [Theorem 6.1](#), given by $(K\boxplus_\mu - K\boxminus_\mu) - (K\boxtimes_\mu - K\boxminus_\mu)$, disappears.

Hence, if μ is an OC measure, we have

$$q(z, w)\overline{q(Z, \bar{W})} - \tilde{q}(z, w)\overline{\tilde{q}(Z, \bar{W})} = (1 - z\bar{Z})K\boxtimes_\mu((z, w), (Z, W)) + (1 - w\bar{W})K\boxtimes_\mu((z, w), (Z, W)),$$

where q is any unit norm polynomial in $\text{Max}_\mu = \boxplus_\mu$.

Evaluating on the diagonal $(z, w) = (Z, W)$ we have

$$\begin{aligned} |q(z, w)|^2 &\geq |q(z, w)|^2 - |\tilde{q}(z, w)|^2 \\ &= (1 - |z|^2)K\boxtimes_\mu((z, w), (z, w)) + (1 - |w|^2)K\boxtimes_\mu((z, w), (z, w)) \geq 0, \end{aligned} \tag{6-3}$$

for all $(z, w) \in \overline{\mathbb{D}^2}$. If we scrutinize this inequality, we can prove something quite strong.

Proposition 6.3. *Let μ be an OC measure and let q be any unit norm polynomial in Max_μ . If $q(z_0, w_0)$ vanishes for some $(z_0, w_0) \in \overline{\mathbb{D}^2}$, every element of \square_μ vanishes at (z_0, w_0) .*

Proof. Recall two formulas from above. By [Proposition 5.5](#)

$$K_{\square_\mu} - K_{\square_\mu} = (1 - |z|^2)K_{\square_\mu} \tag{6-4}$$

and by [\(4-5\)](#)

$$K_{\square_\mu} = K_{\square_\mu} + K_{\square_\mu} + K_{\square_\mu} + \overleftarrow{q\overline{q}}, \tag{6-5}$$

where every reproducing kernel is evaluated on the diagonal $(z, w) = (Z, W)$.

First, suppose $(z_0, w_0) \in \mathbb{D}^2$. We write $v = (z_0, w_0)$ for short. From [\(6-3\)](#), it is immediate that $q(v) = 0$ implies

$$\overleftarrow{q}(v) = K_{\square_\mu}(v, v) = K_{\square_\mu}(v, v) = 0. \tag{6-6}$$

Then, $K_{\square_\mu}(v, v) = 0$ by formulas [\(6-4\)](#) and [\(6-5\)](#). Indeed, $K_{\square_\mu}(v, v) = 0$ implies $K_{\square_\mu}(v, v) = K_{\square_\mu}(v, v) = 0$ by [\(6-4\)](#) (using the fact that reproducing kernels are nonnegative on the diagonal). Then, [\(6-5\)](#) implies $K_{\square_\mu}(v, v) = 0$ since $K_{\square_\mu} = K_{\square_\mu}$ by assumption. If $K_{\square_\mu}(v, v) = 0$ then every element of \square_μ must vanish at v .

To prove the claim for $v = (z_0, w_0) \in \overline{\mathbb{D}^2} \setminus \mathbb{D}^2$, notice that the left hand side of [\(6-3\)](#) vanishes to order at least two at v , and the terms $(1 - |z|^2)$ and $(1 - |w|^2)$ can vanish to order at most one. This again implies [\(6-6\)](#) and by a similar argument $K_{\square_\mu}(v, v) = 0$.

Therefore, every element of \square_μ vanishes at a zero of q in $\overline{\mathbb{D}^2}$. □

Remark 6.4. If μ is a finite measure, then $1 \in \square_\mu$ and this implies q has no zeros on the closed bidisk. Hence, this proves stability in the case of probability measures, as in [[Geronimo and Woerdeman 2004](#); [Knese 2008](#)].

Corollary 6.5. *Suppose μ is an OC measure and let q be any unit norm polynomial in Max_μ . Then, q can be factored into $q = q_1q_2$ where*

- q_1 divides every element of \square_μ ;
- every irreducible factor of q_1 is \mathbb{T}^2 -symmetric, has infinitely many zeros in $\overline{\mathbb{D}^2}$, and vanishes somewhere on \mathbb{T}^2 ; and
- q_2 has no zeros in $\overline{\mathbb{D}^2} \setminus \mathbb{T}^2$ and finitely many zeros in \mathbb{T}^2 .

Proof. It is clear q may be factored into the form $q = q_1q_2$ where every irreducible factor of q_1 has infinitely many zeros in $\overline{\mathbb{D}^2}$ and q_2 has finitely many zeros in $\overline{\mathbb{D}^2}$ (we of course allow for the case where q_1 or q_2 is a constant).

Suppose f is an irreducible factor of q possessing infinitely many zeros in $\overline{\mathbb{D}^2}$; that is, a factor of q_1 . By [Proposition 6.3](#), every element of \square_μ has infinitely many zeros in common with f and hence f divides every element of \square_μ . So, f can be divided out of both sides of the inequality [\(6-3\)](#) and using the resulting inequality one can then show that if f occurs in the factorization of q with multiplicity, it then divides every element of \square_μ with the same multiplicity. Hence, q_1 divides every element of \square_μ . By [Proposition 5.7](#), any such f necessarily is \mathbb{T}^2 -symmetric and vanishes somewhere on \mathbb{T}^2 . This proves the first two items in the statement of the corollary.

Finally, if q_2 has finitely many zeros in $\overline{\mathbb{D}^2}$, q_2 can have no zeros in the bidisk. By [Lemma 3.5](#), q_2 can have no zeros on the sides: $\mathbb{D} \times \mathbb{T}$ and $\mathbb{T} \times \mathbb{D}$. This proves the third item. \square

Since the factor q_1 in the above corollary divides every element of \square_μ , the study of μ and \square_μ can be separated into the study of q_1 and the study of $|q_1|^2 d\mu$ and the set \square_μ/q_1 (which is nothing more than all $p \in L^2(|q_1|^2 d\mu)$ of degree less than or equal to $(n - n_1, m - m_1)$, where (n_1, m_1) is the degree of q_1). Indeed, the map sending

$$f \in \square_\mu \mapsto f/q_1 \in \square_\mu/q_1$$

is an isometry (using the inner product of $L^2(\mu)$ on the left and the inner product of $L^2(|q_1|^2 d\mu)$ on the right). Although this is a somewhat trivial observation, we now feel justified in making the assumption that Min_μ and Max_μ have no common factor, a statement equivalent to saying q and \tilde{q} have no common factor. A statement which is in turn equivalent to saying q_1 is a constant. The following proposition is immediate, since its hypotheses imply $q = q_2$ in [Corollary 6.5](#).

Proposition 6.6. *If μ is an OC measure and if Max_μ and Min_μ are one-dimensional and have no factor in common, then any $q \in \text{Max}_\mu$ is almost stable.*

Lemma 6.7. *Suppose Min_μ is one-dimensional and has no factor in common with Max_μ , and suppose μ is an OC measure. Then,*

$$\dim \mathfrak{A}_\mu = n \quad \text{and} \quad \dim \mathfrak{B}_\mu = m.$$

Proof. Let h be a unit norm polynomial in Min_μ . The polynomial h necessarily has degree exactly (n, m) , otherwise it would be orthogonal to itself. Set $q = \tilde{h}$, where the reflection is performed at the (n, m) level. By [Theorem 6.1](#) with $\epsilon = 0$,

$$\begin{aligned} q(z, w)\overline{q(Z, W)} - \tilde{q}(z, w)\overline{\tilde{q}(Z, W)} \\ = (1 - z\bar{Z})K_{\mathfrak{A}_\mu}((z, w), (Z, W)) + (1 - w\bar{W})K_{\mathfrak{B}_\mu}((z, w), (Z, W)). \end{aligned}$$

Let $d_1 = \dim \mathfrak{A}_\mu$ and $d_2 = \dim \mathfrak{B}_\mu$; let e_1, \dots, e_{d_1} be an orthonormal basis for \mathfrak{A}_μ and f_1, \dots, f_{d_2} an orthonormal basis for \mathfrak{B}_μ . We write these vectorially as

$$\mathbf{E}(z, w) = \begin{pmatrix} e_1(z, w) \\ \vdots \\ e_{d_1}(z, w) \end{pmatrix} \quad \text{and} \quad \mathbf{F}(z, w) = \begin{pmatrix} f_1(z, w) \\ \vdots \\ f_{d_2}(z, w) \end{pmatrix},$$

and then the formula above becomes

$$q(z, w)\overline{q(Z, W)} - \tilde{q}(z, w)\overline{\tilde{q}(Z, W)} = (1 - z\bar{Z})\langle \mathbf{E}(z, w), \mathbf{E}(Z, W) \rangle + (1 - w\bar{W})\langle \mathbf{F}(z, w), \mathbf{F}(Z, W) \rangle.$$

Upon rearranging we have

$$\begin{aligned} q(z, w)\overline{q(Z, W)} + z\bar{Z}\langle \mathbf{E}(z, w), \mathbf{E}(Z, W) \rangle + w\bar{W}\langle \mathbf{F}(z, w), \mathbf{F}(Z, W) \rangle \\ = \tilde{q}(z, w)\overline{\tilde{q}(Z, W)} + \langle \mathbf{E}(z, w), \mathbf{E}(Z, W) \rangle + \langle \mathbf{F}(z, w), \mathbf{F}(Z, W) \rangle. \end{aligned}$$

The map defined by

$$\begin{pmatrix} q(z, w) \\ z\mathbf{E}(z, w) \\ w\mathbf{F}(z, w) \end{pmatrix} \mapsto \begin{pmatrix} \tilde{q}(z, w) \\ \mathbf{E}(z, w) \\ \mathbf{F}(z, w) \end{pmatrix}$$

for each $(z, w) \in \mathbb{C}^2$ defines a unitary on the span of the elements in $\mathbb{C}^{1+d_1+d_2}$ of the form on the left to the span of the elements in $\mathbb{C}^{1+d_1+d_2}$ of the form on the right, which can be extended to a unitary matrix U of dimensions $(1+d_1+d_2) \times (1+d_1+d_2)$. We write U in block form as

$$U = \begin{matrix} & \mathbb{C} & \mathbb{C}^{d_1+d_2} \\ \mathbb{C} & & \\ \mathbb{C}^{d_1+d_2} & & \end{matrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

We also define a $\mathbb{C}^{d_1+d_2}$ -valued polynomial \mathbf{G} by

$$\mathbf{G}(z, w) := \begin{pmatrix} \mathbf{E}(z, w) \\ \mathbf{F}(z, w) \end{pmatrix},$$

and define the $(d_1 + d_2) \times (d_1 + d_2)$ diagonal matrix

$$\Delta(z, w) := \begin{pmatrix} zI_{d_1} & 0 \\ 0 & wI_{d_2} \end{pmatrix}.$$

Then,

$$\begin{aligned} Aq(z, w) + B\Delta(z, w)\mathbf{G}(z, w) &= \tilde{q}(z, w), \\ Cq(z, w) + D\Delta(z, w)\mathbf{G}(z, w) &= \mathbf{G}(z, w). \end{aligned}$$

The latter formula implies

$$\mathbf{G}(z, w) = q(z, w)(I - D\Delta(z, w))^{-1}C,$$

and in turn the former formula implies

$$A + B\Delta(z, w)(I - D\Delta(z, w))^{-1}C = \frac{\tilde{q}(z, w)}{q(z, w)}.$$

Since \tilde{q}/q is already in reduced terms we must have $d_1 \geq n$ and $d_2 \geq m$. We already know $d_1 \leq n$ and $d_2 \leq m$ (see [Proposition 5.3](#)). Therefore, $n = \dim \square_\mu$ and $m = \dim \square_\mu$, and the result follows. \square

Theorem 6.8 (spectral matching). *Let μ and ρ be two OC measures. Suppose $\text{Max}_\mu = \text{Max}_\rho \neq \{0\}$ and let $q \in \text{Max}_\mu$. Assume q and \tilde{q} have no common factor. Then, $\square_\mu = \square_\rho$ and the inner products $\langle \cdot, \cdot \rangle_\mu$ and $\langle \cdot, \cdot \rangle_\rho$ agree up to a constant multiple on \square_μ ; that is,*

$$\frac{1}{\|q\|_{L^2(\mu)}^2} \langle f, g \rangle_\mu = \frac{1}{\|q\|_{L^2(\rho)}^2} \langle f, g \rangle_\rho \quad \text{for all } f, g \in \square_\mu.$$

In other words,

$$\frac{1}{\|q\|_{L^2(\mu)}^2} K_{\square_\mu} = \frac{1}{\|q\|_{L^2(\rho)}^2} K_{\square_\rho}.$$

Proof. We may renormalize μ and ρ so that $1 = \|q\|_{L^2(\mu)} = \|q\|_{L^2(\rho)}$.

By choosing orthonormal bases for the n -dimensional subspaces (by [Lemma 6.7](#)) \boxplus_μ and \boxplus_ρ , we may write

$$K_{\boxplus_\mu}((z, w), (Z, W)) = \langle \mathbf{E}_\mu(z, w), \mathbf{E}_\mu(Z, W) \rangle,$$

$$K_{\boxplus_\rho}((z, w), (Z, W)) = \langle \mathbf{E}_\rho(z, w), \mathbf{E}_\rho(Z, W) \rangle,$$

for $\mathbf{E}_\mu, \mathbf{E}_\rho \in \mathbb{C}^n[z, w]$.

Likewise, we may write the m -dimensional subspaces \boxminus_μ and \boxminus_ρ as

$$K_{\boxminus_\mu}((z, w), (Z, W)) = \langle \mathbf{F}_\mu(z, w), \mathbf{F}_\mu(Z, W) \rangle,$$

$$K_{\boxminus_\rho}((z, w), (Z, W)) = \langle \mathbf{F}_\rho(z, w), \mathbf{F}_\rho(Z, W) \rangle,$$

where $\mathbf{F}_\mu, \mathbf{F}_\rho \in \mathbb{C}^m[z, w]$.

By [Proposition 5.3](#), both $\mathbf{E}_\mu, \mathbf{F}_\mu$ and $\mathbf{E}_\rho, \mathbf{F}_\rho$ satisfy the hypotheses of [Lemma 3.4](#) (in place of \mathbf{E}, \mathbf{F} and $\tilde{\mathbf{E}}, \tilde{\mathbf{F}}$), since by [Theorem 6.1](#), we have

$$\begin{aligned} (1 - z\bar{z})K_{\boxplus_\mu}((z, w), (Z, W)) + (1 - w\bar{w})K_{\boxminus_\mu}((z, w), (Z, W)) \\ = (1 - z\bar{z})K_{\boxplus_\rho}((z, w), (Z, W)) + (1 - w\bar{w})K_{\boxminus_\rho}((z, w), (Z, W)). \end{aligned}$$

Therefore, \mathbf{E}_μ is a unitary multiple of \mathbf{E}_ρ and \mathbf{F}_μ is a unitary multiple of \mathbf{F}_ρ . In other words,

$$K_{\boxplus_\mu}((z, w), (Z, W)) = K_{\boxplus_\rho}((z, w), (Z, W)),$$

$$K_{\boxminus_\mu}((z, w), (Z, W)) = K_{\boxminus_\rho}((z, w), (Z, W)). \quad (6-7)$$

Now we will see that this is all that is needed to reassemble the two inner products on \square_μ or \square_ρ .

By reflection

$$K_{\boxplus_\mu}((z, w), (Z, W)) = K_{\boxplus_\rho}((z, w), (Z, W)),$$

and by the formulas (which hold for both μ and ρ)

$$K_{\boxplus_\mu} - K_{\boxminus_\mu} = (1 - |z|^2)K_{\square_\mu} \quad (\text{Proposition 5.5})$$

and

$$K_{\square_\mu} = K_{\boxplus_\mu} + K_{\boxminus_\mu} + K_{\boxminus_\mu} + \overleftarrow{q\bar{q}} \quad (\text{Equation (4-5)})$$

where every reproducing kernel is evaluated on the diagonal $(z, w) = (Z, W)$, we see that

$$K_{\square_\mu} = K_{\square_\rho}.$$

(This is similar to the argument in the proof of [Proposition 6.3](#).) □

7. Bernstein–Szegő measures

Converse to the previous section, we now study Bernstein–Szegő measures, which will be shown to be OC measures. Bernstein–Szegő measures are measures on \mathbb{T}^2 of the form

$$d\mu = \frac{1}{|q(z, w)|^2} d\sigma(z, w),$$

where $q \in \mathbb{C}[z, w]$ has no zeros on \mathbb{D}^2 . (Recall $d\sigma$ is normalized Lebesgue measure on \mathbb{T}^2 .)

The following proposition looks innocuous, but it addresses the main technical difficulty *not present* in the case of polynomials with no zeros on the entire *closed* bidisk. Note this proposition does not require the polynomial to have finitely many zeros on \mathbb{T}^2 .

Proposition 7.1. *Let $q \in \mathbb{C}[z, w]$ have degree at most (n, m) and no zeros on \mathbb{D}^2 . Define a measure on \mathbb{T}^2 by*

$$d\mu = \frac{1}{|q(z, w)|^2} d\sigma(z, w).$$

Then, $q \perp_{\mu}$ and more generally

$$q \perp_{\mu} \{f \in L^2(\mu) : \hat{f}(j, k) = 0 \text{ for } k < 0 \text{ and for } k = 0 \text{ and } j \leq 0\}.$$

Proof. Let $f \in L^2(\mu)$ satisfy

$$\hat{f}(j, k) = 0 \quad \text{for } k < 0 \text{ and for } k = 0 \text{ and } j \leq 0.$$

It is necessarily true that $f \in L^2(\mathbb{T}^2)$. For almost every $z \in \mathbb{T}$, the function $f_{(z)}(w) = f(z, w)$ is in $L^2(\mathbb{T})$ and since $\hat{f}(j, k) = 0$ for $k < 0$, $f_{(z)}$ is actually in $H^2(\mathbb{T})$ for almost every $z \in \mathbb{T}$.

So, the function (of w)

$$g_{(z)}(w) := \frac{f(z, w)}{q(z, w)}$$

is in the Smirnov class N^+ (which consists of all ratios of bounded analytic functions with outer denominator; see [Duren 1970, Section 2.5]), for almost every $z \in \mathbb{T}$: $q(z, \cdot)$ has no zeros in the disk for all but finitely many $z \in \mathbb{T}$ (by Lemma 3.5) and is therefore *outer* for almost every $z \in \mathbb{T}$. Since $f \in L^2(\mu)$, Fubini's theorem says that for almost every $z \in \mathbb{T}$, we have $g_{(z)} \in L^2(\mathbb{T})$. By Theorem 2.11 in [Duren 1970], $N^+ \cap L^2(\mathbb{T}) = H^2(\mathbb{T})$, and therefore $g_{(z)} \in H^2(\mathbb{T})$ for almost every $z \in \mathbb{T}$.

Owing to the fact that $g_{(z)}$ is orthogonal to w^j for $j < 0$,

$$f(z, 0) = \int_{\mathbb{T}} f(z, w) \frac{dw}{2\pi i w} = \int_{\mathbb{T}} \frac{f(z, w)}{q(z, w)} q(z, w) \frac{dw}{2\pi i w} = \int_{\mathbb{T}} \frac{f(z, w)}{q(z, w)} q(z, 0) \frac{dw}{2\pi i w}$$

for almost every $z \in \mathbb{T}$, and so

$$\int_{\mathbb{T}^2} \frac{f(z, w)}{q(z, w)} \frac{dw}{2\pi i w} \frac{dz}{2\pi i z} = \int_{\mathbb{T}} \frac{f(z, 0)}{q(z, 0)} \frac{dz}{2\pi i z}.$$

Now, the function defined by $h(z) = f(z, 0)/q(z, 0)$ is in $L^2(\mathbb{T})$ by Fubini's theorem. Also, h is in the Smirnov class N^+ because $f(\cdot, 0)$ is in $H^2(\mathbb{T})$ (by the assumption that $\hat{f}(j, 0) = 0$ for $j \leq 0$), and $q(\cdot, 0)$ is outer since $q(z, 0)$ has no zeros in the disk. Therefore, h is in $H^2(\mathbb{T})$. Thus, we may conclude

$$\int_{\mathbb{T}^2} \frac{f(z, w)}{q(z, w)} \frac{dw}{2\pi i w} \frac{dz}{2\pi i z} = \int_{\mathbb{T}} \frac{f(z, 0)}{q(z, 0)} \frac{dz}{2\pi i z} = \frac{f(0, 0)}{q(0, 0)} = 0,$$

since $\hat{f}(0, 0) = 0$.

Since

$$\langle f, q \rangle_\mu = \int_{\mathbb{T}^2} \frac{f(z, w) \overline{q(z, w)}}{|q(z, w)|^2} d\sigma(z, w) = \int_{\mathbb{T}^2} \frac{f(z, w)}{q(z, w)} d\sigma(z, w),$$

we have shown $\langle f, q \rangle_\mu = 0$, or in other words $f \perp_\mu q$. \square

From here, the proofs follow the stable case, as in [Knese 2008], with some minor changes.

Corollary 7.2. *If $f \in L^2(\mu) \cap H^2(\mathbb{T}^2)$ and*

$$\hat{f}(j, k) = 0 \text{ for } k > m \text{ and for } k = m \text{ and } j \geq n,$$

then $\langle f, \tilde{q}g \rangle_\mu = 0$ for any $g \in H^\infty(\mathbb{T}^2)$.

Proof. Notice that $\langle \tilde{q}g, f \rangle_\mu = \langle \tilde{f}gz^n w^m, q \rangle_\mu$. Also, notice that $\tilde{f}gz^n w^m$ satisfies the hypotheses of the previous proposition (it helps to draw a picture of the frequency support of f and $\tilde{f}gz^n w^m$). Therefore, $\langle f, \tilde{q}g \rangle_\mu = 0$. \square

Lemma 7.3. *Define*

$$L_{(Z, W)}(z, w) = L((z, w), (Z, W)) = (z\bar{Z})^n \frac{q(z, w) \overline{q(1/\bar{z}, W)} - \tilde{q}(z, w) \overline{\tilde{q}(1/\bar{z}, W)}}{(1 - z\bar{Z})(1 - w\bar{W})}. \quad (7-1)$$

Suppose $f \in L^2(\mu) \cap H^2(\mathbb{T}^2)$, with $\hat{f}(j, k) = 0$ for $k > m$ and for $k = m$ and $j \geq n$. Then, for $(Z, W) \in \mathbb{D}^2$,

$$\sum_{k=0}^{m-1} \sum_{j=n}^{\infty} \hat{f}(j, k) Z^j W^k = \langle f, L_{(Z, W)} \rangle_\mu.$$

Proof. By Corollary 7.2, f is orthogonal to the function

$$G_{(Z, W)}(z, w) = \frac{\tilde{q}(z, w) z^n \overline{\tilde{q}(1/\bar{z}, W)}}{(1 - z\bar{Z})(1 - w\bar{W})}$$

for each $(Z, W) \in \mathbb{D}^2$.

Therefore,

$$\begin{aligned} \langle f, L_{(Z, W)} \rangle_\mu &= \int_{\mathbb{T}^2} \frac{f(z, w) \overline{q(z, w)} q(z, W) (\bar{z}Z)^n}{(1 - \bar{z}Z)(1 - \bar{w}W) |q(z, w)|^2} \frac{dw dz}{(2\pi i)^2 z w} \\ &= \int_{\mathbb{T}} \int_{\mathbb{T}} \frac{f(z, w) q(z, W) (\bar{z}Z)^n}{(1 - \bar{z}Z)(w - W) q(z, w)} \frac{dw}{2\pi i} \frac{dz}{2\pi i} \end{aligned} \quad (7-2)$$

$$= \int_{\mathbb{T}} \frac{f(z, W)}{q(z, W)} q(z, W) \frac{(\bar{z}Z)^n}{(1 - \bar{z}Z)} \frac{dz}{2\pi i z} \quad (7-3)$$

$$= \sum_{j=n}^{\infty} \sum_{k=0}^{m-1} \hat{f}(j, k) Z^j W^k. \quad (7-4)$$

Going from (7-2) to (7-3) is an application of the Cauchy integral formula and going from (7-3) to (7-4) involves cancellation and another application of the Cauchy integral formula. \square

Theorem 7.4. *Let q be a nonzero polynomial of degree at most (n, m) with no zeros on \mathbb{D}^2 . Define a measure on \mathbb{T}^2 by*

$$d\mu = \frac{1}{|q(z, w)|^2} d\sigma(z, w).$$

Then, μ is an OC measure.

Proof. Let

$$\text{HS} = \{f \in L^2(\mu) \cap H^2(\mathbb{T}^2) : \hat{f}(j, k) = 0 \text{ for } k \geq m\},$$

(HS = half strip) and let

$$\text{NHS} = \{f \in L^2(\mu) \cap H^2(\mathbb{T}^2) : \hat{f}(j, k) = 0 \text{ for } k > m \text{ and when } k = m \text{ and } j \geq n\},$$

(NHS = notched half strip).

We claim that $\text{NHS} \ominus_\mu \text{HS} = \mathfrak{I}_\mu$. To prove $\text{NHS} \ominus_\mu \text{HS} \subset \mathfrak{I}_\mu$, notice that $L_{(Z, W)}$ from Lemma 7.3 is in HS since the numerator of $L_{(Z, W)}$ vanishes when $w = 1/\bar{W}$, and hence $L_{(Z, W)}$ is a polynomial of degree at most $m-1$ in w . So, if $f \in \text{NHS} \ominus_\mu \text{HS}$, then

$$0 = \langle f, L_{(Z, W)} \rangle_\mu = \sum_{j=n}^{\infty} \sum_{k=0}^{m-1} \hat{f}(j, k) Z^j W^k,$$

which means $f \in \mathfrak{I}_\mu$ and therefore $f \in \mathfrak{I}_\mu$. So, $\text{NHS} \ominus_\mu \text{HS} \subset \mathfrak{I}_\mu$.

To prove that $\mathfrak{I}_\mu \subset \text{NHS} \ominus_\mu \text{HS}$, let $P_{\text{HS}} : L^2(\mu) \rightarrow \text{HS}$ denote the orthogonal projection onto HS, a necessarily closed subspace of $L^2(\mu)$ (the topology on $L^2(\mu)$ is finer than the topology on $L^2(\mathbb{T}^2)$). If $f \in \mathfrak{I}_\mu$ then

$$f - P_{\text{HS}} f \in \text{NHS} \ominus_\mu \text{HS} \subset \mathfrak{I}_\mu,$$

and this implies

$$P_{\text{HS}} f \in \mathfrak{I}_\mu \cap \text{HS} = \{0\}.$$

Hence, $P_{\text{HS}} f = 0$ which means $f \perp_\mu \text{HS}$. In other words, $f \in \text{NHS} \ominus_\mu \text{HS}$. Hence, $\text{NHS} \ominus_\mu \text{HS} = \mathfrak{I}_\mu$.

Now, since $\mathfrak{I}_\mu \subset \text{NHS} \ominus_\mu \text{HS}$, it follows that $\mathfrak{I}_\mu \subset \mathfrak{I}_\mu$. A similar argument to the above (using the projection P_{HS}) proves $\mathfrak{I}_\mu \subset \text{NHS} \ominus_\mu \text{HS} = \mathfrak{I}_\mu$. This implies $\mathfrak{I}_\mu = \mathfrak{I}_\mu$; namely, μ is an OC measure. \square

Corollary 7.5. *Let q be a nonzero polynomial of degree at most (n, m) with no zeros on \mathbb{D}^2 . Define a measure on \mathbb{T}^2 by*

$$d\mu = \frac{1}{|q(z, w)|^2} d\sigma(z, w).$$

Then,

$$q(z, w)\overline{q(Z, W)} - \tilde{q}(z, w)\overline{\tilde{q}(Z, W)} = (1 - z\bar{Z})K_{\mathfrak{I}_\mu}((z, w), (Z, W)) + (1 - w\bar{W})K_{\mathfrak{I}_\mu}((z, w), (Z, W)).$$

Proof. Proposition 7.1 says $q \in \text{Max}_\mu$ and Theorem 7.4 says $\mathfrak{I}_\mu = \mathfrak{I}_\mu$. Since $\|q\|_{L^2(\mu)} = 1$, the conclusion follows from Theorem 6.1 since $\mathfrak{I}_\mu = \mathfrak{I}_\mu$ says $\epsilon = 0$. \square

Corollary 7.6 (“Bernstein–Szegő approximation”). *Let ρ be an OC measure and suppose $q \in \text{Max}_\rho$ has no factors in common with \tilde{q} . Define*

$$d\mu = \frac{1}{|q(z, w)|^2} d\sigma(z, w).$$

If we normalize ρ so that $\|q\|_{L^2(\rho)} = 1$, then $\square_\rho = \square_\mu$ and

$$K\square_\rho = K\square_\mu,$$

that is, the inner products on \square_μ and \square_ρ from $L^2(\mu)$ and $L^2(\rho)$ agree.

Proof. By Proposition 7.1 $q \in \text{Max}_\mu$ and by Theorem 7.4, μ is an OC measure. We have assumed q has no factors in common with \tilde{q} and this allows us to apply Theorem 6.8, from which the conclusion follows immediately. □

One final lemma will make the proof of the main theorem a matter of bookkeeping. We use the following notations:

$$Z_q = \{(z, w) \in \mathbb{C}^2 : q(z, w) = 0\}, \tag{7-5}$$

$$\pi_1(z, w) = z \text{ and } \pi_2(z, w) = w. \tag{7-6}$$

Lemma 7.7. *If μ is the Bernstein–Szegő measure associated to $q \in \mathbb{C}[z, w]$, that is,*

$$d\mu = \frac{1}{|q(z, w)|^2} d\sigma(z, w),$$

then $J(z, w) = (w - w_0)$ and $L(z, w) = (z - z_0)$ will be divisors of the ideal \mathcal{F}_μ whenever $w_0 \notin \pi_2(Z_q \cap \mathbb{T}^2)$ and $z_0 \notin \pi_1(Z_q \cap \mathbb{T}^2)$, respectively.

Proof. If $(z - z_0)f(z, w) \in L^2(\mu)$ for some $f \in \mathbb{C}[z, w]$ and $z_0 \notin \pi_1(Z_q \cap \mathbb{T}^2)$, then let U be a neighborhood of $Z_{z-z_0} \cap \mathbb{T}^2$ which does not intersect Z_q . Then, $|z - z_0|^2$ is bounded below on $\mathbb{T}^2 \setminus U$ and $|q|^2$ is bounded below on U , say by a constant c . Then,

$$\infty > \int_{\mathbb{T}^2} \frac{|z - z_0|^2 |f(z, w)|^2}{|q(z, w)|^2} d\sigma \geq \int_{\mathbb{T}^2 \setminus U} \frac{c |f(z, w)|^2}{|q(z, w)|^2} d\sigma$$

and

$$\infty > \int_U |f(z, w)|^2 d\sigma \geq \int_U \frac{c |f(z, w)|^2}{|q(z, w)|^2} d\sigma$$

together imply

$$\|f\|_{L^2(\mu)}^2 = \int_U \frac{|f(z, w)|^2}{|q(z, w)|^2} d\sigma + \int_{\mathbb{T}^2 \setminus U} \frac{|f(z, w)|^2}{|q(z, w)|^2} d\sigma < \infty.$$

So, L is a divisor of \mathcal{F}_μ . The proof for J is similar. □

8. Proof of the main theorem

We have all of the pieces in place to prove the theorem from the introduction. Here is the main theorem with extra details filled in. When we use the inner product notation $\langle \cdot, \cdot \rangle$ below with no subscript, we are taking inner products in \mathbb{C}^N (where the N is taken from context) and not taking any kind of Hilbert function space inner product.

Theorem 8.1. *Let $q \in \mathbb{C}[z, w]$ be almost stable with $\deg q \leq (n, m)$. Then, there exist vector polynomials $\mathbf{E} \in \mathbb{C}^n[z, w]$ and $\mathbf{F} \in \mathbb{C}^m[z, w]$, with $\deg \mathbf{E} \leq (n-1, m)$, and $\deg \mathbf{F} \leq (n, m-1)$, satisfying the following conditions:*

- (1) \mathbf{E} is horizontally $\mathbb{D} \cup X$ -stable where $X = \mathbb{T} \setminus (\pi_2(Z_q))$.
- (2) $\tilde{\mathbf{F}}$ is vertically $\mathbb{D} \cup Y$ -stable where $Y = \mathbb{T} \setminus (\pi_1(Z_q))$.
- (3) $q(z, w)\overline{q(Z, W)} - \tilde{q}(z, w)\overline{\tilde{q}(Z, W)}$
 $= (1 - z\bar{Z})(\mathbf{E}(z, w), \mathbf{E}(Z, W)) + (1 - w\bar{W})(\mathbf{F}(z, w), \mathbf{F}(Z, W)).$ (8-1)
- (4) If $\tilde{\mathbf{E}} \in \mathbb{C}^n[z, w]$ and $\tilde{\mathbf{F}} \in \mathbb{C}^m[z, w]$ satisfy items (1) and (3) above in place of \mathbf{E} and \mathbf{F} , then there exist unitary matrices U_1, U_2 such that

$$\mathbf{E}(z, w) = U_1 \tilde{\mathbf{E}}(z, w) \quad \text{and} \quad \mathbf{F}(z, w) = U_2 \tilde{\mathbf{F}}(z, w).$$

Proof. We use the setup (and conclusion) of [Corollary 7.5](#). By [Lemma 6.7](#), \mathbb{E}_μ has dimension n and \mathbb{F}_μ has dimension m . Let $\{e_1, \dots, e_n\}$ be an orthonormal basis of \mathbb{E}_μ and $\{f_1, \dots, f_m\}$ an orthonormal basis of \mathbb{F}_μ . Define $\mathbf{E} = (e_1, \dots, e_n)^t \in \mathbb{C}^n[z, w]$ and $\mathbf{F} = (f_1, \dots, f_m)^t \in \mathbb{C}^m[z, w]$. [Corollary 7.5](#) now proves item (3).

Write $\mathbf{E}(z, w) = E(w)\Lambda_n(z)$ and $\mathbf{F}(z, w) = F(z)\Lambda_m(w)$. With these choices, [Proposition 5.3](#) says $E(w)$ is invertible for all $w \in \overline{\mathbb{D}}$ with the exception of $w_0 \in \mathbb{T}$ with the property that $w - w_0$ is not a divisor of \mathcal{F}_μ . [Lemma 7.7](#) says $(w - w_0)$ is a divisor of \mathcal{F}_μ when $w_0 \notin \pi_2(Z_q \cap \mathbb{T}^2)$. So, $E(w)$ is invertible when $w \in \overline{\mathbb{D}} \setminus \pi_2(Z_q \cap \mathbb{T}^2)$. The entries of

$$\tilde{\mathbf{F}}(z, w) = z^n w^{m-1} \overline{\mathbf{F}(1/\bar{z}, 1/\bar{w})}$$

form an orthonormal basis for \mathbb{F}_μ and

$$\tilde{\mathbf{F}}(z, w) = z^n \overline{F(1/\bar{z})} w^{m-1} \overline{\Lambda_m(1/\bar{w})} = z^n \overline{F(1/\bar{z})} \chi \Lambda_m(w),$$

where χ is the $m \times m$ matrix with ones on the antidiagonal (entries $(j, m - j)$) and zeros elsewhere. By [Proposition 5.3](#) and [Lemma 7.7](#) $z^n \overline{F(1/\bar{z})} \chi$ is invertible for $z \in \overline{\mathbb{D}} \setminus \pi_1(Z_q \cap \mathbb{T}^2)$. Of course, χ is invertible, so the same statement holds for $z^n \overline{F(1/\bar{z})}$. This proves items (1) and (2) of [Theorem 8.1](#).

[Lemma 3.4](#) proves item (4). □

9. Polynomials with unique decompositions

In this section we give a characterization of the polynomials with no zeros on the bidisk that have a unique sums of squares decomposition.

Proof of Theorem 1.15. Suppose q is almost stable with $\deg p = (n, m)$.

To prove item (1) implies (2) in [Theorem 1.15](#), suppose there are unique Γ_1 and Γ_2 , sums of squared moduli of two-variable polynomials, such that

$$|q(z, w)|^2 - |\tilde{q}(z, w)|^2 = (1 - |z|^2)\Gamma_1(z, w) + (1 - |w|^2)\Gamma_2(z, w).$$

By [Corollary 7.5](#), if μ is the Bernstein–Szegő measure associated to q then

$$\begin{aligned} |q(z, w)|^2 - |\tilde{q}(z, w)|^2 &= (1 - |z|^2)K_{\boxminus_\mu}((z, w), (z, w)) + (1 - |w|^2)K_{\boxplus_\mu}((z, w), (z, w)) \\ &= (1 - |z|^2)K_{\boxminus_\mu}((z, w), (z, w)) + (1 - |w|^2)K_{\boxminus_\mu}((z, w), (z, w)). \end{aligned}$$

These reproducing kernels can be written as sums of squares of two variable polynomials. Since we are assuming such decompositions are unique we have

$$K_{\boxminus_\mu}((z, w), (z, w)) = K_{\boxplus_\mu}((z, w), (z, w)).$$

Because of the formula ([Proposition 5.5](#))

$$K_{\boxplus_\mu}((z, w), (z, w)) - K_{\boxminus_\mu}((z, w), (z, w)) = (1 - |w|^2)K_{\square_\mu}((z, w), (z, w)), \tag{9-1}$$

we see that

$$K_{\square_\mu}((z, w), (z, w)) = 0.$$

This implies $\square_\mu = \{0\}$. In other words, there are no nonzero $f \in \square \cap L^2(\mu) = \square \cap L^2(1/|q|^2 d\sigma)$ and this just says there are no nonzero $f \in \square$ such that

$$f/q \in L^2(\mathbb{T}^2).$$

This proves that item (1) implies item (2) in [Theorem 1.15](#).

To prove item (2) implies (3) in the theorem, assume there are no nonzero $f \in \square$ such that

$$f/q \in L^2(\mathbb{T}^2).$$

Notationally, $\square_\mu = \{0\}$ and again by ([9-1](#)) we have

$$K_{\boxminus_\mu}((z, w), (z, w)) = K_{\boxplus_\mu}((z, w), (z, w)).$$

The two subspaces \boxplus_μ and \boxminus_μ are reflections of one another. So, if we write

$$K_{\boxplus_\mu}((z, w), (z, w)) = K_{\boxminus_\mu}((z, w), (z, w)) = |\mathbf{E}(z, w)|^2,$$

where $\mathbf{E}(z, w) = (E_1(z, w), \dots, E_n(z, w))^t \in \mathbb{C}^n[z, w]$ and E_1, \dots, E_n are an orthonormal basis for $\boxplus_\mu = \boxminus_\mu$, then the entries of

$$\tilde{\mathbf{E}}(z, w) := z^{n-1} w^m \overline{\mathbf{E}(1/\bar{z}, 1/\bar{w})}$$

also form an orthonormal basis for $\boxplus_\mu = \boxminus_\mu$. So,

$$|\mathbf{E}(z, w)|^2 = |\tilde{\mathbf{E}}(z, w)|^2,$$

and by [Lemma 3.2](#) there is an $n \times n$ unitary matrix U such that

$$U\mathbf{E}(z, w) = \tilde{\mathbf{E}}(z, w).$$

(As we commented there [Lemma 3.2](#) holds for two-variable polynomials just as well.) If we reflect both sides of this equation (take conjugates, replace (z, w) with $(1/\bar{z}, 1/\bar{w})$, and multiply through by $z^{n-1}w^m$) we see that

$$\bar{U}\bar{\mathbf{E}}(z, w) = \mathbf{E}(z, w).$$

Note that \bar{U} is the matrix obtained by taking complex conjugates of each entry of U and is not the adjoint of U . In fact, $\bar{U}^{-1} = U^t$ and therefore

$$U^t\mathbf{E}(z, w) = \bar{\mathbf{E}}(z, w) = U\mathbf{E}(z, w).$$

Hence, $U = U^t$ since the vectors $\mathbf{E}(z, w)$ span all of \mathbb{C}^n as (z, w) varies over \mathbb{D}^2 (by [Proposition 5.3](#)). The matrix U is therefore symmetric unitary. Symmetric unitaries can be factored as $U = V^tV$ where V is a unitary — this is the so-called Takagi factorization. The vector polynomial

$$V\mathbf{E}(z, w)$$

is then symmetric since its reflection is

$$\bar{V}\bar{\mathbf{E}}(z, w) = (V^t)^{-1}U\mathbf{E}(z, w) = V\mathbf{E}(z, w)$$

as $U = V^tV$. So we replace \mathbf{E} with $V\mathbf{E}$ and this proves there exists a symmetric vector polynomial \mathbf{E} such that

$$K_{\square_{\mu}}((z, w), (z, w)) = K_{\square_{\mu}}((z, w), (z, w)) = |\mathbf{E}(z, w)|^2.$$

By [Proposition 5.3](#), \mathbf{E} is horizontally $\mathbb{D} \cup \mathbb{E}$ -stable, since \mathbf{E} and $\bar{\mathbf{E}}$ are both horizontally \mathbb{D} -stable.

Similar arguments show that when $\square_{\mu} = \{0\}$, there exists a symmetric vector polynomial $\mathbf{F} \in \mathbb{C}^m[z, w]$ of degree $(n, m-1)$ which is vertically $\mathbb{D} \cup \mathbb{E}$ -stable, and

$$K_{\square_{\mu}}((z, w), (z, w)) = K_{\square_{\mu}}((z, w), (z, w)) = |\mathbf{F}(z, w)|^2.$$

By [Corollary 7.5](#), we have that

$$|q(z, w)|^2 - |\bar{q}(z, w)|^2 = (1 - |z|^2)|\mathbf{E}(z, w)|^2 + (1 - |w|^2)|\mathbf{F}(z, w)|^2, \tag{9-2}$$

where \mathbf{E} and \mathbf{F} satisfy all of the desired properties. This proves item (2) implies item (3).

To prove item (3) implies (1) assume (9-2) holds where \mathbf{E} is horizontally \mathbb{D} -stable and \mathbf{F} is vertically \mathbb{D} -stable. We must show this is the only sums of squares decomposition for q .

Suppose there are vector polynomials $\mathbf{A} \in \mathbb{C}^N[z, w]$, $\mathbf{B} \in \mathbb{C}^M[z, w]$ such that

$$|q(z, w)|^2 - |\bar{q}(z, w)|^2 = (1 - |z|^2)|\mathbf{A}(z, w)|^2 + (1 - |w|^2)|\mathbf{B}(z, w)|^2.$$

Setting $|w| = 1$, [Equation \(9-2\)](#) implies

$$|\mathbf{E}(z, w)|^2 = |\mathbf{A}(z, w)|^2 \quad \text{for } (z, w) \in \mathbb{C} \times \mathbb{T}.$$

Since \mathbf{E} is horizontally \mathbb{D} -stable, [Lemma 3.3](#) applies: $n \leq N$ and there exists a one variable $N \times n$ matrix valued rational inner function Ψ_1 such that

$$\mathbf{A}(z, w) = \Psi_1(w)\mathbf{E}(z, w) \quad \text{for } (z, w) \in \mathbb{D}^2.$$

By similar reasoning, $m \leq M$ and there exists an $M \times m$ matrix valued rational inner function Ψ_2 such that

$$\mathbf{B}(z, w) = \Psi_2(z)\mathbf{F}(z, w).$$

So,

$$|\mathbf{A}(z, w)|^2 \leq |\mathbf{E}(z, w)|^2, \quad |\mathbf{B}(z, w)|^2 \leq |\mathbf{F}(z, w)|^2 \quad \text{for all } (z, w) \in \mathbb{D}^2.$$

However, we must have equality at every point in both of these inequalities because otherwise

$$(1 - |z|^2)|\mathbf{E}(z, w)|^2 + (1 - |w|^2)|\mathbf{F}(z, w)|^2 = (1 - |z|^2)|\mathbf{A}(z, w)|^2 + (1 - |w|^2)|\mathbf{B}(z, w)|^2$$

would be violated. Hence, the sums of squares terms for q are unique:

$$|\mathbf{A}(z, w)|^2 = |\mathbf{E}(z, w)|^2, \quad |\mathbf{B}(z, w)|^2 = |\mathbf{F}(z, w)|^2 \quad \text{for all } (z, w) \in \mathbb{C}^2.$$

This proves (3) implies (1) and concludes the proof. □

Corollary 1.16 says that the only stable polynomials with a unique decomposition are one variable polynomials. We prove this now.

