

# ANALYSIS & PDE

Volume 5

No. 2

2012



mathematical sciences publishers



# Analysis & PDE

[msp.berkeley.edu/apde](http://msp.berkeley.edu/apde)

## EDITORS

EDITOR-IN-CHIEF

Maciej Zworski  
University of California  
Berkeley, USA

## BOARD OF EDITORS

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

## PRODUCTION

[contact@msp.org](mailto:contact@msp.org)

Silvio Levy, Scientific Editor

Sheila Newbery, Senior Production Editor

---

See inside back cover or [msp.berkeley.edu/apde](http://msp.berkeley.edu/apde) for submission instructions.

---

The subscription price for 2012 is US \$140/year for the electronic version, and \$240/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 are managed by EditFLOW™ from Mathematical Sciences Publishers.

PUBLISHED BY

 **mathematical sciences publishers**  
<http://msp.org/>

A NON-PROFIT CORPORATION

Typeset in L<sup>A</sup>T<sub>E</sub>X

Copyright ©2012 by Mathematical Sciences Publishers

## THE GEODESIC X-RAY TRANSFORM WITH FOLD CAUSTICS

PLAMEN STEFANOV AND GUNTHER UHLMANN

We give a detailed microlocal study of X-ray transforms over geodesic-like families of curves with conjugate points of fold type. We show that the normal operator is the sum of a pseudodifferential operator and a Fourier integral operator. We compute the principal symbol of both operators and the canonical relation associated to the Fourier integral operator. In two dimensions, for the geodesic transform, we show that there is always a cancellation of singularities to some order, and we give an example where that order is infinite; therefore the normal operator is not microlocally invertible in that case. In the case of three dimensions or higher if the canonical relation is a local canonical graph we show microlocal invertibility of the normal operator. Several examples are also studied.

### 1. Introduction

In this paper we study the X-ray type of transforms over geodesic-like families of curves with caustics (conjugate points). We concentrate on the most common type of caustics — those of fold type. Let  $\gamma_0$  be a fixed geodesic segment on a Riemannian manifold, and let  $f$  be a function whose support does not contain the endpoints of  $\gamma_0$ . The question we are trying to answer is the following: What information about the wave front set  $\text{WF}(f)$  of  $f$  can be obtained from the assumption that (possibly weighted) integrals

$$Xf(\gamma) = \int_{\gamma} f \, ds \tag{1-1}$$

of  $f$  along all geodesics  $\gamma$  close enough to  $\gamma_0$  vanish (or depend smoothly on  $\gamma$ )? We actually study more general geodesic-like curves. Since  $X$  has a Schwartz kernel with singularities of conormal type,  $Xf$  could only provide information for  $\text{WF}(f)$  near the conormal bundle  $\mathcal{N}^*\gamma_0$  of  $\gamma_0$ . If there are no conjugate points along  $\gamma_0$ , then we know that  $\text{WF}(f) \cap \mathcal{N}^*\gamma_0 = \emptyset$ . This has been shown, among the other results, in [Frigyik et al. 2008; Stefanov and Uhlmann 2008] in this context. It also follows from the microlocal approach to Radon transforms initiated by Guillemin [1985] when the Bolker condition (in our case that means no conjugate points) is satisfied. Then the localized normal operator  $N_{\chi} := X^* \chi X$ , where  $\chi$  is a standard cut-off near  $\gamma_0$  is a pseudodifferential operator ( $\Psi$ DO), elliptic at conormal directions to  $\gamma_0$ . If there are conjugate points along  $\gamma_0$ , then  $N_{\chi}$  is no longer a  $\Psi$ DO. One goal of this work is to study the microlocal structure of  $N_{\chi}$  in presence of fold conjugate points, and then use it to see what singularities can be recovered. That would also allow us to tell whether the problem of inverting  $X$  is Fredholm or

---

Stefanov is partly supported by NSF grant DMS-0800428. Uhlmann is partly supported by NSF FRG grant 0554571 and a Walker Family Endowed Professorship.

*MSC2000:* 53C65.

*Keywords:* caustics, conjugate points, geodesic X-ray transform, integral geometry.

not, and would help us to determine the size of the kernel, and to analyze the stability and the possible instability of this problem.

In some applications like geophysics, recovery of singularities is actually the primary goal. The effect of possible conjugate points is treated there as “artifacts” in the reconstruction, creating multiple images of the same object. Our analysis provides in particular a microlocal way to understand those artifacts, and in some cases, to shed light on the possibility of resolving the singularities. We are also motivated by the nonlinear boundary and lens rigidity problems and their applications to seismology, where the X-ray transform appears as a linearization; see e.g., [Michel 1981/82; Croke 1991; Croke et al. 2000; Stefanov and Uhlmann 2005; Stefanov 2008; Stefanov and Uhlmann 2009].

The simplest possible X-ray transform is that over lines in  $\mathbb{R}^n$ :

$$Xf(x, \theta) = \int f(x + t\theta) dt,$$

where  $\theta \in S^{n-1}$ . Parametrization by  $x \in \mathbb{R}^n$  is overdetermined, of course, and we need to think of  $(x, \theta)$  as a way to parametrize a line. It is well known to be injective, on  $L^1_{\text{comp}}(\mathbb{R}^n)$ , for example. It is easy to see, for example by the Fourier slice theorem, that  $Xf$ , known for a fixed  $\theta_0$  and all  $x$ , determines the Fourier transform  $\hat{f}(\xi)$  for  $\xi \perp \theta_0$ . We refer to [Helgason 1980; Natterer 1986] for more details about Euclidean X-ray and Radon transforms. Using relatively simple microlocal techniques, one can show that  $Xf$ , known in a neighborhood of some line  $\ell$ , determines  $\text{WF}(f)$  near  $\mathcal{N}^*\ell$ . A positive smooth weight in the definition of  $X$  would not change that. Those facts are well known and serve as a basis for local tomography methods; see e.g., [Faridani et al. 1992a; 1992b], where the microlocal point of view is implicit.

Geodesic X-ray transforms have a long history, generalizing the Radon type X-ray transform in the Euclidean space; see, e.g., [Helgason 1980]. When the weight is constant and  $(M, g)$  is a simple manifold with boundary, uniqueness and nonsharp stability estimates have been proven in [Muhometov 1981; Muhometov and Romanov 1978; Bernštejn and Gerver 1978], using the energy method. Simple manifolds are compact manifolds diffeomorphic to a ball with convex boundary and no conjugate points. The uniqueness result has been extended to not necessarily convex manifolds under the no-conjugate-points assumption in [Dairbekov 2006]. The authors used microlocal methods to prove a sharp stability estimate in [Stefanov and Uhlmann 2004] for simple manifolds, and uniqueness and stability estimates for more general weighted geodesic-like transforms without conjugate points in [Frigyik et al. 2008]. The X-ray transform over magnetic geodesics with the simplicity assumption was studied in [Dairbekov et al. 2007]. Many of those and other works study integrals of tensors as well, but the results for tensors of order two or higher are less complete. For an overview of the microlocal approach to the geodesic X-ray transform, see [Stefanov 2008].

We considered in [Stefanov and Uhlmann 2008] the X-ray transform of functions and tensors on manifolds with possible conjugate points. Using the overdeterminacy of the problem in dimensions  $n \geq 3$ , we showed that if there exists a family of geodesics without conjugate points with a conormal bundle covering  $T^*M$ , then we still have generic uniqueness and stability. In dimension two, however,

that family has to be the set of all geodesics, and even in higher dimensions, we did not determine the contribution of the conjugate points to  $Xf$ .

We first show in Theorem 2.1 that the normal operator  $N_\chi$  can be represented as a sum of a  $\Psi$ DO and a Fourier integral operator (FIO). The FIO part comes from the conjugate point and represents the “artifact”. An essential part of the proof of Theorem 2.1 is to understand well the geometry of the conjugate locus  $\Sigma$  of pairs  $(p, q) \in M \times M$  conjugate to each other. We show that the Lagrangian of the FIO is  $N^*\Sigma$ . To prove Theorem 2.1, we analyze the singularities of the Schwartz kernel of  $N_\chi$  in Theorem 6.1, which is interesting by itself.

In Section 9, we study whether we can invert  $N_\chi$  microlocally when the curves are geodesics. We find that in some cases we can and in others we cannot. In two dimensions, some cancellation of singularities always occurs, at least to a finite order; see Theorem 9.2. In dimensions three and higher, there are examples (not all geodesic though) where we cannot resolve singularities, and others where we can. We can if the canonical relation of the FIO part is a local graph, but that is not always the case.

In Section 10, we present a few examples, some of them mentioned above. Most of them are based on the transform of integrating a function over circles of a fixed radius in  $\mathbb{R}^2$ . Those circles are actually geodesics of a magnetic system with a Euclidean metric and a constant magnetic field. This example has the advantage that we can compute explicitly the kernel of  $X^*X$ , and we can get an explicit full expansion of the latter as an FIO, etc. In this case, the singularities cancel to infinite order. We can construct more or less explicit singular distributions  $f$  with the property that their singularities are invisible for  $X$  localized near a single circle, that is,  $Xf \in C^\infty$  locally.

## 2. Formulation of the problem

Let  $(M, g)$  be an  $n$ -dimensional Riemannian manifold. Let  $\exp_p(v)$ , where  $(p, v) \in TM$ , be a regular exponential map; see Section 3, where we recall the definition given in [Warner 1965]. The main example is the exponential map of  $g$  or that of another metric on  $M$  or other geodesic-like curves, for example magnetic geodesics; see also [Dairbekov et al. 2007]. Let  $\kappa$  be a smooth function on  $TM \setminus 0$ . We define the weighted X-ray transform  $Xf$  by

$$Xf(p, \theta) = \int \kappa(\exp_p(t\theta), \dot{\exp}_p(t\theta)) f(\exp_p(t\theta)) dt \quad \text{for } (p, \theta) \in SM, \quad (2-1)$$

where we used the notation  $\dot{\exp}(tv) = d\exp(tv)/dt$ . The  $t$  integral above is carried over the maximal interval, including  $t = 0$ , where  $\exp(t\theta)$  is defined. The assumptions that we make below guarantee that this interval remains bounded.

Let  $(p_0, v_0) \in TM$  be such that  $v = v_0$  is a critical point for  $\exp_{p_0}(v)$  (which we call a conjugate vector) of fold type; see the definition below. Let  $q_0 = \exp_{p_0}(v_0)$ . Then our goal is to study  $Xf$  for  $p$  close to  $p_0$  and  $\theta$  close to  $\theta_0 := v_0/|v_0|$  under the assumption that the support of  $f$  is such that  $v_0$  is the only conjugate vector  $v$  at  $p_0$  such that  $\exp_{p_0}(v) \in \text{supp } f$ . Note that  $v_0$  can be written in two different ways as  $t\theta_0$ , where  $|\theta_0| = 1$  and  $\pm t > 0$ , and we choose the first one. The contribution of the second one can be easily derived from our results by replacing  $\theta_0$  by  $-\theta_0$ .

Instead of studying  $X$  directly, we study the operator

$$\begin{aligned}
 Nf(p) &= \int_{S_p M} \kappa^\sharp(p, \theta) Xf(p, \theta) \, d\sigma_p(\theta) \\
 &= \int_{S_p M} \int \kappa^\sharp(p, \theta) \kappa(\exp_p(t\theta), \dot{\exp}_p(t\theta)) f(\exp_p(t\theta)) \, dt \, d\sigma_p(\theta)
 \end{aligned}
 \tag{2-2}$$

for some smooth  $\kappa^\sharp$  localized in a neighborhood of  $(p_0, \theta_0)$ . Here  $d\sigma_p(\theta)$  is the induced Riemannian surface measure on  $S_p(M)$ . When  $\exp$  is the geodesic exponential map, there is a natural way to give a structure of a manifold to all nontrapping geodesics with a natural choice of a measure; see Section 5. The operator  $X$  can be viewed as a map from functions or distributions on  $M$  to functions or distributions on the geodesics manifold. Then one can define the adjoint  $X^*$  with respect to that measure. Then the operator  $X^*X$  is of the form (2-2) with  $\kappa^\sharp = \bar{\kappa}$ ; see (5-1). The condition that  $\text{supp } \kappa^\sharp$  should be contained in a small enough neighborhood of  $(p_0, \theta_0)$  can be easily satisfied by localizing  $p$  near  $p_0$  and choosing  $\text{supp } \kappa$  to be near  $(\gamma_{p_0, \theta_0}, \dot{\gamma}_{p_0, \theta_0})$ . For general regular exponential maps,  $N$  is not necessarily  $X^*X$ .

A direct calculation — see [Stefanov and Uhlmann 2004] and Theorem 5.1 — shows that the Schwartz kernel of  $X^*X$  in the geodesic case (see also [Frigyik et al. 2008] for general families of curves), is singular at the diagonal, as can be expected, and that singularity defines a  $\Psi$ DO of order  $-1$  similarly to the integral geometry problem for geodesics without conjugate points. See Section 5 for more details. Next, singularities away from the diagonal exist at pairs  $(p, q)$  such that  $q = \exp_p(v)$  for some  $v$ , and  $d_v \exp_p$  is not an isomorphism ( $p$  and  $q$  are conjugate points). The main goal of this paper is to study the contribution of those conjugate points to the structure of  $X^*X$  and the consequences of that. We actually study a localized version of this; for a global version on a larger open set, under the assumption that all conjugate points are of fold type, one can use a partition of unity.

Let  $\mathcal{U}$  be a small enough neighborhood of  $(p_0, \theta_0)$  in  $SM$ . Let  $U$  be a small neighborhood of  $p_0$  such that  $U \subset \pi(\mathcal{U})$ , where  $\pi$  is the natural projection on the base. Fix  $\kappa^\sharp \in C_0^\infty(\mathcal{U})$ . Let  $Nf$  be as in (2-2), related to  $\kappa^\sharp$ , where  $\kappa$  is a smooth weight. We will apply  $X$  to functions  $f$  supported in an open set  $V \ni p_0$  satisfying the conjugacy assumption of the theorem below; see Figure 1. Our goal is to study the contribution of a single fold type of singularity. Let  $\Sigma \subset M \times M$  be the conjugate locus in a neighborhood of  $(p_0, q_0)$ ; see Section 3. Finally, let  $\gamma_0 = \gamma_{p_0, \theta_0}(t)$  for  $t \in I$  be the geodesic through  $(p_0, \theta_0)$  defined in the interval  $I \ni 0$ , with endpoints outside  $V$ .

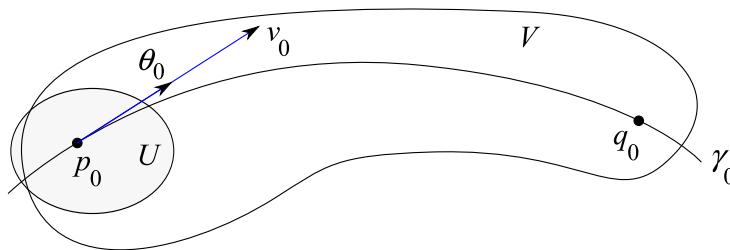


Figure 1

The first main result of this paper is the following.

**Theorem 2.1.** *Let  $v_0 = |v_0|\theta_0$  be a fold conjugate vector at  $p_0$ , and let  $N$  be as in (2-2). Let  $v_0$  be the only singularity of  $\exp_{p_0}(v)$  on the ray  $\{\exp_p(t\theta_0), t \in I\} \cap V$ . Then if  $\mathcal{U}$  (and therefore,  $U$ ) is small enough, the operator  $N : C_0^\infty(V) \rightarrow C_0^\infty(U)$  admits the decomposition*

$$N = A + F, \tag{2-3}$$

where  $A$  is a  $\Psi DO$  of order  $-1$  with principal symbol

$$\sigma_p(A)(x, \xi) = 2\pi \int_{S_x M} \delta(\xi(\theta)) (\kappa^\sharp \kappa)(x, \theta) d\sigma_x(\theta) \tag{2-4}$$

and  $F$  is an FIO of order  $-n/2$  associated to the Lagrangian  $\mathcal{N}^*\Sigma$ . In particular, the canonical relation  $\mathcal{C}$  of  $F$  in local coordinates is given by

$$\mathcal{C} = \{(p, \xi, q, \eta), (p, q) \in \Sigma, \xi = -\eta_i \partial \exp_p^i(v) / \partial p, \eta \in \text{Coker } d_v \exp_p(v), \det d_v \exp_p(v) = 0\}. \tag{2-5}$$

If  $\exp$  is the exponential map of  $g$ , then  $\mathcal{C}$  can also be characterized as  $\mathcal{N}^*\Sigma'$ , where  $\mathcal{N}^*\Sigma$  is as in (4-17) and the prime means that we replace  $\eta$  by  $-\eta$ .

It is easy to check that  $\mathcal{C}$  above is invariantly defined.

In Section 9 we show in dimension 3 or higher that the operator  $N$  is microlocally invertible if  $\mathcal{C}$  is a local canonical graph. In two dimensions, we show in the geodesic case that there is always a loss of some derivatives at least when the curves are geodesics. We study in detail the case of the circular Radon transform in two dimensions in Section 10, and show that then  $N$  is not microlocally invertible.

### 3. Regular exponential maps and their generic singularities

**3a. Regular exponential maps.** Let  $M$  be a fixed  $n$ -dimensional manifold. We will recall the definition of Warner [1965] of a regular exponential map at  $p \in M$ . We think of it as a generalization of the exponential map on a Riemannian manifold, by requiring only those properties that are really necessary for what follows. For that reason, we use the notation  $\exp_p(v)$ . In addition to the requirements of Warner, we will require  $\exp_p(v)$  to be smooth in  $p$ . Let  $N_p(v) \subset T_v T_p M$  denote the kernel of  $d \exp_p$ . Unless specifically indicated,  $d$  is the differential with respect to  $v$ . The radial tangent space at  $v$  will be denoted by  $r_v$ . It can be identified with  $\{sv, s \in \mathbb{R}\}$ , where  $v$  is considered as an element of  $T_v T_p M$ .

**Definition 3.1.** A map  $\exp_p(v)$  that maps  $v \ni T_p M$  into  $M$  for each  $p \in M$  is called a *regular exponential map* if the following hold.

- (R1)  $\exp$  is smooth in both variables, except possibly at  $v = 0$ . Next,  $d \exp_p(tv) / dt \neq 0$  when  $v \neq 0$ .
- (R2) The Hessian  $d^2 \exp_p(v)$  isomorphically maps  $r_v \times N_p(v)$  onto  $T_{\exp_p(v)} M / d \exp_p(T_v T_p M)$  for any  $v \neq 0$  in  $T_p M$  for which  $\exp_p(v)$  is defined.
- (R3) For each  $v \in T_p M \setminus 0$ , there is a convex neighborhood  $U$  of  $v$  such that the number of singularities of  $\exp_p$ , counted with multiplicities, on the ray  $tv$  for  $t \in \mathbb{R}$  in  $U$ , for each such ray that intersects  $U$ , is constant and equal to the order of  $v$  as a singularity of  $\exp_p$ .

An example is the exponential map on a Riemannian (or more generally on a Finsler manifold); see [Warner 1965]. Then (R1) is clearly true. Next, (R2) follows from the following well-known property. Fix  $p$  and a geodesic through it. Consider all Jacobi fields vanishing at  $p$ . Then at any  $q$  on that geodesic, the values of those Jacobi fields that do not vanish at  $q$  and the covariant derivatives of those that vanish at  $q$  span  $T_q M$ . Also, those two spaces are orthogonal. Finally, (R3) represents the well-known continuity property of the conjugate points, counted with their multiplicities that follows from the Morse index theorem; see, e.g., [Jost 1998, Theorem 4.3.2].

We will need also an assumption about the behavior of the exponential map at  $v = 0$ .

(R4)  $\exp_p(tv)$  is smooth in  $p, t, v$  for all  $p \in M$ ,  $|t| \ll 1$ , and  $v \neq 0$ . Moreover,

$$\exp_p(0) = p, \quad \text{and} \quad \frac{d}{dt} \exp_p(tv) = v \quad \text{for } t = 0.$$

Given a regular exponential map, we define the “geodesic”  $\gamma_{p,v}(t)$ , with  $v \neq 0$ , by  $\gamma_{p,v}(t) = \exp_p(tv)$ . We will often use the notation

$$q = \exp_p(v) = \gamma_{p,v}(1), \quad w = -\dot{\exp}_p(v) := -\dot{\gamma}_{p,v}(1), \quad \theta = v/|v|. \quad (3-1)$$

Note that the “geodesic flow” does not necessarily obey the group property. We will assume that

(R5) For  $q$  and  $w$  as in (3-1), we have  $\exp_q(w) = p$  and  $\dot{\exp}_q(w) = -v$ .

This implies that in particular,  $(p, v) \mapsto (q, w)$  is a diffeomorphism. If  $\exp$  is the exponential map of a Riemannian metric, then (R5) is automatically true and that map is actually a symplectomorphism (on  $T^*M$ ).

**Remark 3.1.** In case of magnetic geodesics, or more general Hamiltonian flows, (R5) is equivalent to time reversibility of the “geodesics”. This is not true in general. On the other hand, one could define the reverse exponential map  $\exp_q^-(w) = \gamma_{q,-w}(-1)$  in that case (see e.g., [Dairbekov et al. 2007]) near  $(q_0, w_0)$ , and replace  $\exp$  by  $\exp^-$  in that neighborhood. Then (R5) would hold. In other words, (R5) really says that  $(p, v) \mapsto (q, w)$  is assumed to be a local diffeomorphism with an inverse satisfying (R1)–(R4).

**3b. Generic properties of the conjugate locus.** We recall here the main result by Warner [1965] about the regular points of the conjugate locus of a fixed point  $p$ . The *tangent conjugate locus*  $S(p)$  of  $p$  is the set of all vectors  $v \in T_p M$  such that  $d\exp_p(v)$  (the differential of  $\exp_p(v)$  with respect to  $v$ ) is not an isomorphism. We call such vectors conjugate vectors at  $p$  (called conjugate points in [Warner 1965]). The kernel of  $d\exp_p(v)$  is denoted by  $N_p(v)$ . It is a part of  $T_v T_p M$ , which we identify with  $T_p M$ . In the Riemannian case,  $N_p(v)$  is orthogonal to  $v$  by the Gauss lemma. In the general case, it is always transversal to  $v$  by (R1). The images of the conjugate vectors under the exponential map  $\exp_p$  will be called the *conjugate points* of  $p$ . The image of  $S(p)$  under the exponential map  $\exp_p$  will be denoted by  $\Sigma(p)$  and called the *conjugate locus* of  $p$ . Note that  $S(p) \subset T_p M$ , while  $\Sigma(p) \subset M$ . We always work with  $p$  near a fixed  $p_0$  and with  $v$  near a fixed  $v_0$ . Set  $q_0 = \exp_{p_0}(v_0)$ . Then we are interested in  $S(p)$  restricted to a small neighborhood of  $v_0$ , and in  $\Sigma(p)$  near  $q_0$ . Note that  $\Sigma(p)$  may not contain all points



near  $q_0$  conjugate to  $p$  along some “geodesic”; and may not contain even all of those along  $\exp_{p_0}(tv_0)$  if the latter self-intersects — it contains only those that are of the form  $\exp_p(v)$  with  $v$  close enough to  $v_0$ .

Normally,  $d\exp_p(v)$  stands for the differential of  $\exp_p(v)$  with respect to  $v$ . When we need to take the differential with respect to  $p$ , we will use the notation  $d_p$  for it. We write  $d_v$  for the differential with respect to  $v$  when we want to distinguish between the two.

We denote by  $\Sigma$  the set of all conjugate pairs  $(p, q)$  localized as above. In other words,  $\Sigma = \{(p, q) : q \in \Sigma(p)\}$ , where  $p$  runs over a small neighborhood of  $p_0$ . Also, we denote by  $S$  the set  $(p, v)$ , where  $v \in S(p)$ .

A *regular conjugate vector*  $v$  is defined by the requirement that there exists a neighborhood of  $v$  such that any radial ray of  $T_pM$  contains at most one conjugate point there. The regular conjugate locus then is an everywhere-dense open subset of the conjugate locus that has a natural structure of an  $(n - 1)$ -dimensional manifold. The order of a conjugate vector as a singularity of  $\exp_p$  (the dimension of the kernel of the differential) is called an order of the conjugate vector.

In [1965, Theorem 3.3], Warner characterized the conjugate vectors at a fixed  $p_0$  of order at least 2, and some of those of order 1, as described below. Note that in  $B_1$ , one needs to postulate that  $N_{p_0}(v)$  remains tangent to  $S(p_0)$  at points  $v$  close to  $v_0$  since the latter is not guaranteed by just assuming that it holds at  $v_0$  only.

**(F) Fold conjugate vectors:** Let  $v_0$  be a regular conjugate vector at  $p_0$ , and let  $N_{p_0}(v_0)$  be one-dimensional and transversal to  $S(p_0)$ . Such singularities are known as fold singularities. Then one can find local coordinates  $\xi$  near  $v_0$  and  $y$  near  $q_0$  such that in those coordinates,  $\exp_{p_0}$  is given by

$$y' = \xi' \quad y^n = (\xi^n)^2. \tag{3-2}$$

Then

$$S(p_0) = \{\xi^n = 0\}, \quad N_{p_0}(v_0) = \text{span}\{\partial/\partial\xi^n\}, \quad \Sigma(p_0) = \{y^n = 0\}. \tag{3-3}$$

Since the fold condition is stable under small  $C^\infty$  perturbations, as follows directly from the definition, those properties are preserved under a small perturbation of  $p_0$ .

**(B<sub>1</sub>) Blowdown of order 1:** Let  $v_0$  be a regular conjugate vector at  $p_0$  and let  $N_{p_0}(v)$  be one-dimensional. Assume also that  $N_{p_0}(v)$  is tangent to  $S(p_0)$  for all regular conjugate  $v$  near  $v_0$ . We call such singularities blowdown of order 1. Then locally,  $\exp_{p_0}$  is represented in suitable coordinates by

$$y' = \xi', \quad y^n = \xi^1 \xi^n. \tag{3-4}$$

Then

$$S(p_0) = \{\xi^1 = 0\}, \quad N_{p_0}(v_0) = \text{span}\{\partial/\partial\xi^n\}, \quad \Sigma(p_0) = \{y^1 = y^n = 0\}. \tag{3-5}$$

Even though we postulated that the tangency condition is stable under perturbations of  $v_0$ , it is not stable under a small perturbation of  $p_0$ , and the type of the singularity may change then. In some symmetric cases, one can check directly that the type is locally preserved.

**(B<sub>k</sub>) Blowdown of higher order:** Those are regular conjugate vectors in the case where  $N_{p_0}(v_0)$  is  $k$ -dimensional, with  $2 \leq k \leq n - 1$ . Then in some coordinates,  $\exp_{p_0}$  is represented as

$$y^i = \begin{cases} \xi^i & \text{if } i = 1, \dots, n - k, \\ \xi^1 \xi^i & \text{if } i = n - k + 1, \dots, n. \end{cases} \tag{3-6}$$

Then

$$\begin{aligned} S(p_0) &= \{\xi^1 = 0\}, & N_{p_0}(v_0) &= \text{span}\{\partial/\partial \xi^{n-k+1}, \dots, \partial/\partial \xi^n\}, \\ \Sigma(p_0) &= \{y^1 = y^{n-k+1} = \dots = y^n = 0\}. \end{aligned} \tag{3-7}$$

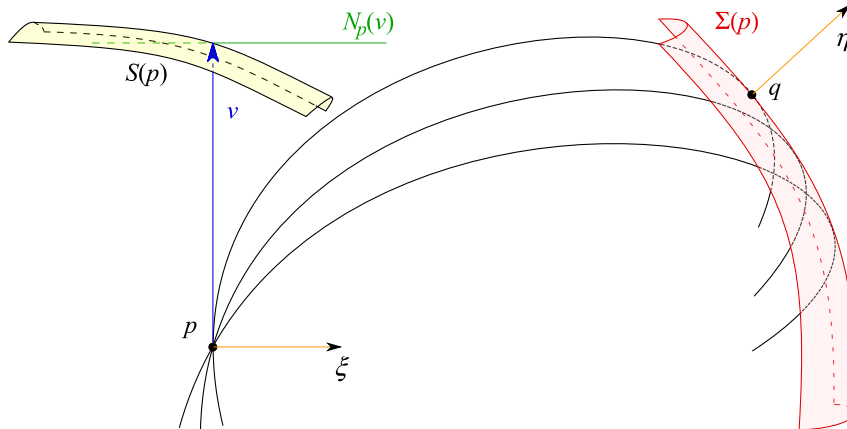
In particular,  $N_{p_0}(v_0)$  must be tangent to  $S(p_0)$ ; see also [Warner 1965, Theorem 3.2]. This singularity is also unstable under perturbations of  $p_0$ . A typical example is the antipodal points on  $S^n$  for  $n \geq 3$ ; then  $k = n - 1$ .

The purpose of this paper is to study the effect of fold conjugate points of  $X$ .

#### 4. Geometry of the fold conjugate locus

In this section, we study the geometry of the tangent conjugate loci  $S(p)$  and  $S$ , and the conjugate loci  $\Sigma(p)$  and  $\Sigma$ , respectively. Recall that we work locally, and everywhere below, even if not stated explicitly,  $(p, v)$  belongs to a small enough neighborhood of  $(p_0, v_0)$ , and  $(q, w)$  is near  $(q_0, w_0)$ . We assume throughout the section that  $v_0$  is conjugate vector at  $p_0$  of fold type. We also fix a nonzero covector  $\eta_0$  at  $q_0$  as in (2-5), and let  $\xi_0$  be the corresponding  $\xi$  as in (2-5). We will see later that  $\xi_0 \neq 0$ . We refer to Figure 2, where  $w$  is not shown, and the zero subscripts are omitted.

- Lemma 4.1.** (a) *Let  $v \in S(p)$  be a fold conjugate vector. Then,  $\Sigma(p)$  near  $q = \exp_p(v)$  is a smooth surface of codimension one, tangent to  $w := -\dot{\gamma}_{p,v}(1)$ .*  
 (b) *The locus  $S$  is a smooth  $(2n - 1)$ -dimensional surface in  $TM$  that can be considered as the bundle  $\{S(p) : p \in M\}$  with fibers  $S(p)$ .*



**Figure 2.** A typical fold conjugate locus.

*Proof.* Consider (a) first. The representation (3-2) implies that locally  $\Sigma(p) = \exp_p(S(p))$  is a smooth surface of codimension one (given by  $y^n = 0$ ). Next, for  $v \in S(p)$ , the differential  $d \exp_p$  sends any vector to a vector tangent to  $S(p)$ , which follows from (3-2) again. In particular, this is true for the radial vector  $v$  (considered as a vector in  $T_v T_p M$ ). This proves that  $w$  is tangent to  $\Sigma(p)$ .

The statement (b) follows from the fact that  $S$  is defined by  $\det d \exp_p(v) = 0$ , and that  $\det d \exp_p(v)$  has a nonvanishing differential with respect to  $v$ .  $\square$

**Remark 4.1.** It is easy to show that in (a),  $\gamma_{p,v}$  is tangent to  $\Sigma(p)$  of order 1 only.

We define ‘‘Jacobi fields’’ along  $\gamma_{p,v}$  vanishing at  $p$  as follows. For any  $\alpha \in T_v T_p M$ , set

$$J(t) = d(\exp_p(tv))(\alpha) = \alpha^k \frac{\partial}{\partial v^k} \exp_p(tv).$$

Then  $J(0) = 0$  and  $\dot{J}(0) = \alpha$ , where  $\dot{J}(T) = dJ(t)/dt$ . If  $J(1) = 0$ , then a direct computation shows that

$$\dot{J}(1) = d^2 \exp_p(v)(\alpha \times v). \tag{4-1}$$

When  $\exp$  is the exponential map of a Riemannian metric, it is natural to work with the covariant derivative  $D_t J(t) =: J'(t)$  instead of  $\dot{J}(t)$ . While they are different in general, they coincide at points where  $J(t) = 0$ .

The next lemma shows that the fold/blowdown conditions are symmetric with respect to  $p$  and  $q$ .

**Lemma 4.2.** *The vector  $v_0$  is a conjugate vector at  $p_0$  of fold type if and only if  $w_0$  is a conjugate vector at  $q_0$  of fold type.*

*Proof.* Set  $w_0 = -\dot{\gamma}_{p_0, v_0}(1)$  as in (3-1). Then  $p_0 = \exp_{q_0}(w_0)$ . Assume now that  $\alpha \in N_{p_0}(v_0)$ . In some local coordinates, differentiate  $p = \exp_q(w)$  with respect to  $v$  in the direction of  $\alpha$ ; here  $q$  and  $w$  are viewed as functions of  $p$  and  $v$ . Then, using the Jacobi field notation introduced above in (4-1), we get

$$0 = d \exp_{q_0}(w_0) \left( \alpha^k \frac{\partial w}{\partial v^k}(p_0, v_0) \right) = d \exp_{q_0}(w_0) \dot{J}(1)$$

because

$$\alpha^k \frac{\partial w}{\partial v^k}(p_0, v_0) = \alpha^k \frac{\partial}{\partial v^k} \frac{d}{dt} \Big|_{t=1} \exp_p(tv)(p_0, v_0) = \dot{J}(1).$$

By (R2),  $\dot{J}(1) \neq 0$ , so in particular, this shows that  $w_0$  is conjugate at  $q_0$ , and  $\dot{J}(1) \in N_{q_0}(w_0)$ . Moreover, by (R2), the linear map

$$N_p(v) \ni \alpha = \dot{J}(0) \mapsto \dot{J}(1) := \beta \in N_q(w), \quad \text{with } J(0) = J(1) = 0, \tag{4-2}$$

defines an isomorphism between  $N_p(v)$  and  $N_q(w)$ . Then (4-2) shows that  $w_0$  is conjugate at  $q_0$  of multiplicity one. By (R3), applied to  $w_0$ , it is also regular.

We will prove now that  $w_0$  is of fold type. Since it is regular and of multiplicity one,  $S(q_0)$  near  $w_0$  is a smooth  $(n - 1)$ -dimensional surface either of type  $F$ , as in (3-3) or of type  $B_1$ , as in (3-5). Assume the latter case first; then  $\Sigma(q_0)$  is of codimension two, as follows from (3-5). In particular, using the normal form (3-4), we see that in this case, one can find a nontrivial one-parameter family of vectors

$w(s)$  such that  $w(0) = w_0$  and  $\exp_{q_0}(w(s)) = p_0$ . Then the corresponding tangent vectors at  $p_0$  would form a nontrivial one-parameter family of vectors  $v(s)$  such that  $\exp_{p_0}(v(s)) = q_0$ . That cannot happen, if  $v_0$  is of type  $F$  (see (3-2)), since the equation  $\exp_{p_0}(v) = q_0$  has (near  $v_0$ ) at most two solutions.  $\square$

For  $(p, v) \in S$ , let  $\alpha = \alpha(p, v) \in N_p(v)$  be a unit vector. To fix the direction, assume that the derivative of  $\det d\exp_p(v)$  in the direction of  $\alpha$  is positive for  $v$  a conjugate vector. Here we identify  $T_v T_p M$  and  $T_p M$ . In the fold case,  $N_p(v)$  is clearly a smooth vector bundle on  $TM$  near  $(p_0, v_0)$ , and  $\alpha$  is a smooth vector field.

**Lemma 4.3.** *For any fixed  $p$  near  $p_0$ , the map*

$$S(p) \ni v \mapsto \alpha(p, v) \in N_p(v) \quad (4-3)$$

*is a local diffeomorphism, smoothly depending on  $p$  if and only if*

$$d^2 \exp_{p_0}(v_0)(N_{p_0}(v_0) \setminus 0 \times \cdot) \Big|_{T_{v_0} S(p_0)} \text{ is of full rank.} \quad (4-4)$$

*Proof.* In local coordinates, we want to find a condition such that the equation

$$\alpha^i \partial_{v^i} \exp_p(v) = 0$$

can be solved for  $v$  so that  $v = v_0$  for  $(p, \alpha) = (p_0, \alpha_0)$ , where  $\alpha_0 = \alpha(p_0, v_0)$ . Then  $v$  would automatically be in  $S(p)$ . By the implicit function theorem, this is equivalent to

$$\det(\partial_v \alpha_0^i \partial_{v^i} \exp_{p_0}(v)) \neq 0 \quad \text{at } v = v_0.$$

Choose a coordinate system near  $v_0$  such that  $\partial/\partial v^n$  spans  $N_{p_0}(v_0)$ , and  $\{\partial/\partial v^1, \dots, \partial/\partial v^{n-1}\}$  span  $T_{v_0} S(p_0)$ . Let  $F(v) = \exp_{p_0}(v)$  and denote by  $F_i$  and  $F_{ij}$  the corresponding partial derivatives. Greek indices below run from 1 to  $n-1$ . We have

$$\partial_n F(v_0) = 0 \quad \text{because } \partial/\partial v^n \in N_{p_0}(v_0), \quad (4-5)$$

$$\partial_\alpha \det(\partial F)(v_0) = 0 \quad \text{because } \partial/\partial v^\alpha \text{ is tangent to } S(p_0) \text{ at } v_0, \quad (4-6)$$

$$\partial_n \det(\partial F)(v_0) \neq 0 \quad \text{by the fold condition,} \quad (4-7)$$

$$c^\alpha \partial_\alpha F(v_0) \neq 0 \quad \text{for all } c \neq 0 \quad \text{because } c^\alpha \partial/\partial v^\alpha \notin N_{p_0}(v_0). \quad (4-8)$$

We want to prove that  $\det(\partial_n \partial F)(v_0) \neq 0$  if and only if (4-4) holds. That determinant equals

$$\det(F_{1n}, F_{2n}, \dots, F_{nn})(v_0). \quad (4-9)$$

Perform the differentiation in (4-6). By (4-5) and (4-8),

$$\det(F_1, \dots, F_{n-1}, F_{n\alpha})(v_0) = 0 \quad \text{for all } \alpha \quad \text{implies} \quad F_{n\alpha}(v_0) \in \text{span}(F_1(v_0), \dots, F_{n-1}(v_0)).$$

Similarly, (4-7) shows that

$$\det(F_1, \dots, F_{n-1}, F_{nn})(v_0) \neq 0 \quad \text{implies} \quad 0 \neq F_{nn}(v_0) \notin \text{span}(F_1(v_0), \dots, F_{n-1}(v_0)). \quad (4-10)$$



Those two relations show that (4-9) vanishes if and only if  $(F_{n1}(v_0), \dots, F_{n,n-1}(v_0))$  form a linearly dependent system that is equivalent to (4-4).  $\square$

We study the structure of the conjugate loci  $\Sigma(p)$ ,  $\Sigma(q)$  and  $\Sigma$  next. Recall again that we work locally near  $p_0$ ,  $v_0$  and  $q_0$ .

**Theorem 4.1.** *Let  $v_0$  be a fold conjugate vector at  $p_0$ .*

- (a) *For any  $p$  near  $p_0$ ,  $\Sigma(p)$  is a smooth hypersurface of dimension  $n - 1$  smoothly depending on  $p$ . Moreover for any  $q = \exp_p(v) \in \Sigma(p)$ ,  $T_q M$  is a direct sum of the linearly independent spaces*

$$T_q M = T_q \Sigma(p) \oplus N_q(w), \quad (4-11)$$

and

$$T_q \Sigma(p) = \text{Im } d \exp_p(v) \quad \text{and} \quad N_q^* \Sigma(p) = \text{Coker } d_v \exp_p(v).$$

*Next, those statements remain true with  $p$  and  $q$  exchanged.*

- (b)  *$\Sigma$  is a smooth  $(2n - 1)$ -dimensional hypersurface in  $M \times M$  near  $(p_0, q_0)$  that is also a fiber bundle  $\Sigma = \{\Sigma(p) : p \in M\}$  with fibers  $\Sigma(p)$  (and also  $\Sigma = \{\Sigma(q) : q \in M\}$ ). Moreover, the conormal bundle  $\mathcal{N}^* \Sigma$  is given by*

$$\mathcal{N}^* \Sigma = \{(p, q, \xi, \eta) : (p, q) \in \Sigma, \xi = \eta_i \partial \exp_p^i(v) / \partial p, \eta \in \text{Coker } d_v \exp_p(v), \text{ where } v = \exp_p^{-1}(q) \text{ with } \exp_p \text{ restricted to } S(p)\}. \quad (4-12)$$

*Proof.* We start with (a). By the normal form (3-2), also clear from the fold condition, the image of  $S(p)$  under  $d \exp_p(v)$  coincides with  $T_q \Sigma(p)$ . In particular,  $d \exp_p(v)$ , restricted to  $S(p)$  is a diffeomorphism to its image. Relation (4-11) follows from (4-2) and (R2).

Consider (b). We have  $(p, q) \in \Sigma$  if and only if there exists  $v$  (near  $v_0$ ) such that

$$q = \exp_p(v) \quad \text{and} \quad \det d_v \exp_p(v) = 0. \quad (4-13)$$

In some local coordinates, we view this as  $n + 1$  equations for the  $3n$ -dimensional variable  $(p, q, v)$  near  $(p_0, q_0, v_0)$ . We show first that the solution, which we denote by  $L$ , is a  $(2n - 1)$ -dimensional submanifold. To this end, we need to show that the following differential has rank  $n + 1$  at  $(p_0, q_0, v_0)$ :

$$\begin{pmatrix} d_p \exp_p(v) & -\text{Id} & d_v \exp_p(v) \\ d_p \det d_v \exp_p(v) & 0 & d_v \det d_v \exp_p(v) \end{pmatrix}. \quad (4-14)$$

The elements of the first row are  $n \times n$  matrices, while the second row consists of three  $n$ -vectors. That the rank of the differential above is full follows from the fact that  $d_v \det d_v \exp_p(v) \neq 0$  at  $(p_0, v_0)$ , guaranteed by the fold condition.

Set  $\pi(p, q, v) = (p, q)$ . We show next that  $\pi(L)$  is a  $(2n - 1)$ -dimensional submanifold too. To this end, we need to show that  $d\pi$  is injective on  $TL$ . The tangent space to  $L$  is given by the orthogonal complement to the rows of (4-14). Denote any vector in  $TL$  by  $\rho = (\rho_p, \rho_q, \rho_v)$ . Then  $d\pi(\rho) = (\rho_p, \rho_q)$ .

Our goal is therefore to show that  $\rho_p = \rho_q = 0$  implies  $\rho_v = 0$ . Then  $(0, 0, \rho_v)$  is orthogonal to the rows of (4-14), and therefore

$$\rho_v^i \partial_{v^i} \exp_p^k(v) = 0 \quad \text{for } k = 1, \dots, n, \quad \text{and} \quad \rho_v^i \partial_{v^i} \det d_v \exp_p(v) = 0.$$

The latter identity shows that  $\rho_v \in N_p(v)$ , while the first one shows that  $\rho_v \in \text{Ker } d_v \exp_p(v)$ . By the fold condition,  $\rho_v = 0$ .