Proof of Corollary 1.16. Suppose $p \in \mathbb{C}[z, w]$ is stable and $\deg p = (n, m)$. It is implicit in most of this paper that $n, m > 0$. By **Theorem 1.15**, since $1/|p|^2$ is integrable, it follows that p does not have a unique sums of squares decomposition. If $n = 0$ or $m = 0$ then p is really just a one variable polynomial with no zeros on closed disk. It is well known that the decomposition in the one variable Christoffel–Darboux formula is unique, since the sums of squares term can just be solved for; it equals

$$\frac{|p(z)|^2 - |\tilde{p}(z)|^2}{1 - |z|^2}$$

in the case where $m = 0$. □

10. Application: Fejér–Riesz factorization

The classical Fejér–Riesz theorem says that a nonnegative one variable trigonometric polynomial t can be factored as $|p(z)|^2$ where $p \in \mathbb{C}[z]$ has no zeros in the disk \mathbb{D} . It is false that all nonnegative *two* variable trig polynomials can be factored as $|p(z, w)|^2$ where $p \in \mathbb{C}[z, w]$ has no zeros on the bidisk. Indeed, Geronimo and Woerdeman [2004] give a characterization of which *strictly positive* trig polynomials have a “Fejér–Riesz type factorization”. We reprove and extend this result to certain cases of *nonnegative* trigonometric polynomials. Our proof does not make use of a certain “maximal entropy result” and is therefore self-contained.

We emphasize that requiring a *finite* measure μ to be an OC measure is a condition on its moments [Knese 2008, Appendix]. First, let us establish the strictly positive result.

Theorem 10.1 [Geronimo and Woerdeman 2004]. *Let $t : \mathbb{T}^2 \rightarrow \mathbb{C}$ be a positive trigonometric polynomial of two variables with Fourier coefficients $\hat{t}(j, k)$ supported on the set $|j| \leq n, |k| \leq m$. Then, there exists a stable $p \in \mathbb{C}[z, w]$, $\deg p \leq (n, m)$ satisfying $t(z, w) = |p(z, w)|^2$ for all $(z, w) \in \mathbb{T}^2$ if and only if the measure $d\mu = (1/t)d\sigma$ is an OC measure.*

Proof. The “only if” direction follows from [Theorem 7.4](#). To prove the “if” direction, observe that if μ is an OC measure, then by [Corollary 7.6](#), if p is a unit norm polynomial in Max_μ , then p is stable (see [Remark 6.4](#)) and defining

$$d\rho = \frac{1}{|p(z, w)|^2} d\sigma,$$

we have that the inner products on $L^2(\mu)$ and $L^2(\rho)$ agree when restricted to \square . So, the moments agree:

$$\int_{\mathbb{T}^2} z^j w^k d\mu = \int_{\mathbb{T}^2} z^j w^k d\rho \quad \text{for } |j| \leq n, |k| \leq m.$$

Here is where we deviate from the Geronimo–Woerdeman proof. Observe that

$$1 = \int_{\mathbb{T}^2} \frac{|p(z, w)|}{\sqrt{t(z, w)}} \frac{\sqrt{t(z, w)}}{|p(z, w)|} \sigma \leq \sqrt{\int_{\mathbb{T}^2} \frac{|p(z, w)|^2}{t(z, w)} d\sigma} \sqrt{\int_{\mathbb{T}^2} \frac{t(z, w)}{|p(z, w)|^2} d\sigma} = \|p\|_{L^2(\mu)} \sqrt{\|t\|_{L^1(\rho)}} \quad (10-1)$$

by Cauchy–Schwarz. Now, $\|p\|_{L^2(\mu)} = 1$ since p was chosen to have unit norm, and since the moments of μ and ρ agree,

$$\|t\|_{L^1(\rho)} = \|t\|_{L^1(\mu)} = \int_{\mathbb{T}^2} \frac{t(z, w)}{t(z, w)} d\sigma = 1.$$

Therefore, we have equality in the above application of Cauchy–Schwarz ([Equation \(10-1\)](#)). So, $|p|/\sqrt{t}$ and $\sqrt{t}/|p|$ are multiples of one another, implying $|p|^2 = ct$ for some constant c . The constant c must be 1 since p has unit norm in $L^2(\mu)$. Hence, $t(z, w) = |p(z, w)|^2$ for $(z, w) \in \mathbb{T}^2$. \square

We would like to extend this result to the case of nonnegative trigonometric polynomials, and we have some results in this direction. Work on characterizing when a nonnegative operator-valued two variable polynomial has a Fejér–Riesz type factorization was done in [\[Dritschel and Woerdeman 2005\]](#). (Although the subtleties of all of the different candidates for the notion of *outerness* in several variables seem to have prevented getting a necessary and sufficient condition for a Fejér–Riesz factorization in that paper.)

We believe that any Fejér–Riesz type factorization for nonnegative two-variable trigonometric polynomials should take into account the notions of *toral* and *atoral* polynomials. These notions were alluded to in [Remark 1.11](#).

Example 10.2. Consider the nonnegative trigonometric polynomial $t(z, w) = |z - w|^2$. It cannot be factored as $|p(z, w)|^2$ where $p \in \mathbb{C}[z, w]$ has no zeros on the bidisk, because p would necessarily vanish on the set $\{(z, w) \in \mathbb{T}^2 : z = w\}$ and therefore $z - w$ would divide p . So, the polynomial $zwt(z, w) = 2zw - z^2 - w^2$ associated to t has a toral factor, and since this toral factor has zeros in the bidisk, there is no hope for such a Fejér–Riesz type of factorization. So, the question of whether a Fejér–Riesz factorization exists depends on the properties of the toral factors of t . This is true more generally.

Let $t : \mathbb{T}^2 \rightarrow \mathbb{C}$ be a nonnegative trigonometric polynomial of two variables, given by

$$t(z, w) = \sum_{j=-N}^N \sum_{k=-M}^M t_{jk} z^j w^k \geq 0,$$

and let $q(z, w) := z^N w^M t(z, w) \in \mathbb{C}[z, w]$.

Lemma 10.3. *If q has an irreducible toral factor p , then p^2 divides q , and $t/|p|^2$ is a nonnegative trigonometric polynomial.*

Proof. Write $q = hp$ for some $h \in \mathbb{C}[z, w]$. By definition of toral, p has infinitely many zeros on \mathbb{T}^2 . The lemma is not difficult in the case where p is a linear polynomial in one variable alone, so we assume this is not the case. Suppose p has degree (n, m) . Let $(z_0, w_0) \in \mathbb{T}^2 \cap Z_p$ with the property that $p(\cdot, w_0)$ has a zero of multiplicity one at z_0 and $t(\cdot, w_0)$ is not identically zero; this will be the case for all but finitely many of the $(z, w) \in \mathbb{T}^2 \cap Z_p$. Now, $t(z, w_0) = z^{-N}w_0^{-M}h(z, w_0)p(z, w_0)$, and as $t(\cdot, w_0)$ is a nonnegative trig polynomial of one variable, it must have zeros of even order on \mathbb{T} . Hence, $h(z_0, w_0) = 0$. Therefore, h and p share infinitely many zeros, and this implies p divides h by irreducibility of p . Hence, p^2 divides q . Toral polynomials are \mathbb{T}^2 -symmetric in the sense that

$$\bar{p} = cp$$

for some unimodular constant c . So,

$$t(z, w) = z^{-N}w^{-M}p(z, w)^2g(z, w) = z^{-N+n}w^{-M+m}|p(z, w)|^2g(z, w) \quad \text{for some } g \in \mathbb{C}[z, w].$$

Thus, $t/|p|^2$ is a nonnegative trig polynomial. □

Corollary 10.4. *If t is a nonnegative trigonometric polynomial, then t can be factored into $t(z, w) = |p(z, w)|^2s(z, w)$ where $p \in \mathbb{C}[z, w]$ is a toral polynomial (or is a constant) and s is a nonnegative trigonometric polynomial with finitely many zeros on \mathbb{T}^2 .*

The corollary divides the study of characterizing trig polynomials with a Fejér–Riesz factorization into the question of when a toral polynomial has no zeros on the bidisk and when a nonnegative trig polynomial finitely many zeros on the torus has a Fejér–Riesz factorization.

To introduce the next result we recall that every positive two variable trigonometric polynomial can be written as a sum of squares of two-variable polynomials. This was proved in [Dritschel 2004] and reproved in [Geronimo and Lai 2006] (the latter paper has a summary of related known results). It is unknown if all nonnegative trigonometric polynomials can be written as a sum of squares of two variable polynomials. The above corollary says that it is enough to address this question for trig polynomials with finitely many zeros. On the other hand, if it is true that all nonnegative trig polynomials are equal to a sum of squares of polynomials, then our approach allows us to characterize when they can be written as a single square of a polynomial with no zeros on the bidisk.

Theorem 10.5. *Suppose $p_1, \dots, p_N \in \mathbb{C}[z, w]$ have degree at most (n, m) and no common factor. Also, assume that for some j , $p_j(0, 0) \neq 0$. Let*

$$t(z, w) = \sum_{j=1}^N |p_j(z, w)|^2 \text{ for } (z, w) \in \mathbb{T}^2,$$

and define $d\mu = (1/t)d\sigma$. The trigonometric polynomial t can be written as $t(z, w) = |p(z, w)|^2$, where p has no zeros on the bidisk, if and only if μ is an OC measure.

If every p_j vanishes at the origin, we could apply a Möbius transformation to make sure not all of the polynomials vanish at the origin and then apply the above theorem to check whether the trig polynomial has the desired factorization.

Proof. Our proof in the case of a strictly positive trig polynomial carries over with some modifications. The “only if” direction again follows from [Theorem 7.4](#). Let us prove that if μ is an OC measure then t has a Fejér–Riesz type of decomposition.

Since t is of the given form it is clear that each $p_j \in L^2(\mu)$, as $|p_j|^2/t \leq 1$ on the torus. The assumption that $p_j(0, 0) \neq 0$ guarantees that Max_μ is nonempty (since we then know $\square_\mu \neq \square_\mu$). Let q be a unit norm polynomial in Max_μ . By [Corollary 6.5](#), q is almost stable. To see this, note the corollary says q can be factored as q_1q_2 where q_1 divides every element of \square_μ and q_2 is almost stable, but we assumed p_1, \dots, p_N have no common factor. Hence, q_1 must be a constant.

Define

$$d\rho = \frac{1}{|q(z, w)|^2} d\sigma.$$

By [Corollary 7.6](#), $\square_\mu = \square_\rho$ and the inner products of $L^2(\mu)$ and $L^2(\rho)$ agree on \square_μ . In particular,

$$p_j/q \in L^2(\mathbb{T}^2) \quad \text{for each } j.$$

Just as in the proof in the strictly positive case, we can prove

$$1 \leq \|q\|_{L^2(\mu)} \sqrt{\|t\|_{L^1(\rho)}}$$

by an application of Cauchy–Schwarz. Since q has unit norm, $\|q\|_{L^2(\mu)} = 1$, and since the inner products agree, we have

$$\|t\|_{L^1(\rho)} = \sum_{j=1}^N \|p_j\|_{L^2(\rho)}^2 = \sum_{j=1}^N \|p_j\|_{L^2(\mu)}^2 = \|t\|_{L^1(\mu)} = 1.$$

Therefore, just as in the proof for the strictly positive case, we have equality in Cauchy–Schwarz, which implies $t = |q|^2$ on the torus. □

So, the above theorem addresses nonnegative trig polynomials of a specific form. The above proof would also work if we could decompose t as

$$t(z, w) = \sum_{j=1}^N p_j(z, w) \overline{q_j(z, w)},$$

where $p_j, q_j \in L^2((1/t)d\sigma)$ have no common factor and not all vanish at $(0, 0)$.

Question 10.6. Can every nonnegative two variable trigonometric polynomial t be decomposed as

$$t(z, w) = \sum_{j=1}^N p_j(z, w) \overline{q_j(z, w)},$$

where p_j, q_j are in $L^2(\frac{1}{t}d\sigma)$ and have no common factor?

Next, we tackle toral factors of nonnegative trig polynomials.

Theorem 10.7. An irreducible toral polynomial $p \in \mathbb{C}[z, w]$ has no zeros in the bidisk if and only if

$$\frac{\overleftarrow{\partial} p}{\partial z} + \frac{\overleftarrow{\partial} p}{\partial w}$$

is almost stable. In this case, all of the zeros on \mathbb{T}^2 occur at singularities of Z_p (i.e., common zeros of $\partial p/\partial z$ and $\partial p/\partial w$).

The above reflections are performed at the degrees of $\partial p/\partial z$ and $\partial p/\partial w$ that would generically be expected. Namely, if p has degree (n, m) , we reflect $\partial p/\partial z$ at the degree $(n-1, m)$.

Proof. If p is toral, then p is necessarily \mathbb{T}^2 symmetric, meaning p is a unimodular constant times \overleftarrow{p} (and in fact we may assume $p = \overleftarrow{p}$ by multiplying by an appropriate constant). It is proved in [Knese 2009] that if p is \mathbb{T}^2 symmetric and has no zeros in the bidisk, then

$$\frac{\overleftarrow{\partial p}}{\partial z} + \frac{\overleftarrow{\partial p}}{\partial w}$$

has no zeros in the set $\overline{\mathbb{D}^2}$ except possibly at singularities of Z_p (and there can be at most finitely many singularities).

Conversely, suppose $\overleftarrow{\partial p/\partial z} + \overleftarrow{\partial p/\partial w}$ is almost stable. This implies

$$\phi(z, w) = \frac{z(\partial p/\partial z)(z, w) + w(\partial p/\partial w)(z, w)}{(\overleftarrow{\partial p/\partial z})(z, w) + (\overleftarrow{\partial p/\partial w})(z, w)}$$

is a (nonconstant) inner function on the bidisk, and must be bounded by 1 in modulus on the bidisk.

It is also proved in [Knese 2009] that if p is \mathbb{T}^2 symmetric, then

$$(n + m)p(z, w) = z \frac{\partial p}{\partial z}(z, w) + w \frac{\partial p}{\partial w}(z, w) + \frac{\overleftarrow{\partial p}}{\partial z}(z, w) + \frac{\overleftarrow{\partial p}}{\partial w}(z, w).$$

So, if $p(z, w) = 0$ for some $(z, w) \in \mathbb{D}^2$, then $|\phi(z, w)| = 1$, which is a contradiction. Therefore, p has no zeros in the bidisk. □

Remark 10.8. We view this as progress on determining which nonnegative trig polynomials have a Fejér–Riesz decomposition for the following reasons. A nonnegative trig polynomial has a unique toral factor $|p|^2$ and determining whether p has no zeros in the bidisk can be approached by looking at each factor of p . For the factors f whose zero sets have no singularities on the torus, the above theorem says we can check whether $\overleftarrow{\partial f/\partial z} + \overleftarrow{\partial f/\partial w}$ is stable. A two-variable Schur–Cohn test, such as the one presented in [Geronimo and Woerdeman 2004], can be used to check this condition. For factors with singularities on the torus, one would need to adapt the Schur–Cohn test to the *almost* stable case. We leave this for future work.

To summarize, given a nonnegative trig polynomial t we can factor it into $t(z, w) = |p(z, w)|^2 s(z, w)$ where p is a toral polynomial and s is a nonnegative trig polynomial with finitely many zeros on \mathbb{T}^2 . The above remark addresses cases where we can determine whether p has no zeros in the bidisk. If s has no zeros on the torus, the Geronimo–Woerdeman theorem characterizes whether it can be factored as $|q|^2$ where q is stable. We have extended this characterization to a class of nonnegative trig polynomials with a special form, for which it is unknown whether this is all nonnegative trig polynomials.

11. Application: distinguished varieties

One of our main applications is a bounded analytic extension theorem for distinguished varieties, which we now define.

Definition 11.1. A nonempty subset $V \subset \mathbb{C}^2$ is a *distinguished variety* if V is an algebraic curve: there exists $p \in \mathbb{C}[z, w]$ such that

$$V = \{(z, w) \in \mathbb{C}^2 : p(z, w) = 0\}$$

and V exits the bidisk through the distinguished boundary

$$\partial(V \cap \overline{\mathbb{D}^2}) \subset \mathbb{T}^2.$$

Our goal is to prove the following result. (This is a more qualitative version of [Theorem 11.4](#) below.)

Theorem 11.2. *Let $V \subset \mathbb{C}^2$ be a distinguished variety. Then, there is a rational function of z , $C(z)$, with no poles in \mathbb{D} , such that for every $f \in \mathbb{C}[z, w]$, there is a rational function $F \in \mathbb{C}(z, w)$, holomorphic on \mathbb{D}^2 , which agrees with f on V :*

$$F(z, w) = f(z, w) \quad \text{for all } (z, w) \in V \cap \mathbb{D}^2$$

and satisfies the estimate

$$|F(z, w)| \leq |C(z)| \sup_{V \cap \mathbb{D}^2} |f| \quad \text{for all } (z, w) \in \mathbb{D}^2.$$

If V has no singularities on \mathbb{T}^2 , $C(z)$ can be taken to be a constant.

The last statement is already proved in [\[Knese 2009\]](#). Essentially, the purpose of this section is to inject the work of this paper into the work of [\[Knese 2009\]](#). The use of the Cole–Wermer sums of squares formula is essential to the work in [\[Knese 2009\]](#), and if we use [Theorem 1.3](#) in its place, the following lengthy theorem can be proved by slightly modifying the proofs in [\[Knese 2009\]](#).

Theorem 11.3. *Let V be a distinguished variety given as the zero set of a square-free polynomial $p \in \mathbb{C}[z, w]$ of degree (n, m) . Let $a, b > 0$ be positive real numbers. Then, there exist $\mathbf{P} \in \mathbb{C}^n[z, w]$, $\deg \mathbf{P} \leq (n-1, m)$, and $\mathbf{Q} \in \mathbb{C}^m[z, w]$, $\deg \mathbf{Q} \leq (n, m-1)$ such that*

- \mathbf{P} is horizontally $\mathbb{D} \cup X_2$ -stable and \mathbf{Q} is vertically $\mathbb{D} \cup X_1$ -stable where $X_2 = \mathbb{T} \setminus \pi_2(S)$, $X_1 = \mathbb{T} \setminus \pi_1(S)$ and S is the set of singularities of V ;
- $(bm - an)|p(z, w)|^2 + 2 \operatorname{Re} \left[\left(az \frac{\partial p}{\partial z}(z, w) - bw \frac{\partial p}{\partial w}(z, w) \right) \overline{p(z, w)} \right] + (1 - |z|^2)|\mathbf{P}(z, w)|^2 = (1 - |w|^2)|\mathbf{Q}(z, w)|^2$;
- there is a $m \times m$ matrix-valued rational inner function $\Phi : \mathbb{D} \rightarrow \mathbb{C}^{m \times m}$ such that V has the following representation

$$V \cap \mathbb{D}^2 = \{(z, w) \in \mathbb{D}^2 : \det(wI_m - \Phi(z)) = 0\},$$

and \mathbf{Q} is a “polynomial eigenvector” for Φ :

$$\Phi(z)\mathbf{Q}(z, w) = w\mathbf{Q}(z, w) \quad \text{for all } (z, w) \in V.$$

Guide to the proof. Everything above is contained in a theorem in [Knese 2009] except for the horizontal and vertical stability of \mathbf{P} , \mathbf{Q} , respectively. So let us briefly outline how all of this can be done. All of the following are proved in [Knese 2009]:

- (1) If $p \in \mathbb{C}[z, w]$ has degree (n, m) and defines a distinguished variety, then the polynomial

$$q(z, w) = z^n p\left(\frac{1}{z}, w\right)$$

is \mathbb{T}^2 -symmetric and has no zeros on the bidisk.

- (2) Such a q has the property that for each $a, b > 0$

$$a \frac{\overleftarrow{\partial} q}{\partial z} + b \frac{\overleftarrow{\partial} q}{\partial w}$$

has no zeros on the closed bidisk $\overline{\mathbb{D}^2}$ except possibly at the finite number of singularities of Z_q , which necessarily occur on \mathbb{T}^2 .

- (3) Such a q satisfies

$$\begin{aligned} (an + bm)^2 |q(z, w)|^2 - 2 \operatorname{Re}[(azq_z(z, w) + bwq_w(z, w))(an + bm)\overline{q(z, w)}] \\ = \left| a \frac{\overleftarrow{\partial} q}{\partial z}(z, w) + b \frac{\overleftarrow{\partial} q}{\partial w}(z, w) \right|^2 - \left| az \frac{\partial q}{\partial z}(z, w) + bw \frac{\partial q}{\partial w}(z, w) \right|^2. \end{aligned} \quad (11-1)$$

By Theorem 8.1, this last item (11-1) can be written as

$$(1 - |z|^2)|\mathbf{E}(z, w)|^2 + (1 - |w|^2)|\mathbf{F}(z, w)|^2$$

where \mathbf{E} is horizontally $\mathbb{D} \cup Y_2$ -stable and $\overleftarrow{\mathbf{F}}$ is vertically $\mathbb{D} \cup Y_1$ -stable; here $Y_2 = \mathbb{T} \setminus \pi_2(S_q)$, $Y_1 = \mathbb{T} \setminus \pi_1(S_q)$, and S_q is the set of singularities of q . If we convert back to statements involving the polynomial p (by replacing z with $1/z$ and multiplying by z^n) we get

$$\begin{aligned} (bm - an)|p(z, w)|^2 + 2 \operatorname{Re} \left[\left(az \frac{\partial p}{\partial z}(z, w) - bw \frac{\partial p}{\partial w}(z, w) \right) \overline{p(z, w)} \right] + (1 - |z|^2)|\mathbf{P}(z, w)|^2 \\ = (1 - |w|^2)|\mathbf{Q}(z, w)|^2, \end{aligned}$$

where \mathbf{P} is horizontally $\mathbb{D} \cup X_2$ -stable, \mathbf{Q} is vertically $\mathbb{D} \cup X_1$ -stable, $X_2 = \mathbb{T} \setminus \pi_2(S)$, $X_1 = \mathbb{T} \setminus \pi_1(S)$, and S is the set of singularities of V .

For the rest of the theorem, the proofs in [Knese 2009] can be applied unchanged. \square

Here is the promised *bounded analytic extension* theorem. The proof is identical to the proof in [Knese 2009] for distinguished varieties with no singularities on the torus. The only difference is that we did not have Theorem 1.3 to tell us that \mathbf{Q} as above is vertically $\mathbb{D} \cup X_1$ -stable, where $X_1 = \mathbb{T} \setminus \pi_1(S)$. (In the case of no singularities we already knew \mathbf{Q} is vertically $\overline{\mathbb{D}}$ -stable.)

Let us write

$$\mathbf{Q}(z, w) = Q(z)\Lambda_m(w),$$

where the matrix polynomial $Q(z)$ is invertible for all $z \in \mathbb{D} \cup X_1$ (by definition of vertical $\mathbb{D} \cup X_1$ -stability).

Theorem 11.4. *Let V be a distinguished variety and let Φ , Q , and \mathbf{Q} be as in [Theorem 11.3](#). Then, for any polynomial $f \in \mathbb{C}[z, w]$, the rational function*

$$F(z, w) := (1, 0, \dots, 0)Q(z)^{-1}f(zI_m, \Phi(z))\mathbf{Q}(z, w)$$

is equal to f on $V \cap \mathbb{D}^2$ and we have the estimates

$$|F(z, w)| \leq \|Q(z)^{-1}\| \|\mathbf{Q}(z, w)\| \sup_{V \cap \mathbb{D}^2} |f| \leq \sqrt{m} \|Q(z)^{-1}\| \|Q(z)\| \sup_{V \cap \mathbb{D}^2} |f| \quad \text{for all } (z, w) \in \mathbb{D}^2.$$

Here we are taking the operator norm of the matrices $Q(z)$ and $Q(z)^{-1}$.

In words, the growth of the extension F is controlled by a rational function of one variable. When V has no singularities on \mathbb{T}^2 , $Q(z)$ is invertible for $z \in \overline{\mathbb{D}}$ and

$$\sup_{\mathbb{D}} \|Q(z)^{-1}\| \|Q(z)\|$$

is a finite constant.

The following is an example of the above two theorems.

Example 11.5. Consider the following reducible distinguished variety in \mathbb{C}^2

$$V = \{(z, w) \in \mathbb{C}^2 : (z - w)(z^2 - w) = 0\}.$$

Like all distinguished varieties it has a *determinantal representation* of the following form:

$$V \cap \mathbb{D}^2 = \{(z, w) \in \mathbb{D}^2 : \det(wI - \Phi(z)) = 0\},$$

where Φ is a rational matrix valued inner function. One choice of Φ is

$$\Phi(z) = \frac{1}{2} \begin{pmatrix} z(1+z) & z^2(1-z) \\ (1-z) & z(1+z) \end{pmatrix}.$$

As can easily be checked

$$\det(wI_2 - \Phi(z)) = w^2 - zw - z^2w + z^3 = (w - z)(w - z^2).$$

The variety V is simple yet instructive because it has a singularity at the origin and more importantly a singularity on the torus at the point $(1, 1)$.

One choice for \mathbf{Q} as above is

$$\mathbf{Q}(z, w) = \begin{pmatrix} 2w - z - z^2 \\ 1 - z \end{pmatrix}.$$

Writing

$$\mathbf{Q}(z, w) = Q(z) \begin{pmatrix} 1 \\ w \end{pmatrix},$$

where

$$Q(z) = \begin{pmatrix} -z - z^2 & 2 \\ 1 - z & 0 \end{pmatrix},$$

we note that $Q(z)$ is invertible in $\overline{\mathbb{D}} \setminus \{1\}$; that is, \mathbf{Q} is vertically $\overline{\mathbb{D}} \setminus \{1\}$ -stable.

The analytic extension theorem now works as follows.

Let $f \in \mathbb{C}[z, w]$ which we think of as a function on V . Then, the rational function

$$F(z, w) = (1, 0)Q(z)^{-1}f(zI, \Phi(z))Q(z, w)$$

agrees with f on V because Q is a polynomial eigenvector for Φ on V . Furthermore, the size of F on the bidisk can be estimated purely in terms of a fixed rational function of z and the supremum of f on $V \cap \mathbb{D}^2$.

Indeed,

$$|F(z, w)| \leq |(1, 0)Q(z)^{-1}| |Q(z, w)| \sup_{V \cap \mathbb{D}^2} |f| \leq \sqrt{1 + \frac{16}{|1-z|^2}} \sup_{V \cap \mathbb{D}^2} |f|.$$

12. Application: Agler’s Pick interpolation theorem

As another application we give a simple proof of necessity in the Pick interpolation theorem on the bidisk. The proof below sidesteps the use of Andô’s inequality and cone-separation arguments found in most proofs. (The proof of sufficiency can be accomplished with a “lurking isometry” argument; see [Lemma 6.7](#) for something similar.) The proof is very similar to the argument in [\[Cole and Wermer 1999\]](#) for establishing Andô’s inequality from the sum of squares decomposition.

Theorem 12.1 (Agler). *Given distinct points*

$$(z_1, w_1), \dots, (z_N, w_N) \in \mathbb{D}^2$$

and complex numbers

$$c_1, \dots, c_N \in \mathbb{D},$$

there exists a holomorphic function $f : \mathbb{D}^2 \rightarrow \mathbb{D}$ which interpolates

$$f(z_j, w_j) = c_j \text{ for } j = 1, 2, \dots, N$$

if and only if there exist positive semidefinite $N \times N$ matrices Γ and Δ such that

$$1 - c_j \bar{c}_k = (1 - z_j \bar{z}_k) \Gamma_{jk} + (1 - w_j \bar{w}_k) \Delta_{jk}.$$

Proof of necessity. We first prove the theorem for rational inner functions and then use an approximation theorem to prove necessity in general. So, let f be a rational inner function on the bidisk. Every rational inner function can be written as $f = \tilde{p}/p$ for some $p \in \mathbb{C}[z, w]$ of degree at most (n, m) having no zeros on the bidisk [\[Rudin 1969, Section 5.5.1\]](#). Decomposing p as in [\(8-1\)](#) and setting $(z, w) = (z_j, w_j)$ and $(Z, W) = (z_k, w_k)$ we have

$$\begin{aligned} p(z_j, w_j) \overline{p(z_k, w_k)} - \tilde{p}(z_j, w_j) \overline{\tilde{p}(z_k, w_k)} \\ = (1 - z_j \bar{z}_k) \langle \mathbf{E}(z_j, w_j), \mathbf{E}(z_k, w_k) \rangle + (1 - w_j \bar{w}_k) \langle \mathbf{F}(z_j, w_j), \mathbf{F}(z_k, w_k) \rangle. \end{aligned}$$

Therefore, if $f(z_j, w_j) = (\tilde{p}/p)(z_j, w_j) = c_j$, then

$$\Gamma_{jk} = \frac{1}{p(z_j, w_j) \overline{p(z_k, w_k)}} \langle \mathbf{E}(z_j, w_j), \mathbf{E}(z_k, w_k) \rangle, \quad \Delta_{jk} = \frac{1}{p(z_j, w_j) \overline{p(z_k, w_k)}} \langle \mathbf{F}(z_j, w_j), \mathbf{F}(z_k, w_k) \rangle$$

are both positive semidefinite matrices and they satisfy

$$1 - c_j \bar{c}_k = (1 - z_j \bar{z}_k) \Gamma_{jk} + (1 - w_j \bar{w}_k) \Delta_{jk}, \tag{12-1}$$

as desired.

In general, suppose $f : \mathbb{D}^2 \rightarrow \mathbb{D}$ is holomorphic and $f(z_j, w_j) = c_j$. Rudin’s extension of Carathéodory’s theorem to the polydisk [Rudin 1969, Theorem 5.5.1] says that f is the pointwise limit of a sequence of rational inner functions: $f_\alpha \rightarrow f$ as $\alpha \rightarrow \infty$, where α is used to index the positive integers. Corresponding to each such rational inner function f_α , we write $f_\alpha(z_j, w_j) = c_{\alpha,j}$ and we choose positive semidefinite matrices $\Gamma_\alpha, \Delta_\alpha$ so that an equation analogous to (12-1) holds:

$$1 - c_{\alpha,j} \bar{c}_{\alpha,k} = (1 - z_j \bar{z}_k) (\Gamma_\alpha)_{jk} + (1 - w_j \bar{w}_k) (\Delta_\alpha)_{jk}. \tag{12-2}$$

The set of positive semidefinite matrices (of a fixed size) with diagonal entries bounded by some constant is compact (their operator norms are bounded by their traces which are uniformly bounded). The diagonal entries of Γ_α and Δ_α are bounded independently of α (e.g., it is not hard to prove

$$\frac{1}{1 - |z_j|^2} \geq (\Gamma_\alpha)_{jj}$$

for $j = 1, \dots, N$) and therefore we may choose a subsequence so that Γ_α converges to some positive semidefinite matrix Γ and Δ_α converges to some positive semidefinite matrix Δ . Therefore, if we take the limit as $\alpha \rightarrow \infty$ in Equation (12-2) we have proved

$$1 - c_j \bar{c}_k = (1 - z_j \bar{z}_k) \Gamma_{jk} + (1 - w_j \bar{w}_k) \Delta_{jk},$$

which proves necessity in general. □

Question 12.2. Can the uniqueness in Theorem 1.3 be carried over in some way to the above theorem?

Solutions to extremal Pick problems in two variables (those solvable with a function of norm one but no less) are not unique as they are in one variable, so we are necessarily vague in our question.

13. Questions

We have already asked three questions: Questions 5.8, 10.6, and 12.2. Here are two others. One of the most fundamental questions to come out of our research is the following:

Question 13.1. When is a rational function p/q in $L^2(\mathbb{T}^2)$?

Here we may as well assume $p, q \in \mathbb{C}[z, w]$ are relatively prime but we are otherwise not imposing any conditions on their zero sets. If we impose restrictions, we can ask a more concrete question.

Suppose $q \in \mathbb{C}[z, w]$ is almost stable, $\deg q = (n, m)$, and suppose $p \in \mathbb{C}[z, w]$ has degree $\leq (n - 1, m - 1)$. If $p/q \in L^2(\mathbb{T}^2)$, then the sums of squares decomposition (as in Theorem 6.1) tells us that there is a constant c such that

$$|q(z, w)|^2 - |\tilde{q}(z, w)|^2 \geq c(1 - |z|^2)(1 - |w|^2)|p(z, w)|^2 \quad \text{for } (z, w) \in \mathbb{D}^2, \tag{13-1}$$

since p will be in \square_μ for the Bernstein–Szegő measure μ associated to q .

Question 13.2. Is the converse true? Does the estimate (13-1) imply $p/q \in L^2(\mathbb{T}^2)$?

Notational index and conventions

In this section we index where various notations and terms are defined in the paper. We also list our notational conventions.

stable/almost stable	Definition 1.4	horizontally stable	Definition 1.6
vertically stable	Definition 1.6	reflection $\tilde{q}(z, w)$	Definition 1.9
$\Lambda_n(z), \Lambda_m(w)$	Equation (1-3)	toral	Definition 1.12
atoral	Definition 1.13	divisor of ideal	Definition 5.1
distinguished variety	Definition 11.1	$d\sigma = d\sigma(z, w)$	Equation (4-1)
degree (n, m)	Definition 1.8	$\hat{q}(j, k)$	Equation (4-4)
$\square, \square, \square, \square, \square, \square$	Notation 4.3	$\langle f, g \rangle_\mu$	Equation (4-3)
$KV, K\square_\mu$, etc.	Notation 4.1	$w\square_\mu, z\square_\mu, \boxplus_\mu, \boxminus_\mu$, etc.	Notation 4.4
$\text{Max}_\mu, \text{Min}_\mu$	Equations (4-7) and (4-8)	\mathcal{I}_μ	Equation (4-2)
OC measure	Definition 6.2	\mathbb{T}^2 -symmetric	Definition 5.6
$L(z, w)$	Equation (7-1)	π_1, π_2	Equation (7-6)
Z_q	Equation (7-5)	$\mathbb{C}^N[z, \mathbb{C}^N[z, w], \mathbb{E}$	Notation 1.5

n, m	fixed positive integers (see Remark 4.2)
p, q	elements of $\mathbb{C}[z, w]$
E, F, G, A, B, Q	vector polynomials
E, F, A, B, Q	matrix polynomials in one variable
$\langle \cdot, \cdot \rangle$ with no subscript	inner product on \mathbb{C}^N (N determined from context)
$L^2(\mathbb{T}^2)$	L^2 on the torus with respect to Lebesgue measure
$L^2(\mu), L^2(\rho)$	L^2 on the torus with respect to the measure μ or ρ
$H^2(\mathbb{T}), H^2(\mathbb{T}^2)$	classical Hardy space on \mathbb{T} or \mathbb{T}^2
Φ, Ψ	one variable matrix valued inner functions

Acknowledgements

The author would like to sincerely thank John McCarthy and the anonymous referee for sharing their thoughts on this paper.

References

- [Agler et al. 2006] J. Agler, J. E. McCarthy, and M. Stankus, “Toral algebraic sets and function theory on polydisks”, *J. Geom. Anal.* **16**:4 (2006), 551–562. [MR 2007j:32002](#) [Zbl 1103.14019](#)
- [Agler et al. 2008] J. Agler, J. E. McCarthy, and M. Stankus, “Local geometry of zero sets of holomorphic functions near the torus”, *New York J. Math.* **14** (2008), 517–538. [MR 2010a:32014](#) [Zbl 1153.14002](#)
- [Cole and Wermer 1999] B. J. Cole and J. Wermer, “Ando’s theorem and sums of squares”, *Indiana Univ. Math. J.* **48**:3 (1999), 767–791. [MR 2000m:47014](#) [Zbl 0945.47010](#)
- [D’Angelo 1993] J. P. D’Angelo, *Several complex variables and the geometry of real hypersurfaces*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1993. [MR 94i:32022](#) [Zbl 0854.32001](#)
- [Dritschel 2004] M. A. Dritschel, “On factorization of trigonometric polynomials”, *Integral Equations Operator Theory* **49**:1 (2004), 11–42. [MR 2005d:47034](#) [Zbl 1070.47011](#)

- [Dritschel and Woerdeman 2005] M. A. Dritschel and H. J. Woerdeman, “Outer factorizations in one and several variables”, *Trans. Amer. Math. Soc.* **357**:11 (2005), 4661–4679. [MR 2006e:47039](#) [Zbl 1073.47024](#)
- [Duren 1970] P. L. Duren, *Theory of H^p spaces*, Pure and Applied Mathematics **38**, Academic Press, New York, 1970. [MR 42 #3552](#) [Zbl 0215.20203](#)
- [Geronimo and Lai 2006] J. S. Geronimo and M.-J. Lai, “Factorization of multivariate positive Laurent polynomials”, *J. Approx. Theory* **139**:1-2 (2006), 327–345. [MR 2007a:47023](#) [Zbl 1099.12001](#)
- [Geronimo and Woerdeman 2004] J. S. Geronimo and H. J. Woerdeman, “Positive extensions, Fejér–Riesz factorization and autoregressive filters in two variables”, *Ann. of Math. (2)* **160**:3 (2004), 839–906. [MR 2006b:42036](#)
- [Knese 2008] G. Knese, “Bernstein–Szegő measures on the two dimensional torus”, *Indiana Univ. Math. J.* **57**:3 (2008), 1353–1376. [MR 2009h:46054](#)
- [Knese 2009] G. Knese, “Polynomials defining distinguished varieties”, preprint, 2009. To appear in *Trans. Amer. Math. Soc.* [arXiv 0909.1818v1](#)
- [Rudin 1969] W. Rudin, *Function theory in polydiscs*, W. A. Benjamin, New York, 1969. [MR 41 #501](#) [Zbl 0177.34101](#)
- [Simon 2005] B. Simon, *Orthogonal polynomials on the unit circle. Part I*, American Mathematical Society Colloquium Publications **54**, American Mathematical Society, Providence, RI, 2005. [MR 2006a:42002a](#) [Zbl 1082.42020](#)
- [Simon 2008] B. Simon, “The Christoffel–Darboux kernel”, pp. 295–335 in *Perspectives in partial differential equations, harmonic analysis and applications*, Proc. Sympos. Pure Math. **79**, Amer. Math. Soc., Providence, RI, 2008. [MR 2010d:42045](#) [Zbl 1159.42020](#)

Received 23 Oct 2008. Revised 20 Oct 2009. Accepted 3 Dec 2009.

GREG KNESE: gknese@uci.edu

University of California, Irvine, Department of Mathematics, Irvine CA 92697-3875, United States

<http://www.math.uci.edu/~gknese>

LOCAL WELLPOSEDNESS FOR THE 2+1-DIMENSIONAL MONOPOLE EQUATION

MAGDALENA CZUBAK

The space-time monopole equation on \mathbb{R}^{2+1} can be derived by a dimensional reduction of the antiselfdual Yang–Mills equations on \mathbb{R}^{2+2} . It can be also viewed as the hyperbolic analog of Bogomolny equations. We uncover null forms in the nonlinearities and employ optimal bilinear estimates in the framework of wave–Sobolev spaces. As a result, we show the equation is locally wellposed in the Coulomb gauge for initial data sufficiently small in H^s for $s > \frac{1}{4}$.

1. Introduction

In this paper we study local wellposedness of the Cauchy problem for the monopole equation on \mathbb{R}^{2+1} Minkowski space in the Coulomb gauge. The space-time monopole equation can be derived by a dimensional reduction from the antiselfdual Yang–Mills equations on \mathbb{R}^{2+2} , and is given by

$$F_A = *D_A\phi, \tag{ME}$$

where F_A is the curvature of a one-form connection A on \mathbb{R}^{2+1} , $D_A\phi$ is a covariant derivative of the Higgs field ϕ , and $*$ is the Hodge star operator with respect to the Minkowski \mathbb{R}^{2+1} metric. (ME) is a hyperbolic analog of Bogomolny equations, and was first introduced by Ward [1989] and discussed from the point of view of twistors. Ward [1999] also studied its soliton solutions. Recently, Dai, Terng and Uhlenbeck [2006] gave a broad survey on the space-time monopole equation. In particular, using the inverse scattering transform they have shown global existence and uniqueness up to a gauge transformation for small initial data in $W^{2,1}$. However, L^2 based wellposedness theory for this equation has not been investigated. The objective of this paper is to fill this gap by specifically treating the Cauchy problem for rough initial data in H^s .

Written in coordinates, (ME) is a system of first order hyperbolic partial differential equations. The unknowns are a pair (A, ϕ) . If (A, ϕ) solve the equation, then so do

$$\lambda A(\lambda t, \lambda x) \quad \text{and} \quad \lambda\phi(\lambda t, \lambda x), \quad \text{for any } \lambda > 0.$$

This results in the critical exponent $s_c = 0$. Since in general one expects local wellposedness for $s > s_c$ the goal would be to show (ME) is wellposed for $s > 0$. Nevertheless, the two dimensions create an obstacle, which so far only allows $s > \frac{1}{4}$. We explain this now. In Section 4 we choose a Coulomb gauge, and

MSC2000: 35L70, 70S15.

Keywords: monopole, null form, Coulomb gauge, wellposedness.

reformulate (ME) as a system of semilinear wave equations coupled with an elliptic equation, to which we refer as auxiliary monopole equations (aME). Schematically it looks as follows:

$$\begin{aligned}\square u &= \mathfrak{B}_+(\partial u, \partial v, A_0), \\ \square v &= \mathfrak{B}_-(\partial u, \partial v, A_0), \\ \Delta A_0 &= \mathcal{C}(\partial u, \partial v, A_0),\end{aligned}\tag{aME}$$

where $\mathfrak{B}_\pm, \mathcal{C}$ are bilinear forms,¹ A_0 is the temporal part of the connection A , $\partial u, \partial v$ denote space-time derivatives of u and v respectively, and are given in terms of ϕ and the spatial part of A . As a result, showing wellposedness of (ME) for $s > 0$ can follow from showing (aME) is wellposed for $s > 1$ (see Theorem 4.1). Also, the most difficult nonlinearity that we have to handle is contained in $\mathfrak{B}_\pm(\partial u, \partial v, A_0)$. Luckily, it exhibits a structure of a null form. There are two standard null forms:

$$Q_0(u, v) = -\partial_t u \partial_t v + \nabla u \cdot \nabla v, \quad Q_{\alpha\beta}(u, v) = \partial_\alpha u \partial_\beta v - \partial_\beta u \partial_\alpha v.$$

The null condition was introduced by Klainerman [1984], and it was first applied to produce better local wellposedness results for wave equations with a null form by Klainerman and Machedon [1993]. Indeed, in low dimensions, for these kind of nonlinearities one can assume much less regularity of the initial data than for the general products. Counterexamples for general products were shown by Lindblad [1996]. We uncover the null form $Q_{\alpha\beta}$ in our system of wave equations as well as a new type of a null form which is related to $Q_{\alpha\beta}$. Unfortunately, the results in two spatial dimensions for $Q_{\alpha\beta}$ are not as optimal as they are in higher dimensions or as they are for Q_0 . In fact, the best result in literature so far for $Q_{\alpha\beta}$ in two dimensions is due to Zhou [1997]. He establishes local wellposedness for initial data in $H^s \times H^{s-1}$ for $s > \frac{5}{4}$. In addition, by examining the first iterate Zhou shows that this is as close as one can get to the critical level using iteration methods.² On the other hand, for dimensions $n \geq 3$ Klainerman and Machedon [1996] showed almost optimal local wellposedness in $H^s \times H^{s-1}$ for $s > n/2$. The articles [Klainerman and Machedon 1995; Klainerman and Selberg 2002] give equally satisfying results for Q_0 , and in all dimensions $n \geq 2$.

Now, one of the nonlinearities in the system (aME) is $Q_{\alpha\beta}$, so showing (aME) is locally wellposed for $s > \frac{5}{4}$ would be sharp by iteration methods. This is what we do, and as a result we obtain local wellposedness of (ME) in the Coulomb gauge for $s > \frac{1}{4}$ (see the Main Theorem below). However, (aME) is not exactly (ME), so we hope to treat (ME) directly in the near future and improve the results. What should be mentioned here is that we have considered other traditional gauges such as Lorentz and Temporal, but they have not been as nearly useful as the Coulomb gauge. Perhaps other, less traditional gauges could be used. Moreover, we note that even the estimates involving the temporal variable A_0 seem to require $s > \frac{1}{4}$.

We include a brief discussion about global wellposedness. As already mentioned, in [Dai et al. 2006] the inverse scattering transform is used to show global existence and uniqueness up to a gauge transformation for small data in $W^{2,1}$. To extend it to global wellposedness in L^2 based theory, we would like to benefit from the local result in this article. It is not immediately clear how this can be accomplished

¹ See Section 4 for the precise formula for \mathfrak{B}_\pm and \mathcal{C} .

² The discussion of the first iterate can be also found in the appendix of [Klainerman and Selberg 2002], and it can be deduced from the estimates and counterexamples found within [Foschi and Klainerman 2000].

since, for example, the energy functional is not positive definite and, in fact, it vanishes for the solutions of (ME) [Ward 1989]. Global wellposedness is an interesting question, and we would like to investigate it in the future.

The main result of this paper is contained in the following theorem.