This analysis also shows that the covectors  $\nu$  orthogonal to  $\Sigma$  are of the form  $\nu = (v_p, v_q)$  with the property that  $(v_p, v_q, 0)$  is conormal to  $L$ . Since the conormals to  $L$  are spanned by the rows of (4-14), to get the third component to vanish, we have to take a linear combination with coefficients  $a_i$  for  $i = 1, \dots, n$  and  $b$  such that

$$a_i \frac{\partial q^i}{\partial v^j} + b \frac{\partial \det d_v \exp_p(v)}{\partial v^j} = 0 \quad \text{for all } j, \quad (4-15)$$

where  $q = \exp_p(v)$ . Let  $0 \neq \alpha \in N_p(v)$ . Multiply by  $\alpha^j$  and sum over  $j$  above to get that the  $v$ -derivative of  $b \det d_v \exp_p(v)$  in the direction of  $N_p(v)$  vanishes. According to the fold assumption, this is only possible if  $b = 0$ . Then we get that  $a \in \text{Coker } d_v \exp_p(v)$ . Therefore the covectors normal to  $\Sigma$  are of the form

$$\nu = \left( \left\{ a_i \frac{\partial q^i}{\partial p^j} \right\}, -a \right) \quad \text{for } a \in \text{Coker } d_v \exp_p(v), \quad (4-16)$$

which proves (4-12). □

**Theorem 4.2.** *Let  $v_0$  be a fold conjugate vector at  $p_0$ . Let  $\exp_p$  be the exponential map of a Riemannian metric.*

(a) *The sum in (4-11) is an orthogonal one, that is,*

$$N_q \Sigma(p) = N_q(w).$$

(b) *Next, (4-17) also admits the representation*

$$\mathcal{N}\Sigma = \left\{ (p, q, \alpha, \beta); (p, q) \in \Sigma, \alpha = J'(0), \beta = -J'(1), \text{ where } J \text{ is any Jacobi field} \right. \\ \left. \text{along the locally unique geodesic connecting } p \text{ and } q \text{ with } J(0) = J(1) = 0 \right\}. \quad (4-17)$$

(c)  *$\mathcal{N}\Sigma$  is a graph of a smooth map  $(p, \alpha) \mapsto (q, \beta)$  if and only if condition (4-4) is fulfilled. Then that map is a local diffeomorphism.*

**Remark 4.2.** Note that for  $(p, q) \in \Sigma$ , the geodesic connecting  $p$  and  $q$  is unique, as follows from the normal form (3-2), only among the geodesics with  $\dot{\gamma}(0)$  close to  $v_0$ . Also,  $J$  is determined uniquely up to a multiplicative constant. Next, once we prove that  $\Sigma$  is smooth, then  $\alpha \in N_p(v)$  and  $\beta \in N_q(w)$  by (a) (see also (3-2)), but (4-17) gives something more than that — it restricts  $(\alpha, \beta)$  to an one-dimensional space.

**Remark 4.3.** It is natural to ask whether  $|J'(0)| = |J'(1)|$ . One can show that generically this is not so.

*Proof.* By [Lang 1995, Lemma IX.3.5], the conjugate of  $d \exp_p(v)$  with respect to the metric form is given by

$$(d \exp_p(v))^* = d \exp_q(w), \quad (4-18)$$

where we use the notation of (3-1). The normal to  $\Sigma(p)$  at  $q$  is in the orthogonal complement to the image of  $d \exp_p(v)$ , which by (4-18) is  $\text{Ker } d \exp_q(w) = N_q(w)$ . This proves (a).

Then we get by (4-18) and (4-15) (where  $b = 0$ ) that  $a \in N_q(w)$ , where we identify the covector  $a$  with a vector by the metric.

We will use now [Lang 1995, Lemma IX.3.4]: For any two Jacobi fields  $J_1$  and  $J_2$  along a fixed geodesic, the Wronskian  $\langle J'_1, J_2 \rangle - \langle J_1, J'_2 \rangle$  is constant. Along the geodesic connecting  $p$  and  $q$ , in fixed coordinates near  $p$ , let  $\tilde{J}$  be determined by  $\tilde{J}(0) = e_j$  and  $\tilde{J}'(0) = 0$ . Here  $e_j$  has components  $\delta_j^i$ . If  $p$  and  $q$  are conjugate to each other, then  $\tilde{J}(1)$  is equal to the variation  $\partial q / \partial p^j$ , and this is independent of the choice of local coordinates as long as  $e_j$  is considered as a fixed vector at  $p$ . Define another Jacobi field by  $J(1) = 0$  and  $J'(1) = a$ , where  $a$  is as in (4-16) but considered as a vector. Denote the field in the brackets in (4-16) by  $X_j$ . Then

$$\begin{aligned} X_j &= \langle a, \tilde{J}(1) \rangle = \langle J'(1), \tilde{J}(1) \rangle \\ &= \langle J'(1), \tilde{J}(1) \rangle - \langle J(1), \tilde{J}'(1) \rangle \\ &= \langle J'(0), \tilde{J}(0) \rangle - \langle J(0), \tilde{J}'(0) \rangle = J'_j(0). \end{aligned}$$

This proves (4-17).

The proof of (c) follows directly from Lemma 4.3.  $\square$

## 5. The Schwartz kernel of $N$ near the diagonal and mapping properties of $X$ and $N$

**5a. The geodesic case.** Let  $\exp$  be the exponential map of the metric  $g$ . Then  $X$  is the weighted geodesic ray transform. One way to parametrize the geodesics is the following. Let  $H$  be any orientable hypersurface with the property that it intersects transversally, at one point only, any geodesic in  $M$  issued from a point in  $\mathcal{U}$ . For our local analysis,  $H$  can be an arbitrarily small surface intersecting transversally  $\gamma_{p_0, v_0}$ , so let us fix that choice. Let  $d\text{Vol}_H$  be the induced measure in  $H$ , and let  $\nu$  be a smooth unit normal vector field on  $H$  consistent with the orientation of  $H$ . Let  $\mathcal{H}$  consist of all  $(p, \theta) \in SM$  with the property that  $p \in H$  and  $\theta$  is not tangent to  $H$ , and positively oriented, that is,  $\langle \nu, \theta \rangle > 0$ . Introduce the measure  $d\mu = \langle n, \theta \rangle d\text{Vol}_H(p) d\sigma_p(\theta)$  on  $\mathcal{H}$ . Then one can parametrize all geodesics intersecting  $H$  transversally by their intersection  $p$  with  $H$  and the corresponding direction, that is, by elements in  $\mathcal{H}$ . An important property of  $d\mu$  is that it introduces by Liouville's theorem a measure on that geodesics set that is invariant under a different choice of  $H$ ; see e.g., [Stefanov and Uhlmann 2004].

The weighted geodesic transform  $X$  can be defined as in (2-1) for  $(p, \theta) \in \mathcal{H}$  instead of  $(p, \theta) \in \mathcal{U}$  because transporting  $(p, v)$  along the geodesic flow does not change the integral. Since we assumed originally that  $\kappa$  is localized near a small enough neighborhood of  $\gamma_{p_0, v_0}$ , we get that  $\kappa$  is supported in a

small neighborhood of  $(p_0, \theta_0)$  in  $\mathcal{H}$ . We view  $X$  as the map

$$X : L^2(M) \rightarrow L^2(\mathcal{H}, d\mu)$$

restricted to a neighborhood of  $(p_0, \theta_0)$ . This map is bounded [Sharafutdinov 1994], and this also follows from our analysis of  $N$ . By the proof of [Stefanov and Uhlmann 2004, Proposition 1],  $X^*X$  is given by

$$X^*Xf(p) = \frac{1}{\sqrt{\det g(p)}} \int_{S_p M} \int \bar{\kappa}(p, \theta) \kappa(\exp_p(t\theta), \dot{\exp}_p(t\theta)) f(\exp_p(t\theta)) dt d\sigma_p(\theta). \quad (5-1)$$

We therefore proved the following.

**Proposition 5.1.** *Let  $\exp$  be the geodesic exponential map. Let  $X$  be the weighted geodesic ray transform (2-1), and let  $N$  be as in (2-2), depending on  $\kappa^\sharp$ . Then*

$$X^*X = N \quad \text{with } \kappa^\sharp = \bar{\kappa}.$$

Split the  $t$  integral in (5-1) into the regions  $t > 0$  and  $t < 0$ , and make a change of variables  $(t, \theta) \mapsto (-t, -\theta)$  in the second one to get

$$X^*Xf(p) = \frac{1}{\sqrt{\det g(p)}} \int_{T_p M} W(p, v) f(\exp_p(v)) d\text{Vol}(v), \quad (5-2)$$

where

$$W = |v|^{-n+1} (\bar{\kappa}(p, v/|v|) \kappa(\exp_p(v), \dot{\exp}_p(v)/|v|) + \bar{\kappa}(p, -v/|v|) \kappa(\exp_p(v), -\dot{\exp}_p(v)/|v|)). \quad (5-3)$$

Note that  $|\dot{\exp}_p(v)| = |v|$  in this case.

Next we recall a result in [Stefanov and Uhlmann 2004]. Part (a) is based on formula (5-2) after a change of variables. We denote by  $\rho$  the distance in the metric  $g$ .

**Theorem 5.1.** *Let  $\exp$  be the exponential map of  $M$ . Assume that  $\exp_p : \exp_p^{-1}(M) \rightarrow M$  is a diffeomorphism for  $p$  near  $p_0$ .*

(a) *For  $p$  in the same neighborhood of  $p_0$ ,*

$$X^*Xf(p) = \frac{1}{\sqrt{\det g(p)}} \int A(p, q) \frac{f(q)}{\rho(p, q)^{n-1}} \left| \det \frac{\partial^2(\rho^2/2)}{\partial p \partial q} \right| dq, \quad (5-4)$$

where

$$A(p, q) = \bar{\kappa}(p, -\text{grad}_p \rho) \kappa(q, \text{grad}_q \rho) + \bar{\kappa}(p, \text{grad}_p \rho) \kappa(q, -\text{grad}_q \rho).$$

(b)  *$X^*X$  is a classical  $\Psi$ DO of order  $-1$  with principal symbol*

$$\sigma_p(X^*X)(x, \xi) = 2\pi \int_{S_x M} \delta(\xi(\theta)) |\kappa(x, \theta)|^2 d\sigma_x(\theta), \quad (5-5)$$

where  $\xi(\theta) = \xi_i \theta^i$  and  $\delta$  is the Dirac delta function.



The integral (5-4) is not written in an invariant form but one can easily check by writing it with respect to the volume form that the kernel is invariant. We also note that in the proof of Theorem 2.1, we apply the theorem above by restricting  $\text{supp } f$  and the region where we study  $Nf$  to a small enough neighborhood of  $p_0$ , where there will be no conjugate points. This gives the  $\Psi\text{DO}$  part  $A$  of  $N$  in Theorem 2.1. Finally, note that Theorem 5.2 provides a proof of part (b) even in the context of general exponential maps.

**5b. Mapping properties of  $X$ .** Let  $(x', x^n)$  be semigeodesic coordinates on  $H$  near  $x_0$ . Then  $(x', \xi')$  parametrize the vectors near  $(x_0, \theta_0)$ . We define the Sobolev space  $H^1(\mathcal{H})$  of functions constant along the flow, supported near the flow-out of  $(x_0, \theta_0)$  as the  $H^s$  norm in those coordinates with respect to the measure  $d\mu$ . We can choose another such surface  $H$  near  $q_0$  with some fixed coordinates on it; the resulting norm will be equivalent to that on  $\mathcal{H}$ .

**Proposition 5.2.** *With the notation and the assumptions above, for any  $s \geq 0$ , the operators*

$$X : H_0^s(V) \rightarrow H^{s+1/2}(\mathcal{H}), \tag{5-6}$$

$$X^*X : H_0^s(V) \rightarrow H^{s+1}(V) \tag{5-7}$$

are bounded.

*Proof.* Recall first that the weight  $\kappa$  localizes in a small neighborhood of  $(\gamma_0, \dot{\gamma}_0)$ . Let first  $f$  have small enough support in a set that we will call  $M_0$ . Then  $M_0$  will be a simple manifold if small enough. Then we can replace  $H$  by another surface  $H_0$  that lies in  $M_0$ , and we denote by  $\mathcal{H}_0$  the corresponding  $\mathcal{H}$ . This changes the original parametrization to a new one, which will give us an equivalent norm.

Then, if  $s$  is a half-integer,

$$\|Xf\|_{H^{s+1/2}(\mathcal{H}_0)}^2 \leq C \sum_{|\alpha| \leq 2s+1} |(\partial_{x', \xi'}^\alpha Xf, Xf)_{L^2(\mathcal{H}_0)}| = C \sum_{|\alpha| \leq 2s+1} |(X^* \partial_{x', \xi'}^\alpha Xf, f)_{L^2(\mathcal{H}_0)}|.$$

The term  $\partial_{x', \xi'}^\alpha Xf$  is a sum of weighted ray transforms of derivatives of  $f$  up to order  $|\alpha|$ . Then  $X^* \partial_{x', \xi'}^\alpha X$  is a  $\Psi\text{DO}$  of order  $|\alpha| - 1$  because  $M_0$  is a simple manifold. That easily implies

$$\|Xf\|_{H^{s+1/2}(\mathcal{H}_0)} \leq C \|f\|_{H^s}.$$

The case of general  $s \geq 0$  follows by interpolation; see, e.g., [Taylor 1996, Section 4.2].

To finish that proof, we cover  $\gamma_0$  with open sets so that the closure of each one is a simple manifold. Choose a finite subset and a partition of unity  $1 = \sum \chi_j$  related to that. Then we apply the estimate above to each  $X\chi_j f$  on the corresponding  $\mathcal{H}_j$ . We then have finitely many Sobolev norms that are equivalent, and in particular equivalent to the one on  $\mathcal{H}$ . This proves (5-6).

To prove the continuity of  $X^*X$ , we need to estimate the derivatives of  $X^*X$ . We have that  $\partial^\alpha X^*Xf$  is sum of operators  $X_{\kappa_\alpha}$  of the same kind but with possibly different weights applied to derivatives of  $Xf$  up to order  $|\alpha|$ ; see (5-1). Let first  $s = 0$ . For  $f, h$  in  $C_0^\infty(V)$  and  $|\beta| = 1$ , we have

$$|(f, X_{\kappa_\beta}^* \partial_{x', \xi'}^\beta Xh)_{L^2(V)}| \leq C \|X_{\kappa_\beta} f\|_{H^{1/2}} \|Xh\|_{H^{1/2}} \leq C \|f\|_{L^2(V)} \|h\|_{L^2(V)}.$$

In the last inequality, we used (5-6), which we proved already. This proves (5-7) for  $s = 0$ .

For  $s \geq 1$ , integer, we can “commute” the derivative in  $\partial^\alpha X^* X$  with  $X^* X$  by writing it as a finite sum of operators of the type  $X_\beta^* X_\beta P_\beta f$ , with  $|\beta| \leq |\alpha|$ , where  $P_\beta$  are differential operators of order  $\beta$ . To this end, we first commute it with  $X^*$ , as above, and then with  $X$ . Then we apply (5-7) with  $s = 0$ . The case of general  $s \geq 0$  follows by interpolation.  $\square$

**Remark 5.1.** We did not use the fold condition here. In fact, Proposition 5.2 holds without any assumptions on the type of the conjugate points as long as  $V$  is contained in a small enough neighborhood of a fixed geodesic segment that extends to a larger one with both endpoints outside  $V$ . Proving the mapping properties of  $X^* X$  based on its FIO characterization is not straightforward, and we would get the same conclusion only under some assumptions, for example that the canonical relation is a canonical graph; that is not always true.

**Remark 5.2.** A global version of Proposition 5.2 can easily be derived by a partition of unity in the phase space. Let  $(M, g)$  be a compact nontrapping Riemannian manifold with boundary, that is, one in which all maximal geodesics in  $M$  have a uniform finite bound on their length. Let  $M_1$  be another such manifold whose interior includes  $M$ , and assume that  $\partial M_1$  is strictly convex. Such  $M_1$  always exists if  $\partial M$  is strictly convex. Let  $\partial_- SM_1$  denote the vectors with base point on  $\partial M$  pointing into  $M_1$ . Then we can parametrize all (directed) geodesics with points in  $\partial_- SM_1$ , which plays the role of  $\mathcal{H}$  above. Then for  $s \geq 0$ ,

$$X : H_0^s(M) \rightarrow H^{s+1/2}(\partial_- SM_1), \quad X^* X : H_0^s(M) \rightarrow H^{s+1}(M_1)$$

are bounded.

**5c. General regular exponential maps.** Let now  $\exp$  be a regular exponential map. As above, we split the  $t$ -integral in the second line below into two parts to get

$$\begin{aligned} Nf(p) &= \int \kappa^\sharp(p, \theta) Xf(p, \theta) d\sigma_p(\theta) \\ &= \int_{S_p M} \int \kappa^\sharp(p, \theta) \kappa(\exp_p(t\theta), \dot{\exp}_p(t\theta)) f(\exp_p(t\theta)) dt d\sigma_p(\theta) \\ &= \int_{T_p M} W(p, v) f(\exp_p(v)) d\text{Vol}(v), \end{aligned} \tag{5-8}$$

where

$$W = |v|^{-n+1} (\kappa^\sharp(p, v/|v|) \kappa(\exp_p(v), \dot{\exp}_p(v)/|v|) + \kappa^\sharp(p, -v/|v|) \kappa(\exp_p(v), -\dot{\exp}_p(v)/|v|)). \tag{5-9}$$

**Theorem 5.2.** Let  $\exp_p(v)$  satisfy (R1) and (R4) and assume for any  $(p, \theta) \in \text{supp } \kappa^\sharp$  that  $t\theta$  is not a conjugate vector at  $p$  for  $t$  such that  $\exp_p(t\theta) \in \text{supp } f$ . Then  $N$  is a classical  $\Psi DO$  of order  $-1$  with principal symbol

$$\sigma_p(N)(x, \xi) = 2\pi \int_{S_x M} \delta(\xi(\theta)) (\kappa^\sharp \kappa)(x, \theta) d\sigma_x(\theta), \tag{5-10}$$

where  $\xi(\theta) = \xi_i \theta^j$  and  $\delta$  is the Dirac delta function.

*Proof.* The theorem is essentially proved in [Frigyik et al. 2008, Section 4], where the exponential map is related to a geodesic like family of curves. We will repeat the arguments there in this more general situation.

Notice first that it is enough to study small enough  $|t|$ . Fix local coordinates  $x$  near  $p_0$ . By (R4),

$$\exp_x(t\theta) = x + tm(t, \theta; x), \quad \text{where } m(0, \theta; x) = \theta,$$

with a smooth function  $m$  near  $(0, \theta_0, p_0)$ . Introduce new variables  $(r, \omega) \in \mathbb{R} \times S_x M$  by

$$r = t|m(t, \theta; x)| \quad \text{and} \quad \omega = m(t, \theta; x)/|m(t, \theta; x)|,$$

where  $|\cdot|$  is the norm in the metric  $g(x)$ . Then  $(r, \omega)$  are polar coordinates for  $\exp_x(t\theta) - x = r\omega$  with  $r$  that can be negative as well, that is,

$$\exp_x(t\theta) = x + r\omega.$$

The functions  $(r, \omega)$  are clearly smooth for  $|t| \ll 1$  and  $x$  close to  $p_0$ . Let

$$J(t, \theta; x) = \det d_{t,v}(r, \omega)$$

be the Jacobi determinant of the map  $(t, v) \mapsto (r, \omega)$ . By (R4),  $J|_{t=0} = 1$ ; therefore that map is a local diffeomorphism from  $(-\varepsilon, \varepsilon) \times S_x M$  to its image for  $0 < \varepsilon \ll 1$ . It is not hard to see that for  $0 < \varepsilon \ll 1$  it is also a global diffeomorphism because it is clearly injective. Let  $t = t(x, r, \omega)$  and  $\theta = \theta(x, r, \omega)$  be the inverse functions defined by that map. Then

$$t = r + O(|r|), \quad \theta = \omega + O(|r|), \quad \exp(t\theta) = \omega + O(|r|).$$

Assume that the weight  $\kappa$  in (2-2) vanishes for  $p$  outside some small neighborhood of  $p_0$ . Then after a change of variables, we get

$$Nf(x) = \int_{S_x M} \int A(x, r, \omega) f(x + r\omega) dr d\sigma_x(\omega),$$

where

$$A(x, r, \omega) = \kappa^\sharp(x, \theta(x, r, \omega))\kappa(x + r\omega, \omega + rO(1))J^{-1}(x, r, \omega)$$

with  $J$  as before, but written in the variables  $(x, r, \omega)$ . By [Frigyik et al. 2008, Lemma 4.2],  $N$  is a classical  $\Psi$ DO of order  $-1$  with principal symbol

$$2\pi \int_{S_x M} \delta(\xi(\omega))A(x, 0, \omega) d\sigma_x(\omega) = 2\pi \int_{S_x M} \delta(\xi(\omega))\kappa^\sharp(x, \omega)\kappa(x, \omega) d\sigma_x(\omega). \quad (5-11)$$

The proof in [Frigyik et al. 2008] starts with the change of variables  $y = x + r\omega$ . Then we write the Schwartz kernel of  $N$  as a singular one with leading part  $2A_{\text{even}}(x, 0, \omega)|x - y|^{-1}$ , where  $\omega = (y - x)/|y - x|$  and  $A_{\text{even}}$  is the even part of  $A$  with respect to  $\omega$ . It then follows that  $N$  is a  $\Psi$ DO of order  $-1$  with a principal symbol as claimed.  $\square$

**Remark 5.3.** Formulas (5-2) and (5-8) are valid regardless of possible conjugate points. In our setup, the supports of  $\kappa$  and  $\kappa^\sharp$  guarantee that  $\exp_p(t\theta)$  for  $(p, \theta)$  close to  $(p_0, \theta_0)$  reaches a conjugate point for  $t > 0$  but not for  $t < 0$ . Therefore, near the conjugate point  $q$  of  $p$ , the second term on the right sides of (5-3), and (5-9), respectively, vanishes.

## 6. The Schwartz kernel of $N$ near the conjugate locus $\Sigma$

We will introduce first three invariants. Let  $F : M \rightarrow N$  be a smooth orientation-preserving map between two orientable Riemannian manifolds  $(M, g)$  and  $(N, h)$ . Then one defines  $\det dF$  invariantly by

$$F^*(d\text{Vol}_N) = (\det dF) d\text{Vol}_M; \quad (6-1)$$

see also [Lang 1995, X.3]. In local coordinates,

$$\det dF(x) = \sqrt{\frac{\det h(F(x))}{\det g(x)}} \det \frac{\partial F(x)}{\partial x}. \quad (6-2)$$

We choose an orientation of  $S(p_0)$  near  $v_0$ , as a surface in  $T_{p_0}M$ , by choosing a unit normal field so that the derivative of  $\det d\exp_{p_0}(v)$  along it is positive on  $S(p)$ . Then we extend this orientation to  $S(p)$  for  $p$  close to  $p_0$  by continuity. In Figure 2, the positive side is the one below  $S(p)$  if  $v$  is the first conjugate vector along the geodesic through  $(p, v)$ . Then we choose an orientation of  $\Sigma(p)$  such that the positive side is that in the range of  $\exp_p$ . In Figure 2, the positive side is to the left of  $\Sigma(p)$ . The so chosen orientations conform with the signs of  $\xi^n$  and  $y^n$  in the normal form (3-2).

Next we synchronize the orientations of  $T_pM$  and  $M$  near  $q$  by postulating that  $\exp_p$  is an orientation-preserving map from the positive side of  $S(p)$ , as described above, to the positive side of  $\Sigma(p)$ .

For each  $p \in M$ , the transformation laws in  $TT_pM$  under coordinate changes on the base show that  $T_pM$  has the natural structure of a Riemannian manifold with the constant metric  $g(p)$ . Then one can define  $\det d\exp_p$  invariantly as above. Let  $d\text{Vol}_p$  be the volume form in  $T_pM$ , and let  $d\text{Vol}$  be the volume form in  $M$ . Then  $\det d\exp_p$  is defined invariantly by

$$\exp_p^* d\text{Vol} = (\det d\exp_p) d\text{Vol}_p. \quad (6-3)$$

In local coordinates,

$$\det d\exp_p = \sqrt{\frac{\det g(\exp_p v)}{\det g(p)}} \det \frac{\partial}{\partial v} \exp_p(v),$$

where, with some abuse of notation,  $g(p)$  is the metric  $g$  in fixed coordinates near a fixed  $p_0$ , and  $g(\exp_p v)$  is the metric  $g$  in a possibly different system of fixed coordinates near  $q_0 = \exp_{p_0} v_0$ . Set

$$A(p, v) := |\det d\exp_p(v)|. \quad (6-4)$$

Since  $\det d\exp_p(v)$  is a defining function for  $S(p)$ , its differential is conormal to it. By the fold condition,  $A \neq 0$ . One can check directly that  $A$  is invariantly defined on  $\Sigma$ .



By (3-3), for  $(p, v) \in S$ , the differential of  $\exp_p$  maps isomorphically  $T_v S(p)$  (equipped with the metric on that plane induced by  $g(p)$ ) into  $T_q \Sigma$ , with the induced metric. Let  $D$  be the determinant of  $\exp_p|_{S(p)}$ , that is,

$$D := \det(d \exp_p|_{T_v S(p)}), \quad (6-5)$$

defined invariantly by (6-1). We synchronize the orientations of  $S(p)$  and  $\Sigma(p)$  so that  $D > 0$ .

We express next the weight  $W(p, v)$  restricted to  $S$  in terms of the variables  $(p, q)$ . For  $(p, q) \in \Sigma$ ,  $v = \exp_p^{-1}(q)$ , where we inverted  $\exp_p$  restricted to  $S$ . Let  $w = w(p, q)$  be defined as in (3-1) with  $v$  as above. Then we set (see also (5-9), and Remark 5.3)

$$W_\Sigma(p, q) := W(p, \exp_p^{-1}(q))|_\Sigma = |v|^{1-n} \kappa^\sharp(p, v/|v|) \kappa(q, -w/|v|). \quad (6-6)$$

For  $p$  close to  $p_0$ ,  $\Sigma(p)$  divides  $M$  in a neighborhood of  $q_0$  into two parts: One of them is in the range of  $\exp_p(v)$  for  $v$  near  $v_0$  (this is the positive one with respect to the chosen orientation); the other is not. Let  $z'(p, q)$  be the distance from  $q$  to  $\Sigma(p)$  with a positive sign in the first region, and with a negative sign in the second. Then  $z' = z'(p, q)$ , for fixed  $p$ , is a normal coordinate to  $\Sigma(p)$  depending smoothly on  $p$ , and  $\Sigma$  is given locally by  $z' = 0$ . Then  $z'$  is a defining function for  $\Sigma$ , that is,  $\Sigma = \{z' = 0\}$  and  $d_{p,q} z' \neq 0$  because  $d_q z' \neq 0$ . Let  $z'' = z''(p, q) \in \mathbb{R}^{2n-1}$  be such that its differential restricted to  $T\Sigma$  is an isomorphism at  $(p_0, q_0)$ . Since  $dz''$  and  $dz'$  are linearly independent,  $z = z(z', z'')$  are coordinates near  $(p_0, q_0)$ . One way to construct  $z''$  is the following. Choose  $(z_{n+1}, \dots, z_{2n})$ , depending on  $p$  only, to be local coordinates for  $p$ , and to choose  $(z', z_2, \dots, z_n)$ , depending on  $p$  and  $q$ , to be semigeodesic coordinates of  $q$  near  $\Sigma(p)$ .

The next theorem shows that near  $\Sigma$ , the operator  $N$  has a singular but integrable kernel with a conormal singularity of the type  $1/\sqrt{z'}$ .

**Theorem 6.1.** *Near  $\Sigma(p)$ , the Schwartz kernel  $N(p, q)$  of  $N$  (with respect to the volume measure) near  $(p_0, q_0)$  is of the form*

$$N = W_\Sigma \frac{\sqrt{2}}{\sqrt{ADz'}} (1 + \sqrt{z'} R(\sqrt{z'}, z'')), \quad (6-7)$$

where  $W_\Sigma = W_\Sigma(z'')$ ,  $A = A(z'')$ ,  $D = D(z'')$ , and  $R$  is a smooth function.

*Proof.* We start with the representation (5-8). We will make the change of variables  $y = \exp_p(v)$  for  $(p, v)$  close to  $(p_0, v_0)$  as always. Then  $y$  will be on the positive side of  $\Sigma(p)$ , and the exponential map is 2-to-1 there. We split the integration in (5-8) in two parts: One, where  $v$  is on the positive side of  $S(p)$ , that we call  $N_+ f$ , and the other one we denote by  $N_- f$ . Then

$$N_\pm f(p) = \int_{S_p M} \int W f(y) (\det d \exp_p^\pm(v))^{-1} d\text{Vol}(y), \quad (6-8)$$

where  $W$  is as in (6-6) but not restricted to  $\Sigma$ , and  $(\exp_p^\pm)^{-1}$  there is the corresponding inverse in each of the two cases.

To prove the theorem, we need to analyze the singularity of the Jacobian determinant  $\det d \exp_p(v)$  near  $\Sigma(p)$ . It is enough to do this at  $(p_0, v_0)$ .

Let  $y = (y', y^n)$  be semigeodesic coordinates near  $\Sigma(q_0)$ , with  $q_0 = \exp_{p_0}(v_0)$ , and let  $y_0$  correspond to  $q_0$ . We assume that  $y^n > 0$  on the positive side of  $\Sigma(p)$ . In other words,  $y^n = z'(p_0, q)$ .

We have

$$d\text{Vol}(y) = \det(d_v \exp_p(v)) d\text{Vol}(v).$$

The form on the left can be written as  $d\text{Vol}_{\Sigma(p)}(y') dy^n$ , while the one on the right, restricted to  $S(p)$ , equals  $d\text{Vol}_{S(p)}(v') dv^n$  in boundary normal coordinates to  $S(p)$ , where  $v^n > 0$  gives the positive side of  $S(p)$ . On the other hand, by (6-5),

$$d\text{Vol}_{\Sigma(p)}(y') = D d\text{Vol}_{S(p)}(v').$$

We therefore get

$$D dy^n = \det(d \exp_p(v)) dv^n.$$

By the definition of  $A$ , we have

$$\det d_v \exp_p(v) = Av^n(1 + O(v^n)). \quad (6-9)$$

Therefore,

$$D dy^n = A(1 + O(v^n)) v^n dv^n.$$

Since  $y^n = 0$  for  $v^n = 0$ , we get

$$y^n = (v^n)^2 \frac{A}{2D} (1 + O(v^n)).$$

Solve this for  $v^n$  and plug into (6-9) to get

$$\det d \exp_p(v) = \pm \sqrt{2ADy^n} (1 + O_{\pm}(\sqrt{y^n})). \quad (6-10)$$

Here  $O_{\pm}(\sqrt{y^n})$  denotes a smooth function of  $\sqrt{y^n}$  near the origin with coefficients smooth in  $y'$ , which vanishes at  $y^n = 0$ . The positive/negative sign corresponds to  $v$  belonging to the positive/negative side of  $S(p)$ . By (6-8),

$$N_{\pm} f(p) = \int W f(y) \frac{1}{\sqrt{2ADy^n}} (1 + O_{\pm}(\sqrt{y^n})) d\text{Vol}(y). \quad (6-11)$$

We replace  $A_0$  and  $D_0$  in (6-11) by their values at  $y^n = 0$ ; the error will then just replace the remainder term above by another one of the same type. Similarly,  $W = W(p, v)$ , where  $\exp_p(v) = q$ . Solving the latter for  $v = v(p, q)$  provides a function having a finite Taylor expansion in powers of  $\sqrt{y^n}$  of any order, with smooth coefficients. The leading term is what we denoted by  $W_{\Sigma}$ , a smooth function on  $\Sigma$ .

With the aid of (6-2), it is easy to see that (6-11) is a coordinate representation of the formula (6-7) at the so fixed  $p$ . When  $p$  varies near  $p_0$ , it is enough to notice that since we already wrote the integral in invariant form,  $y^n$  then becomes the function  $z'(p, q)$  introduced above. For  $z''$  we then have  $z''(p, q) = (x(p), y'(p, q))$ . Finally, we note that another choice of  $z''$  such that  $(z', z'')$  are coordinates would preserve (6-7) with a possibly different  $R$ .  $\square$

**7.  $N$  as a Fourier integral operator: Proof of Theorem 2.1**

We are ready to finish the proof of Theorem 2.1. By Theorem 6.1, near  $\Sigma$ , the Schwartz kernel of  $N$  has a conormal singularity at  $\Sigma$ , supported on one side of it, that admits a singular expansion in powers of  $\sqrt{z'_+}$ , with leading singularity  $1/\sqrt{z'_+}$ . The Fourier transform of the latter is

$$\sqrt{\pi} e^{-i\pi/4} (\zeta_+^{-1/2} + i\zeta_-^{-1/2}), \tag{7-1}$$

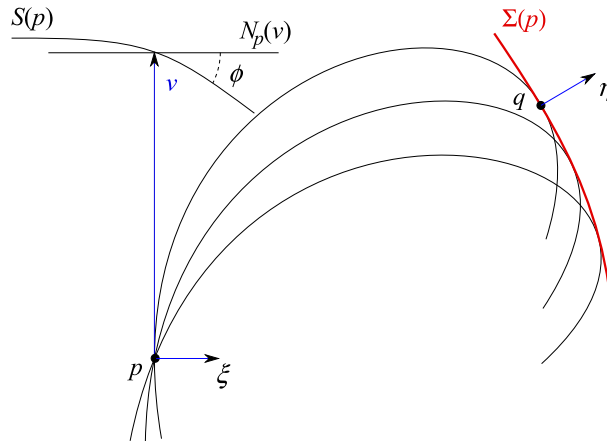
where  $\zeta_+ = \max(\zeta, 0)$  and  $\zeta_- = (-\zeta)_+$ . The singularity near  $\zeta = 0$  can be cut off, and we then get a symbol of order  $-1/2$ , depending smoothly on the other  $2n - 1$  variables. Therefore, near  $\Sigma$ , the kernel of  $N$  belongs to the conformal class  $I^{-n/2}(M \times M, \Sigma; \mathbb{C})$ ; see e.g., [Hörmander 1985a, 18.2]. It is elliptic when  $\kappa^\sharp(p_0, \theta_0)\kappa^\sharp(q_0, -w_0) \neq 0$  by (5-9) and (6-6). Therefore, the kernel of  $N$  near  $\Sigma$  is a kernel of an FIO associated to the Lagrangian  $T^*\Sigma$ . Moreover, the amplitude of the conormal singularity at  $\Sigma$  is in the class  $S_{\text{phg}}^{-1/2, 1/2}$  (polyhomogeneous of order  $-1/2$ , having an asymptotic expansion in integer powers of  $|\zeta|^{1/2}$ ); see also (9-12) and (9-13).

**8. The two-dimensional case**

**Theorem 8.1.** *Let  $\dim M = 2$ . Assume that (R1)–(R5) are fulfilled. Then  $\mathcal{N}^*\Sigma \setminus 0$ , near  $(p_0, \xi_0, q_0, \eta_0)$ , is the graph of a local diffeomorphism  $T^*M \setminus 0 \in (p, \xi) \mapsto (q, \eta) \in T^*M \setminus 0$ , homogeneous of order one in its second variable (a canonical graph).*

*Proof.* For  $(p, \xi)$  near  $(p_0, \xi_0)$ , there are exactly two smooth maps that map  $\xi$  to a unit normal vector. We choose the one that maps  $\xi_0$  to  $v_0/|v_0|$ . Then we map the latter to  $v \in S(p)$ . Since the radial ray through  $v$  is transversal to  $S(p)$ , that map is smooth. Knowing  $v$ , then we can express  $q = \exp_p(v) \in \Sigma(p)$  and  $w = -\exp_p(v)$  as smooth functions of  $(p, \xi)$  as well. Then in local coordinates,  $\eta = \xi_i \partial \exp_q^i(w) / \partial q$  (see (4-12)), which in particular proves the homogeneity.

By (R5), this map is invertible. □



**Figure 3.** The 2D case.

The principal symbol of  $X^*X$  in the geodesics case (see Theorem 5.1, and (5-5)) is given by

$$\sigma_p(X^*X)(x, \xi) = 2\pi |\kappa(x, \xi^\perp / |\xi^\perp|)|^2, \quad (8-1)$$

where  $\xi^\perp$  is a continuous choice of a vector field normal to  $\xi$  and of the same length such that  $\xi_0^\perp / |\xi_0^\perp| = \theta_0$  and  $-\xi_0^\perp / |-\xi_0^\perp| = \theta_0$  at  $p = p_0$ ; therefore, the sign of the angle of rotation is different near  $\xi_0$  and near  $-\xi_0$ . Notice that (5-5) in the two-dimensional case is a sum of two terms but we assumed that  $\kappa$  is supported near  $(p_0, \theta_0)$ ; therefore only one of the terms is nontrivial. A similar remark applies to (5-10).

Theorem 6.1 takes the following form in two dimensions, in the Riemannian case.

**Corollary 8.1.** *Let  $n = 2$  and let  $\exp$  be the exponential map of a Riemannian metric. With the notation of Theorem 6.1, we then have*

$$N = W_\Sigma \frac{\sqrt{2}}{\sqrt{Bz'}} (1 + \sqrt{z'} R(\sqrt{z'}, z'')), \quad (8-2)$$

where

$$B = \left| \frac{d}{dN} \det d \exp_p(v) \right|$$

is evaluated at  $v \in S(p)$  such that  $q = \exp_p(v)$ , and  $d/dN$  stands for the derivative in the direction of  $N_p(v)$ .

*Proof.* Note first that  $B \neq 0$  by the fold condition. Let  $\phi$  be the (acute) angle between  $S(p)$  and  $N_p(v)$  at  $v$ . Since  $N_p(v)$  is orthogonal to the radial ray at  $v$ , we can introduce an orthonormal coordinate system at  $v$  with the first coordinate vector being  $v/|v|$ , and the second one the positively oriented unit vector along  $N_p(v)$ , which we call  $\xi$ . Let us parallel transport this frame along the geodesic  $\gamma_{p,v}$  and invert the direction of the tangent vector to conform with our choice of  $w$  at  $q$ . In particular, this introduces a similar coordinate system near the corresponding vector  $w$  at  $q$  in the conjugate locus. In these coordinates then

$$d \exp_p(v) = \begin{pmatrix} -1 & 0 \\ 0 & j/|v| \end{pmatrix}, \quad (8-3)$$

where  $j$  is uniquely determined by  $J(t) = j(t)\Xi(t)$ , where  $J(t)$  is the Jacobi field with  $J(0) = 0$ ,  $J'(0) = \xi$ , and  $\Xi(t)$  is the parallel transport of  $\xi$ ; compare that with (4-1). The extra factor  $1/|v|$  comes from the fact that we normalize  $v$  now in our basis, so that the result would be the Jacobian determinant. Then the Jacobi determinant  $\det d \exp_p(v)$  is given by  $-j/|v|$ . In particular, for  $(p, v) \in S$  we have  $d \exp_p(v) = \text{diag}(-1, 0)$ . Note that  $j$  depends on  $v$  as well; therefore its differential, which essentially gives  $d \det d \exp_p(v)$ , depends on the properties of the Jacobi field under a variation of the geodesic.

Now, it easily follows from the definition (6-5) of  $D$  that  $D = \sin \phi$ . On the other hand,  $d \det d \exp_p(v)$  is conormal to  $S(p)$ ; therefore, the derivative of  $\det d \exp_p(v)$  in the direction of  $N_p(v)$  satisfies

$$\left| \frac{d}{dN} \det d \exp_p(v) \right| = |d \det d \exp_p(v)| \sin \phi = A \sin \phi = AD. \quad \square$$



**9. Resolving the singularities in the geodesic case**

As before, let  $(p_0, q_0)$  be a pair of fold conjugate points along  $\gamma_0$ , and  $X$  be the ray transform with a weight that localizes near  $\gamma_0$ . We want to see whether we can resolve the singularities of  $f$  near  $p_0$  and near  $q_0$  knowing that  $Xf \in C^\infty$ , and more generally, whether we can invert  $X$  microlocally. Assume for simplicity that  $p_0 \neq q_0$ .

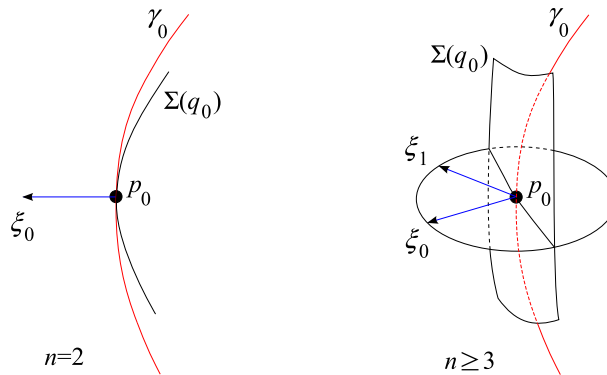
We will restrict ourselves to the geodesic case only but the same analysis holds without changes to the case of magnetic geodesics as well. We avoid the formal introduction of magnetic geodesics for simplicity of the exposition. Assume also that

$$\kappa(p, \theta)\kappa(q, -w/|w|) \neq 0 \quad \text{for } (p, \theta) \in \mathcal{U}_0, \tag{9-1}$$

where  $(q, w)$  are given by (3-1), and  $\mathcal{U} \ni \mathcal{U}_0 \ni (p_0, \theta_0)$ . This guarantees the microlocal ellipticity of the  $\Psi$ DO  $A$  near  $\mathcal{N}^*(p_0, v_0)$  and  $\mathcal{N}^*(q_0, w_0)$  in Theorem 2.1; see Theorem 5.1.

**9a. Sketch of the results.** We explain the results before first in an informal way. As we pointed out in the introduction,  $Xf(\gamma)$  for geodesics near  $\gamma_0$  can only provide information for  $\text{WF}(f)$  near  $\mathcal{N}^*\gamma_0$ , and does not “see” the other singularities. The analysis below, based on Theorem 2.1, shows that on a principal symbol level, the operator  $|D|^{1/2}F$  behaves as a Radon type of transform on the curves (when  $n = 2$ ) or the surfaces (when  $n \geq 3$ )  $\Sigma(p)$ . Similarly, its adjoint behaves as a Radon transform on the curves/surfaces  $\Sigma(q)$ . Therefore, there are two geometric objects that can detect singularities at  $p_0$  conormal to  $v_0$ : the geodesic  $\gamma_0 = \gamma_{p_0, v_0}$  (and those close to it) and the conjugate locus  $\Sigma(q_0)$  through  $p_0$  (and those corresponding to perturbations of  $v_0$ ). We refer to Figure 4.

When  $n = 2$ , the information coming from integrals along the two curves (and their neighborhoods) may in principle cancel; and we show in Theorem 9.2 that this actually happens, at least to order one. When  $n \geq 3$ , the Radon transform over  $\Sigma(q) \ni p$  competes with the geodesic transform over geodesics through  $p$ . Depending on the properties of that Radon transform, the information that we get for  $\pm\xi_0$



**Figure 4.** Two geometric objects can detect singularities at  $p_0$  in the geodesic case: a geodesic  $\gamma_0$  through  $p_0$ , and the conjugate locus  $\Sigma(q_0)$  of  $q_0$  conjugate to  $p_0$ . By Theorem 4.2,  $\gamma_0$  is parallel to  $\Sigma(q_0)$ .

may or may not cancel because  $\xi_0$  is conormal both to  $\gamma_0$  and  $\Sigma(q_0)$ . On the other hand, for any other  $\xi_1$  conormal to  $v_0$  but not parallel to  $\xi_0$ , the geodesic  $\gamma_0$  (and those close to it) can detect whether it is in  $\text{WF}(f)$  but the Radon transform restricted to small perturbations of  $v_0$  (and therefore of  $q_0$ ) will not. Thus, we can invert  $N$  microlocally at such  $(p_0, \xi_1)$ .

Now, when  $n \geq 3$ , we may try to invert  $N$  even at  $\xi_0$  by choosing  $v$ 's close to  $v_0$  but normal to  $\xi_0$ . If  $\xi_0$  happens not to be conormal to the corresponding conjugate locus  $\Sigma(q(p_0, v))$  at  $p_0$ , we can just use the argument above with the new  $v$ . In particular, if the map (4-3) is a local diffeomorphism, this can be done.

This suggests the following sufficient condition for inverting  $N$  at  $(p_0, \xi_1)$ :

$$\begin{aligned} &\text{There is some } \theta_1 \in S_{p_0}M \text{ such that } \kappa(p_0, \theta_1) \neq 0, \quad \xi_1(\theta_1) = 0, \\ &\text{and } \xi_1 \text{ is not conormal to } \Sigma(q(p_0, \theta_1)) \text{ at } p_0. \end{aligned} \quad (9-2)$$

Above,  $\Sigma(q(p_0, \theta_1))$  is the conjugate locus to the point  $q$  that is conjugate to  $p_0$  along  $\gamma_{p_0, \theta_1}$ . We normally denote that point by  $q(p_0, v_1)$ , where  $v_1 \in S(p_0)$  has the same direction as  $\theta_1$ .

In case of the geodesic transform, one could formulate (9-2) in terms of the map (4-3) as follows:

$$\begin{aligned} &\text{There exists } v_1 \in S(p_0) \text{ such that } \kappa(p_0, v_1/|v_1|) \neq 0, \quad \xi_1(v_1) = 0, \\ &\text{and } \xi_1 \text{ is not the image of } v_1 \text{ under the map (4-3) at } p_0. \end{aligned} \quad (9-3)$$

In Section 10c, we present an example where (4-3) is a local diffeomorphism, and therefore (9-2) holds. In Section 10d we present another example where (9-2) fails.

**9b. Recovery of singularities in all dimensions.** We proceed next with analysis of the recovery of singularities.

Let  $\chi_{1,2}$  be smooth functions on  $M$  that localize near  $p_0$ , and  $q_0$ , respectively, that is,  $\text{supp } \chi_1 \subset U_1$  and  $\text{supp } \chi_2 \subset U_2$ , where  $U_{1,2}$  are small enough neighborhoods of  $p$  and  $q$ , respectively. Assume that  $\chi_1, \chi_2$  equal 1 in smaller neighborhoods of  $p_0, q_0$ , where  $f_1, f_2$  are supported. Then  $f := f_1 + f_2$  is supported in  $U_1 \cup U_2$  and we can write

$$\chi_1 Nf = A_1 f_1 + F_{12} f_2, \quad (9-4)$$

where  $A_1 = \chi_1 N \chi_1$  is a  $\Psi$ DO by Theorem 5.2, while  $F_{12} = \chi_1 N \chi_2$  is the FIO that we denoted by  $F$  in Theorem 2.1. By (R5), we can do the same thing near  $q_0$  to get

$$\chi_2 Nf = A_2 f_2 + F_{21} f_1, \quad (9-5)$$

where  $A_2 = \chi_2 N \chi_2$ ,  $F_{21} = \chi_2 N \chi_1$ . It follows immediately that  $F_{21} = F_{12}^*$ . Recall that  $F_{12} = F$  in the notation of Theorem 2.1. Assuming  $X^* X f \in C^\infty$ , we get

$$A_1 f_1 + F f_2 \in C^\infty \quad \text{and} \quad A_2 f_2 + F^* f_1 \in C^\infty. \quad (9-6)$$

Solve the first equation for  $f_2$ , and plug into the second one to get

$$(\text{Id} - A_2^{-1} F^* A_1^{-1} F) f_2 \in C^\infty \quad \text{near } (q_0, \pm \eta_0), \quad (9-7)$$

where  $A_1^{-1}$  and  $A_2^{-1}$ , denote parametrices of  $A_1$  and  $A_2$  near  $(p_0, \pm\xi_0)$  and  $(q_0, \pm\eta_0)$ , respectively. The operator in the parentheses is a  $\Psi$ DO of order 0 if the canonical relation is a graph, which is true in particular when  $n = 2$ , by Theorem 8.1. In that case, if  $\text{Id} - A_2^{-1}F^*A_1^{-1}F$  is an elliptic (as a  $\Psi$ DO of order 0) near  $(q_0, \pm\eta_0)$ , then we can recover the singularities. Without the canonical graph assumption, if it is hypoelliptic, then we still can.

Another way to express the arguments above is the following. Since  $\chi_{1,2}$  together with  $\kappa$  restrict to conic neighborhoods of  $(p_0 \pm \xi_0)$  and  $(q_0 \pm \eta_0)$ , respectively, and  $A_{1,2}, F, F^*$  have canonical relations of graph type that preserve the union of those neighborhoods, we may think of  $f = f_1 + f_2$  as a vector  $f = (f_1, f_2)$ , and then

$$F = \begin{pmatrix} A_1 & F \\ F^* & A_2 \end{pmatrix}. \tag{9-8}$$

The operator  $\text{Id} - A_2^{-1}F^*A_1^{-1}F$  can be considered then as the ‘‘determinant’’ of  $F$ , up to elliptic factors.

**Theorem 9.1.** *Let the canonical relation of  $F$  be a canonical graph. With the assumptions and the notation above, if the zeroth order  $\Psi$ DO*

$$\text{Id} - A_2^{-1}F^*A_1^{-1}F \tag{9-9}$$

*is elliptic in a conic neighborhood of  $(q_0, \pm\eta_0)$ , then  $Xf \in C^\infty$  near  $(p_0, \theta_0)$  (or more generally,  $Nf \in C^\infty$  near  $p_0$  and  $q_0$ ) implies  $f \in C^\infty$ .*

In the geodesic case in two dimensions, the principal symbol of  $A_2^{-1}F^*A_1^{-1}F$  is always 1; see Proposition 9.1 below.

When  $n \geq 3$  and  $F$  is of graph type, then  $A_2^{-1}F^*A_1^{-1}F$  is of negative order; therefore we can resolve the singularities.

**Corollary 9.1.** *Let  $n \geq 3$  and assume that the canonical relation of  $F$  is a canonical graph. Then the conclusions of Theorem 9.1 hold, that is,  $Xf \in C^\infty$  near  $(p_0, \theta_0)$  (or more generally,  $Nf \in C^\infty$  near  $p_0$  and  $q_0$ ) implies  $f \in C^\infty$ .*

*Proof.* In this case,  $A_1^{-1}F$  is an FIO of order  $1 - n/2$  with the same canonical relation as  $F$ . Similarly  $A_2^{-1}F^*$  is an FIO of order  $1 - n/2$  with a canonical relation that is a graph of the inverse canonical map. Their composition is therefore a  $\Psi$ DO of order  $2 - n < 0$ . Its principal symbol as a  $\Psi$ DO of order 0 is zero. The corollary now follows from Theorem 9.1.  $\square$

In Section 10c, we give an example where the assumptions of the corollary hold. Note that those assumptions are stable under small perturbations of the dynamical system.

When the graph condition does not hold, the analysis is harder. Then (4-3) is not a local diffeomorphism. If its range is a lower dimension submanifold, for example, we can at least recover the conormal singularities to  $\theta_0$  away from it, as the corollary below implies. Note that below, (b) implies (a). Also, (9-1) is not needed; only ellipticity of  $\kappa$  at  $(p_0, \theta_0)$  suffices.

**Corollary 9.2.** *Let  $Xf \in C^\infty$  for  $\gamma$  near  $\gamma_0$ .*

(a) If  $\xi_1 \in T_{p_0}M \setminus 0$  is conormal to  $v_0$  but not conormal to  $\Sigma(q_0)$  (not parallel to  $\xi_0$ ), then

$$(p_0, \xi_1) \notin \text{WF}(f).$$

(b) The same conclusion holds if condition (9-2) or the equivalent (9-3) is fulfilled.

*Proof.* Note first that  $A_1$  is elliptic at  $(p_0, \zeta)$  by (9-1) and Theorem 5.1(b). By the first relation in (9-6),  $(p_0, \xi_1) \in \text{WF}(f_1)$  if and only if  $(p_0, \xi_1) \in \text{WF}(Ff_2)$ . To analyze the latter, we will use the relation  $\text{WF}(Ff_2) \subset \text{WF}'(F) \circ \text{WF}(f_2)$ ; see [Hörmander 1983, Theorem 8.5.5], noting that in the notation there,  $\text{WF}(F)_X$  is empty. By Theorem 6.1,  $\text{WF}'(F)$  consists of those points in the canonical relation  $\mathcal{C}$ ; see (2-5), for which the conormal singularity in (6-7) is not canceled by a zero weight.

Now, let  $\xi_1$  be as in (a). Since  $\xi_1$  is separated by  $\pm\xi_0$  by a conic neighborhood, one can choose a weight  $\chi$  on  $SM$  that is constant along the geodesic flow, nonzero at  $(p_0, \theta_0)$  and supported in a flow-out of a neighborhood  $\mathcal{V}$  of it small enough such that the conormals to the corresponding conjugate loci at  $p_0$  stay away from a neighborhood of  $\xi_1$ . In the geodesics case, the condition is that the map (4-3) restricted to  $\mathcal{V}$  does not intersect a chosen small enough conic neighborhood of  $\pm\xi_0$ . This can always be done by continuity arguments. Then left projection of  $\text{WF}'(F)$  will not be singular at  $(p_0, \xi_1)$ , and therefore,  $Ff_2$  will have the same property regardless of the singularities of  $f_2$ .

Statement (b) follows from (a) by varying  $v$  near  $v_0$  in directions normal to  $\xi_1$ .  $\square$

**9c. Calculating the principal symbol of (9-9) in case of Riemannian surfaces.** Let  $\exp$  be the exponential map of  $g$ , and let  $n \geq 2$ . We will take  $n = 2$  later. Recall that the leading singularity of the kernel of  $N$  near  $\Sigma$  is of the type  $(z'_+)^{-1/2}$ , by Theorem 6.1. We will compose  $F$  with a certain  $\Psi\text{DO}$   $R$  so that this singularity becomes of the type  $\delta(z')$ . Then modulo lower order terms,  $FRf(p)$  will be a weighted Radon transform over the surface  $\Sigma(p)$ . In 2D, that will be an X-ray type of transform. We are only interested in this composition acting on distributions with wave front sets in a small conic neighborhood  $\mathcal{W}$  of  $(q_0, \pm\eta_0)$ .

The Fourier transform of  $(z'_+)^{-1/2}$  is given by (7-1). Its reciprocal is

$$\pi^{-1/2} e^{i\pi/4} (h(\zeta)\zeta^{1/2} - ih(-\zeta)(-\zeta)^{1/2}) = \pi^{-1/2} e^{i\pi/4} (h(\zeta) - ih(-\zeta)) |\zeta|^{1/2},$$

where  $h$  is the Heaviside function and  $|\zeta|$  is the norm in  $T_y^*M$ . We fix  $p$  near  $p_0$  and local coordinates  $x = x(p)$  there, and we work in semigeodesic coordinates  $y = y(p, q)$  near  $q_0$  normal to  $\Sigma(p)$  oriented as in Section 6. Let  $x$  denote local coordinates near  $q_0$ . Let  $R$  be a properly supported  $\Psi\text{DO}$  of order  $1/2$  with principal symbol equal to

$$r(y, \eta) = \pi^{-1/2} e^{i\pi/4} (h(\eta_n) - ih(-\eta_n)) |\eta|^{1/2} r_0(y, \eta) \quad (9-10)$$

in  $\mathcal{W}$ , outside some neighborhood of the zero section, where  $r_0$  is a homogeneous symbol of order 0, an even function of  $\eta$ . Note that

$$|r|^2 = \pi^{-1} |\eta| r_0^2. \quad (9-11)$$

The appearance of the Heaviside function here can be explained by the fact that  $N^*\Sigma$  has two connected components: near  $(p_0, q_0, -\xi_0, \eta_0)$  and near  $(p_0, q_0, \xi_0, -\eta_0)$ ; the constants need to be chosen differently in each component.

We start with computing the composition  $FR$ .

Since the kernel of  $FR$  is the transpose of that of  $RF'$ , we will compute the latter; and we only need those singularities that belong to  ${}^oW$ . Denote by  $F(p, q)$  the Schwartz kernel of  $F$ . Then the kernel  $F'(q, p) = F(p, q)$  of  $F'$  (with the notation  $F'f(q) = \int F'(q, p)f(p) d\text{Vol}(p)$ ) can be written as  $F'(q(x, y), p(x))$ , which with some abuse of notation we denote again by  $F'(y, x)$ . Then

$$F'(y, x) := (2\pi)^{-1} \int e^{iy^n \eta_n} \tilde{F}'(y', \eta_n, x) d\eta_n, \tag{9-12}$$

where  $\tilde{F}'$  is the partial Fourier transform of  $F$  with respect to  $y^n$ , and there is no summation in  $y^n \eta_n$ . By Theorem 6.1 and (7-1),

$$\tilde{F}'(y', \eta_n, x) = \pi^{1/2} e^{-i\pi/4} (h(\eta_n) + ih(-\eta_n)) |\eta_n|^{-1/2} G(x, y', \eta_n), \tag{9-13}$$

where  $G$  is a symbol with respect to  $\eta_n$ , smoothly depending on  $(x, y')$  with principal part

$$G_0 := W_\Sigma \frac{\sqrt{2}}{\sqrt{AD}}.$$

Moreover, by Theorem 6.1,  $G$  has an expansion in terms of positive powers of  $|\eta_n|^{-1/2}$ . In particular,  $G - G_0$  is an amplitude of order  $-1/2$  that contributes a conormal distribution in the class

$$I^{-n/2-1/2}(M \times M, \Sigma; \mathbb{C});$$

see e.g., [Hörmander 1985a, Theorem 18.2.8]. By the calculus of conormal singularities, e.g., [ibid., Theorem 18.2.12], the kernel of  $FR$  is of conormal type at  $y^n = 0$  as well, with a principal symbol given by that of  $F$  multiplied by  $r|_{y^n=0, \eta'=0}$ . That principal symbol coincides with the full one modulo conormal kernels of order 1 less than the former; see the expansions in [ibid.] preceding Theorem 18.2.12. Since we assumed that  $r_0$  is an even homogeneous function of  $\eta$  of order 0,  $r_0(y', 0, 0, \eta_n)$  is a function of  $y'$  only for  $\eta$  in a conic neighborhood of  $(0, \pm 1)$ , equal to  $r(y, 0, 0, 1)$ . Therefore, the principal part of  $r(y, D_y)F'(\cdot, x)$  is

$$(2\pi)^{-1} \int e^{iy^n \eta_n} G_0(x, \eta') r_0(y', 0, 0, 1) d\eta_n = W_\Sigma \frac{\sqrt{2}}{\sqrt{AD}} r_0(y', 0, 0, 1) \delta(y^n), \tag{9-14}$$

and the latter is in  $I^{-n/2+1/2}(M \times M, \Sigma; \mathbb{C})$ . The “error” is determined by the next term of the principal symbol of the composition  $FR$  with  $G$  replaced by  $G_0$ , which is of order 1 lower, and by the contribution of  $G = G_0$ , which is of order  $-1/2$  lower. Since the coordinates  $(y', y^n)$  depend on  $p$  as well,  $r_0(y', 0, 0, 1)$  is actually the restriction of  $r_0$  to  $N^*\Sigma(p)$ . So we proved the following.

**Lemma 9.1.** *Let  $r_0$  be as in (9-10). Then  $FR \in I^{1/2-n/2}(M \times M, \Sigma; \mathbb{C})$ , modulo  $I^{-n/2}(M \times M, \Sigma; \mathbb{C})$ , reduces to the Radon transform*

$$FRf(p) \simeq \int_{\Sigma(p)} af \, dS, \quad \text{where } a := r_0|_{\mathcal{N}^*\Sigma(p)} W_\Sigma \frac{\sqrt{2}}{\sqrt{AD}},$$

where  $dS$  is the Riemannian surface measure on  $\Sigma(p)$  that we previously denoted by  $d\text{Vol}_{\Sigma(p)}$ .

In two dimensions, this is an X-ray type of transform. In higher dimensions, this is a Radon type of transform on the family of codimension one surfaces  $\Sigma(p)$ .

In what follows,  $n = 2$ .

We will compute  $RF^*FR$  next. We have

$$\int FRf \overline{FRh} \, d\text{Vol} \simeq \int_M \int_{\Sigma(p)} (af)(z') \, dS(z') \int_{\Sigma(p)} (\bar{a}\bar{h})(q) \, dS(q) \, d\text{Vol}(p) \quad (9-15)$$

modulo terms of the kind  $(Pf, h)$ , where  $P$  is a  $\Psi$ DO of order  $-3/2$  or less.

In the latter integral,  $p$  parametrizes the curve  $\Sigma(p)$ , while  $q \in \Sigma(p)$  parametrizes a point on it. Another parametrization is by  $p$  and  $\xi \in S_p^*M$  with  $\xi$  oriented positively; then  $q = \exp_p(v)$ , where  $v \in \Sigma(p)$  and  $\xi(v) = 0$ . For the Jacobian of that change we have

$$dS(q) \, d\text{Vol}(p) = D \, d\text{Vol}_{S(p)}(v) \, d\text{Vol}(p) = \frac{|v|D}{\cos \phi} \, d\sigma_p(\xi) \, d\text{Vol}(p), \quad (9-16)$$

and we recall that  $d\sigma_p$  denotes the surface measure on  $S_pM$ , which in this case is a circle. The canonical map  $(p, \xi) \rightarrow (q, \eta)$  is symplectic and therefore preserves the volume form  $dp \, d\xi$ . Set

$$K := |\eta(p, \xi)|/|\xi|. \quad (9-17)$$

Then this map takes  $S^*M$  into  $\{(q, \eta) \in T^*M : |\eta| = K\}$ . Project that bundle to the unit circle one, and set  $\hat{\eta} = \eta/|\eta|$ . Then we have the map  $(p, \xi) \rightarrow (q, \hat{\eta})$ , and  $d\text{Vol}(p) \, d\sigma_p(\xi) = K^2 \, d\text{Vol}(q) \, d\sigma_q(\hat{\eta})$ .

When we perform those changes of variables in (9-15), we will have

$$dS(q) \, d\text{Vol}(p) = \frac{|w|DK^2}{\cos \phi} \, d\text{Vol}(q) \, d\sigma_q(\eta), \quad (9-18)$$

where  $p \in M$ ,  $q \in \Sigma(p)$ ,  $(q, \eta) \in S^*M$ , and we removed the hat over  $\eta$ . Let  $w$  be the corresponding vector in  $S(q)$  normal to  $\eta$ . That parametrizes the curves  $\Sigma(p)$  over which we integrate by initial points  $q$  and unit conormal vectors  $\eta$ . The latter can be replaced by unit tangent vectors  $\hat{w} = w/|w|$ ; then  $d\text{Vol}(q) \, d\sigma_q(\eta) = d\text{Vol}(q) \, d\sigma_q(\hat{w})$ . Let us denote the so parametrized curves by  $c_{q, \hat{w}}(s)$ , where  $s$  is an arc-length parameter.

It remains to notice that the integral with respect to  $z' \in \Sigma(p)$  is an integral with respect to the arc-length measure on  $\Sigma(p)$ , which we denote by  $s$ . Then performing the change of the variables  $(p, q, z') \mapsto (q, \hat{w}, z')$  in (9-15), we get

$$\int FRf \overline{FRh} \, d\text{Vol} \simeq \int_{\mathbb{R} \times S_q M \times M} (af)(c_{q, \hat{w}}(s)) \bar{a}(q, -\hat{w}) \bar{h}(q) \, ds \frac{|w|DK^2}{\cos \phi} \, d\sigma_q(\hat{w}) \, d\text{Vol}(q). \quad (9-19)$$

Therefore, we get as in (5-2), (5-4),

$$\begin{aligned} R^* F^* F R f(q) &\simeq \frac{1}{\sqrt{\det(g(q))}} \int a\bar{a} \frac{|w|DK^2}{\cos \phi} \frac{f(q')}{\rho(q, q')} d\text{Vol}(q') \\ &\simeq \frac{1}{\sqrt{\det(g(q))}} \int |r_0|_{\mathcal{N}^*\Sigma(p)}|^2 |W_\Sigma|^2 \frac{2|w|K^2}{A \cos \phi} \frac{f(q')}{\rho(q, q')} d\text{Vol}(q'). \end{aligned} \quad (9-20)$$

For the directional derivatives of  $\det d \exp_p(v) = -J'/|v|$  (see (8-3)), the derivative along the radial ray is  $|J'(1)|/|v|$  by absolute value, while the derivative in the direction of  $S(p)$  vanishes. That implies

$$A \cos \phi = |J'(1)|/|w| = K/|w|.$$

Therefore,

$$R^* F^* F R f(q) \simeq \frac{1}{\sqrt{\det(g(q))}} \int 2K |r_0|_{\mathcal{N}^*\Sigma(p)}|^2 |W_\Sigma|^2 |w|^2 \frac{f(q')}{\rho(q, q')} d\text{Vol}(q'). \quad (9-21)$$

Here  $(p, v)$  is defined as follows. It is the point in  $SM$  that lies on the continuation of the geodesic through  $q, q'$  to its conjugate point near  $p_0$ , The weight  $\kappa$  restricts  $q'$  to a small neighborhood of  $\gamma_0$ . Next,  $A_2$  restricts  $q'$  near  $q_0$ .

We compare (9-21) with (5-4) and (5-5). Notice that the Jacobian term in (5-4) at the diagonal equals  $\sqrt{\det g}$  and therefore cancels the factor in front of the integral in the calculation of the principal symbol. We therefore proved the following.

**Lemma 9.2.** *Let  $n = 2$ . Then  $R^* F^* F R$  is a  $\Psi DO$  of order  $-1$  with principal symbol modulo  $S^{-3/2}$  at  $(q, \eta)$  near  $(q_0, \eta_0)$  given by*

$$4\pi K |\eta|^{-1} |r_0|_{\mathcal{N}^*\Sigma(p)}|^2 |\kappa(p, v/|v|)|^2 |\kappa(q, -w/|w|)|^2.$$

Here  $w/|w|$  is a continuous choice of a unit vector normal to  $\eta$  at  $q$ , so that  $(q, w/|w|) = (q_0, w_0/|w_0|)$  when  $(q, \eta) = (q_0, \eta_0)$ , and  $v/|v|$  is a parallel transport of  $-w/|w|$  from  $q$  to its conjugate point  $p$  along the geodesic  $\gamma_{q,w}$ .

Later we use the notation  $w = \eta^\perp/|\eta^\perp|$ , and  $v = \xi^\perp/|\xi^\perp|$ .

**Proposition 9.1.** *Let  $n = 2$ . Then*

$$\text{Id} - A_2^{-1} F^* A_1^{-1} F$$

*is a  $\Psi DO$  of order  $-1/2$ .*

*Proof.* We apply Lemma 9.2 with  $\pi^{-1/2} e^{i\pi/4} |\eta|^{1/2} r_0$  being the principal symbol of  $A_2^{-1/2}$  (see (9-10)), where  $A_2^{-1/2}$  is a parametrix of  $A_2^{1/2}$  near  $(q_0, \pm\eta_0)$ . To this end, choose

$$\pi^{-1/2} e^{i\pi/4} (2\pi)^{-1/2} r_0(q, \eta) = (2\pi)^{-1/2} |\kappa(q, \eta^\perp/|\eta^\perp|)|^{-1};$$

see (8-1). Note that  $\kappa(q, w/|w|) = \kappa(p, -v/|v|) = 0$  because of the assumption on  $\text{supp } \kappa$ . Then  $|r_0|_{\mathcal{N}^*\Sigma(p)} = 2^{-1/2} |\kappa(q, -w/|w|)|^{-1}$ , where  $w$  is as in (3-1). The choice of  $r_0$  yields  $RR^* = A_2^{-1/2} \text{ mod}$



$\Psi^{-1}$ . So Lemma 9.2 implies that  $R^*F^*FR$ , and therefore  $RR^*F^*F$  and  $A_2^{-1}F^*F$  have principal symbol

$$\sigma_p(A_2^{-1}F^*F)(q, \eta) = 2\pi K |\kappa(p, \xi^\perp/|\xi^\perp|)|^2/|\eta|$$

We only need to insert  $A_1^{-1}$  between  $F^*$  and  $F$ . By [Hörmander 1985b, Theorem 25.3.5], modulo  $\Psi$ DOs of order 1 lower, the principal symbol of  $A_2^{-1}F^*A_1^{-1}F$  is given by that of  $A_2^{-1}F^*F$  multiplied by the principal symbol  $(2\pi|\kappa(p, v)|^2/|\xi|)^{-1}$  of  $A_1^{-1}$  pushed forward by the canonical map of  $F$ . In other words,

$$\sigma_p(A_2^{-1}F^*A_1^{-1}F)(q, \eta) = \frac{2\pi|\kappa(p, \xi^\perp/|\xi^\perp|)|^2}{|\eta|} K [2\pi|\kappa((p, \xi^\perp/|\xi^\perp|)|^2/|\xi(q, \eta)|)]^{-1} = 1. \quad \square$$

The following lemma is needed below for the proof of Theorem 9.2.

**Lemma 9.3.** *Let  $\kappa_1$  and  $\kappa$  both satisfy the assumptions for  $\kappa$  in the introduction, and let  $\kappa(p_0, \theta_0) \neq 0$ . Let  $\chi \in \Psi^0$  have essential support near  $(p_0, \pm\xi_0) \cup (q_0, \pm\eta_0)$  and Schwartz kernel in  $(U_1 \times U_1) \cup (U_2 \times U_2)$ . Then there exists a zero order classical  $\Psi$ DO  $Q$  with the same support properties such that*

$$QX_\kappa^*X_\kappa\chi = X_{\kappa_1}^*X_\kappa\chi \text{ mod } I^{-3/2}(M \times M, \Delta \cup \mathcal{N}^*\Sigma, \mathbb{C}),$$

where  $\Delta$  is the diagonal. In particular,  $QX_\kappa^*X_\kappa\chi - X_{\kappa_1}^*X_\kappa\chi : H^s \rightarrow H^{s+3/2}$  is bounded for any  $s$ .

*Proof.* We define  $Q = Q_1 + Q_2$  where  $Q_{1,2}$  have Schwartz kernels in  $U_1 \times U_1$  and  $U_2 \times U_2$ , respectively. Following the notational convention in (9-8),  $Q = \text{diag}(Q_1, Q_2)$ .

Then we choose  $Q_1$  to have principal symbol

$$\bar{\kappa}_1(p, \xi^\perp/|\xi^\perp|)/\bar{\kappa}(p, \xi^\perp/|\xi^\perp|) \tag{9-22}$$

in a conic neighborhood of  $(p_0, \pm\xi_0)$  with the same choice of  $\xi^\perp$  as in (8-1). Next, we choose  $Q_2$  with a principal symbol

$$\bar{\kappa}_1(q, \eta^\perp/|\eta^\perp|)/\bar{\kappa}(q, \eta^\perp/|\eta^\perp|) \tag{9-23}$$

in a conic neighborhood of  $(q_0, \pm\eta_0)$ . Then

$$QX_\kappa^*X_\kappa = \begin{pmatrix} Q_1A_1 & Q_1F \\ Q_2F^* & Q_2A_2 \end{pmatrix}.$$

Then (see (8-1))

$$\sigma_p(Q_1A_1) = 2\pi(\bar{\kappa}_1\kappa)(p, \xi^\perp/|\xi^\perp|) \quad \text{and} \quad \sigma_p(Q_2A_2) = 2\pi(\bar{\kappa}_1\kappa)(q, \eta^\perp/|\eta^\perp|).$$

For  $Q_1F$  and  $Q_2F^*$ , we use the arguments used in the proof of Lemma 9.1. A representation of the Schwartz kernel of  $F'$  as a conormal distribution is given by (9-12). The composition  $Q_2F^*$  then is of the same conormal type with a principal symbol equal to the complex conjugate of that of  $F'$  multiplied by the symbol (9-23) restricted to  $\mathcal{N}^*\Sigma$ . This replaces  $\kappa^\sharp = \bar{\kappa}$  in (6-6) by  $\bar{\kappa}_1$ . Since  $\kappa^\sharp = \bar{\kappa}$  in (6-6), we get that  $Q_2F^*$  is of the same conormal type with leading singularity as in Theorem 6.1, with

$$W_\Sigma = |v|^{-1}\bar{\kappa}(p, v/|v|)\kappa_1(q, -w/|w|).$$

This is however the leading singularity of  $\chi_2 X_{\kappa_1}^* X_{\kappa} \chi_1$ .

The proof for  $Q_1 F$  is the same with the roles of  $p$  and  $q$  replaced. □

**9d. Cancellation of singularities on Riemannian surfaces.** Assume in all dimensions that there are no conjugate points on the geodesics in  $M$ , and that  $\partial M$  is strictly convex. Let  $M_1 \supset M$  be an extension of  $M$  such that the interior of  $M_1$  contains  $M$  be as in Remark 5.2. Then if  $\kappa \neq 0$ ,

$$\|f\|_{L^2(M)} \leq C \|X^* X f\|_{H^1(M_1)} + C_k \|f\|_{H^{-k}(M)} \quad \text{for all } f \in L^2(M), \tag{9-24}$$

for all  $k \geq 0$ ; see [Stefanov and Uhlmann 2004; Frigiyik et al. 2008], and [Stefanov and Uhlmann 2008] for a class of manifolds with conjugate points. When we know that  $X$  is injective, for example when the weight is constant; then we can remove the  $H^{-k}$  term. The same arguments there show that for any  $s \geq 0$ ,

$$\|f\|_{H^s(M)} \leq C \|X^* X f\|_{H^{s+1}(M_1)} + C_k \|f\|_{H^{-k}(M)} \quad \text{for all } f \in H_0^s(M). \tag{9-25}$$

Consider  $Xf$  parametrized by points in  $\partial_+ S M_1$ , which defines Sobolev spaces for  $Xf$  as in Section 5b. Then

$$\|f\|_{H^s(M)} \leq C \|Xf\|_{H^{s+1/2}(\partial_+ S M_1)} + C_k \|f\|_{H^{-k}(M)} \quad \text{for all } f \in H_0^s(M) \text{ and } s \geq 0. \tag{9-26}$$

Indeed, in Proposition 5.2, one can complete  $M_1$  and  $\mathcal{H}$  to closed manifolds, and then we would get that  $X^* : H^s \rightarrow H^{s+1/2}$  is bounded. Then (9-26) follows by (9-25). Estimate (9-26) is sharp in view of Proposition 5.2. In the following theorem, we show that (9-24) and (9-26) fail in the 2D case, with a loss at least of one derivative in the first one, and 1/2 derivative in the second.

**Theorem 9.2.** *Let  $n = 2$ , and let  $\gamma_0$  be a geodesic of  $g$  with conjugate points satisfying the assumptions in Section 2. Then for each  $f_2 \in H^s(M)$ , with  $s \geq 0$ , with  $\text{WF}(f_2)$  in a small neighborhood of  $(q_0, \pm\eta^0)$ , there exists  $f_1 \in H^s(M)$  with  $\text{WF}(f_1)$  in a some neighborhood of  $(p_0, \pm\xi^0)$  such that*

$$Xf \in H^{s+3/4} \quad \text{and} \quad X^* X f \in H^{s+3/2}, \quad \text{where } f := f_1 + f_2.$$

*In particular, if  $(M, g)$  is a nontrapping Riemannian surface with boundary with fold type of conjugate points on some geodesics, neither of the inequalities (9-24) and (9-26) can hold.*

**Remark 9.1.** It is an open problem whether we can replace  $H^{s+3/4}$  and  $H^{s+3/2}$  above with  $C^\infty$ . See Section 10a for an example where this can be done.

**Remark 9.2.** If there are no conjugate points, one has  $Xf \in H^{s+1/2}$ ,  $X^* X f \in H^{s+1}$ . Therefore, the conjugate points are responsible for a 1/4 derivative smoothing for  $Xf$ , and a 1/2 derivative smoothing for  $X^* X f$

*Proof.* Let  $f_2$  be as in the theorem. Set

$$f_1 = -A_1^{-1} F f_2,$$

where, as before,  $A_1^{-1}$  and  $A_2^{-1}$  are parametrices of  $A_{1,2}$  in conic neighborhoods of  $(p_0, \pm\xi_0)$  and  $(q_0, \pm\eta_0)$ , respectively. Then  $f_1$  belongs to  $H^s$  and has a wave front set in small neighborhood of

$(p_0 \pm, \xi_0)$ , by Theorem 2.1. By construction and by (9-4),

$$\chi_1 X^* X f \in C^\infty. \quad (9-27)$$

Next, by (9-27),

$$A_2 f_2 + F^* f_1 = A_2 f_2 - F^* A_1^{-1} F f_2 = (A_2 - F^* A_1^{-1} F) f_2.$$

The operator in the parentheses is a  $\Psi$ DO of order  $-3/2$  by Proposition 9.1. Therefore (see (9-5))

$$\chi_2 X^* X f = A_2 f_2 + F^* f_1 \in H^{s+3/2}.$$

We therefore get  $X^* X f \in H^{s+3/2}(U_1 \cup U_2)$ .

To prove  $X f \in H^{s+3/4}$ , note first that above we actually proved that

$$X^* X (\text{Id} - A_1^{-1} F) \chi : H^s(U_2) \rightarrow H^{s+3/2}(U_1 \cup U_2) \quad (9-28)$$

is bounded, being a  $\Psi$ DO of order  $-3/2$ , where  $\chi$  denotes a zero order  $\Psi$ DO with essential support in a small neighborhood of  $(p_0, \pm \eta_0)$  and Schwartz kernel supported in  $U_2 \times U_2$ .

Our goal is to show that

$$X (\text{Id} - A_1^{-1} F) \chi : H^s(U_2) \rightarrow H_0^{s+3/4}(\mathcal{H})$$

is bounded. It is enough to prove that

$$\chi^* (\text{Id} - A_1^{-1} F)^* X^* P_{2s+3/2} X (\text{Id} - A_1^{-1} F) \chi : H^s(U_2) \rightarrow H^{-s}(U_2) \quad (9-29)$$

for any  $\Psi$ DO  $P_{2s+3/2}$  of order  $2s + 3/2$  on  $\mathcal{H}$ . All adjoints here are in the corresponding  $L^2$  spaces. By (9-28),

$$Q_{2s+3/2} X^* X (\text{Id} - A_1^{-1} F) \chi : H^s(U_2) \rightarrow H^{-s}(U_2) \quad (9-30)$$

is bounded for any  $\Psi$ DO  $Q_{2s+3/2}$  of order  $2s + 3/2$ .

To deduce (9-29) from (9-30), it is enough to “commute”  $X^*$  with  $P_{2s+3/2}$  in (9-29). Let  $2s + 3/2$  be a nonnegative integer first. As in the proof of Proposition 5.2, we use the fact that  $X^* P_{2s+3/2} = (P_{2s+3/2}^* X)^*$ , and  $P_{2s+2}^* X f$  is a finite sum of X-ray transforms with various weights of derivatives of  $f$  of order not exceeding  $2s + 2$ . Thus we can write

$$X^* P_{2s+2} = \sum \tilde{Q}_j X_j^*, \quad (9-31)$$

where  $Q_j$  are differential operators on  $\mathcal{H}$  of degree  $2s + 3/2$  or less, and  $X_j$  are like  $X$  in (2-1) but with different weights still supported where  $\kappa$  is supported. By Lemma 9.3,  $\tilde{Q}_j X_j^* X = R_j X^* X$ , where  $R_j$  is a  $\Psi$ DO of the same order as  $\tilde{Q}_j$ . The proof of (9-29) is then completed by the observation that  $\chi^* (\text{Id} - A_1^{-1} F)^*$  maps continuously  $H^s$  into itself, since the canonical relation of  $F$  is canonical graph.  $\square$

### 10. Examples

In this section, we present a few examples. We start in Section 10a with the fixed radius circular transform in the plane, where we can have cancellation of singularities similarly to Theorem 9.2 but we show that this happens to any order. Then we consider in Section 10b the geodesic X-ray transform on the sphere, where the conjugacy is not of fold type, but a similar result holds. Next, in Section 10c, we study an example of magnetic geodesics in the Euclidean space  $\mathbb{R}^3$  with a constant magnetic field. We show that then the canonical relation of  $F$  is a canonical graph, and therefore, one can resolve the singularities. Finally, in Section 10d, we present an example of a Riemannian manifold of product type where the graph condition is violated.

**10a. The fixed radius circular transform in the plane.** Let  $R$  be the integral transform in  $\mathbb{R}^2$  of integrating functions over circles of radius 1. We fix the negative orientation on those circles; then for each  $(x, \xi) \in S\mathbb{R}^2$ , there is a unique unit circle passing through  $x$  in the direction of  $\theta$ . It is very easy to see (below) that the first conjugate point appears at “time”  $\pi$ . The next one is at  $2\pi$ , which equals the period of the curve. If one originally chooses  $f$  supported near, say  $(0, 0)$  and  $(2, 0)$ ; and chooses  $\gamma_0$  to be the arc of the circle that is a small extension of  $\{|x_1 - 1|^2 + x_2^2 = 1, x_2 \geq 0\}$ , then we are in the situation studied above. On the other hand, if we do not impose any assumptions on  $\text{supp } f$ , we will get contributions that are smoothing operators only. Therefore, we do not need to restrict  $\text{supp } f$ .

Those circles are also magnetic geodesics with respect to the Euclidean metric and a constant nonzero magnetic field; see e.g., [Dairbekov et al. 2007]. Let us use the following parametrization first. We temporarily denote vectors  $\theta$  by  $\vec{\theta} := (\sin \theta, \cos \theta)$  to reserve  $\theta$  for their (nonstandard) polar angles. The circle through  $x$  in the direction of  $\vec{\theta}$  is given by

$$\gamma_{x,\theta}(t) = x + (\cos \theta - \cos(\theta + t), -\sin \theta + \sin(\theta + t)). \tag{10-1}$$

Then  $\gamma_{x,\theta}(0) = x, \dot{\gamma}_{x,\theta}(0) = \vec{\theta}$ . Let  $J_1$  be the Jacobi matrix  $\partial\gamma_{x,\theta}(t)/\partial(t, \theta)$ . We have

$$J_1 = \begin{pmatrix} \sin(\theta + t) & -\sin \theta + \sin(\theta + t) \\ \cos(\theta + t) & -\cos \theta + \cos(\theta + t) \end{pmatrix}. \tag{10-2}$$

Then  $\det J_1 = -\sin(\theta + t) \cos \theta + \sin \theta \cos(\theta + t) = -\sin t$ . It vanishes when  $t = \pi$  (see the remarks above why the other zeros do not matter). Therefore, in the  $(t, \theta)$  coordinates, the tangent conjugate locus  $S(x)$  is given by  $\{t = \pi\}$  for any  $x$ . The conjugate locus of  $x$  then is the circle  $\Sigma(x) = \{\gamma_{x,\theta}(\pi)\} = \{x + 2(\cos \theta, -\sin \theta); \theta \in \mathbb{R}\}$ , that is,

$$\Sigma(x) = \{y : |y - x| = 2\},$$

which is the envelope of all circles of radius 1 passing through  $x$ ; see Figure 5. Next,

$$J_1|_{t=\pi} = \begin{pmatrix} -\sin \theta & -2 \sin \theta \\ -\cos \theta & -2 \cos \theta \end{pmatrix}. \tag{10-3}$$

The null space consist of multiples of  $2\partial/\partial t - \partial/\partial \theta$ . That null space is transversal to  $\{t = \pi\}$ ; therefore, we have a fold conjugate locus.

To write this in the Cartesian coordinates  $x = (x^1, x^2)$ , set

$$v = t(\sin \theta, \cos \theta),$$

that is,  $v = t\vec{\theta}$ . Set also  $\exp_x(v) = \gamma_{x,\theta}(t)$ , that is, the endpoint of the magnetic geodesic originating at  $x$  in the direction  $v/|v|$ , of length  $|v|$ . Then

$$S(x) = \{v : |v| = \pi\}.$$

We compute next  $N_x(v)$  for  $v = (0, \pi)$ . By the rotational symmetry, this would determine  $N_x(v)$  for any  $v \in S_x(v)$  in a trivial way. For the Jacobi matrix  $J_2 := \partial v / \partial(t, \theta)$  we get

$$J_2 = \begin{pmatrix} \sin \theta & t \cos \theta \\ \cos \theta & -t \sin \theta \end{pmatrix}. \quad (10-4)$$

To find the Jacobi matrix  $J := \partial \exp_x(v) / \partial v = \partial \gamma_{x,\theta}(t) / \partial v$  at  $v = (0, \pi)$ , we write  $J = J_1 J_2^{-1}$  at  $\theta = 0$ ,  $t = \pi$ , to get

$$J|_{v=(0,\pi)} = \begin{pmatrix} 0 & 0 \\ -2/\pi & -1 \end{pmatrix}. \quad (10-5)$$

The null space is spanned by  $(-1, 2/\pi)$ . For general  $\theta$  it follows immediately that

$$N_x(v) = \mathbb{R}e^{-i\theta}(-1, 2/\pi),$$

where we used complex identification to denote rotation by the angle  $-\theta$ . We could have obtained this as  $J = J_1 J_2^{-1}$  for  $t = \pi$ , and general  $\theta$ 's, of course. In particular, for  $\theta = 0$ , that is, for  $v = (0, \pi)$ , we get  $N_x(v) = \mathbb{R}(-1/2, \pi)$ . We see again that  $S$  is a fold conjugate locus. The other assumptions of the dynamical system are easy to check.