Main Theorem. *Let $\frac{1}{4} < s < \frac{1}{2}$ and $r \in (0, 2s]$ and consider the space-time monopole equation*

$$F_A = *D_A\phi, \quad (\text{ME})$$

with initial data

$$(A_1, A_2, \phi)|_{t=0} = (a_1, a_2, \phi_0),$$

then (ME) in the Coulomb gauge is locally wellposed for initial data sufficiently small in $H^s(\mathbb{R}^2)$ in the following sense.

- **Local existence:** For all $a_1, a_2, \phi_0 \in H^s(\mathbb{R}^2)$ sufficiently small there exist $T > 0$ depending continuously on the norm of the initial data, and functions

$$A_0 \in C_b([0, T], \dot{H}^r), \quad A_1, A_2, \phi \in C_b([0, T], H^s),$$

which solve (ME) in the Coulomb gauge on $[0, T] \times \mathbb{R}^2$ in the sense of distributions and such that the initial conditions are satisfied.

- **Uniqueness:** If $T > 0$ and (A, ϕ) and (A', ϕ') are two solutions of (ME) in the Coulomb gauge on $(0, T) \times \mathbb{R}^2$ belonging to

$$C_b([0, T], \dot{H}^r) \times (H_T^{s,\theta})^3$$

with the same initial data, then $(A, \phi) = (A', \phi')$ on $(0, T) \times \mathbb{R}^2$.

- **Continuous dependence on the initial data:** For any $a_1, a_2, \phi_0 \in H^s(\mathbb{R}^2)$ there is a neighborhood U of a_1, a_2, ϕ_0 in $(H^s(\mathbb{R}^2))^3$ such that the solution map $(a, \phi_0) \rightarrow (A, \phi)$ is continuous from U into $C_b([0, T], \dot{H}^r) \times (C_b([0, T], H^s))^3$.

Remark 1.1. The spaces $H_T^{s,\theta}$ are defined in Section 2B.

Remark 1.2. The initial data does not have to be given in the Coulomb gauge. See Theorem 3.3.

Remark 1.3. There are two reasons for the requirement of the small initial data. First, the construction of the global Coulomb gauge requires an assumption on the size of the data (see Section 3B). The second obstacle comes from the elliptic equation for A_0 in (aME), and including A_0 in the Picard iteration. See Remark 4.2 for further discussion.

Remark 1.4. We do not prescribe initial data for A_0 , because when A is in the Coulomb gauge, $A_0(t)$ can be determined at any time by solving the elliptic equation. See Section 4 for more details.

Remark 1.5. To simplify the exposition, in this paper we assume $\frac{1}{4} < s < \frac{1}{2}$. See [Czubak 2008] for all $s > \frac{1}{4}$. In general, the higher the value of s , the less delicate the estimates have to be. We have a uniform way to handle all $s > \frac{1}{4}$ for the estimates involving the null forms (see Section 5B1 for a discussion). Therefore the reason for restricting the range of s is rather due to the technicalities of the estimates for A_0 (Theorems 5.3 and 5.5 and Corollaries 5.4 and 5.6) and the regularity of the gauge transformations (Lemma 3.1). The technicalities are not very interesting and are handled in [Czubak 2008] with similar arguments as those presented here.

The outline of the paper is as follows. [Section 2](#) sets notation, introduces spaces, and estimates used throughout the paper. In [Section 3](#) we take a closer look at the equations and discuss gauge transformations. In [Section 4](#) we rewrite (ME) as a system of wave equations coupled with an elliptic equation. We also show local wellposedness of the new system implies local wellposedness of (ME) in the Coulomb gauge. [Section 5](#) is devoted to the proof of the [Main Theorem](#), which is reduced to establishing estimates (5-4)–(5-8).

2. Preliminaries

First we establish notation, then we introduce function spaces as well as estimates used.

2A. Notation. The expression $a \lesssim b$ means $a \leq Cb$ for some positive constant C . A point in the 2+1-dimensional Minkowski space is written as $(t, x) = (x^\alpha)_{0 \leq \alpha \leq 2}$. Greek indices range from 0 to 2, and Roman indices range from 1 to 2. We raise and lower indices with the Minkowski metric $\text{diag}(-1, 1, 1)$. We write $\partial_\alpha = \partial_{x^\alpha}$ and $\partial_t = \partial_0$, and we also use the Einstein notation. Therefore, $\partial^i \partial_i = \Delta$, and $\partial^\alpha \partial_\alpha = -\partial_t^2 + \Delta = \square$. When we refer to spatial and time derivatives of a function f , we write ∂f , and when we consider only spatial derivatives of f , we write ∇f . Finally, d denotes the exterior differentiation operator and d^* its dual given by $d^* = (-1)^k ***d*$, where $*$ is the Hodge $*$ operator (see, for example, [Roe 1998]) and k comes from d^* acting on some given k -form. It will be clear from the context, when $*$ and d^* operators act with respect to the Minkowski metric and when with respect to the Euclidean metric. For the convenience of the reader we include the following: with respect to the Euclidean metric on \mathbb{R}^2 we have

$$*dx = dy, \quad *dy = -dx, \quad *1 = dx \wedge dy,$$

and with respect to the $\text{diag}(-1, 1, 1)$ metric on \mathbb{R}^{2+1} ,

$$*dt = dx \wedge dy, \quad *dx = dt \wedge dy, \quad *dy = -dt \wedge dx.$$

2B. Function spaces. We use Picard iteration. Here we introduce the spaces, in which we are going to perform the iteration³. First we define the following Fourier multiplier operators

$$\begin{aligned} \widehat{\Lambda^\alpha f}(\xi) &= (1 + |\xi|^2)^{\alpha/2} \widehat{f}(\xi), & \widehat{\Lambda_+^\alpha u}(\tau, \xi) &= (1 + \tau^2 + |\xi|^2)^{\alpha/2} \widehat{u}(\tau, \xi), \\ \widehat{\Lambda_-^\alpha u}(\tau, \xi) &= \left(1 + \frac{(\tau^2 - |\xi|^2)^2}{1 + \tau^2 + |\xi|^2}\right)^{\alpha/2} \widehat{u}(\tau, \xi), \end{aligned}$$

where the symbol of Λ_-^α is comparable to $(1 + ||\tau| - |\xi||)^\alpha$. The corresponding homogeneous operators are denoted by D^α , D_+^α , D_-^α , respectively.

Now, the spaces of interest are the wave-Sobolev spaces, $H^{s,\theta}$ and $\mathcal{H}^{s,\theta}$, given by⁴

$$\|u\|_{H^{s,\theta}} = \|\Lambda^s \Lambda_-^\theta u\|_{L^2(\mathbb{R}^{2+1})}, \quad \|u\|_{\mathcal{H}^{s,\theta}} = \|u\|_{H^{s,\theta}} + \|\partial_t u\|_{H^{s-1,\theta}}.$$

³We are also going to employ a combination of the standard $L_t^p W_x^{s,q}$ spaces for A_0 . See [Section 5C](#).

⁴These spaces, together with results in [Selberg 2002b], allowed Klainerman and Selberg to present a unified approach to local wellposedness for wave maps, Yang–Mills and Maxwell–Klein–Gordon types of equations in [Klainerman and Selberg 2002], and are now the natural choice for low regularity subcritical local wellposedness for wave equations. See also [Tao 2006].

An equivalent norm for $\mathcal{H}^{s,\theta}$ is $\|u\|_{\mathcal{H}^{s,\theta}} = \|\Lambda^{s-1}\Lambda_+\Lambda_-u\|_{L^2(\mathbb{R}^{2+1})}$. By results in [Selberg 1999] if $\theta > 1/2$, we have

$$H^{s,\theta} \hookrightarrow C_b(\mathbb{R}, H^s), \quad (2-1)$$

$$\mathcal{H}^{s,\theta} \hookrightarrow C_b(\mathbb{R}, H^s) \cap C_b^1(\mathbb{R}, H^{s-1}). \quad (2-2)$$

This is a crucial fact needed to localize our solutions in time. We denote the restrictions to the time interval $[0, T]$ by

$$H_T^{s,\theta} \quad \text{and} \quad \mathcal{H}_T^{s,\theta},$$

respectively.

2C. Estimates used. Throughout the paper we use the following estimates:

$$\|D^{-\sigma}(uv)\|_{L_t^p L_x^q} \lesssim \|u\|_{H^{s,\theta}} \|v\|_{H^{s,\theta}} \quad (2-3)$$

is a theorem established by Klainerman and Tataru [1999] for the space-time operator D_+ . The proof for the spatial operator D is in [Selberg 1999]. There are several conditions that σ, p, q have to satisfy, and they are listed in Section 5D, where we discuss the application of the estimate. Further,

$$\|u\|_{L_t^p L_x^2} \lesssim \|u\|_{H^{0,\theta}} \leq \infty \quad \text{if } 2 \leq p \leq \infty, \theta > \frac{1}{2}, \quad (2-4)$$

$$\|u\|_{L_t^p L_x^q} \lesssim \|u\|_{H^{1-2/q-1/p,\theta}} \quad \text{if } 2 \leq p \leq \infty, 2 \leq q < \infty, 2/p \leq \frac{1}{2} - 1/q, \theta > \frac{1}{2}, \quad (2-5)$$

$$\|uv\|_{L_{t,x}^2} \lesssim \|u\|_{H^{a,\alpha}} \|v\|_{H^{b,\beta}} \quad \text{if } a, b, \alpha, \beta \geq 0, a+b > 1, \alpha+\beta > \frac{1}{2}. \quad (2-6)$$

Estimate (2-4) can be proved by interpolation between $H^{0,\theta} \hookrightarrow L_{t,x}^2$ and (2-1) with $s=0$. Estimate (2-5) is a two-dimensional case of Theorem D in [Klainerman and Selberg 2002]. Finally, (2-6) is a special case of the proposition in [Klainerman and Selberg 2002, Appendix A.2].

3. A closer look at the monopole equations

3A. Derivation and background. Electric charge is quantized, which means that it appears in integer multiples of an electron. This is called the principle of quantization and has been observed in nature. The only theoretical proof so far was presented by Paul Dirac [1931]. In the proof Dirac introduced the concept of a magnetic monopole, of an isolated point-source of a magnetic charge. Despite extensive research, magnetic monopoles have not been (yet) found in nature. We refer to magnetic monopoles as Euclidean monopoles. The Euclidean monopole equation has exactly the same form as our space-time monopole equation (ME),

$$F_A = *D_A\phi,$$

with the exception that $*$ acts here with respect to the Euclidean metric and the base manifold is \mathbb{R}^3 instead of \mathbb{R}^{2+1} . The Euclidean monopole equations are also referred to as Bogomolny equations. For more on Euclidean monopoles we refer the reader to [Jaffe and Taubes 1980] and [Atiyah and Hitchin 1988]. In this paper we study the space-time monopole equation, which was first introduced by Ward [1989]. Both the Euclidean and the space-time monopole equations are examples of integrable systems

and have an equivalent formulation as a Lax pair. This and much more can be found in [Dai et al. 2006].

Given a space-time monopole equation

$$F_A = *D_A\phi, \quad (\text{ME})$$

the unknowns are a pair (A, ϕ) . A is a connection 1-form given by

$$A = A_0 dt + A_1 dx + A_2 dy, \quad \text{where } A_\alpha : \mathbb{R}^{2+1} \rightarrow \mathfrak{g}. \quad (3-1)$$

Here \mathfrak{g} is the Lie algebra of a Lie group G , which is typically taken to be a matrix group $SU(n)$ or $U(n)$. In this paper we consider $G = SU(n)$, but everything we say here should generalize to any compact Lie group.

To be more general we could say A is a connection on a principal G -bundle. Then observe that the G -bundle we deal here with is a trivial bundle $\mathbb{R}^{2+1} \times G$.

Next, ϕ is a section of a vector bundle associated to the G -bundle by a representation. We use the adjoint representation. Since we have a trivial bundle, we can just think of the Higgs field ϕ as a map from $\mathbb{R}^{2,1} \rightarrow \mathfrak{g}$.

F_A is the curvature of A . It is a Lie algebra valued 2-form on \mathbb{R}^{2+1}

$$F_A = \frac{1}{2} F_{\alpha\beta} dx^\alpha \wedge dx^\beta, \quad \text{where } F_{\alpha\beta} = \partial_\alpha A_\beta - \partial_\beta A_\alpha + [A_\alpha, A_\beta], \quad (3-2)$$

where $[\cdot, \cdot]$ denotes the Lie bracket, which for matrices can be thought of simply as $[X, Y] = XY - YX$.

When we write $[\phi, B]$, where B is a 1-form, we mean

$$[\phi, B] = [\phi, B_i] dx^i \quad \text{and} \quad [B, C] = \frac{1}{2} [B_i, C_j] dx^i \wedge dx^j, \quad \text{for two 1-forms } B, C. \quad (3-3)$$

In the physics language, frequently adopted by the mathematicians, A is called a gauge potential, ϕ a scalar field and F_A is called an electromagnetic field.

Next, D_A is the covariant exterior derivative associated to A , and $D_A\phi$ is given by

$$D_A\phi = D_\alpha\phi dx^\alpha, \quad \text{where } D_\alpha\phi = \partial_\alpha\phi + [A_\alpha, \phi]. \quad (3-4)$$

The space-time monopole equation (ME) is obtained by a dimensional reduction of the antiselfdual Yang–Mills equations on \mathbb{R}^{2+2} , given by

$$F_A = -*F_A. \quad (\text{ASDYM})$$

If the curvature of a connection A satisfies (ASDYM), then A is called an antiselfdual connection. (The corresponding selfdual Yang–Mills equation is

$$F_A = *F_A; \quad (\text{SDYM})$$

in either case F_A satisfies the Yang–Mills equation $D_A *F = 0$, since $F_A = \pm *F_A$ implies

$$D_A *F = \pm D_A F = 0,$$

as can be seen from the second Bianchi identity, or by direct computation from (3-2).)

Both (ASDYM) and (SDYM) are known to give rise to many different integrable equations (see [Ward 1985; Ablowitz et al. 2003] and references therein). In particular, the 2+2 signature is used to derive both KDV and NLS in [Mason and Sparling 1989], and harmonic maps from \mathbb{R}^2 and \mathbb{R}^{1+1} into a Lie

group in [Uhlenbeck 1992]. Also see [Ward 1989], where Einstein's vacuum equation for cylindrically symmetric space-times, and the sine-Gordon equation are derived from (ME).

We now present the details of the derivation of (ME) from (ASDYM) outlined in [Dai et al. 2006]. Let

$$dx_1^2 + dx_2^2 - dx_3^2 - dx_4^2$$

be a metric on \mathbb{R}^{2+2} , then in coordinates (ASDYM) is

$$F_{12} = -F_{34}, \quad F_{13} = -F_{24}, \quad F_{23} = F_{14}. \quad (3-5)$$

The next step is the dimensional reduction, where we assume the connection A is independent of x_3 , and we let $A_3 = \phi$. Then (3-5) becomes

$$D_0\phi = F_{12}, \quad D_1\phi = F_{02}, \quad D_2\phi = F_{10}, \quad (3-6)$$

where we use index 0 instead of 4. This is exactly (ME) written out in components.

Remark 3.1. Equivalently we could write (ME) as

$$F_{\alpha\beta} = -\varepsilon_{\alpha\beta\gamma} D^\gamma \phi, \quad (3-7)$$

where $\varepsilon_{\alpha\beta\gamma}$ is a completely antisymmetric tensor with $\varepsilon_{012} = 1$, and where we raise the index γ using the Minkowski metric. We choose to work with the Hodge operator $*$ as it simplifies our task in Section 4.

Following [Dai et al. 2006], there is another way to write (ME), which is very useful for computations. (ME) is an equation involving 2-forms on both sides. By taking the parts corresponding to $dt \wedge dx$ and $dt \wedge dy$ on the one hand, and the parts corresponding to $dx \wedge dy$ on the other, we obtain, respectively,

$$\partial_t A + [A_0, A] - dA_0 = *d\phi + [*A, \phi], \quad (3-8)$$

$$dA + [A, A] = *(\partial_t \phi + [A_0, \phi]). \quad (3-9)$$

Observe that now operators d and $*$ act only with respect to the spatial variables. Similarly, A now denotes only the spatial part of the connection, i.e., $A = (A_1, A_2)$. Moreover, (3-8) is an equation involving 1-forms, and (3-9) involves 2-forms.

3B. Gauge transformations. (ME) is invariant under gauge transformations. Indeed, if we have a smooth map g , with compact support such that $g : \mathbb{R}^{2+1} \rightarrow G$, and

$$A \rightarrow A_g = gAg^{-1} + gdg^{-1}, \quad \phi \rightarrow \phi_g = g\phi g^{-1}, \quad (3-10)$$

then a computation shows $F_A \rightarrow gF_A g^{-1}$ and $D_A \phi \rightarrow gD_A \phi g^{-1}$. Therefore if a pair (A, ϕ) solves (ME), so does (A_g, ϕ_g) .

We would like to discuss the regularity of the gauge transformations. If $A \in X$, $\phi \in Y$ where X, Y are some Banach spaces, the smoothness and compact support assumption on g can be lowered just enough so the gauge transformation defined above is a continuous map from X back into X , and from Y back into Y . First note that since we are mapping into a compact Lie group, we can assume $g \in L_{t,x}^\infty$ and $\|g\|_{L_{t,x}^\infty} = \|g^{-1}\|_{L_{t,x}^\infty}$. Next, note that the Main Theorem produces a solution so that ϕ and the spatial parts of the connection $A_1, A_2 \in C_b(I, H^s)$, $\frac{1}{4} < s < \frac{1}{2}$, and $A_0 \in C_b(I, \dot{H}^r)$, $r \in (0, 2s]$.

Lemma 3.1. *Let $0 < \alpha < 1$, and $Y = C_b(I, \dot{H}^1 \cap \dot{H}^{\alpha+1}) \cap L^\infty$, then the gauge action is a continuous map from*

$$C_b(I, H^\alpha) \times Y \rightarrow C_b(I, H^\alpha), \quad (h, g) \mapsto ghg^{-1} + gdg^{-1}, \quad (3-11)$$

and the following estimate holds:

$$\|h_g\|_{C_b(I, H^\alpha)} \lesssim (\|h\|_{C_b(I, H^\alpha)} + 1)\|g\|_Y^2. \quad (3-12)$$

Proof. The continuity of the map easily follows from the inequalities we obtain below. Next, for fixed t we have

$$\|g(t)h(t)g^{-1}(t) + g(t)dg^{-1}(t)\|_{H^\alpha} \lesssim \|ghg^{-1}\|_{L^2} + \|D^\alpha(ghg^{-1})\|_{L^2} + \|gdg^{-1}\|_{H^\alpha},$$

where for the ease of notation we eliminated writing of the variable t on the right side of the inequality. The first term is bounded by $\|h(t)\|_{H^\alpha}\|g\|_{L^\infty}^2$. For the second one we have

$$\|D^\alpha(ghg^{-1})\|_{L^2} \lesssim \|D^\alpha gh\|_{L^2}\|g\|_{L^\infty} + \|hD^\alpha g^{-1}\|_{L^2}\|g\|_{L^\infty} + \|h\|_{\dot{H}^\alpha}\|g\|_{L^\infty}^2.$$

It is enough to only look at the first term since g and g^{-1} have the same regularity. By Hölder's inequality and Sobolev embedding

$$\|D^\alpha gh\|_{L^2} \leq \|D^\alpha g\|_{L^{2/\alpha}}\|h\|_{L^{(1/2-\alpha/2)^{-1}}} \lesssim \|g\|_{\dot{H}^1}\|h\|_{\dot{H}^\alpha}, \quad (3-13)$$

where we use that $\frac{\alpha}{2} = \frac{1}{2} - \frac{1-\alpha}{2}$. Finally for the last term we have

$$\|gdg^{-1}\|_{H^\alpha} \lesssim \|g\|_{\dot{H}^1}\|g\|_{L^\infty} + \|D^\alpha gdg^{-1}\|_{L^2} + \|g\|_{\dot{H}^{\alpha+1}}\|g\|_{L^\infty}, \quad (3-14)$$

and we are done if we observe that the second term can be handled exactly as in (3-13). \square

Remark 3.2. We assume $0 < \alpha < 1$ since this is the case we need. However it is not difficult to see the lemma still holds with $\alpha = 0$ or $\alpha \geq 1$ [Czubak 2008].

From the lemma, we trivially obtain the following corollary.

Corollary 3.2. *Let $0 < r, s < 1$,*

$$X = C_b(I, \dot{H}^r) \times C_b(I, H^s) \times C_b(I, H^s), \quad \text{and} \quad Y = C_b(I, \dot{H}^1 \cap \dot{H}^{s+1} \cap \dot{H}^{r+1}) \cap L^\infty.$$

Then the gauge action is a continuous map from

$$X \times Y \rightarrow X, \quad (A_0, A_1, A_2) \mapsto A_g, \quad (3-15)$$

as well as from

$$C_b(I, H^s) \times Y \rightarrow C_b(I, H^s), \quad \phi \mapsto \phi_g = g\phi g^{-1}, \quad (3-16)$$

and the following estimates hold:

$$\begin{aligned} \|A_g\|_X &\lesssim (1 + \|A\|_X)\|g\|_Y^2, \\ \|\phi_g\|_{C_b(I, H^s)} &\lesssim \|\phi\|_{C_b(I, H^s)}\|g\|_Y^2. \end{aligned} \quad (3-17)$$

Since in this paper we work in the Coulomb gauge, we ask: given any initial data $a_1, a_2, \phi_0 \in H^s(\mathbb{R}^2)$, can we find a gauge transformation so that the initial data is placed in the Coulomb gauge? Dell'Antonio and Zwanziger [1991] produce a global \dot{H}^1 Coulomb gauge using variational methods. Here, we also require $g \in \dot{H}^{s+1}$, and two dimensions are tricky. Fortunately, if the initial data is small, we can obtain a global gauge with the additional regularity as needed. This has been studied by the author and Uhlenbeck for two dimensions and higher; the result in two dimensions is the following:

Theorem 3.3 [Czubak and Uhlenbeck \geq 2010]. *Let $s > 0$. Given $A(0) = a$ sufficiently small in*

$$H^s(\mathbb{R}^2) \times H^s(\mathbb{R}^2),$$

there exists a gauge transformation $g \in \dot{H}^{s+1}(\mathbb{R}^2) \cap \dot{H}^1(\mathbb{R}^2) \cap L^\infty$ such that $\partial^i(ga_i g^{-1} + g\partial_i g^{-1}) = 0$.

4. The monopole equation in the Coulomb gauge as a system of wave and elliptic equations

We begin by rewriting the monopole equation in the Coulomb gauge as a system of wave equations coupled with an elliptic equation. We refer to the new system as the auxiliary monopole equations (aME). Then we establish that local wellposedness (LWP) for (ME) in the Coulomb gauge can be obtained from LWP of (aME).

4A. Derivation of (aME) from (ME). Suppose we have initial data

$$A_i|_{t=0} = a_i \quad \text{for } i = 1, 2 \quad \text{and} \quad \phi|_{t=0} = \phi_0, \quad (4-1)$$

where $\partial^i a_i = 0$. Recall equations (3-8) and (3-9):

$$\partial_t A + [A_0, A] - dA_0 = *d\phi + [*A, \phi], \quad (4-2)$$

$$dA + [A, A] = *(\partial_t \phi + [A_0, \phi]), \quad (4-3)$$

where d and $*$ act only with respect to the spatial variables, and A denotes only the spatial part of the connection. If we impose the Coulomb gauge condition, then

$$d^*A = 0. \quad (4-4)$$

By equivalence of closed and exact forms on \mathbb{R}^n , we can further suppose that

$$A = *df, \quad (4-5)$$

for some $f : \mathbb{R}^{2+1} \rightarrow \mathfrak{g}$. Observe that

$$d * df = \Delta f dx \wedge dy, \quad [*df, *df] = [df, df] = 1/2[\partial_i f, \partial_j f] dx^i \wedge dx^j, \quad (4-6)$$

and $**\omega = -\omega$ for a one-form on \mathbb{R}^2 . Hence (4-2) and (4-3) become

$$\partial_t * df + [A_0, *df] - dA_0 = *d\phi - [df, \phi], \quad (4-7)$$

$$\Delta f + [\partial_1 f, \partial_2 f] = \partial_t \phi + [A_0, \phi]. \quad (4-8)$$

Take d^* of (4-7) to obtain

$$\Delta A_0 = -d^*[A_0, *df] - d^*[df, \phi].$$

This is the elliptic equation in (aME). To uncover the wave equations we proceed as follows. Take d of (4-7)

$$\partial_t \Delta f + \partial^j [A_0, \partial_j f] = \Delta \phi + \partial_2 [\partial_1 f, \phi] - \partial_1 [\partial_2 f, \phi]. \quad (4-9)$$

Consider (4-9) and (4-8) on the spatial Fourier transform side:

$$-\partial_t |\xi|^2 \hat{f} + |\xi|^2 \hat{\phi} = i(\xi_2 [\widehat{\partial_1 f, \phi}] - \xi_1 [\widehat{\partial_2 f, \phi}] - \xi_j [\widehat{A_0, \partial_j f}]), \quad (4-10)$$

$$-|\xi|^2 \hat{f} - \partial_t \hat{\phi} = -[\widehat{\partial_1 f, \partial_2 f}] + [\widehat{A_0, \phi}]. \quad (4-11)$$

This allows us to write (4-10) and (4-11) as a system for ϕ and df :

$$(\partial_t - i|\xi|)(\hat{\phi} + i|\xi|\hat{f}) = -\hat{\mathcal{B}}_+(\phi, df, A_0), \quad (4-12)$$

$$(\partial_t + i|\xi|)(\hat{\phi} - i|\xi|\hat{f}) = -\hat{\mathcal{B}}_-(\phi, df, A_0), \quad (4-13)$$

where

$$\hat{\mathcal{B}}_{\pm} = -[\widehat{\partial_1 f, \partial_2 f}] + [\widehat{A_0, \phi}] \pm \left(\frac{\xi_1}{|\xi|} [\widehat{\partial_2 f, \phi}] - \frac{\xi_2}{|\xi|} [\widehat{\partial_1 f, \phi}] + \frac{\xi_j}{|\xi|} [\widehat{A_0, \partial_j f}] \right). \quad (4-14)$$

Indeed, multiply (4-10) by $i/|\xi|$, and first add the resulting equation to (4-11) to obtain (4-12), and then subtract it from (4-11) to obtain (4-13). Now we let

$$\hat{\phi} + i|\xi|\hat{f} = (\partial_t + i|\xi|)\hat{u} \quad \text{and} \quad \hat{\phi} - i|\xi|\hat{f} = (\partial_t - i|\xi|)\hat{v}, \quad (4-15)$$

where $u, v : \mathbb{R}^{2+1} \rightarrow \mathfrak{g}$. This gives

$$\square u = \mathcal{B}_+(\phi, df, A_0), \quad \square v = \mathcal{B}_-(\phi, df, A_0).$$

See Remark 4.1 below.

Now we discuss initial data. From (4-15), we have

$$\partial_t \widehat{u}(0) = \hat{\phi}_0 + i|\xi|\widehat{f}(0) - i|\xi|\widehat{u}(0), \quad (4-16)$$

$$\partial_t \widehat{v}(0) = \hat{\phi}_0 - i|\xi|\widehat{f}(0) + i|\xi|\widehat{v}(0). \quad (4-17)$$

We are free to choose any data for u and v , as long as in the end we can recover the original data for ϕ and A . Hence we just let $u(0) = v(0) = 0$. We still need to say what $|\xi|\widehat{f}(0)$ is in terms of the initial data (a, ϕ_0) . Let $\hat{h} = |\xi|\widehat{f}(0)$. By (4-1) and (4-5) we have $a_1 = A_1(0) = -\partial_2 f(0)$ and $a_2 = A_2(0) = \partial_1 f(0)$. Therefore

$$R_1 h = a_2, \quad R_2 h = -a_1,$$

where R_j denotes the Riesz transform, $(-\Delta)^{-1/2} \partial_j$. Differentiate the first equation with respect to x , the second with respect to y , and add them together to obtain

$$\Delta D^{-1} h = \partial_1 a_2 - \partial_2 a_1.$$

So

$$h = R_2 a_1 - R_1 a_2, \quad (4-18)$$

It follows that the initial data for u and v are

$$u(0) = v(0) = 0, \quad \partial_t u(0) = \phi_0 + ih, \quad \partial_t v(0) = \phi_0 - ih, \quad (4-19)$$

with h defined by (4-18).

In summary, the monopole equation in the Coulomb gauge

$$F_A = *D_A\phi, \quad d^*A = 0,$$

with initial data (4-1) can be rewritten as the system

$$\begin{aligned} \square u &= \mathcal{B}_+(\phi, \nabla f, A_0), \\ \square v &= \mathcal{B}_-(\phi, \nabla f, A_0), \\ \Delta A_0 &= \mathcal{C}(\phi, \nabla f, A_0), \end{aligned} \tag{aME}$$

where

$$\mathcal{C} = -\partial_1[A_0, \partial_2 f] + \partial_2[A_0, \partial_1 f] + \partial_j[\partial_j f, \phi], \tag{4-20}$$

$$\mathcal{B}_\pm = -\mathcal{B}_1 \mp i\mathcal{B}_2 + \mathcal{B}_3 \mp i\mathcal{B}_4, \tag{4-21}$$

and

$$\mathcal{B}_1 = [\partial_1 f, \partial_2 f], \quad \mathcal{B}_2 = R_1[\partial_2 f, \phi] - R_2[\partial_1 f, \phi], \quad \mathcal{B}_3 = [A_0, \phi], \quad \mathcal{B}_4 = R_j[A_0, \partial_j f]. \tag{4-22}$$

The initial data for (aME) is given by (4-19).

Remark 4.1. u and v are our new unknowns, but we are really interested in ϕ and df . Therefore, we observe that once we know what u and v are, we can determine ϕ and df by using

$$\hat{\phi} = \frac{(\partial_t + i|\xi|)\hat{u} + (\partial_t - i|\xi|)\hat{v}}{2}, \quad i|\xi|\hat{f} = \frac{(\partial_t + i|\xi|)\hat{u} - (\partial_t - i|\xi|)\hat{v}}{2}, \tag{4-23}$$

or equivalently

$$\phi = \frac{(\partial_t + iD)u + (\partial_t - iD)v}{2}, \quad \partial_j f = -iR_j \left(\frac{(\partial_t + iD)u - (\partial_t - iD)v}{2} \right). \tag{4-24}$$

From df we get A by letting $A = *df$. Finally, for simplicity we usually keep the nonlinearities in terms of ϕ and df . However, since ϕ and df can be written in terms of derivatives of u and v we sometimes write $\mathcal{B}_\pm(\phi, df, A_0)$ as $\mathcal{B}_\pm(\partial u, \partial v, A_0)$.

Remark 4.2. (aME) has some resemblance to a system considered by Selberg [2002a] for the Maxwell–Klein–Gordon (MKG) equations, where he successfully obtains almost optimal local wellposedness in dimensions $1 + 4$. Besides the dimension considered, there are two fundamental technical differences applicable to our problem. First comes from the fact that the monopole equation we consider here is an example of a system in the nonabelian gauge theory whereas MKG is an example of a system in the abelian gauge theory. The existence of a global Coulomb gauge requires smallness of initial data in the nonabelian gauge theories, but is not needed in the abelian theories. Another technical difference arises from Selberg being able to solve the elliptic equation for his temporal variable A_0 using the Riesz representation theorem, where he does not require smallness of the initial data. The elliptic equation in (aME) is more difficult, so we include A_0 in the Picard iteration. As a result we are not able to allow large data by taking a small time interval, which we could do if we only had the two wave equations. Finally, we point out that the proof of our estimates involving A_0 is modeled after Selberg’s proof in [Selberg 2002a] (see Remark 5.1 and Section 5C).

4B. Return to the monopole equation. Now we have a theorem, where we show how LWP for (aME) implies LWP for (ME) in the Coulomb gauge. For completeness, we first state exactly what we mean by LWP of (aME).

Let $r \in (0, \min(2s, 1 + s)]$, $s > 0$. Consider the system (aME) with initial data

$$(u, u_t)|_{t=0} = (u_0, u_1) \quad \text{and} \quad (v, v_t)|_{t=0} = (v_0, v_1)$$

in $H^{s+1} \times H^s$, then (aME) is LWP if the following conditions are satisfied:

Local existence. There exist $T > 0$ depending continuously on the norm of the initial data, and functions

$$A_0 \in C_b([0, T], \dot{H}^r), \quad u, v \in \mathcal{H}_T^{s+1, \theta} \hookrightarrow C_b([0, T], H^{s+1}) \cap C_b^1([0, T], H^s),$$

which solve (aME) on $[0, T] \times \mathbb{R}^2$ in the sense of distributions and such that the initial conditions are satisfied.

Uniqueness. If $T > 0$ and (A_0, u, v) and (A'_0, u', v') are two solutions of (aME) on $(0, T) \times \mathbb{R}^2$ belonging to

$$C_b([0, T], \dot{H}^r) \times \mathcal{H}_T^{s+1, \theta} \times \mathcal{H}_T^{s+1, \theta},$$

with the same initial data, then $(A_0, u, v) = (A'_0, u', v')$ on $(0, T) \times \mathbb{R}^2$.

Continuous dependence on initial data. For any $(u_0, u_1), (v_0, v_1) \in H^{s+1} \times H^s$ there is a neighborhood U of the initial data such that the solution map $(u_0, u_1), (v_0, v_1) \rightarrow (A_0, u, v)$ is continuous from U into $C_b([0, T], \dot{H}^r) \times (C_b([0, T], H^{s+1}) \cap C_b^1([0, T], H^s))^2$.

In fact, by the results in [Selberg 2002b] combined with estimates for the elliptic equation, we can show the stronger estimates

$$\begin{aligned} \|u - u'\|_{\mathcal{H}_T^{s+1, \theta}} + \|v - v'\|_{\mathcal{H}_T^{s+1, \theta}} + \|A_0 - A'_0\|_{C_b([0, T], \dot{H}^r)} \\ \lesssim \|u_0 - u'_0\|_{H^{s+1}} + \|u_1 - u'_1\|_{H^s} + \|v_0 - v'_0\|_{H^{s+1}} + \|v_1 - v'_1\|_{H^s}, \end{aligned} \quad (4-25)$$

where $(u'_0, u'_1), (v'_0, v'_1)$ are sufficiently close to $(u_0, u_1), (v_0, v_1)$.

Remark 4.3. Note that below we have no restriction on s , that is, if we could show (aME) is LWP in $H^{s+1} \times H^s$, $s > 0$, we would get LWP of (ME) in the Coulomb gauge in H^s for $s > 0$ as well.

Theorem 4.1. Consider (ME) in the Coulomb gauge with the following initial data in H^s for $s > 0$:

$$A_i|_{t=0} = a_i, \quad i = 1, 2, \quad \phi|_{t=0} = \phi_0, \quad \text{with } \partial^i a_i = 0. \quad (4-26)$$

Then local wellposedness of (aME) with initial data as in (4-19) implies local wellposedness of (ME) in the Coulomb gauge with initial data given by (4-26).

Proof. First, in view of Section 4A it is clear that if u, v satisfy (aME) with initial data as in (4-19), then solutions of (ME) in the Coulomb gauge satisfy the initial data as given in (4-26).

Local existence. From (4-24), if

$$u, v \in \mathcal{H}_T^{s+1, \theta}, \quad \text{then} \quad \phi, A = *df \in H_T^{s, \theta},$$

as needed. We now verify that if (u, v, A_0) solve (aME), then (ϕ, df, A_0) solve (ME) in the Coulomb gauge. Since $A = *df$, A is in the Coulomb gauge: $d*A = -*d*(*df) = 0$. Next note that (ME) in the

Coulomb gauge is equivalent to (4-7) and (4-8). Suppose u, v, A_0 solve (aME). It follows (df, ϕ) solve (4-12) and (4-13). Add (4-12) to (4-13) to recover (4-11), which is equivalent to (4-8).

Next given (aME) we need to show (4-7) holds. Write (4-7) in coordinates,

$$\partial_1 A_0 - \partial_2 \phi + \partial_t \partial_2 f = [\partial_1 f, \phi] - [A_0, \partial_2 f], \quad (4-27)$$

$$\partial_2 A_0 + \partial_1 \phi - \partial_t \partial_1 f = [\partial_2 f, \phi] + [A_0, \partial_1 f]. \quad (4-28)$$

From the elliptic equation in (aME) we have:

$$A_0 = \Delta^{-1}(-\partial_1[A_0, \partial_2 f] + \partial_2[A_0, \partial_1 f] + \partial_1[\partial_1 f, \phi] + \partial_2[\partial_2 f, \phi]). \quad (4-29)$$

Also subtract (4-12) from (4-13) and multiply by $|\zeta|$ on both sides to obtain (4-9), which implies

$$\phi - \partial_t f = \Delta^{-1}(\partial_j[A_0, \partial_j f] - \partial_2[\partial_1 f, \phi] + \partial_1[\partial_2 f, \phi]). \quad (4-30)$$

In order to recover (4-27), first use (4-29) to get

$$\partial_1 A_0 = \Delta^{-1}(-\partial_1^2[A_0, \partial_2 f] + \partial_1 \partial_2[A_0, \partial_1 f] + \partial_1^2[\partial_1 f, \phi] + \partial_1 \partial_2[\partial_2 f, \phi]). \quad (4-31)$$

Next use (4-30) to get

$$\partial_2(\phi - \partial_t f) = \Delta^{-1}(\partial_2 \partial_1[A_0, \partial_1 f] + \partial_2^2[A_0, \partial_2 f] - \partial_2^2[\partial_1 f, \phi] + \partial_2 \partial_1[\partial_2 f, \phi]), \quad (4-32)$$

and subtract this from (4-31) to get (4-27) as needed. We recover (4-28) in the exactly same way.

Continuous dependence on initial data. We would like to show that

$$\begin{aligned} \|A_0 - A'_0\|_{C_b([0, T], \dot{H}^r)} + \|A_1 - A'_1\|_{H_T^{s, \theta}} + \|A_2 - A'_2\|_{H_T^{s, \theta}} + \|\phi - \phi'\|_{H_T^{s, \theta}} \\ \lesssim \|a_1 - a'_1\|_{H^s} + \|a_2 - a'_2\|_{H^s} + \|\phi_0 - \phi'_0\|_{H^s} \end{aligned} \quad (4-33)$$

for any a'_1, a'_2, ϕ'_0 sufficiently close to a_1, a_2, ϕ_0 . In view of LWP for (aME) with data given by

$$u(0) = v(0) = 0, \quad \partial_t u(0) = \phi_0 + ih, \quad \partial_t v(0) = \phi_0 - ih, \quad h = R_2 a_1 - R_1 a_2,$$

and by (4-25) we have

$$\begin{aligned} \|u - u'\|_{\mathcal{Y}_T^{s+1, \theta}} + \|v - v'\|_{\mathcal{Y}_T^{s+1, \theta}} + \|A_0 - A'_0\|_{C_b([0, T], \dot{H}^r)} \\ \lesssim \|u'_0\|_{H^{s+1}} + \|\phi_0 + ih - u'_1\|_{H^s} + \|v'_0\|_{H^{s+1}} + \|\phi_0 - ih - v'_1\|_{H^s}, \end{aligned} \quad (4-34)$$

for all u'_0, v'_0, u'_1, v'_1 satisfying

$$\|u'_0\|_{H^{s+1}} + \|\phi_0 + ih - u'_1\|_{H^s} + \|v'_0\|_{H^{s+1}} + \|\phi_0 - ih - v'_1\|_{H^s} \leq \delta, \quad \text{for some } \delta > 0. \quad (4-35)$$

In particular choose

$$u'_0 = v'_0 = 0, \quad u'_1 = \phi'_0 + ih', \quad v'_1 = \phi'_0 - ih', \quad h' = R_2 a'_1 - R_1 a'_2, \quad (4-36)$$

such that

$$\begin{aligned} \|\phi_0 + ih - \phi'_0 - ih'\|_{H^s} + \|\phi_0 - ih - \phi'_0 + ih'\|_{H^s} &\lesssim \|\phi_0 - \phi'_0\|_{H^s} + \|R_1(a_2 - a'_2)\|_{H^s} + \|R_2(a_1 - a'_1)\|_{H^s} \\ &\leq \|\phi_0 - \phi'_0\|_{H^s} + \|a_1 - a'_1\|_{H^s} + \|a_2 - a'_2\|_{H^s} \\ &\leq \delta. \end{aligned} \quad (4-37)$$

Then, by (4-34)–(4-37), $\|A_0 - A'_0\|_{C_b([0,T],\dot{H}^r)}$ is bounded by the right side of (4-33). Next observe that

$$\begin{aligned} \|A_1 - A'_1\|_{H_T^{s,\theta}} &\lesssim \|R_2(\partial_t + iD)(u - u')\|_{H_T^{s,\theta}} + \|R_2(\partial_t - iD)(v - v')\|_{H_T^{s,\theta}} \\ &\leq \|u - u'\|_{\mathcal{H}_T^{s+1,\theta}} + \|v - v'\|_{\mathcal{H}_T^{s+1,\theta}}. \end{aligned}$$

So again by (4-34)–(4-37) $\|A_1 - A'_1\|_{H_T^{s,\theta}}$ is bounded by the right side of (4-33). We bound the difference for A_2 and ϕ in a similar fashion.

Uniqueness. By LWP of (aME), A_0 is unique in the required class. We need to show A and ϕ are unique in $H_T^{s,\theta}$. However, by (4-33) this is obvious. \square

5. Proof of the Main Theorem

By Theorem 4.1 it is enough to show LWP for (aME). We start by explaining how we are going to perform our iteration.