It is much more natural to parametrize those circles by their centers; we use the notation  $C(x)$ . Then the circular integral transform is defined by

$$Xf(y) = \int_{C(y)} f \, d\ell = \int_{|\omega|=1} f(y + \omega) \, d\ell_\omega = \int_0^{2\pi} f(z + e^{i\alpha}) \, d\alpha. \quad (10-6)$$

The connection to the natural parametrization by  $x$  and  $\theta$  that we used above is as follows. As in [Dairbekov et al. 2007], for all circles in neighborhood of a given one, for example the one with  $x = 0$  and  $\theta = 0$ , we choose a curve  $S$  through  $x = 0$ , transversal to that circle. Let  $z$  be the point of intersection of those circles with  $S$ , close to 0. Then we use  $z$  and  $\theta$  as parameters, and the natural measure is  $d\mu = |\theta \cdot \nu(z)| \, d\ell_z \, d\theta$ , where  $d\ell_z$  is the Euclidean length measure on  $S$ , and  $\nu(z)$  is the unit normal at  $z$ . This measure is independent of the choice of  $S$ . Choose  $S = \{x^2 = 0\}$ . Then the natural measure on those circles is  $d\mu = \cos \theta \, dz^1 \, d\theta$ , near  $z^1 = 0$  and  $\theta = 0$ . The center of each such circle is given by  $y := (z^1 + \cos \theta, -\sin \theta)$ ; see (10-1). Using  $y$  as a new parameter, and computing the Jacobian of the map  $(z^1, \theta) \mapsto y$ , we see that  $d\mu = dy$  in the new variables. Therefore, with the parametrization by its center as in (10-6),  $X$  is unitarily equivalent to its previous definition, and  $X^*X$  will not change if we define  $X^*$  with respect to the inner product  $L^2(\mathbb{R}^2, dy)$ .

**10a1.** *X as a convolution.* It is well known and easy to see that  $X$  is a convolution with the delta function  $\delta_{S^1}$  of the unit circle

$$Xf = \delta_{S^1} * f.$$

Fourier transforming, we get

$$X = 2\pi \mathcal{F}^{-1} J_0(|\xi|) \mathcal{F}, \tag{10-7}$$

where  $J_0$  is the Bessel function of order 0. This shows that

$$X^* X = (2\pi)^2 \mathcal{F}^{-1} J_0^2(|\xi|) \mathcal{F}. \tag{10-8}$$

Note that  $J_0^2(|\xi|)$  is not a symbol because it oscillates. In principle, one can use this representation to analyze  $X^* X$  but this is not so convenient when we want to analyze  $X$  locally.

**10a2.** *Integral representation.* We write

$$\begin{aligned} (Xf, Xh) &= \int \int_{|\omega|=1} f(x + \omega) d\ell_\omega \int_{|\theta|=1} \bar{h}(x + \theta) d\ell_\theta dx \\ &= \int \int_{|\omega|=1} \int_{|\theta|=1} f(y + \omega - \theta) \bar{h}(y) d\ell_\omega d\ell_\theta dy. \end{aligned} \tag{10-9}$$

Therefore,

$$X^* Xf(x) = \int_{|\omega|=1} \int_{|\theta|=1} f(x + \omega + \theta) d\ell_\omega d\ell_\theta; \tag{10-10}$$

compare with (5-1).

We will make the change of variables  $z = \omega + \theta$ . For  $0 < |z| < 2$ , there are exactly two ways  $z$  can be represented this way. Write  $\omega = e^{i\alpha}$  and  $\theta = e^{i\beta}$ . Since  $d\ell_\omega = d\alpha$ ,  $d\ell_\theta = d\beta$ , and  $dz_1 \wedge dz_2 = (-2i)^{-1} dz \wedge d\bar{z}$ , we get

$$\begin{aligned} dz_1 \wedge dz_2 &= \frac{1}{-2i} (ie^{i\alpha} d\alpha + ie^{i\beta} d\beta) \wedge (-ie^{-i\alpha} d\alpha - ie^{-i\beta} d\beta) = \sin(\beta - \alpha) d\alpha \wedge d\beta \\ &= \sin(\beta - \alpha) d\ell_\omega \wedge d\ell_\theta. \end{aligned}$$

It is easy to see that  $|\beta - \alpha|$  equals twice the angle between  $z = \omega + \theta$  and  $\theta$ . Let  $r = |z|$ . Then  $r/2 = \cos \frac{1}{2}|\alpha - \beta|$ . Elementary calculations then lead to

$$\sin|\alpha - \beta| = \frac{r}{2} \sqrt{4 - r^2}.$$

Therefore, (10-10) yields the following.

**Proposition 10.1.** *Let  $X$  be the circular transform defined above. Then*

$$X^* Xf(x) = \int_{r < 2} \frac{4}{r\sqrt{4-r^2}} f(y) dy, \quad \text{where } r := |x - y|. \tag{10-11}$$

**10a3.**  $X^*X$  as an FIO. The kernel has singularities near the diagonal  $x = y$ , and also near

$$\Sigma = \{|x - y| = 2\}.$$

That singularity is of the type  $(2 - |x - y|)^{-1/2}$ , and for a fixed  $x$  the expression  $2 - |x - y|$  measures the distance from the circle  $\Sigma(x)$  to the point  $y$  inside that circle. We therefore get the same singularity as in Theorem 6.1. Note also that

$$\mathcal{N}^*\Sigma = \{(x, x \pm 2\xi/|\xi|, \xi, -\xi) : \xi \in \mathbb{R}^2 \setminus 0\}. \tag{10-12}$$

Based on Proposition 10.1, and Theorem 2.1, we conclude that  $X^*X$  is an FIO of order  $-1$  with a canonical relation  $\mathcal{C}$  of the following type. We have  $(x, \xi, y, \eta) \in \mathcal{C}$  if and only if  $(y, \eta) = (x, \xi)$  (that gives us the  $\Psi$ DO part), or  $(y, \eta) = (x \pm 2\xi/|\xi|, \xi)$ .

This can also be formulated also in the following form.

**Theorem 10.1.** *Let  $X$  be the circular transform defined above. Then, modulo  $\Psi^{-\infty}$ ,*

$$X^*X = A_0 + F_+ + F_-, \tag{10-13}$$

where  $A_0, F_+$  and  $F_-$  are Fourier multipliers with the properties

- (a)  $A_0 = 4\pi|D|^{-1} \text{ mod } \Psi^{-2}$ ;
- (b)  $F_{\pm}$  are elliptic FIOs of order  $-1$  with canonical relations of a graph type given by

$$\mathcal{F}_{\pm} : (x, \xi) \mapsto (x \pm 2\xi/|\xi|, \xi); \tag{10-14}$$

- (c)  $F_- = F_+^*$ .

*Proof.* We start with the Fourier multiplier representation (10-7). The leading term of  $(2\pi)^2 J_0^2(|\xi|)$  is

$$\frac{8\pi}{|\xi|} \cos^2(|\xi| - \pi/4) = \frac{8\pi}{|\xi|} (1 + \sin(2|\xi|)) = 2\pi \left( \frac{2}{|\xi|} + \frac{e^{2i|\xi|}}{i|\xi|} - \frac{e^{-2i|\xi|}}{i|\xi|} \right). \tag{10-15}$$

Those three terms are the principal parts of the operators in (10-13). The first one gives  $4\pi|D|^{-1}$ , while the second and the third one are FIOs with phase functions  $\phi_{\pm} = (x - y) \cdot \xi \pm 2|\xi|$ . A direct calculation show that the canonical relations of  $F_{\pm}$  are given by (10-14), indeed. For the complete proof of the theorem, we need the full asymptotic expansion of  $J_0$ .

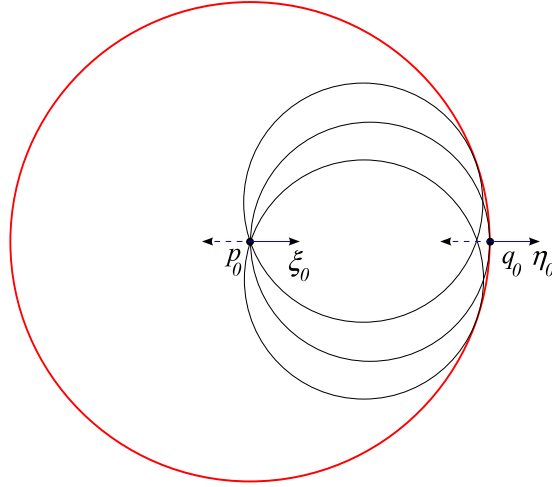
We recall the well-known expansion of  $J_0(z)$  for  $z \rightarrow \infty$ :

$$J_0(z) \sim \sqrt{2/(\pi z)} (P(z) \cos(z - \pi/4) - Q(z) \sin(z - \pi/4)),$$

where

$$P(z) \sim \sum_{k=0}^{\infty} p_k z^{-2k} \quad \text{and} \quad Q(z) \sim \sum_{k=0}^{\infty} q_k z^{-2k-1},$$





**Figure 5**

with some (explicit) coefficients  $p_k$  and  $q_k$ . In particular,  $p_1 = 1$  and  $q_1 = -1/8$ . Then

$$\begin{aligned} (2\pi)^2 J_0^2(z) &\sim \frac{2\pi}{z} \left( (P + iQ)e^{i(z-\pi/4)} + (P - iQ)e^{-i(z-\pi/4)} \right)^2 \\ &\sim \frac{2\pi}{z} \left( -i(P + iQ)^2 e^{2iz} + i(P - iQ)^2 e^{-2iz} + 2P^2 + 2Q^2 \right). \end{aligned}$$

We set

$$A_0 = 4\pi |D|^{-1} (P^2(|D|) + Q^2(|D|)) \quad \text{and} \quad F_{\pm} = \mp 2\pi i |D|^{-1} (P(|D|) \pm iQ(|D|))^2 e^{\pm 2i|D|}. \quad (10-16)$$

This completes the proof. □

We will now connect this to Theorem 2.1. Let  $p_0 = (0, 0)$ ,  $q_0 = (2, 0)$ ,  $v_0 = (0, \pi)$ ,  $w_0 = (0, \pi)$ . Then  $v_0 \in S(p_0)$ . Choose  $\xi_0 = (1, 0)$ , conormal to the conjugate locus  $\Sigma(q_0) = \{|x - q_0| = 2\}$  at  $p_0$ ; choose  $\eta_0 = (1, 0)$ , conormal to the conjugate locus  $\Sigma(p_0) = \{|x - p_0| = 2\}$  at  $q_0$ . The directions of  $\xi_0$  and  $\eta_0$  reflect the choice of the orientation we made earlier. We refer to Figure 5.

If we localize  $X$  near  $v = v_0$ , then the pseudodifferential part of  $X^* \chi X$  is  $(1/2)A_0$ , see (5-10). Therefore, in the notation of Theorem 2.1,

$$A = \frac{1}{2}A_0 \quad \text{and} \quad F = F_+ + F_-.$$

The canonical relation of  $F_+$  maps  $(p_0, \xi_0)$  into  $(q_0, \eta_0)$ , (see Figure 5), while that of  $F_-$  maps  $(p_0, -\xi_0)$  into  $(q_0, -\eta_0)$ . This is consistent with the results in Theorem 2.1, where the Lagrangian has two disconnected components located near  $(p_0, q_0, \pm\xi_0, \mp\eta_0)$ .

To analyze the operator (9-9), note first that  $A_1 = A_2 = A_0/2$ . Let us first analyze this operator applied to distributions with wave front set near  $(q_0, \eta_0)$  but not near  $(q_0, -\eta_0)$ . Then  $F$  reduces to  $F_+$  only, and we have, modulo  $\Psi^{-\infty}$ ,

$$A_2^{-1} F^* A_1^{-1} F = \frac{1}{4} A^{-2} F_+^* F_+ = \text{Id};$$

see (10-16). The analysis near  $(q_0, -\eta_0)$  is similar. Therefore, we have a stronger version of Theorem 9.2 in this case: Singularities can cancel to any order.

**Theorem 10.2.** *Let  $f_1$  be any distribution with  $\text{WF}(f_1)$  supported in a small conic neighborhood of some  $(x_0, \xi^0) \in T^*\mathbb{R}^2 \setminus 0$ . Then there exists a distribution  $f_2$  with  $\text{WF}(f_2)$  supported in a small conic neighborhood of  $(x_0 \pm 2\xi^0/|\xi^0|, \xi^0)$ , which is an image of  $\text{WF}(f_1)$  under the map  $\mathcal{F}_\pm$ , such that  $X(f_1 + f_2) \in C^\infty$  for all unit circles in a neighborhood of the unit circle  $C(x_0 \pm \xi^0)$ .*

In other words, for a fixed circle  $C_0$  of radius 1, there is a rich set of distributions  $f$ , with any order of singularity at  $\mathcal{N}^*C_0$ , such that those singularities are invisible by  $X$  localized near  $C_0$ , that is,  $Xf \in C^\infty$ . Explicit examples can be constructed by choosing  $f_2(x) = \delta(x - q_0)$ . Then  $Ff_2$  near  $p_0$  is just given by the Schwartz kernel of  $X^*X$ ; see (10-11). To obtain  $f_1$ , we apply  $2A_0^{-1}$  to the result.

We emphasize that the theorem provides an example of cancellation of singularities for the localized transform only. As we will see below,  $Xf \in C^\infty$  (globally) for  $f \in \mathcal{E}'$  implies  $f \in C^\infty$ . On the other hand, without the compact support assumption, one can construct singular distributions in the kernel of  $X$ , using the Fourier transform.

**10a4.** *The wave front set of a distribution in  $\text{Ker } X$ .* Now, if  $Xf = 0$  or, more generally, if  $Xf \in C^\infty$ , one easily gets that

$$\text{For all } f \in \text{Ker } X, \text{ WF}(f) \text{ is invariant under the action of the group } \{\mathcal{F}_\pm^m \text{ for } m \in \mathbf{Z}\}. \quad (10-17)$$

Then, if  $f$  is compactly supported (or more generally, smooth outside some compact set), we get that  $\text{WF}(f)$  must be empty, that is,  $f \in C^\infty(\mathbb{R}^2)$ . In other words, even though recovery of  $\text{WF}(f)$  is impossible by knowing  $Xf$  locally, as we saw above, the condition  $Xf \in C^\infty$  globally, together with the compact support assumption, yielded a global recovery of singularities. Here an important role is played by the fact that  $X$  is translation invariant, and in particular, our assumptions are valid for any  $(p_0, \theta_0) \in TS\mathbb{R}^2$  that cannot be guaranteed in the general case. Also, the dynamics is not time reversible; therefore for each  $(x_0, \xi^0) \in T^*M \setminus 0$  there are two different curves through  $x_0$  in our family. The latter is true for general magnetic systems with a nonzero magnetic field; see [Dairbekov et al. 2007].

**Remark 10.1.** One can see that  $X$  is invertible on  $L^2(M)$  by using Fourier transform; see (10-7). The formal inverse is  $1/J_0(|\xi|)$ , and conjugating a compactly supported  $\chi$  with the Fourier transform, one gets a convolution in the  $\xi$  variable that will smoothen out the zeros of  $J_0(|\xi|)$ , thus producing a Fourier multiplier with asymptotic  $\sim |\xi|^{1/2}$ . However, in  $L^p(\mathbb{R}^2)$  with  $p > 4$  it is not invertible, and elements of the kernel include functions with Fourier transforms supported on the circles  $J_0(|\xi|) = 0$ ; see also [Thangavelu 1994; Agranovsky and Kuchment 2011].

Finally, we remark that in this case, one can study  $X$  directly, instead of  $X^*X = X^2$ , with the same methods. Our goal however is to connect the analysis of this transform with our general results.

**10b. The X-ray transform on the sphere.** Consider the geodesic ray transform on the sphere  $S^n$ . The conjugate points are not of fold type; instead they are of blow-down type. Let  $J$  be the antipodal map.

Without going into details, we will just mention that then (2-3) still holds with

$$CN = |D|^{-1} - |D|^{-1}J,$$

with some constant  $C$ , where the canonical relation of  $F$  is the graph of the antipodal map, lifted to  $T^*S^2$ . Then  $CN|D| = \text{Id} - J$ . The canonical graph is an involution, however (its square is identity), so arguments similar to that in the previous example do not apply. That means that singularities may cancel. In fact, it is known that  $X$  has an infinite-dimensional kernel — all odd functions with respect to  $J$ .

In this case  $\Sigma$  consists of all antipodal pairs  $(x, y)$ , and has dimension 2 (and codimension 2), unlike the case above (dimension 3 and codimension 1). On the other hand,  $\mathcal{N}^*\Sigma$  still has the same dimension (that is  $2n = 4$ , and this is always the case as long as  $\Sigma$  is smooth submanifold). One can see that the Lagrangian in this case is still  $\mathcal{N}^*\Sigma$ .

**10c. Magnetic geodesics in  $\mathbb{R}^3$ .** Consider the magnetic geodesic system in the Euclidean space  $\mathbb{R}^3$  with a constant magnetic potential  $(0, 0, \alpha)$ ,  $\alpha > 0$ . The geodesic equation is then given by

$$\ddot{\gamma} = \dot{\gamma} \times (0, 0, \alpha), \quad (10-18)$$

where  $\times$  denotes the vector product in  $\mathbb{R}^3$ . The right hand side above is the Lorentz force, which is always normal to the trajectory and therefore does not affect the speed. We restrict to trajectories on energy level 1, which is preserved under the flow. Then we get

$$\ddot{\gamma}^1 = \alpha \dot{\gamma}^2, \quad \ddot{\gamma}^2 = -\alpha \dot{\gamma}^1, \quad \ddot{\gamma}^3 = 0.$$

The magnetic geodesics are then given by

$$\gamma(t) = \gamma(0) + \left( \frac{r}{\alpha} (\sin(\alpha t + \theta) - \sin \theta), \frac{r}{\alpha} (-\cos(\alpha t + \theta) + \cos \theta), tz \right),$$

where  $(r, \theta, z)$  are the cylindrical coordinates of  $\dot{\gamma}(0)$ . The unit speed requirement means that

$$r^2 + z^2 = 1.$$

The geodesics are then spirals; when  $z = 0$  then they reduce to closed circles, and when  $r = 0$  they are vertical lines.

The parametrization by cylindrical coordinates is singular when  $r = 0$ . Away from that we can use  $\theta$  and  $z$  to parametrize unit speeds. Then in  $\exp_p(v)$ , we use the coordinates  $(t, \theta, z)$  to parametrize  $v$ , that is,

$$v = t(\sqrt{1 - z^2}(\cos \theta, \sin \theta), z).$$

At  $t = 0$  we may have an additional singularity but this is irrelevant for our analysis since we know that the exponential map has an injective differential near  $v = 0$ . An easy computation yields that the conjugate locus is given by the condition  $\alpha t = \pi$ , that is,

$$S_p(v) = \{v : |v| = \pi/\alpha\},$$

and this is true for any  $p \in \mathbb{R}^3$ . This is a sphere in  $T\mathbb{R}^3$ . For  $\Sigma(p)$  we then get

$$\gamma(\pi/\alpha) = p + \alpha^{-1}(-2r \sin \theta, 2r \cos \theta, \pi z) \quad (10-19)$$

with  $p = \gamma(0)$ . This shows that  $\Sigma(p)$  is an ellipsoid

$$\Sigma = \{(p, q) : \frac{1}{4}(q_1 - p_1)^2 + \frac{1}{4}(q_2 - p_2)^2 + \pi^{-2}(q_3 - p_3)^2 = \alpha^{-2}\}.$$

Then

$$\mathcal{N}^*\Sigma = \{(p, q, \xi, \eta) : (p, q) \in \Sigma, \xi = c(p_1 - q_1, p_2 - q_2, 4\pi^{-2}(p_3 - q_3)), \eta = -\xi, 0 \neq c \in \mathbb{R}\}. \quad (10-20)$$

Therefore, given  $p, \xi$ , we can immediately get  $q$  as a smooth function of  $(p, \xi)$ , and we can obtain  $v$  such that  $\exp_p(v) = q$  by (10-19), where the left hand side is  $q$ . Therefore,  $(p, \xi) \mapsto v$  is a smooth map, and therefore so is  $(p, \xi) \mapsto (q, \eta)$ . The later also directly follows from (10-20), since  $\eta = -\xi$ .

We therefore get that  $F$  is an FIO of order  $-3/2$  with a canonical relation

$$(p, \xi) \mapsto (q, \xi), \quad (10-21)$$

where  $q$  can be determined as described above. A geometric description of  $q$  is the following:  $q$  is one of the two points on the ellipsoid  $\Sigma$ , where the normal is given by  $\xi$ . The choice of one out of the two points is determined by the choice of the initial velocity  $v_0$  near which we localize; changing  $v_0$  to  $-v_0$  would alter that choice. Since (10-21) is a diffeomorphism,  $F$  is of canonical graph type and therefore maps  $H^s$  to  $H^{s+3/2}$ . In contrast,  $A_{1,2}$  are elliptic of order  $-1$ ; thus they dominate over  $F$ . By Corollary 9.1,  $X$  can be inverted microlocally in the setup described in Section 2.

**10d. Fold caustics on product manifolds.** Let  $(M, g) = (M', g') \times (M'', g'')$  be a product of two Riemannian manifolds. The geodesics on  $M$  then have the form

$$\gamma_{p,v}(t) = (\gamma'_{p',v'}(t), \gamma''_{p'',v''}(t)).$$

Consequently,

$$\exp_p(v) = (\exp'_{p'}(v'), \exp''_{p''}(v'')).$$

Assume that in  $(M', g')$ ,  $v'_0$  is conjugate at  $p_0$  of fold type, and assume that  $v''_0$  is not conjugate at  $p''_0$  in  $(M'', g'')$ . Then

$$d \exp_p(v) = \text{diag}(d \exp'_{p'}(v'), d \exp''_{p''}(v'')).$$

The kernel of  $d \exp_p(v)$  then consists of  $N_p(v) = N_{p'}(v') \times 0$ . Next,  $S(p) = S(p') \times T_{p''}M''$ , and  $\Sigma(p) = \Sigma'(p') \times M''$ . Then  $N_p(v_0)$  is transversal to  $S(p)$  at  $v = v_0$ ; therefore  $(v', v'')$  is a fold conjugate vector for  $v' \in S'(p)$  close to  $v_0$  and for any  $v''$ . Then the left projection  $\pi_L$  of the Lagrangian  $\mathcal{N}^*\Sigma$  consists of  $(p, \xi)$  with  $(p', \xi') \in \pi_L(\Sigma')$  and  $\xi'' = 0$ . Thus the rank drops at least by  $n'' = \dim(M'')$ . We get the same conclusion for  $\pi_R(\mathcal{N}^*\Sigma)$ . Therefore,  $\mathcal{N}^*\Sigma$  is not a canonical graph in this case.

Let  $n' = \dim(M') = 2$ . Then the canonical relation in  $(M', g')$  is a canonical graph, and we get that  $\pi_{L,R}(\mathcal{N}^*\Sigma)$  have rank  $2n' + n'' = 4 + n''$  instead of the maximal possible  $2n = 4 + 2n''$ ; that is, the loss is exactly  $n''$ .

Assume now that  $n' = 2$ , that  $n'' = 1$ , and that the metric in  $M$  is given by

$$\sum_{\alpha, \beta=1}^2 g_{\alpha\beta}(x^1, x^2) dx^\alpha dx^\beta + (dx^3)^2.$$

Assume also that in  $M'$ , we have a fold conjugate vector  $v_0 = (0, 1)$  at  $x^1 = x^2 = 0$ . Then all possible conormals to the conjugate loci at  $(0, 0)$  corresponding to small perturbations of  $v_0$  will lie in the plane  $v^3 = 0$ . This is an example where Corollary 9.2 can be applied. We can recover singularities of the kind  $\xi = (\xi_1, \xi_2, \xi_3)$  at  $p_0 = (0, 0, 0)$  with  $\xi_3 \neq 0$  and  $(\xi_1, \xi_2)$  in a conic neighborhood of  $(1, 0)$ . The ones with  $\xi_3 = 0$  are the problematic ones.

### References

- [Agranovsky and Kuchment 2011] M. Agranovsky and P. Kuchment, “The support theorem for the single radius spherical mean transform”, *Mem. Differential Equations Math. Phys.* **52** (2011), 1–16. MR 2883793 Zbl 05984457
- [Bernšteĭn and Gerver 1978] I. N. Bernšteĭn and M. L. Gerver, “A problem of integral geometry for a family of geodesics and an inverse kinematic seismics problem”, *Dokl. Akad. Nauk SSSR* **243**:2 (1978), 302–305. In Russian. MR 80g:53050
- [Croke 1991] C. B. Croke, “Rigidity and the distance between boundary points”, *J. Differential Geom.* **33**:2 (1991), 445–464. MR 92a:53053 Zbl 0729.53043
- [Croke et al. 2000] C. B. Croke, N. S. Dairbekov, and V. A. Sharafutdinov, “Local boundary rigidity of a compact Riemannian manifold with curvature bounded above”, *Trans. Amer. Math. Soc.* **352**:9 (2000), 3937–3956. MR 2000m:53054 Zbl 0958.53027
- [Dairbekov 2006] N. S. Dairbekov, “Integral geometry problem for nontrapping manifolds”, *Inverse Problems* **22**:2 (2006), 431–445. MR 2007i:53084 Zbl 1093.53077
- [Dairbekov et al. 2007] N. S. Dairbekov, G. P. Paternain, P. Stefanov, and G. Uhlmann, “The boundary rigidity problem in the presence of a magnetic field”, *Adv. Math.* **216**:2 (2007), 535–609. MR 2008m:37107 Zbl 1131.53047
- [Faridani et al. 1992a] A. Faridani, E. L. Ritman, and K. T. Smith, “Examples of local tomography”, *SIAM J. Appl. Math.* **52**:4 (1992), 1193–1198. MR 1174054 Zbl 0777.65076
- [Faridani et al. 1992b] A. Faridani, E. L. Ritman, and K. T. Smith, “Local tomography”, *SIAM J. Appl. Math.* **52**:2 (1992), 459–484. MR 93b:92008 Zbl 0758.65081
- [Frigyik et al. 2008] B. Frigyik, P. Stefanov, and G. Uhlmann, “The X-ray transform for a generic family of curves and weights”, *J. Geom. Anal.* **18**:1 (2008), 89–108. MR 2008j:53128 Zbl 1148.53055
- [Guillemin 1985] V. Guillemin, “On some results of Gel’fand in integral geometry”, pp. 149–155 in *Pseudodifferential operators and applications* (Notre Dame, IN, 1984), edited by F. Trèves, Proc. Sympos. Pure Math. **43**, Amer. Math. Soc., Providence, RI, 1985. MR 87d:58137
- [Helgason 1980] S. Helgason, *The Radon transform*, Progress in Mathematics **5**, Birkhäuser, Mass., 1980. MR 83f:43012 Zbl 0453.43011
- [Hörmander 1983] L. Hörmander, *The analysis of linear partial differential operators, I: Distribution theory and Fourier analysis*, Grundlehren der Mathematischen Wissenschaften **256**, Springer, Berlin, 1983. MR 85g:35002a Zbl 0521.35001
- [Hörmander 1985a] L. Hörmander, *The analysis of linear partial differential operators, III: Pseudodifferential operators*, Grundlehren der Mathematischen Wissenschaften **274**, Springer, Berlin, 1985. MR 87d:35002a Zbl 0601.35001
- [Hörmander 1985b] L. Hörmander, *The analysis of linear partial differential operators, IV: Fourier integral operators*, Grundlehren der Mathematischen Wissenschaften **275**, Springer, Berlin, 1985. MR 87d:35002b Zbl 0612.35001
- [Jost 1998] J. Jost, *Riemannian geometry and geometric analysis*, 2nd ed., Springer, Berlin, 1998. MR 99g:53025 Zbl 0997.53500

- [Lang 1995] S. Lang, *Differential and Riemannian manifolds*, 3rd ed., Graduate Texts in Mathematics **160**, Springer, New York, 1995. MR 96d:53001 Zbl 0824.58003
- [Michel 1981/82] R. Michel, “Sur la rigidité imposée par la longueur des géodésiques”, *Invent. Math.* **65**:1 (1981/82), 71–83. MR 83d:58021 Zbl 0471.53030
- [Muhometov 1981] R. G. Muhometov, “On a problem of reconstructing Riemannian metrics”, *Sibirsk. Mat. Zh.* **22**:3 (1981), 119–135, 237. In Russian; translated in *Siberian Math. J.* **22**:3 (1981), 420–433. MR 82m:53071
- [Muhometov and Romanov 1978] R. G. Muhometov and V. G. Romanov, “On the problem of finding an isotropic Riemannian metric in an  $n$ -dimensional space”, *Dokl. Akad. Nauk SSSR* **243**:1 (1978), 41–44. In Russian; translated in *Soviet Math. Dokl.* **19**:6 (1978), 1330–1333. MR 81a:53059
- [Natterer 1986] F. Natterer, *The mathematics of computerized tomography*, B. G. Teubner, Stuttgart, 1986. MR 88m:44008 Zbl 0617.92001
- [Sharafutdinov 1994] V. A. Sharafutdinov, *Integral geometry of tensor fields*, VSP, Utrecht, 1994. MR 97h:53077 Zbl 0883.53004
- [Stefanov 2008] P. Stefanov, “Microlocal approach to tensor tomography and boundary and lens rigidity”, *Serdica Math. J.* **34**:1 (2008), 67–112. MR 2009d:53051 Zbl 1199.53099
- [Stefanov and Uhlmann 2004] P. Stefanov and G. Uhlmann, “Stability estimates for the X-ray transform of tensor fields and boundary rigidity”, *Duke Math. J.* **123**:3 (2004), 445–467. MR 2005h:53130 Zbl 1058.44003
- [Stefanov and Uhlmann 2005] P. Stefanov and G. Uhlmann, “Boundary rigidity and stability for generic simple metrics”, *J. Amer. Math. Soc.* **18**:4 (2005), 975–1003. MR 2006h:53031 Zbl 1079.53061
- [Stefanov and Uhlmann 2008] P. Stefanov and G. Uhlmann, “Integral geometry on tensor fields on a class of non-simple Riemannian manifolds”, *Amer. J. Math.* **130**:1 (2008), 239–268. MR 2009e:53051 Zbl 1151.53033
- [Stefanov and Uhlmann 2009] P. Stefanov and G. Uhlmann, “Local lens rigidity with incomplete data for a class of non-simple Riemannian manifolds”, *J. Differential Geom.* **82**:2 (2009), 383–409. MR 2011d:53081 Zbl 05604775
- [Taylor 1996] M. E. Taylor, *Partial differential equations, I: Basic theory*, Applied Mathematical Sciences **115**, Springer, New York, 1996. MR 98b:35002b Zbl 1206.35002
- [Thangavelu 1994] S. Thangavelu, “Spherical means and CR functions on the Heisenberg group”, *J. Anal. Math.* **63** (1994), 255–286. MR 95c:43008 Zbl 0822.43001
- [Warner 1965] F. W. Warner, “The conjugate locus of a Riemannian manifold”, *Amer. J. Math.* **87** (1965), 575–604. MR 34#8344 Zbl 0129.36002

Received 20 Apr 2010. Revised 16 Feb 2011. Accepted 2 Mar 2011.

PLAMEN STEFANOV: stefanov@math.purdue.edu

Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2607, United States

<http://www.math.purdue.edu/~stefanov/>

GUNTHER UHLMANN: gunther@math.washington.edu

Department of Mathematics, University of Washington, Seattle, WA 98195-4350, United States

and

Department of Mathematics, University of California at Irvine, Irvine, CA 92697, United States

<http://www.math.washington.edu/~gunther/>



# EXISTENCE OF EXTREMALS FOR A FOURIER RESTRICTION INEQUALITY

MICHAEL CHRIST AND SHUANGLIN SHAO

The adjoint Fourier restriction inequality of Tomas and Stein states that the mapping  $f \mapsto \widehat{f\sigma}$  is bounded from  $L^2(\mathbb{S}^2)$  to  $L^4(\mathbb{R}^3)$ . We prove that there exist functions that extremize this inequality, and that any extremizing sequence of nonnegative functions has a subsequence that converges to an extremizer.

1. Introduction	261
2. Outline of the proof and definitions	265
3. Step 2: $\mathbf{S} \geq (3/2)^{1/4}\mathbf{P}$	270
4. Step 3: $\mathbf{S} \geq 2^{1/4}\mathbf{P}$	273
5. Step 4: Symmetrization	274
6. Step 5: Big pieces of caps	276
7. Analytic preliminaries	278
8. Step 6A: A decomposition algorithm	283
9. Step 6B: A geometric property of the decomposition	285
10. Step 6C: Upper bounds for extremizing sequences	286
11. Preliminaries for Step 7	288
12. Step 7: Precompactness after rescaling	290
13. Step 8: Excluding small caps	294
14. Estimation of the cross term $\ F_\nu^0\sigma * F_\nu^\infty\sigma\ _2^2$	296
15. Step 9: Large caps	299
16. Constants are local maxima	301
17. A variational calculation	304
18. Proof of Lemma 6.1	309
Acknowledgement	311
References	311

## 1. Introduction

Let  $\mathbb{S}^2$  denote the unit sphere in  $\mathbb{R}^3$ , equipped with surface measure  $\sigma$ . The adjoint Fourier restriction inequality of Tomas and Stein, for  $\mathbb{S}^2$ , states that there exists  $C < \infty$  such that

$$\|\widehat{f\sigma}\|_{L^4(\mathbb{R}^3)} \leq C\|f\|_{L^2(\mathbb{S}^2,\sigma)} \tag{1-1}$$

Christ was supported in part by NSF grant DMS-0901569. Shao was supported by the National Science Foundation under agreement DMS-0635607.

MSC2000: 42A38.

Keywords: extremals, adjoint Fourier restriction inequality.



for all  $f \in L^2(\mathbb{S}^2)$ . With the Fourier transform defined to be  $\hat{g}(\xi) = \int e^{-ix \cdot \xi} g(x) dx$ , denote by

$$\mathcal{R} = \sup_{0 \neq f \in L^2(\mathbb{S}^2)} \frac{\|\widehat{f\sigma}\|_{L^4(\mathbb{R}^3)}}{\|f\|_{L^2(\mathbb{S}^2, \sigma)}}$$

the optimal constant in the inequality (1-1).

**Definition 1.1.** An extremizing sequence for the inequality (1-1) is a sequence  $\{f_\nu\}$  of functions in  $L^2(\mathbb{S}^2)$  satisfying  $\|f_\nu\|_2 \leq 1$ , such that  $\|\widehat{f_\nu\sigma}\|_{L^4(\mathbb{R}^3)} \rightarrow \mathcal{R}$  as  $\nu \rightarrow \infty$ .

An extremizer for the inequality (1-1) is a function  $f \neq 0$  that satisfies  $\|\widehat{f\sigma}\|_4 = \mathcal{R}\|f\|_2$ .

The main result of this paper is this:

**Theorem 1.2.** *There exists an extremizer in  $L^2(\mathbb{S}^2)$  for the inequality (1-1).*

The inequality dual to (1-1) is  $\|\hat{h}\|_{L^2(\mathbb{S}^2, \sigma)} \leq C\|h\|_{L^{4/3}(\mathbb{R}^3)}$ . If  $f$  extremizes (1-1), then  $\widehat{f\sigma} \cdot |\widehat{f\sigma}|^2$  extremizes the dual inequality.

Our inequality is one of endpoint type. That is, it becomes false if either of the exponents 2, 4 is decreased. An analogue of Theorem 1.2 has more recently been obtained by Fanelli, Vega, and Visciglia [Fanelli et al. 2011], for adjoint restriction inequalities not of endpoint type.

**Definition 1.3.** A sequence of functions in  $L^2(\mathbb{S}^2)$  is precompact if any subsequence has a sub-subsequence that is Cauchy in  $L^2(\mathbb{S}^2)$ .

Nonnegative functions play a special role in our analysis, because

$$\|\widehat{|f|\sigma}\|_4 \geq \|\widehat{f\sigma}\|_4 \quad \text{for all } f \in L^2(\mathbb{S}^2).$$

Therefore if  $\{f_\nu\}$  is an extremizing sequence, so is  $\{|f_\nu|\}$ . Any limit, in the  $L^2$  norm, of an extremizing sequence is of course an extremizer. Thus the following implies Theorem 1.2.

**Theorem 1.4.** *Any extremizing sequence of nonnegative functions in  $L^2(\mathbb{S}^2)$  for the inequality (1-1) is precompact.*

In particular, the set of all nonnegative extremizers is itself compact. We do not know whether nonnegative extremizers are unique modulo rotations of  $\mathbb{S}^2$  and multiplication by constants. They do possess the following symmetry, which will be useful in our analysis.

**Theorem 1.5.** *Every extremizer satisfies  $|f(-x)| = |f(x)|$  for almost every  $x \in \mathbb{S}^2$ .*

Proposition 2.7 below states that more generally, the quantity  $\|\widehat{f\sigma}\|_4$  never decreases under  $L^2$  norm-preserving symmetrization of  $f$  with respect to the map  $x \mapsto -x$ .

For complex-valued extremizers and near extremizers, the situation regarding precompactness of extremizing sequences is different, due to the presence of a noncompact group of symmetries of the inequality. For  $\xi \in \mathbb{C}^3$ , define  $e_\xi(x) = e^{x \cdot \xi}$ . Then  $\|\widehat{f e_{i\xi}\sigma}\|_4 = \|\widehat{f\sigma}\|_4$  for arbitrary  $\xi \in \mathbb{R}^3$ , where  $f \in L^2(\mathbb{S}^2)$ . Consequently complex-valued extremizing sequences need not be precompact. However, we show in a sequel [Christ and Shao 2012] that this simple obstruction is the only one; if  $\{f_\nu\}$  is any complex-valued extremizing sequence, then there exists a sequence  $\{\xi_\nu\} \subset \mathbb{R}^3$  such that  $e^{-ix \cdot \xi_\nu} f_\nu(x)$  is precompact.

The symmetries  $f \mapsto f \cdot e^{ix \cdot \xi}$  merit further discussion. Matters are clearer for the paraboloid  $\mathbb{P}^2 = \{(y_1, y_2, y_3) : y_3 = \frac{1}{2}y_1^2 + \frac{1}{2}y_2^2\}$  than for  $\mathbb{S}^2$ . For  $\mathbb{P}^2$ , the analogues of these unimodular exponentials are quadratic exponentials  $e^{ix \cdot \eta + i\tau|x|^2}$  with  $(\eta, \tau) \in \mathbb{R}^{2+1}$ ; compare with  $\mathbb{S}^2$ , where  $\xi \in \mathbb{R}^3$  also ranges over a three-dimensional space. To see the analogy, consider a small neighborhood of  $(0, 0, 1) \in \mathbb{S}^2$ , equipped with coordinates  $x' \in \mathbb{R}^2$  such that  $x = (x', (1 - |x'|^2)^{1/2})$ . Then for  $\xi = (0, 0, \lambda)$ , we have  $e^{ix \cdot \xi} = \exp(i\lambda(1 - \frac{1}{2}|x'|^2 + O(|x'|^4)))$  for small  $x'$ ; thus for small  $x'$  one has essentially quadratic oscillation. The presence of these symmetries among the extremizers for  $\mathbb{P}^2$  implies that, in the language of concentration compactness theory [Kunze 2003], an extremizer  $f$  can be tight at a scale  $r$ , and  $\hat{f}$  can simultaneously be tight at a scale  $\hat{r}$ , with the product  $r \cdot \hat{r}$  arbitrarily large.

Define

$$\mathbf{S} := \sup_{0 \neq f \in L^2(\mathbb{S}^2, \sigma)} \frac{\|f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)}^{1/2}}{\|f\|_{L^2(\mathbb{S}^2, \sigma)}}.$$

Then  $\mathcal{R} = (2\pi)^{3/4}\mathbf{S}$  by Plancherel’s theorem and the connection between the Fourier transform and convolution.

$\mathbf{S}$  is the supremum of a functional, whose critical points are characterized by the generalized Euler–Lagrange equation

$$(f\sigma * f\sigma * \tilde{f}\sigma)|_{\mathbb{S}^2} = \lambda f \quad \text{almost everywhere on } \mathbb{S}^2, \tag{1-2}$$

where  $\tilde{f}(x) = \overline{f(-x)}$  and  $\lambda$  is a Lagrange multiplier determined by  $f$ . This follows from a routine variational argument; see for instance [Christ and Quilodrán 2010], where more general results of this type are justified. Equation (1-2) will be used in a forthcoming paper [Christ and Shao 2012] to prove that all critical points are infinitely differentiable. By taking the  $L^2(\mathbb{S}^2)$  inner product of both sides with  $f$ , one obtains an alternative characterization of extremizers.

**Proposition 1.6.** *A complex-valued function  $f \in L^2(\mathbb{S}^2)$  is an extremizer if and only if*

$$(f\sigma * f\sigma * \tilde{f}\sigma)|_{\mathbb{S}^2} = \mathbf{S}^4 \|f\|_2^2 f \quad \text{almost everywhere on } \mathbb{S}^2,$$

where  $\tilde{f}(x) = \overline{f(-x)}$ .

Since the numerical value of  $\mathbf{S}$  has not been determined, this equation is not entirely explicit and provides only a negative test for extremizers.

Fundamental questions remain open, among them these:

**Questions 1.7.** Are extremizers unique modulo rotations and multiplication by constants? Are constant functions extremizers?

In this context, it is interesting to observe that constant functions are *local* maxima. Let  $\mathbf{1}$  denote the constant function  $f(x) \equiv 1$ .

**Theorem 1.8.** *There exists  $\delta > 0$  such that whenever  $\|f - \mathbf{1}\|_{L^2(\mathbb{S}^2)} < \delta$ ,*

$$\frac{\|\widehat{f\sigma}\|_4^4}{\|f\|_2^4} \leq \frac{\|\widehat{\sigma}\|_4^4}{\|\mathbf{1}\|_2^4},$$

with equality only if  $f$  is constant.

Let  $\mathbb{P}^2$  be the paraboloid introduced above. Let  $\sigma_P$  be the measure  $d\sigma_P = dx_1 dx_2$  on  $\mathbb{P}^2$ .<sup>1</sup> Then the mapping  $f \mapsto \widehat{f\sigma_P}$  is likewise bounded from  $L^2(\mathbb{P}^2, \sigma_P)$  to  $L^4(\mathbb{R}^3)$ . Denote by  $\mathcal{R}_{\mathbb{P}^2}$  the optimal constant in the inequality

$$\|\widehat{f\sigma_P}\|_{L^4(\mathbb{R}^3)} \leq \mathcal{R}_{\mathbb{P}^2} \|f\|_{L^2(\mathbb{P}^2, \sigma_P)}. \quad (1-3)$$

Foschi [2007] has proved that extremals exist for this inequality, and moreover, that every radial Gaussian  $f(x', x_3) = e^{-c|x'|^2}$  is an extremal, where  $x' = (x_1, x_2)$ , and that  $\mathcal{R}_{\mathbb{P}^2} = 2^{3/4}\pi$ . Alternative proofs were given by Hundertmark and Zharnitsky [2006] and by Bennett, Bez, Carbery, and Hundertmark [Bennett et al. 2009]. The simple relation  $\mathcal{R} \geq \mathcal{R}_{\mathbb{P}^2}$  is of significance for our discussion. This relation follows from examination of a suitable sequence of trial functions  $f_\nu$ , such that  $f_\nu(x)^2$  converges weakly to a Dirac mass on  $\mathbb{S}^2$ , and  $f_\nu$  is approximately a Gaussian in suitably rescaled coordinates, depending on  $\nu$ . It is essential for this comparison that  $\mathbb{P}^2$  has the same curvature at 0 as  $\mathbb{S}^2$ , which explains the factors of  $\frac{1}{2}$  in the definition of  $\mathbb{P}^2$ .

The first author to discuss existence of extremizers for Strichartz/Fourier restriction inequalities was apparently Kunze [2003], who proved the existence of extremizers for the parabola in  $\mathbb{R}^2$ , and showed that (in our notation) any nonnegative extremizing sequence is precompact modulo the action of the natural symmetry group of the inequality. Several papers have subsequently dealt with related problems, in some cases determining all extremizers explicitly [Foschi 2007; Hundertmark and Zharnitsky 2006; Bennett et al. 2009; Carneiro 2009], in other cases merely proving existence [Shao 2009]. A powerful result [Shao 2009] that leads easily to existence of extremizers is the profile decomposition; see [Bégout and Vargas 2007]. Of these works, the one most closely related to ours is that of Kunze. One difficulty that we face is the lack of exact scaling symmetries. In some facets of the analysis this is merely a technical obstacle, but it is bound up with the most essential obstacle, which is the possibility that the optimal constant might be achieved only in a limit where  $|f_\nu|^2$  tends to a Dirac mass, or a sum of two Dirac masses.

Our analysis follows the general concentration compactness framework developed by Lions [1984a; 1984b; 1985a; 1985b]. We have elected to make the exposition self-contained in this respect, not drawing on that theory; to do so would apparently not dramatically shorten the exposition, since most of our labor is lavished on specific issues raised by the character of a particular nonlocal operator.

Existence of extremals for another scale-invariant convolution inequality in which curvature plays an essential role, as it does here, was proved in [Christ 2011a]. There the underlying geometry is more subtle, but the operator analyzed is merely linear, while the analysis of this paper is bilinear. Despite differences in details, that analysis and the method of this paper have much in common. The role of an inequality of Moyua, Vargas, and Vega [Moyua et al. 1999] used here was played in [Christ 2011a] by [Christ 2011b].

---

<sup>1</sup>See [Christ 2011a] for a brief discussion of the naturality of this measure from a geometric perspective.

### 2. Outline of the proof and definitions

The following overview of the proof includes notations, definitions, and statements of intermediate results that are not repeated subsequently, and thus is an integral part of the presentation.

**Step 1.** The first step is quite simple, but in it a critical distinction appears between our problem for  $\mathbb{S}^2$ , and for higher-dimensional spheres. The inequality  $\|\widehat{f\sigma}\|_{L^4(\mathbb{R}^3)} \leq \mathcal{R}\|f\|_{L^2(\mathbb{S}^2, \sigma)}$  is equivalent, by squaring and Plancherel’s theorem, to

$$\|f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)} \leq \mathbf{S}^2\|f\|_{L^2(\mathbb{S}^2)}^2, \tag{2-1}$$

where

$$\mathcal{R} = (2\pi)^{3/4}\mathbf{S}$$

and  $*$  denotes convolution of measures. This has been exploited in [Kunze 2003; Foschi 2007; Hundertmark and Zharnitsky 2006; Bennett et al. 2009]. In higher dimensions, the exponent 4 is replaced by an exponent that is no longer an even integer, and no such equivalence is available.

Now the pointwise inequality  $|f\sigma * f\sigma| \leq |f|\sigma * |f|\sigma$ , the relation  $\widehat{\mu * \nu} = \widehat{\mu}\widehat{\nu}$ , and Plancherel’s theorem imply this:

**Lemma 2.1.** *For any complex-valued function  $f \in L^2(\mathbb{S}^2)$ ,*

$$\|\widehat{f\sigma}\|_{L^4(\mathbb{R}^3)} \leq \|\widehat{|f|\sigma}\|_{L^4(\mathbb{R}^3)}.$$

*Therefore if  $f$  is an extremizer for inequality (1-1), then so is  $|f|$ ; if  $\{f_\nu\}$  is an extremizing sequence, so is  $\{|f_\nu|\}$ .*

This permits us to work with nonnegative functions throughout the analysis. For much of our analysis this makes no difference, but nonnegativity will be useful in Step 7, allowing an elementary approach to a step whose analogue in higher dimensions seems to require more sophisticated techniques.

**Step 2.** A potential obstruction to the existence of extremizers, and certainly to the precompactness of arbitrary extremizing sequences, is the possibility that for an extremizing sequence satisfying  $\|f_\nu\|_2 = 1$ ,  $|f_\nu|^2$  could conceivably converge weakly to a Dirac mass at a point of  $\mathbb{S}^2$ . Straightforward analysis of a sequence  $\{f_\nu\}$  chosen so that  $|f_\nu|^2$  converges in this way, disregarding the question of whether  $\{f_\nu\}$  is extremizing, reveals that  $\mathcal{R} \geq \mathcal{R}_{\mathbb{P}^2}$ ; see Lemma 3.1. Now if  $\mathcal{R}$  were to equal  $\mathcal{R}_{\mathbb{P}^2}$ , any such sequence would be extremizing, yet would not be precompact. Therefore an unavoidable step in our analysis is to demonstrate a strict inequality  $\mathcal{R} > \mathcal{R}_{\mathbb{P}^2}$ .

In fact, as will be explained below, this is true in two distinct ways. The more superficial is this:

**Lemma 2.2.** *Let  $g \in L^2(\mathbb{S}^2)$  be supported in  $\{x \in \mathbb{S}^2 : x_3 > \frac{1}{2}\}$ . Define  $f(x) = 2^{-1/2}g(x) + 2^{-1/2}\overline{g(-x)}$ . Then  $\|f\|_2 = \|g\|_2$ , and*

$$\|f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)} = (3/2)^{1/2}\|g\sigma * g\sigma\|_{L^2(\mathbb{R}^3)}.$$

Define the optimal constant in the corresponding inequality for the paraboloid to be

$$\mathbf{P} = \sup_{0 \neq g \in L^2(\mathbb{P}^2, \sigma_P)} \frac{\|g\sigma_P * g\sigma_P\|_{L^2(\mathbb{R}^3)}^{1/2}}{\|g\|_{L^2(\mathbb{P}^2, \sigma_P)}}.$$

By Lemma 2.2, the optimal constants for  $\mathbb{S}^2$  and  $\mathbb{P}^2$  satisfy the following.

**Corollary 2.3.**  $\mathbf{S} \geq (3/2)^{1/4}\mathbf{P}.$

**Step 3.** The simplest possibility left open by Step 2 is that an extremizing sequence might concentrate at a pair of antipodal points, that is,  $|f_\nu|^2$  might converge weakly to a linear combination of two Dirac masses, at antipodal points  $z$  and  $-z$ . This scenario is indeed the crux of the problem. The crucial ingredient in excluding it is an improved inequality  $\mathbf{S} > (3/2)^{1/4}\mathbf{P}$ . We will give two independent proofs of this inequality. The first gives a precise improvement:

**Lemma 2.4.**  $\mathbf{S} \geq 2^{1/4}\mathbf{P}.$

Equivalently,  $\mathcal{R} \geq 2^{1/4}\mathcal{R}_{\mathbb{P}^2}$ . This is proved by an exact computation of  $\|f\sigma * f\sigma\|_2$  for  $f \equiv 1$ . We do not know whether constant functions are in fact extremal for (1-1), or equivalently, whether  $\mathbf{S} = 2^{1/4}\mathbf{P}$ . Constants are indeed critical points of the associated functional, and thus satisfy a (possibly) modified Euler–Lagrange equation (1-2), in which  $\mathbf{S}$  is replaced by  $2^{1/4}\mathbf{P}$ .

An alternative proof that  $\mathbf{S} > (3/2)^{1/4}\mathbf{P}$ , along perturbative lines, is given in Section 17.

**Step 4. Definition 2.5.** A complex-valued function  $f \in L^2(\mathbb{S}^2)$  is said to be even if  $f(-x) = \overline{f(x)}$  for almost every  $x \in \mathbb{S}^2$ .

We will be working almost exclusively with nonnegative functions, for which this condition becomes  $f(-x) \equiv f(x)$ .

**Definition 2.6.** Let  $f \in L^2(\mathbb{S}^2)$  be nonnegative. The antipodally symmetric rearrangement  $f_\star$  is the unique nonnegative element of  $L^2(\mathbb{S}^2)$  that satisfies

$$\begin{aligned} f_\star(-x) &= f_\star(x) && \text{for all } x \in \mathbb{S}^2, \\ f_\star(x)^2 + f_\star(-x)^2 &= f(x)^2 + f(-x)^2 && \text{for all } x \in \mathbb{S}^2. \end{aligned}$$

In other words,  $f_\star(x) = \sqrt{(f(x)^2 + f(-x)^2)/2}$  for all  $x \in \mathbb{S}^2$ .

**Proposition 2.7.** For any nonnegative  $f \in L^2(\mathbb{S}^2)$ ,

$$\|f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)} \leq \|f_\star\sigma * f_\star\sigma\|_{L^2(\mathbb{R}^3)},$$

with strict inequality unless  $f = f_\star$  almost everywhere. Consequently any extremizer for the inequality (1-1) satisfies  $|f(-x)| = |f(x)|$  for almost every  $x \in \mathbb{S}^2$ .

An equivalent formulation is that  $\|\widehat{f\sigma}\|_4 \leq \|\widehat{f_\star\sigma}\|_4$ .

This allows us to restrict attention from nonnegative functions to even nonnegative functions throughout the discussion. This simplification is more convenient than essential.

**Step 5.** A first key step towards gaining control of near-extremals has already been essentially accomplished by Moyua, Vargas, and Vega [Moyua et al. 1999].

**Definition 2.8.** The cap  $\mathcal{C} = \mathcal{C}(z, r)$  with center  $z \in \mathbb{S}^2$  and radius  $r \in (0, 1]$  is the set of all points  $y \in \mathbb{S}^2$  that lie in the same hemisphere, centered at  $z$ , as  $z$  itself, and that satisfy  $|\pi_{H_z}(y)| < r$ , where the subspace  $H_z \subset \mathbb{R}^3$  is the orthogonal complement of  $z$  and  $\pi_{H_z}$  denotes the orthogonal projection onto  $H_z$ .

**Lemma 2.9.** For any  $\delta > 0$  there exist  $C_\delta < \infty$  and  $\eta_\delta > 0$  with the following property. If  $f \in L^2(\mathbb{S}^2)$  satisfies  $\|f\sigma * f\sigma\|_2 \geq \delta^2 \mathbf{S}^2 \|f\|_2^2$ , then there exist a decomposition  $f = g + h$  and a cap  $\mathcal{C}$  satisfying

$$\begin{aligned} 0 &\leq |g|, |h| \leq |f|, \\ g \text{ and } h &\text{ have disjoint supports,} \\ |g(x)| &\leq C_\delta \|f\|_2 |\mathcal{C}|^{-1/2} \chi_{\mathcal{C}}(x) \quad \text{for all } x, \\ \|g\|_2 &\geq \eta_\delta \|f\|_2. \end{aligned}$$

The first conclusion is of course redundant. If  $f \geq 0$  then it follows that  $g, h \geq 0$  almost everywhere.

Lemma 2.9 is a corollary [Moyua et al. 1999, Theorem 4.2]. It can also be proved via arguments closely related to those in [Christ 2011b].

**Step 6.** This step is related to the techniques used in [Christ 2011a].

**Definition 2.10.** Let  $\mathcal{C} = \mathcal{C}(z, r)$  be a cap. For  $z \in \mathbb{S}^2$ , define  $\psi_z(x) = r^{-1}L(\pi_{H_z}(x))$  for  $x$  in the hemisphere  $\{x : x \cdot z > 0\}$ , where  $\pi_{H_z}$  is the orthogonal projection onto  $H_z$ , and  $L = L_z : H_z \rightarrow \mathbb{R}^2$  is an arbitrary linear isometry. The rescaling map associated with  $\mathcal{C}$  is defined by  $\phi_{\mathcal{C}(z,r)} = \psi_z^{-1}$ .

The map  $\phi_{\mathcal{C}(z,r)}$  is a bijection from  $B(0, r^{-1}) \subset \mathbb{R}^2$  to the indicated hemisphere. For  $z = (0, 0, 1)$ ,  $\phi_{\mathcal{C}(z,r)}(y_1, y_2) = (ry_1, ry_2, (1 - r^2|y|^2)^{1/2})$  for  $y \in B(0, r^{-1})$ .

**Definition 2.11.** Let  $\mathcal{C} = \mathcal{C}(z, r)$  be a cap. For  $f \in L^2(\mathbb{S}^2)$ , define the pullback of  $f$  by

$$\phi_{\mathcal{C}}^* f(y) = r \cdot (f \circ \phi_{\mathcal{C}})(y).$$

These pullbacks preserve norms up to uniformly bounded factors provided that  $r \leq r_0 < 1$ ; we have  $\|\phi_{\mathcal{C}}^* f\|_{L^2(\mathbb{R}^2)} \asymp \|f\|_{L^2(\mathbb{S}^2, \sigma)}$ , with the ratio of these norms bounded above and below by positive, finite constants, uniformly in  $f, r, z$ . For the sake of definiteness only, we will sometimes set  $r_0 = \frac{1}{2}$ .

**Definition 2.12.** Let  $\Theta : [1, \infty) \rightarrow (0, \infty)$  satisfy  $\Theta(R) \rightarrow 0$  as  $R \rightarrow \infty$ , and  $\mathcal{C} = \mathcal{C}(z, r) \subset \mathbb{S}^2$  be a cap of radius  $r$  and center  $z$ . A function  $f \in L^2(\mathbb{S}^2)$  is said to be upper normalized, with gauge function  $\Theta$ , with respect to  $\mathcal{C}$ , if

$$\|f\|_2 \leq C < \infty, \tag{2-2}$$

$$\int_{|f(x)| \geq Rr^{-1}} |f(x)|^2 d\sigma(x) \leq \Theta(R) \quad \text{for all } R \geq 1, \tag{2-3}$$

$$\int_{|x-z| \geq Rr} |f(x)|^2 d\sigma(x) \leq \Theta(R) \quad \text{for all } R \geq 1. \tag{2-4}$$

An even function  $f$  is said to be upper even-normalized with respect to  $\Theta, \mathcal{C}(z, r)$  if, when  $f$  is decomposed as  $f = f_+ + f_-$ , where  $f_+$  is the restriction of  $f$  to the hemisphere  $\{x \in \mathbb{S}^2 : x \cdot z > 0\}$ , the summand  $f_+$  is upper normalized with respect to  $\Theta, \mathcal{C}(z, r)$ .

A function  $f \in L^2(\mathbb{R}^2)$  is said to be upper normalized with respect to the unit ball in  $\mathbb{R}^2$  if  $\|f\|_2 \leq C < \infty$ ,  $\int_{|f(x)| \geq R} |f(x)|^2 dx \leq \Theta(R)$  for all  $R \geq 1$ , and  $\int_{|x| \geq R} |f(x)|^2 dx \leq \Theta(R)$  for all  $R \geq 1$ .

For an even function  $f$ , we have  $f_-(x) \equiv \overline{f_+(-x)}$ , for almost every  $x \in \mathbb{S}^2$ . We will usually omit the phrase “with gauge function  $\Theta$ ”, and will say that a function is upper normalized if it satisfies the required inequalities with respect to some appropriate function  $\Theta$  which has been, in principle, specified earlier in the discussion.

**Definition 2.13.** A nonzero function  $f \in L^2(\mathbb{S}^2)$  is said to be  $\delta$ -nearly extremal for the inequality (2-1) if

$$\|f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)} \geq (1 - \delta)^2 \mathbf{S}^2 \|f\|_2^2.$$

**Proposition 2.14.** *There exists a function  $\Theta : [1, \infty) \rightarrow (0, \infty)$  satisfying  $\Theta(R) \rightarrow 0$  as  $R \rightarrow \infty$  with the following property. For any  $\varepsilon > 0$  there exists  $\delta > 0$  such that any nonnegative even function  $f \in L^2(\mathbb{S}^2)$  satisfying  $\|f\|_2 = 1$  that is  $\delta$ -nearly extremal may be decomposed as  $f = F + G$ , where  $F$  and  $G$  are even and nonnegative with disjoint supports,  $\|G\|_2 < \varepsilon$ , and there exists a cap  $\mathcal{C}$  such that  $F$  is upper even-normalized with respect to  $\mathcal{C}$ .*

The proof is a largely formal argument that rests on two inputs: Lemma 2.9, and the observation that  $\|\chi_{\mathcal{C}}\sigma * \chi_{\mathcal{C}'}\sigma\|_2 \ll |\mathcal{C}|^{1/2} |\mathcal{C}'|^{1/2}$  for two caps  $\mathcal{C}$  and  $\mathcal{C}'$ , unless they have comparable radii and nearby centers.

**Step 7.** In this step we establish *a priori* bounds for extremizing sequences, which include a limited but uniform smoothness after suitable rescaling. Step 7 and the closely related Step 9 are the only ones that require nonnegative extremizing sequences.

**Proposition 2.15.** *Let  $\{f_\nu\} \subset L^2(\mathbb{S}^2)$  be an extremizing sequence of nonnegative even functions for the inequality (2-1), satisfying  $\|f_\nu\|_2 \equiv 1$ . Suppose that each  $f_\nu$  is upper even-normalized with respect to a cap  $\mathcal{C}_\nu = \mathcal{C}(z_\nu, r_\nu)$ , with constants uniform in  $\nu$ . Assume that  $\lim_{\nu \rightarrow \infty} r_\nu = 0$ . Then for any  $\varepsilon > 0$  there exists  $C_\varepsilon < \infty$  such that each  $\phi_\nu^*(f_\nu)$  may be decomposed as  $\phi_\nu^*(f_\nu) = G_\nu + H_\nu$  where*

$$\begin{aligned} \|H_\nu\|_2 &< \varepsilon, \\ G_\nu &\text{ is supported where } |x| \leq C_\varepsilon, \\ \|G_\nu\|_{C^1} &\leq C_\varepsilon. \end{aligned}$$

Here  $\phi_\nu^* = \phi_{\mathcal{C}_\nu}^*$ .

Proposition 2.15 expresses a weak form of equicontinuity, after rescaling. In outline: If  $g \in L^2(\mathbb{R}^2)$  satisfies  $\|g\|_2 \sim 1$ , if  $g$  is upper normalized with respect to the unit ball, and if  $g$  is nonnegative, then  $\int_{|\xi| \leq 1} |\widehat{g}(\xi)|^2 d\xi$  is bounded below by a universal strictly positive constant. If the conclusions of the proposition were to fail, then  $g_\nu = \phi_\nu^*(f_\nu)$  would have to satisfy  $\int_{|\xi| \geq \Lambda_\nu} |\widehat{g}_\nu(\xi)|^2 d\xi \geq \eta > 0$ , with

$\limsup \Lambda_\nu = \infty$ . Thus in an appropriately rescaled sense, for some subsequence,  $f_\nu$  would be a superposition of a slowly varying part and a highly oscillatory part, with perhaps some intermediate portion of arbitrarily small norm for large  $\nu$ . For the bilinear expression  $f\sigma * f\sigma$ , we show that the cross term resulting from the high and low frequency parts is small, and that this contradicts extremality.

**Step 8. Proposition 2.16.** *Let  $\{f_\nu\} \subset L^2(\mathbb{S}^2)$  be an extremizing sequence of nonnegative even functions for the inequality (2-1), satisfying  $\|f_\nu\|_2 \equiv 1$ . Suppose that each  $f_\nu$  is upper even-normalized with respect to a cap  $\mathcal{C}_\nu = \mathcal{C}(z_\nu, r_\nu)$ , with constants uniform in  $\nu$ . Then  $\inf_\nu r_\nu > 0$ .*

Thus the situation considered in the hypotheses of Proposition 2.15 cannot arise. The proof of Proposition 2.16 proceeds by contradiction. One can assume that  $r_\nu \rightarrow 0$ . A natural rescaling and transference procedure constructs a corresponding sequence of functions  $\{f_\nu^+\}$  on  $\mathbb{P}^2$ , which possesses a weak form of equicontinuity, as a consequence of Proposition 2.15. In coordinates rescaled according to  $r_\nu$ , each  $f_\nu^+$  is acted upon by an adjoint Fourier restriction operator associated to a hypersurface that depends on  $r_\nu$ , and that approaches  $\mathbb{P}^2$  as  $r_\nu \rightarrow 0$ . The weak equicontinuity of  $\{f_\nu^+\}$ , combined with the convergence of these hypersurfaces, can be used to construct a new sequence  $F_\nu \in L^2(\mathbb{P}^2)$  that satisfies  $\limsup_{\nu \rightarrow \infty} \|\widehat{F_\nu \sigma_P}\|_4 / \|F_\nu\|_2 \geq (3/2)^{-1/4} \lim_{\nu \rightarrow \infty} \|\widehat{f_\nu \sigma}\|_4 / \|f_\nu\|_2$ . It follows that  $\mathcal{R}_{\mathbb{P}^2} \geq (3/2)^{-1/4} \mathcal{R}$ . But this contradicts the inequality  $\mathcal{R} \geq 2^{1/4} \mathcal{R}_{\mathbb{P}^2}$  of Step 3.

**Step 9.** The following variant of Proposition 2.15 is proved by essentially the same reasoning, with one small modification.

**Proposition 2.17.** *Let  $\{f_\nu\} \subset L^2(\mathbb{S}^2)$  be an extremizing sequence of nonnegative even functions for the inequality (2-1), satisfying  $\|f_\nu\|_2 \equiv 1$ . Suppose that each  $f_\nu$  is upper even-normalized with respect to a cap  $\mathcal{C}_\nu = \mathcal{C}(z_\nu, r_\nu)$ , with constants uniform in  $\nu$ . Let  $\rho > 0$ , and suppose that  $r_\nu \geq \rho$  for every  $\nu$ . Then after passing to some subsequence of  $\{r_\nu\}$ , each  $f_\nu$  may be decomposed as  $f_\nu = g_\nu + h_\nu$ , where  $\|h_\nu\|_2 < \varepsilon$  and  $\|g_\nu\|_{C^1} \leq C_{\varepsilon, \rho}$ , where  $C_{\varepsilon, \rho}$  depends only on  $\varepsilon, \rho$ , not on  $\nu$ .*

An application of Rellich’s lemma yields precompactness:

**Corollary 2.18.** *Let  $\{f_\nu\} \subset L^2(\mathbb{S}^2)$  be an extremizing sequence of even nonnegative functions for the inequality (2-1), which are upper even-normalized with respect to a sequence of caps  $\{\mathcal{C}_\nu = \mathcal{C}(z_\nu, r_\nu)\}$ . Then  $\{f_\nu\}$  is precompact in  $L^2(\mathbb{S}^2)$ .*

**Conclusion.** Extremizing sequences exist. We have shown that there exists an extremizing sequence that consists of even, nonnegative functions. Such a sequence is upper even-normalized with respect to a sequence of caps. By Proposition 2.16, the radii of these caps cannot tend to zero. By Corollary 2.18, such a sequence has a subsequence that converges in  $L^2(\mathbb{S}^2)$ . The limit of such a subsequence is an extremal. □

**Not a Step.** As explained above in Step 2, the fundamental potential obstruction to the precompactness of (nonnegative) extremizing sequences was the possibility that  $|f_\nu|^2$  could converge weakly to a Dirac mass, or to a sum of two Dirac masses at a pair of antipodal points. Exclusion of this possibility relied on a suitable lower bound for **S** relative to **P**. The following result examines a natural one-parameter family of candidate trial functions, which provide an alternative source for a lower bound for **S**.



**Proposition 2.19.** *For all  $\xi \in \mathbb{R}^3$  with  $|\xi|$  sufficiently large,*

$$\|\widehat{e_\xi \sigma}\|_{L^4(\mathbb{R}^3)} > \mathcal{R}_{\mathbb{P}^2} \|e_\xi\|_{L^2(\mathbb{S}^2)}.$$

If  $\xi = (0, 0, \lambda)$ , then  $e_\xi^2/\|e_\xi\|_2^2$  does converge weakly as  $\lambda \rightarrow +\infty$  to a constant multiple of a Dirac mass at  $(0, 0, 1)$ . Proposition 2.19 is proved in Section 17 via a perturbative calculation.

By taking the considerations of Step 2 involving even functions into account, Proposition 2.19 provides an alternative route to the essential comparison  $\mathbf{S} > (3/2)^{1/4}\mathbf{P}$ . Although Proposition 2.19 is not strictly necessary for the main lines of our proof, the calculation that underlies it is a natural tool for the investigation of manifolds more general than  $\mathbb{S}^2$ . However, both routes rely on specific properties of the sphere and paraboloid, whose generalization to related problems is not certain.

### 3. Step 2: $\mathbf{S} \geq (3/2)^{1/4}\mathbf{P}$

We begin by establishing the comparison  $\mathbf{S} \geq \mathbf{P}$ . This is based directly on the fact that a sphere is osculated to second order by an appropriate paraboloid.

**Lemma 3.1.** *The optimal constants  $\mathbf{S}$  and  $\mathbf{P}$ , for  $\mathbb{S}^2$  and  $\mathbb{P}^2$  respectively, satisfy  $\mathbf{S} \geq \mathbf{P}$ . Moreover, for any  $r, \varepsilon > 0$  and any  $z \in \mathbb{S}^2$ , there exists a function  $g$  supported in a cap  $\mathcal{C}(z, r) \subset \mathbb{S}^2$  satisfying*

$$\|g\sigma * g\sigma\|_{L^2(\mathbb{R}^3)} \geq (\mathbf{P} - \varepsilon)^2 \|g\|_{L^2(\sigma)}^2,$$

where  $L^2(\sigma)$  denotes  $L^2(\mathbb{S}^2, \sigma)$ .

*Proof.* Rotations are symmetries of the inequality (2-1). That is, for any rotation  $A$  of  $\mathbb{R}^3$  and any  $g \in L^2(\sigma)$ , the function  $g_A = g \circ A$  satisfies  $\|g_A\|_{L^2(\sigma)} = \|g\|_{L^2(\sigma)}$  and  $\|g_A\sigma * g_A\sigma\|_{L^2(\mathbb{R}^3)} = \|g\sigma * g\sigma\|_{L^2(\mathbb{R}^3)}$ . Therefore it is no loss of generality to assume that  $z = (0, 0, 1)$ .

Write  $x = (x', x_3) \in \mathbb{R}^2 \times \mathbb{R}$  as coordinates for  $\mathbb{R}^3$ . Each of the two convolution inequalities under consideration here (one for  $\mathbb{S}^2$ , one for  $\mathbb{P}^2$ ) is equivalent to a corresponding adjoint Fourier restriction inequality, with optimal constants  $\mathcal{R}$  and  $\mathcal{R}_{\mathbb{P}^2}$  respectively. It suffices to prove that for each  $\varepsilon > 0$ , there exists  $f_\varepsilon$  supported in the set of all  $(x', x_3) \in \mathbb{S}^2$  such that  $|x'| < \varepsilon$  and  $x_3 > 0$ , such that  $\|f_\varepsilon\|_{L^2(\sigma)} \leq 1 + \varepsilon$  and  $\|\widehat{f_\varepsilon \sigma}\|_{L^4(\mathbb{R}^3)} \geq (\mathcal{R}_{\mathbb{P}^2} - \varepsilon)$ .

By definition of  $\mathcal{R}_{\mathbb{P}^2}$ , for any  $\varepsilon > 0$  there exists a compactly supported  $C^\infty$  function  $F_\varepsilon : \mathbb{R}^2 \rightarrow \mathbb{R}$  satisfying

$$\int_{\mathbb{R}^2} |F_\varepsilon(x_1, x_2)|^2 dx_1 dx_2 = 1 \quad \text{and} \quad \int_{\mathbb{R}^3} |\widehat{F_\varepsilon \sigma_P}|^4 \geq (\mathcal{R}_{\mathbb{P}^2} - \varepsilon)^4.$$

Here we have mildly abused notation in that the domain of  $F_\varepsilon$  is not  $\mathbb{P}^2$ ; by  $\widehat{F_\varepsilon \sigma_P}(y', y_3)$  we mean  $\int_{\mathbb{R}^2} F_\varepsilon(x') e^{-ix' \cdot y'} e^{-iy_3 |x'|^2/2} dx'$ .

Suppose that  $F_\varepsilon$  is supported in  $\{x' \in \mathbb{R}^2 : |x'| \leq \rho_\varepsilon\}$ , where  $\rho_\varepsilon \geq 1$ . For  $\delta \in (0, \varepsilon \rho_\varepsilon^{-1}]$  and  $(x', x_3) \in \mathbb{S}^2$ , define

$$f_{\varepsilon, \delta}(x', x_3) = \delta^{-1} F_\varepsilon(\delta^{-1} x').$$

Then  $f_{\varepsilon, \delta}$  is supported in  $\mathcal{C}((0, 0, 1), \delta \rho_\varepsilon) \subset \mathcal{C}((0, 0, 1), \varepsilon)$ . Because  $d\sigma(x) = (1 + O(\varepsilon^2))dx'$  in  $\mathcal{C}((0, 0, 1), \varepsilon)$ , we have  $\|f_{\varepsilon, \delta}\|_{L^2(\sigma)} = (1 + O(\varepsilon))$ .

Now

$$\begin{aligned} \widehat{f_{\varepsilon,\delta}\sigma}(y) &= \int_{\mathbb{R}^2} f_{\varepsilon,\delta}(x', \sqrt{1-|x'|^2}) e^{-iy'\cdot x'} e^{-iy_3\sqrt{1-|x'|^2}} h(x') dx' \\ &= \int_{\mathbb{R}^2} \delta^{-1} F_\varepsilon(\delta^{-1}x') e^{-iy'\cdot x'} e^{-iy_3\sqrt{1-|x'|^2}} h(x') dx' \\ &= \delta e^{-iy_3} \int_{\mathbb{R}^2} F_\varepsilon(x') e^{-i\delta y'\cdot x'} e^{-iy_3(\sqrt{1-\delta^2|x'|^2}-1)} h(\delta x') dx', \end{aligned}$$

where  $h = d\sigma/dx'$  satisfies  $h(x') = 1 + O(|x'|^2)$ . Substitute  $(u', u_3) = (\delta y', -\delta^2 y_3)$  and let  $g_{\varepsilon,\delta}(u) = \delta^{-1} e^{iy_3} \widehat{f_{\varepsilon,\delta}\sigma}(y)$ . Then  $\|\widehat{f_{\varepsilon,\delta}\sigma}\|_{L^4(\mathbb{R}^3, dy)} = \|g_{\varepsilon,\delta}\|_{L^4(\mathbb{R}^3 du)}$ , and

$$g_{\varepsilon,\delta}(u) = \int_{\mathbb{R}^2} F_\varepsilon(x') e^{-iu'\cdot x'} e^{i\delta^{-2}u_3(\sqrt{1-\delta^2|x'|^2}-1)} h(\delta x') dx'.$$

Expanding as

$$\delta^{-2}(\sqrt{1-\delta^2|x'|^2}-1) = -\frac{1}{2}|x'|^2 + O(\delta^2|x'|^4)$$

gives

$$g_{\varepsilon,\delta}(u) = \int_{\mathbb{R}^2} F_\varepsilon(x') e^{-iu'\cdot x'} e^{-iu_3|x'|^2/2} (1 + O(\delta^2|x'|^2 + \delta^2|x'|^4)) dx'.$$

Let  $\lambda < \infty$  be another parameter. Then uniformly for all  $u$  satisfying  $|u| \leq \lambda$ ,

$$g_{\varepsilon,\delta}(u) = \widehat{F_\varepsilon\sigma_P}(u) + O(\delta^2\rho_\varepsilon^4).$$

Therefore with  $\varepsilon, \lambda$  fixed,

$$\limsup_{\delta \rightarrow 0} \|\widehat{f_{\varepsilon,\delta}\sigma}\|_{L^4(\mathbb{R}^3)}^4 \geq \int_{|u| \leq \lambda} |\widehat{F_\varepsilon\sigma_P}(u)|^4 du.$$

Therefore

$$\limsup_{\delta \rightarrow 0} \|\widehat{f_{\varepsilon,\delta}\sigma}\|_{L^4(\mathbb{R}^3)} \geq \|\widehat{F_\varepsilon\sigma_P}\|_{L^4(\mathbb{R}^3)} \geq (\mathcal{R}_{\mathbb{P}^2} - \varepsilon),$$

while

$$\|f_{\varepsilon,\delta}\|_{L^2(\sigma)} = 1 + O(\varepsilon). \quad \square$$

Improvement by the factor  $(3/2)^{1/4}$  is based on the reflection symmetry  $x \mapsto -x$  of  $\mathbb{S}^2$ . Recall  $\tilde{f}(x) = \overline{f(-x)}$ , which simplifies to  $\tilde{f}(x) = f(-x)$  for real-valued functions. Denote by  $\langle F, G \rangle$  the pairing of two functions in  $L^2(\mathbb{R}^3)$ , that is,  $\langle F, G \rangle = \int_{\mathbb{R}^3} F\overline{G} dx$ .

**Lemma 3.2.** *For any four real-valued functions  $f_j \in L^2(\mathbb{S}^2)$ ,*

$$\langle f_1\sigma * f_2\sigma, f_3\sigma * f_4\sigma \rangle = \langle f_1\sigma * \tilde{f}_3\sigma, \tilde{f}_2\sigma * f_4\sigma \rangle \tag{3-1}$$

and

$$\|f_1\sigma * f_2\sigma\|_{L^2(\mathbb{R}^3)} = \|f_1\sigma * \tilde{f}_2\sigma\|_{L^2(\mathbb{R}^3)}. \tag{3-2}$$

*Proof.* The inequality  $\|f\sigma * g\sigma\|_{L^2(\mathbb{R}^3)} \leq \mathbf{S}^2 \|f\|_{L^2(\sigma)} \|g\|_{L^2(\sigma)}$  ensures that these quantities are well defined, and that the first identity holds for all  $L^2$  functions provided that it holds for all nonnegative continuous functions  $f_j$ . In that case  $f_3\sigma * f_4\sigma(x) \leq C|x|^{-1}$  for all  $x \in \mathbb{R}^3$ , where  $C < \infty$  depends on  $f_3, f_4$ , and  $f_3\sigma * f_4\sigma$  is continuous except at  $x = 0$ . For real-valued functions  $F \in C^0(\mathbb{R}^3)$  and  $f_j \in C^0(\mathbb{S}^2)$ ,

$$\langle f_1\sigma * f_2\sigma, F \rangle = \int (\tilde{f}_2\sigma * F) f_1 d\sigma,$$

a consequence of the definition of convolution of measures and Fubini's theorem. Limiting arguments then lead to (3-1).

Equation (3-2) now follows:

$$\begin{aligned} \|f_1\sigma * f_2\sigma\|_{L^2(\mathbb{R}^3)}^2 &= \langle f_1\sigma * f_2\sigma, f_1\sigma * f_2\sigma \rangle = \langle f_1\sigma * f_2\sigma, f_2\sigma * f_1\sigma \rangle \\ &= \langle f_1\sigma * \tilde{f}_2\sigma, \tilde{f}_2\sigma * f_1\sigma \rangle = \langle f_1\sigma * \tilde{f}_2\sigma, f_1\sigma * \tilde{f}_2\sigma \rangle = \|f_1\sigma * \tilde{f}_2\sigma\|_{L^2}^2. \quad \square \end{aligned}$$

*Proof of Lemma 2.2.* Let  $g \in L^2(\mathbb{S}^2)$  be supported in  $\{x : x_3 > \frac{1}{2}\}$ . Set  $d\mu = g d\sigma$ . Let  $f(x) = 2^{-1/2}(g(x) + \overline{g(-x)})$  and  $d\nu = f d\sigma = 2^{-1/2}(\mu + \tilde{\mu})$ . The two terms  $g(x)$  and  $g(-x)$  have disjoint supports, so

$$\|f\|_{L^2(\mathbb{S}^2)}^2 = \|g\|_{L^2(\mathbb{S}^2)}^2.$$

Now

$$\nu * \nu = \frac{1}{2}(\mu + \tilde{\mu}) * (\mu + \tilde{\mu}) = \frac{1}{2}((\mu * \mu) + (\tilde{\mu} * \tilde{\mu}) + 2(\mu * \tilde{\mu})).$$

The three summands on the right side have pairwise disjoint supports; the first is supported where  $x_3 > 1$ , the second where  $x_3 < -1$ , and the third where  $|x_3| < 1$ . Therefore

$$\|\nu * \nu\|_{L^2(\mathbb{R}^3)}^2 = \frac{1}{4}(\|\mu * \mu\|_{L^2}^2 + \|\tilde{\mu} * \tilde{\mu}\|_{L^2}^2 + 4\|\mu * \tilde{\mu}\|_{L^2}^2).$$

There holds  $\|\mu * \mu\|_{L^2} = \|\tilde{\mu} * \tilde{\mu}\|_{L^2}$ , since one is the reflection about the origin of the other. By Lemma 3.2, it is also the case that  $\|\mu * \tilde{\mu}\|_{L^2}^2 = \|\mu * \mu\|_{L^2}^2$ . Thus

$$\|\nu * \nu\|_{L^2(\mathbb{R}^3)}^2 = \frac{3}{2}\|\mu * \mu\|_{L^2}^2,$$

establishing Lemma 2.2. □

*Proof of Corollary 2.3.* Let  $\varepsilon > 0$ . Choose  $g \in L^2(\mathbb{S}^2)$ , supported in  $\{x \in \mathbb{S}^2 : x_3 > \frac{1}{2}\}$ , satisfying  $\|g\sigma * g\sigma\|_2^2 \geq (\mathbf{P} - \varepsilon)^4 \|g\|_{L^2(\mathbb{S}^2)}^4$ . By replacing  $g$  by  $|g|$ , we may assume that  $g \geq 0$ .

Consider once more  $f(x) = 2^{-1/2}(g(x) + g(-x))$ . By Lemma 2.2,

$$\|f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)}^2 = \frac{3}{2}\|g\sigma * g\sigma\|_{L^2(\mathbb{R}^3)}^2 \geq \frac{3}{2}(\mathbf{P} - \varepsilon)^4 \|g\|_{L^2(\mathbb{S}^2)}^4 = \frac{3}{2}(\mathbf{P} - \varepsilon)^4 \|f\|_{L^2(\mathbb{S}^2)}^4.$$

Letting  $\varepsilon \rightarrow 0$  yields Corollary 2.3. □

**4. Step 3:  $S \geq 2^{1/4}P$**

*Proof of Lemma 2.4.* We will obtain a lower bound for  $S$  by calculating  $\|f\sigma * f\sigma\|_2^2$  for  $f \equiv 1$ . The following facts are well known: The unit ball in  $\mathbb{R}^3$  has volume  $4\pi/3$ ,  $\sigma(\mathbb{S}^2) = 4\pi$ , and the volume form in  $\mathbb{R}^3$  in polar coordinates is  $r^2 dr d\sigma(\theta)$ .

One calculates that

$$\sigma * \sigma(x) = a|x|^{-1}\chi_{|x|\leq 2} \tag{4-1}$$

for a certain constant  $a > 0$ . We will not need to evaluate  $a$ , which will cancel out at the end of the calculation. Let  $\sigma_P$  denote the measure  $dx'$  on the paraboloid  $\mathbb{P}^2 = \{x \in \mathbb{R}^3 : x_3 = \frac{1}{2}|x'|^2\}$ . What we do need to know is that

$$\sigma_P * \sigma_P(z) \equiv \frac{1}{2}a\chi_\Omega$$

where  $\Omega$  denotes the support of  $\sigma_P * \sigma_P$  and this constant  $a$  is the same as that in (4-1). This factor of  $\frac{1}{2}$  in the definition of  $\mathbb{P}^2$  is required to make the curvature of  $\mathbb{P}^2$  equal to the curvature of  $\mathbb{S}^2$ ; one sees that they are equal by writing the equation for  $\mathbb{S}^2$  near the north pole as  $x_3 - 1 = (1 - |x'|^2)^{1/2} - 1$  and Taylor expanding the right side. Note that the factor  $a/2$  in the formula for  $\sigma_P * \sigma_P$  agrees with the limit as  $|x| \rightarrow 2$  of the function  $a/|x|$ , which appears in the formula for  $\sigma * \sigma$ . This asymptotic equality must hold since the two surfaces have equal curvatures; hence the two convolutions must agree on the diagonal of the maps  $(x, y) \mapsto x + y$ . We will not prove that  $\sigma_P * \sigma_P$  is constant on its support; this is a reflection of the symmetry of the paraboloid (including appropriate dilation symmetry) and invariance of curvature under mappings of the form  $(x', x_3) \mapsto (x', x_3 - L(x'))$  where  $L : \mathbb{R}^2 \rightarrow \mathbb{R}^1$  is linear.

The support of  $\sigma_P * \sigma_P$  is

$$\Omega = \{z : z_3 > \frac{1}{4}|z'|^2\}.$$

It is known [Foschi 2007; Hundertmark and Zharnitsky 2006] that any Gaussian is an extremizer for the paraboloid, and conversely. Another proof that Gaussians extremize the inequality is in [Bennett et al. 2009]. Set  $F(x', x_3) = e^{-|x'|^2/2} \equiv e^{-x_3}$  on the paraboloid. Observe that if  $x + y = z \in \mathbb{R}^3$ , then

$$F(x)F(y) = e^{-x_3 - y_3} = e^{-z_3}.$$

Therefore

$$(F\sigma_P * F\sigma_P)(z) = \frac{1}{2}ae^{-z_3}\chi_{z_3 > |z'|^2/4}.$$

Consequently

$$\begin{aligned} \|F\sigma_P * F\sigma_P\|_2^2 &= \frac{1}{4}a^2 \int_{z' \in \mathbb{R}^2} \int_{z_3 > |z'|^2/4} e^{-2z_3} dz \\ &= \frac{1}{4}a^2 \int_0^\infty 2\pi \int_{r^2/4}^\infty e^{-2s} ds r dr = \frac{1}{4}a^2 2\pi \int_0^\infty \frac{1}{2}e^{-r^2/2} r dr = \frac{1}{4}\pi a^2. \end{aligned}$$

On the other hand,

$$\|\sigma * \sigma\|_{L^2(\mathbb{R}^3)}^2 = \int_{|x|\leq 2} a^2|x|^{-2} dx = a^2 \int_0^2 r^{-2} 4\pi r^2 dr = 4\pi a^2 \int_0^2 dr = 8\pi a^2.$$

Meanwhile

$$\|1\|_{L^2(\sigma)}^2 = \sigma(\mathbb{S}^2) = 4\pi,$$

and

$$\|F\|_{L^2(\sigma_P)}^2 = \int_{\mathbb{R}^2} e^{-2|x|^2/2} dx = \int_0^\infty e^{-r^2} 2\pi r dr = \pi.$$

Putting this all together,

$$\frac{\|F\sigma_P * F\sigma_P\|_2^2}{\|F\|_{L^2(\sigma_P)}^4} = \frac{a^2\pi/4}{\pi^2} = \frac{a^2}{4\pi},$$

while

$$\frac{\|1\sigma * 1\sigma\|_2^2}{\|1\|_{L^2(\sigma)}^4} = \frac{8\pi a^2}{(4\pi)^2} = \frac{a^2}{2\pi}.$$

The second ratio is equal to twice the first, as claimed.  $\square$

#### 5. Step 4: Symmetrization

Proposition 2.7 states that  $\|f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)} \leq \|f_\star\sigma * f_\star\sigma\|_{L^2(\mathbb{R}^3)}$  for any nonnegative function  $f \in L^2(\mathbb{S}^2)$ , where  $f_\star$  denotes the antipodally symmetric rearrangement of  $f$ , defined in Definition 2.6.