5A. Set up of the iteration. Equations (aME) are written for functions u and v . Nevertheless, functions u and v are only our auxiliary functions, and we are really interested in solving for df and ϕ . In addition, the nonlinearities \mathcal{B}_\pm are a linear combination of terms \mathcal{B}_i , $i = 1, 2, 3, 4$, given by (4-22), and the \mathcal{B}_i are written in terms of ϕ , df and A_0 . Also, when we do our estimates, it is easier to keep the \mathcal{B}_i in terms of ϕ and df with the exception of \mathcal{B}_2 , which we rewrite in terms of ∂u and ∂v (see Section 5B2 for the details). These comments motivate the following procedure for our iteration. Start with $\phi_{-1} = df_{-1} = 0$. Then $\mathcal{B}_\pm \equiv 0$. Solve the homogeneous wave equations for u_0, v_0 with the initial data given by (4-19). Then to solve for df_0, ϕ_0 , use (4-24). Then feed ϕ_0 and df_0 into the elliptic equation,

$$\Delta A_{0,0} = -d^*([A_{0,0}, *df_0] + [df_0, \phi_0]), \quad (5-1)$$

and solve for $A_{0,0}$. Next we take df_0, ϕ_0 and $A_{0,0}$ plug them into $\mathcal{B}_1, \mathcal{B}_3, \mathcal{B}_4$, but rewrite \mathcal{B}_2 in terms of $\partial u_0, \partial v_0$. We continue in this manner, so at the j th step of the iteration, $j \geq 1$, we solve

$$\begin{aligned} \square u_j &= -\mathcal{B}_1(\nabla f_{j-1}) - i\mathcal{B}_2(\partial u_{j-1}, \partial v_{j-1}) + \mathcal{B}_3(A_{0,j-1}, \phi_{j-1}) - i\mathcal{B}_4(A_{0,j-1}, \nabla f_{j-1}), \\ \square v_j &= -\mathcal{B}_1(\nabla f_{j-1}) + i\mathcal{B}_2(\partial u_{j-1}, \partial v_{j-1}) + \mathcal{B}_3(A_{0,j-1}, \phi_{j-1}) + i\mathcal{B}_4(A_{0,j-1}, \nabla f_{j-1}), \\ \Delta A_{0,j} &= -d^*([A_{0,j}, *df_j] + [df_j, \phi_j]). \end{aligned}$$

5B. Estimates needed. The elliptic equation is discussed in Section 5C. Therefore we begin by discussing the inversion of the wave operator in $\mathcal{H}^{s+1,\theta}$ spaces. The main idea is that for the purposes of local in time estimates \square^{-1} can be replaced with $\Lambda_\pm^{-1} \Lambda_\mp^{-1}$. The first estimates, leading to wellposedness for small initial data, were proved by Klainerman and Machedon [1995]. The small data assumption was removed by Selberg [2002b], who showed that by introducing ε small enough in the invertible version

of the wave operator, that is, $\Lambda_+^{-1}\Lambda_-^{-1+\varepsilon}$, we can use initial data as large as we wish.⁵ Selberg [2002b] also gave a very useful, general framework for local wellposedness of wave equations, which reduces the proof of the [Main Theorem](#) to establishing the estimates below, for the nonlinearities \mathcal{B}_\pm , and to combining them with appropriate elliptic estimates from [Section 5C](#). The needed estimates for \mathcal{B}_\pm are

$$\|\Lambda_+^{-1}\Lambda_-^{-1+\varepsilon}\mathcal{B}_\pm(\partial u, \partial v, A_0)\|_{\mathcal{H}^{s+1,\theta}} \lesssim \|u\|_{\mathcal{H}^{s+1,\theta}} + \|v\|_{\mathcal{H}^{s+1,\theta}}, \quad (5-2)$$

$$\|\Lambda_+^{-1}\Lambda_-^{-1+\varepsilon}(\mathcal{B}_\pm(\partial u, \partial v, A_0) - \mathcal{B}_\pm(\partial u', \partial v', A'_0))\|_{\mathcal{H}^{s+1,\theta}} \lesssim \|u - u'\|_{\mathcal{H}^{s+1,\theta}} + \|v - v'\|_{\mathcal{H}^{s+1,\theta}}, \quad (5-3)$$

where the suppressed constants depend continuously on the $\mathcal{H}^{s+1,\theta}$ norms of u, u', v, v' . Since \mathcal{B}_\pm are bilinear, (5-3) can follow from (5-2). In this paper small initial data is necessary (see [Theorem 3.3](#) and [Section 5C](#)), so we do not need ε , but we keep it to make the estimates general. Let $\frac{1}{4} < s < \frac{1}{2}$ and set θ, ε as follows:

$$\frac{3}{4} - \frac{\varepsilon}{2} < \theta \leq s + \frac{1}{2} - \varepsilon, \quad \text{and} \quad \theta < 1 - \varepsilon, \quad 0 \leq \varepsilon < \min\left(2s - \frac{1}{2}, \frac{1}{2}\right).$$

Next observe $\Lambda_+\Lambda_-^{-1-\varepsilon}\mathcal{H}^{s+1,\theta} = H^{s,\theta-1+\varepsilon}$, as well as that

$$\|\nabla f\|_{H^{s,\theta}}, \|\phi\|_{H^{s,\theta}} \lesssim \|u\|_{\mathcal{H}^{s+1,\theta}} + \|v\|_{\mathcal{H}^{s+1,\theta}}.$$

Therefore, using (4-21) and (4-22), it is enough to prove the following:

$$\|\mathcal{B}_1\|_{H^{s,\theta-1+\varepsilon}} = \|[\partial_1 f, \partial_2 f]\|_{H^{s,\theta-1+\varepsilon}} \lesssim \|\nabla f\|_{H^{s,\theta}}^2, \quad (5-4)$$

$$\|\mathcal{B}_2\|_{H^{s,\theta-1+\varepsilon}} \lesssim \|[\partial_j f, \phi]\|_{H^{s,\theta-1+\varepsilon}} \lesssim \|\partial_j f\|_{H^{s,\theta}} \|\phi\|_{H^{s,\theta}} \quad \text{for } j = 1, 2, \quad (5-5)$$

$$\|\mathcal{B}_3\|_{H^{s,\theta-1+\varepsilon}} \lesssim \|A_0\phi\|_{H^{s,\theta-1+\varepsilon}} \lesssim \|A_0\| \|\phi\|_{H^{s,\theta}}, \quad (5-6)$$

$$\|\mathcal{B}_4\|_{H^{s,\theta-1+\varepsilon}} \lesssim \|A_0\partial_j f\|_{H^{s,\theta-1+\varepsilon}} \lesssim \|A_0\| \|\partial_j f\|_{H^{s,\theta}} \quad \text{for } j = 1, 2, \quad (5-7)$$

where the norm used for A_0 is immaterial, mainly because in [Section 5C](#) we show that

$$\|A_0\| \lesssim \|\nabla f\|_{H^{s,\theta}} \|\phi\|_{H^{s,\theta}}. \quad (5-8)$$

A few remarks are in order. Estimate (5-4) corresponds to estimates for the null form Q_{ij} , and estimate (5-5) gives rise to a new null form Q (this is discussed in the next two sections). A_0 in estimates (5-6) and (5-7) solves the elliptic equation in (aME), which results in a quite good regularity for A_0 . As a result, we do not have to look for any special structures to get (5-6) and (5-7) to hold, so we can drop the brackets, and also treat these estimates as equivalent since ϕ and df exhibit the same regularity. Finally, since Riesz transforms are clearly bounded on L^2 , we ignore them in the estimates needed in (5-5) and (5-7). The estimates (5-4) and (5-5) for the null forms are the most interesting. Hence we discuss them first, and then we consider the elliptic terms.

5B1. *Null forms: proof of estimate (5-4).* $[\partial_1 f, \partial_2 f]$ has a structure of a null form Q_{ij} :

$$[\partial_1 f, \partial_2 f] = \partial_1 f \partial_2 f - \partial_2 f \partial_1 f = Q_{12}(f, f).$$

⁵See also [Klainerman and Selberg 2002, Section 5] for an excellent discussion and motivation of the issues involved in the Picard iteration.

It follows that (5-4) is equivalent to

$$\|Q_{12}(f, f)\|_{H^{s,\theta-1+\varepsilon}} \lesssim \|\nabla f\|_{H^{s,\theta}}^2.$$

Fortunately the hard work for null forms of type $Q_{\alpha,\beta}$ in two dimensions is already carried out by Zhou [1997]. His proof is done using spaces $N^{s+1,\theta}$ with the norm given by⁶

$$\|u\|_{N^{s+1,\theta}} = \|\Lambda_+^{s+1} \Lambda_-^\theta u\|_{L^2}. \tag{5-9}$$

In his work $\theta = s + \frac{1}{2}$. We state Zhou’s result.

Theorem [Zhou 1997]. Consider in \mathbb{R}^{2+1} the space-time norms (5-9) and functions φ, ψ defined on \mathbb{R}^{2+1} . The estimates

$$\|Q_{\alpha\beta}(\varphi, \psi)\|_{N^{s,s-1/2}} \lesssim \|\varphi\|_{N^{s+1,s+1/2}} \|\psi\|_{N^{s+1,s+1/2}}$$

hold for any $\frac{1}{4} < s < \frac{1}{2}$.

Our iteration is done using spaces $\mathcal{H}^{s+1,\theta}$. Inspection of Zhou’s proof shows that it could be easily modified to be placed in the context of $\mathcal{H}^{s+1,\theta}$ spaces. However, even though our auxiliary functions’ iterates u_j and v_j belong to $\mathcal{H}^{s+1,\theta}$, from (4-24) we only have

$$df \in H^{s,\theta} \Rightarrow \|\Lambda^s \Lambda_-^\theta Df\|_{L^2(\mathbb{R}^{2+1})} < \infty, \tag{5-10}$$

but again inspection of Zhou’s proof shows we can still handle $Q_{12}(f, f)$ given only that (5-10) holds. Zhou’s proof works for $\frac{1}{4} < s < \frac{1}{2}$. However, it motivates an alternate proof that uses $\mathcal{H}^{s+1,\theta}$ and works for all values of $s > \frac{1}{4}$. The proof is closely related to the original proof in [Zhou 1997], but on the surface it seems more concise. The reason for this is that we use Theorem F from [Klainerman and Selberg 2002], which involves all the technicalities. See [Czubak 2008] for the details.

5B2. Null forms: proof of estimate (5-5). We need

$$\|[\partial_j f, \phi]\|_{H^{s,\theta-1+\varepsilon}} \lesssim \|\partial_j f\|_{H^{s,\theta}} \|\phi\|_{H^{s,\theta}}, \quad j = 1, 2.$$

However, analysis of the first iterate shows that for this estimate to hold we need $s > \frac{3}{4}$, so we need to work a little bit harder, and use (4-24)⁷

$$[\partial_j f, \phi] = -\frac{i}{4} [R_j(\partial_t u + i D u - \partial_t v + i D v), \partial_t u + i D u + \partial_t v - i D v]. \tag{5-11}$$

If we use the bilinearity of the bracket, we can group (5-11) by terms involving brackets of u with itself, v with itself, and then also by the terms that are mixed, that is, involve both u and v . So we have

$$\begin{aligned} 4i[\partial_j f, \phi] = & [R_j(\partial_t + i D)u, (\partial_t + i D)u] - [R_j(\partial_t - i D)v, (\partial_t - i D)v] \\ & + [R_j(\partial_t + i D)u, (\partial_t - i D)v] - [R_j(\partial_t - i D)v, (\partial_t + i D)u]. \end{aligned}$$

Since u and v are matrix-valued and do not commute we need to combine the last two brackets to take advantage of a null form structure. This corresponds to (5-13) below (note the plus sign in the formula).

The needed estimates are contained in the following theorem.

⁶See [Selberg 1999, Section 3.5] for a comparison with $\mathcal{H}^{s+1,\theta}$ spaces.

⁷The obvious way is to just substitute for ϕ and leave $\partial_j f$ the same, but it is an exercise to see that this does not work — for several reasons!

Theorem 5.1. *Suppose $s > \frac{1}{4}$ and $\frac{3}{4} - \frac{\varepsilon}{2} < \theta \leq s + \frac{1}{2}$, with $\theta < 1 - \varepsilon$ and $0 \leq \varepsilon < \min\left(2s - \frac{1}{2}, \frac{1}{2}\right)$. Let $Q(\varphi, \psi)$ be given by*

$$Q(\varphi, \psi) = (\partial_t \pm iD)R_j\varphi(\partial_t \pm iD)\psi - (\partial_t \pm iD)\varphi(\partial_t \pm iD)R_j\psi, \quad \text{or} \quad (5-12)$$

$$Q(\varphi, \psi) = (\partial_t \pm iD)R_j\varphi(\partial_t \mp iD)\psi + (\partial_t \pm iD)\varphi(\partial_t \mp iD)R_j\psi. \quad (5-13)$$

Then

$$Q(\mathcal{H}^{s+1,\theta}, \mathcal{H}^{s+1,\theta}) \hookrightarrow H^{s,\theta-1+\varepsilon} \quad (5-14)$$

or, equivalently,

$$\|Q(\varphi, \psi)\|_{H^{s,\theta-1+\varepsilon}} \lesssim \|\varphi\|_{\mathcal{H}^{s+1,\theta}} \|\psi\|_{\mathcal{H}^{s+1,\theta}}. \quad (5-15)$$

Proof. We show the details only for

$$(\partial_t + iD)R_j\varphi(\partial_t - iD)\psi + (\partial_t + iD)\varphi(\partial_t - iD)R_j\psi,$$

as the rest follows similarly. Observe that the symbol of Q is then

$$q(\tau, \zeta, \lambda, \eta) = \left(\frac{\zeta_j}{|\zeta|} + \frac{\eta_j}{|\eta|}\right)(\tau + |\zeta|)(\lambda - |\eta|).$$

Suppose $\tau\lambda \geq 0$, then

$$q \leq 2|(\tau + |\zeta|)(\lambda - |\eta|)| \leq \begin{cases} 2|\tau + |\zeta||\lambda - |\eta|| & \text{if } \tau, \lambda \geq 0, \\ 2|\tau - |\zeta||\eta| + |\lambda| & \text{if } \tau, \lambda \leq 0. \end{cases}$$

It follows

$$\iint_{\tau\lambda \geq 0} |\Lambda^s \Lambda_-^{\theta-1+\varepsilon} Q(\varphi, \psi)|^2 d\tau d\xi \lesssim \|D_+\varphi D_-\psi\|_{H^{s,\theta-1+\varepsilon}}^2 + \|D_-\varphi D_+\psi\|_{H^{s,\theta-1+\varepsilon}}^2, \quad (5-16)$$

and the estimate follows by [Theorem 5.2](#) below.

Suppose $\tau\lambda < 0$. If we break down the computations into the region

$$\{(\tau, \zeta), (\lambda, \eta) : |\tau| \geq 2|\zeta| \text{ or } |\lambda| \geq 2|\eta|\} \quad (5-17)$$

and its complement, then in the region [\(5-17\)](#), we bound q by

$$q \leq 2(|\tau| + |\zeta|)(|\lambda| + |\eta|),$$

since there we do not need any special structure. (This is a simple exercise in this region; see [\[Czubak 2008, Appendix B\]](#).)

In the complementary region, we have

$$q \leq 4|\zeta||\eta| \left| \frac{\zeta_i}{|\zeta|} + \frac{\eta_i}{|\eta|} \right|,$$

which is the absolute value of the symbol of the null form Q_{ij} in the first iterate. It has received a lot of attention, but we have not seen a reference where it was discussed in a context other than that of the initial data in $H^{s+1} \times H^s$. This may be because it has not come up as a nonlinearity before, and/or because it can be handled in the same way as the null form Q_{ij} . The details are in [\[Czubak 2008\]](#). \square

Now we prove an estimate needed to show (5-16) is bounded by the square of the right side of (5-15).

Theorem 5.2. *Let $s > 0$, $\max(\frac{1}{2}, 1 - s) < \theta < 1$, and $0 \leq \varepsilon \leq 1 - \theta$. Then*

$$\|D_+ \varphi D_- \psi\|_{H^{s, \theta-1+\varepsilon}} \lesssim \|\varphi\|_{\mathcal{H}^{s+1, \theta}} \|\psi\|_{\mathcal{H}^{s+1, \theta}}.$$

Proof. We would like to show that $\|\Lambda^s \Lambda_-^{\theta-1+\varepsilon} (D_+ \varphi D_- \psi)\|_{L^2(\mathbb{R}^{2+1})} \lesssim \|\varphi\|_{\mathcal{H}^{s+1, \theta}} \|\psi\|_{\mathcal{H}^{s+1, \theta}}$. This follows from showing that

$$H^{s, \theta} \cdot \mathcal{H}^{s+1, \theta-1} \hookrightarrow H^{s, \theta-1+\varepsilon},$$

which by the product rule⁸ for the operator Λ^s in turn follows from

$$H^{0, \theta} \cdot \mathcal{H}^{s+1, \theta-1} \hookrightarrow H^{0, \theta-1+\varepsilon}, \quad H^{s, \theta} \cdot \mathcal{H}^{1, \theta-1} \hookrightarrow H^{0, \theta-1+\varepsilon}.$$

It is easy to check that $\mathcal{H}^{s+1, \theta-1} \hookrightarrow H^{s+1+\theta-1, 0}$ and $\mathcal{H}^{1, \theta-1} \hookrightarrow H^{\theta, 0}$, so we just need to show

$$H^{0, \theta} \cdot H^{s+\theta, 0} \hookrightarrow H^{0, \theta-1+\varepsilon}, \quad H^{s, \theta} \cdot H^{\theta, 0} \hookrightarrow H^{0, \theta-1+\varepsilon},$$

which are weaker than

$$H^{0, \theta} \cdot H^{s+\theta, 0} \hookrightarrow L^2, \quad H^{s, \theta} \cdot H^{\theta, 0} \hookrightarrow L^2,$$

but those follow from the Klainerman–Selberg estimate (2-6) as long as $s + \theta > 1$, which holds by the conditions we impose on s and θ . \square

An alternate approach could be to follow the set up used by [Klainerman and Machedon 1995] and estimate the integral directly.

5B3. Elliptic piece: proof of estimate (5-6). Recall we wish to show

$$\|A_0 w\|_{H^{s, \theta-1+\varepsilon}} \lesssim \|A_0\| \|w\|_{H^{s, \theta}}. \tag{5-18}$$

We need this estimate during our iteration, so we really mean $A_{0, j}$, but for simplicity we omit writing of the index j . Now we choose a norm for A_0 to be anything that makes (5-18) possible to establish. This results in

$$\|A_0\| = \|A_0\|_{L_t^{\tilde{p}} L_x^\infty} + \|D^s A_0\|_{L_t^p L_x^q},$$

where

$$\tilde{p} \in \left(1 - 2s, \frac{1}{2}\right), \quad \frac{2}{p} = 1 - \frac{1}{q}, \quad \max\left(\frac{1}{3}(1 - 2s), \frac{s}{2}\right) < \frac{1}{q} < \frac{2}{3}s. \tag{5-19}$$

For now we assume we can show $A_0 \in L_t^{\tilde{p}} L_x^\infty \cap L_t^p \dot{W}_x^{s, q}$ and delay the proof to Section 5C, where the reasons for our choices of \tilde{p}, p, q should become clear. We start by using $\theta - 1 + \varepsilon < 0$:

$$\|A_0 w\|_{H^{s, \theta-1+\varepsilon}} \leq \|\Lambda^s (A_0 w)\|_{L^2(\mathbb{R}^{2+1})} \lesssim \|A_0 w\|_{L^2(\mathbb{R}^{2+1})} + \|D^s (A_0 w)\|_{L^2(\mathbb{R}^{2+1})}. \tag{5-20}$$

For the first term, by Hölder’s inequality,

$$\begin{aligned} \|A_0 w\|_{L^2(\mathbb{R}^{2+1})} &\leq \|A_0\|_{L_t^{\tilde{p}} L_x^\infty} \|w\|_{L_t^{\tilde{p}'} L_x^2}, \quad \text{for } \frac{1}{\tilde{p}} + \frac{1}{\tilde{p}'} = \frac{1}{2}, \quad p \text{ as in (5-19)} \\ &\lesssim \|A_0\| \|w\|_{H^{0, \theta}}, \quad \text{by (2-4)} \\ &\leq \|A_0\| \|w\|_{H^{s, \theta}}. \end{aligned} \tag{5-21}$$

⁸On L^2 this is very easy to establish using triangle inequality. See [Klainerman and Selberg 2002].

We bound the second term in (5-20) by

$$\|D^s(A_0 w)\|_{L^2(\mathbb{R}^{2+1})} \lesssim \underbrace{\|A_0\|_{L_t^{\tilde{p}} L_x^\infty} \|D^s w\|_{L_t^{\tilde{p}'} L_x^2}}_I + \underbrace{\|D^s A_0\|_{L_t^p L_x^q} \|w\|_{L_t^{p'} L_x^{q'}}}_{II},$$

where

$$\frac{1}{p} + \frac{1}{p'} = \frac{1}{2} = \frac{1}{q} + \frac{1}{q'}$$

and p, q are as in (5-19) and \tilde{p} as in (5-21). I is handled similarly to (5-21) as follows. Apply (2-4) with $u = D^s w$ to obtain

$$I \lesssim \|A_0\| \|D^s w\|_{H^{0,\theta}} \leq \|A_0\| \|w\|_{H^{s,\theta}}. \tag{5-22}$$

We now consider II . By the choices of p, q , the Klainerman–Selberg estimate (2-5) applies (see the discussion in Section 5D for an explanation) and gives

$$II \leq \|A_0\| \|w\|_{L_t^{p'} L_x^{q'}} \lesssim \|A_0\| \|w\|_{H^{1-2/q'-1/p',\theta}}. \tag{5-23}$$

From (5-19) we also have

$$II \lesssim \|A_0\| \|w\|_{H^{1-2/q'-1/p',\theta}} \lesssim \|A_0\| \|w\|_{H^{s,\theta}}. \tag{5-24}$$

and (5-18) follows now from (5-21), (5-22) and (5-24).

Remark 5.1. The above proof illustrates other difficulties due to working in two dimensions. Initially, we wanted to follow the proof of estimate (38) in [Selberg 2002a], and just use the $\|\Lambda^s A_0\|_{L_t^p L_x^q}$ norm. Unfortunately in two dimensions, the condition $sq > 2$ needed to show that $A_0 \in L_t^p L_x^\infty$ is disjoint from the conditions needed to use Klainerman–Tataru estimate (2-3) and establish that $\Lambda^s A_0 \in L_t^p L_x^q$ in the first place. This resulted in the $L_t^{\tilde{p}} L_x^\infty \cap L_t^p \dot{W}_x^{s,q}$ space above and also having to employ the Klainerman–Selberg estimate (2-5), which was not needed for the proof of [Selberg 2002a, estimate (38)].

5C. Elliptic regularity: estimates for A_0 . Here we present a variety of a priori estimates for the non-dynamical variable A_0 . At each point we could add the index j to A_0, df and ϕ . Therefore the presentation also applies to the iterates $A_{0,j}$. It is an exercise to show that the estimates we obtain here are enough to solve for $A_{0,j}$ at each step as well as to close the iteration for A_0 . Let A_0 solve

$$\Delta A_0 = -d^*[A_0, *df] - d^*[df, \phi] = -\partial_1[A_0, \partial_2 f] + \partial_2[A_0, \partial_1 f] + \partial_j[\partial_j f, \phi].$$

There is a wide range of estimates A_0 satisfies. Nevertheless, the two spatial dimensions limit our “range of motion.” For example, it does not seem possible to place $A_0(t)$ in L^2 . We state the general results and only show the cases we need to prove $A_0 \in L_t^{\tilde{p}} L_x^\infty \cap L_t^p \dot{W}_x^{s,q}$ as required in the last section. The rest of the cases can be found in [Czubak 2008]. We add that the proofs of both of the following theorems were originally inspired by the proof of estimate (45) in [Selberg 2002a]. We start with the homogeneous estimates.

Theorem 5.3. *Let $s > 0$, and let $0 \leq a \leq s + 1$ be given, and suppose $1 \leq p \leq \infty$ and $1 < q < \infty$ satisfy*

$$\max\left(\frac{1}{3}(1 + 2a - 4s), \frac{1}{2}(1 + a - 4s), \frac{1}{2} \min(a, 1)\right) < \frac{1}{q} < \frac{1+a}{2}, \tag{5-25}$$

$$1 - \frac{2}{q} + a - 2s \leq \frac{1}{p} \leq \frac{1}{2}\left(1 - \frac{1}{q}\right), \quad \frac{1}{p} < \left(1 - \frac{2}{q} + a\right). \tag{5-26}$$

(i) If $0 \leq a \leq 1$ and the $H^{s,\theta}$ norm of ∇f is sufficiently small, then $A_0 \in L_t^p \dot{W}_x^{a,q}$ and we have the estimate

$$\|A_0\|_{L_t^p \dot{W}_x^{a,q}} \lesssim \|\phi\|_{H^{s,\theta}} \|\nabla f\|_{H^{s,\theta}}. \quad (5-27)$$

(ii) If $1 < a \leq s+1$ and $A_0 \in L_t^p L_x^{(1/q-1/2)^{-1}}$, then $A_0 \in L_t^p \dot{W}_x^{a,q}$ and we have

$$\|A_0\|_{L_t^p \dot{W}_x^{a,q}} \lesssim (\|A_0\|_{L_t^p L_x^{(1/q-1/2)^{-1}}} + \|\phi\|_{H^{s,\theta}}) \|\nabla f\|_{H^{s,\theta}}. \quad (5-28)$$

Corollary 5.4. Let $s > 0$, then $A_0 \in C_b(I : \dot{H}_x^a)$, where

$$0 < a \leq \begin{cases} 2s & \text{if } 0 < s \leq 1, \\ 1+s & \text{if } 1 < s. \end{cases}$$

Proof of Corollary 5.4. Suppose $0 < s < \frac{1}{2}$. Then use part (i) of the theorem with $q = 2$ and $p = \infty$ to obtain $A_0 \in L_t^\infty \dot{H}_x^a$ for $a \leq 2s$. A_0 continuous as a function of time easily follows from a contraction argument in $C_b(I : \dot{H}_x^a)$ using $L_t^\infty \dot{H}_x^a$ estimates. The case $s \geq \frac{1}{2}$ is considered in [Czubak 2008]. \square

So far we just need $s > 0$ in order to make the estimates work. The requirement for $s > \frac{1}{4}$ does not come in until we start looking at the nonhomogeneous spaces, where also the range of p and q is smaller. However, we can distinguish two cases, $aq < 2$ and $aq > 2$.

Theorem 5.5. Let $s > 0$, and suppose the $H^{s,\theta}$ norm of ∇f is sufficiently small.

(i) If $aq < 2$ for $0 < a < \min(2s, 1)$ and if p and q satisfy

$$\max\left(\frac{1}{2} + a - 2s, \frac{a}{2}\right) < \frac{1}{q} < \frac{1}{2}, \quad (5-29)$$

$$1 - \frac{2}{q} + a - 2s \leq \frac{1}{p} < \frac{1}{2} - \frac{1}{q}, \quad (5-30)$$

then $A_0 \in L_t^p W_x^{a,q}$ and we have the estimate

$$\|A_0\|_{L_t^p W_x^{a,q}} \lesssim \|\phi\|_{H^{s,\theta}} \|\nabla f\|_{H^{s,\theta}}. \quad (5-31)$$

(ii) If $aq > 2$, we need $s > \frac{1}{4}$ and $0 < a < \min(4s-1, 1+s, 2s)$. Suppose p and q also satisfy

$$\max\left(\frac{a-s}{2}, \frac{1}{2} + a - 2s\right) < \frac{1}{q} < \frac{1}{2} \min(a, 1), \quad (5-32)$$

$$1 - \frac{2}{q} + a - 2s \leq \frac{1}{p} < \frac{1}{2} - \frac{1}{q}; \quad (5-33)$$

then $A_0 \in L_t^p W_x^{a,q}$ and we have the estimate

$$\|A_0\|_{L_t^p W_x^{a,q}} \lesssim \|\phi\|_{H^{s,\theta}} \|\nabla f\|_{H^{s,\theta}}. \quad (5-34)$$

Corollary 5.6. *If $s > \frac{1}{4}$ and the $H^{s,\theta}$ norm of ∇f is sufficiently small, we have in particular $A_0 \in L_t^p L_x^\infty$ for p satisfying*

$$1 - 2s < \frac{1}{p} < \frac{1}{2}, \quad (5-35)$$

and we have the estimate

$$\|A_0\|_{L_t^p L_x^\infty} \lesssim \|\phi\|_{H^{s,\theta}} \|\nabla f\|_{H^{s,\theta}}. \quad (5-36)$$

Proof of Corollary 5.6. For each $p \in (1 - 2s, \frac{1}{2})$ we can find some a and q , which satisfy the conditions of [Theorem 5.5](#), part (ii). The corollary then follows from the Sobolev embedding: $W^{a,q}(\mathbb{R}^2) \hookrightarrow L^\infty(\mathbb{R}^2)$ for $aq > 2$. \square

Remark 5.2. Estimate (5-34) is where $s > \frac{1}{4}$ is needed. Conditions on $\frac{1}{p}$ in (5-33) are needed so we can use (below) the Klainerman–Tataru estimate (2-3). In order to be able to choose such $\frac{1}{p}$, obviously $1 - \frac{2}{q} + a - 2s$ must be strictly less than $\frac{1}{2} - \frac{1}{q}$. This forces $\frac{1}{q}$ to be strictly greater than $\frac{1}{2} + a - 2s$. We also need $aq > 2$ to use the Sobolev embedding in [Corollary 5.6](#), so if we want to be able to find q between $\frac{1}{2} + a - 2s$ and $\frac{a}{2}$, a is forced to be strictly less than $4s - 1$. Therefore s must be greater than $\frac{1}{4}$. See below for another instance of requiring $s > \frac{1}{4}$.

5D. Proof of estimates needed in 5B3. Recall we would like to show $A_0 \in L_t^{\tilde{p}} L_x^\infty \cap L_t^p \dot{W}_x^{s,q}$. Therefore, we are interested in part (i) of [Theorem 5.3](#) and part ii) in [Theorem 5.5](#), so we can conclude [Corollary 5.6](#). Moreover, we need a specific case of part (i) in [Theorem 5.3](#), because we need $A_0 \in L_t^p \dot{W}_x^{s,q}$, where p, q in addition satisfy

$$1 - \frac{2}{p} \leq \frac{1}{q} < \frac{1}{2} \quad \text{and} \quad \frac{2}{q} - \frac{1}{2} + \frac{1}{p} \leq s, \quad (5-37)$$

so we can use the embedding

$$H^{s,\theta} \hookrightarrow H^{1-(1-2/q)-(1/2-1/p),\theta} \hookrightarrow L_t^{(1/2-1/p)^{-1}} L_x^{(1/2-1/q)^{-1}} \quad (5-38)$$

in (5-23) and (5-24). When we put (5-37) together with (5-25) and (5-26) with $a = s$, we obtain the second line of (5-19), namely

$$\frac{2}{p} = 1 - \frac{1}{q}, \quad \max\left(\frac{1}{3}(1 - 2s), \frac{s}{2}\right) < \frac{1}{q} < \frac{2}{3}s. \quad (5-39)$$

Remark 5.3. Observe that in order to be able to find such q we must have $s > \frac{1}{4}$.

Consider

$$\begin{aligned} \|A_0\|_{L_t^p \dot{W}_x^{s,q}} &= \|\Delta^{-1}(d^*[A_0, *df] + d^*[df, \phi])\|_{L_t^p \dot{W}_x^{s,q}} \\ &\lesssim \|D^{-1}(A_0 \nabla f)\|_{L_t^p \dot{W}_x^{s,q}} + \|D^{-1}(\nabla f \phi)\|_{L_t^p \dot{W}_x^{s,q}} \\ &\lesssim \|D^{s-1}(A_0 \nabla f)\|_{L_t^p L_x^q} + \|D^{s-1}(\nabla f \phi)\|_{L_t^p L_x^q} \\ &\lesssim \|A_0 \nabla f\|_{L_t^p L_x^r} + \|D^{s-1}(\nabla f \phi)\|_{L_t^p L_x^q}, \end{aligned} \quad (5-40)$$

where we use the Sobolev embedding with $\frac{1}{q} = \frac{1}{r} - \frac{1-s}{2}$. The latter term is bounded by $\|\nabla f\|_{H^{s,\theta}} \|\phi\|_{H^{s,\theta}}$ using the Klainerman–Tataru estimate (2-3), whose application we discuss in the section below. For the former we use $\frac{1}{r} = \frac{1}{q} + \frac{1-s}{2} = (\frac{1}{q} - \frac{s}{2}) + \frac{1}{2}$:

$$\|A_0 \nabla f\|_{L_t^p L_x^r} \leq \|A_0\|_{L_t^p L_x^{(1/q-s/2)^{-1}}} \|\nabla f\|_{L_t^\infty L_x^2} \lesssim \|A_0\|_{L_t^p \dot{W}_x^{s,q}} \|\nabla f\|_{H^{s,\theta}}. \quad (5-41)$$

Then if the $H^{s,\theta}$ norm of ∇f is sufficiently small, we obtain

$$\|A_0\|_{L_t^p \dot{W}_x^{s,q}} \lesssim \|\nabla f\|_{H^{s,\theta}} \|\phi\|_{H^{s,\theta}}, \tag{5-42}$$

as needed.

For the nonhomogeneous estimate, since here $\frac{1}{4} < s < \frac{1}{2}$ the upper bound for a is simply $4s - 1$. In addition, for our purposes right now it suffices to show the estimate for one particular a . Therefore we set $0 < a < \min(s, 4s - 1)$ for $\frac{1}{4} < s < \frac{1}{2}$, and we let p, q satisfy (5-32) and (5-33). We have

$$\begin{aligned} \|A_0\|_{L_t^p W_x^{a,q}} &\lesssim \|D^{-1}(A_0 \nabla f)\|_{L_t^p W_x^{a,q}} + \|D^{-1}(\nabla f \phi)\|_{L_t^p W_x^{a,q}} \\ &\lesssim \|D^{-1}(A_0 \nabla f)\|_{L_t^p L_x^q} + \|D^{-1}(\nabla f \phi)\|_{L_t^p L_x^q} + \|D^{a-1}(A_0 \nabla f)\|_{L_t^p L_x^q} + \|D^{a-1}(\nabla f \phi)\|_{L_t^p L_x^q}. \end{aligned} \tag{5-43}$$

The Klainerman–Tataru estimate (2-3) handles the second and the last term (see below). Consider the first term:

$$\begin{aligned} \|D^{-1}(A_0 \nabla f)\|_{L_t^p L_x^q} &\lesssim \|A_0 \nabla f\|_{L_t^p L_x^r}, && \text{for } \frac{1}{q} = \frac{1}{r} - \frac{1}{2}, \\ &\leq \|A_0\|_{L_t^p L_x^q} \|\nabla f\|_{L_t^\infty L_x^2} \\ &\leq \|A_0\|_{L_t^p W_x^{a,q}} \|\nabla f\|_{H^{s,\theta}}. \end{aligned} \tag{5-44}$$

For the third term we have

$$\begin{aligned} \|D^{a-1}(A_0 \nabla f)\|_{L_t^p L_x^q} &\lesssim \|A_0 \nabla f\|_{L_t^p L_x^r}, && \text{for } \frac{1}{q} = \frac{1}{r} - \frac{1-a}{2} \\ &\lesssim \|A_0\|_{L_t^p L_x^q} \|D^a \nabla f\|_{L_t^\infty L_x^2}, && \text{for } \frac{1}{r} = \frac{1}{q} + \left(\frac{1}{2} - \frac{a}{2}\right) \\ &\lesssim \|A_0\|_{L_t^p W_x^{a,q}} \|\nabla f\|_{H^{s,\theta}}, \end{aligned}$$

Then as before, this completes the proof if the $H^{s,\theta}$ norm of ∇f is sufficiently small.

Applying the Klainerman–Tataru theorem. We said that several of the above estimates follow from the Klainerman–Tataru estimate (2-3). We need to check that this is in fact the case. We begin by stating the theorem. We state it for two dimensions only, and as it is given in [Klainerman and Selberg 2002] (the original result holds for $n \geq 2$).

Theorem [Klainerman and Tataru 1999]. *Let $1 \leq p \leq \infty, 1 \leq q < \infty$. Assume that*

$$\frac{1}{p} \leq \frac{1}{2} \left(1 - \frac{1}{q}\right), \tag{5-45}$$

$$0 < \sigma < 2 \left(1 - \frac{1}{q} - \frac{1}{p}\right), \tag{5-46}$$

$$s_1, s_2 < 1 - \frac{1}{q} - \frac{1}{2p}, \tag{5-47}$$

$$s_1 + s_2 + \sigma = 2 \left(1 - \frac{1}{q} - \frac{1}{2p}\right). \tag{5-48}$$

Then

$$\|D^{-\sigma}(uv)\|_{L_t^p L_x^q(\mathbb{R}^2)} \lesssim \|u\|_{H^{s_1,\theta}} \|v\|_{H^{s_2,\theta}},$$

provided $\theta > \frac{1}{2}$.

The first time we use the theorem is in (5-40) for the term $\|D^{s-1}(\nabla f\phi)\|_{L_t^p L_x^q}$. Note that $\sigma = 1 - s$. Clearly $1 \leq p \leq \infty$, $1 \leq q < \infty$. Next by (5-39) $\frac{2}{p} = 1 - \frac{1}{q}$, so (5-45) holds. Since $s < \frac{1}{2}$, $\sigma > 0$, and we can see (5-46) holds when we substitute $\frac{1}{2} - \frac{1}{2q}$ for $\frac{1}{p}$ in the right side and use $\frac{1}{q} < \frac{2}{3}s$. Next we let $s_1 = s_2$ and with $\sigma = 1 - s > 0$, (5-48) implies (5-47), so we only check (5-48). To that end we must be able to choose s_1 so that

$$2s_1 = 1 - \frac{2}{q} - \frac{1}{p} + s \leq 2s,$$

which is equivalent to our condition on p and one of the lower bounds on $\frac{1}{q}$.

The next place we use the theorem is in (5-43) for $\|D^{-1}(\nabla f\phi)\|_{L_t^p L_x^q}$, $\|D^{a-1}(\nabla f\phi)\|_{L_t^p L_x^q}$, where p and q are as in (5-32) and (5-33) with $0 < a < \min(s, 4s - 1) < 1$. Then for $\sigma = 1$, by the right side of (5-33), (5-46) holds and implies (5-45). Note, since (5-46) is true with $\sigma = 1$, it is true with $\sigma = 1 - a$. Next, for $\sigma = 1$ (5-48) gives (5-47) and also for $\sigma = 1 - a$ as long as $0 < a < 1$. So again it is sufficient to see we can have s_1 defined by (5-48) such that $s_1 \leq s$, but for $\sigma = 1 - a$ that follows from the left side of (5-33), and shows we can find it for $\sigma = 1$ as well.

Acknowledgment

The author expresses deep gratitude to her thesis advisor Karen Uhlenbeck for her time, many helpful discussions, and in particular for suggesting the problem with the reformulation (4-12)–(4-13). The author also thanks the referees for many insightful comments and their careful reading of the manuscript.

References

- [Ablowitz et al. 2003] M. J. Ablowitz, S. Chakravarty, and R. G. Halburd, “Integrable systems and reductions of the self-dual Yang–Mills equations”, *J. Math. Phys.* **44**:8 (2003), 3147–3173. [MR 2004h:70034](#) [Zbl 1062.70050](#)
- [Atiyah and Hitchin 1988] M. Atiyah and N. Hitchin, *The geometry and dynamics of magnetic monopoles*, Princeton University Press, 1988. [MR 89k:53067](#) [Zbl 0671.53001](#)
- [Czubak 2008] M. Czubak, *Well-posedness for the space-time monopole equation and Ward wave map*, Ph.D. thesis, University of Texas, Austin, 2008.
- [Czubak and Uhlenbeck \geq 2010] M. Czubak and K. Uhlenbeck, “On the existence of Coulomb gauges”, In preparation.
- [Dai et al. 2006] B. Dai, C.-L. Terng, and K. Uhlenbeck, “On the space-time monopole equation”, pp. 1–30 in *Surveys in differential geometry*, edited by S.-T. Yau, Surv. Differ. Geom. **10**, Int. Press, Somerville, MA, 2006. [MR 2009f:53033](#) [Zbl 1157.53016](#)
- [Dell’Antonio and Zwanziger 1991] G. Dell’Antonio and D. Zwanziger, “Every gauge orbit passes inside the Gribov horizon”, *Comm. Math. Phys.* **138**:2 (1991), 291–299. [MR 92i:58029](#) [Zbl 0726.53067](#)
- [Dirac 1931] P. A. M. Dirac, “Quantised singularities in the electromagnetic field”, *Proceedings of the Royal Society of London, Series A* **133**:821 (1931), 60–72.
- [Foschi and Klainerman 2000] D. Foschi and S. Klainerman, “Bilinear space-time estimates for homogeneous wave equations”, *Ann. Sci. École Norm. Sup. (4)* **33**:2 (2000), 211–274. [MR 2001g:35145](#) [Zbl 0959.35107](#)
- [Jaffe and Taubes 1980] A. Jaffe and C. Taubes, *Vortices and monopoles: Structure of static gauge theories*, Progress in Physics **2**, Birkhäuser, Mass., 1980. [MR 82m:81051](#) [Zbl 0457.53034](#)
- [Klainerman 1984] S. Klainerman, “Long time behaviour of solutions to nonlinear wave equations”, pp. 1209–1215 in *Proceedings of the International Congress of Mathematicians* (Warsaw, 1983), vol. 2, PWN, 1984. [MR 804771](#) [Zbl 0581.35052](#)
- [Klainerman and Machedon 1993] S. Klainerman and M. Machedon, “Space-time estimates for null forms and the local existence theorem”, *Comm. Pure Appl. Math.* **46**:9 (1993), 1221–1268. [MR 94h:35137](#) [Zbl 0803.35095](#)

- [Klainerman and Machedon 1995] S. Klainerman and M. Machedon, “Smoothing estimates for null forms and applications”, *Duke Math. J.* **81**:1 (1995), 99–133. [MR 97h:35022](#) [Zbl 0909.35094](#)
- [Klainerman and Machedon 1996] S. Klainerman and M. Machedon, “Estimates for null forms and the spaces $H_{s,\delta}$ ”, *Internat. Math. Res. Notices* **17** (1996), 853–865. [MR 98j:46028](#) [Zbl 0909.35095](#)
- [Klainerman and Selberg 2002] S. Klainerman and S. Selberg, “Bilinear estimates and applications to nonlinear wave equations”, *Commun. Contemp. Math.* **4**:2 (2002), 223–295. [MR 2003d:35182](#) [Zbl 1146.35389](#)
- [Klainerman and Tataru 1999] S. Klainerman and D. Tataru, “On the optimal local regularity for Yang–Mills equations in \mathbf{R}^{4+1} ”, *J. Amer. Math. Soc.* **12**:1 (1999), 93–116. [MR 2000c:58052](#) [Zbl 0924.58010](#)
- [Lindblad 1996] H. Lindblad, “Counterexamples to local existence for semi-linear wave equations”, *Amer. J. Math.* **118**:1 (1996), 1–16. [MR 97b:35124](#) [Zbl 0855.35080](#)
- [Mason and Sparling 1989] L. J. Mason and G. A. J. Sparling, “Nonlinear Schrödinger and Korteweg–de Vries are reductions of self-dual Yang–Mills”, *Phys. Lett. A* **137**:1-2 (1989), 29–33. [MR 90d:58169](#)
- [Roe 1998] J. Roe, *Elliptic operators, topology and asymptotic methods*, 2nd ed., Pitman Research Notes in Mathematics Series **395**, Longman, Harlow, 1998. [MR 99m:58182](#) [Zbl 0919.58060](#)
- [Selberg 1999] S. Selberg, *Multilinear spacetime estimates and applications to local existence theory for nonlinear wave equations*, Ph.D. thesis, Princeton University, 1999.
- [Selberg 2002a] S. Selberg, “Almost optimal local well-posedness of the Maxwell–Klein–Gordon equations in 1 + 4 dimensions”, *Comm. Partial Differential Equations* **27**:5-6 (2002), 1183–1227. [MR 2003f:35247](#) [Zbl 1013.35077](#)
- [Selberg 2002b] S. Selberg, “On an estimate for the wave equation and applications to nonlinear problems”, *Differential Integral Equations* **15**:2 (2002), 213–236. [MR 2002h:35204](#) [Zbl 1032.35121](#)
- [Tao 2006] T. Tao, *Nonlinear dispersive equations: Local and global analysis*, CBMS Regional Conference Series in Mathematics **106**, Amer. Math. Soc., Providence, 2006. [MR 2008i:35211](#) [Zbl 1106.35001](#)
- [Uhlenbeck 1992] K. Uhlenbeck, “On the connection between harmonic maps and the self-dual Yang–Mills and the sine-Gordon equations”, *J. Geom. Phys.* **8**:1-4 (1992), 283–316. [MR 93f:58050](#) [Zbl 0747.58025](#)
- [Ward 1985] R. S. Ward, “Integrable and solvable systems, and relations among them”, *Philos. Trans. Roy. Soc. London Ser. A* **315**:1533 (1985), 451–457. [MR 87e:58105](#) [Zbl 0579.35078](#)
- [Ward 1989] R. S. Ward, “Twistors in 2 + 1 dimensions”, *J. Math. Phys.* **30**:10 (1989), 2246–2251. [MR 90k:32089](#) [Zbl 0699.58065](#)
- [Ward 1999] R. S. Ward, “Two integrable systems related to hyperbolic monopoles”, *Asian J. Math.* **3**:1 (1999), 325–332. [MR 2000j:37098](#) [Zbl 0986.37061](#)
- [Zhou 1997] Y. Zhou, “Local existence with minimal regularity for nonlinear wave equations”, *Amer. J. Math.* **119**:3 (1997), 671–703. [MR 98e:35119](#) [Zbl 0881.35077](#)

Received 10 Feb 2009. Revised 15 Sep 2009. Accepted 21 Jan 2010.

MAGDALENA CZUBAK: czubak@math.toronto.edu

Department of Mathematics, University of Toronto, 40 Saint George Street, Toronto, Ontario M5S 2E4, Canada

<http://www.math.toronto.edu/czubak/>

REGULARITY OF ALMOST PERIODIC MODULO SCALING SOLUTIONS FOR MASS-CRITICAL NLS AND APPLICATIONS

DONG LI AND XIAOYI ZHANG

We consider the L_x^2 solution u to mass-critical NLS $iu_t + \Delta u = \pm|u|^{4/d}u$. We prove that in dimensions $d \geq 4$, if the solution is spherically symmetric and is *almost periodic modulo scaling*, then it must lie in $H_x^{1+\varepsilon}$ for some $\varepsilon > 0$. Moreover, the kinetic energy of the solution is localized uniformly in time. One important application of the theorem is a simplified proof of the scattering conjecture for mass-critical NLS without reducing to three enemies. As another important application, we establish a Liouville type result for L_x^2 initial data with ground state mass. We prove that if a radial L_x^2 solution to focusing mass-critical problem has the ground state mass and does not scatter in both time directions, then it must be global and coincide with the solitary wave up to symmetries. Here the ground state is the unique, positive, radial solution to elliptic equation $\Delta Q - Q + Q^{1+4/d} = 0$. This is the first rigidity type result in scale invariant space L_x^2 .