*Proof of Proposition 2.7.* Let  $\sigma$  denote surface measure on  $\mathbb{S}^2$ . For  $h \geq 0$ ,

$$\|h\sigma * h\sigma\|_{L^2}^2 = \int h(a)h(b)h(c)h(d) d\lambda(a, b, c, d) \quad (5-1)$$

for a certain nonnegative measure  $\lambda$  that is supported on the set where  $a + b = c + d$ , and that is invariant under the transformations

$$\begin{aligned} (a, b, c, d) &\mapsto (b, a, c, d), & (a, b, c, d) &\mapsto (a, -c, -b, d) \\ (a, b, c, d) &\mapsto (c, d, a, b), & (a, b, c, d) &\mapsto (-a, -b, -c, -d). \end{aligned}$$

This invariance, which is essential to the discussion, follows from the identities

$$\begin{aligned} f\sigma * g\sigma &= g\sigma * f\sigma, \\ \langle f\sigma * g\sigma, h\sigma * k\sigma \rangle &= \langle h\sigma * k\sigma, f\sigma * g\sigma \rangle, \\ \langle f\sigma * g\sigma, h\sigma * k\sigma \rangle &= \langle f\sigma * \tilde{h}\sigma, \tilde{g}\sigma * k\sigma \rangle \end{aligned}$$

for arbitrary real-valued functions, where  $\tilde{F}(x) = F(-x)$ .

Denote by  $G$  the finite group of symmetries of  $(\mathbb{R}^3)^4$  that these generate.  $G$  has cardinality 48. Indeed, exactly one of  $a$  and  $-a$  appears; suppose that  $a$  appears. There are 4 places in which it can go. Then  $\pm b$  can go into any of 3 slots, but whether it is  $+b$  or  $-b$  is determined by which slot. There remain two slots into which  $\pm c$  can go; again, the  $\pm$  sign is determined by the slot. Then  $\pm d$  goes into the remaining slot, with the  $\pm$  sign again determined. The analysis is parallel if  $-a$  appears. Thus there are  $2 \times 4 \times 3 \times 2 = 48$  possibilities.

By the orbit of a point we mean its image under  $G$ ; by a generic point we mean one whose orbit has cardinality 48. In (5-1), it suffices to integrate only over all *generic* 4-tuples  $(a, b, c, d)$  satisfying  $a + b = c + d$ , since these form a set of full  $\lambda$ -measure.

To the orbit  $\mathbb{O}$  we associate the functions

$$\mathcal{F}(\mathbb{O}) = \sum_{(a,b,c,d) \in \mathbb{O}} f(a)f(b)f(c)f(d) \quad \text{and} \quad \mathcal{F}_*(\mathbb{O}) = \sum_{(a,b,c,d) \in \mathbb{O}} f_*(a)f_*(b)f_*(c)f_*(d).$$

Let  $\Omega$  denote the set of all orbits of generic points. We can write

$$\|f * f\|_{L^2}^2 = \int_{\Omega} \mathcal{F}(\mathbb{O}) d\tilde{\lambda}(\mathbb{O}) \quad \text{and} \quad \|f_* * f_*\|_{L^2}^2 = \int_{\Omega} \mathcal{F}_*(\mathbb{O}) d\tilde{\lambda}(\mathbb{O})$$

for a certain nonnegative measure  $\tilde{\lambda}$ . Therefore it suffices to prove that for any generic orbit  $\mathbb{O}$ ,

$$\sum_{(a,b,c,d) \in \mathbb{O}} f(a)f(b)f(c)f(d) \leq \sum_{(a,b,c,d) \in \mathbb{O}} f_*(a)f_*(b)f_*(c)f_*(d). \tag{5-2}$$

Fix any generic ordered 4-tuple  $(a, b, c, d)$  satisfying  $a + b = c + d$ . We prove (5-2) for its orbit. By homogeneity, it is no loss of generality to assume that  $f^2(a) + f^2(-a) = 1$  and that the same holds simultaneously for  $b, c, d$ . Thus we may write

$$f(a) = \cos(\varphi), \quad f(b) = \cos(\psi), \quad f(c) = \cos(\alpha), \quad f(d) = \cos(\beta)$$

for some  $\varphi, \psi, \alpha, \beta \in [0, \pi/2]$  with  $f(-a) = \sin(\varphi), \dots, f(-d) = \sin(\beta)$ . This means that

$$f_*(x) = 2^{-1/2} \quad \text{for each } x \in \{\pm a, \pm b, \pm c, \pm d\}.$$

Now

$$\begin{aligned} \frac{1}{8} \sum_{(a',b',c',d') \in \mathbb{O}} f(a')f(b')f(c')f(d') &= \cos(\varphi)\cos(\psi)\cos(\alpha)\cos(\beta) + \sin(\varphi)\sin(\psi)\sin(\alpha)\sin(\beta) \\ &\quad + \cos(\varphi)\sin(\psi)\cos(\alpha)\sin(\beta) + \cos(\varphi)\sin(\psi)\sin(\alpha)\cos(\beta) \\ &\quad + \sin(\varphi)\cos(\psi)\cos(\alpha)\sin(\beta) + \sin(\varphi)\cos(\psi)\sin(\alpha)\cos(\beta) \\ &= \Gamma(\varphi, \psi, \alpha, \beta), \end{aligned}$$

where

$$\Gamma(\varphi, \psi, \alpha, \beta) = \cos(\varphi)\cos(\psi)\cos(\alpha)\cos(\beta) + \sin(\varphi)\sin(\psi)\sin(\alpha)\sin(\beta) + \sin(\varphi + \psi)\sin(\alpha + \beta).$$

Therefore the following lemma will complete the proof of Proposition 2.7. □

**Lemma 5.1.**  $\max_{\varphi, \psi, \alpha, \beta \in [0, \pi/2]} \Gamma(\varphi, \psi, \alpha, \beta) = \frac{3}{2}$ . Moreover, this maximum value is attained only at  $(\frac{\pi}{4}, \frac{\pi}{4}, \frac{\pi}{4}, \frac{\pi}{4})$ .

Since

$$\Gamma(\frac{\pi}{4}, \frac{\pi}{4}, \frac{\pi}{4}, \frac{\pi}{4}) = 1 + (1/\sqrt{2})^4 + (1/\sqrt{2})^4 = \frac{3}{2},$$

the maximum value of  $\Gamma$  is at least  $\frac{3}{2}$ . This point corresponds to the values taken by  $f_\star$ . Compare this with  $\Gamma(0, 0, 0, 0) = 1$ , which represents the extreme case when  $f$  vanishes at one of each pair of antipodal points; this ratio  $(3/2)/1$  is the same  $3/2$  that appears in Corollary 2.3.

*Proof.* We write  $\Gamma$  as

$$\begin{aligned}\Gamma &= \cos(\phi + \psi) \cos(\alpha + \beta) + \sin(\phi + \psi) \sin(\alpha + \beta) + \cos \phi \cos \psi \sin \alpha \sin \beta + \sin \phi \sin \psi \cos \alpha \cos \beta \\ &= \cos((\phi + \psi) - (\alpha + \beta)) + \cos \phi \cos \psi \sin \alpha \sin \beta + \sin \phi \sin \psi \cos \alpha \cos \beta.\end{aligned}$$

Now

$$\begin{aligned}\cos \phi \cos \psi &= \frac{\cos(\phi + \psi) + \cos(\phi - \psi)}{2} \leq \frac{1 + \cos(\phi + \psi)}{2}, \\ \sin \alpha \sin \beta &= \frac{-\cos(\alpha + \beta) + \cos(\alpha - \beta)}{2} \leq \frac{1 - \cos(\alpha + \beta)}{2}\end{aligned}$$

with equality only if  $\phi = \psi$  and  $\alpha = \beta$ , and there are similar identities for  $\sin \phi \sin \psi$  and  $\cos \alpha \cos \beta$ .

Therefore

$$\begin{aligned}\Gamma &\leq \cos((\phi + \psi) - (\alpha + \beta)) + \frac{1}{4}(1 + \cos(\phi + \psi))(1 - \cos(\alpha + \beta)) \\ &\quad + \frac{1}{4}(1 - \cos(\phi + \psi))(1 + \cos(\alpha + \beta)) \\ &= \cos((\phi + \psi) - (\alpha + \beta)) + \frac{1}{2}(1 - \cos(\phi + \psi) \cos(\alpha + \beta)) \\ &= \cos((\phi + \psi) - (\alpha + \beta)) - \frac{1}{2}(\cos((\phi + \psi) + (\alpha + \beta)) + \cos((\phi + \psi) - (\alpha + \beta))) + \frac{1}{2} \\ &= \frac{1}{2}(\cos((\phi + \psi) - (\alpha + \beta)) - \cos((\phi + \psi) + (\alpha + \beta))) + \frac{1}{2} \leq \frac{3}{2}.\end{aligned}$$

The value  $\frac{3}{2}$  can only be attained if all inequalities in this derivation are equalities. Equality in the final inequality forces  $\phi + \psi + \alpha + \beta = \pi$  and  $\phi + \psi = \alpha + \beta$ . Together with the equalities  $\phi = \psi$  and  $\alpha = \beta$  already noted, these force  $\phi = \psi = \alpha = \beta = \pi/4$ .  $\square$

## 6. Step 5: Big pieces of caps

In this section we prove Lemma 2.9. While we are ultimately interested in establishing strong structural control of near-extremal functions, here we establish a weak connection between functions satisfying modest lower bounds  $\|\widehat{f\sigma}\|_4 \geq \delta \|f\|_2$ , with  $\delta > 0$  arbitrarily small, and characteristic functions of caps.

For each integer  $k \geq 0$  choose a maximal subset  $\{z_k^j\} \subset \mathbb{S}^2$  satisfying  $|z_k^j - z_k^i| \geq 2^{-k}$  for all  $i \neq j$ . Then for any  $x \in \mathbb{S}^2$  there exists  $z_k^i$  such that  $|x - z_k^i| \leq 2^{-k}$ ; otherwise  $x$  could be adjoined to  $\{z_k^j\}$ , contradicting maximality. Therefore the caps  $\mathcal{C}_k^j = \mathcal{C}(z_k^j, 2^{-k+1})$  cover  $\mathbb{S}^2$  for each  $k$ , and there exists  $C < \infty$  such that for any  $k$ , no point of  $\mathbb{S}^2$  belongs to more than  $C$  of the caps  $\mathcal{C}_k^j$ . The constant  $C$  is independent of  $k$ .

For  $p \in [1, \infty)$ , the  $X_p$  norm is defined by

$$\|f\|_{X_p}^4 = \sum_{k=0}^{\infty} \sum_j 2^{-4k} \left( |\mathcal{C}_k^j|^{-1} \int_{\mathcal{C}_k^j} |f|^p \right)^{4/p}.$$

The factor  $2^{-4k}$  can alternatively be written as  $|\mathcal{C}_k^j|^2$ .

Define also

$$\Lambda_{k,j}(f) = \left( |\mathcal{C}_k^j|^{-1} \int_{\mathcal{C}_k^j} |f| \right) \left( |\mathcal{C}_k^j|^{-1} \int_{\mathbb{S}^2} |f|^2 \right)^{-1/2}.$$

By Hölder’s inequality,

$$\Lambda_{k,j}(f) \leq \left( |\mathcal{C}_k^j|^{-1} \int_{\mathcal{C}_k^j} |f|^2 \right)^{1/2} \left( |\mathcal{C}_k^j|^{-1} \int_{\mathbb{S}^2} |f|^2 \right)^{-1/2} = \|f\|_{L^2(\mathcal{C}_k^j)} / \|f\|_{L^2(\mathbb{S}^2)} \leq 1.$$

It is shown in [Moyua et al. 1999, Lemma 4.4] that  $L^2 \subset X_p$  for any  $p < 2$ . We will exploit the following refinement, which is very closely related to a result in Bégout and Vargas [2007], and whose somewhat tedious proof is deferred to Section 18.

**Lemma 6.1.** *For any  $p \in [1, 2)$ , there exist  $C < \infty$  and  $\gamma > 0$  such that for any  $f \in L^2(\mathbb{S}^2)$ ,*

$$\|f\|_{X_p} \leq C \|f\|_2 \sup_{k,j} (\Lambda_{k,j}(f))^\gamma.$$

Thus  $\|f\|_{X_p} \leq C_p \|f\|_2$  for any  $f \in L^2(\mathbb{S}^2)$ . Moreover, when the  $X_p$  norm is not significantly smaller than the  $L^2$  norm,  $\sup_{k,j} \Lambda_{k,j}(f)$  cannot be small.

**Proposition 6.2** (Moyua, Vargas, and Vega [1999]). *There exist  $C < \infty$  and  $p \in (1, 2)$  such that for any  $f \in L^2(\mathbb{S}^2)$ ,*

$$\|\widehat{f\sigma}\|_{L^4(\mathbb{R}^3)} \leq C \|f\|_{X_p}.$$

This result contains Lemma 2.9 by an elementary argument, but we give the details for the sake of completeness.

*Proof of Lemma 2.9.* Let  $\delta > 0$ . Let  $0 \neq f \in L^2(\mathbb{S}^2)$  and suppose that  $\|\widehat{f\sigma}\|_{L^4(\mathbb{R}^3)} \geq \delta \|f\|_2$ . For convenience, normalize so that  $\|f\|_2 = 1$ . The hypothesis, combined with the proposition and the lemma above, yields

$$\sup_{k,j} \Lambda_{k,j}(f) \geq c \delta^{1/\gamma}.$$

Fix  $k$  and  $j$  such that  $\Lambda_{k,j}(f) \geq \frac{1}{2} c \delta^{1/\gamma}$ . Henceforth write  $\mathcal{C} = \mathcal{C}_k^j$ . Thus

$$\int_{\mathcal{C}} |f| \geq c_0 \delta^{1/\gamma} |\mathcal{C}|^{1/2},$$

where  $c_0 > 0$  is a constant independent of  $f$ .

Let  $R \geq 1$ . Define  $E = \{x \in \mathcal{C} : |f(x)| \leq R\}$ . Set  $g = f\chi_E$  and  $h = f - f\chi_E$ . Then  $g$  and  $h$  have disjoint supports,  $g + h = f$ ,  $g$  is supported on  $\mathcal{C}$ , and  $\|g\|_\infty \leq R$ . Now  $|h(x)| \geq R$  for almost every  $x \in \mathcal{C}$  for which  $h(x) \neq 0$ , so

$$\int_{\mathcal{C}} |h| \leq R^{-1} \int_{\mathcal{C}} |h|^2 \leq R^{-1} \|f\|_2^2 = R^{-1}.$$

Define  $R$  by  $R^{-1} = \frac{1}{2} c_0 \delta^{1/\gamma} |\mathcal{C}|^{1/2}$ . Then

$$\int_{\mathcal{C}} |g| = \int_{\mathcal{C}} |f| - \int_{\mathcal{C}} |h| \geq \frac{1}{2} c_0 \delta^{1/\gamma} |\mathcal{C}|^{1/2}.$$



By Hölder's inequality, since  $g$  is supported on  $\mathcal{C}$ ,

$$\|g\|_2 \geq |\mathcal{C}|^{-1+1/2} \|g\|_{L^1(\mathcal{C})} \geq c\delta^{1/\gamma} = c\delta^{1/\gamma} \|f\|_2.$$

Thus the decomposition  $f = g + h$  satisfies the conclusions of Lemma 2.9, with  $\eta_\delta$  proportional to  $\delta^{1/\gamma}$ , and  $C_\delta$  proportional to  $\delta^{-1/\gamma}$ .  $\square$

## 7. Analytic preliminaries

*On near-extremals.*

**Lemma 7.1.** *Let  $f = g + h \in L^2(\mathbb{S}^2)$ . Suppose that  $g \perp h$ ,  $g \neq 0$ , and that  $f$  is  $\delta$ -nearly extremal for some  $\delta \in (0, \frac{1}{4}]$ . Then*

$$\frac{\|h\|_2}{\|f\|_2} \leq C \max\left(\frac{\|h\sigma * h\sigma\|_2^{1/2}}{\|h\|_2}, \delta^{1/2}\right). \quad (7-1)$$

Here  $C < \infty$  is a constant independent of  $g$  and  $h$ .

*Proof.* The inequality is invariant under multiplication of  $f$  by a positive constant, so we may assume without loss of generality that  $\|g\|_2 = 1$ . We may assume that  $\|h\|_2 > 0$ , since otherwise the conclusion is trivial. Define  $y = \|h\|_2$  and

$$\eta = \|h\sigma * h\sigma\|_2^{1/2} / \mathbf{S}\|h\|_2.$$

If  $\eta > \frac{1}{2}$ , then (7-1) holds trivially with  $C = 2/\mathbf{S}$ , for the left side cannot exceed 1 since  $f = g + h$  with  $g \perp h$ .

Since  $\|f\sigma * f\sigma\|_2^{1/2}$  is a constant multiple of  $\|\widehat{f\sigma}\|_4$ , the functional  $f \mapsto \|f\sigma * f\sigma\|_2^{1/2}$  satisfies the triangle inequality. Therefore

$$(1 - \delta)^4 \mathbf{S}^4 \|f\|_2^4 \leq \|f\sigma * f\sigma\|_2^2 \leq (\|g\sigma * g\sigma\|_2^{1/2} + \|h\sigma * h\sigma\|_2^{1/2})^4 \leq \mathbf{S}^4 (1 + \eta y)^4.$$

Since  $g \perp h$ ,  $\|f\|_2^2 = 1 + y^2$  and therefore

$$(1 - \delta)(1 + y^2)^{1/2} \leq 1 + \eta y.$$

Squaring gives

$$(1 - 2\delta)(1 + y^2) \leq 1 + 2\eta y + \eta^2 y^2.$$

Since  $\delta \in (0, \frac{1}{4}]$  and  $\eta \leq \frac{1}{2}$ ,

$$\frac{1}{2}y^2 \leq 2\delta + 2\eta y + \eta^2 y^2 \leq 2\delta + 2\eta y + \frac{1}{4}y^2,$$

whence either  $y^2 \leq 16\delta$  or  $y \leq 16\eta$ .

Substituting the definitions of  $y$  and  $\eta$  and majorizing  $\|h\|_2/\|f\|_2$  by  $\|h\|_2/\|g\|_2$  yields the stated conclusion.  $\square$

**Simple bilinear convolution estimates.**

**Lemma 7.2.** *Let  $f \in L^2(\mathbb{S}^2)$  be nonnegative and satisfy  $\|f\|_2 \leq 1$ . Let  $z \in \mathbb{S}^2$  and  $\varepsilon > 0$ . Let  $R \geq 1$  and  $0 < \rho \leq 1$ . Then*

$$\|f\sigma * f\sigma\|_{L^2(\{|x|>2-\varepsilon\})} \leq CR^2\varepsilon^{1/2}\rho + C\left(\int_{f(x)\geq R} f^2(x) d\sigma(x)\right)^{1/2} + C\left(\int_{|x-z|\geq\rho} f^2(x) d\sigma(x)\right)^{1/2}.$$

*Proof.* Decompose  $f = g + h$  where  $g$  and  $h$  are nonnegative,

$$\|h\|_2 \leq \left(\int_{f(x)\geq R} f^2(x) d\sigma(x)\right)^{1/2} + \left(\int_{|x-z|\geq\rho} f^2(x) d\sigma(x)\right)^{1/2}$$

and  $\|g\|_2 \leq 1$  and  $\|g\|_\infty \leq R$ , and  $g$  is supported on  $\{x \in \mathbb{S}^2 : |x - z| \leq \rho\}$ . Then

$$g\sigma * g\sigma(x) \leq R^2\sigma * \sigma(x) \leq CR^2|x|^{-1}$$

for  $|x| < 2$ , and equals 0 otherwise. Moreover,  $g\sigma * g\sigma$  is supported in  $\{x : |x - 2z| < 2\rho\}$ . The  $L^2(\mathbb{R}^3)$  norm of  $|x|^{-1}1_{|x|\leq 2}$  over the intersection of this region with  $\{x : |x| > 2 - \varepsilon\}$  is  $\leq C\rho\varepsilon^{1/2}$ . This gives the bound  $CR^2\rho\varepsilon^{1/2}$  for  $\|g\sigma * g\sigma\|_2$ . Since  $\|g\|_2 \leq 1$ , the general inequality

$$\|F\sigma * G\sigma\|_{L^2(\mathbb{R}^3)} \leq C\|F\|_2\|G\|_2$$

gives the required bound for both  $g\sigma * h\sigma$  and  $h\sigma * h\sigma$ . □

**Corollary 7.3.** *Let  $\{f_\nu\}$  be a sequence of real-valued functions that are upper even-normalized above with respect to a sequence of caps  $\mathcal{C}_\nu$  of radii  $r_\nu$ . If*

$$\delta_\nu/r_\nu^2 \rightarrow 0,$$

*then*

$$\int_{|x|>2-\delta_\nu} (|f_\nu|\sigma * |f_\nu|\sigma)^2 dx \rightarrow 0 \quad \text{as } \nu \rightarrow \infty.$$

**Lemma 7.4.** *Let  $f \in L^2(\mathbb{S}^2)$  be a function that is upper even-normalized with respect to a cap  $\mathcal{C}$  of radius  $r$ . Then for all  $R \geq 1$ ,*

$$\int_{R^{1/2}r \leq |x| \leq 2-Rr^2} |(f\sigma * f\sigma)(x)|^2 dx \leq \Psi(R),$$

*where  $\Psi(R) \rightarrow 0$  as  $R \rightarrow \infty$ , and  $\Psi$  depends only on the function  $\Theta$  in the normalization inequalities (2-3) and(2-4), not on  $r$ .*

*Proof.* It suffices to prove this for  $r$  small,  $R$  large, and  $Rr^2$  uniformly bounded. Let  $\mathcal{C} = \mathcal{C}(z, r)$  have center  $z \in \mathbb{S}^2$ . Let  $A \in [1, \infty)$  and decompose  $f = g_+ + h_+ + g_- + h_-$ , where  $g_+, g_-$  are supported respectively in  $\mathcal{C}(z, Ar)$  and  $\mathcal{C}(-z, Ar)$ ,  $\|h_+\|_2 \leq \Theta(A)$  and  $\|h_-\|_2 \leq \Theta(A)$ , where  $\Theta(A) \rightarrow 0$  as  $A \rightarrow \infty$ .

Expand  $f\sigma * f\sigma$  as a sum of the resulting 16 terms. The terms  $g_+\sigma * g_+\sigma$  and  $g_-\sigma * g_-\sigma$  are supported where  $|x| > 2 - CA^2r^2$ . If we choose  $A$  so that  $CA^2 < R$ , then these vanish identically in the

region  $|x| \leq 2 - Rr^2$ . The (two) terms  $g_+\sigma * g_-\sigma$  are supported where  $|x| \leq CAr$ . Therefore they also contribute nothing, provided that  $CAr \leq R^{1/2}r$ .

Each of the remaining terms involves at least one factor of  $h_+$  or of  $h_-$ . Since  $\|F\sigma * G\sigma\|_{L^2(\mathbb{R}^3)} \leq C\|F\|_2\|G\|_2$  for all  $F, G \in L^2(\mathbb{S}^2)$ , and since  $g_\pm, h_\pm = O(1)$  in  $L^2(\mathbb{S}^2)$  norm, each of these terms is  $O(\|h_\pm\|_2)$ . Therefore

$$\int_{R^{1/2}r \leq |x| \leq 2 - Rr^2} |f\sigma * f\sigma(x)|^2 dx \leq C\Theta(A)^2$$

for any  $A$  that satisfies  $CA^2 < R$ . This completes the proof, provided that  $Rr^2 = O(1)$ .  $\square$

The set of all caps can be made into a metric space. Define the distance  $\rho$  from  $\mathcal{C}(y, r)$  to  $\mathcal{C}(y', r')$  to be the Euclidean distance from  $(y/r, \log(1/r))$  to  $(y'/r', \log(1/r'))$  in  $\mathbb{R}^3 \times \mathbb{R}^+$ . Note that for instance when  $r = r'$ , the distance is  $r^{-1}|y - y'|$ , so this distance has the natural scaling. If  $y = y'$ , then the distance is  $|\log(r/r')|$ ; this has the natural property that it depends only on the *ratio* of the two radii. The definition ensures that this is truly a metric.

For any metric space  $(X, \rho)$  and any equivalence relation  $\equiv$  on  $X$ , the function

$$\varrho([x], [y]) = \inf_{x' \in [x], y' \in [y]} \rho(x', y')$$

is a metric on the set of equivalence classes  $X/\equiv$ . Let  $\mathcal{M}$  be the set of all caps  $\mathcal{C} \subset \mathbb{S}^2$  modulo the equivalence relation  $\mathcal{C} \equiv -\mathcal{C}$ , where  $-\mathcal{C} = \{-z : z \in \mathcal{C}\}$ . Then the following defines a metric on  $\mathcal{M}$ .

**Definition 7.5.** For any two caps  $\mathcal{C}, \mathcal{C}' \subset \mathbb{S}^2$ ,

$$\varrho([\mathcal{C}], [\mathcal{C}']) = \min(\rho(\mathcal{C}, \mathcal{C}'), \rho(-\mathcal{C}, \mathcal{C}')),$$

where  $[\mathcal{C}]$  denotes the equivalence class  $[\mathcal{C}] = \{\mathcal{C}, -\mathcal{C}\} \in \mathcal{M}$ .

We will also write  $\varrho(\mathcal{C}, \mathcal{C}') = \varrho([\mathcal{C}], [\mathcal{C}'])$ .

**Lemma 7.6.** For any  $\varepsilon > 0$  there exists  $\rho < \infty$  such that

$$\|\chi_{\mathcal{C}}\sigma * \chi_{\mathcal{C}'}\sigma\|_{L^2(\mathbb{R}^3)} < \varepsilon|\mathcal{C}|^{1/2}|\mathcal{C}'|^{1/2}, \quad \text{whenever } \varrho(\mathcal{C}, \mathcal{C}') > \rho.$$

*Proof.* Let  $\mathcal{C} = \mathcal{C}(z, r)$  and  $\mathcal{C}' = \mathcal{C}(z', r')$ . Set  $f = |\mathcal{C}|^{-1/2}\chi_{\mathcal{C}} \leq Cr^{-1}\chi_{\mathcal{C}}$  and  $f' = |\mathcal{C}'|^{-1/2}\chi_{\mathcal{C}'} \leq Cr'^{-1}\chi_{\mathcal{C}'}$ . Without loss of generality,  $r' \leq r$ . We may suppose that  $r' \ll 1$ ; otherwise the caps are not far apart. We will also assume at first that no points are nearly antipodal, that is, that  $|x + x'| \geq \delta$  for all  $x \in \mathcal{C}$  and  $x' \in \mathcal{C}'$ , for some fixed constant  $\delta > 0$ ; we will return to this point later.

Consider first the case where  $r \sim r'$ . Then we may assume that  $|z - z'| \geq 10r$ , say. Then  $f\sigma * f'\sigma$  has  $L^\infty$  norm  $\leq Cr^{-2} \cdot r/|z - z'|$ , and is supported in a three-dimensional cylinder whose base has radius  $Cr$  and whose height is  $\leq Cr^2 + Cr|z - z'| \leq Cr|z - z'|$ . The volume of this cylinder is  $\leq Cr^3|z - z'|$ . In all,

$$\|f\sigma * f'\sigma\|_{L^2(\mathbb{R}^3)} \leq Cr^{-1}|z - z'|^{-1} \cdot r^{3/2}|z - z'|^{1/2} = C(r/|z - z'|)^{1/2},$$

which is small precisely when the caps are far apart.

Consider next the case where  $r' \ll r$ , and still  $|z - z'| \geq 10r$ . Then the  $L^\infty$  norm is no more than  $Cr^{-1}r'^{-1} \cdot r'|z - z'|^{-1}$ . The support is contained in a tubular neighborhood of a (translated) cap of radius  $Cr$ ; this tubular neighborhood has width  $\leq Cr'|z - z'|$ . Hence the volume of the support is  $\leq Cr^2r'|z - z'|$ . Consequently

$$\|f\sigma * f'\sigma\|_{L^2(\mathbb{R}^3)} \leq Cr^{-1}r'^{-1}|z - z'|^{-1} \cdot rr'^{1/2}|z - z'|^{1/2} = C(r'/|z - z'|)^{1/2} \leq C(r'/r)^{1/2}.$$

Consider next the case where  $r' \ll r$  and  $|z - z'| \leq 10r$ . It suffices to replace  $f$  by its restriction  $F$  to the complement of the cap  $\mathcal{C}^*$  centered at  $z'$  of radius  $10r^{3/4}r'^{1/4}$ , since

$$\|f - F\|_2 \leq Cr^{-1}r^{3/4}r'^{1/4} = C(r'/r)^{1/4} \ll 1.$$

$F\sigma * f'\sigma$  is supported in a region of volume  $\leq Cr^3r'$ , and as is easily verified,

$$\|F\sigma * f'\sigma\|_\infty \leq Cr^{-1}r'^{-1} \cdot (r'/r^{3/4}r'^{1/4}) = Cr^{-7/4}r'^{-1/4}.$$

Therefore

$$\|F\sigma * f'\sigma\|_2 \leq Cr^{-7/4}r'^{-1/4} \cdot (r^3r')^{1/2} = Cr^{-1/4}r'^{1/4} \ll 1.$$

It only remains to handle caps that are nearly antipodal. But this follows from the nonantipodal case by the identity

$$\|f\sigma * g\sigma\|_2 = \|\tilde{f}\sigma * g\sigma\|_2, \quad \text{where } \tilde{f}(x) \equiv \overline{f(-x)}. \quad \square$$

**Fourier integral operators.** Here we discuss another ingredient required for the proof of Lemma 12.2, certain estimates that rely on cancellation, in contrast to those in the preceding section.

For  $0 < \rho \lesssim 1$ , define  $T_\rho : L^2(\mathbb{S}^2) \rightarrow L^2(\mathbb{S}^2)$  by

$$T_\rho f(x) = \int f(y) d\mu_{x,\rho}(y),$$

where  $\mu_{x,\rho}$  is arc-length measure on the circle  $\{y \in \mathbb{S}^2 : |y - x| = \rho\}$ , normalized to be a probability measure.

Let  $\Delta$  denote the spherical Laplacian.

**Lemma 7.7.** *We have*

$$\|T_\rho f\|_{L^2(\mathbb{S}^2)} \leq C\|(I - \rho^2\Delta)^{-1/4}f\|_{L^2(\mathbb{S}^2)} \tag{7-2}$$

uniformly for all  $\rho > 0$  and all  $f \in L^2(\mathbb{S}^2)$ .

*Sketch of proof.* There are three elements in the proof of (7-2).

(i) Consider any fixed  $\rho \in (0, 2)$ . Define  $\Phi_\rho(x, y) = |x - y|^2 - \rho^2$ . Then the  $3 \times 3$  matrix

$$\begin{pmatrix} 0 & \partial\Phi_\rho/\partial x \\ \partial\Phi_\rho/\partial y & \partial^2\Phi_\rho/\partial x\partial y \end{pmatrix} \tag{7-3}$$

is nonsingular for any  $(x, y)$  satisfying  $\Phi_\rho(x, y) = 0$ . This is a straightforward computation, easily done by taking advantage of rotational symmetry to reduce to a computation of Taylor expansions about  $x = (0, 0, 1)$  and  $y = (\cos(\theta), 0, \sin(\theta))$ .

(ii)  $T_\rho$  is defined by integration against a smooth density on  $\{(x, y) \in \mathbb{S}^2 \times \mathbb{S}^2 : \Phi_\rho(x, y) = 0\}$ . As discussed on [Sogge 1993, pages 188–9], the nonsingularity of the matrix (7-3) implies that  $T_\rho$  is a Fourier integral operator of order  $-(n - 1)/2 = -1/2$  on  $\mathbb{S}^n = \mathbb{S}^2$ . Any such operator is smoothing of order  $1/2$  in the scale of  $L^2$  Sobolev spaces [Sogge 1993].

(iii) If  $T_\rho$  is rewritten with appropriate normalizations in coordinates adapted to any cap  $\mathcal{C}(z, \rho)$ , then the inequality holds uniformly in  $\rho$ . The only issue here is as  $\rho \rightarrow 0$ , but plainly in that situation there is a limiting operator on  $\mathbb{R}^2$ ,  $f \mapsto \int_{\mathbb{S}^1} f(x - y) d\mu(y)$ , where  $\mu$  is arc length measure on  $\mathbb{S}^1 \subset \mathbb{R}^2$ . This limiting operator is again a Fourier integral operator of order  $-1/2$ . It follows that the bounds are uniform after rescaling. Reversal of the rescaling introduces the factor  $\rho^2$  to  $\Delta$  in the inequality.  $\square$

The operators  $T_\rho$  are related to our bilinear convolutions: For  $f \in L^2(\mathbb{S}^2)$  and  $x \in \mathbb{R}^3$  satisfying  $0 < |x| < 2$ ,

$$(f\sigma * \sigma)(x) = c|x|^{-1}T_\rho f(x/|x|),$$

where  $\rho^2 + |x/2|^2 = 1$ . Define  $e_\xi(x) = e^{x \cdot \xi}$ , for  $x \in \mathbb{R}^3$  and  $\xi \in \mathbb{C}$  (and in particular for  $x \in \mathbb{S}^2$ ). There is the more general identity

$$(f\sigma * e_{i\xi}\sigma)(x) = e_{i\xi}(x)(e_{-i\xi}f\sigma * \sigma)(x) = c|x|^{-1}e_{i\xi}(x)T_\rho(e_{-i\xi}f)(x). \tag{7-4}$$

Suppose that  $g \in L^2(\mathbb{S}^2)$  takes the form  $g(x) = \int_H a(\xi)e_{i\xi}(x) d\nu(\xi)$ , where  $H \subset \mathbb{R}^3$  is a two-dimensional subspace,  $\nu$  is Lebesgue measure on  $H$ , and  $a \in L^2(H)$ . Then

$$(f\sigma * g\sigma)(x) = c|x|^{-1} \int_H a(\xi)e_{i\xi}(x)T_\rho(e_{-i\xi}f)(x) d\nu(\xi).$$

For  $t \in (0, 2)$ , define  $\rho(t) > 0$  by

$$\rho(t)^2 + (t/2)^2 = 1.$$

Then for any interval  $I \subset (0, 2)$ ,

$$\begin{aligned} \int_{|x| \in I} |(f\sigma * g\sigma)(x)|^2 dx &\leq C \int_I t^{-2} \left\| \int_H |a(\zeta)| \cdot |T_{\rho(t)}(e_{-i\zeta}f)| d\zeta \right\|_{L^2(\mathbb{S}^2)}^2 t^2 dt \\ &= C \int_I \left\| \int_H |a(\zeta)| \cdot |T_{\rho(t)}(e_{-i\zeta}f)| d\zeta \right\|_{L^2(\mathbb{S}^2)}^2 dt. \end{aligned} \tag{7-5}$$

**Fourier coefficient estimates in terms of the spherical Laplacian.** The following routine lemma is convenient because it provides an intrinsic characterization of expressions that arise in the analysis. The proof relies on the machinery of pseudodifferential operators, and is left to the reader.

**Lemma 7.8.** *Let  $\mathcal{C}$  be a cap of radius  $\varrho \leq \frac{1}{2}$ . Let  $\phi$  be the rescaling map associated with  $\mathcal{C}$ . Let  $f$  be supported in  $\mathcal{C} \cup (-\mathcal{C})$ . Then for any  $t \in \mathbb{R}$  and  $0 < r \leq \varrho$ ,*

$$C^{-1} \|(I - r^2\Delta)^{t/2} f\|_{L^2(\mathbb{S}^2)}^2 \leq \int_{\mathbb{R}^2} |\widehat{\phi^* f}(\xi)|^2 (1 + |r\varrho^{-1}\xi|^2)^t d\xi \leq C \|(I - r^2\Delta)^{t/2} f\|_{L^2(\mathbb{S}^2)}^2.$$

Here  $C \in (0, \infty)$  depends on  $t$  but not on  $f, r, \varrho, \mathcal{C}$ .

### 8. Step 6A: A decomposition algorithm

The following iterative procedure may be applied to any nonnegative function  $f \in L^2(\mathbb{S}^2)$  of positive norm.

**Decomposition algorithm.** Initialize by setting  $G_0 = f$ , and  $\varepsilon_0 = 1/2$ .

*Step  $\nu$ .* The inputs for Step  $\nu$  are a nonnegative function  $G_\nu \in L^2(\mathbb{S}^2)$  and a positive number  $\varepsilon_\nu$ . Its outputs are functions  $f_\nu$  and  $G_{\nu+1}$  and nonnegative numbers  $\varepsilon_\nu^*$  and  $\varepsilon_{\nu+1}$ . If  $\|G_\nu \sigma * G_\nu \sigma\|_2 = 0$ , then  $G_\nu = 0$  almost everywhere. The algorithm then terminates, and we define  $\varepsilon_\nu^* = 0$  and  $f_\nu = 0$ , and  $G_\mu = f_\mu = 0$  and  $\varepsilon_\mu = 0$  for all  $\mu > \nu$ .

If  $0 < \|G_\nu \sigma * G_\nu \sigma\|_2 < \varepsilon_\nu^2 \mathbf{S}^2 \|f\|_2^2$ , then replace  $\varepsilon_\nu$  by  $\varepsilon_\nu/2$ , and repeat until the first time that  $\|G_\nu \sigma * G_\nu \sigma\|_2 \geq \varepsilon_\nu^2 \mathbf{S}^2 \|f\|_2^2$ . Define  $\varepsilon_\nu^*$  to be this value of  $\varepsilon_\nu$ . Then

$$(\varepsilon_\nu^*)^2 \mathbf{S}^2 \|f\|_2^2 \leq \|G_\nu \sigma * G_\nu \sigma\|_2 \leq 4(\varepsilon_\nu^*)^2 \mathbf{S}^2 \|f\|_2^2.$$

Apply Lemma 2.9 to obtain a cap  $\mathcal{C}_\nu$  and a decomposition  $G_\nu = f_\nu + G_{\nu+1}$  with disjointly supported nonnegative summands satisfying  $f_\nu \leq C_\nu \|f\|_2 |\mathcal{C}_\nu|^{-1/2} \chi_{\mathcal{C}_\nu}$ , and  $\|f_\nu\|_2 \geq \eta_\nu \|f\|_2$ . Here  $C_\nu, \eta_\nu$  are bounded above and below, respectively, by quantities that depend only on  $\|G_\nu \sigma * G_\nu \sigma\|_2^{1/2} / \|G_\nu\|_2 \geq \varepsilon_\nu^*$ . Define  $\varepsilon_{\nu+1} = \varepsilon_\nu^*$ , and move on to step  $\nu + 1$ . □

It is important for our application to observe that if  $f$  is even then at every step,  $f_\nu$  may likewise be chosen to be even. The upper bound for  $f_\nu$  then becomes

$$f_\nu \leq C_\nu |\mathcal{C}_\nu|^{-1/2} \chi_{\mathcal{C}_\nu \cup -\mathcal{C}_\nu}.$$

Henceforth the algorithm will be applied only to even functions, and we will always choose all  $f_\nu$  to be even.

If the algorithm terminates at some finite step  $\nu$ , then a finite decomposition  $f = \sum_{k=0}^\nu f_k$  results.

**Lemma 8.1.** *Let  $f \in L^2(\mathbb{S}^2)$  be a nonnegative function with positive norm. If the decomposition algorithm never terminates for  $f$ , then  $\varepsilon_\nu^* \rightarrow 0$  as  $\nu \rightarrow \infty$ , and  $\sum_{\nu=0}^N f_\nu \rightarrow f$  in  $L^2$  as  $N \rightarrow \infty$ .*

*Proof.* Assume without loss of generality that  $\|f\|_2 = 1$ . The functions  $f_\nu$  have disjoint supports and hence are pairwise orthogonal, and  $\sum_\nu f_\nu \leq f$ , so  $\sum_\nu \|f_\nu\|_2^2 \leq \|f\|_2^2$ . Since the sequence  $\varepsilon_\nu^*$  is nonincreasing and  $\|f_\nu\|_2 / \|f\|_2$  is bounded below by a function of  $\varepsilon_\nu^*$ , this forces  $\varepsilon_\nu^* \rightarrow 0$ .

The second conclusion is equivalent to  $\|G_N\|_2 \rightarrow 0$ . According to Lemma 2.9,  $\|f_\nu\|_2$  is bounded below by a function of  $\|G_\nu \sigma * G_\nu \sigma\|_2$ . Since  $\sum_\nu \|f_\nu\|_2^2 < \infty$ , we have  $\|f_\nu\|_2 \rightarrow 0$  and therefore  $\|G_\nu \sigma * G_\nu \sigma\|_2 \rightarrow 0$ . By construction,  $G_{\nu+1}(x) \leq G_\nu(x)$  for every  $x \in \mathbb{S}^2$ , so  $G(x) = \lim_{\nu \rightarrow \infty} G_\nu(x)$  exists and  $\|G \sigma * G \sigma\|_2 \leq \|G_\nu \sigma * G_\nu \sigma\|_2$  for all  $\nu$ . Thus  $G \sigma * G \sigma \equiv 0$ , so  $G \equiv 0$ . This forces  $\|G_\nu\|_2 \rightarrow 0$ , by the dominated convergence theorem. □

For general  $f$ , this decomposition may be highly inefficient. But if  $f$  is nearly extremal for the inequality (2-1), then more useful properties hold.

**Lemma 8.2.** *There exists a continuous function  $\theta : (0, 1] \rightarrow (0, \infty)$  such that for any  $\varepsilon > 0$  there exists  $\delta > 0$  such that for any  $\delta$ -nearly extremal nonnegative function  $f \in L^2(\mathbb{S}^2)$  satisfying  $\|f\|_2 = 1$ , the functions  $f_\nu$  and  $G_\nu$  associated by the decomposition algorithm to  $f$  satisfy*

$$\|f_\nu\|_2 \geq \theta(\|G_\nu\|_2) \quad \text{for any index } \nu \text{ such that } \|G_\nu\|_2 \geq \varepsilon.$$

This is a direct consequence of Lemmas 2.9 and 7.1. It is essential for applications below that  $\theta$  be independent of  $\varepsilon$ .

If  $f$  is nearly extremal, then the norms of  $f_\nu$  and  $G_\nu$  enjoy upper bounds independent of  $f$ , for all except very large  $\nu$ .

**Lemma 8.3.** *There exist a sequence of positive constants  $\gamma_\nu \rightarrow 0$  and a function  $N : (0, \frac{1}{2}] \rightarrow \mathbb{Z}^+$  satisfying  $N(\delta) \rightarrow \infty$  as  $\delta \rightarrow 0$  such that for any nonnegative  $f \in L^2(\mathbb{S}^2)$ , if  $f$  is  $\delta$ -nearly extremal then the quantities  $\varepsilon_\nu^*$  obtained when the decomposition algorithm is applied to  $f$  satisfy*

$$\begin{aligned} \varepsilon_\nu^* &\leq \gamma_\nu && \text{for all } \nu \leq N(\delta), \\ \|G_\nu\|_2 &\leq \gamma_\nu \|f\|_2 && \text{for all } \nu \leq N(\delta), \\ \|f_\nu\|_2 &\leq \gamma_\nu \|f\|_2 && \text{for all } \nu \leq N(\delta). \end{aligned}$$

This holds whether or not the algorithm terminates for  $f$ .

*Proof.*  $\mathbf{S}^2 \|G_\nu\|_2^2 \geq \|G_\nu \sigma * G_\nu \sigma\|_2 \geq \varepsilon_\nu^{*2} \mathbf{S}^2 \|f\|_2^2 = (\varepsilon_\nu^{*2} \|f\|_2^2 / \|G_\nu\|_2^2) \mathbf{S}^2 \|G_\nu\|_2^2,$

so  $\varepsilon_\nu^* \leq \|G_\nu\|_2 / \|f\|_2$ . Thus the second conclusion implies the first. Since  $\|f_\nu\|_2 \leq \|G_\nu\|_2$ , it also implies the third.

We recall two facts. First, Lemma 7.1, applied to  $h = G_\nu$  and  $g = f_0 + \dots + f_{\nu-1}$ , asserts that there are constants  $c_0, C_1 \in \mathbb{R}^+$  such that if  $f \in L^2$  is  $\delta$ -nearly extremal, either  $\|G_\nu \sigma * G_\nu \sigma\|_2 \geq c_0 \|G_\nu\|_2^4 \|f\|_2^{-2}$  or  $\|G_\nu\|_2 \leq C_1 \delta^{1/2} \|f\|_2$ . Second, according to Lemma 2.9, there exists a nondecreasing function  $\rho : (0, \infty) \rightarrow (0, \infty)$  satisfying  $\rho(t) \rightarrow 0$  as  $t \rightarrow 0$  such that for every nonzero  $f \in L^2$  and any  $\nu$ , if  $\|G_\nu \sigma * G_\nu \sigma\|_2 \geq t \|G_\nu\|_2^2$ , then  $\|f_\nu\|_2^2 \geq \rho(t) \|G_\nu\|_2^2$ .

Choose a sequence  $\{\gamma_\nu\}$  of positive numbers that tends monotonically to zero, but does so sufficiently slowly to satisfy

$$\nu \gamma_\nu^2 \rho(c_0 \gamma_\nu^2) > 1 \quad \text{for all } \nu.$$

Define  $N(\delta)$  to be the largest integer satisfying  $\gamma_{N(\delta)} \geq C_1 \delta^{1/2}$ . This  $N(\delta) \rightarrow \infty$  as  $\delta \rightarrow 0$  because  $\gamma_\nu > 0$  for all  $\nu$ .

Let  $f$  and  $\delta$  be given. Suppose that  $\nu \leq N(\delta)$ . We argue by contradiction, supposing that  $\|G_\nu\|_2 > \gamma_\nu \|f\|_2$ . Then  $\|G_\nu\|_2 > C_1 \delta^{1/2} \|f\|_2$  by definition of  $N(\delta)$ . By the dichotomy above,

$$\|G_\nu \sigma * G_\nu \sigma\|_2 \geq c_0 \|G_\nu\|_2^4 \|f\|_2^{-2} \geq c_0 \gamma_\nu^2 \|G_\nu\|_2^2.$$

By the second fact reviewed above,

$$\|f_\nu\|_2^2 \geq \rho(c_0 \gamma_\nu^2) \|G_\nu\|_2^2 \geq \gamma_\nu^2 \rho(c_0 \gamma_\nu^2) \|f\|_2^2.$$

Since  $\|G_\mu\|_2 \geq \|G_\nu\|_2$  for all  $\mu \leq \nu$ , the same lower bound follows for  $\|f_\nu\|_2^2$  for all  $\mu \leq \nu$ . Since the functions  $f_\mu$  are pairwise orthogonal,  $\sum_{\mu \leq \nu} \|f_\mu\|_2^2 \leq \|f\|_2^2$ , and consequently  $\nu\gamma_\nu^2 \rho(c_0\gamma_\nu^2) \leq 1$ , a contradiction.  $\square$

The next lemma also follows directly from the decomposition algorithm coupled with Lemma 2.9.

**Lemma 8.4.** *For any  $\varepsilon > 0$  there exist  $\delta_\varepsilon > 0$  and  $C_\varepsilon < \infty$  such that for every  $\delta_\varepsilon$ -nearly extremal nonnegative function  $f \in L^2$ , the functions  $f_\nu$  and  $G_\nu$  associated to  $f$  by the decomposition algorithm satisfy*

(i) *For any  $\nu$ , if  $\|G_\nu\|_2 \geq \varepsilon\|f\|_2$ , then there exists a cap  $\mathcal{C}_\nu \subset \mathbb{S}^2$  such that*

$$f_\nu \leq C_\varepsilon \|f\|_2 |\mathcal{C}_\nu|^{-1/2} \chi_{\mathcal{C}_\nu \cup -\mathcal{C}_\nu}.$$

(ii) *If  $\|G_\nu\|_2 \geq \varepsilon\|f\|_2$ , then  $\|f_\nu\|_2 \geq \delta_\varepsilon\|f\|_2$ .*

### 9. Step 6B: A geometric property of the decomposition

We have established inequalities concerning the  $L^2$  norms of the functions  $f_\nu$  and  $G_\nu$  that the decomposition algorithm yields, based on quite general principles and a single analytic fact, Lemma 2.9, concerning the particular inequality that we are studying. We next establish an additional inequality of a geometric nature, based on a single additional fact, the weak interaction of distant caps in the sense of Lemma 7.6.

**Lemma 9.1.** *In any metric space, for any  $N$  and  $r$ , any finite set  $S$  of cardinality  $N$  and diameter equal to  $r$  may be partitioned into two disjoint nonempty subsets  $S = S' \cup S''$  such that  $\text{distance}(S', S'') \geq r/2N$ . Moreover, given two points  $s', s'' \in S$  satisfying  $\text{distance}(s', s'') = r$ , this partition can be constructed so that  $s' \in S'$  and  $s'' \in S''$ .*

*Proof.* Consider the metric balls  $B_k$  centered at  $s'$  of radii  $kr/2N$  for  $k = 1, 2, \dots, 2N$ . By the pigeonhole principle, there exists  $k$  such that  $(B_{k+1} \setminus B_k) \cap S = \emptyset$ . Set  $S' = B_k \cap S$  and  $S'' = S \setminus S'$ . The triangle inequality yields the conclusion.  $\square$

**Lemma 9.2.** *For any  $\varepsilon > 0$  there exist  $\delta > 0$  and  $\lambda < \infty$  such that for any  $0 \leq f \in L^2(\mathbb{S}^2)$  that is  $\delta$ -nearly extremal, the summands  $f_\nu$  produced by the decomposition algorithm and the associated caps  $\mathcal{C}_\nu$  satisfy*

$$\varrho(\mathcal{C}_j, \mathcal{C}_k) \leq \lambda \quad \text{whenever } \|f_j\|_2 \geq \varepsilon\|f\|_2 \text{ and } \|f_k\|_2 \geq \varepsilon\|f\|_2.$$

Here  $\varrho$  is the distance between  $\mathcal{C}_j \cup -\mathcal{C}_j$  and  $\mathcal{C}_k \cup -\mathcal{C}_k$ , as defined in Definition 7.5.

*Proof.* It suffices to prove this for all sufficiently small  $\varepsilon$ . Let  $f$  be a nonnegative  $L^2$  function that satisfies  $\|f\|_2 = 1$  and is  $\delta$ -nearly extremal for a sufficiently small  $\delta = \delta(\varepsilon)$ , and let  $\{G_\nu, f_\nu\}$  be associated to  $f$  via the decomposition algorithm. Set  $F = \sum_{\nu=0}^N f_\nu$ .

Suppose that  $\|f_{j_0}\|_2 \geq \varepsilon$  and  $\|f_{k_0}\|_2 \geq \varepsilon$ . Let  $N$  be the smallest integer such that  $\|G_{N+1}\|_2 < \varepsilon^3$ . Since  $\|G_\nu\|_2$  is a nonincreasing function of  $\nu$ , and since  $\|f_\nu\|_2 \leq \|G_\nu\|_2$ , necessarily  $j_0, k_0 \leq N$ . Moreover, by Lemma 8.3, there exists  $M_\varepsilon < \infty$  depending only on  $\varepsilon$  such that  $N \leq M_\varepsilon$ . By Lemma 8.4, if  $\delta$  is chosen to be a sufficiently small function of  $\varepsilon$ , then since  $\|G_\nu\|_2 \geq \varepsilon^3$  for all  $\nu \leq N$ , we have  $f_\nu \leq \theta(\varepsilon)|\mathcal{C}_\nu|^{-1/2} \chi_{\mathcal{C}_\nu \cup -\mathcal{C}_\nu}$  for all such  $\nu$ , where  $\theta$  is a continuous, strictly positive function on  $(0, 1]$ .



Now let  $\lambda < \infty$  be a large quantity to be specified. It suffices to show that if  $\delta(\varepsilon)$  is sufficiently small, an assumption that  $\varrho(\mathcal{C}_j, \mathcal{C}_k) > \lambda$  implies an upper bound, which depends only on  $\varepsilon$ , for  $\lambda$ .

Lemma 9.1 yields a decomposition  $F = F_1 + F_2 = \sum_{v \in S_1} f_v + \sum_{v \in S_2} f_v$ , where  $[0, N] = S_1 \cup S_2$  is a partition of  $[0, N]$ ,  $j_0 \in S_1$ ,  $k_0 \in S_2$ , and  $\varrho(\mathcal{C}_j, \mathcal{C}_k) \geq \lambda/2N \geq \lambda/2M_\varepsilon$  for all  $j \in S_1$  and  $k \in S_2$ . Certainly  $\|F_1\|_2 \geq \|f_{j_0}\|_2 \geq \varepsilon$  and similarly  $\|F_2\|_2 \geq \varepsilon$ . The convolution cross term satisfies

$$\|F_1\sigma * F_2\sigma\|_2 \leq \sum_{j \in S_1} \sum_{k \in S_2} \|f_j\sigma * f_k\sigma\|_2 \leq M_\varepsilon^2 \gamma(\lambda/2M_\varepsilon) \theta(\varepsilon)^2,$$

where  $\gamma(\lambda) \rightarrow 0$  as  $\lambda \rightarrow \infty$  by Lemma 7.6. Therefore

$$\begin{aligned} \|F\sigma * F\sigma\|_2^2 &\leq \|F_1\sigma * F_1\sigma\|_2^2 + \|F_2\sigma * F_2\sigma\|_2^2 + C\|f\|_2^2 \|F_1\sigma * F_2\sigma\|_2 \\ &\leq \mathbf{S}^4 \|F_1\|_2^4 + \mathbf{S}^4 \|F_2\|_2^4 + M_\varepsilon^2 \gamma(\lambda/2M_\varepsilon) \theta(\varepsilon)^2. \end{aligned}$$

Since  $F_1$  and  $F_2$  have disjoint supports,  $\|F_1\|_2^2 + \|F_2\|_2^2 \leq \|f\|_2^2 = 1$  and consequently

$$\|F_1\|_2^4 + \|F_2\|_2^4 \leq \max(\|F_1\|_2^2, \|F_2\|_2^2) \cdot (\|F_1\|_2^2 + \|F_2\|_2^2) \leq (1 - \varepsilon^2) \cdot 1 \leq 1 - \varepsilon^2.$$

Thus

$$\|F\sigma * F\sigma\|_2^2 \leq \mathbf{S}^4(1 - \varepsilon^2) + M_\varepsilon^2 \gamma(\lambda/2M_\varepsilon) \theta(\varepsilon)^2.$$

Therefore

$$\begin{aligned} (1 - \delta)^2 \mathbf{S}^2 &\leq \|f\sigma * f\sigma\|_2 \leq \|F\sigma * F\sigma\|_2 + C\|f\|_2 \|f - F\|_2 \\ &\leq \|F\sigma * F\sigma\|_2 + C\varepsilon^3, \end{aligned}$$

so by transitivity

$$(1 - \delta)^4 \mathbf{S}^4 \leq C\varepsilon^3 + \mathbf{S}^4(1 - \varepsilon^2) + M_\varepsilon^2 \gamma(\lambda/2M_\varepsilon) \theta(\varepsilon)^2.$$

Since  $\gamma(t) \rightarrow 0$  as  $t \rightarrow \infty$ , for all sufficiently small  $\varepsilon > 0$  this implies an upper bound, which depends only on  $\varepsilon$ , for  $\lambda$ , as was to be proved.  $\square$

## 10. Step 6C: Upper bounds for extremizing sequences

Proposition 2.14 states that any nearly extremal function satisfies appropriately scaled upper bounds relative to some cap. It is convenient for the proof to first observe that a superficially weaker statement implies the version stated.

**Lemma 10.1.** *There exists a function  $\Theta : [1, \infty) \rightarrow (0, \infty)$  satisfying  $\Theta(R) \rightarrow 0$  as  $R \rightarrow \infty$  with the following property. For any  $\varepsilon > 0$  and  $\bar{R} \in [1, \infty)$  there exists  $\delta > 0$  such that any nonnegative even function  $f$  that has  $\|f\|_2 = 1$  and is  $\delta$ -nearly extremal may be decomposed as  $f = F + G$ , where  $F$  and  $G$  are even and nonnegative with disjoint supports,  $\|G\|_2 < \varepsilon$ , and there exists a cap  $\mathcal{C} = \mathcal{C}(z, r)$*

such that for any  $R \in [1, \bar{R}]$ ,

$$\int_{\min(|x-z|, |x+z|) \geq Rr} F^2(x) d\sigma(x) \leq \Theta(R), \tag{10-1}$$

$$\int_{F(x) \geq Rr^{-1}} F^2(x) d\sigma(x) \leq \Theta(R). \tag{10-2}$$

*Proof that Lemma 10.1 implies Proposition 2.14.* Let  $\Theta$  be the function promised by the lemma. Let  $\varepsilon$  and  $f$  be given, and assume without loss of generality that  $\varepsilon$  is small. Assuming as we may that  $\Theta$  is a continuous, strictly decreasing function, define  $\bar{R} = \bar{R}(\varepsilon)$  by the equation  $\Theta(\bar{R}) = \varepsilon^2/2$ . Let  $\mathcal{C} = \mathcal{C}(z, r)$  and suppose  $\delta = \delta(\varepsilon, \bar{R}(\varepsilon))$  along with  $F$  and  $G$  satisfy the conclusions of the lemma relative to  $\varepsilon$  and  $\bar{R}(\varepsilon)$ . Define  $\chi$  to be the characteristic function of the set of all  $x \in \mathbb{S}^2$  that satisfy either  $\min(|x-z|, |x+z|) \geq \bar{R}r$  or  $F(x) > \bar{R}|\mathcal{C}|^{-1/2}$ . Redecompose  $f = \tilde{F} + \tilde{G}$ , where  $\tilde{F} = (1 - \chi)F$  and  $\tilde{G} = G + \chi F$ . Then  $\|\tilde{G}\|_2 < 2\varepsilon$ , while  $\tilde{F}$  satisfies the required inequalities. For instance, if  $R \leq \bar{R}$  then

$$\int_{\tilde{F}(x) \geq R|\mathcal{C}|^{-1/2}} \tilde{F}(x)^2 d\sigma(x) \leq \int_{F(x) \geq R|\mathcal{C}|^{-1/2}} F(x)^2 d\sigma(x) \leq \Theta(R),$$

while the integrand vanishes if  $R > \bar{R}$ . □

*Proof of Lemma 10.1.* Let  $\eta : [1, \infty) \rightarrow (0, \infty)$  be a function to be chosen below, satisfying  $\eta(t) \rightarrow 0$  as  $t \rightarrow \infty$ . This function will not depend on the quantity  $\bar{R}$ .

Let  $\bar{R} \geq 1$ ,  $R \in [1, \bar{R}]$ , and  $\varepsilon > 0$  be given. Let  $\delta = \delta(\varepsilon, \bar{R}) > 0$  be a small quantity to be chosen below. Let  $0 \leq f \in L^2(\mathbb{S}^2)$  be even and  $\delta$ -nearly extremal. It is no loss of generality in normalizing such that  $\|f\|_2 = 1$ .

Let  $\{f_\nu\}$  be the sequence of functions obtained by applying the decomposition algorithm to  $f$ . Choose  $\delta = \delta(\varepsilon) > 0$  sufficiently small and  $M = M(\varepsilon)$  sufficiently large to guarantee that  $\|G_{M+1}\|_2 < \varepsilon/2$  and that  $f_\nu$  and  $G_\nu$  satisfy all conclusions of Lemma 8.4 and Lemma 8.3 for  $\nu \leq M$ . Set  $F = \sum_{\nu=0}^M f_\nu$ . Then  $\|f - F\|_2 = \|G_{M+1}\|_2 < \varepsilon/2$ .

Let  $N \in \{0, 1, 2, \dots\}$  be the minimum of  $M$  and the smallest number such that  $\|f_{N+1}\|_2 < \eta$ .  $N$  is majorized by a quantity that depends only on  $\eta$ . Set  $\mathcal{F} = \mathcal{F}_N = \sum_{k=0}^N f_k$ . It follows from Lemma 8.4(ii) that

$$\|F - \mathcal{F}\|_2 < \gamma(\eta), \quad \text{where } \gamma(\eta) \rightarrow 0 \text{ as } \eta \rightarrow 0. \tag{10-3}$$

This function  $\gamma$  is independent of  $\varepsilon$  and  $\bar{R}$ .

To prove the lemma, we must produce an appropriate cap  $\mathcal{C} = \mathcal{C}(z, r)$ , and must establish the existence of  $\Theta$ . To do the former is simple: To  $f_0$  is associated a cap  $\mathcal{C}_0 = \mathcal{C}(z_0, r_0)$  such that  $f_0 \leq C|\mathcal{C}_0|^{-1/2}(\chi_{\mathcal{C}_0} + \chi_{-\mathcal{C}_0})$ . Then  $\mathcal{C} = \mathcal{C}_0$  is the required cap. Note that by Lemma 2.9,  $\|f_0\|_2 \geq c$  for some positive universal constant  $c$ .

Suppose that functions  $R \mapsto \eta(R)$  and  $R \mapsto \Theta(R)$  are chosen so that

$$\eta(R) \rightarrow 0 \text{ as } R \rightarrow \infty \quad \text{and} \quad \gamma(\eta(R)) \leq \Theta(R) \text{ for all } R.$$

Then by (10-3),  $F - \mathcal{F}$  already satisfies the desired inequalities in  $L^2(\mathbb{S}^2)$ , so it suffices to show that  $\mathcal{F}(x) \equiv 0$  whenever  $\min(|x - z|, |x + z|) > Rr_0$ , and that  $\|\mathcal{F}\|_\infty \leq R|\mathcal{C}_0|^{-1/2}$ .

Each summand satisfies  $f_k \leq C(\eta)|\mathcal{C}_k|^{-1/2}\chi_{\mathcal{C}_k \cup -\mathcal{C}_k}$ , where  $C(\eta) < \infty$  depends only on  $\eta$ , and in particular,  $f_k$  is supported in  $\mathcal{C}_k \cup -\mathcal{C}_k$ . Now  $\|f_k\|_2 \geq \eta$  for all  $k \leq N$  by definition of  $N$ . Therefore by Lemma 9.2, there exists a function  $\eta \mapsto \lambda(\eta) < \infty$  such that if  $\delta$  is sufficiently small as a function of  $\eta$ , then  $\varrho(\mathcal{C}_k, \mathcal{C}_0) \leq \lambda(\eta)$  for all  $k \leq N$ . This is needed for  $\eta = \eta(R)$  for all  $R$  in the compact set  $[1, \bar{R}]$ , so such a  $\delta$  may be chosen as a function of  $\bar{R}$  alone; conditions already imposed on  $\delta$  above make it a function of both  $\varepsilon$  and  $\bar{R}$ .

In the region of all  $x \in \mathbb{S}^2$  satisfying  $\min(|x - z_0|, |x + z_0|) > Rr_0$ , either  $f_k \equiv 0$ , or  $\mathcal{C}_k$  has radius no less than  $\frac{1}{4}Rr_0$ , or the center  $z_k$  of  $\mathcal{C}_k$  satisfies  $\max(|z_k - z_0|, |z_k + z_0|) \geq \frac{1}{4}Rr_0$ . Choose a function  $R \mapsto \eta(R)$  that tends to 0 sufficiently slowly as  $R \rightarrow \infty$  to ensure that  $\lambda(\eta(R)) \rightarrow \infty$  sufficiently slowly that the latter two cases would contradict the inequality  $\varrho(\mathcal{C}_k, \mathcal{C}_0) \leq \lambda$ , and therefore cannot arise. Then  $\mathcal{F}(x) \equiv 0$  when  $\min(|x - z_0|, |x + z_0|) > Rr_0$ .

With the function  $\eta$  specified,  $\Theta$  can be defined by

$$\Theta(R) = \gamma(\eta(R)). \quad (10-4)$$

Then (10-1) holds for all  $R \in [1, \bar{R}]$ .

We claim next that  $\|\mathcal{F}\|_\infty < R|\mathcal{C}_0|^{-1/2}$  if  $R$  is sufficiently large as a function of  $\eta$ . Indeed, because the summands  $f_k$  have pairwise disjoint supports, it suffices to control  $\max_{k \leq N} \|f_k\|_\infty$ . Again, by Lemma 8.4,  $\|f_k\|_\infty \leq C(\eta)|\mathcal{C}_k|^{-1/2}$ . If  $\eta(R)$  is chosen to tend to zero sufficiently slowly as  $R \rightarrow \infty$  to ensure that  $C(\eta(R))\lambda(\eta(R)) < R$  for all  $k \leq N$ , then inequality (10-2) holds provided that  $\Theta$  is defined by (10-4).

The final function  $\eta$  must be chosen to tend to zero slowly enough to satisfy the requirements of the proofs of both (10-1) and (10-2).  $\square$

## 11. Preliminaries for Step 7

**Lemma 11.1.** *Let  $\Theta : [1, \infty) \rightarrow (0, \infty)$  satisfy  $\Theta(R) \rightarrow 0$  as  $R \rightarrow \infty$ . Let  $\delta > 0$ . Then there exists  $c > 0$  such that any nonnegative function  $g \in L^2(\mathbb{R}^2)$  satisfying  $\|g\|_2 = 1$  and the upper bounds*

$$\int_{|x| \geq R} g(x)^2 dx + \int_{g(x) \geq R} g(x)^2 dx \leq \Theta(R) \quad \text{for all } R \geq 1,$$

*has Fourier transform satisfying the lower bound*

$$\int_{|\xi| \leq \delta} |\hat{g}(\xi)|^2 d\xi \geq c.$$

*Proof.* Let  $g \in L^2(\mathbb{R}^2)$  satisfy the hypotheses. For  $t > 0$ , let  $\varphi_t(y) = e^{-t|y|^2/2}$ . Then

$$\int g \varphi_t dy = (2\pi)^{-2} \int \hat{g}(\xi) \widehat{\varphi}_t(\xi) d\xi = (2\pi)^{-1} t^{-1} \int \hat{g}(\xi) e^{-|\xi|^2/2t} d\xi.$$

For any  $R, \rho \geq 1$ , let  $S = \{y : |y| \leq R \text{ and } g(y) \leq \rho\}$ . Provided that  $R$  and  $\rho$  are chosen to be sufficiently large that  $\Theta(R) + \Theta(\rho) \leq \frac{1}{2}$ ,

$$\begin{aligned} \int_{\mathbb{R}^2} g \varphi_t \, dy &\geq e^{-tR^2/2} \int_S g(y) \, dy \geq e^{-tR^2/2} \rho^{-1} \int_S g^2(y) \, dy \\ &= e^{-tR^2/2} \rho^{-1} \left( \|g\|_2^2 - \int_{\mathbb{R}^2 \setminus S} g^2(y) \, dy \right) \geq \frac{1}{2} e^{-tR^2/2} \rho^{-1} \end{aligned}$$

for any  $t > 0$ . On the other hand, by the Cauchy–Schwarz inequality

$$\begin{aligned} \int_{|\xi| \geq \delta} |\hat{g}(\xi)| t^{-1} e^{-|\xi|^2/2t} \, d\xi &\leq \pi^{1/2} t^{-1} \|\hat{g}\|_2 \left( \int_{r=\delta}^{\infty} e^{-r^2/t} 2r \, dr \right)^{1/2} \\ &= \pi^{1/2} t^{-1} \left( t \int_{s=\delta^2/t}^{\infty} e^{-s} \, ds \right)^{1/2} = \pi^{1/2} t^{-1/2} e^{-\delta^2/2t}. \end{aligned}$$

The Cauchy–Schwarz inequality also gives

$$\begin{aligned} \int_{|\xi| \leq \delta} |\hat{g}(\xi)| t^{-1} e^{-|\xi|^2/2t} \, d\xi &\leq \left( \int_{|\xi| \leq \delta} |\hat{g}(\xi)|^2 \, d\xi \right)^{1/2} (2\pi)^{1/2} \left( \int_0^{\infty} t^{-2} e^{-r^2/t} r \, dr \right)^{1/2} \\ &= \pi^{1/2} t^{-1/2} \left( \int_{|\xi| \leq \delta} |\hat{g}(\xi)|^2 \, d\xi \right)^{1/2}. \end{aligned}$$

Therefore

$$\begin{aligned} \pi^{1/2} t^{-1/2} \left( \int_{|\xi| \leq \delta} |\hat{g}(\xi)|^2 \, d\xi \right)^{1/2} &\geq \int_{\mathbb{R}^2} \hat{g}(\xi) t^{-1} e^{-|\xi|^2/2t} \, d\xi - \int_{|\xi| \geq \delta} |\hat{g}(\xi)| t^{-1} e^{-|\xi|^2/2t} \, d\xi \\ &\geq \pi e^{-tR^2/2} \rho^{-1} - \pi^{1/2} t^{-1/2} e^{-\delta^2/2t}. \end{aligned}$$

Now substitute  $t = \delta^2/\gamma$ , where  $\gamma = \gamma(\delta) \geq 1$ , to obtain

$$\pi^{1/2} \gamma^{1/2} \delta^{-1} \left( \int_{|\xi| \leq \delta} |\hat{g}(\xi)|^2 \, d\xi \right)^{1/2} \geq \pi e^{-\delta^2 R^2/2\gamma} \rho^{-1} - \pi^{1/2} \gamma^{1/2} \delta^{-1} e^{-\gamma/2}.$$

The quantities  $R$  and  $\rho$  have already been fixed, independent of  $\delta$ . As  $\delta$  also remains fixed while  $\gamma \rightarrow \infty$ , this last lower bound tends to  $\pi \rho^{-1} - 0 > 0$ . Thus choosing  $\gamma$  sufficiently large yields the desired lower bound. □

**Lemma 11.2.** *Let  $c_0 > 0$ . Let  $\{g_\nu\}$  be any sequence of functions in  $L^2(\mathbb{R}^2)$  satisfying  $\|g_\nu\|_{L^2} = 1$  and  $\int_{|\xi| \leq 1} |\hat{g}_\nu(\xi)|^2 \, d\xi \geq c_0$ . Then either there exists a function  $\theta : [1, \infty) \rightarrow (0, \infty)$  satisfying*

$$\theta(s) \rightarrow 0 \quad \text{as } s \rightarrow \infty$$

such that

$$\int_{|\xi| \geq s} |\hat{g}_\nu(\xi)|^2 \, d\xi \leq \theta(s) \quad \text{for all } s \in [1, \infty) \text{ and all } \nu,$$

or there exist a subsequence  $\nu_k \rightarrow \infty$  and real constants  $\delta > 0$ ,  $\varepsilon_k > 0$ , and  $S_k \geq s_k \geq 1$  such that  $s_k \rightarrow \infty$ ,  $\varepsilon_k \rightarrow 0$ ,  $S_k = s_k^3$ ,

$$\int_{|\xi| \leq s_k} |\widehat{g_{\nu_k}}(\xi)|^2 d\xi \geq \delta, \quad \int_{|\xi| \geq S_k} |\widehat{g_{\nu_k}}(\xi)|^2 d\xi \geq \delta, \quad \int_{s_k \leq |\xi| \leq S_k} |\widehat{g_{\nu_k}}(\xi)|^2 d\xi < \varepsilon_k.$$

In this lemma,  $\delta$  is permitted, in principle, to depend on  $\{g_\nu\}$ , and  $\varepsilon_k$  and  $s_k$  are permitted to depend on  $\{g_\nu\}$  and on  $k$  in an arbitrary manner, provided only that they satisfy the stated conditions. The relation  $S_k = s_k^3$  is chosen simply because it is convenient for the proof of Lemma 12.2 below; one could arrange to have  $S_k$  equal to any function of  $s_k$  that might be desired.

*Proof.* Define a sequence  $\rho_1, \rho_2, \dots$  by  $\rho_1 = 2$  and by induction,  $\rho_{j+1} = \rho_j^3$ . If the first conclusion does not hold, then after passing to a subsequence and renumbering, we have

$$\int_{|\xi| \geq \rho_\nu} |\widehat{g_\nu}(\xi)|^2 d\xi \geq \delta \quad \text{for all } \nu.$$

Consider a large  $\nu$ . Since

$$\sum_{j=1}^{\nu-1} \int_{\rho_j \leq |\xi| \leq \rho_{j+1}} |\widehat{g_\nu}(\xi)|^2 d\xi \leq (2\pi)^2 \|g_\nu\|_2^2 \leq (2\pi)^2$$

and there are  $\nu - 1$  summands, there must exist  $j(\nu)$  satisfying

$$\int_{\rho_j \leq |\xi| \leq \rho_{j+1}} |\widehat{g_\nu}(\xi)|^2 d\xi \leq C\nu^{-1}.$$

It suffices to set  $s_\nu = \rho_{j(\nu)}$ ,  $S_\nu = \rho_{j(\nu)+1} = s_\nu^3$ , and  $\varepsilon_\nu = C\nu^{-1}$ . □

## 12. Step 7: Precompactness after rescaling

We begin the proof of Proposition 2.15. Let  $\{f_\nu\}$  be as in Proposition 2.15. Set  $g_\nu = \phi_\nu^*(f_\nu)$ , where  $\phi_\nu$  is the rescaling map associated to  $\mathcal{C}_\nu$ . Let  $r_\nu \rightarrow 0$ . Then by definition of  $g_\nu$ ,

$$\|g_\nu\|_{L^2(\mathbb{R}^2)}^2 \rightarrow \frac{1}{2} \quad \text{as } \nu \rightarrow \infty,$$

so the results of the preceding section apply to  $2^{1/2}g_\nu$ , and hence to  $g_\nu$  itself, uniformly in  $\nu$ .

If the first alternative in the conclusion of Lemma 11.2 holds, then we obtain the conclusion of Proposition 2.15. Therefore we may assume, by passing to a subsequence, that  $\{g_\nu\}$  satisfies the conclusions of the second alternative of Lemma 11.2.

Split

$$g_\nu = g_\nu^0 + g_\nu^\infty + g_\nu^b,$$

where

$$\begin{aligned} \|g_v^0\|_2 &\geq \delta, & \widehat{g_v^0}(\xi) \text{ is supported where } |\xi| &\leq 2s_v, \\ \|g_v^\infty\|_2 &\geq \delta, & \widehat{g_v^\infty}(\xi) \text{ is supported where } |\xi| &\geq \frac{1}{2}s_v, \\ \|g_v^b\|_2 &< \varepsilon_v, \end{aligned}$$

$g_v^0, g_v^\infty$  are upper normalized with respect to  $\mathcal{B}$ , and  $\varepsilon_v \rightarrow 0$  as  $\nu \rightarrow \infty$ .

Here  $\delta > 0$  is a certain constant independent of  $\nu$ , and  $\mathcal{B}$  denotes the unit ball in  $\mathbb{R}^2$ . This splitting is accomplished via an appropriate  $C^\infty$  three term partition of unity in the Fourier space  $\mathbb{R}_\xi^2$ .

Write  $\mathcal{C}_\nu = \mathcal{C}(z_\nu, r_\nu)$ . The decomposition above of  $g_\nu = \phi_\nu^*(f_\nu)$  induces a corresponding decomposition

$$f_\nu = F_\nu^0 + F_\nu^\infty + F_\nu^b,$$

where all three summands are real-valued and even and for all sufficiently large  $\nu$ ,

$$\left. \begin{aligned} &F_\nu^0, F_\nu^\infty, F_\nu^b \text{ are upper even-normalized with respect to } \mathcal{C}_\nu, \\ &\|F_\nu^b\|_2 \rightarrow 0 \text{ as } \nu \rightarrow \infty, \quad \|F_\nu^0\|_2 \geq \delta/2, \quad \|F_\nu^\infty\|_2 \geq \delta/2, \\ &F_\nu^0 \text{ and } F_\nu^\infty \text{ are supported in } \mathcal{C}(z_\nu, \frac{1}{2}) \cup -\mathcal{C}(z_\nu, \frac{1}{2}). \end{aligned} \right\} \tag{12-1}$$

Moreover:

**Lemma 12.1.** *The decomposition  $f_\nu = F_\nu^0 + F_\nu^\infty + F_\nu^b$  may be carried out so that the conditions above are satisfied, and moreover, for certain constants  $C, C_N < \infty$ , the summands  $F_\nu^0$  and  $F_\nu^\infty$  are real-valued, even, and admit representations*

$$F_\nu^0(y) = \int_{H_\nu} a_\nu^{0,\pm}(\xi) e^{iy \cdot \xi} d\xi, \quad \text{and} \quad F_\nu^\infty(y) = \int_{H_\nu} a_\nu^{\infty,\pm}(\xi) e^{iy \cdot \xi} d\xi, \tag{12-2}$$

where the representations with plus signs are valid for  $y \in \mathcal{C}(z_\nu, \frac{1}{2})$ , and those with minus signs are valid for  $y \in -\mathcal{C}(z_\nu, \frac{1}{2})$ , with Fourier coefficients  $a_\nu^{0,\pm}$  and  $a_\nu^{\infty,\pm}$  satisfying

$$\int_{r_\nu|\xi| \leq s_\nu/4} |a_\nu^{\infty,\pm}(\xi)|^2 d\xi \leq C S_\nu^{-1} \quad \text{for all } \nu, \tag{12-3}$$

$$\int_{r_\nu|\xi| \geq 4s_\nu} |a_\nu^{0,\pm}(\xi)|^2 d\xi \leq C_N s_\nu^{-N} \quad \text{for all } \nu, \text{ for any } N < \infty. \tag{12-4}$$

*Proof.* By rotational symmetry, it suffices to prove this under the assumption that  $z_\nu = (0, 0, 1)$  for all  $\nu$ . Then  $\phi_\nu^*(f_\nu)(x') = r_\nu f_\nu(r_\nu x', (1 - r_\nu^2|x'|^2)^{1/2})$  for  $x' \in \mathbb{R}^2$ , and  $H_\nu = \{x = (x', 0) \in \mathbb{R}^2 \times \mathbb{R}^1\}$ .

Once a representation of the required form is established for the restriction of  $f_\nu$  to the hemisphere  $\mathbb{S}_+^2 = \{y \in \mathbb{S}^2 : y_3 > 0\}$ , the symmetry  $f_\nu(-y) \equiv f_\nu(y)$  leads immediately to the desired representation for  $y_3 < 0$ . So we restrict attention to  $\mathbb{S}_+^2$ . For the remainder of this proof, we identify  $(\xi', 0) \in \mathbb{R}^{2+1}$  with  $\xi' \in \mathbb{R}^2$ , and denote elements of  $\mathbb{R}^2$  by  $\xi$  rather than by  $\xi'$ .

Fix a compactly supported  $C^\infty$  function  $\zeta : \mathbb{R}^2 \rightarrow \mathbb{R}$  that is supported in  $\{y' : |y'| < \frac{1}{2}\}$  and is  $\equiv 1$  in  $\{y' : |y'| \leq \frac{1}{4}\}$ . For  $y' \in \mathbb{R}^2$ , define

$$G_\nu^0(y') = (2\pi)^{-1} \zeta(y') r_\nu^{-1} \int_{\mathbb{R}^2} e^{ir_\nu^{-1}y' \cdot \xi} \widehat{g_\nu^0}(\xi) d\xi, \quad (12-5)$$

$$G_\nu^\infty(y') = (2\pi)^{-1} \zeta(y') r_\nu^{-1} \int_{\mathbb{R}^2} e^{ir_\nu^{-1}y' \cdot \xi} \widehat{g_\nu^\infty}(\xi) d\xi. \quad (12-6)$$

Then from the fact that  $g_\nu^0$  and  $g_\nu^\infty$  are upper normalized with respect to  $\mathcal{B}$ , it follows that

$$\|r_\nu^{-1} g_\nu^0(r_\nu^{-1} \cdot) - G_\nu^0(\cdot)\|_{L^2(\mathbb{R}^2)} \rightarrow 0 \quad \text{as } \nu \rightarrow \infty,$$

and likewise

$$\|r_\nu^{-1} g_\nu^\infty(r_\nu^{-1} \cdot) - G_\nu^\infty(\cdot)\|_{L^2(\mathbb{R}^2)} \rightarrow 0 \quad \text{as } \nu \rightarrow \infty,$$

using the hypothesis that  $r_\nu \rightarrow 0$  coupled with the fact that the support of  $\zeta$  is independent of  $r_\nu$ . It can of course be arranged that  $G_\nu^0$  and  $G_\nu^\infty$  are real-valued.

Define

$$F_\nu^b|_{\mathbb{S}_+^2}(y', y_3) = f_\nu(y', y_3) - G_\nu^0(y') - G_\nu^\infty(y').$$

where  $y_3 = \sqrt{1 - |y'|^2}$ . The function  $F_\nu^b|_{\mathbb{S}_+^2}$  is upper normalized with respect to  $\mathcal{C}_\nu$  because all three summands in its definition are upper normalized. Since  $\phi_\nu^*(f_\nu) = g_\nu^0 + g_\nu^\infty + g_\nu^b$ , since  $\|g_\nu^b\|_{L^2(\mathbb{R}^2)} \rightarrow 0$  as  $\nu \rightarrow \infty$ , since  $f_\nu$  is upper normalized with respect to  $\mathcal{C}_\nu$  and  $r_\nu \rightarrow 0$ , and since  $\phi_\nu^*$  is essentially an isometry from  $L^2(\mathbb{S}_+^2)$  to  $L^2(\mathbb{R}^2)$  for large  $\nu$  (again because  $r_\nu \rightarrow 0$ ), it follows that

$$\|F_\nu^b\|_{L^2(\mathbb{S}_+^2)} \rightarrow 0 \quad \text{as } \nu \rightarrow \infty.$$

When regarded in this way as functions of  $y = (y', y_3) \in \mathbb{S}_+^2$ , the summands  $G_\nu^0(y')$  and  $G_\nu^\infty(y')$  are each upper normalized with respect to the caps  $\mathcal{C}_\nu$ , because  $g_\nu^0$  and  $g_\nu^\infty$  are upper normalized with respect to  $\mathcal{B}$ . It remains only to show that  $G_\nu^0(y')$  can be represented in the form  $\int_{\mathbb{R}^2} e^{iy' \cdot \xi} a_\nu^{0,+}(\xi) d\xi$ , where  $a_\nu^{0,+}$  satisfies the required bound (12-4), and likewise for  $G_\nu^\infty$ . To prove this for  $G_\nu^0$ , it suffices to rewrite the product of  $\zeta(y')$  with the inverse Fourier transform in (12-5) as the inverse Fourier transform of a convolution, and to combine the bound  $|\widehat{\zeta}(\xi)| \leq C_N(1 + |\xi|)^{-N}$  for all  $N$  with the fact that  $\widehat{g_\nu^0}(\xi) \equiv 0$  for  $\{\xi : |\xi| > 2s_\nu\}$ . The analysis of  $G_\nu^\infty$  is essentially identical, using the given fact that  $\widehat{g_\nu^\infty}(\xi) \equiv 0$  for  $\{\xi : |\xi| < \frac{1}{2}S_\nu\}$ .  $\square$

As  $\nu \rightarrow \infty$ ,

$$\|f_\nu \sigma * f_\nu \sigma\|_2 \leq \|(F_\nu^0 \sigma * F_\nu^0 \sigma) + (F_\nu^\infty \sigma * F_\nu^\infty \sigma)\|_2 + 2\|F_\nu^0 \sigma * F_\nu^\infty \sigma\|_2 + o(1)$$

where  $o(1)$  denotes a function that tends to zero as  $\nu \rightarrow \infty$ . Applying the triangle inequality to the first term does not lead to a useful bound. Instead,

$$\begin{aligned} & \| (F_\nu^0 \sigma * F_\nu^0 \sigma) + (F_\nu^\infty \sigma * F_\nu^\infty \sigma) \|_2^2 \\ & \leq \| F_\nu^0 \sigma * F_\nu^0 \sigma \|_2^2 + \| F_\nu^\infty \sigma * F_\nu^\infty \sigma \|_2^2 + 2|\langle F_\nu^0 \sigma * F_\nu^0 \sigma, F_\nu^\infty \sigma * F_\nu^\infty \sigma \rangle| \\ & = \| F_\nu^0 \sigma * F_\nu^0 \sigma \|_2^2 + \| F_\nu^\infty \sigma * F_\nu^\infty \sigma \|_2^2 + 2|\langle F_\nu^0 \sigma * F_\nu^\infty \sigma, F_\nu^0 \sigma * F_\nu^\infty \sigma \rangle| \end{aligned}$$

since  $F_\nu^0$  and  $F_\nu^\infty$  are real and even. Therefore and since  $F_\nu^0, F_\nu^\infty$  have uniformly bounded  $L^2$  norms,

$$\| f_\nu \sigma * f_\nu \sigma \|_2^2 \leq \| F_\nu^0 \sigma * F_\nu^0 \sigma \|_2^2 + \| F_\nu^\infty \sigma * F_\nu^\infty \sigma \|_2^2 + C \| F_\nu^0 \sigma * F_\nu^\infty \sigma \|_2^2 + o(1). \quad (12-7)$$

The following key lemma will be proved below, in Section 14.