1. Introduction

Main results. We consider the d -dimensional mass-critical nonlinear Schrödinger equation

$$iu_t + \Delta u = \mu|u|^{4/d}u =: F(u). \tag{1-1}$$

Here, $\mu = \pm 1$, with $\mu = +1$ known as the *defocusing* and $\mu = -1$ as the *focusing* case. The name “mass-critical” refers to the fact that the scaling symmetry

$$u(t, x) = \lambda^{d/2}u(\lambda^2 t, \lambda x) \tag{1-2}$$

leaves both the equation and the mass invariant. Here the mass is defined as

$$M(u(t)) = \int_{\mathbb{R}^d} |u(t, x)|^2 dx = M(u_0). \tag{1-3}$$

The precise meaning of the solution we discuss throughout the paper is the following:

Definition 1.1 (solution). A function $u : I \times \mathbb{R}^d \rightarrow \mathbb{C}$ on a nonempty time interval $I \subset \mathbb{R}$ is a strong $L_x^2(\mathbb{R}^d)$ solution (or solution for short) if it lies in the class $C_t^0 L_x^2(K \times \mathbb{R}^d) \cap L_{t,x}^{2(d+2)/d}(K \times \mathbb{R}^d)$ for all compact $K \subset I$, and we have the Duhamel formula

$$u(t_1) = e^{i(t_1-t_0)\Delta}u(t_0) - i \int_{t_0}^{t_1} e^{i(t_1-t)\Delta} F(u(t)) dt \tag{1-4}$$

MSC2000: 35Q55.

Keywords: Schrödinger equation, mass-critical.

for all $t_0, t_1 \in I$. Here $e^{it\Delta}$ is the propagator for free Schrödinger equation. We say that u is a maximal-lifespan solution if the solution can not be extended to any strictly larger interval. We say that u is global if $I = \mathbb{R}$.

The standard local theory for such solutions was worked out by Cazenave and Weissler [2003]. They constructed the local in time solution for arbitrary initial data in $L_x^2(\mathbb{R}^d)$. They also showed that the solution depends continuously on the initial data in the same space. However due to the criticality of the problem, the lifespan of the local solution depends on the profile of the initial data instead of the mere L_x^2 -norm. When the initial data is small enough, they proved the solution exists globally and scatters in the following sense: there exist unique $u_{\pm} \in L_x^2(\mathbb{R}^d)$ such that

$$\lim_{t \rightarrow \infty} \|u(t) - e^{it\Delta}u_+\|_{L_x^2} = \lim_{t \rightarrow -\infty} \|u(t) - e^{it\Delta}u_-\|_{L_x^2} = 0. \tag{1-5}$$

Whilst the local theory is fairly complete, the understanding of the global theory for large solutions is still only partial. Briefly speaking, the global theory for large solutions amounts to proving the global wellposedness and scattering for generic L_x^2 initial data in the defocusing case; investigating the long time behavior of global solutions, characterizing the structure and profile of finite time blowup solutions in the focusing case and so on. In recent years, by using concentration compactness tools developed and used in [Merle 1993; Kenig and Merle 2006; Keraani 2001; 2006; Bégout and Vargas 2007; Killip et al. 2008; 2009a; 2009b; Li and Zhang 2009b; 2009a], one can address part of these problems by exploring the properties of a large class of solutions which have certain compactness properties. To this end, following [Tao et al. 2008], we introduce:

Definition 1.2 (almost periodic modulo symmetry solutions). Let u be the maximal-lifespan solution of (1-1) on time interval I . Let $I_0 \subset I$ be a subinterval. We say u is *almost periodic modulo symmetries* on I_0 if there exists functions $x(t), N(t), \zeta(t), \theta(t)$ with $t \in I_0$ such that the orbit

$$\left\{ e^{i\theta(t)} e^{ix \cdot \zeta(t)} N(t)^{-d/2} u\left(t, \frac{x-x(t)}{N(t)}\right), t \in I_0 \right\}$$

is precompact in $L_x^2(\mathbb{R}^d)$. By the Arzelà–Ascoli Theorem, an equivalent way to write this definition is the following: there exists a function $C(\eta)$ such that for any $\eta > 0$,

$$\int_{|x-x(t)| > C(\eta)/N(t)} |u(t, x)|^2 dx \leq \eta, \quad \int_{|\zeta-\zeta(t)| > C(\eta)N(t)} |\hat{u}(t, \zeta)|^2 d\zeta \leq \eta.$$

In particular, we call u is *almost periodic modulo scaling* on I_0 if, in this situation, $x(t) = \zeta(t) \equiv 0$ for all $t \in I_0$.

The parameter $N(t)$ is the frequency scale. In the physical space, its reciprocal corresponds to the concentration size of the solution. The parameter $x(t), \zeta(t)$ correspond to the center of mass at physical and frequency spaces respectively. Basically we have no a priori control on these parameters, which is the main source of the difficulty of establishing useful properties for almost periodic modulo symmetry solutions. However, under the spherical symmetry assumption, one is allowed to fix the center of mass, thus leaving only one parameter $N(t)$ which can still vary arbitrarily. This case turns out to be treatable in high dimensions $d \geq 4$. Here is the main theorem of this paper:

Theorem 1.3. *Let $d \geq 4$. Let u be a maximal-lifespan solution on I and is spherically symmetric. Suppose u is almost periodic modulo scaling on I . Then there exists $\varepsilon = \varepsilon(d) < 4/d$ such that*

$$u(t) \in H_x^{1+\varepsilon} \quad \text{for all } t \in I. \quad (1-6)$$

Moreover, the kinetic energy of the solution is localized uniformly in time: for any $\eta > 0$, there exists $C(\eta)$ such that for any $t \in I$

$$\int_{|x| \geq C(\eta)} |\nabla u(t, x)|^2 dx \leq \eta. \quad (1-7)$$

Here, ε only depends on the dimension d , while $C(\eta)$ depends also on the solution u .

Remark 1.4. This result seems a bit surprising in view of the fact that the scaling parameter $N(t)$ can vary arbitrarily and the solution is only assumed to be in the scale invariant space L_x^2 . On the other hand, [Theorem 1.3](#) bears similarities with previous works [[Killip et al. 2008](#); [2009a](#); [2009b](#); [Li and Zhang 2009b](#)], where they were able to deal with dimensions two and higher. However in [[Killip et al. 2009a](#); [Li and Zhang 2009b](#)], the solution is assumed to have H_x^1 regularity and this latter fact allows one to treat solutions being almost periodic modulo scaling in only one time direction. In [[Killip et al. 2008](#); [2009b](#)], the additional regularity is only established for three typical solutions known as *three enemies*. Namely, these are *almost periodic modulo scaling* solutions with a priori control on $N(t)$:

- (a) The self-similar solution. This solution is defined on maximal time interval $(0, \infty)$ and $N(t) = t^{-1/2}$ for any $t \in (0, \infty)$.
- (b) The soliton-like solution. This solution is global and $N(t) = 1$.
- (c) The high to low cascade. This solution is also global with $N(t)$ satisfying the conditions $N(t) \leq 1$ and $\liminf_{t \rightarrow \pm\infty} N(t) = 0$.

On the other hand, the technique in this paper allows us to deal with all enemies *with no a priori assumption* on $N(t)$ in dimensions $d \geq 4$.

Remark 1.5. The dependence on the dimension comes from the fact that in dimension $d \geq 4$, the non-linearity $|u|^{4/d}u$ can be put in Lebesgue space $L_x^p(\mathbb{R}^d)$ for some $p \geq 1$ only knowing that $u \in L_x^2(\mathbb{R}^d)$. This property is not available in low dimensions $d = 2, 3$. So in these dimensions, it is still open proving the additional regularity for solutions other than the three enemies.

Remark 1.6. Besides the spherical symmetry, we can also consider other symmetries that can freeze the center of mass at the origin. For example, one can consider the splitting spherical symmetry introduced in [[Li and Zhang 2009b](#)]. In [[Li and Zhang 2009a](#)], we select the six dimensions as a sample case to show how the technique can be extended to deal with the solution with splitting spherical symmetry and is almost periodic modulo scaling. There the main difficulty comes from the fact that the waves can propagate anisotropically along splitting subspaces. As shown in the proofs of [Proposition 4.4](#) and [Proposition 4.6](#), the spherical symmetry is mainly used to treat the part where the plane waves travel away from the origin. For this part, one uses the weighted Strichartz estimate for radial functions to get the decay. In the splittingly spherical symmetric case, we develop tools such as weighted Strichartz estimate (see [[Li and Zhang 2009b](#)]) for splittingly spherical symmetric functions to make use of the decay property.

Remark 1.7. To prove [Theorem 1.3](#) we need to control the parts of the solution both near the spatial origin and away from it. To control the part away from the origin, we use the techniques from [\[Killip et al. 2009a\]](#) where we need the radial assumption on the solution. To control the part near the origin, we introduce a novel *local iteration* scheme which actually does not need the radial assumption provided we already have the control on the piece away from the origin. We should also stress that our proof uses the almost periodicity in a very light way. Instead of assuming the solution is almost periodic modulo scaling on the whole time interval, one could assume the following *sequential almost periodicity*: there exist $t_n^+ \rightarrow \sup I, t_n^- \rightarrow \inf I$ and scaling parameters $N(t_n^+), N(t_n^-)$, such that both of the sets

$$\{N(t_n^+)^{-d/2}u(t_n^+, \cdot / N(t_n^+))\}, \{N(t_n^-)^{-d/2}u(t_n^-, \cdot / N(t_n^-))\}$$

are precompact in $L_x^2(\mathbb{R}^d)$.

Applications of Theorem 1.3. The applications of [Theorem 1.3](#) are related to the scattering conjecture and the rigidity conjecture which we now explain. In the defocusing case, the scattering conjecture says that all solutions with finite mass exist globally and scatter in both time directions. In the focusing case, besides scattering solutions, there exist finite time blowup solutions as shown in [\[Glassey 1977\]](#) and the solitary wave solutions of the form $e^{it}R(x)$. Here R solves the elliptic equation

$$\Delta R - R + |R|^{4/d}R = 0.$$

There are infinitely many solutions to this equation, but only one positive solution which is spherically symmetric (up to translations) and whose mass is minimal among all these R 's. This solution is usually called the ground state:

Definition 1.8 (ground state [\[Berestycki and Lions 1979; Kwong 1989\]](#)). The ground state Q refers to the unique positive radial Schwartz solution to the elliptic equation

$$\Delta Q - Q + |Q|^{4/d}Q = 0.$$

It is believed that the mass of Q serves as the minimal mass among all the nonscattering solutions in the focusing case. To summarize, we have:

Conjecture 1.9 (scattering conjecture). *Let $u_0 \in L_x^2(\mathbb{R}^d)$. In the focusing case, we also assume $M(u_0) < M(Q)$. Then the corresponding solution to (1-1) exists globally and scatters in both time directions.*

This conjecture has been proved in dimensions $d \geq 2$ when the initial data is spherically symmetric; see [\[Killip et al. 2008; 2009b\]](#).¹ We now give a high level overview of the proof which is based on a contradiction argument. Assuming the scattering conjecture is not true, one can then use concentration compactness tools to obtain minimal mass nonscattering² solutions which are almost periodic modulo scaling (due to the spherical symmetry) with scaling parameter $N(t)$. To obtain better control of $N(t)$, another limiting procedure is performed to reduce the consideration to three typical solutions alluded as to “three enemies”. To kill three enemies and thereby obtaining the contradiction, one can use the

¹In the defocusing case and $d \geq 3$, one can take advantage of Morawetz estimate to prove the additional regularity; see [\[Tao et al. 2007\]](#) for more details.

²Here by “nonscattering”, we mean that the $L_{t,x}^{2(d+2)/d}$ norm of the solution is infinite. Obviously, a “nonscattering” solution may blow up at finite time or exist globally with infinite $L_{t,x}^{2(d+2)/d}$ norm.

information of $N(t)$ to obtain additional regularity of these solutions which together with a truncated virial argument establishes the claim.

Thanks to [Theorem 1.3](#), we can simplify the argument by directly working with all enemies whose scaling parameter $N(t)$ can vary arbitrarily in dimensions $d \geq 4$. In other words, the limiting procedure of picking three enemies is not needed here. We record the result as:

Corollary 1.10 (scattering in dimension $d \geq 4$ with spherical symmetry). *Let $d \geq 4$. Let $u_0 \in L_x^2(\mathbb{R}^d)$ be spherically symmetric. In the focusing case, we assume $M(u_0) < M(Q)$. Then the solution to (1-1) with this initial data exists globally and satisfies*

$$\|u\|_{L_{t,x}^{2(d+2)/d}(\mathbb{R} \times \mathbb{R}^d)} \leq C(\|u_0\|_{L_x^2}).$$

We turn now to the rigidity conjecture.

In the focusing case, a main issue is to understand the large time behavior of nonscattering solutions. This problem has only been addressed in the case when the mass of u is equal to or slightly bigger than that of the ground state; see [[Merle 1993](#); [Merle and Raphael 2005](#); [Killip et al. 2009a](#); [Li and Zhang 2009b](#)] and the references therein. In this paper, we are primarily concerned with the case when the solution has the ground state mass. Our main focus is to characterize and classify all such solutions. At the level of ground state mass, there are two explicit examples of nonscattering solutions: the solitary wave SW which exists globally and the pseudoconformal ground state $\text{Pc}(Q)$ which blows up at $t = 0$:

$$\text{SW} = e^{it} Q(x), \quad \text{Pc}(Q) = |t|^{-d/2} e^{(t|x|^2 - 4)/(4t)} Q\left(\frac{x}{t}\right).$$

It is conjectured that, up to symmetries, these are the only two threshold solutions for scattering at the level of minimal mass. Associated with this is the following rigidity conjecture which identifies all solutions with ground state mass as either SW or $\text{Pc}(Q)$ if they do not scatter. Since both mass and the equation are invariant under a couple of symmetries, the coincidence of the solutions with the examples only hold modulo these symmetries. Specifically, the symmetries are: translation, phase rotation, scaling and the Galilean boost.

Conjecture 1.11 (rigidity conjecture at the ground state mass). *Let $u_0 \in L_x^2(\mathbb{R}^d)$ satisfy $M(u_0) = M(Q)$. Then only the following cases can occur:*

- (1) *The solution u blows up at finite time, then in this case u must coincide with $\text{Pc}(Q)$ up to symmetries of the equation.*
- (2) *The solution u is a global solution. Then in this case, u either scatters in both time directions or u must coincide with SW up to symmetries of the equation.*

[Merle \[1993\]](#) considered the first part of the conjecture, where he identified all finite time blowup solutions as $\text{Pc}(Q)$ under an additional H_x^1 assumption on the initial data. See also [[Weinstein 1986](#)] for the preliminary result and [[Hmidi and Keraani 2005](#)] for a simplified proof of Merle's argument. By Merle's result and pseudoconformal transformation, the second part of the conjecture, which characterizes all global solutions with ground state mass, still holds if we make the strong assumption that the initial data $u_0 \in \Sigma = \{f \in H_x^1, xf \in L_x^2\}$. Finally it is worthwhile noticing that Merle's argument works for all dimensions without any symmetry assumption on the initial data.

Without the Σ assumption on the initial data, it is not clear at all how to deal with the case when u_0 is merely in L_x^2 and the corresponding solution is global. Recently in [Killip et al. 2009a; Li and Zhang 2009b], we proved the second part of the conjecture when the initial data $u_0 \in H_x^1(\mathbb{R}^d)$, $d \geq 2$ and is spherically symmetric. In dimensions $d \geq 4$, the results hold even under a weaker symmetry assumption, namely, the initial data is only required to be splitting-spherical symmetric (see [Li and Zhang 2009b] for more details).

As stated, all the results concerning the rigidity conjecture require the H_x^1 regularity on the initial data since it is the minimal regularity to define the energy and to carry out the spectral analysis. Here the energy refers to

$$E(u(t)) = \frac{1}{2} \|\nabla u(t)\|_{L_x^2}^2 - \frac{d}{2(d+2)} \|u(t)\|_{L_x^{2(d+2)/d}}^{2(d+2)/d} = E(u_0).$$

To prove the rigidity results for pure L_x^2 solutions, a reasonable strategy is to upgrade the regularity of the solution to H_x^1 or better by taking advantage of certain compactness properties of the solutions. This is where Theorem 1.3 has to be used. We can then use known H_x^1 results to classify these solutions. Therefore as a direct consequence of Theorem 1.3, we have:

Theorem 1.12 (rigidity for two-way nonscattering solutions with ground state mass). *Let $d \geq 4$. Let $u_0 \in L_x^2(\mathbb{R}^d)$ be spherically symmetric and $M(u_0) = M(Q)$. Let u be the maximal lifespan solution on I which does not scatter on both sides:*

$$\|u\|_{L_{t,x}^{2(d+2)/d}([t_0, \sup I) \times \mathbb{R}^d)} = \|u\|_{L_{t,x}^{2(d+2)/d}((\inf I, t_0] \times \mathbb{R}^d)} = \infty, \quad t_0 \in I.$$

Then $I = \mathbb{R}$ and $u = e^{it} Q$ up to phase rotation and scaling.

For technical reasons, we need to impose the condition that the solution does not scatter in both time directions. It is an interesting problem to extend our techniques to the case when the solution scatters only in one time direction, but does not scatter in the other.

We give the proof of these two results in Section 3. Now we briefly sketch the proof of Theorem 1.3.

Main idea of the proof of Theorem 1.3: a local iteration scheme. We will work with each single dyadic frequency of u :

$$\|P_N u(t)\|_{L_x^2}.$$

The decay in N will correspond to the regularity of the solution. First we observe that when restricted to the region away from the origin, the argument in [Killip et al. 2009a] gives us

$$\|\phi_{>1} P_N u(t)\|_{L_x^2} \lesssim N^{-1-\varepsilon} \tag{1-8}$$

with a uniform in time bound. Here $\phi_{>1}$ is a smooth cut-off function supported in the region $|x| > 1$. This reduces matters to estimating the part of the solution near the spatial origin, that is, $\|\phi_{\leq 1} P_N u(t)\|_{L_x^2}$. This piece is trivially bounded by

$$A_N = \|P_N u\|_{S([t, t+1/\sqrt{N}])},$$

that is, the Strichartz norm of $P_N u$ on a local time interval $[t, t + 1/\sqrt{N}]$. It turns out, after some technical manipulations, that this latter quantity is better suited for iteration and bootstrapping. Indeed

we shall establish recurrent relations for A_N and we will iterate our estimates only finitely many (but sufficiently many) steps. The crucial point is that during the iteration process, we shall never need more than the information of the solution on a unit time interval $[t, t + 1]$. Therefore we do not need to use the full control of $N(t)$. We remark that although as a sacrifice the $H_x^{1+\varepsilon}$ norm of $u(t)$ depends on t , this information combined with the kinetic energy localization in [Section 3](#) suffice to prove [Corollary 1.10](#) and [Theorem 1.12](#).

2. Preliminaries

Some notation. We write $X \lesssim Y$ or $Y \gtrsim X$ to indicate $X \leq CY$ for some constant $C > 0$. We use $O(Y)$ to denote any quantity X such that $|X| \lesssim Y$. We use the notation $X \sim Y$ whenever $X \lesssim Y \lesssim X$. The fact that these constants depend upon the dimension d will be suppressed. If C depends upon some additional parameters, we will indicate this with subscripts; for example, $X \lesssim_u Y$ denotes the assertion that $X \leq C_u Y$ for some C_u depending on u . Sometimes when the context is clear, we will suppress the dependence on u and write $X \lesssim_u Y$ as $X \lesssim Y$. We will write $C = C(Y_1, \dots, Y_n)$ to stress that the constant C depends on quantities Y_1, \dots, Y_n . We denote by $X \pm$ any quantity of the form $X \pm \varepsilon$ for any $\varepsilon > 0$.

We use the ‘‘Japanese bracket’’ convention: $\langle x \rangle := (1 + |x|^2)^{1/2}$.

We write $L_t^q L_x^r$ to denote the Banach space with norm

$$\|u\|_{L_t^q L_x^r(\mathbb{R} \times \mathbb{R}^d)} := \left(\int_{\mathbb{R}} \left(\int_{\mathbb{R}^d} |u(t, x)|^r dx \right)^{q/r} dt \right)^{1/q},$$

with the usual modifications when q or r are equal to infinity, or when the domain $\mathbb{R} \times \mathbb{R}^d$ is replaced by a smaller region of spacetime such as $I \times \mathbb{R}^d$. When $q = r$ we abbreviate $L_t^q L_x^q$ as $L_{t,x}^q$.

Throughout this paper, we will use $\phi \in C^\infty(\mathbb{R}^d)$ for a radial bump function supported in the ball $\{x \in \mathbb{R}^d : |x| \leq 25/24\}$ and equal to 1 on the ball $\{x \in \mathbb{R}^d : |x| \leq 1\}$. For any constant $C > 0$, we set $\phi_{\leq C}(x) := \phi(x/C)$ and $\phi_{> C} := 1 - \phi_{\leq C}$.

Basic harmonic analysis. For each number $N > 0$, we define the Fourier multipliers

$$\begin{aligned} \widehat{P_{\leq N} f}(\xi) &:= \phi_{\leq N}(\xi) \hat{f}(\xi), \\ \widehat{P_{> N} f}(\xi) &:= \phi_{> N}(\xi) \hat{f}(\xi), \\ \widehat{P_N f}(\xi) &:= (\phi_{\leq N} - \phi_{\leq N/2})(\xi) \hat{f}(\xi), \end{aligned}$$

and similarly $P_{< N}$ and $P_{\geq N}$. We also define

$$P_{M < \dots \leq N} := P_{\leq N} - P_{\leq M} = \sum_{M < N' \leq N} P_{N'}$$

whenever $M < N$. We will usually use these multipliers when M and N are *dyadic numbers* (that is, of the form 2^n for some integer n); in particular, all summations over N or M are understood to be over dyadic numbers. Nevertheless, it will occasionally be convenient to allow M and N not to be powers of 2. Since P_N is not truly a projection ($P_N^2 \neq P_N$), we will occasionally need to use fattened Littlewood–Paley

operators:

$$\tilde{P}_N := P_{N/2} + P_N + P_{2N}. \quad (2-1)$$

These obey $P_N \tilde{P}_N = \tilde{P}_N P_N = P_N$.

Like all Fourier multipliers, the Littlewood–Paley operators commute with the propagator $e^{it\Delta}$, as well as with differential operators such as $i\partial_t + \Delta$. We will use basic properties of these operators many times, including:

Lemma 2.1 (Bernstein estimates). *For $1 \leq p \leq q \leq \infty$,*

$$\begin{aligned} \left\| |\nabla|^{\pm s} P_N f \right\|_{L_x^p(\mathbb{R}^d)} &\sim N^{\pm s} \|P_N f\|_{L_x^p(\mathbb{R}^d)}, \\ \|P_{\leq N} f\|_{L_x^q(\mathbb{R}^d)} &\lesssim N^{d/p-d/q} \|P_{\leq N} f\|_{L_x^p(\mathbb{R}^d)}, \\ \|P_N f\|_{L_x^q(\mathbb{R}^d)} &\lesssim N^{d/p-d/q} \|P_N f\|_{L_x^p(\mathbb{R}^d)}. \end{aligned}$$

While it is true that spatial cutoffs do not commute with Littlewood–Paley operators, we still have the following:

Lemma 2.2 (mismatch estimates in real space). *Let $R, N > 0$. Then*

$$\begin{aligned} \left\| \phi_{>R} \nabla P_{\leq N} \phi_{\leq R/2} f \right\|_{L_x^p(\mathbb{R}^d)} &\lesssim_m N^{1-m} R^{-m} \|f\|_{L_x^p(\mathbb{R}^d)}, \\ \left\| \phi_{>R} P_{\leq N} \phi_{\leq R/2} f \right\|_{L_x^p(\mathbb{R}^d)} &\lesssim_m N^{-m} R^{-m} \|f\|_{L_x^p(\mathbb{R}^d)} \end{aligned}$$

for any $1 \leq p \leq \infty$ and $m \geq 0$.

Proof. We will only prove the first inequality; the second follows similarly.

It is not hard to obtain kernel estimates for the operator $\phi_{>R} \nabla P_{\leq N} \phi_{\leq R/2}$. Indeed, an exercise in nonstationary phase shows

$$\left| \phi_{>R} \nabla P_{\leq N} \phi_{\leq R/2}(x, y) \right| \lesssim N^{d+1-2k} |x-y|^{-2k} \phi_{|x-y|>R/2}$$

for any $k \geq 0$. An application of Young’s inequality yields the claim. \square

Similar estimates hold when the roles of the frequency and physical spaces are interchanged. The proof is easiest when working on L_x^2 , which is the case we will need; nevertheless, the following statement holds on L_x^p for any $1 \leq p \leq \infty$.

Lemma 2.3 (mismatch estimates in frequency space). *For $R > 0$ and $N, M > 0$ such that $\max\{N, M\} \geq 4 \min\{N, M\}$,*

$$\begin{aligned} \left\| P_N \phi_{\leq R} P_M f \right\|_{L_x^2(\mathbb{R}^d)} &\lesssim_m \max\{N, M\}^{-m} R^{-m} \|f\|_{L_x^2(\mathbb{R}^d)}, \\ \left\| P_N \phi_{\leq R} \nabla P_M f \right\|_{L_x^2(\mathbb{R}^d)} &\lesssim_m M \max\{N, M\}^{-m} R^{-m} \|f\|_{L_x^2(\mathbb{R}^d)} \end{aligned}$$

for any $m \geq 0$. The same estimates hold if we replace $\phi_{\leq R}$ by $\phi_{>R}$.

Proof. The first claim follows from Plancherel’s Theorem and [Lemma 2.2](#) and its adjoint. To obtain the second claim from this, we write

$$P_N \phi_{\leq R} \nabla P_M = P_N \phi_{\leq R} P_M \nabla \tilde{P}_M$$

and note that $\|\nabla \tilde{P}_M\|_{L_x^2 \rightarrow L_x^2} \lesssim M$. \square

Some analysis tools. We will need the following radial Sobolev embedding to exploit the decay property of a radial function. For the proof and the more complete version, see [Tao et al. 2007].

Lemma 2.4 (radial Sobolev embedding [Tao et al. 2007]). *Let the dimension d be at least 2. Let $s > 0$, $\alpha > 0$ and $1 < p, q < \infty$ obey the scaling restriction $\alpha + s = d(1/q - 1/p)$. Then the following holds:*

$$\||x|^\alpha f\|_{L^p(\mathbb{R}^d)} \lesssim \||\nabla|^s f\|_{L^q(\mathbb{R}^d)},$$

where the implicit constant depends on s, α, p, q . When $p = \infty$, we have

$$\||x|^{(d-1)/2} P_N f\|_{L^\infty(\mathbb{R}^d)} \lesssim N^{1/2} \|P_N f\|_{L_x^2(\mathbb{R}^d)}.$$

We will need the following fractional chain rule lemma.

Lemma 2.5 (fractional chain rule for a C^1 function [Christ and Weinstein 1991; Staffilani 1997; Taylor 2000]). *Let $G \in C^1(\mathbb{C})$, $\sigma \in (0, 1)$, and $1 < r, r_1, r_2 < \infty$ such that $1/r = 1/r_1 + 1/r_2$. Then we have*

$$\||\nabla|^\sigma G(u)\|_r \lesssim \|G'(u)\|_{r_1} \||\nabla|^\sigma u\|_{r_2}.$$

Proof. See [Christ and Weinstein 1991; Staffilani 1997; Taylor 2000]. □

Lemma 2.6 [Killip et al. 2008]. *Let $0 < s < 1 + 4/d$ and $F(u) = |u|^{4/d}u$. Then*

$$\||\nabla|^s F(u)\|_{L_x^{(2(d+2))/(d+4)}} \lesssim \||\nabla|^s u\|_{L_x^{2(d+2)/d}} \|u\|_{L_x^{2(d+2)/d}}^{4/d}.$$

We will need the following sharp Gagliardo–Nirenberg inequality:

Lemma 2.7 [Weinstein 1983]. *Let Q be the ground state in the Definition 1.8. Then for any $f \in H_x^1(\mathbb{R}^d)$, we have*

$$\|f\|_{L_x^{2(d+2)/d}}^{2(d+2)/d} \leq \frac{d+2}{d} \left(\frac{M(f)}{M(Q)} \right)^{2/d} \|\nabla f\|_{L_x^2}^2. \quad (2-2)$$

The equality holds only and if only

$$f = ce^{i\theta} \lambda^{d/2} Q(\lambda(x - x_0)) \quad (2-3)$$

for $(c, \theta, \lambda) \in (\mathbb{R}^+, \mathbb{R}, \mathbb{R}^+)$.

Strichartz estimates. The free Schrödinger flow has the explicit expression

$$e^{it\Delta} f(x) = \frac{1}{(4\pi t)^{d/2}} \int_{\mathbb{R}^d} e^{i|x-y|^2/4t} f(y) dy,$$

from which we can derive the kernel estimate of the frequency localized propagator.

Lemma 2.8 (kernel estimates [Killip et al. 2008; 2009b]). *For any $m \geq 0$, we have*

$$|(P_N e^{it\Delta}(x, y))| \lesssim_m \begin{cases} |t|^{-d/2} & \text{if } |x - y| \sim Nt, \\ \frac{N^d}{|N^2 t|^m \langle N|x - y| \rangle^m} & \text{otherwise,} \end{cases}$$

for $|t| \geq N^{-2}$ and

$$|(P_N e^{it\Delta})(x, y)| \lesssim_m N^d \langle N|x - y| \rangle^{-m} \quad \text{for } |t| \leq N^{-2}.$$

We will frequently use the standard Strichartz estimate. Let $d \geq 3$. Let I be a time interval. We define the Strichartz space on I :

$$S(I) = L_t^\infty L_x^2(I \times \mathbb{R}^d) \cap L_t^2 L_x^{2d/(d-2)}(I \times \mathbb{R}^d).$$

We also define $N(I)$ to be $L_t^1 L_x^2(I \times \mathbb{R}^d) + L_t^2 L_x^{2d/(d+2)}(I \times \mathbb{R}^d)$. Then the standard Strichartz estimate reads:

Lemma 2.9 (Strichartz). *Let $d \geq 3$. Let I be an interval, $t_0 \in I$, and let $u_0 \in L_x^2(\mathbb{R}^d)$ and $f \in N(I)$. Then, the function u defined by*

$$u(t) := e^{i(t-t_0)\Delta} u_0 - i \int_{t_0}^t e^{i(t-t')\Delta} f(t') dt'$$

obeys the estimate

$$\|u\|_{S(I)} \lesssim \|u_0\|_{L_x^2} + \|f\|_{N(I)},$$

where all spacetime norms are over $I \times \mathbb{R}^d$.

Proof. See, for example, [Ginibre and Velo 1992; Strichartz 1977]. For the endpoint see [Keel and Tao 1998]. □

We will also need a weighted Strichartz estimate, which exploits heavily the spherical symmetry in order to obtain spatial decay.

Lemma 2.10 (weighted Strichartz [Killip et al. 2008; 2009b]). *Let I be an interval, $t_0 \in I$, and let $F : I \times \mathbb{R}^d \rightarrow \mathbb{C}$ be spherically symmetric. Then,*

$$\left\| \int_{t_0}^t e^{i(t-t')\Delta} F(t') dt' \right\|_{L_x^2} \lesssim \left\| |x|^{-2(d-1)/q} F \right\|_{L_t^{q/(q-1)} L_x^{2q/(q+4)}(I \times \mathbb{R}^d)}$$

for all $4 \leq q \leq \infty$.

The in/out decomposition. We will need an incoming/outgoing decomposition; we will use the one developed in [Killip et al. 2008; 2009b]. As there, we define operators P^\pm by

$$[P^\pm f](r) := \frac{1}{2} f(r) \pm \frac{i}{\pi} \int_0^\infty \frac{r^{2-d} f(\rho) \rho^{d-1} d\rho}{r^2 - \rho^2},$$

where the radial function $f : \mathbb{R}^d \rightarrow \mathbb{C}$ is written as a function of radius only. We will denote by P^+ the projection onto outgoing spherical waves; however, it is not a true projection as it is neither idempotent nor self-adjoint. Similarly, P^- plays the role of a projection onto incoming spherical waves; its kernel is the complex conjugate of the kernel of P^+ as required by time-reversal symmetry.

For $N > 0$ let P_N^\pm denote the product $P^\pm P_N$, where P_N is the Littlewood–Paley projection. We record the following properties of P^\pm :

Proposition 2.11 (properties of P^\pm [Killip et al. 2008; 2009b]).

- (i) $P^+ + P^-$ represents the projection from L^2 onto L_{rad}^2 .
- (ii) Fix $N > 0$. Then

$$\|\chi_{\geq 1/N} P_{\geq N}^\pm f\|_{L^2(\mathbb{R}^d)} \lesssim \|f\|_{L^2(\mathbb{R}^d)}$$

with an N -independent constant.

(iii) If the dimension $d = 2$, then P^\pm are bounded on $L^2(\mathbb{R}^2)$.

(iv) For $|x| \gtrsim N^{-1}$ and $t \gtrsim N^{-2}$, the integral kernel obeys

$$|[P_N^\pm e^{\mp it\Delta}](x, y)| \lesssim \begin{cases} (|x||y|)^{-(d-1)/2} |t|^{-1/2} & |y| - |x| \sim Nt, \\ \frac{N^d}{(N|x|)^{(d-1)/2} (N|y|)^{(d-1)/2} \langle N^2t + N|x| - N|y| \rangle^{-m}} & \text{otherwise,} \end{cases}$$

for all $m \geq 0$.

(v) For $|x| \gtrsim N^{-1}$ and $|t| \lesssim N^{-2}$, the integral kernel obeys

$$|[P_N^\pm e^{\mp it\Delta}](x, y)| \lesssim \frac{N^d}{(N|x|)^{(d-1)/2} (N|y|)^{(d-1)/2} \langle N|x| - N|y| \rangle^{-m}}$$

for any $m \geq 0$.

3. Theorem 1.3 implies Corollary 1.10 and Theorem 1.12

In this section, we assume Theorem 1.3 holds momentarily and prove the scattering and the rigidity results Corollary 1.10 and Theorem 1.12.

Proof of Corollary 1.10. Suppose by contradiction that Corollary 1.10 does not hold. Then there exists minimal mass M_c for which $M_c < \infty$ in the defocusing case, $M_c < M(Q)$ in the focusing case and maximal-lifespan solution $u(t, x)$ on $I = (-T_*, T^*)$ such that

- (1) u is spherically symmetric and $M(u) = M_c$;
- (2) u is almost periodic modulo scaling on I .

See for instance [Tao et al. 2008] for this part of the argument which is by now standard. Applying Theorem 1.3, we know that $u \in H_x^{1+\varepsilon}$. We now detail the rest of the argument in the focusing case, since the defocusing case is even simpler. By the sharp Gagliardo–Nirenberg inequality and the fact that $M(u) < M(Q)$ we have

$$\|u(t)\|_{H_x^1} \lesssim_{M(u)} 1.$$

From this and the standard local theory in H_x^1 we know that u exists globally, that is, $T_* = T^* = \infty$. In this situation, the contradiction will come from the truncated virial and the kinetic energy localization as we explain now. Let $\phi_{\leq R}$ be the smooth cutoff function, we define the truncated virial as

$$V_R(t) = \int \phi_{\leq R}(x) |x|^2 |u(t, x)|^2 dx.$$

Obviously

$$V_R(t) \lesssim R^2 \quad \text{for all } t \in \mathbb{R}. \tag{3-1}$$

On the other hand, we compute the second derivative of virial with respect to t ; this gives

$$\partial_{tt} V_R(t) = 8E(u) + O\left(\int_{|x|>R} |\nabla u(t, x)|^2 + |u(t, x)|^{2(d+2)/d} + \frac{1}{R^2} \int_{|x|>R} |u(t, x)|^2 dx\right). \tag{3-2}$$

Since $M(u) < M(Q)$ and $u \in H_x^1$, from the sharp Gagliardo–Nirenberg inequality (2-2) we have

$$E(u) > 0.$$

Now we can use the kinetic energy localization (1-7) and the Gagliardo–Nirenberg inequality to control the $O(\cdot)$ term in (3-2) and finally get

$$\partial_{tt} V_R(t) \geq 4E(u) > 0$$

by taking R sufficiently large. This obviously contradicts (3-1), finishing the proof of Corollary 1.10. \square

Proof of Theorem 1.12. Let $d \geq 4$ and let u be the solution of (1-1) satisfying the following:

- (1) $M(u) = M(Q)$ and u is spherically symmetric.
- (2) u does not scatter in both time directions.

By [Killip et al. 2008] or Corollary 1.10, $M(Q)$ is the minimal mass, and the compactness argument in [Keraani 2006; Bégout and Vargas 2007; Tao et al. 2008] shows that u is *almost periodic modulo scaling* in both time directions. Now we can apply Theorem 1.3 to deduce that $u \in H_x^1$. Since from Merle’s result, the only finite-time blowup solution must be $Pc(Q)$ up to symmetries and $Pc(Q)$ scatters in one time direction, we know from condition (2) that u must be a global solution.

From (2-2), this global solution u satisfies $E(u) \geq 0$. Moreover, the same virial argument as in the proof of Corollary 1.10 precludes the case $E(u) > 0$, thus $E(u) = 0$. From here the coincidence of the solution with solitary wave follows immediately, again by the sharp Gagliardo–Nirenberg inequality. \square

4. The proof of Theorem 1.3

The proof of Theorem 1.3 proceeds in two steps. In the first step, we prove that away from the origin, the solution has $H_x^{1+\epsilon}$ regularity. Moreover, a similar (but more refined) argument establishes the spatial decay estimate. These two pieces together suffice for us to establish the kinetic localization estimate. However, in this step, the total kinetic energy does not need to be finite.

In the second step, we prove the total kinetic energy is actually finite by controlling the piece near the spatial origin. Thanks to the first step, we only need to consider a single frequency $P_N u$ with spatial cutoff $\phi_{\leq 1}$. We can bound this quantity by the Strichartz norm of $P_N u$ on a short time interval $[t, t + 1/\sqrt{N}]$. We then establish a recurrent relation for this local Strichartz norm. Iterating the estimates finitely many times then yields the desired bound. More details are given below.

Before proceeding, we remark that in all of the arguments that follow, the only property we use for an *almost periodic modulo scaling* solution is that it satisfies the improved Duhamel formula. This was first derived in [Tao et al. 2008].

Proposition 4.1 (improved Duhamel formula [Tao et al. 2008]). *Let u be the solution of (1-1) and is almost periodic modulo scaling on the time interval I . Then*

$$u(t) = \lim_{T \rightarrow \inf I} -i \int_T^t e^{i(t-\tau)\Delta} F(u(\tau)) d\tau = \lim_{T \rightarrow \sup I} i \int_t^T e^{i(t-\tau)\Delta} F(u(\tau)) d\tau. \tag{4-1}$$

Here the limit is in weak L_x^2 sense.

Remark 4.2. As was already mentioned in [Remark 1.7](#), we actually only need the *sequential almost periodicity* of the solution for the later proof to work. This would imply the following sequence version of improved Duhamel formula:

$$u(t) = \lim_{n \rightarrow \infty} -i \int_{T_n^-}^t e^{i(t-\tau)\Delta} F(u(\tau)) d\tau = \lim_{n \rightarrow \infty} i \int_t^{T_n^+} e^{i(t-\tau)\Delta} F(u(\tau)) d\tau.$$

Here again the limit is in weak L_x^2 sense.

In what follows, we shall only assume that

$$\left. \begin{array}{l} u \text{ is a maximal lifespan solution on } I; \\ u \text{ is spherically symmetric in space;} \\ u \text{ satisfies the improved Duhamel formula (4-1).} \end{array} \right\} \quad (4-2)$$

By time translation invariance and without loss of generality we also assume $[0, 1] \subset I$.

Localization for kinetic energy. The purpose of this section is to establish the uniform in time localization of the kinetic energy for solutions satisfying the conditions (4-2). More precisely, we will prove:

Proposition 4.3 (kinetic energy localization). *Suppose u satisfies (4-2). Then there exists a function $C(\eta)$ such that*

$$\|\phi_{>C(\eta)} \nabla u(t)\|_{L_x^2} \leq \eta \quad \text{for all } \eta > 0, t \in I.$$

As shown in the proof of [\[Li and Zhang 2009b, Theorems 1.14–1.15, page 31\]](#), [Proposition 4.3](#) will follow immediately from the following two propositions which concern the decay of each single frequency.

Proposition 4.4 (frequency decay estimate). *Suppose u satisfies (4-2). Let $\varepsilon = (d-1)/d$. Then for any $t \in I$ and $N \geq 1$, we have*

$$\|\phi_{>1} P_N u(t)\|_{L_x^2} \lesssim N^{-1-\varepsilon}. \quad (4-3)$$

Remark 4.5. The decay $N^{-1-(d-1)/d}$ may seem a bit surprising since the exponent $1+(d-1)/d$ is bigger than the regularity of the nonlinearity $1+4/d$ for dimension $d > 5$. However this is not contradictory since in (4-3) we are only considering the part of the solution away from the origin. In this regime the additional regularity of the solution comes from the smoothing effects of the Schrödinger equation and the radial symmetry. On the other hand for the part of the solution near the origin, we only obtain Sobolev regularity H^s for some $s < 1+4/d$ (see (4-24)).

Proposition 4.6 (spatial decay estimate). *Suppose u satisfies (4-2). Let N_0, N_1 be two dyadic numbers. Then there exist $R_0 = R_0(N_0, N_1)$ and $\delta = \delta(d)$ such that for all $R \geq R_0$, $N \in [N_0, N_1]$ and $t \in I$, we have*

$$\|\phi_{>R} P_N u(t)\|_{L_x^2} \lesssim R^{-\delta}.$$

The proofs of both propositions have been presented, in various forms, in [\[Killip et al. 2009a; Li and Zhang 2009b\]](#). We sketch the proofs here for the sake of completeness. The proof of [Proposition 4.3](#) will be skipped since it follows directly from [Proposition 4.4](#) and [Proposition 4.6](#).

Proof of Proposition 4.4. We first use the in/out decomposition and triangle inequality for the bound

$$\|\phi_{>1} P_N u(t)\|_2 \leq \|\phi_{>1} P_N^+ u(t)\|_2 + \|\phi_{>1} P_N^- u(t)\|_2.$$

Since the two terms give the same contribution, we only estimate, for instance, the outgoing piece. For this piece, we use the forward Duhamel formula. Moreover, we will split the integral into different time regimes and introduce the spatial cutoffs. We have

$$\begin{aligned} \|\phi_{>1} P_N^+ u(t)\|_2 &\lesssim \left\| \phi_{>1} P_N^+ \int_t^{\sup I} e^{i(t-s)\Delta} F(u(s)) ds \right\|_2 \\ &\lesssim \left\| \phi_{>1} P_N^+ \int_0^{\sup I-t} e^{-is\Delta} F(u(t+s)) d\tau \right\|_2 \\ &\lesssim \left\| \phi_{>1} P_N^+ \int_0^{1/N} e^{-is\Delta} \phi_{>1/2} F(u(t+s)) ds \right\|_2 \end{aligned} \tag{4-4}$$

$$+ \left\| \phi_{>1} P_N^+ \int_0^{1/N} e^{-is\Delta} \phi_{\leq 1/2} F(u(t+s)) ds \right\|_2 \tag{4-5}$$

$$+ \left\| \phi_{>1} P_N^+ \int_{1/N}^{\sup I-t} e^{-is\Delta} \phi_{>Ns/2} F(u(t+s)) ds \right\|_2 \tag{4-6}$$

$$+ \left\| \phi_{>1} P_N^+ \int_{1/N}^{\sup I-t} e^{-is\Delta} \phi_{\leq Ns/2} F(u(t+s)) ds \right\|_2. \tag{4-7}$$

The main contribution comes from (4-4) and (4-6). To estimate (4-4), we drop the bounded operator $\phi_{>1} P_N^+$ and commute the frequency cutoff \tilde{P}_N with the spatial cutoff $\phi_{>1}$ (this produces a harmless high order term by the mismatch estimate Lemma 2.3). Thus we have

$$(4-4) \lesssim \left\| \int_0^{1/N} e^{-is\Delta} \phi_{>1/2} P_{N/8 < \dots \leq 8N} F(\phi_{>1/4} u(t+s)) d \right\|_2 + N^{-10}. \tag{4-8}$$

We now use the weighted Strichartz lemma (Lemma 2.10) to estimate the last term:

$$(4-4) \lesssim \|P_{N/8 < \dots \leq 8N} F(\phi_{>1/4} u(t+s))\|_{L_s^{d/(d-1)} L_x^{2d/(d+4)}([0, 1/N])} + N^{-10} \lesssim N^{-(d-1)/d}.$$

The estimate of (4-6) follows in a similar way. Applying the mismatch estimate and weighted Strichartz inequality, we have

$$\begin{aligned} (4-6) &\lesssim \left\| \int_{1/N}^{\sup I-t} e^{-is\Delta} \phi_{>Ns/2} P_{N/8 < \dots \leq 8N} F(\phi_{>Ns/4} u(t+s)) ds \right\|_2 + N^{-10} \\ &\lesssim \|(Ns)^{-2(d-1)/d} P_{N/8 < \dots \leq 8N} F(\phi_{>Ns/4} u(t+s))\|_{L_s^{d/(d-1)} L_x^{2d/(d+4)}([1/N, \sup I-t])} + N^{-10} \\ &\lesssim N^{-2(d-1)/d} \|s^{-2(d-1)/d} F(\phi_{>Ns/4} u(t+s))\|_{L_x^{2d/(d+4)} L_s^{d/(d-1)}([1/N, \sup I-t])} + N^{-10} \\ &\lesssim N^{-(d-1)/d}. \end{aligned}$$

Finally we consider the contribution from the tail terms (4-7) and (4-5). Applying Proposition 2.11, we bound the kernel as follows:

$$|(\phi_{>1} P_N^+ e^{-is\Delta} \phi_{\leq 1/2})(x, y)| \lesssim N^{-9d} \langle N(x-y) \rangle^{-10d} \quad \text{for } 0 < s \leq \frac{1}{N},$$

$$|(\phi_{>1} P_N^+ e^{-is\Delta} \phi_{\leq Ns/2})(x, y)| \lesssim N^d \langle N^2 s \rangle^{-10d} \langle N(x-y) \rangle^{-10d} \lesssim N^{-9d} \langle N(x-y) \rangle^{-10d} \quad \text{for } s > \frac{1}{N}.$$

The desired decay then follows from the kernel estimate and a simple use of Young's inequality. Combining the estimates of these four pieces together, we obtain

$$\|\phi_{>1} P_N u(t)\|_{L_x^2} \lesssim N^{-(d-1)/d} \quad \text{for all } t \in I.$$

Moreover it is easy to check that, after notational change, the same analysis establishes

$$\|\phi_{>c} P_N u(t)\|_{L_x^2} \lesssim_c N^{-(d-1)/d} \quad \text{for all } t \in I. \quad (4-9)$$

This implies

$$\||\nabla|^{(d-1)/d-} (\phi_{>c} u(t))\|_{L_x^2} \lesssim_c 1 \quad \text{for all } t \in I. \quad (4-10)$$

Now we can upgrade the decay (4-9) by inserting (4-10) when we repeat the same argument as above. For example, using Bernstein and (4-10), the term (4-4) can be re-estimated as follows:

$$\begin{aligned} (4-4) &\lesssim \|P_{N/8 < \dots \leq 8N} F(\phi_{>1/4} u(t+s))\|_{L_x^{d/(d-1)} L_x^{2d/(d+4)}([0, 1/N])} \\ &\lesssim N^{-2d/(d-1)+} \||\nabla|^{(d-1)/d-} F(\phi_{>1/4} u(t+s))\|_{L_x^\infty L_x^{2d/(d+4)}([0, 1/N])} \\ &\lesssim N^{-2(d-1)/d+}. \end{aligned}$$

The same computation applies to (4-6), so we get

$$\|\phi_{>c} P_N u(t)\|_2 \lesssim_c N^{-2(d-1)/d+} \quad \text{for all } t \in I.$$

Another repetition of the argument yields (4-3) for $\varepsilon = (d-1)/d$. \square

The proof of Proposition 4.6 has the same spirit as the proof of Proposition 4.4. So here we only briefly sketch the proof.

Proof sketch of Proposition 4.6. Using the in/out decomposition, it suffices to consider the piece

$$\|\phi_{>R} P_N^+ u(t)\|_2,$$

for which we use the forward Duhamel formula to express $u(t)$. This further reduces our consideration to the integral

$$\|\phi_{>R} P_N^+ \int_0^{\sup I-t} e^{-is\Delta} F(u(t+s)) ds\|_2.$$

Now we split the time integral into regimes where $0 < s < R/(100N)$, and $s > R/(100N)$. For the small time regime, we insert the spatial cutoff $\phi_{>R/2}$ and $\phi_{\leq R/2}$. For the large time regime, we insert the spatial cutoff $\phi_{>Ns/2}$ and $\phi_{\leq Ns/2}$. As indicated in the proof of Proposition 4.4, the pieces with cutoff near the origin will give arbitrary decay in R by using the kernel estimate Proposition 2.11. The pieces with cutoff away from the origin can be dealt with by the weighted Strichartz estimate. The point here

is that since the frequencies are fixed in the dyadic interval $[N_0, N_1]$, we can take R sufficiently large to cancel any N dependent quantity. \square

Local iteration to prove H_x^{1+} regularity. In this part, we prove $u(0) = u_0 \in H_x^{1+}$. This amounts to showing $\|P_{\geq N}u_0\|_{L_x^2} \lesssim N^{-1-}$ for $N \geq 1$. Using [Proposition 4.4](#), we first show the quantity $\|P_{\geq N}u_0\|_{L_x^2}$ is determined by the dual Strichartz norm of the nonlinearity on the local time interval $[0, 1/\sqrt{N}]$.

Lemma 4.7. *Let u satisfy (4-2). Let $\varepsilon = (d-1)/d$. Then for any $N \geq 1$, we have*

$$\|P_{\geq N}u_0\|_{L_x^2} \leq C(d, \|u_0\|_{L_x^2}) \left(N^{-1-\varepsilon} + \|P_{\geq N}F(u)\|_{L_{t,x}^{2(d+2)/(d+4)}([0, 1/\sqrt{N}] \times \mathbb{R}^d)} \right). \quad (4-11)$$

Remark 4.8. Here the choice of the time interval cutoff at $N^{-1/2}$ is not special. Perhaps a more natural choice is $1/N$ since the solution propagates at speed N and one is localizing to spacial scale $O(1)$. This latter choice would also work for our iteration scheme.

Proof. Since by [Proposition 4.4](#), we have that $\|\phi_{>1}P_{\geq N}u_0\|_{L_x^2} \lesssim N^{-1-\varepsilon}$, we only need to estimate the piece $\|\phi_{\leq 1}P_{\geq N}u_0\|_{L_x^2}$. In the following, the implicit constants are allowed to depend on d and $\|u_0\|_{L_x^2}$. By the improved Duhamel formula we get

$$\begin{aligned} \|\phi_{\leq 1}P_{\geq N}u_0\|_{L_x^2} &\leq \|\phi_{\leq 1}P_{\geq N} \int_0^{\sup I} e^{-i\tau\Delta} F(u(\tau)) d\tau\|_{L_x^2} \\ &\leq \|\phi_{\leq 1}P_{\geq N} \int_0^{1/\sqrt{N}} e^{-i\tau\Delta} F(u(\tau)) d\tau\|_{L_x^2} \end{aligned} \quad (4-12)$$

$$+ \|\phi_{\leq 1}P_{\geq N} \int_{1/\sqrt{N}}^{\sup I} e^{-i\tau\Delta} \phi_{\leq N\tau/8} F(u(\tau)) d\tau\|_{L_x^2} \quad (4-13)$$

$$+ \|\phi_{\leq 1}P_{\geq N} \int_{1/\sqrt{N}}^{\sup I} e^{-i\tau\Delta} \phi_{>N\tau/8} F(u(\tau)) d\tau\|_{L_x^2}. \quad (4-14)$$

For (4-12), we use Strichartz to bound it by

$$\|P_{\geq N}F(u)\|_{L_{t,x}^{2(d+2)/(d+4)}([0, 1/\sqrt{N}] \times \mathbb{R}^d)}.$$

For (4-13), using the kernel estimate with $m = 10d$, we have

$$\begin{aligned} (4-13) &\leq \sum_{M \geq N} \left\| \phi_{\leq 1}P_M \int_{1/\sqrt{N}}^{\sup I} e^{-i\tau\Delta} \phi_{\leq N\tau/8} F(u(\tau)) d\tau \right\|_{L_x^2} \\ &\lesssim \sum_{M \geq N} M^{d-20d} \int_{1/\sqrt{N}}^{\sup I} \tau^{-10d} \|\langle M | \cdot \rangle^{-10d} * F(u)\|_{L_x^2} d\tau \\ &\lesssim \sum_{M \geq N} M^{d-20d} M^{(1/2)(10d-1)} \|F(u)\|_{L_\tau^\infty L_x^{2d/(d+4)}} \|\langle M | \cdot \rangle^{-10d}\|_{L_x^{d/(d-2)}} \\ &\lesssim \sum_{M \geq N} M^{(3/2)(1-10d)} \\ &\lesssim N^{-10}. \end{aligned}$$

For (4-14), by the triangle inequality, we have

$$(4-14) \lesssim \left\| P_{\geq N} \int_{1/\sqrt{N}}^{\sup I} e^{-i\tau\Delta} \phi_{>N\tau/8} F(u\phi_{>1/8})(\tau) d\tau \right\|_{L_x^2} \\ \lesssim \left\| P_{\geq N} \int_{1/\sqrt{N}}^{\sup I} e^{-i\tau\Delta} \phi_{>N\tau/8} P_{\leq N/8} F(u\phi_{>1/8})(\tau) d\tau \right\|_{L_x^2} \quad (4-15)$$

$$+ \left\| P_{\geq N} \int_{1/\sqrt{N}}^{\sup I} e^{-i\tau\Delta} \phi_{>N\tau/8} P_{>N/8} F(u\phi_{>1/8})(\tau) d\tau \right\|_{L_x^2} \quad (4-16)$$

For the term (4-15), we use the mismatch estimate Lemma 2.3 and Bernstein to bound it as

$$(4-15) \lesssim \int_{1/\sqrt{N}}^{\sup I} (N^2\tau)^{-10d} \|P_{\leq N/8} F(u\phi_{>1/8})\|_{L_x^2} d\tau \lesssim \int_{1/\sqrt{N}}^{\sup I} (N^2\tau)^{-10d} N^2 d\tau \lesssim N^{-5}.$$

For the term (4-16), we use weighted Strichartz to estimate and Proposition 4.3 to get

$$(4-16) \lesssim \|(N\tau)^{-2(d-1)/d} P_{>N/8} F(u\phi_{>1/8})\|_{L_\tau^{d/(d-1)} L_x^{2d/(d+4)}([1/\sqrt{N}, \sup I] \times \mathbb{R}^d)} \\ \lesssim N^{-2(d-1)/d} \|\tau^{-2(d-1)/d} P_{>N/8} F(u\phi_{>1/8})\|_{L_\tau^{d/(d-1)}([1/\sqrt{N}, \sup I])} \cdot N^{-1} \|\nabla P_{>N/8} F(u\phi_{>1/8})\|_{L_\tau^\infty L_x^{2d/(d+4)}} \\ \lesssim N^{-1-3(d-1)/(2d)}.$$

This finishes the proof of Lemma 4.7. \square

Now we further estimate the dual Strichartz norm of the nonlinearity.

Lemma 4.9 (dual Strichartz norm control). *Let u satisfy (4-2). Let $\beta > 0$, $N_0 \geq 1$, $N > (1/\beta)N_0$. Then for any $0 < s < 1 + 4/d$, we have*

$$\|P_{\geq N} F(u)\|_{L_{t,x}^{2(d+2)/(d+4)}([0, 1/\sqrt{N}] \times \mathbb{R}^d)} \\ \lesssim \|u\|_{S([0, 1/\sqrt{N}])}^{4/d} \sum_{M \leq \beta N} \left(\frac{M}{N}\right)^s \|P_M u\|_{S([0, 1/\sqrt{N}])} \\ + \|u_{>\beta N}\|_{S([0, 1/\sqrt{N}])} \left(N_0^{4/(d+2)} N^{-1/(d+2)} + \|u_{>N_0}\|_{L_\tau^\infty L_x^2}^{8/(d(d+2))} \|u_{>N_0}\|_{S([0, 1/\sqrt{N}])}^{4/(d+2)}\right). \quad (4-17)$$

Proof. By splitting u into low, medium and high frequencies, $u = u_{\leq N_0} + u_{N_0 < \dots \leq \beta N} + u_{>\beta N}$, we write

$$F(u) = F(u_{\leq \beta N}) + O(u_{>\beta N} |u_{\leq N_0}|^{4/d}) + O(u_{>\beta N} |u_{>N_0}|^{4/d}). \quad (4-18)$$

The contribution due to the first term can be estimated as follows. By using Lemma 2.6, we have

$$\|P_{\geq N} F(u_{\leq \beta N})\|_{L_{t,x}^{2(d+2)/(d+4)}([0, 1/\sqrt{N}] \times \mathbb{R}^d)} \\ \lesssim N^{-s} \| |\nabla|^s P_{\geq N} F(u_{\leq \beta N}) \|_{L_{t,x}^{2(d+2)/(d+4)}([0, 1/\sqrt{N}] \times \mathbb{R}^d)} \\ \lesssim N^{-s} \| |\nabla|^s u_{\leq \beta N} \|_{L_{t,x}^{2(d+2)/d}([0, 1/\sqrt{N}] \times \mathbb{R}^d)} \|u_{\leq \beta N}\|_{L_{t,x}^{2(d+2)/d}([0, 1/\sqrt{N}] \times \mathbb{R}^d)}^{4/d} \\ \lesssim \|u\|_{S([0, 1/\sqrt{N}])}^{4/d} \sum_{M \leq \beta N} \left(\frac{M}{N}\right)^s \|P_M u\|_{S([0, 1/\sqrt{N}])}.$$

For the contribution due to the second part of (4-18), we use Bernstein to get

$$\begin{aligned} \|u_{>\beta N} |u_{\leq N_0}|^{4/d} \|_{L_{t,x}^{2(d+2)/(d+4)}([0,1/\sqrt{N}] \times \mathbb{R}^d)} &\lesssim \|u_{>\beta N}\|_{L_{t,x}^{2(d+2)/d}([0,1/\sqrt{N}] \times \mathbb{R}^d)} \|u_{\leq N_0}\|_{L_{t,x}^{2(d+2)/d}([0,1/\sqrt{N}] \times \mathbb{R}^d)}^{4/d} \\ &\lesssim \|u_{>\beta N}\|_{S([0,1/\sqrt{N}])} N_0^{4/(d+2)} N^{-1/(d+2)} \|u_{\leq N_0}\|_{L_t^\infty L_x^2}^{4/d} \\ &\lesssim \|u_{>\beta N}\|_{S([0,1/\sqrt{N}])} N_0^{4/(d+2)} N^{-1/(d+2)}. \end{aligned}$$

For the third term in (4-18), we use Hölder and interpolation to get

$$\begin{aligned} \|u_{>\beta N} |u_{>N_0}|^{4/d} \|_{L_{t,x}^{2(d+2)/(d+4)}([0,1/\sqrt{N}] \times \mathbb{R}^d)} &\lesssim \|u_{>\beta N}\|_{L_{t,x}^{2(d+2)/d}([0,1/\sqrt{N}] \times \mathbb{R}^d)} \|u_{>N_0}\|_{L_{t,x}^{2(d+2)/d}([0,1/\sqrt{N}] \times \mathbb{R}^d)}^{4/d} \\ &\lesssim \|u_{>\beta N}\|_{S([0,1/\sqrt{N}])} \|u_{>N_0}\|_{L_t^\infty L_x^2([0,1/\sqrt{N}] \times \mathbb{R}^d)}^{8/(d(d+2))} \|u_{>N_0}\|_{S([0,1/\sqrt{N}])}^{4/(d+2)}. \end{aligned}$$

Collecting the three pieces together, we get (4-17). \square

Now by Strichartz estimate,

$$\|P_{\geq N} u\|_{S([0,1/\sqrt{N}])} \lesssim \|P_{\geq N} u_0\|_{L_x^2} + \|P_{\geq N} F(u)\|_{L_{t,x}^{2(d+2)/(d+4)}([0,1/\sqrt{N}] \times \mathbb{R}^d)},$$

and the latter is in turn determined by $\|P_{\geq N} u\|_{S([0,1/\sqrt{N}])}$ due to Lemma 4.7 and Lemma 4.9. This enables us to set up a recurrent relation for $\|P_{\geq N} u\|_{S([0,1/\sqrt{N}])}$.

We define

$$A_N = \|P_{\geq N} u\|_{S([0,1/\sqrt{N}])}.$$

Since locally the Strichartz norm of u is bounded, we can write

$$A := \|u\|_{S([0,1])} + 1 < \infty.$$

Using the Strichartz inequality, Lemma 4.7, Lemma 4.9 and taking $s = 1 + 2/d$, we obtain

$$\begin{aligned} A_N &\leq C(d) (\|P_{\geq N} u_0\|_{L_x^2} + \|P_{\geq N} F(u)\|_{L_{t,x}^{2(d+2)/(d+4)}([0,1/\sqrt{N}] \times \mathbb{R}^d)}) \\ &\leq C(d, \|u_0\|_{L_x^2}) \left(N^{-1-\varepsilon} \right. \\ &\quad \left. + A^{4/d} \sum_{M \leq \beta N} \left(\frac{M}{N}\right)^{1+2/d} \|P_M u\|_{S([0,1/\sqrt{N}])} \right) \end{aligned} \quad (4-19)$$

$$\left. + \|P_{\geq \beta N} u\|_{S([0,1/\sqrt{N}])} (N_0^{4/(d+2)} N^{-1/(d+2)} + A^{4/(d+2)} \|u_{\geq N_0}\|_{L_t^\infty L_x^2([0,1/\sqrt{N}])}^{8/(d(d+2))}) \right). \quad (4-20)$$

For (4-19), we do a little modification. Noting $P_M = P_M P_{\geq M/2}$, we have

$$\begin{aligned} (4-19) &\lesssim A^{4/d} \sum_{M \leq \beta N} \left(\frac{M}{N}\right)^{1+2/d} \|P_{\geq M/2} u\|_{S([0,1/\sqrt{N}])} \\ &\lesssim A^{4/d} \sum_{M \leq 2\beta N} \left(\frac{M}{N}\right)^{1+2/d} \|P_{\geq M} u\|_{S([0,1/\sqrt{N}])}. \end{aligned}$$

We shall take β to be sufficiently small. The constraint on β will be specified later.

Now we absorb (4-20) into (4-19) through taking suitable parameters. First we take $N_0 = N_0(\beta, A)$ such that

$$A^{4/(d+2)} \|u_{>N_0}\|_{L_t^\infty L_x^2([0,1])}^{8/(d(d+2))} \leq \frac{1}{100} \beta^{1+2/d}.$$

This is certainly possible since $u \in C([0, 1], L_x^2)$ and $[0, 1]$ is a compact interval. Then we assume $N \geq M_0$ where

$$M_0^{-1/(d+2)} N_0^{4/(d+2)} \leq \frac{1}{100} \beta^{1+2/d}. \quad (4-21)$$

Under these restrictions we have

$$(4-20) \leq \frac{1}{2} \beta^{1+2/d} \|P_{\geq \beta N} u\|_{S([0,1/\sqrt{N}])}. \quad (4-22)$$

Therefore we get for all $N \geq M_0$ that

$$\begin{aligned} A_N &\leq C(d, \|u_0\|_{L_x^2}) \left(N^{-1-\varepsilon} + \sum_{M \leq 2\beta N} \left(\frac{M}{N} \right)^{1+2/d} \|P_{\geq M} u\|_{S([0,1/\sqrt{N}])} \right) \\ &\leq N^{-1-\varepsilon/2} + \sum_{M \leq 2\beta N} \left(\frac{M}{N} \right)^{1+1/d} \|P_{\geq M} u\|_{S([0,1/\sqrt{N}])}, \end{aligned}$$

where in the last inequality we have killed the constant $C(d, \|u_0\|_{L_x^2})$. This is possible by first taking β sufficiently small, then taking M_0 large enough.

We split the summation into $M \leq M_0$ and $M > M_0$. For large M , we trivially bound the summand by

$$\left(\frac{M}{N} \right)^{1+1/d} A_M.$$

Then we sum all the pieces for small M , which gives that

$$\sum_{M \leq M_0} \left(\frac{M}{N} \right)^{1+1/d} \|P_{\geq M} u\|_{S([0,1/\sqrt{N}])} \lesssim A M_0^{1+1/d} N^{-1-1/d}.$$

Finally we establish the following recurrence relation for A_N : Let $s = 1/d + 1$. Then there exists $C_1 > 0$ such that for all $N \geq M_0$,

$$A_N \leq C_1 M_0^s N^{-s} + \sum_{M_0 < M \leq 2\beta N} \left(\frac{M}{N} \right)^s A_M. \quad (4-23)$$

This combined with the trivial bound $A_N \leq A$ will give us the final control on A_N ,

$$A_N \leq C(A, M_0) N^{-s+} \quad \text{for all } N \geq M_0, \quad (4-24)$$

if we apply the following lemma:

Lemma 4.10 (recursive control). *Let $s > 1$, $\gamma > 0$ and $s - \gamma > 1$. Let $C_1 > 0$ be such that for all $N \geq M_0$,*

$$A_N \leq C_1 M_0^s N^{-s} + \sum_{M_0 \leq M \leq \beta' N} \left(\frac{M}{N} \right)^s A_M, \quad (4-25)$$

$$A_N \leq A. \quad (4-26)$$

Then there exists a constant $c(s, \gamma, A) > 0$ such that for all $0 < \beta' < c(s, \gamma, A)$, we have

$$A_N \leq 2C_1 M_0^s N^{-s+\gamma} \quad \text{for all } N \geq M_0. \quad (4-27)$$

Proof. We will inductively prove that

$$A_N \leq 2C_1 M_0^s N^{-s+\gamma} + (\beta')^j. \quad (4-28)$$

First, plugging the bound (4-26) into (4-25), we get

$$A_N \leq C_1 M_0 N^{-s} + C(s)A(\beta')^s \leq 2C_1 M_0 N^{-s+\gamma} + \beta',$$

by requiring $(\beta')^{s-1} < 1/(100C(s)A)$. This establishes (4-28) for $j = 1$.

Now assuming (4-28) holds for j -th step, we plug this bound into (4-25) to compute

$$\begin{aligned} A_N &\leq C_1 M_0^s N^{-s} + 2C(s)(\beta')^\gamma \cdot C_1 M_0^s N^{-s+\gamma} + C(s)(\beta')^{s-1} \cdot (\beta')^{j+1} \\ &\leq 2C_1 M_0^s N^{-s+\gamma} + (\beta')^{j+1}, \end{aligned}$$

by requiring $(\beta')^\gamma < 1/(100C(s))$. This establishes (4-28) for $j + 1$.

Finally, (4-27) follows by taking $j \rightarrow \infty$ in (4-28). □

Acknowledgements

Both authors were supported by start-up funding from the Mathematics Department of the University of Iowa. Li was also supported by NSF grant DMS-0908032 and an Old Gold summer fellowship from the University of Iowa. Zhang was also supported by an Alfred P. Sloan Research Fellowship and Project 973 in China.

References

- [Bégout and Vargas 2007] P. Bégout and A. Vargas, “Mass concentration phenomena for the L^2 -critical nonlinear Schrödinger equation”, *Trans. Amer. Math. Soc.* **359**:11 (2007), 5257–5282. [MR 2008g:35190](#)
- [Berestycki and Lions 1979] H. Berestycki and P.-L. Lions, “Existence d’ondes solitaires dans des problèmes nonlinéaires du type Klein–Gordon”, *C. R. Acad. Sci. Paris Sér. A-B* **288**:7 (1979), A395–A398. [MR 80i:35076](#) [Zbl 0397.35024](#)
- [Cazenave 2003] T. Cazenave, *Semilinear Schrödinger equations*, Courant Lecture Notes in Math. **10**, NYU Courant Inst. of Math. Sciences, New York, 2003. [MR 2004j:35266](#) [Zbl 1055.35003](#)
- [Christ and Weinstein 1991] F. M. Christ and M. I. Weinstein, “Dispersion of small amplitude solutions of the generalized Korteweg–de Vries equation”, *J. Funct. Anal.* **100**:1 (1991), 87–109. [MR 92h:35203](#) [Zbl 0743.35067](#)
- [Ginibre and Velo 1992] J. Ginibre and G. Velo, “Smoothing properties and retarded estimates for some dispersive evolution equations”, *Comm. Math. Phys.* **144**:1 (1992), 163–188. [MR 93a:35065](#) [Zbl 0762.35008](#)
- [Glassey 1977] R. T. Glassey, “On the blowing up of solutions to the Cauchy problem for nonlinear Schrödinger equations”, *J. Math. Phys.* **18**:9 (1977), 1794–1797. [MR 57 #842](#)
- [Hmidi and Keraani 2005] T. Hmidi and S. Keraani, “Blowup theory for the critical nonlinear Schrödinger equations revisited”, *Int. Math. Res. Not.* **46** (2005), 2815–2828. [MR 2007k:35464](#) [Zbl 1126.35067](#)
- [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](#)
- [Kenig and Merle 2006] C. E. Kenig and F. Merle, “Global well-posedness, scattering and blow-up for the energy-critical, focusing, non-linear Schrödinger equation in the radial case”, *Invent. Math.* **166**:3 (2006), 645–675. [MR 2007g:35232](#) [Zbl 1115.35125](#)

- [Keraani 2001] S. Keraani, “On the defect of compactness for the Strichartz estimates of the Schrödinger equations”, *J. Differential Equations* **175**:2 (2001), 353–392. [MR 2002j:35281](#) [Zbl 1038.35119](#)
- [Keraani 2006] S. Keraani, “On the blow up phenomenon of the critical nonlinear Schrödinger equation”, *J. Funct. Anal.* **235**:1 (2006), 171–192. [MR 2007e:35260](#) [Zbl 1099.35132](#)
- [Killip et al. 2008] R. Killip, M. Visan, and X. Zhang, “The mass-critical nonlinear Schrödinger equation with radial data in dimensions three and higher”, *Anal. PDE* **1**:2 (2008), 229–266. [MR MR2472890](#) [Zbl 1171.35111](#)
- [Killip et al. 2009a] R. Killip, D. Li, M. Visan, and X. Zhang, “Characterization of minimal-mass blowup solutions to the focusing mass-critical NLS”, *SIAM J. Math. Anal.* **41**:1 (2009), 219–236. [MR MR2505858](#)
- [Killip et al. 2009b] R. Killip, T. Tao, and M. Visan, “The cubic nonlinear Schrödinger equation in two dimensions with radial data”, *J. Eur. Math. Soc.* **11**:6 (2009), 1203–1258. [MR MR2557134](#) [Zbl 05641373](#)
- [Kwong 1989] M. K. Kwong, “Uniqueness of positive solutions of $\Delta u - u + u^p = 0$ in \mathbf{R}^n ”, *Arch. Rational Mech. Anal.* **105**:3 (1989), 243–266. [MR 90d:35015](#) [Zbl 0676.35032](#)
- [Li and Zhang 2009a] D. Li and X. Zhang, “On the focusing mass critical problem in six dimensions with splitting spherically symmetric initial data”, preprint, 2009. Submitted to *Discrete Contin. Dynam. Systems*.
- [Li and Zhang 2009b] D. Li and X. Zhang, “On the rigidity of solitary waves for the focusing mass-critical NLS in dimensions $d \geq 2$ ”, preprint, 2009. [arXiv 0902.0802](#)
- [Merle 1993] F. Merle, “Determination of blow-up solutions with minimal mass for nonlinear Schrödinger equations with critical power”, *Duke Math. J.* **69**:2 (1993), 427–454. [MR 94b:35262](#) [Zbl 0808.35141](#)
- [Merle and Raphael 2005] F. Merle and P. Raphael, “The blow-up dynamic and upper bound on the blow-up rate for critical nonlinear Schrödinger equation”, *Ann. of Math. (2)* **161**:1 (2005), 157–222. [MR 2006k:35277](#) [Zbl 02204253](#)
- [Staffilani 1997] G. Staffilani, “On the generalized Korteweg–de Vries-type equations”, *Differential Integral Equations* **10**:4 (1997), 777–796. [MR 2001a:35005](#) [Zbl 0891.35135](#)
- [Strichartz 1977] R. S. Strichartz, “Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations”, *Duke Math. J.* **44**:3 (1977), 705–714. [MR 58 #23577](#) [Zbl 0372.35001](#)
- [Tao et al. 2007] T. Tao, M. Visan, and X. Zhang, “Global well-posedness and scattering for the defocusing mass-critical nonlinear Schrödinger equation for radial data in high dimensions”, *Duke Math. J.* **140**:1 (2007), 165–202. [MR 2010a:35249](#) [Zbl 05208403](#)
- [Tao et al. 2008] T. Tao, M. Visan, and X. Zhang, “Minimal-mass blowup solutions of the mass-critical NLS”, *Forum Math.* **20**:5 (2008), 881–919. [MR 2009m:35495](#) [Zbl 1154.35085](#)
- [Taylor 2000] M. E. Taylor, *Tools for PDE. Pseudodifferential operators, paradifferential operators, and layer potentials*, Math. Surveys and Monogr. **81**, Amer. Math. Soc., Providence, RI, 2000. [MR 2001g:35004](#) [Zbl 0963.35211](#)
- [Weinstein 1983] M. I. Weinstein, “Nonlinear Schrödinger equations and sharp interpolation estimates”, *Comm. Math. Phys.* **87**:4 (1983), 567–576. [MR 84d:35140](#) [Zbl 0527.35023](#)
- [Weinstein 1986] M. I. Weinstein, “On the structure and formation of singularities in solutions to nonlinear dispersive evolution equations”, *Comm. Partial Differential Equations* **11**:5 (1986), 545–565. [MR 87i:35026](#) [Zbl 0596.35022](#)

Received 10 Aug 2009. Revised 18 Nov 2009. Accepted 17 Dec 2009.

DONG LI: mpdongli@gmail.com

Department of Mathematics, University of Iowa, 14 MacLean Hall, Iowa City, IA 52240, United States

XIAOYI ZHANG: zh.xiaoyi@gmail.com

Academy of Mathematics and System Sciences, Beijing, China

and

Department of Mathematics, University of Iowa, 14 MacLean Hall, Iowa City, IA 52240, United States

ESTIMÉES DES NOYAUX DE GREEN ET DE LA CHALEUR SUR LES ESPACES SYMÉTRIQUES

GILLES CARRON

On majore les noyaux de Green et de la chaleur au dehors de la diagonale pour des opérateurs de type laplacien sur les espaces symétriques.

We provide an upper bound for the off-diagonal entries of the Green and heat kernel for Laplace-type operators on symmetric spaces.

1. Introduction

On considère ici un espace symétrique $X = G/K$ de type non compact. À une représentation unitaire (ρ, V) de dimension finie de K , on associe le fibré vectoriel $G \times_K V$ au dessus de X , dont l'espace des sections lisses s'identifie à

$$C^\infty(E) \simeq \{f \in C^\infty(G, V) : g \in G, k \in K \Rightarrow f(gk) = \rho(k^{-1})f(g)\}$$

L'objet de cet article est un opérateur de type laplacien G -invariant agissant sur les sections de E

$$L = \nabla^* \nabla + R \tag{1-1}$$

où ∇ est une connexion hermitienne G -invariante sur E et R une section G -invariante du fibré des endomorphismes hermitiens de E . Nous donnons quelques estimations de la résolvante et du noyau de la chaleur associé à L . Notre premier résultat est le suivant :

Théorème A. *Notons λ_0 le bas du spectre de l'opérateur L , $o = \text{Id} \cdot K \in X$ et $G_s(x, y)$ le noyau de Schwartz de la résolvante $(L - \lambda_0 + s^2)^{-1}$ où s est un nombre complexe tel que s et s^2 aient leurs parties réelles strictement positives. Il y a une constante C telle que pour tout $x \in X$ tel que $d(x, o) \geq 2$, on ait alors :*

$$|G_s(x, o)| \leq C e^{-\rho(x^+) - \text{Re}(s)d(x, o)},$$

où on a noté x^+ la composante suivant $\bar{\alpha}^+$ de $x = gK$ dans la décomposition de Cartan $G = K e^{\bar{\alpha}^+} K$ et $\rho \in \mathfrak{a}_\mathbb{C}^*$ la demi somme des racines restreintes positives associées à $(\mathfrak{g}_\mathbb{C}, \mathfrak{a})$.

Le calcul explicite de λ_0 est en général difficile. Le bas du spectre du laplacien agissant sur les fonctions est égal à $\|\rho\|^2$. Concernant le laplacien de Hodge–de Rham sur les formes différentielles des calculs explicites sont faits par H. Donnelly [1981] et E. Pedon [1999; 2005] en rang 1 et par N. Lohoué et S. Medhi [2007, Appendix A] pour certains espaces hermitiens.

MSC2000: primary 53C35, 58J50; secondary 22E40.

Mots-clefs: espace symétrique, noyau de Green, noyau de la chaleur, laplacien, propagation à vitesse finie, symmetric space, Green kernel, heat kernel, laplacian, finite-speed propagation.

La preuve de notre estimation n'utilise que deux ingrédients, à savoir une estimation du volume de $KB(x, 1) \subset X$ et un argument désormais classique, introduit par J. Cheeger, M. Gromov et M. Taylor [Cheeger et al. 1982], de propagation à vitesse finie. Pour certains espaces localement symétriques ou à géométrie bornée, Taylor [1989] a utilisé la technique de propagation à vitesse finie pour obtenir des résultats optimaux sur les normes $L^p \rightarrow L^p$ de fonctions du laplacien.

Notre résultat est sensiblement meilleur que celui obtenu récemment par Lohoué et Mehdi [2007] à propos du laplacien de Hodge–de Rham ; en utilisant un théorème de Paley–Wiener de P. Delorme [2005] et la théorie des représentations de G , ils obtiennent pour tout $\varepsilon \in]0, 1[$ l'existence d'une constante C_ε telle que pour tout $x \in X$ tel que $d(x, o) \geq 1$,

$$|G_s(x, o)| \leq C_\varepsilon \Phi_0(x) e^{-(1-\varepsilon)\operatorname{Re}(s)d(x,o)},$$

où Φ_0 est la fonction sphérique élémentaire de Harish-Chandra de G . On sait qu'il y a une constante telle que $\Phi_0(x) \geq C e^{-\rho(x^+)}$, en fait la fonction $\Phi_0(x)e^{\rho(x^+)}$ croît polynomialement sur $\bar{\mathfrak{a}}^+$ [Anker 1987].

L'approche développée par R. Mazzeo et A. Vasy utilise elle la géométrie de l'espace symétrique et une construction de paramétrice reliée à cette géométrie. Il s'agit d'une méthode beaucoup plus élaborée que la nôtre mais elle fournit beaucoup plus d'informations que l'estimation ponctuelle obtenue ici. Dans le cas de l'espace symétrique $SL_3(\mathbb{R})/SO_3(\mathbb{R})$, Mazzeo et Vasy [2007] ont obtenu un développement asymptotique complet de la résolvante ; de plus cette méthode pourrait être généralisée à toutes les géométries asymptotiquement symétriques.

Cependant notre estimation n'est pas, en général, optimale. Par exemple pour les fonctions, on sait grâce au travail de J-P. Anker et L. Ji [1999, Theorem 4.22(i)] que pour $s > 0$, l'on a une estimation de la forme

$$C^{-1}d(x, o)^{-\beta} \Phi_0(x)e^{-sd(x,o)} \leq G_s(x, o) \leq C d(x, o)^{-\beta} \Phi_0(x)e^{-sd(x,o)}$$

où si on note Σ^{++} les racines positives indivisibles et l le rang de X alors $\beta = |\Sigma^{++}| + (l - 1)/2$. En fait, on a l'estimation $d(x, o)^{-\beta} \Phi_0(x) \leq C d(x, o)^{-(l-1)/2} e^{-\rho(x^+)}$. Sur les fonctions, notre estimation est donc optimale en rang 1, et en rang supérieur, elle est optimale à un facteur polynomial près ; notons également que grâce à [Carron et Pedon 2004, Theorem 3.6], notre résultat est optimal pour le laplacien de Hodge–de Rham en rang 1.

Nous avons également obtenu une estimation du noyau de Schwartz de l'opérateur de la chaleur e^{-tL} par la même méthode. Pour énoncer ce résultat, on rappelle quelques notations sur la structure algébrique de X . On note $\mathfrak{k} \subset \mathfrak{g}$ les algèbres de Lie de K et G et

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$$

la décomposition en espaces propres de l'involution de Cartan θ . Soit $\mathfrak{a} \subset \mathfrak{p}$ une sous-algèbre abélienne maximale et $\Sigma \subset \mathfrak{a}_\mathbb{C}^*$ le système restreint des racines de $(\mathfrak{g}, \mathfrak{a})$. On fixe alors $\mathfrak{a}^+ \subset \mathfrak{a}$ une chambre de Weyl et on note $\Sigma^+ \subset \Sigma$ le système des racines restreintes positives associées. Le rang de l'espace symétrique X est $l = \dim \mathfrak{a}$; la dimension de l'espace symétrique X est notée d . L'espace \mathfrak{p} se décompose en

$$\mathfrak{p} = \mathfrak{a} \oplus \bigoplus_{\alpha \in \Sigma^+} \mathfrak{p}_\alpha.$$

où si on introduit

$$\mathfrak{n}_\alpha = \{n \in \mathfrak{g} : a \in \mathfrak{a} \Rightarrow \operatorname{ad}(a)n = \alpha(a)n\}$$

alors

$$\mathfrak{p}_\alpha = \{x + \theta(x) : x \in \mathfrak{n}_\alpha\}.$$

On note aussi $m_\alpha = \dim \mathfrak{n}_\alpha$ et donc $\rho = \frac{1}{2} \sum_{\alpha \in \Sigma^+} m_\alpha \alpha \in \mathfrak{a}_\mathbb{C}^*$. Dans la décomposition de Cartan de $G = K e^{\bar{\mathfrak{a}}^+} K$, si $x = gK \in X = G/K$, on note x^+ l'unique élément de $\bar{\mathfrak{a}}^+$ tel que $g \in K e^{x^+} K$. Notre estimation est alors la suivante :

Théorème B. *Notons $h_t(x, y)$ le noyau de Schwartz de l'opérateur de la chaleur e^{-tL} . Il existe une constante C telle que pour tout $x \in X$ tel que $d(x, o) \geq 2$ on ait :*

$$|h_t(x, o)| \leq C e^{-\lambda_0 t - \rho(x^+) - d(x, o)^2/4t} \phi_t(x)$$

où

$$\phi_t(x) = \begin{cases} \frac{\sqrt{t}}{d(x, o) + \sqrt{t}} & \text{si } d(x, o) \leq t, \\ \frac{d(x, o)^{(d+l)/2-1}}{t^{(d+l-1)/2}} \prod_{\alpha \in \Sigma^+} \left(\frac{1 + \alpha(x^+)}{t} + \alpha(x^+) \right)^{m_\alpha/2} & \text{si } d(x, o) \geq t. \end{cases}$$

Cette majoration n'est également pas optimale. On peut comparer notre estimation avec celle obtenue par Lohoué et Mehdi [2007] à propos du laplacien de Hodge–de Rham ; ils obtiennent pour tout $\varepsilon \in]0, 1[$ des constantes C_ε et A_ε telles que si $d(x, o) \geq A_\varepsilon$ alors

$$|h_t(x, o)| \leq C_\varepsilon \Phi_0(x) e^{-\lambda_0 t} e^{-(1-\varepsilon)d(x, o)^2/4t} t^{-\varepsilon \gamma}.$$

Notre estimation est donc meilleure lorsque $d(x, o)$ tend vers l'infini mais bien plus mauvaise lorsque t tend vers $+\infty$. Dans le cas de l'espace hyperbolique réel et du laplacien scalaire, on peut vérifier avec l'estimation de E. Davies et N. Mandouvalos [1988] que notre estimée est optimale dans le régime où $d(x, o) \geq \max\{2, t\}$.

2. Une estimée de volume

Proposition 2.1. *Il y a des constantes strictement positives c, C telles que pour tout $\varepsilon \in [0, 1[$ et $x \in X$*

$$c \varepsilon^l e^{2\rho(x^+)} \prod_{\alpha \in \Sigma^+} \left(\frac{\varepsilon + \alpha(x^+)}{1 + \alpha(x^+)} \right)^{m_\alpha} \leq \text{vol } KB(x, \varepsilon) \leq C \varepsilon^l e^{2\rho(x^+)} \prod_{\alpha \in \Sigma^+} \left(\frac{\varepsilon + \alpha(x^+)}{1 + \alpha(x^+)} \right)^{m_\alpha}.$$

Démonstration. Grâce à [Anker et Ji 1999, lemme 2.1.2], nous savons que

$$KB(x, \varepsilon) \simeq K \exp(B(x^+, \varepsilon) \cap \bar{\mathfrak{a}}^+)$$

dans la décomposition de Cartan $X = K e^{\bar{\mathfrak{a}}^+}$. Ainsi si $J(h) dk dh$ est la forme volume de X dans les coordonnées $(k, h) \mapsto k e^h K$ nous avons :

$$\text{vol } KB(x, \varepsilon) = \int_{B(x^+, \varepsilon) \cap \bar{\mathfrak{a}}^+} J(h) dh.$$

Cependant nous avons pour une constante positive C :

$$J(h) = C \prod_{\alpha \in \Sigma^+} \sinh^{m_\alpha}(\alpha(h)) \approx C e^{2\rho(h)} \prod_{\alpha \in \Sigma^+} \left(\frac{\alpha(h)}{1 + \alpha(h)} \right)^{m_\alpha};$$

c'est à dire qu'il y a une constante $\eta > 0$ tel que pour tout h :

$$\eta e^{2\rho(h)} \prod_{\alpha \in \Sigma^+} \left(\frac{\alpha(h)}{1 + \alpha(h)} \right)^{m_\alpha} \leq J(h) \leq \eta^{-1} e^{2\rho(h)} \prod_{\alpha \in \Sigma^+} \left(\frac{\alpha(h)}{1 + \alpha(h)} \right)^{m_\alpha}$$

Grâce à la preuve de [Anker et Ji 1999, lemme 2.1.6(i)], on en déduit que pour $\varepsilon \in]0, 1[$ et $h \in B(x^+, \varepsilon) \cap \mathfrak{a}^+$, on a

$$\rho(x^+) - |\rho| \leq \rho(h) \leq \rho(x^+) + |\rho|$$

et pour $\alpha \in \Sigma^+$,

$$|\alpha(h - x^+)| \leq \varepsilon/\sqrt{2}, \quad \left(1 - \frac{1}{\sqrt{2}}\right)(1 + \alpha(x^+)) \leq 1 + \alpha(h) \leq 2(1 + \alpha(x^+)), \quad \alpha(h) \leq \alpha(x^+) + \varepsilon.$$

On en déduit aisément la majoration annoncée.

Pour la minoration, on considère $\Sigma^{+++} = \{\alpha_1, \dots, \alpha_l\}$ un système de racines réduites qui est une base de $\mathfrak{a}_\mathbb{C}^*$ et E_1, \dots, E_l la base de \mathfrak{a} duale à Σ^{+++} . On pose alors

$$v = \sum_i E_i.$$

Ainsi pour $\alpha \in \Sigma^+$, on a $\alpha(v) \geq 1$. On a bien évidemment

$$B\left(x^+ + \frac{\varepsilon}{10 + 10|v|}v, \frac{\varepsilon}{20 + 20|v|}\right) \subset B(x^+, \varepsilon);$$

or sur la boule de gauche on a pour $\alpha \in \Sigma^+$

$$\alpha \geq \alpha(x^+) + \frac{\varepsilon}{10 + 10|v|}\alpha(v) - \frac{\varepsilon}{20 + 20|v|} \geq \alpha(x^+) + \frac{\varepsilon}{20 + 20|v|}.$$

On obtient ainsi facilement une minoration du volume de $KB\left(x^+ + \frac{\varepsilon}{10 + 10|v|}v, \frac{\varepsilon}{20 + 20|v|}\right)$ et donc du volume de $KB(x^+, \varepsilon)$. \square

3. Estimation du noyau de Green

Ici, on étudie le noyau de Schwartz de l'opérateur $(L - \lambda_0 + s^2)^{-1}$ au dehors de la diagonale où s est un nombre complexe de partie réelle strictement positive. On commence par une estimée classique induite par la propriété de propagation à vitesse finie de l'opérateur $\cos(t\sqrt{L - \lambda_0})$; cf. [Cheeger et al. 1982, Proposition 3.1] et aussi [Ma et Marinescu 2007, appendice D]. On considère $x \in X$ vérifiant $d(x, o) \geq 2$ et on note $A := KB(x, 1)$.

Lemme 3.1. *Soit $\sigma \in L^2(A, E)$ et $u := (L - \lambda_0 + s^2)^{-1}\sigma$. Alors on a*

$$\|u\|_{L^2(B(o,1))} \leq \frac{1}{(\operatorname{Re} s)^2} e^{-\operatorname{Re}(s)(d(x,o)-2)} \|\sigma\|_{L^2}.$$

Démonstration. En effet, on a

$$u = \int_0^\infty \frac{e^{-s\xi}}{s} \cos(\xi\sqrt{L-\lambda_0}) \sigma \, d\xi.$$

Les hypothèses faites sur x et σ et la propriété de propagation à vitesse finie impliquent que dès que $0 \leq \xi \leq d(x, o) - 2$, on a $\|\cos(\xi\sqrt{L-\lambda_0}) \sigma\|_{L^2(B(o,1))} = 0$. D'où

$$\begin{aligned} \|u\|_{L^2(B(o,1))} &\leq \int_{d(x,o)-2}^\infty \frac{e^{-\operatorname{Re}(s)\xi}}{|s|} \left\| \cos(\xi\sqrt{L-\lambda_0}) \sigma \right\|_{L^2(B(o,1))} \, d\xi \\ &\leq \int_{d(x,o)-2}^\infty \frac{e^{-\operatorname{Re}(s)\xi}}{|s|} \left\| \cos(\xi\sqrt{L-\lambda_0}) \sigma \right\|_{L^2(X)} \, d\xi \\ &\leq \int_{d(x,o)-2}^\infty \frac{e^{-\operatorname{Re}(s)\xi}}{|s|} \|\sigma\|_{L^2(X)} \, d\xi \leq \frac{1}{(\operatorname{Re} s)^2} e^{-\operatorname{Re}(s)(d(x,o)-2)} \|\sigma\|_{L^2}. \quad \square \end{aligned}$$

On utilise alors l'estimée elliptique standard suivante :

Proposition 3.2. *Soit $\lambda \in \mathbb{C}$. Il y a une constante C qui dépend de X, λ, L telle que si $r \in]0, 1]$ et si $v \in L^2(B(x, r), E)$ vérifie $Lv = \lambda v$ alors*

$$|v(x)| \leq \frac{C}{r^{d/2}} \|v\|_{L^2(B(x,r))}.$$

On en déduit l'existence d'une constante C qui ne dépend que de X, L, s telle que

$$|u(o)| \leq C e^{-\operatorname{Re}(s)d(x,o)} \|\sigma\|_{L^2(A)}.$$

Cette estimation fournit une majoration de la norme de l'application linéaire

$$T : L^2(A, E) \mapsto E_o \simeq V, \quad \sigma \rightarrow (L - \lambda_0 - s^2)^{-1} \sigma(o)$$

Et donc cela induit la même majoration de la norme de l'opérateur adjoint $T^* : V \simeq E_o \rightarrow L^2(A, E)$ qui est défini pour $v \in E_o$ par :

$$(T^*v)(y) = G_{\bar{s}}(y, o)v$$

On obtient donc :

$$\int_{KB(x,1)} |G_{\bar{s}}(y, o)|^2 \, dy \leq C e^{-2\operatorname{Re}(s)d(x,o)}.$$

Lemme 3.3. *il y a une constante C qui ne dépend que de X telle que si $f \in L^1(A)$ alors il existe $k \in K$ tel que*

$$\|f\|_{L^1(B(kx, \frac{1}{4}))} \leq C e^{-2\rho(x^+)} \|f\|_{L^1(A)}.$$

Démonstration. En effet,

$$\begin{aligned} \int_{KB(x, \frac{1}{4})} \left(\int_{B(y, \frac{1}{2})} |f|(z) \, dz \right) \, dy &= \int_{KB(x,1)} \operatorname{vol} \left(B(z, \frac{1}{2}) \cap KB(x, \frac{1}{4}) \right) |f|(z) \, dz \\ &\leq \operatorname{vol} B(0, \frac{1}{2}) \int_{KB(x,1)} |f|(z) \, dz. \end{aligned}$$

On en déduit donc l'existence d'un $y \in KB(x, \frac{1}{4})$ tel que

$$\text{vol } KB(x, \frac{1}{4}) \|f\|_{L^1(B(y, 1/2))} \leq C \|f\|_{L^1}.$$

Il y a donc un $k \in K$ tel que $d(y, kx) \leq \frac{1}{4}$, d'où $B(kx, \frac{1}{4}) \subset B(y, \frac{1}{2})$. Maintenant, grâce à l'estimée donnée par la [proposition 2.1](#), soit $\text{vol}(KB(x, \frac{1}{4})) \approx e^{2\rho(x^+)}$, on obtient bien le résultat annoncé. \square

On en déduit donc l'existence d'un $k \in K$ tel que

$$\int_{B(kx, 1/4)} |G_{\bar{s}}(y, o)|^2 dy \leq C e^{-2\text{Re}(s)d(x, o) - 2\rho(x^+)}.$$

Les mêmes estimées elliptiques entraînent alors

$$|G_{\bar{s}}(kx, o)| = |G_s(o, x)| \leq C e^{-\text{Re}(s)d(x, o) - \rho(x^+)},$$

ce qui démontre le [théorème A](#).

4. Estimation du noyau de la chaleur

On étudie maintenant de la même façon le noyau de Schwartz de l'opérateur de la chaleur e^{-tL} au dehors de la diagonale. On considère donc $x \in X$ et $\varepsilon > 0$ tels que $d(x, o) \geq 2\varepsilon$. On note $A := KB(x, \varepsilon)$. Nous commençons par le même type d'estimations :

Lemme 4.1. *Soit $\sigma \in L^2(A, E)$ et $f_t := e^{-tL}\sigma$. Alors on a*

$$\|f_t\|_{L^2(B(o, \varepsilon))} \leq \frac{e^{-\lambda_0 t}}{\sqrt{\pi t}} \int_{d(x, o) - 2\varepsilon}^{\infty} e^{-\zeta^2/4t} d\zeta \|\sigma\|_{L^2}.$$

Cette estimation se montre de la même façon que l'estimée du [lemme 3.1](#), en partant de la formule :

$$f_t = \frac{e^{-\lambda_0 t}}{\sqrt{\pi t}} \int_0^{\infty} e^{-\zeta^2/4t} \cos(\zeta \sqrt{L - \lambda_0}) \sigma d\zeta.$$

Maintenant, on utilise l'estimation parabolique suivante :

Proposition 4.2. *Il y a une constante C (qui ne dépend que de X, L) telle que si $r \in]0, 1]$ et si $v \in L^2([t - r^2, t] \times B(x, r), E)$ est une solution de l'équation :*

$$\frac{\partial}{\partial t} v + Lv = 0$$

alors

$$|v(t, x)|^2 \leq \frac{C}{r^{d+2}} \int_{t-r^2}^t \left(\int_{B(x, r)} |v(\tau, y)|^2 dy \right) d\tau.$$

Ceci provient du fait que si on note μ la plus petite valeur propre de l'opérateur R dans la formule (1-1) alors la fonction u définie par

$$u(t, x) = |v(t, x)| e^{\mu t}$$

vérifie

$$\frac{\partial}{\partial t} u + \Delta u \leq 0.$$

Les inégalités paraboliques de J. Moser impliquent alors ce résultat ; voir [Moser 1964], [Grigor'yan 1994, Theorem 3.1] ou [Saloff-Coste 1992, Theorem 5.1].

On en déduit l'existence d'une constante C telle que

$$|f_t(o)|^2 \leq \frac{C}{\varepsilon^{d+2}} \int_{[t-\varepsilon^2, t] \times B(o, \varepsilon)} |f_\tau(y)|^2 d\tau dy.$$

Avec l'estimation

$$\int_A^\infty e^{-\xi^2/4t} d\xi = \sqrt{t} \int_{A^2/4t}^\infty e^{-v} \frac{dv}{\sqrt{v}} \leq \frac{C\sqrt{t}}{A/\sqrt{t} + 1} e^{-A^2/4t},$$

on obtient, pour $\varepsilon \in]0, 1]$ et $t \geq 2\varepsilon^2$ et $d(x, o) \geq 2$, l'estimée suivante :

$$|f_t(o)| \leq C\varepsilon^{-d/2} \frac{\sqrt{t}}{d(x, o) + \sqrt{t}} e^{-\lambda_0 t - (d(x, o) - 2\varepsilon)^2/4t} \|\sigma\|_{L^2}.$$

Avec les mêmes arguments que ceux utilisés dans la preuve du [théorème A](#), on en déduit :

$$\left(\int_{KB(x, \varepsilon)} |h_t(y, o)|^2 dy \right)^{1/2} \leq C\varepsilon^{-d/2} \frac{\sqrt{t}}{d(x, o) + \sqrt{t}} e^{-\lambda_0 t - (d(x, o) - 2\varepsilon)^2/4t}.$$

La même argumentation basée sur le [lemme 3.3](#) permet de trouver $k \in K$ tel que

$$\left(\int_{B(kx, \varepsilon/4)} |h_t(y, o)|^2 dy \right)^{1/2} \leq C (\text{vol } KB(x, \varepsilon))^{-1/2} \frac{\sqrt{t}}{d(x, o) + \sqrt{t}} e^{-\lambda_0 t - (d(x, o) - 2\varepsilon)^2/4t}.$$

Les mêmes estimées paraboliques donnent alors la majoration suivante : pour $\varepsilon \in]0, 1/2]$, $t \geq 3\varepsilon^2$ et $d(x, o) \geq 2$, on a

$$|h_t(kx, o)| = |h_t(x, o)| \leq C(\varepsilon^d \text{vol } KB(x, \varepsilon))^{-1/2} \frac{\sqrt{t}}{d(x, o) + \sqrt{t}} e^{-\lambda_0 t - (d(x, o) - 2\varepsilon)^2/4t}.$$

Or nous avons

$$\frac{(d(x, o) - 2\varepsilon)^2}{4t} = \frac{d(x, o)^2}{4t} - \frac{d(x, o)\varepsilon}{t} + \frac{\varepsilon^2}{t}.$$

Donc lorsque $d(x, o) \leq t$ on choisit $\varepsilon = \frac{1}{100}$ et on obtient la majoration :

$$|h_t(x, o)| \leq C \frac{\sqrt{t}}{d(x, o) + \sqrt{t}} e^{-\lambda_0 t - d(x, o)^2/4t - \rho(x^+)}$$

Lorsque $d(x, o) \geq t$ on choisit $\varepsilon = \frac{t}{100d(x, o)}$ et on obtient pour $d(x, o) \geq 2$

$$\begin{aligned} |h_t(x, o)| &\leq \frac{C\sqrt{t}}{d(x, o)} \left(\frac{d(x, o)}{t} \right)^{(d+l)/2} \prod_{\alpha \in \Sigma^+} \left(\frac{1 + \alpha(x^+)}{\frac{t}{100d(x, o)} + \alpha(x^+)} \right)^{m_\alpha/2} e^{-\lambda_0 t - d(x, o)^2/4t - \rho(x^+)} \\ &\leq \frac{Cd(x, o)^{(d+l)/2-1}}{t^{(d+l-1)/2}} \prod_{\alpha \in \Sigma^+} \left(\frac{1 + \alpha(x^+)}{\frac{t}{d(x, o)} + \alpha(x^+)} \right)^{m_\alpha/2} e^{-\lambda_0 t - d(x, o)^2/4t - \rho(x^+)}, \end{aligned}$$

ce qui termine la démonstration du [théorème B](#).

5. Applications

Dans [Carron et Pedon 2004], une estimation du prolongement analytique de la résolvante avait été obtenue ; cependant les méthodes rudimentaires utilisées ici ne permettent pas d'obtenir un tel résultat. Néanmoins, nos estimées, comme celles de Lohoué et Mehdi, permettent une estimation inférieure du bas du spectre de l'opérateur L sur des espaces localement symétriques $\Gamma \backslash G/K$ où $\Gamma \subset G$ est un sous-groupe discret sans torsion.

Définition. (cf. [Carron et Pedon 2004, Theorem 2.7]) Soit $\Gamma \subset G$ est un sous-groupe discret sans torsion. On note $\tilde{\delta}(\Gamma)$ l'exposant critique modifié de Γ qui est l'exposant critique de la série de Poincaré :

$$\sum_{\gamma \in \Gamma} e^{-\rho(\gamma^+) - sd(\gamma(o), o)},$$

c'est à dire :

$$\tilde{\delta}(\Gamma) = \inf \left\{ s \in \mathbb{R} : \sum_{\gamma \in \Gamma} e^{-\rho(\gamma^+) - sd(\gamma(o), o)} < \infty \right\}.$$

Notons $G_s^0(x, y)$ le noyau de Green de l'opérateur $(\Delta - |\rho|^2 + s^2)^{-1}$ agissant sur les fonctions. Grâce à notre estimation et à l'estimation inférieure de G^0 dans [Anker et Ji 1999, Theorem 4.2.2], on sait que pour tout $s > 0$ et $\eta \in]0, s[$, il y a une constante $C_{s,\eta}$ tel que pour tout $x, y \in X$,

$$|G_s(x, y)| \leq C_{s,\eta} G_{s-\eta}^0(x, y).$$

Le même raisonnement que celui utilisé pour démontrer [Carron et Pedon 2004, Theorem 2.7] montre que :

Théorème 5.1. *Notons toujours λ_0 le bas du spectre de l'opérateur L sur $X = G/K$.*

- (i) *Si $\tilde{\delta}(\Gamma) > 0$ alors le bas du spectre de L sur $\Gamma \backslash G/K$ est minoré par $\lambda_0 - (\tilde{\delta}(\Gamma))^2$.*
- (ii) *Si $\tilde{\delta}(\Gamma) \leq 0$ alors le bas du spectre de L sur $\Gamma \backslash G/K$ est minoré par λ_0 .*
- (iii) *Si $\tilde{\delta}(\Gamma) \leq 0$ et si le rayon d'injectivité de $\Gamma \backslash G/K$ est non-majoré, i.e., $\sup_{x \in X} \inf_{\gamma \in \Gamma} d(x, \gamma(x)) = \infty$, alors le bas du spectre de L est λ_0 .*

Remarques 5.2. (i) Cet exposant critique modifié se compare aisément à l'exposant critique de Γ , à savoir à $\delta(\Gamma)$, l'exposant critique de la série

$$\sum_{\gamma \in \Gamma} e^{-sd(\gamma(o), o)}.$$

Si on note $\rho_{\min} = \inf_{H \in \mathfrak{a}_+} \rho(h)/|h|$ alors

$$\rho_{\min} + \tilde{\delta}(\Gamma) \leq \delta(\Gamma) \leq |\rho| + \tilde{\delta}(\Gamma)$$

ce qui permet de ré-obtenir le résultat dans [Lohoué et Mehdi 2007, Theorem 6.1].

- (ii) Lorsque $\tilde{\delta}(\Gamma) < \sqrt{\lambda_0}$, ce résultat implique que le noyau L^2 de L sur $\Gamma \backslash G/K$ est trivial. Il est cependant difficile d'obtenir des calculs explicites de λ_0 . Concernant le laplacien de Hodge–de Rham sur les formes différentielles des calculs explicites se trouvent dans [Donnelly 1981 ; 1999 ; 2005] en rang 1 et dans [Lohoué et Mehdi 2007, Appendix A] pour certains espaces hermitiens.

Remerciements

Je remercie ici E. Pedon et M. Olbrich pour leurs commentaires très utiles. Je remercie également le rapporteur pour ses remarques perspicaces. Je bénéficie du support partiel du projet ANR GeomEinstein 06-BLAN-0154.

References

- [Anker 1987] J.-P. Anker, “La forme exacte de l’estimation fondamentale de Harish-Chandra”, *C. R. Acad. Sci. Paris Sér. I Math.* **305**:9 (1987), 371–374. [MR 89i:22016](#) [Zbl 0636.22005](#)
- [Anker et Ji 1999] J.-P. Anker et L. Ji, “Heat kernel and Green function estimates on noncompact symmetric spaces”, *Geom. Funct. Anal.* **9**:6 (1999), 1035–1091. [MR 2001b:58038](#) [Zbl 0942.43005](#)
- [Carron et Pedon 2004] G. Carron et E. Pedon, “On the differential form spectrum of hyperbolic manifolds”, *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **3**:4 (2004), 705–747. [MR 2005j:58041](#) [Zbl 1170.53309](#)
- [Cheeger et al. 1982] J. Cheeger, M. Gromov et M. Taylor, “Finite propagation speed, kernel estimates for functions of the Laplace operator, and the geometry of complete Riemannian manifolds”, *J. Differential Geom.* **17**:1 (1982), 15–53. [MR 84b:58109](#) [Zbl 0493.53035](#)
- [Davies et Mandouvalos 1988] E. B. Davies et N. Mandouvalos, “Heat kernel bounds on hyperbolic space and Kleinian groups”, *Proc. London Math. Soc.* (3) **57**:1 (1988), 182–208. [MR 89i:58137](#) [Zbl 0643.30035](#)
- [Delorme 2005] P. Delorme, “Sur le théorème de Paley–Wiener d’Arthur”, *Ann. of Math. (2)* **162**:2 (2005), 987–1029. [MR 2006g:22009](#) [Zbl 1121.22002](#)
- [Donnelly 1981] H. Donnelly, “The differential form spectrum of hyperbolic space”, *Manuscripta Math.* **33**:3-4 (1981), 365–385. [MR 82f:58085](#) [Zbl 0464.58020](#)
- [Grigor’yan 1994] A. Grigor’yan, “Heat kernel upper bounds on a complete non-compact manifold”, *Rev. Mat. Iberoamericana* **10**:2 (1994), 395–452. [MR 96b:58107](#)
- [Lohoué et Mehdi 2007] N. Lohoué et S. Mehdi, “Estimées du noyau de la chaleur pour les formes différentielles sur les espaces symétriques et L^2 -cohomologie des espaces localement symétriques”, *C. R. Math. Acad. Sci. Paris* **345**:3 (2007), 119–122. [MR 2009f:58037](#) [Zbl 05182217](#)
- [Ma et Marinescu 2007] X. Ma et G. Marinescu, *Holomorphic Morse inequalities and Bergman kernels*, Progress in Mathematics **254**, Birkhäuser, Basel, 2007. [MR 2008g:32030](#) [Zbl 1135.32001](#)
- [Mazzeo et Vasy 2007] R. Mazzeo et A. Vasy, “Scattering theory on $SL(3)/SO(3)$: connections with quantum 3-body scattering”, *Proc. Lond. Math. Soc.* (3) **94**:3 (2007), 545–593. [MR 2008g:43018](#) [Zbl 1117.43009](#)
- [Moser 1964] J. Moser, “A Harnack inequality for parabolic differential equations”, *Comm. Pure Appl. Math.* **17** (1964), 101–134. [MR 28 #2357](#) [Zbl 0149.06902](#)
- [Pedon 1999] E. Pedon, “Harmonic analysis for differential forms on complex hyperbolic spaces”, *J. Geom. Phys.* **32**:2 (1999), 102–130. [MR 2000j:22013](#) [Zbl 0935.22010](#)
- [Pedon 2005] E. Pedon, “The differential form spectrum of quaternionic hyperbolic spaces”, *Bull. Sci. Math.* **129**:3 (2005), 227–265. [MR 2005m:58065](#) [Zbl 1070.22005](#)
- [Saloff-Coste 1992] L. Saloff-Coste, “Uniformly elliptic operators on Riemannian manifolds”, *J. Differential Geom.* **36**:2 (1992), 417–450. [MR 93m:58122](#) [Zbl 0735.58032](#)
- [Taylor 1989] M. E. Taylor, “ L^p -estimates on functions of the Laplace operator”, *Duke Math. J.* **58**:3 (1989), 773–793. [MR 91d:58253](#) [Zbl 0691.58043](#)

Received 15 Sep 2009. Revised 25 Jan 2010. Accepted 22 Feb 2010.

GILLES CARRON: Gilles.Carron@univ-nantes.fr

Laboratoire de Mathématiques Jean Leray (UMR 6629), Université de Nantes, 2, rue de la Houssinière, B.P. 92208, 44322 Nantes Cedex 3, France

<http://www.math.sciences.univ-nantes.fr/~carron/>

LOWER BOUNDS FOR RESONANCES OF INFINITE-AREA RIEMANN SURFACES

DMITRY JAKOBSON AND FRÉDÉRIC NAUD

For infinite-area, geometrically finite surfaces $X = \Gamma \backslash \mathbb{H}^2$, we prove new omega lower bounds on the local density of resonances $\mathcal{D}(z)$ when z lies in a logarithmic neighborhood of the real axis. These lower bounds involve the dimension δ of the limit set of Γ . The first bound is valid when $\delta > \frac{1}{2}$ and shows logarithmic growth of the number $\mathcal{D}(z)$ of resonances at high energy, that is, when $|\operatorname{Re}(z)| \rightarrow +\infty$. The second bound holds for $\delta > \frac{3}{4}$ and if Γ is an infinite-index subgroup of certain arithmetic groups. In this case we obtain a polynomial lower bound. Both results are in favor of a conjecture of Guillopé and Zworski on the existence of a fractal Weyl law for resonances.

1. Introduction and results

Resonances arise in spectral theory on noncompact Riemannian manifolds when one tries to figure out what the natural replacement data should be for the missing eigenvalues of the Laplacian. The basic problem of the mathematical theory of resonances is to relate the resonances spectrum (which is a discrete set of complex numbers) to the geometry of the underlying manifold and its geodesic flow. In this paper we will focus on a particular setting where the spectral and scattering theory are already well developed: infinite-area surfaces with constant negative curvature. For a detailed account of the spectral theory of infinite-area surfaces, we refer the reader to [Borthwick 2007]. Let \mathbb{H}^2 be the hyperbolic plane endowed with its standard metric of constant Gaussian curvature -1 . Let Γ be a geometrically finite discrete group of isometries acting on \mathbb{H}^2 . This means that Γ admits a finite sided polygonal fundamental domain in \mathbb{H}^2 . We will require that Γ has no *elliptic* elements different from the identity and that the quotient $\Gamma \backslash \mathbb{H}^2$ is of *infinite hyperbolic area*. Under these assumptions, the quotient space $X = \Gamma \backslash \mathbb{H}^2$ is a nice Riemann surface whose geometry can be described as follows. The surface X can be decomposed into a finite area surface with geodesic boundary N , called the Nielsen region, on which infinite-area ends F_i are glued: the funnels. We assume throughout that the number of funnels f is not zero. Each funnel F_i is isometric to a half-cylinder

$$F_i = (\mathbb{R}/l_i\mathbb{Z})_\theta \times (\mathbb{R}^+)_t,$$

where $l_i > 0$, with the warped metric

$$ds^2 = dt^2 + \cosh^2 t d\theta^2.$$

MSC2000: 11F72, 58J50.

Keywords: Laplacian, resonances, arithmetic fuchsian groups.

Jakobson was partially supported by NSERC, FQRNT and a Dawson fellowship. Naud was partially supported by ANR grants JC05-52556 and 09-JCJC-0099-01.

The Nielsen region N is itself decomposed into a compact surface K with geodesic and horocyclic boundary on which c noncompact, finite area ends C_i are glued: the cusps. A cusp C_i is isometric to a half-cylinder

$$C_i = (\mathbb{R}/h_i\mathbb{Z})_\theta \times ([1, +\infty))_y,$$

where $h_i > 0$, endowed with the familiar Poincaré metric

$$ds^2 = \frac{d\theta^2 + dy^2}{y^2}.$$

Let Δ_X be the hyperbolic Laplacian on X . Its spectrum on $L^2(X)$ has been described by Lax and Phillips [1984a; 1984b; 1985]: $[\frac{1}{4}, +\infty)$ is the continuous spectrum, and there are no embedded eigenvalues. The rest of the spectrum is made of a (possibly empty) finite set of eigenvalues, starting at $\delta(1-\delta)$, where $0 \leq \delta < 1$ is the Hausdorff dimension of the limit set of Γ . The fact that the bottom of the spectrum is related to the dimension δ was first pointed out by Patterson [1976] for convex cocompact groups (which amounts to saying that there are no cusps on X or equivalently, no *parabolic* elements in Γ). This result was later extended for geometrically finite groups by Sullivan [1979; 1984].

The dimension δ has another important interpretation. Let S_1X denote the unit tangent bundle; then the *trapped set* is defined as the set of points in S_1X whose orbit under the geodesic flow remains (after projection on X) in the Nielsen region N in the past and future. The Liouville measure of this set is always zero, but its Hausdorff dimension is actually $2\delta + 1$.

By the preceding description of the spectrum, the resolvent

$$R(\lambda) = (\Delta_X - \frac{1}{4} - \lambda^2)^{-1} : L^2(X) \rightarrow L^2(X),$$

is therefore well defined and analytic on the lower half-plane $\{\text{Im } \lambda < 0\}$ except at a possible finite set of poles corresponding to the finite point spectrum. *Resonances* are then defined as poles of the meromorphic continuation of

$$R(\lambda) : C_0^\infty(X) \rightarrow C^\infty(X)$$

to the whole complex plane. The set of poles is denoted by \mathcal{R}_X . This continuation is usually performed via the analytic Fredholm theorem after the construction of an adequate parametrix. The first result of this kind in the more general setting of asymptotically hyperbolic manifolds is due to Mazzeo and Melrose [1987]. A more precise parametrix for surfaces was constructed by Guillopé and Zworski [1995; 1997]; it allowed them to obtain global counting results for resonances of the following type. Let $N(R)$ be the number of resonances (counted with multiplicity) of modulus smaller than R . We have for all $R \geq 0$,

$$C^{-1}R^2 \leq N(R) \leq C + CR^2,$$

for some $C > 0$. Hence the set of resonances satisfies a quadratic growth law similar to the usual Weyl law for finite area surfaces. These bounds are actually valid for compact perturbations of the hyperbolic metric [Borthwick 2008], and in particular are not sensitive to the geometry of the trapped set. It is therefore necessary to examine finer properties of \mathcal{R}_X to recover some geometrical information on X . The most natural thing to do is to look at resonances that are close to the real axis. Physically, these are the most relevant resonances, because they correspond to metastable states that live the longest (the imaginary part corresponding to the decay rate). In the case of Schottky groups (equivalently, convex cocompact quotients

in dimension 2), a “fractal” upper bound was obtained in [Zworski 1999; Guillopé et al. 2004], namely

$$N_C(T) = O(T^{1+\delta}), \tag{1}$$

where

$$N_C(T) := \#\{z \in \mathcal{R}_X : \text{Im } z \leq C, |\text{Re } z| \leq T\}.$$

The first proof of a geometric bound of this type involving fractal dimension is due to Sjöstrand [1990] for potential scattering. This upper bound, together with numerical experiments, has led to the following conjecture, known as the *fractal Weyl law*.

Conjecture 1.1 (Guillopé–Zworski). There exist $C > 0$ and $A > 0$ such that for all T large enough,

$$A^{-1}T^{1+\delta} \leq N_C(T) \leq AT^{1+\delta}.$$

The only existing lower bound can be found in [Guillopé and Zworski 1999], where the authors show that for all $\epsilon > 0$, one can find $C_\epsilon > 0$ such that

$$N_{C_\epsilon}(T) = \Omega(T^{1-\epsilon}),$$

where $\Omega(\cdot)$ means being not a $O(\cdot)$; in other words, one can find a sequence $(T_i)_{i \in \mathbb{N}}$ with $T_i \rightarrow \infty$ such that

$$\lim_{i \rightarrow \infty} \frac{N_{C_\epsilon}(T_i)}{T_i^{1-\epsilon}} = +\infty.$$

This is a frustrating lower bound: not only it does not involve δ but it is not even optimal in the computable case of elementary groups where $N_C(T)$ grows linearly. Guillopé et al. [2004] actually prove a stronger statement than (1). Let $\mathcal{D}(z)$ be the number of resonances in the disc centered at z and radius one:

$$\mathcal{D}(z) := \#\{\lambda \in \mathcal{R}_X : |\lambda - z| \leq 1\}.$$

Then if $\text{Im } z \leq C$, we have $\mathcal{D}(z) = O(|\text{Re } z|^\delta)$, the implied constant depending solely on C . A similar statement for semiclassical Schrödinger operators can be found in [Sjöstrand and Zworski 2007]. Note that *if the Guillopé–Zworski conjecture holds*, then by the box principle, for all $\epsilon > 0$, one can find a sequence (z_i) with $|\text{Re } z_i| \rightarrow +\infty$ and $\text{Im } z_i \leq C$ such that for all $i \in \mathbb{N}$,

$$\mathcal{D}(z_i) \geq |\text{Re } z_i|^{\delta-\epsilon}. \tag{2}$$

To state our results, we need one more piece of notation. Let $A > 0$ and set

$$W_A = \{\lambda \in \mathbb{C} : \text{Im } \lambda \leq A \log(1 + |\text{Re } \lambda|)\}.$$

Guillopé and Zworski [1997] have shown that in logarithmic regions W_A , the density of resonances grows at least linearly. We shall prove the following thing.

Theorem 1.2. *Let Γ be a geometrically finite group as above. Assume that $\delta > \frac{1}{2}$, and fix arbitrarily small $\epsilon > 0$ and $A > 0$. Then there exists a sequence $(z_i)_{i \in \mathbb{N}}$ with $z_i \in W_A$ and $|\text{Re } z_i| \rightarrow +\infty$, such that for all $i \geq 0$,*

$$\mathcal{D}(z_i) \geq (\log |\text{Re } z_i|)^{(\delta-1/2)/\delta-\epsilon}.$$

In other words, the local density $\mathcal{D}(z)$ of resonances in logarithmic regions W_A is not bounded, and sensitive to the dimension of the trapped set. This implies in particular that the resonance set $\mathcal{R}_X \cap W_A$

is different from a lattice when $\delta > \frac{1}{2}$, which clearly could not follow from the existing lower bound in strips nor the global counting results. Building groups with $\delta > \frac{1}{2}$ is easy: if there is a parabolic element this is always the case and if one wants to consider only convex-cocompact groups, pinching a pair of pants will do it; see [Section 4](#). We point out that the proof is based on Dirichlet box arguments, a technique that has proved useful to obtain lower bounds for the remainder in Weyl's law on compact negatively curved manifolds; see [[Jakobson et al. 2008](#); [Jakobson and Polterovich 2007](#)].

It is possible to obtain significantly better lower bounds that are closer to (2), by using infinite-index subgroups of *arithmetic groups*. Arithmetic groups are algebraically defined discrete groups of isometries of \mathbb{H}^2 , the most celebrated being the modular group $\mathrm{PSL}_2(\mathbb{Z})$. For more details on definitions and references, see [Section 3](#). Our result is as follows.

Theorem 1.3. *Let Γ be a geometrically finite group as above, and assume that Γ is an infinite-index subgroup of an arithmetic group Γ_0 derived from a quaternion algebra. Suppose $\delta > \frac{3}{4}$, and fix arbitrarily small $\epsilon > 0$ and $A > 0$. Then there exists a sequence $(z_k) \in W_A$ with $|\mathrm{Re} z_k| \rightarrow +\infty$, such that for all $k \geq 0$,*

$$\mathcal{D}(z_k) \geq |\mathrm{Re} z_k|^{2\delta-3/2-\epsilon}.$$

This improvement is based on the very special structure of closed geodesics on arithmetic surfaces: the set of lengths has high multiplicities and good separation (see [Section 3](#) for more details). We point out that these techniques due to Selberg have been used recently by Anantharaman [[2009](#)] to obtain some results on the spectral deviations for the damped wave equation on compact arithmetic surfaces. This lower bound is clearly in favor of the Guillopé–Zworski conjecture, at least for the class of groups considered above. One may wonder at this point if [Theorem 1.3](#) is not empty: Gamburd [[2002](#)] has shown in (see [Section 4](#) for details) the existence of several geometrically finite subgroups Γ of $\mathrm{PSL}_2(\mathbb{Z})$ with dimension $\delta > \frac{3}{4}$. Another natural question is can we give a bound on the sequence $|\mathrm{Re} z_k|$? We explain at the end of [Section 3](#) how one can obtain a polynomial upper bound: for each $\epsilon > 0$ one can find an exponent $p_\epsilon > 0$ such that $|\mathrm{Re} z_k| = O(k^{p_\epsilon})$.

The lower bounds obtained above are to our knowledge the first examples in the literature which are related to the dimension of the trapped set, at least for fractal dimensions. Similar results should hold for higher dimensional convex-compact manifolds, by applying a similar strategy of proof based on the trace formula in [[Guillarmou and Naud 2006](#)].

The plan of the paper is as follows: in [Section 2](#) we recall the necessary material for the proofs, including the wave trace formula which is at the basis of our results. We then prove [Theorem 1.2](#) by a Dirichlet box-principle argument. [Section 3](#) is devoted to the case of arithmetically built groups. The heart of the proof is based on a trick of Selberg and Hejhal on mean square estimates. This is where the high multiplicity and the separation play a key role. In [Section 4](#) we discuss various examples of geometrically finite groups with δ large, and we construct an explicit family of convex cocompact subgroups of $\mathrm{PSL}_2(\mathbb{Z})$ with $\delta > \frac{3}{4}$.

2. Wave trace and log lower bounds

In this section, we prove [Theorem 1.2](#). Some of the technical estimates below will be of some use in the next section. We use the notation of the introduction. The constant $A > 0$ defining the logarithmic region W_A is set once for all.

The variant of Selberg’s trace formula we need here is due to [Guillopé and Zworski 1999]. We denote by \mathcal{P} the set of primitive closed geodesics on the surface $X = \Gamma \backslash \mathbb{H}^2$, and if $\gamma \in \mathcal{P}$, $l(\gamma)$ is the length. In the following, c is the number of cusps, and N is the Nielsen region. Let $\varphi \in C_0^\infty((0, +\infty))$, that is, a smooth function, compactly supported in \mathbb{R}_+^* . We have the identity

$$\sum_{\lambda \in \mathcal{R}_X} \widehat{\varphi}(-\lambda) = -\frac{\text{Vol}(N)}{4\pi} \int_0^{+\infty} \frac{\cosh(x/2)}{\sinh^2(x/2)} \varphi(x) dx + \frac{c}{2} \int_0^{+\infty} \frac{\cosh(x/2)}{\sinh(x/2)} \varphi(x) dx + \sum_{\gamma \in \mathcal{P}} \sum_{k \geq 1} \frac{l(\gamma)}{2 \sinh(kl(\gamma)/2)} \varphi(kl(\gamma)), \quad (3)$$

where $\widehat{\varphi}$ is the usual Fourier transform,

$$\widehat{\varphi}(\xi) = \int_{\mathbb{R}} \varphi(x) e^{-ix\xi} dx.$$

We recall that \mathcal{R}_X (except a possible finite number of term on the imaginary axis starting at $\lambda = i(\frac{1}{2} - \delta)$) is included in the upper half-plane. Note that we have omitted the main singular terms at $t = 0$ which are not relevant for our problem; see [Guillopé and Zworski 1999] for the formula in full detail. Proofs of Theorem 1.2 and 1.3 are based on the use of test functions of the form

$$\varphi_{t,\alpha}(x) = e^{-itx} \varphi_0(x - \alpha),$$

where $t > 0$, $\alpha > 0$ will be large and $\varphi_0 \in C_0^\infty(\mathbb{R})$ is a positive function, supported on the interval $[-1, +1]$ identical to 1 on $[-\frac{1}{2}, +\frac{1}{2}]$. The basic idea is to use the full-length spectrum (the set of lengths of closed geodesics) in the contribution from the geometric side instead of one single, closed primitive geodesic and its iterates as in the proof of Guillopé and Zworski [1999]. The price to pay for that is to lose positivity and deal with oscillating contributions. We start with some useful lemmas that consist mainly of brute force estimates. They will be used to control sums over resonances in the proof of Theorem 1.2 and 1.3. The reader can skip it for its first reading.

Lemma 2.1. *For all $N \geq 0$, one can find $C_N > 0$ such that for all $\xi \in \mathbb{C}$,*

$$|\widehat{\varphi_{t,\alpha}}(\xi)| \leq C_N \frac{e^{\alpha \text{Im } \xi + |\text{Im } \xi|}}{(1 + |t + \xi|)^N}.$$

Proof. Write $\widehat{\varphi_{t,\alpha}}(\xi) = e^{-i\alpha(t+\xi)} \widehat{\varphi_0}(t + \xi)$, and integrate by parts N times. Notice that while estimating $|\widehat{\varphi_0}(u)|$ with $u \in \mathbb{C}$, there is an extra factor $e^{|\text{Im } u|}$ coming out, which explain the presence of the (harmless) extra term $|\text{Im } \xi|$ in the exponents above. □

Lemma 2.2. *Let $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be either $f(x) = (\log(1 + x))^\beta$ or $f(x) = x^\beta$ with $0 < \beta < 1$. Assume that $\mathfrak{D}(z) = O(f(|\text{Re } z|))$ for all $z \in W_A$ with $|\text{Re } z|$ large enough. Then, for all α, t large and all $k \geq 0$, one has*

$$\left| \sum_{\lambda \in W_A \cap \mathcal{R}_X} \widehat{\varphi_{\alpha,t}}(-\lambda) \right| = O\left(\frac{e^{\alpha(\delta-1/2)}}{t^k}\right) + O(f(t)),$$

where the implied constants do not depend on α, t .

Proof. Let us assume that $\mathfrak{D}(z) = O(f(|\operatorname{Re} z|))$ whenever $|\operatorname{Re}(z)| \geq p_0 \geq 1$ and $z \in W_A$. Let $t > 0$ be so large that $t > p_0 + 1$, assume that $\alpha > 1$. By absolute convergence one can write

$$\sum_{\lambda \in W_A \cap \mathfrak{R}_X} \widehat{\varphi_{\alpha,t}}(-\lambda) = \sum_{p \in \mathbb{Z}} \sum_{\substack{p \leq \operatorname{Re} \lambda \leq p+1 \\ \lambda \in W_A \cap \mathfrak{R}_X}} \widehat{\varphi_{\alpha,t}}(-\lambda).$$

Let us set

$$S_p(\alpha, t) = \sum_{\substack{p \leq \operatorname{Re} \lambda \leq p+1 \\ \lambda \in W_A \cap \mathfrak{R}_X}} \widehat{\varphi_{\alpha,t}}(-\lambda).$$

We split the above sum as

$$\sum_{\lambda \in W_A \cap \mathfrak{R}_X} \widehat{\varphi_{\alpha,t}}(-\lambda) = \sum_{p < -p_0} S_p(\alpha, t) + \sum_{-p_0 \leq p \leq p_0} S_p(\alpha, t) + \sum_{p > p_0} S_p(\alpha, t).$$

The middle term involves only finitely many resonances $\lambda \in W_A$, and they satisfy $\operatorname{Im} \lambda \geq \frac{1}{2} - \delta$. Therefore using [Lemma 2.1](#), we have

$$\left| \sum_{-p_0 \leq p \leq p_0} S_p(\alpha, t) \right| \leq C_k \frac{e^{(-\alpha+1)(1/2-\delta)}}{(1+|t-p_0-1|)^k} \sum_{\substack{\lambda \in \mathfrak{R}_X \cap W_A \\ |\operatorname{Re} \lambda| \leq p_0}} 1 = O\left(\frac{e^{\alpha(\delta-1/2)}}{t^k}\right).$$

The first term can be estimated as

$$\left| \sum_{p < -p_0} S_p(\alpha, t) \right| \leq C_2 \sum_{p < -p_0} \frac{1}{(1+|p+1-t|)^2} \sum_{\substack{p \leq \operatorname{Re} \lambda \leq p+1 \\ \lambda \in \mathfrak{R}_X \cap W_A}} e^{(-\alpha+1)\operatorname{Im} \lambda},$$

while the last term is of size

$$\left| \sum_{p > p_0} S_p(\alpha, t) \right| \leq C_2 \sum_{p > p_0} \frac{\tilde{S}_p(\alpha)}{(1 + \min\{|p-t|, |p+1-t|\})^2},$$

where we have set

$$\tilde{S}_p(\alpha) = \sum_{\substack{p \leq \operatorname{Re} \lambda \leq p+1 \\ \lambda \in W_A \cap \mathfrak{R}_X}} e^{(-\alpha+1)\operatorname{Im} \lambda}.$$

The following lemma will be convenient (this is where the hypothesis on $\mathfrak{D}(z)$ is used).

Lemma 2.3. *Under the hypothesis of [Lemma 2.2](#), there exists a constant M , independent of α , p and such that for all $|p| \geq p_0$, we have*

$$\tilde{S}_p(\alpha) \leq Mf(|p|).$$

Let us postpone the proof of this result for a moment and show how to end the proof of [Lemma 2.2](#). Clearly, using [Lemma 2.3](#), the sum of the first and last terms is smaller than

$$C \sum_{p \in \mathbb{Z}} \frac{f(|p|)}{(1+|p-t|)^2},$$

for a constant $C > 0$ large enough. We can now write, denoting by $[t]$ the integer part of t ,

$$\sum_{p \in \mathbb{Z}} \frac{f(|p|)}{(1 + |p - t|)^2} = \sum_{q \in \mathbb{Z}} \frac{f(|q + [t]|)}{(1 + |q + [t] - t|)^2} \leq C' \sum_{q \in \mathbb{Z}} \frac{f(|q| + [t])}{(1 + |q|)^2},$$

again for a well chosen $C' > 0$ (we have used the fact that f is increasing). To end the proof, simply write

$$\sum_{q \in \mathbb{Z}} \frac{f(|q| + [t])}{(1 + |q|)^2} = \sum_{|q| \leq [t]} \frac{f(|q| + [t])}{(1 + |q|)^2} + \sum_{|q| > [t]} \frac{f(|q| + [t])}{(1 + |q|)^2},$$

which yields

$$\sum_{q \in \mathbb{Z}} \frac{f(|q| + [t])}{(1 + |q|)^2} \leq f(2[t]) \sum_{q \in \mathbb{Z}} \frac{1}{(1 + |q|)^2} + \sum_{|q| > [t]} \frac{f(2|q|)}{(1 + |q|)^2}.$$

Since $f(2|q|) = O(|q|^{1-\epsilon})$, the second term is clearly bounded in t and we get the upper bound of size $O(f(2t))$. It remains to prove [Lemma 2.3](#). It will follow from a standard covering argument. It is enough to consider just the case $p > p_0$. We recall that for all $\lambda \in \mathcal{R}_X$, then for $\text{Re } \lambda \neq 0$ we have actually $\text{Im } \lambda \geq 0$ by definition. Let \mathcal{A}_p denote the set

$$\mathcal{A}_p = \{z \in W_A : p \leq \text{Re } z \leq p + 1\},$$

let $D(z)$ denote the unit disc centered at $z \in \mathbb{C}$, and set

$$K(p) = \max\{k \geq 0 : k\sqrt{3} \leq A \log(1 + p)\}.$$

For $1 \leq k \leq K(p)$, we define the rectangle $R(k)$ by

$$R(k) = \{z \in \mathcal{A}_p : (k - 1)\sqrt{3} \leq \text{Im } z \leq k\sqrt{3}\}.$$

Set $l = A \log(1 + p) - K(p)\sqrt{3} < \sqrt{3}$. One can check that, for p large enough,

$$\mathcal{A}_p \subset \left(\bigcup_{k=1}^{K(p)} R(k) \right) \cup D\left(p + \frac{1}{2} + i(K(p) + l/2)\right) \cup D\left(p + \frac{1}{2} + i(K(p) + l)\right).$$

Indeed,

$$\mathcal{A}_p \setminus \left(\bigcup_{k=1}^{K(p)} R(k) \right)$$

is exactly the set

$$\{z \in \mathbb{C} : p \leq \text{Re } z \leq p + 1 \text{ and } K(p)\sqrt{3} \leq \text{Im } z \leq A \log(1 + \text{Re } z)\},$$

which is clearly covered by the union of the two above discs as long as

$$A \log(1 + p + 1) - A \log(1 + p) = A \log\left(1 + \frac{1}{p+1}\right) \leq \frac{\sqrt{3}}{2}.$$

Note that for all $k = 1, \dots, K(p)$, we have $R(k) \subset D(p + \frac{1}{2} + i\sqrt{3}(\frac{1}{2} + k - 1))$. We can now conclude by estimating

$$\begin{aligned} \tilde{S}_p(\alpha) &= \sum_{\lambda \in \mathfrak{A}_p \cap \mathfrak{R}_X} e^{(-\alpha+1)\operatorname{Im} \lambda} \\ &\leq \sum_{j=0}^{K(p)-1} \mathfrak{D}\left(p + \frac{1}{2} + i\sqrt{3}\left(\frac{1}{2} + j\right)\right) e^{(-\alpha+1)j\sqrt{3}} + \mathfrak{D}\left(p + \frac{1}{2} + i(K(p))\right) + \mathfrak{D}\left(p + \frac{1}{2} + i(K(p)) + \frac{1}{2}\right). \end{aligned}$$

Recalling that $\alpha > 1$ and $\mathfrak{D}(z) \leq Cf(|\operatorname{Re} z|)$ for all $z \in W_A$ with $|\operatorname{Re} z| \geq p_0$, we thus obtain

$$\tilde{S}_p(\alpha) \leq 2Cf\left(p + \frac{1}{2}\right) + C \frac{f\left(p + \frac{1}{2}\right)}{1 - e^{(-\alpha+1)\sqrt{3}}},$$

and therefore $\tilde{S}_p(\alpha) = O(f(p))$, uniformly in α . \square

Before we start the proof of [Theorem 1.2](#), we need one more lemma, which is the key observation that motivates the definition of the region W_A .

Lemma 2.4. *There exist some constants $\alpha_0, C_0 > 0$, independent of α, t such that for all $\alpha \geq \alpha_0$,*

$$\left| \sum_{\lambda \in \mathfrak{R}_X \setminus W_A} \widehat{\varphi_{\alpha,t}}(-\lambda) \right| \leq C_0.$$

Proof. We assume first that $\alpha > 1$. If $\lambda \notin W_A$, then $\operatorname{Im} \lambda \geq 0$ and

$$|\lambda|^2 = (\operatorname{Re} \lambda)^2 + (\operatorname{Im} \lambda)^2 \leq e^{(2/A)\operatorname{Im} \lambda} + (\operatorname{Im} \lambda)^2 \leq e^{(3/A)\operatorname{Im} \lambda},$$

whenever $\operatorname{Im} \lambda \geq C_A$ where C_A is a large enough constant depending on A . We can assume in the sequel that $C_A \geq 1$. Using [Lemma 2.1](#) with $N = 0$, we get

$$\left| \sum_{\lambda \in \mathfrak{R}_X \setminus W_A} \widehat{\varphi_{\alpha,t}}(-\lambda) \right| \leq C_0 \#\{\lambda \in \mathfrak{R}_X \setminus W_A : \operatorname{Im} \lambda \leq C_A\} + \sum_{\substack{\lambda \in \mathfrak{R}_X \setminus W_A \\ \operatorname{Im} \lambda \geq C_A}} \frac{1}{|\lambda|^{(\alpha-1)2A/3}}.$$

The first term is clearly independent of α while the second can be bounded by the Stieltjes integral

$$\sum_{\substack{\lambda \in \mathfrak{R}_X \setminus W_A \\ \operatorname{Im} \lambda \geq C_A}} \frac{1}{|\lambda|^{(\alpha-1)2A/3}} \leq \int_1^{+\infty} u^{-(\alpha-1)2A/3} dN(u),$$

where $N(u) = O(u^2)$ is the counting function for resonances in discs defined in [Section 1](#). By integration by parts, the above integral is clearly convergent and bounded in α as long as

$$A(\alpha - 1) > 3.$$

The proof is complete. \square

We can now start the proof of [Theorem 1.2](#). Let's test the trace formula [\(3\)](#) with the family $\varphi_{\alpha,t}$, where α is a large positive number:

$$\sum_{\lambda \in \mathfrak{R}_X} \widehat{\varphi_{\alpha,t}}(-\lambda) = -\frac{\text{Vol}(N)}{4\pi} \int_{\alpha-1}^{\alpha+1} \frac{\cosh(x/2)}{\sinh^2(x/2)} \varphi_{\alpha,t}(x) dx + \frac{c}{2} \int_{\alpha-1}^{\alpha+1} \frac{\cosh(x/2)}{\sinh(x/2)} \varphi_{\alpha,t}(x) dx$$

$$+ \sum_{\alpha-1 \leq kl(\gamma) \leq \alpha+1} \frac{l(\gamma)}{2 \sinh(kl(\gamma)/2)} e^{-itkl(\gamma)} \varphi_0(kl(\gamma) - \alpha).$$

The first two terms on the right side are clearly bounded with respect to α and t . To get an appropriate control on the sum

$$\mathcal{S}_{\alpha,t} := \sum_{\alpha-1 \leq kl(\gamma) \leq \alpha+1} \frac{l(\gamma)}{2 \sinh(kl(\gamma)/2)} e^{-itkl(\gamma)} \varphi_0(kl(\gamma) - \alpha),$$

we will use the following lemma, also known as the Dirichlet box theorem.

Lemma 2.5. *Let $\alpha_1, \dots, \alpha_N \in \mathbb{R}$, and $D \in \mathbb{N}^*$. For all $Q \geq 2$ one can find an integer $q \in \{D, \dots, DQ^N\}$ such that*

$$\max_{1 \leq j \leq N} \|q\alpha_j\| \leq \frac{1}{Q},$$

where $\|x\| = \min_{n \in \mathbb{Z}} |x - n|$.

Proof. We use the box principle. Set

$$N_\alpha := \#\{(k, l(\gamma)) \in \mathbb{N}^* \times \mathcal{P} : kl(\gamma) \in [\alpha - 1, \alpha + 1]\}.$$

Fix a constant $\varepsilon_0 > 0$ and set $D_\alpha = \lceil (4\pi)^{\varepsilon_0 N_\alpha} \rceil$. By [Lemma 2.5](#) with $Q = \lceil 4\pi \rceil$, for all $\alpha \gg 1$, one can find $q_\alpha \in \{D_\alpha, \dots, D_\alpha Q^{N_\alpha}\}$ such that

$$\max_{\alpha-1 \leq kl(\gamma) \leq \alpha+1} \|q_\alpha kl(\gamma)\| \leq \frac{1}{Q}.$$

Set $t_\alpha := 2\pi q_\alpha$, we have for all $\alpha - 1 \leq kl(\gamma) \leq \alpha + 1$,

$$|e^{it_\alpha kl(\gamma)} - 1| \leq \frac{2\pi}{Q} < \frac{2}{3}.$$

Hence we get

$$|\mathcal{S}_{\alpha,t_\alpha}| \geq \frac{1}{3} \left(\sum_{\alpha-1 \leq kl(\gamma) \leq \alpha+1} \frac{l(\gamma)}{2 \sinh(kl(\gamma)/2)} \varphi_0(kl(\gamma) - \alpha) \right) \geq C_0 e^{-\alpha/2} \sum_{\alpha-\frac{1}{2} \leq kl(\gamma) \leq \alpha+\frac{1}{2}} 1,$$

for a well chosen constant $C_0 > 0$. We now recall that by the prime geodesic theorem (see [\[Naud 2005\]](#) for a proof and references in the case of infinite-area surfaces), one has, as $T \rightarrow +\infty$,

$$\#\{(k, l(\gamma)) \in \mathbb{N}^* \times \mathcal{P} : kl(\gamma) \leq T\} = \frac{e^{\delta T}}{\delta T} (1 + o(1)).$$

This yields, for α large,

$$|\mathcal{S}_{\alpha,t_\alpha}| \geq C_1 \frac{e^{(\delta-1/2)\alpha}}{\alpha},$$

where C_1 is again a suitable constant. Using the prime geodesic theorem, one shows also that

$$C_2^{-1} \frac{e^{\delta\alpha}}{\alpha} \leq N_\alpha \leq C_2 e^{\delta\alpha},$$

with $C_2 > 0$ and α large. We have therefore $\log \log t_\alpha \leq \delta\alpha + \text{constants}$, which can be more conveniently restated as

$$\log \log t_\alpha \leq (\delta + \epsilon)\alpha \quad \text{for all } \epsilon > 0 \text{ and } \alpha \text{ large.}$$

Similarly we get the lower bound

$$\log \log t_\alpha \geq (\delta - \epsilon)\alpha.$$

We can now conclude the proof. Assume that $\delta > \frac{1}{2}$. Suppose that for all $z \in W_A$ with $|\operatorname{Re} z| \geq R_0$, one has $\mathfrak{D}(z) \leq (\log |\operatorname{Re} z|)^\beta$, where $\beta > 0$ will be determined later on. Then by [Lemma 2.2](#) with $k = 1$, and [Lemma 2.4](#), one gets as $\alpha \rightarrow +\infty$,

$$C_1 \frac{e^{(\delta-1/2)\alpha}}{\alpha} \leq |\mathcal{S}_{\alpha, t_\alpha}| \leq O(1) + O\left(\frac{e^{\alpha(\delta-1/2)}}{t_\alpha}\right) + O((\log t_\alpha)^\beta).$$

Now recall that

$$\frac{\log \log t_\alpha}{\delta + \epsilon} \leq \alpha \leq \frac{\log \log t_\alpha}{\delta - \epsilon},$$

so that we have

$$\frac{C_1(\delta + \epsilon)}{\log \log t_\alpha} (\log t_\alpha)^{(\delta-1/2)/(\delta+\epsilon)} \leq O(1) + O\left(\frac{(\log t_\alpha)^{(\delta-1/2)/(\delta-\epsilon)}}{t_\alpha}\right) + O((\log t_\alpha)^\beta).$$

We have a contradiction whenever $\beta < (\delta - \frac{1}{2})/(\delta + \epsilon)$. As a conclusion, for all $\epsilon > 0$ and all $R_0 \geq 0$ one can find $z \in W_A$ with $|\operatorname{Re} z| \geq R_0$ and $\mathfrak{D}(z) > (\log |\operatorname{Re} z|)^{((\delta-1/2)/(\delta)-\epsilon)}$. This proves [Theorem 1.2](#). \square

3. Mean square lower bounds and arithmetic length spectrum

The goal of this section is to prove [Theorem 1.3](#). First we need to recall a few basic facts about arithmetic group. Instead of detailing the construction of such groups, we refer the reader to the introductory book [\[Katok 1992\]](#), and will use a characterization of arithmetic groups derived from quaternion algebra due to Takeuchi [\[1975\]](#), which is all we need for this section.

We recall that a discrete group of isometries of the hyperbolic plane \mathbb{H}^2 can be viewed as a discrete subgroup of $\operatorname{PSL}_2(\mathbb{R})$. If $M \in \operatorname{PSL}_2(\mathbb{R})$ corresponds to a hyperbolic isometry, then $\operatorname{Tr} M$ is related to the translation length l of M by the formula $2 \cosh(l/2) = |\operatorname{Tr} M|$. Takeuchi's result is as follows.

Theorem 3.1 (Takeuchi). *Let Γ be a discrete, cofinite subgroup of $\operatorname{PSL}_2(\mathbb{R})$. Set*

$$\operatorname{Tr} \Gamma := \{\operatorname{Tr} T : T \in \Gamma\}.$$

Then Γ is derived from a quaternion algebra if and only if

- (1) *the field $K = \mathbb{Q}(\operatorname{Tr} \Gamma)$ is an algebraic field of finite degree and $\operatorname{Tr} \Gamma$ is a subset of the ring of integers of K , and*
- (2) *for all embeddings $\varphi : K \rightarrow \mathbb{C}$, $\varphi \neq \operatorname{Id}$, the set $\varphi(\operatorname{Tr} \gamma)$ is bounded in \mathbb{C} .*

For a proof of this characterization, see [Katok 1992; Takeuchi 1975]. Condition (2) has some strong implications on the structure of the trace set $\text{Tr } \Gamma$, as the next result shows. A similar statement can be found in [Luo and Sarnak 1994].

Lemma 3.2. *Let Γ_0 be an arithmetic group derived from a quaternion algebra.*

- (1) *There exists a constant $C_0 > 0$ depending solely on Γ_0 such that for all $x, x' \in \text{Tr } \Gamma_0$ with $x \neq x'$, $|x - x'| \geq C_0$.*
- (2) *There exists a constant M_0 depending only on Γ_0 such that for all R large,*

$$\Pi_0(x) := \#\{x \in \text{Tr } \Gamma_0 : |x| \leq R\} \leq M_0 R.$$

Proof. Clearly (1) implies (2) by a box argument. Let us prove (1). The field $K = \mathbb{Q}(\text{Tr } \Gamma_0)$ is a totally real number field of degree say $n = [K : \mathbb{Q}]$. Let $\varphi_1 = \text{Id}, \varphi_2, \dots, \varphi_n$ be the n distinct embeddings of K into \mathbb{C} . The set $\text{Tr } \Gamma_0$ is a subset of the ring of integers O_K of K . We denote by $N_{\mathbb{Q}}^K(\cdot)$ the norm on K . We recall that if $x \in O_K$ then $N_{\mathbb{Q}}^K(x) \in \mathbb{Z}$. Let $x \neq x'$ belong to $\text{Tr } \Gamma_0$, we have

$$1 \leq |N_{\mathbb{Q}}^K(x - x')| = \prod_{i=1}^n |\varphi_i(x - x')| \leq |x - x'| M^{n-1},$$

where $M > 0$ is given by property (2) of Takeuchi’s characterization. □

This important feature of the trace set was noticed by physicists working on quantum chaos [Bogomolny et al. 1997] and was clearly emphasized by Luo and Sarnak [1994] in their work on the number variance of arithmetic surfaces. Selberg and Hejhal [1976], when trying to obtain sharp lower bounds for the error term in Weyl’s law, had already noticed similar properties for some examples of cocompact arithmetic groups.

In the rest of this section we will work with a geometrically finite group Γ as defined in Section 1, and we assume in addition that Γ is an infinite-index subgroup of an arithmetic group Γ_0 , derived from a quaternion algebra. The simplest examples of such groups Γ that one can think of are finitely generated Schottky subgroups of $\text{PSL}_2(\mathbb{Z})$, but there are definitely many other examples, see the next section.

Given such a group Γ , one can define the length spectrum of $X = \Gamma \backslash \mathbb{H}^2$ by

$$\mathcal{L}_\Gamma := \{kl(\gamma) : (k, \gamma) \in \mathbb{N}^* \times \mathcal{P}\},$$

where as in the preceding section, \mathcal{P} is the set of primitive closed geodesics. We have the following key properties.

Proposition 3.3. *Let Γ be a fuchsian group as above, then we have:*

- (1) *Let $l_1, l_2 \in \mathcal{L}_\Gamma$ with $2 \cosh(l_i/2) = \text{Tr } M_i, i \in \{1, 2\}$, then*

$$|l_1 - l_2| \geq e^{-(\max(l_1, l_2))/2} |\text{Tr } M_1 - \text{Tr } M_2|.$$

- (2) *There exists a constant $C_1 > 0$ depending only on Γ_0 such that for all α large,*

$$\#\{l \in \mathcal{L}_\Gamma : \alpha - 1 \leq l \leq \alpha + 1\} \leq C_1 e^{\alpha/2}.$$

Proof. The set of closed geodesics on $X = \Gamma \backslash \mathbb{H}^2$ is in one-to-one correspondence with the set of conjugacy classes of hyperbolic elements in the fundamental group Γ , each closed geodesic γ having its length $l(\gamma)$ given by the formula

$$2 \cosh(l(\gamma)/2) = |\text{Tr } T_\gamma|,$$

where $T_\gamma \in \Gamma$ is an hyperbolic isometry. The length spectrum \mathcal{L}_Γ is therefore in one-to-one correspondence with the trace set $\text{Tr } \Gamma$ via the above formula (except for the conjugacy classes of parabolic elements with trace 2). Since we have $\text{Tr } \Gamma \subset \text{Tr } \Gamma_0$, we can use the preceding Lemma and crude bounds to prove estimate (2). To obtain the first lower bound (1), one simply writes (assuming $l_2 > l_1$),

$$l_2 - l_1 = 2 \int_{x_1}^{x_2} \frac{dt}{t} \geq 2 \frac{x_2 - x_1}{x_2},$$

where we have

$$x_i = e^{l_i/2} = \frac{1}{2}(\text{Tr } M_i + \sqrt{(\text{Tr } M_i)^2 - 4}).$$

Clearly one gets

$$x_2 - x_1 = \frac{1}{2} \int_{\text{Tr } M_1}^{\text{Tr } M_2} \left(1 + \frac{u}{\sqrt{u^2 - 4}}\right) du \geq \frac{1}{2}(\text{Tr } M_2 - \text{Tr } M_1),$$

and the proof is done. □

When compared with the prime geodesic theorem (see Section 2), estimate (2) in the proposition shows that whenever $\delta > \frac{1}{2}$ there must be some exponentially large multiplicities in the length spectrum. This is the key observation of Selberg and Hejhal [Hejhal 1976, Section 18, Chapter 2] that will allow us to produce a better lower bound than in Section 2. More precisely:

Proposition 3.4. *Let Γ be a group as above, δ being the dimension of its limit set. Let $\mathcal{P}_{\alpha,t}$ be the sum defined by*

$$\mathcal{P}_{\alpha,t} := \sum_{\alpha-1 \leq kl(\gamma) \leq \alpha+1} \frac{l(\gamma)}{2 \sinh(kl(\gamma)/2)} e^{-itkl(\gamma)} \varphi_0(kl(\gamma) - \alpha).$$

There exists a constant $A > 0$ such that for all T large, if one sets $\alpha = 2 \log T - A$ then the integral $\mathcal{F}(T)$ defined by

$$\mathcal{F}(T) = \int_T^{3T} \left(1 - \frac{|t - 2T|}{T}\right) |\mathcal{P}_{\alpha,t}|^2 dt,$$

enjoys the lower bound

$$\mathcal{F}(T) \geq C_2 \frac{T^{1+4\delta-3}}{(\log T)^2},$$

for some constant $C_2 > 0$ independent of T .

Let us show how Theorem 1.3 follows from this lower bound. First we assume that for all $z \in W_A$ with $|\text{Re } z| \geq R_0$, we have

$$\mathcal{D}(z) \leq |\text{Re } z|^\beta,$$

for some $0 < \beta < 1$. Set $\alpha = 2 \log T - A$, where A is given by the above proposition. We have

$$C_2 \frac{T^{1+4\delta-3}}{(\log T)^2} \leq \mathcal{F}(T) \leq \int_T^{3T} |\mathcal{P}_{\alpha,t}|^2 dt.$$

By the trace formula (3) applied to $\varphi_{\alpha,t}$, and Lemma 2.2 with $k = 2$, Lemma 2.4, we have

$$|\mathcal{S}_{\alpha,t}| \leq O(1) + O\left(\frac{T^{2\delta-1}}{T^2}\right) + O(t^\beta);$$

therefore $\int_T^{3T} |\mathcal{S}_{\alpha,t}|^2 dt = O(T^{2\beta+1})$, which produces a contradiction whenever $\beta < 2\delta - 3/2$. This proves Theorem 1.3. \square

Proof of Proposition 3.4. We start with an elementary observation. For all $\lambda \in \mathbb{R}$ and $T > 0$ set

$$J(T, \lambda) = \int_T^{3T} \left(1 - \frac{|t - 2T|}{T}\right) e^{-i\lambda t} dt.$$

Lemma 3.5. *With the preceding notation, we have $|J(T, \lambda)| \leq \frac{4}{\lambda^2 T}$ for all $\lambda \neq 0$, while $J(T, 0) = T$.*

Proof. This follows by direct computation. \square

At this point we need some more notation. If $\ell \in \mathcal{L}_\Gamma$, we denote by $\mu(\ell)$ the multiplicity of ℓ as the length of a closed geodesic. If $\ell \in \mathcal{L}_\Gamma$, then let $\tilde{\ell}$ denote the primitive length of ℓ , that is, if $\ell = k\ell(\gamma)$ with γ a primitive closed geodesic, then $\tilde{\ell} = \ell(\gamma)$. Using this notation, we have

$$\mathcal{I}(T) = \sum_{\ell, \ell' \in \mathcal{L}_\Gamma} \frac{\tilde{\ell} \tilde{\ell}' \mu(\ell) \mu(\ell')}{4 \sinh(\ell/2) \sinh(\ell'/2)} J(T, \ell - \ell') \varphi_0(\ell - \alpha) \varphi_0(\ell' - \alpha).$$

We now set $\mathcal{I}(T) = \mathcal{I}_1(T) + \mathcal{I}_2(T)$, where

$$\mathcal{I}_1(T) = T \sum_{\ell \in \mathcal{L}_\Gamma} \frac{(\tilde{\ell} \mu(\ell))^2}{4 \sinh^2(\ell/2)} \varphi_0^2(\ell - \alpha)$$

and

$$\mathcal{I}_2(T) = \sum_{\substack{\ell, \ell' \in \mathcal{L}_\Gamma \\ \ell \neq \ell'}} \frac{\tilde{\ell} \tilde{\ell}' \mu(\ell) \mu(\ell')}{4 \sinh(\ell/2) \sinh(\ell'/2)} J(T, \ell - \ell') \varphi_0(\ell - \alpha) \varphi_0(\ell' - \alpha).$$

By Lemma 3.5, we have

$$|\mathcal{I}_2(T)| \leq \frac{4}{T} \sum_{\substack{\ell, \ell' \in \mathcal{L}_\Gamma \\ \ell \neq \ell'}} \frac{\tilde{\ell} \tilde{\ell}' \mu(\ell) \mu(\ell') \varphi_0(\ell - \alpha) \varphi_0(\ell' - \alpha)}{4 \sinh(\ell/2) \sinh(\ell'/2) (\ell - \ell')^2}.$$

Using the inequality $ab \leq \frac{1}{2}(a^2 + b^2)$ for all $a, b \in \mathbb{R}$, we get by the symmetry of the summation

$$|\mathcal{I}_2(T)| \leq \frac{4}{T} \sum_{\substack{\ell, \ell' \in \mathcal{L}_\Gamma \\ \ell \neq \ell'}} \frac{(\tilde{\ell} \mu(\ell))^2 \varphi_0^2(\ell - \alpha)}{4 \sinh(\ell/2) \sinh(\ell'/2) (\ell - \ell')^2}.$$

Therefore, one can find a constant $C > 0$ such that, for all α and T large,

$$|\mathcal{I}_2(T)| \leq C \frac{e^{-\alpha}}{T} \sum_{\ell \in \mathcal{L}_\Gamma} (\tilde{\ell} \mu(\ell))^2 \varphi_0^2(\ell - \alpha) \sum_{\substack{\ell' \in \mathcal{L}_\Gamma \cap [\alpha-1, \alpha+1] \\ \ell' \neq \ell}} \frac{1}{(\ell - \ell')^2}.$$

By [Proposition 3.3\(1\)](#), we can write $x = 2 \cosh(\ell/2)$, where $x \in \text{Tr } \Gamma$, and thus

$$\sum_{\substack{\ell' \in \mathcal{L}_\Gamma \cap [\alpha-1, \alpha+1] \\ \ell' \neq \ell}} \frac{1}{(\ell - \ell')^2} \leq e^{\alpha+1} \sum_{x' \in \text{Tr } \Gamma} \frac{1}{(x - x')^2}.$$

We can now bound

$$\sum_{x' \in \text{Tr } \Gamma} \frac{1}{(x - x')^2} \leq \int_2^{x-C_0} \frac{d\Pi_0(u)}{(x-u)^2} + \int_{x+C_0}^{+\infty} \frac{d\Pi_0(u)}{(x-u)^2},$$

where Π_0 is the counting function for the trace set of the arithmetic group Γ_0 and the constant C_0 is given by [Lemma 3.2](#). Using the fact that the growth $\Pi_0(u) = O(u)$, two Stieltjes integration by parts show that there exists a constant \tilde{C}_0 depending only on Γ_0 such that for all $x \in \text{Tr } \Gamma$,

$$\sum_{x' \in \text{Tr } \Gamma} \frac{1}{(x - x')^2} \leq \tilde{C}_0.$$

Going back to $\mathcal{F}_2(T)$, we have obtained for T and α large,

$$|\mathcal{F}_2(T)| \leq \frac{C'}{T} \sum_{\ell \in \mathcal{L}_\Gamma} (\tilde{\ell}\mu(\ell))^2 \varphi_0^2(\ell - \alpha).$$

Recall that

$$\mathcal{F}_1(T) = T \sum_{\ell \in \mathcal{L}_\Gamma} \frac{(\tilde{\ell}\mu(\ell))^2}{4 \sinh^2(\ell/2)} \varphi_0^2(\ell - \alpha) \geq C'' e^{-\alpha} T \sum_{\ell \in \mathcal{L}_\Gamma} (\tilde{\ell}\mu(\ell))^2 \varphi_0^2(\ell - \alpha),$$

again for α large and some $C'' > 0$. Therefore $|\mathcal{F}_2| \leq \frac{1}{2} \mathcal{F}_1$ as long as

$$e^\alpha \leq \frac{1}{2} \frac{C''}{C'} T^2,$$

which is definitely achieved if one sets $\alpha = 2 \log T - A$, where $A \gg 1$. We have thus

$$|\mathcal{F}(T)| \geq \frac{1}{2} |\mathcal{F}_1(T)| \geq C'' e^{-\alpha} T \sum_{\ell \in \mathcal{L}_\Gamma \cap [\alpha-1, \alpha+1]} (\tilde{\ell}\mu(\ell))^2 \varphi_0^2(\ell - \alpha) \geq \tilde{C}'' e^{-\alpha} T \sum_{\ell \in \mathcal{L}_\Gamma \cap [\alpha-\frac{1}{2}, \alpha+\frac{1}{2}]} (\mu(\ell))^2,$$

for some $\tilde{C}'' > 0$. By Schwarz inequality we get

$$\left(\sum_{\alpha-\frac{1}{2} \leq kl(\gamma) \leq \alpha+\frac{1}{2}} 1 \right)^2 = \left(\sum_{\ell \in \mathcal{L}_\Gamma \cap [\alpha-\frac{1}{2}, \alpha+\frac{1}{2}]} \mu(\ell) \right)^2 \leq \left(\sum_{\ell \in \mathcal{L}_\Gamma \cap [\alpha-\frac{1}{2}, \alpha+\frac{1}{2}]} (\mu(\ell))^2 \right) \left(\sum_{\ell \in \mathcal{L}_\Gamma \cap [\alpha-\frac{1}{2}, \alpha+\frac{1}{2}]} 1 \right).$$

By [Proposition 3.3\(2\)](#),

$$\sum_{\ell \in \mathcal{L}_\Gamma \cap [\alpha-\frac{1}{2}, \alpha+\frac{1}{2}]} 1 = O(e^{\alpha/2}),$$

while the prime geodesic theorem yields

$$\sum_{\alpha-\frac{1}{2} \leq kl(\gamma) \leq \alpha+\frac{1}{2}} 1 \geq B \frac{e^{\delta\alpha}}{\alpha},$$

where $B > 0$. Hence we have obtained

$$\sum_{\ell \in \mathcal{L}_\Gamma \cap [\alpha - \frac{1}{2}, \alpha + \frac{1}{2}]} (\mu(\ell))^2 \geq B^2 \frac{e^{(2\delta-1/2)\alpha}}{\alpha^2}.$$

Going back to $\mathcal{F}(T)$ and recalling that $\alpha = 2 \log T - A$ we get

$$|\mathcal{F}(T)| \geq B' \frac{T^{1+4\delta-3}}{(\log T)^2}.$$

The proof is now complete. □

It is now time to indicate how to get upper bounds on the sequence $|\operatorname{Re} z_k|$ as $k \rightarrow \infty$. First, we can notice that the above [Proposition 3.4](#) still holds on shorter intervals. Indeed, pick any $0 < \rho < 1$ and set

$$\mathcal{F}_\rho(T) = \int_{2T-T^\rho}^{2T+T^\rho} \left(1 - \frac{|t-2T|}{T^\rho}\right) |\mathcal{G}_{\alpha,t}|^2 dt,$$

then one can show that taking $\alpha = 2\rho \log T - A$, for some $A \gg 1$, there exists a constant $C_\rho > 0$ such that for T large one has

$$\mathcal{F}_\rho(T) \geq C_\rho \frac{T^{(4\delta-3)\rho+\rho}}{(\log T)^2}.$$

The assumption of [Lemma 2.2](#) can be weakened: indeed to get the desired upper bound on $|\mathcal{G}_{\alpha,t}| = O(t^\beta)$, it is enough to assume that

$$\mathfrak{D}(z) = O(|\operatorname{Re} z|^\beta)$$

for all $z \in W_A$ and $\operatorname{Re} z \in [2t - t^\mu, 2t + t^\mu]$, for some $0 < \mu < 1$. These two minor modifications allow to obtain a more precise statement (by following the same line of proof). For all $\epsilon > 0$, one can find an exponent $1 > \rho_\epsilon > 0$ such that for all T large, there exists $z \in W_A$ with the property

$$\operatorname{Re} z \in [2T - T^{\rho_\epsilon}, 2T + T^{\rho_\epsilon}] \text{ and } \mathfrak{D}(z) \geq \operatorname{Re} z^{2\delta-3/2-\epsilon}.$$

Choose $1 > \mu_\epsilon > \rho_\epsilon$ and define by induction a sequence (T_k) by $T_0 \gg 1$ and for all $k \geq 0$, $T_{k+1} = T_k + (T_k)^{\mu_\epsilon}$. For all $k \geq 0$, set

$$I_k = [2T_k - (T_k)^{\rho_\epsilon}, 2T_k + (T_k)^{\rho_\epsilon}].$$

For all $k \geq 0$, one can find $z_k \in W_A$ with

$$\operatorname{Re} z_k \in I_k \text{ and } \mathfrak{D}(z_k) \geq (\operatorname{Re} z_k)^{2\delta-3/2-\epsilon}.$$

Moreover because $\mu_\epsilon > \rho_\epsilon$, we have $D(z_k) \cap D(z_{k+1}) = \emptyset$ for k large. To obtain the leading behavior of T_k as $k \rightarrow +\infty$, one can perform a change of variable $x_k = 1/T_k$ and consider the dynamical system on the real line given by

$$f_{\mu_\epsilon}(x) = \frac{x}{1 + x^{1-\mu_\epsilon}}.$$

Clearly 0 is a neutral fixed point for f_{μ_ϵ} and for all $x_0 > 0$,

$$x_k = f_{\mu_\epsilon}^{(k)}(x_0) > 0$$

tends to 0 as $k \rightarrow +\infty$. Remark that since we have for $x \leq 1$,

$$f_{\mu_\epsilon}(x) \leq \frac{x}{1+x},$$

we get the crude upper bound

$$x_k = O\left(\frac{1}{k}\right).$$

To obtain an asymptotic estimate, we set $u_k = (x_k)^\alpha$, where α will be determined later on. Writing

$$u_N - u_0 = \sum_{k=0}^{N-1} f_\mu(x_k)^\alpha - x_k^\alpha,$$

and since we have the local expansion at $x = 0$

$$f_\mu(x)^\alpha - x^\alpha = -\alpha x^{1-\mu+\alpha} + O(x^{2-2\mu+\alpha}),$$

the choice of $\alpha = \mu - 1$ yields as $N \rightarrow +\infty$,

$$u_N = (1 - \mu)N + O\left(\sum_{k=1}^N \frac{1}{k^{1-\mu}}\right) = (1 - \mu)N + O(N^\mu).$$

Therefore

$$\lim_{k \rightarrow \infty} (1 - \mu_\epsilon)^{1/(1-\mu_\epsilon)} k^{1/(1-\mu_\epsilon)} x_k = 1.$$

Thus we have the polynomial bound $|\operatorname{Re} z_k| = O(k^{1/(1-\mu_\epsilon)})$. Clearly the exponent $p_\epsilon = \frac{1}{1-\mu_\epsilon}$ will tend to infinity as ϵ goes to 0.

4. Examples

In this section we discuss briefly examples of surfaces $X = \Gamma \backslash \mathbb{H}^2$ satisfying the assumptions of Theorems 1.2 and 1.3. We assume that the reader has some basic knowledge in fuchsian groups and hyperbolic geometry, for which we refer to [Katok 1992]. By the work of Patterson [1976], we know that if X has at least one cusp, that is, if Γ has at least one nontrivial parabolic element, then the dimension $\delta > \frac{1}{2}$. If one wants examples without cusps, then δ can be made arbitrarily close to 1 by “pinching” the geodesic boundary of Nielsen’s region. Let us explain what we mean. By [Patterson 1976] and the spectral analysis in [Lax and Phillips 1984a; 1984b; 1985], we have $\delta > \frac{1}{2}$ if and only if $\lambda_0(X) < \frac{1}{4}$, where $\lambda_0(X)$ is the bottom of the spectrum of the Laplacian Δ_X . In that case $\lambda_0(X) = \delta(1 - \delta)$. Hence to get $\delta > \frac{1}{2}$, it is enough to show that the Rayleigh quotient

$$\lambda_0(X) = \inf_{f \neq 0} \frac{\int_X |\nabla f|^2 d\operatorname{Vol}}{\int_X f^2 d\operatorname{Vol}} < \frac{1}{4},$$

where f is an L^2 function on X with an L^2 gradient ∇f . Based on the above formula, Pignataro and Sullivan proved the following, where $\ell(X)$ denotes the maximum length of the closed geodesics which are the boundary of the Nielsen region of X (the convex core):

Proposition 4.1 [Pignataro and Sullivan 1986]. *There exists a constant $C(X) > 0$ depending only the topology of X such that*

$$\lambda_0(X) \leq C(X)\ell(X).$$

Therefore if $\ell(X)$ is small enough, one definitely has $\delta > \frac{1}{2}$. Applying the same strategy to find examples satisfying the hypothesis of [Theorem 1.3](#) is harder. Indeed, the discreteness of arithmetic groups makes it difficult to perform deformations. What we are looking for are geometrically finite, infinite-index subgroups Γ of arithmetic groups derived from quaternion algebras with $\delta(\Gamma) > \frac{3}{4}$. The easiest thing to do is to consider first $\mathrm{PSL}_2(\mathbb{Z})$ and look at some of its subgroups.

Let us first consider the group Λ_N obtained as

$$\Lambda_N := \langle g_0, g_1, \dots, g_N \rangle,$$

where

$$g_0(z) = \frac{-1}{z} \simeq \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad g_k = \tau^k g_0 \tau^{-k}, \quad \tau(z) = z + 2 \simeq \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}.$$

Let D_j , $j = 0, \dots, N$ be the unit closed disc centered at $2j$. A fundamental domain for the action of Λ_N on \mathbb{H}^2 is given by

$$\mathcal{F} = \overline{\mathbb{H}^2 \setminus (D_0 \cup \dots \cup D_N)}.$$

The group Λ_N is therefore geometrically finite and has no parabolic elements, despite the presence of (false) ‘‘cusps’’ in the fundamental domain. The elliptic elements are the conjugacy classes of g_0, \dots, g_N , which are of order 2. Up to a covering of order 2, we can get rid of them.

For $k = 1, \dots, N$, set $h_k = g_0 g_k$, and consider the subgroup

$$\Gamma_N = \langle h_1, \dots, h_N; h_1^{-1}, \dots, h_N^{-1} \rangle,$$

then it is easy to see that Γ_N is a subgroup of Λ_N of index 2 and has no elliptic elements, hence a convex cocompact group. Because Γ_N is of finite index the critical exponents $\delta(\Gamma_N)$ and $\delta(\Lambda_N)$ are the same: the critical exponent is defined as the infimum of positive real numbers σ such that the Poincaré series

$$P(\sigma) := \sum_{\gamma \in \Gamma} e^{-\sigma d(i, \gamma i)},$$

are convergent. Here d is the hyperbolic distance in the half-plane model. A classical result of Sullivan [1984] shows that for geometrically finite groups, the critical exponent is also the Hausdorff dimension of the limit set, hence Λ_N and Γ_N have same dimension for their limit set. The group Λ_N is also considered in [Gamburd 2002], where he shows using a min-max argument and a suitable test function that $\delta(\Lambda_N)$ can be made as close to 1 as we want, provided N is large enough (estimates are effective).

An alternative way to construct similar convex cocompact subgroups of $\mathrm{PSL}_2(\mathbb{Z})$ with δ close to 1 is given in [Bourgain and Kontorovich 2010]. The idea is to start with the free subgroup $\Gamma(2) = \langle A, B, A^{-1}, B^{-1} \rangle$ generated by

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}.$$

Its commutator subgroup is a free, infinitely generated subgroup with critical exponent 1. Moreover it has no parabolic elements. This commutator subgroup contains finitely generated (hence convex cocompact) subgroups with critical exponent δ arbitrarily close to 1.

As a conclusion, we have found several examples of convex cocompact subgroups of $\mathrm{PSL}_2(\mathbb{Z})$ with $\delta > \frac{3}{4}$. By a similar technique, one can produce several examples with cusps. In that direction, let us point out that the Hecke group Γ_3 generated by $g : z \mapsto -1/z$ and $h : z \mapsto z + 3$ is a good candidate: its Hausdorff dimension was estimated by Phillips and Sarnak [1985] to be $\delta = 0.753 \pm 0.003$. Can one prove (or disprove) rigorously that $\delta > 0.75$?

It would be interesting in itself to find similar constructions for arithmetic groups that were not considered in this paper. In a sequel, the authors plan to address the case of arithmetic groups derived from quaternion division algebras (which are cocompact surface groups). It would also be interesting to consider groups acting on higher-dimensional hyperbolic spaces, for example arithmetic Kleinian groups.

Acknowledgments

Both authors thank the Banff International Research Station, where part of this work was done, for its hospitality. This paper benefited from interesting discussion with Nalini Anantharaman. We also thank Iosif Polterovich and Julie Rowlett for the stimulating discussions that led to this paper.

References

- [Anantharaman 2009] N. Anantharaman, “Spectral deviations for the damped wave equation”, preprint, 2009. [arXiv 0904.1736](#)
- [Bogomolny et al. 1997] E. B. Bogomolny, B. Georgeot, M.-J. Giannoni, and C. Schmit, “Arithmetical chaos”, *Phys. Rep.* **291**:5-6 (1997), 219–324. [MR 99c:11062](#)
- [Borthwick 2007] D. Borthwick, *Spectral theory of infinite-area hyperbolic surfaces*, Progress in Mathematics **256**, Birkhäuser, Boston, MA, 2007. [MR 2008h:58056](#) [Zbl 1130.58001](#)
- [Borthwick 2008] D. Borthwick, “Upper and lower bounds on resonances for manifolds hyperbolic near infinity”, *Comm. Partial Differential Equations* **33**:7-9 (2008), 1507–1539. [MR 2009i:58039](#) [Zbl 1168.58012](#)
- [Bourgain and Kontorovich 2010] J. Bourgain and A. Kontorovich, “On representations of integers in thin subgroups of $\mathrm{SL}_2(\mathbb{Z})$ ”, preprint, 2010. [arXiv 1001.4534](#)
- [Gamburd 2002] A. Gamburd, “On the spectral gap for infinite index “congruence” subgroups of $\mathrm{SL}_2(\mathbb{Z})$ ”, *Israel J. Math.* **127** (2002), 157–200. [MR 2003b:11050](#)
- [Guillarmou and Naud 2006] C. Guillarmou and F. Naud, “Wave 0-trace and length spectrum on convex co-compact hyperbolic manifolds”, *Comm. Anal. Geom.* **14**:5 (2006), 945–967. [MR 2008f:58032](#) [Zbl 1127.58028](#)
- [Guillopé and Zworski 1995] L. Guillopé and M. Zworski, “Upper bounds on the number of resonances for non-compact Riemann surfaces”, *J. Funct. Anal.* **129**:2 (1995), 364–389. [MR 96b:58116](#) [Zbl 0841.58063](#)
- [Guillopé and Zworski 1997] L. Guillopé and M. Zworski, “Scattering asymptotics for Riemann surfaces”, *Ann. of Math. (2)* **145**:3 (1997), 597–660. [MR 98g:58181](#) [Zbl 0898.58054](#)
- [Guillopé and Zworski 1999] L. Guillopé and M. Zworski, “The wave trace for Riemann surfaces”, *Geom. Funct. Anal.* **9**:6 (1999), 1156–1168. [MR 2001a:11086](#) [Zbl 0947.58022](#)
- [Guillopé et al. 2004] L. Guillopé, K. K. Lin, and M. Zworski, “The Selberg zeta function for convex co-compact Schottky groups”, *Comm. Math. Phys.* **245**:1 (2004), 149–176. [MR 2005f:11193](#) [Zbl 1075.11059](#)
- [Hejhal 1976] D. A. Hejhal, *The Selberg trace formula for $\mathrm{PSL}(2, R)$, I*, Lecture Notes in Math. **548**, Springer, Berlin, 1976. [MR 55 #12641](#) [Zbl 0347.10018](#)
- [Jakobson and Polterovich 2007] D. Jakobson and I. Polterovich, “Estimates from below for the spectral function and for the remainder in local Weyl’s law”, *Geom. Funct. Anal.* **17**:3 (2007), 806–838. [MR 2009h:35302](#) [Zbl 1161.58012](#)

- [Jakobson et al. 2008] D. Jakobson, I. Polterovich, and J. A. Toth, “A lower bound for the remainder in Weyl’s law on negatively curved surfaces”, *Int. Math. Res. Not.* **2008**:2 (2008). [MR 2009f:58038](#)
- [Katok 1992] S. Katok, *Fuchsian groups*, University of Chicago Press, Chicago, 1992. [MR 93d:20088](#) [Zbl 0753.30001](#)
- [Lax and Phillips 1984a] P. D. Lax and R. S. Phillips, “Translation representation for automorphic solutions of the wave equation in non-Euclidean spaces, I”, *Comm. Pure Appl. Math.* **37**:3 (1984), 303–328. [MR 86c:58148](#) [Zbl 0549.10024](#)
- [Lax and Phillips 1984b] P. D. Lax and R. S. Phillips, “Translation representations for automorphic solutions of the wave equation in non-Euclidean spaces, II”, *Comm. Pure Appl. Math.* **37**:6 (1984), 779–813. [MR 86h:58140](#) [Zbl 0549.10019](#)
- [Lax and Phillips 1985] P. D. Lax and R. S. Phillips, “Translation representations for automorphic solutions of the wave equation in non-Euclidean spaces, III”, *Comm. Pure Appl. Math.* **38**:2 (1985), 179–207. [MR 86j:58150](#) [Zbl 0578.10033](#)
- [Luo and Sarnak 1994] W. Luo and P. Sarnak, “Number variance for arithmetic hyperbolic surfaces”, *Comm. Math. Phys.* **161**:2 (1994), 419–432. [MR 95k:11076](#) [Zbl 0797.58069](#)
- [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](#)
- [Naud 2005] F. Naud, “Precise asymptotics of the length spectrum for finite-geometry Riemann surfaces”, *Int. Math. Res. Not.* **2005** (2005), 299–310. [Zbl 1073.37021](#)
- [Patterson 1976] S. J. Patterson, “The limit set of a Fuchsian group”, *Acta Math.* **136**:3-4 (1976), 241–273. [MR 56 #8841](#) [Zbl 0336.30005](#)
- [Phillips and Sarnak 1985] R. S. Phillips and P. Sarnak, “On the spectrum of the Hecke groups”, *Duke Math. J.* **52**:1 (1985), 211–221. [MR 86j:11042](#) [Zbl 0564.30030](#)
- [Pignataro and Sullivan 1986] T. Pignataro and D. Sullivan, “Ground state and lowest eigenvalue of the Laplacian for noncompact hyperbolic surfaces”, *Comm. Math. Phys.* **104**:4 (1986), 529–535. [MR 87m:58178](#)
- [Sjöstrand 1990] J. Sjöstrand, “Geometric bounds on the density of resonances for semiclassical problems”, *Duke Math. J.* **60**:1 (1990), 1–57. [MR 91e:35166](#) [Zbl 0702.35188](#)
- [Sjöstrand and Zworski 2007] J. Sjöstrand and M. Zworski, “Fractal upper bounds on the density of semiclassical resonances”, *Duke Math. J.* **137**:3 (2007), 381–459. [MR 2008e:35037](#) [Zbl 05154881](#)
- [Sullivan 1979] D. Sullivan, “The density at infinity of a discrete group of hyperbolic motions”, *Inst. Hautes Études Sci. Publ. Math.* **50** (1979), 171–202. [MR 81b:58031](#) [Zbl 0439.30034](#)
- [Sullivan 1984] D. Sullivan, “Entropy, Hausdorff measures old and new, and limit sets of geometrically finite Kleinian groups”, *Acta Math.* **153**:3-4 (1984), 259–277. [MR 86c:58093](#) [Zbl 0566.58022](#)
- [Takeuchi 1975] K. Takeuchi, “A characterization of arithmetic Fuchsian groups”, *J. Math. Soc. Japan* **27**:4 (1975), 600–612. [MR 53 #2842](#) [Zbl 0311.20030](#)
- [Zworski 1999] M. Zworski, “Dimension of the limit set and the density of resonances for convex co-compact hyperbolic surfaces”, *Invent. Math.* **136**:2 (1999), 353–409. [MR 2002d:58038](#) [Zbl 1016.58014](#)

Received 24 Sep 2009. Accepted 10 Feb 2010.

DMITRY JAKOBSON: jakobson@math.mcgill.ca

Department of Mathematics and Statistics, McGill University, Montreal, QC H3A 2K6, Canada

FRÉDÉRIC NAUD: frederic.naud@univ-avignon.fr

Laboratoire d’Analyse Non-linéaire et Géométrie (EA 2151), Université d’Avignon et des pays de Vaucluse, 84018 Avignon, France

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at pjm.math.berkeley.edu/apde.

Originality. Submission of a manuscript acknowledges that the manuscript is original 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 L^AT_EX but submissions in other varieties of T_EX, 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 BibT_EX 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@mathscipub.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 3 No. 2 2010

Polynomials with no zeros on the bidisk	109
GREG KNESE	
Local wellposedness for the 2+1-dimensional monopole equation	151
MAGDALENA CZUBAK	
Regularity of almost periodic modulo scaling solutions for mass-critical NLS and applications	175
DONG LI and XIAOYI ZHANG	
Estimées des noyaux de Green et de la chaleur sur les espaces symétriques	197
GILLES CARRON	
Lower bounds for resonances of infinite-area Riemann surfaces	207
DMITRY JAKOBSON and FRÉDÉRIC NAUD	