**Lemma 12.2.** *Let  $F_\nu^0$  and  $F_\nu^\infty$  be upper even-normalized with respect to a sequence of caps of radii  $r_\nu \rightarrow 0$ . Assume that  $F_\nu^0$  and  $F_\nu^\infty$  admit Fourier representations satisfying the inequalities specified in Lemma 12.1. Then*

$$\| F_\nu^0 \sigma * F_\nu^\infty \sigma \|_{L^2(\mathbb{R}^3)} \rightarrow 0.$$

**Corollary 12.3.** *The second alternative in the conclusion of Lemma 11.2 cannot hold.*

*Proof.* Assume Lemma 12.2. Then by (12-7),

$$\| f_\nu \sigma * f_\nu \sigma \|_2^2 \leq \| F_\nu^0 \sigma * F_\nu^0 \sigma \|_2^2 + \| F_\nu^\infty \sigma * F_\nu^\infty \sigma \|_2^2 + o(1) \leq \mathbf{S}^4 \| F_\nu^0 \|_2^4 + \mathbf{S}^4 \| F_\nu^\infty \|_2^4 + o(1).$$

Since  $S_\nu/s_\nu \rightarrow \infty$  and  $\| F_\nu^b \|_2 \rightarrow 0$ , it follows easily from (12-3) and (12-4) that

$$\| F_\nu^0 \|_2^2 + \| F_\nu^\infty \|_2^2 \leq (1 + o(1)) \| f_\nu \|_2^2 = 1 + o(1).$$

Since  $\min(\| F_\nu^0 \|_2, \| F_\nu^\infty \|_2) \geq \delta/2$ , this forces

$$\max(\| F_\nu^0 \|_2^2, \| F_\nu^\infty \|_2^2) \leq 1 - \rho$$

for all sufficiently large  $\nu$ , for some  $\rho > 0$  independent of  $\nu$ . It follows that

$$\begin{aligned} \mathbf{S}^4 \| F_\nu^0 \|_{L^2(\sigma)}^4 + \mathbf{S}^4 \| F_\nu^\infty \|_{L^2(\sigma)}^4 & \leq \mathbf{S}^4 (\| F_\nu^0 \|_{L^2(\sigma)}^2 + \| F_\nu^\infty \|_{L^2(\sigma)}^2) \max(\| F_\nu^0 \|_2^2, \| F_\nu^\infty \|_2^2) \\ & \leq \mathbf{S}^4 (1 + o(1))(1 - \rho). \end{aligned}$$

We conclude that

$$\limsup_{\nu \rightarrow \infty} \| f_\nu \sigma * f_\nu \sigma \|_{L^2(\mathbb{R}^3)}^2 < \mathbf{S}^4,$$

contradicting the assumption that  $\{f_\nu\}$  was an extremizing sequence.  $\square$

Combining the results above, the proof of Proposition 2.15 is complete except for the proof of Lemma 12.2.



### 13. Step 8: Excluding small caps

In this section we prove Proposition 2.16, which asserts that the radii  $r_\nu$  of the caps  $\mathcal{C}_\nu$  associated to an extremizing sequence  $\{f_\nu\}$  of positive even functions cannot tend to zero.

**Lemma 13.1.** *Let  $\{f_\nu\}$  be any sequence of real-valued, even functions on  $\mathbb{S}^2$  satisfying  $\|f_\nu\|_{L^2} = 1$ . Suppose that  $f_\nu$  is upper even-normalized with respect to a cap  $\mathcal{C}_\nu = \mathcal{C}(z_\nu, r_\nu)$ , uniformly in  $\nu$ . Suppose that the sequence of pullbacks  $\phi_\nu^*(f_\nu)$  satisfies the first alternative in the conclusion of Lemma 11.2. Suppose that  $r_\nu \rightarrow 0$ . Then there exists a sequence of functions  $F_\nu : \mathbb{P}^2 \rightarrow \mathbb{R}$  satisfying  $\|F_\nu\|_2 \rightarrow 1$  such that*

$$\limsup_{\nu \rightarrow \infty} \|F_\nu \sigma_P * F_\nu \sigma_P\|_2 \geq (3/2)^{-1/2} \limsup_{\nu \rightarrow \infty} \|f_\nu \sigma * f_\nu \sigma\|_2.$$

*Proof of Proposition 2.16.* Let  $\{f_\nu\}$  be an extremizing sequence of nonnegative even functions for the inequality (2-1) satisfying  $\|f_\nu\|_2 = 1$ . There exists a sequence of caps  $\mathcal{C}_\nu = \mathcal{C}(z_\nu, r_\nu)$  such that each  $f_\nu$  is upper even-normalized with respect to  $\mathcal{C}_\nu$ . We must prove that  $\inf_\nu r_\nu > 0$ .

If not, then by passing to a subsequence we may assume that  $r_\nu \rightarrow 0$ . By Proposition 2.15, the sequence of pullbacks  $g_\nu = \phi_\nu^*(f_\nu)$  is precompact in  $L^2(\mathbb{R}^2)$ . Thus the hypotheses of Lemma 13.1 are satisfied, so there exists a sequence of functions  $F_\nu \in L^2(\mathbb{P}^2)$  satisfying its conclusions.

Now  $\|F_\nu \sigma_P * F_\nu \sigma_P\|_2 \leq \mathbf{P}^2 \|F_\nu\|_{L^2(\mathbb{P}^2)}^2$  by the definition of  $\mathbf{P}$ . Consequently

$$\limsup_{\nu \rightarrow \infty} \|f_\nu \sigma * f_\nu \sigma\|_2 \leq (3/2)^{1/2} \mathbf{P}^2.$$

The left side tends to  $\mathbf{S}^2$  since  $\{f_\nu\}$  is an extremizing sequence for (2-1), so  $\mathbf{S}^2 \leq (3/2)^{1/2} \mathbf{P}^2$ , contradicting the inequality  $\mathbf{S} \geq 2^{1/4} \mathbf{P}$  of Lemma 2.4. □

*Proof of Lemma 13.1.* Write  $\mathcal{C}_\nu = \mathcal{C}(z_\nu, r_\nu)$ . Decompose  $2^{1/2} f_\nu(x) = f_\nu^+(x) + f_\nu^+(-x) + f_\nu^b(x)$ , where  $f_\nu^+$  is real,  $f_\nu^+$  is supported in  $\mathcal{C}(z_\nu, r_\nu^{1/2})$ ,  $\|f_\nu^b\|_2 \rightarrow 0$ , and the functions  $\phi_\nu^*(f_\nu^+)$  satisfy the first alternative of the conclusions of Lemma 11.2, uniformly in  $\nu$ .

Since  $f_\nu$  is even and  $\|f_\nu\|_2 = 1$ , we have  $\|f_\nu^+\|_2 \rightarrow 1$  as  $\nu \rightarrow \infty$ . Moreover  $g_\nu(x) = f_\nu^+(x) + f_\nu^+(-x)$  satisfies

$$\|g_\nu \sigma * g_\nu \sigma\|_2^2 / \|g_\nu\|_2^4 \equiv \frac{3}{2} \|f_\nu^+ \sigma * f_\nu^+ \sigma\|_2^2 / \|f_\nu^+\|_2^4,$$

and therefore

$$\limsup_{\nu \rightarrow \infty} \|f_\nu^+ \sigma * f_\nu^+ \sigma\|_2^2 = (3/2)^{-1} \limsup_{\nu \rightarrow \infty} \|f_\nu \sigma * f_\nu \sigma\|_2^2.$$

By rotation symmetry, we may suppose that  $z_\nu = (0, 0, 1)$  for all  $\nu$ . Define  $F_\nu : \mathbb{P}^2 \rightarrow [0, \infty)$  by

$$F_\nu(y, |y|^2/2) = r_\nu f_\nu^+(r_\nu y, (1 - r_\nu^2 |y|^2)^{1/2})$$

for  $y \in \mathbb{R}^2$ . The function  $F_\nu$  will also be regarded as an element of  $L^2(\mathbb{R}^2, dy)$  by  $F_\nu(y) = F_\nu(y, |y|^2/2)$ . Then  $\|F_\nu\|_{L^2(\mathbb{P}^2, \sigma_P)} = \|F_\nu\|_{L^2(\mathbb{R}^2)} \rightarrow 1$  as  $\nu \rightarrow \infty$ .

It remains to prove that

$$\limsup_{\nu \rightarrow \infty} \|\widehat{F_\nu \sigma_P}\|_{L^4(\mathbb{R}^3)}^4 \geq \limsup_{\nu \rightarrow \infty} \|\widehat{f_\nu^+ \sigma}\|_{L^4(\mathbb{R}^3)}^4.$$

We have

$$\int_{|y|\geq R} F_v(y)^2 dy + \int_{F_v(y)\geq R} F_v(y)^2 dy + \int_{|\xi|\geq R} |\widehat{F}_v(\xi)|^2 d\xi \rightarrow 0 \tag{13-1}$$

as  $R \rightarrow \infty$ , uniformly in  $v$ .

Thus we must compare  $\widehat{F}_v \sigma_P(x, t) = \int e^{-ix \cdot y - it|y|^2/2} F_v(y) dy$  with

$$\begin{aligned} \widehat{f}_v^+ \sigma(x, t) &= \int_{\mathbb{R}^2} e^{-ix \cdot v - it(1-|v|^2)^{1/2}} f_v^+(v, (1-|v|^2)^{1/2}) d\sigma(v, (1-|v|^2)^{1/2}) \\ &= \int_{\mathbb{R}^2} e^{-ix \cdot v - it(1-|v|^2)^{1/2}} f_v^+(v, (1-|v|^2)^{1/2}) (1-|v|^2)^{-1/2} dv. \end{aligned}$$

In the latter integral, substitute  $v = r_v y$  to obtain

$$\begin{aligned} r_v^{-1} \widehat{f}_v^+ \sigma(r_v^{-1}x, -r_v^{-2}t) &= r_v^{-1} r_v^2 \int_{\mathbb{R}^2} e^{-ix \cdot y + itr_v^{-2}(1-r_v^2|y|^2)^{1/2}} f_v^+(r_v y, (1-r_v^2|y|^2)^{1/2}) (1-r_v^2|y|^2)^{-1/2} dy \\ &= \int_{\mathbb{R}^2} e^{-ix \cdot y + itr_v^{-2}(1-r_v^2|y|^2)^{1/2}} F_v(y) (1-r_v^2|y|^2)^{-1/2} dy \\ &= e^{itr_v^{-2}} \int_{\mathbb{R}^2} e^{-ix \cdot y - it|y|^2/2} F_v(y) h_v(t, y) dy, \end{aligned}$$

where

$$h_v(t, y) = e^{it\psi_v(y)} (1-r_v^2|y|^2)^{-1/2} \quad \text{and} \quad \psi_v(y) = -r_v^{-2} + |y|^2/2 + r_v^{-2}(1-r_v^2|y|^2)^{1/2}.$$

Thus

$$\|\widehat{f}_v^+ \sigma\|_4^4 = \int_{\mathbb{R}} \int_{\mathbb{R}^2} |r_v^{-1} \widehat{f}_v^+ \sigma(r_v^{-1}x, -r_v^{-2}t)|^4 dx dt = \left\| \int_{\mathbb{R}^2} e^{-ix \cdot y - it|y|^2/2} F_v(y) h_v(t, y) dy \right\|_{L^4(\mathbb{R}^3)}^4.$$

It will be important that on any compact subset of  $\mathbb{R}_t^1 \times \mathbb{R}_y^2$ ,

$$h_v(t, y) \rightarrow 1 \text{ in the } C^N \text{ norm as } v \rightarrow \infty, \text{ for all } N < \infty. \tag{13-2}$$

Define

$$u_v(x, t) = \int_{\mathbb{R}^2} e^{-ix \cdot y - it|y|^2/2} F_v(y) h_v(t, y) dy \quad \text{and} \quad \tilde{u}_v(x, t) = \int e^{-ix \cdot y - it|y|^2/2} F_v(y) dy.$$

**Lemma 13.2.** *We have*

$$\begin{aligned} \int_{|(x,t)|\geq R} |u_v(x, t)|^4 dx dt &\rightarrow 0 \quad \text{as } R \rightarrow \infty, \text{ uniformly in } v, \\ \int_{|(x,t)|\geq R} |\tilde{u}_v(x, t)|^4 dx dt &\rightarrow 0 \quad \text{as } R \rightarrow \infty, \text{ uniformly in } v. \end{aligned}$$

*Proof.* Define operators  $T_v$  and  $T$  from  $L^2(\mathbb{R}^2)$  to  $L^4(\mathbb{R}^3)$  by

$$T_v g(x, t) = \int_{\mathbb{R}^2} e^{-ix \cdot y - it|y|^2/2} g(y) \chi_{r_v^{-1}|y|\leq 1/2}(y) h_v(t, y) dy, \quad T g(x, t) = \int e^{-ix \cdot y - it|y|^2/2} g(y) dy.$$

The operator  $T : L^2(\mathbb{R}^2) \rightarrow L^4(\mathbb{R}^3)$  is bounded. Although the operators  $T_\nu$  are written in coordinates that disguise this fact, they are bounded from  $L^2(\mathbb{R}^2)$  to  $L^4(\mathbb{R}^3)$  uniformly in  $\nu$ , being obtained via norm-preserving changes of variables from the single bounded operator  $L^2(\mathbb{S}^2, \sigma) \ni h \mapsto \widehat{h\sigma}$ .

If  $g \in C^2(\mathbb{R}^2)$  has compact support, then  $|T_\nu g(x, t)| \leq C_g |(x, t)|^{-1}$ , where  $C_g$  depends only on the  $C^1$  norm of  $g$  and on the diameter of its support, provided that  $\nu$  is sufficiently large that the support of  $g$  is contained in  $B(0, r_\nu^{-1})$ . This follows from (13-2) together with the method of stationary phase; the phase functions appearing in the definition of  $T_\nu$  have uniformly nondegenerate critical points (if any), uniformly in  $\nu$ .

These two facts, together with the three uniform inequalities (13-1), lead directly to the stated conclusion for  $u_\nu$  by a routine argument.

A slightly simpler application of the same reasoning applies to  $\tilde{u}_\nu$ . □

Therefore it suffices to prove that for any  $R < \infty$ ,

$$\int_{|(x,t)| \leq R} |u_\nu(x, t) - \tilde{u}_\nu(x, t)|^4 dx dt \rightarrow 0 \quad \text{as } \nu \rightarrow \infty. \quad (13-3)$$

If  $g \in L^1$  has compact support, then

$$|T_\nu(g)(x, t) - T(g)(x, t)| \rightarrow 0, \quad \text{uniformly for all } |(x, t)| \leq R. \quad (13-4)$$

Since  $T_\nu$  and  $T$  are uniformly bounded operators from  $L^2$  to  $L^4$ , and since the class of all compactly supported  $g \in L^1$  is dense in  $L^2$ , (13-3) follows from (13-4). □

#### 14. Estimation of the cross term $\|F_\nu^0 \sigma * F_\nu^\infty \sigma\|_2^2$

To prove Lemma 12.2, let  $f_\nu$ ,  $F_\nu^0$  and  $F_\nu^\infty$  be as above. Let  $f_\nu$  be upper even-normalized with respect to a cap  $\mathcal{C}_\nu$  of radius  $r_\nu$ . Since the inequality in question is invariant under rotations of  $\mathbb{R}^3$ , we may suppose without loss of generality that  $\mathcal{C}_\nu$  is centered at the north pole  $z_0 = (0, 0, 1)$ .

Decompose  $F_\nu^0 = F_\nu^{0,+} + F_\nu^{0,-}$ , where both summands are real-valued,  $F_\nu^{0,+}$  is supported in  $\mathcal{C}(z_0, \frac{1}{2})$ ,  $F_\nu^{0,-}(x) = F_\nu^{0,+}(-x)$ ,  $F_\nu^{0,\pm}$  is upper normalized with respect to  $\mathcal{C}(\pm z_0, r_\nu)$ , and  $F_\nu^{0,\pm}$  have the same Fourier representations (12-2) as  $F_\nu^0$ . There is a parallel decomposition  $F_\nu^\infty = F_\nu^{\infty,+} + F_\nu^{\infty,-}$ . By Lemma 3.2,

$$\|F_\nu^{0,+} \sigma * F_\nu^{\infty,+} \sigma\|_2 = \|F_\nu^{0,-} \sigma * F_\nu^{\infty,-} \sigma\|_2 = \|F_\nu^{0,-} \sigma * F_\nu^{\infty,+} \sigma\|_2 = \|F_\nu^{0,+} \sigma * F_\nu^{\infty,-} \sigma\|_2.$$

Therefore it suffices to bound  $\|F_\nu^{0,+} \sigma * F_\nu^{\infty,+} \sigma\|_2$ .

**Lemma 14.1.** *Let  $\delta_\nu, \delta_\nu^* > 0$  be sequences of positive numbers that satisfy*

$$\delta_\nu / r_\nu^2 \rightarrow 0 \quad \text{and} \quad \delta_\nu^* / r_\nu^2 \rightarrow \infty.$$

*Then, with  $A := \{x \in \mathbb{R}^3 : |x| > 2 - \delta_\nu \text{ or } |x| < 2 - \delta_\nu^*\}$ ,*

$$\|F_\nu^{0,+} \sigma * F_\nu^{\infty,+} \sigma\|_{L^2(A)} \rightarrow 0 \quad \text{as } \nu \rightarrow \infty.$$

*Proof.* Since  $F_\nu^{0,+}$  and  $F_\nu^{\infty,+}$  are upper normalized with respect to  $\mathcal{C}_\nu$ , Corollary 7.3 asserts that the region  $|x| > 2 - \delta_\nu$  makes a small contribution for large  $\nu$ . To handle the region  $|x| < 2 - \delta_\nu^*$ , choose a sequence  $t_\nu \geq 1$  tending slowly to infinity. Decompose  $F_\nu^{0,+} = F_\nu^{0,+} \chi_{\mathcal{C}(z_0, t_\nu r_\nu)} + F_\nu^{0,+} \chi_{\mathbb{S}^2 \setminus \mathcal{C}(z_0, t_\nu r_\nu)}$ , and decompose  $F_\nu^{\infty,+}$  in the same way. If  $t_\nu \rightarrow \infty$  sufficiently slowly, then the main term  $F_\nu^{0,+} \chi_{\mathcal{C}(z_0, t_\nu r_\nu)} \sigma * F_\nu^{\infty,+} \chi_{\mathcal{C}(z_0, t_\nu r_\nu)} \sigma$  is supported where  $|x| > 2 - \delta_\nu^*$ . Expanding  $F_\nu^{0,+} \sigma * F_\nu^{\infty,+} \sigma$  according to this decomposition leaves three more terms. Each of these has small norm in  $L^2(\mathbb{R}^3)$  for large  $\nu$ , because  $\|F_\nu^{0,+}\|_{L^2(\mathbb{S}^2 \setminus \mathcal{C}(z_0, t_\nu r_\nu))} \rightarrow 0$  and  $\|F_\nu^{\infty,+}\|_{L^2(\mathbb{S}^2 \setminus \mathcal{C}(z_0, t_\nu r_\nu))} \rightarrow 0$ .  $\square$

If  $h_1$  and  $h_2$  are supported in  $\mathcal{C}(z_0, r)$ , then  $h_1 \sigma * h_2 \sigma$  is supported in  $\{x \in \mathbb{R}^3 : |x - 2z_0| \leq Cr\}$ . Since  $F_\nu^{0,+}$  and  $F_\nu^{\infty,+}$  are upper normalized with respect to  $\mathcal{C}(z_\nu, r_\nu)$ , and since  $r_\nu \rightarrow 0$ , it follows from the inequality  $\|h_1 \sigma * h_2 \sigma\|_{L^2(\mathbb{R}^3)} \leq C \|h_1\|_2 \|h_2\|_2$  that

$$\int_{|x-2z_0| \geq 1/100} |(F_\nu^{0,+} \sigma * F_\nu^{\infty,+} \sigma)(x)|^2 dx \rightarrow 0 \quad \text{as } \nu \rightarrow \infty.$$

On the other hand, if  $|x - 2z_0| \leq 1/100$ , then for all sufficiently large  $\nu$ ,  $(F_\nu^{0,+} \sigma * F_\nu^{\infty,+} \sigma)(x)$  depends only on the restrictions of  $F_\nu^{0,+}$  and  $F_\nu^{\infty,+}$  to  $\mathcal{C}(z_0, 1/10)$ . This has the following significance in terms of the Fourier representations (12-3), (12-4) of Lemma 12.1:

$$F_\nu^{0,+}(x) = \int_{r_\nu|\zeta| \leq 4s_\nu} e^{ix\zeta} a_\nu^{0,+}(\zeta) d\zeta + o(1) \quad \text{in } L^2(\mathcal{C}(z_0, 1/10)) \text{ as } \nu \rightarrow \infty \quad (14-1)$$

by virtue of (12-4); this does not follow for  $L^2(\mathbb{S}^2)$  because surface measure on  $\mathbb{S}^2$  is not approximately equivalent to Lebesgue measure on  $\{(x_1, x_2, 0)\}$  near the equator  $\{x \in \mathbb{S}^2 : x_3 = 0\}$ . Likewise, by (12-3),

$$F_\nu^{\infty,+}(x) = \int_{r_\nu|\zeta| \geq S_\nu/4} e^{ix\zeta} a_\nu^{\infty,+}(\zeta) d\zeta + o(1) \quad \text{in } L^2(\mathcal{C}(z_0, 1/10)) \text{ as } \nu \rightarrow \infty. \quad (14-2)$$

Henceforth we simplify notation by writing  $a_\nu^0$  in place of  $a_\nu^{0,+}$  and  $a_\nu^\infty$  in place of  $a_\nu^{\infty,+}$ , and we will take these functions to be supported in the sets  $r_\nu|\zeta| \leq 4s_\nu$  and  $r_\nu|\zeta| \geq S_\nu/4$ , respectively.

Set  $H = \{\xi \in \mathbb{R}^3 : \xi_3 = 0\}$ , and identify  $(\xi_1, \xi_2, 0) \in H$  with  $(\xi_1, \xi_2) \in \mathbb{R}^2$ . Denote a region  $\mathcal{A}_\nu$  and an interval  $I_\nu$  by

$$\mathcal{A}_\nu = \{x \in \mathbb{R}^3 : 2 - \delta_\nu^* \leq |x| \leq 2 - \delta_\nu \text{ and } |x - 2z_0| < 1/100\} \quad \text{and} \quad I_\nu = [2 - \delta_\nu^*, 2 - \delta_\nu].$$

It remains only to estimate  $\|F_\nu^{\infty,+} \sigma * F_\nu^{0,+} \sigma\|_{L^2(\mathcal{A}_\nu)}$ . For  $x \in \mathcal{A}_\nu$  and for all sufficiently large  $\nu$ ,  $(F_\nu^{0,+} \sigma * F_\nu^{\infty,+} \sigma)(x)$  depends only on the restrictions of  $F_\nu^{0,+}$ ,  $F_\nu^{\infty,+}$  to  $\mathcal{C}(z_0, 1/10)$ . Therefore in majorizing  $\|F_\nu^{\infty,+} \sigma * F_\nu^{0,+} \sigma\|_{L^2(\mathcal{A}_\nu)}$ , we may replace  $F_\nu^{0,+}(x)$  by  $\int_{r_\nu|\zeta| \leq 4s_\nu} e^{ix\zeta} a_\nu^0(\zeta) d\zeta$  and  $F_\nu^{\infty,+}(x)$  by  $\int_{r_\nu|\zeta| \geq S_\nu/4} e^{ix\zeta} a_\nu^\infty(\zeta) d\zeta$ , at the expense of additional terms that are  $o(1)$  as  $\nu \rightarrow \infty$ . We will continue to denote these modified functions by  $F_\nu^{0,+}$ ,  $F_\nu^{\infty,+}$ .

Set  $h_\zeta = e_{-i\zeta} F_\nu^{\infty,+}$  for  $r_\nu|\zeta| \leq 4s_\nu$ . Let

$$H^* = \{\zeta \in H : r_\nu|\zeta| \leq 4s_\nu\}.$$

By (7-4), (7-5), (14-1), and (14-2),

$$\begin{aligned}
\|F_v^{\infty,+} \sigma * F_v^{0,+} \sigma\|_{L^2(\mathcal{A}_v)}^2 &\leq C \int_{I_v} \left\| \int_{H^*} |a_v^0(\zeta)| \cdot |T_{\rho(t)} h_\zeta| d\zeta \right\|_{L^2(\mathbb{S}^2)}^2 dt + o(1) \\
&\leq C \int_{I_v} \left( \int_{H^*} |a_v^0(\zeta)| \cdot \|T_{\rho(t)} h_\zeta\|_{L^2(\mathbb{S}^2)} d\zeta \right)^2 dt + o(1) \\
&\leq C \|a_v^0\|_2^2 \int_{H^*} \int_{I_v} \|T_{\rho(t)} h_\zeta\|_{L^2(\mathbb{S}^2)}^2 dt d\zeta + o(1) \\
&\leq C \int_{H^*} \int_{I_v} \|T_{\rho(t)} h_\zeta\|_{L^2(\mathbb{S}^2)}^2 dt d\zeta + o(1)
\end{aligned}$$

by the Minkowski and Cauchy–Schwarz inequalities. Inserting the Fourier integral operator bound  $\|T_\rho(h_\zeta)\|_2^2 \leq C \|(I - \rho^2 \Delta)^{-1/4} h_\zeta\|_2^2$  yields

$$\begin{aligned}
\|F_v^{\infty,+} \sigma * F_v^{0,+} \sigma\|_{L^2(\mathcal{A}_v)}^2 &\leq C \int_{\zeta \in H^*} \int_{I_v} \int_{\xi \in H} (1 + \rho(t)|\xi|)^{-1} |\widehat{h_\zeta}(\xi)|^2 d\xi dt d\zeta + o(1) \\
&= C \int_{\zeta \in H^*} \int_{I_v} \int_{\xi \in H} (1 + \rho(t)|\xi|)^{-1} |a_v^\infty(\xi - \zeta)|^2 d\xi dt d\zeta \\
&\sim s_v^2 r_v^{-2} \int_{I_v} \int_H (1 + \rho(t)|\xi|)^{-1} |a_v^\infty(\xi)|^2 d\xi dt + o(1), \tag{14-3}
\end{aligned}$$

since  $|\xi| \gg |\zeta|$  for  $\zeta$  in the support of  $a_v^0$  and  $\xi$  in the support of  $a_v^\infty$ . Next,

$$\begin{aligned}
\int_H (1 + \rho|\xi|)^{-1} |a_v^\infty(\xi)|^2 d\xi &\leq C \int_{r_v|\xi| \leq c_0 S_v} |a_v^\infty(\xi)|^2 d\xi + C \int_{r_v|\xi| \geq c_0 S_v} (1 + \rho|\xi|)^{-1} |a_v^\infty(\xi)|^2 d\xi \\
&\leq C S_v^{-1} \|F_v^{\infty,+}\|_2^2 + C \max_{r_v|\xi| \geq c_0 S_v} (1 + \rho|\xi|)^{-1} \cdot \|F_v^{\infty,+}\|_2^2 \\
&\leq C S_v^{-1} + C \rho^{-1} r_v S_v^{-1}.
\end{aligned}$$

The first term after the first inequality was estimated using (12-3). Inserting the final line into (14-3) yields

$$\begin{aligned}
\|F_v^{\infty,+} \sigma * F_v^{0,+} \sigma\|_{L^2(\mathcal{A}_v)}^2 &\leq C s_v^2 r_v^{-2} \int_{I_v} (S_v^{-1} + \rho(t)^{-1} r_v S_v^{-1}) dt \\
&\leq C s_v^2 r_v^{-2} \int_{I_v} (S_v^{-1} + (2-t)^{-1/2} r_v S_v^{-1}) dt \\
&\quad \text{since } (t/2)^2 + \rho(t)^2 = 1 \text{ implies } \rho(t) \geq C(2-t)^{1/2} \\
&= C s_v^2 S_v^{-1} r_v^{-2} \int_{I_v} (1 + r_v(2-t)^{-1/2}) dt \\
&\leq C s_v^2 S_v^{-1} r_v^{-2} |I_v| (1 + \max_{t \in I_v} r_v(2-t)^{-1/2}) \\
&\leq C s_v^2 S_v^{-1} (r_v^{-2} \delta_v^*) (1 + \delta_v^{-1/2} r_v) \leq C s_v^{-1} (r_v^{-2} \delta_v^*) (1 + \delta_v^{-1/2} r_v)
\end{aligned}$$

since  $S_v \geq s_v^3$ .

Combining all terms, we have shown that

$$\|F_v^{0,+}\sigma * F_v^{\infty,+}\sigma\|_2^2 \leq o(1) + Cs_v^{-1}(r_v^{-2}\delta_v^*)(1 + \delta_v^{-1/2}r_v)$$

as  $v \rightarrow \infty$ , provided that  $\delta_v/r_v^2 \rightarrow 0$  and  $\delta_v^*/r_v^2 \rightarrow \infty$ . Since  $s_v \rightarrow \infty$ , it is possible to choose  $\delta_v$  and  $\delta_v^*$  to satisfy the additional constraint

$$s_v^{-1}(r_v^{-2}\delta_v^*)(1 + \delta_v^{-1/2}r_v) \rightarrow 0 \quad \text{as } v \rightarrow \infty.$$

With such a choice, we obtain

$$\|F_v^{0,+}\sigma * F_v^{\infty,+}\sigma\|_2^2 \rightarrow 0 \quad \text{as } v \rightarrow \infty,$$

completing the proof of Lemma 12.2. □

### 15. Step 9: Large caps

We now prove Proposition 2.17. The proof is quite similar to that of Proposition 2.15, without the complication of ensuring uniformity of bounds as  $r_v \rightarrow 0$ . However, the proof of Proposition 2.15 also exploited the condition  $r_v \rightarrow 0$  in a positive way, and substantive modification is therefore required here. Matters here that are essentially identical to corresponding matters in the earlier proof will be treated sketchily.

There is given an extremizing sequence  $\{f_v\}$  of even nonnegative functions satisfying  $\|f_v\|_{L^2(\mathbb{S}^2)} = 1$ , each of which is upper even-normalized with respect to a certain cap  $\mathcal{C}(z_v, r_v)$ . It is given that  $r_* = \inf_v r_v$  is strictly positive.

Introduce a  $C^\infty$  partition of unity of  $\mathbb{S}^2$  by nonnegative functions  $\eta_j$ , each of which is supported in a cap  $\mathcal{C}(z_j, \frac{1}{2})$ . The points  $z_j$  and functions  $\eta_j$  are to be chosen independent of  $v$ . Let  $\phi_j : \mathbb{R}^2 \rightarrow \mathbb{S}^2$  and  $\phi_j^* : L^2(\mathbb{S}^2) \rightarrow L^2(\mathbb{R}^2)$  be the associated mappings.

Since  $r_* \leq r_v \leq 1$ , the uniform upper normalization of  $f_v$  means simply that  $\|f_v\|_{L^2(\mathbb{S}^2)} \leq 1$ , and there exists a function  $\Theta$  that is independent of  $v$  and satisfies  $\Theta(R) \rightarrow \infty$  as  $R \rightarrow \infty$ , such that

$$\int_{|f_v(x)| \geq R} |f_v(x)|^2 d\sigma(x) \leq \Theta(R) \quad \text{for all } v.$$

Thus the radii  $r_v$  no longer enter into the discussion.

Decompose  $f_v = \sum_j f_{v,j}$ , where  $f_{v,j} = \eta_j f_v$ . By identifying the plane tangent to  $\mathbb{S}^2$  at  $z_j$  with a fixed copy of  $\mathbb{R}^2$ , we may regard each  $g_{v,j} = \phi_j^*(f_{v,j})$  as an element of  $L^2(\mathbb{R}^2)$ ; thus the functions  $g_{v,j}$ , and hence their Fourier transforms, have a common domain. The functions  $g_{v,j}$  are supported in  $\{y \in \mathbb{R}^2 : |y| \leq \frac{1}{2}\}$ , and again  $\int_{|g_{v,j}(y)| \geq R} |g_{v,j}(y)|^2 dy \leq \Theta(R)$ , where  $\Theta(R) \rightarrow 0$  as  $R \rightarrow \infty$ .

The analogue of Lemma 11.2 in this simplified situation is the following dichotomy: Either there exists a function  $\theta : [1, \infty) \rightarrow (0, \infty)$  satisfying  $\theta(s) \rightarrow 0$  as  $s \rightarrow \infty$  such that

$$\int_{|\xi| \geq s} \sum_j |\widehat{g_{v,j}}(\xi)|^2 d\xi \leq \theta(s) \quad \text{for all } s \in [1, \infty) \text{ and all } v, \tag{15-1}$$

or there exist  $\delta, \varepsilon_\nu, s_\nu, S_\nu$  as in Lemma 11.2 such that the conclusions in the second case of that lemma hold, with  $|\widehat{g}_\nu|^2$  replaced by  $\sum_j |\widehat{g}_{\nu,j}|^2$ . The proof of this dichotomy is essentially identical to the proof of Lemma 11.2 itself.

If (15-1) holds, then the conclusion of Proposition 2.17 is simply a reformulation of the conjunction of the upper normalization bounds for  $f_\nu$  with (15-1); the desired decomposition of  $f_\nu$  is obtained by expressing each  $g_{\nu,j}$  as an inverse Fourier transform, splitting the resulting integral with respect to  $\xi$  into large  $|\xi|$  and smaller  $|\xi|$  regions, and reversing the mapping  $\phi_j^*$  to transplant both summands to  $\mathbb{S}^2$ . The contribution of sufficiently large  $|\xi|$  will have small  $L^2(\mathbb{S}^2)$  norm, while the contribution of smaller  $|\xi|$  will satisfy an adequate  $C^1$  norm bound.

It remains only to demonstrate that the second case of the dichotomy cannot arise; there cannot exist  $\delta, \varepsilon_\nu, s_\nu, S_\nu$  satisfying all conclusions of the second case of Lemma 11.2. Suppose to the contrary that this situation were to arise. Denote by  $\phi_{j,*}^{-1}$  the left inverse of  $\phi_j^*$ , mapping functions supported in  $\{y \in \mathbb{R}^2 : |y| \leq \frac{3}{4}\}$  to functions supported in  $\mathcal{C}(z_j, \frac{3}{4}) \subset \mathbb{S}^2$ . By summing over  $j$ , one would obtain as in (12-1) a decomposition

$$f_\nu = F_\nu^0 + F_\nu^\infty + F_\nu^b, \quad (15-2)$$

where  $\lim_{\nu \rightarrow \infty} \|F_\nu^b\|_2 = 0$ ,  $F_\nu^\infty$  is highly oscillatory, and  $F_\nu^0$  is slowly varying in comparison with  $F_\nu^\infty$ . Here for instance

$$F_\nu^\infty = \sum_j \phi_{j,*}^{-1}(\zeta \cdot g_{\nu,j}^\infty), \quad \text{where } \widehat{g_{\nu,j}^\infty}(\xi) = (1 - m(\xi/S_\nu)) \widehat{g_{\nu,j}}(\xi)$$

and where the  $C^\infty$  cutoff functions  $\zeta$  and  $m$  have the following properties:  $\zeta \in C^\infty(\mathbb{R}^2)$  is  $\equiv 1$  on the ball  $B(0, \frac{5}{8})$ , and is supported on  $B(0, \frac{3}{4})$ , while  $m(\xi) \equiv 0$  for  $|\xi| \geq \frac{3}{8}$  and  $m(\xi) \equiv 1$  for  $|\xi| \leq \frac{1}{4}$ .  $F_\nu^0$  is defined in the same way, with  $1 - m(\xi/S_\nu)$  replaced by  $m(\xi/8s_\nu)$ .

The decomposition (15-2) can be modified so that  $F_\nu^0, F_\nu^\infty$  and  $F_\nu^b$  remain real-valued and even, without sacrificing any of its desired properties. First replace each summand by its real part. Then replace  $F_\nu^0(x)$  by  $\frac{1}{2}F_\nu^0(x) + \frac{1}{2}F_\nu^0(-x)$ , and similarly for  $F_\nu^\infty$  and  $F_\nu^b$ .

The remainder  $(1 - \zeta)g_{\nu,j}^\infty$ , which is neglected in the construction of  $F_\nu^\infty$ , gives rise to one of several summands which contribute to  $F_\nu^b$ . Because  $S_\nu \rightarrow \infty$ , and because the cutoff function  $m$  is smooth and compactly supported,  $\|(1 - \zeta)g_{\nu,j}^\infty\|_{L^2(\mathbb{R}^2)} \rightarrow 0$  as  $\nu \rightarrow \infty$ .

From the fact that  $s_\nu \rightarrow \infty$  and the relation  $S_\nu \geq s_\nu^3$ , it follows easily that  $\langle F_\nu^0, F_\nu^\infty \rangle \rightarrow 0$  as  $\nu \rightarrow \infty$ . Therefore since  $\|F_\nu^b\|_2 \rightarrow 0$ ,

$$\begin{aligned} \|f_\nu\|_2^2 - \|F_\nu^0\|_2^2 - \|F_\nu^\infty\|_2^2 &\rightarrow 0 \\ \|F_\nu^0\|_2^2 + \|F_\nu^\infty\|_2^2 &\rightarrow 1 = \|f_\nu\|_2^2. \end{aligned}$$

As in Section 14, the relation  $S_\nu \geq s_\nu^3 \rightarrow \infty$  also leads to

$$\|F_\nu^0 \sigma * F_\nu^\infty \sigma\|_{L^2(\mathbb{R}^3)} \rightarrow 0. \quad (15-3)$$

This requires several substeps. These are entirely parallel to those in Section 12 and Section 14, so the details are omitted.

We need to know that

$$\liminf_{\nu \rightarrow \infty} \|F_\nu^\infty\|_2 > 0.$$

This is less apparent than was the analogous statement in the proof of Proposition 2.15, because  $F_\nu^\infty$  is defined here as a sum over  $j$  that recombines different terms resulting from our partition of unity, and it must be shown that this summation does not introduce unwanted cancellation. Indeed, suppose to the contrary that  $\|F_\nu^\infty\|_2 \rightarrow 0$  for a subsequence of values of  $\nu$ . Then there must exist an index  $i$  such that for a certain sub-subsequence,

$$\int_{|\xi| \geq S_\nu} |\widehat{g_{\nu,i}}(\xi)|^2 d\xi \gtrsim 1. \tag{15-4}$$

Pass to such a sub-subsequence, substitute the representation  $f_\nu = F_\nu^0 + F_\nu^\infty + F_\nu^b$  into the definition  $g_{\nu,i} = \phi_i^*(\eta_i f_\nu)$ , and consider  $\widehat{g_{\nu,i}}(\xi)$  for  $|\xi| \gtrsim S_\nu$ . The contribution of  $F_\nu^0$  to this Fourier transform in this regime tends to zero in  $L^2(d\xi)$  norm, because  $S_\nu/s_\nu \rightarrow \infty$ . The contribution of  $F_\nu^b$  tends to zero in  $L^2$  norm, because  $F_\nu^b$  itself does so. Therefore the contribution of  $F_\nu^\infty$  to the integral (15-4) cannot tend to zero. Therefore  $\|F_\nu^\infty\|_{L^2(\mathbb{S}^2)}$  cannot tend to zero.

Since the  $L^2(\mathbb{S}^2)$  norms of both  $F_\nu^\infty$  and  $F_\nu^0$  enjoy strictly positive lower bounds, the small cross-term bound (15-3) implies as in the proof of Lemma 12.2 that

$$\limsup_{\nu \rightarrow \infty} \|f_\nu \sigma * f_\nu \sigma\|_2^2 < \mathbf{S}^4,$$

contradicting the assumption that  $\{f_\nu\}$  is an extremizing sequence.

### 16. Constants are local maxima

Theorem 1.8 asserts that constant functions are local maxima. Define

$$\Psi(f) = \|f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)}^2 \quad \text{and} \quad \Phi(f) = \frac{\Psi(f)}{\|f\|_{L^2(\mathbb{S}^2)}^4}.$$

Denote by  $\mathbf{1}$  the constant function  $\mathbf{1}(x) = 1$  for all  $x \in \mathbb{S}^2$ .

*Proof of Theorem 1.8.* Since  $\Phi(f) = \Phi(tf)$  for all  $t > 0$ , and since  $\Phi(f) \leq \Phi(|f|)$ , we may restrict attention to functions of the form  $f = \mathbf{1} + \varepsilon g$  where  $0 \leq \varepsilon \leq \delta$ ,  $g \perp \mathbf{1}$ ,  $g$  is real-valued, and  $\|g\|_{L^2(\mathbb{S}^2)} = 1$ . We may further assume that  $g(-x) = g(x)$ , by Proposition 2.7.

The constant function  $\mathbf{1}$  is a critical point for  $\Phi$ . Indeed, by rotation symmetry,  $f = \mathbf{1}$  satisfies the generalized Euler–Lagrange equation  $f = \lambda(f\sigma * f\sigma * \tilde{f}\sigma)|_{\mathbb{S}^2}$  that characterizes critical points.

A straightforward calculation gives the Taylor expansion

$$\Phi(\mathbf{1} + \varepsilon g) = \Phi(\mathbf{1}) + \varepsilon^2 \|\mathbf{1}\|_{L^2(\mathbb{S}^2)}^{-4} (6\langle g\sigma * g\sigma, \sigma * \sigma \rangle - 2\Psi(\mathbf{1})\|\mathbf{1}\|_2^{-2}\|g\|_2^2) + O(\varepsilon^3),$$

where  $O(\varepsilon^3)$  denotes a quantity whose absolute value is majorized by  $C\varepsilon^3$ , uniformly for  $g \in L^2(\mathbb{S}^2)$  satisfying  $\|g\|_2 \leq 1$ . Thus it suffices to show that

$$\sup_{\|g\|_2=1} 6\langle g\sigma * g\sigma, \sigma * \sigma \rangle < 2\Psi(\mathbf{1})\|\mathbf{1}\|_2^{-2}.$$



The quantities  $\Psi(\mathbf{1})$  and  $\|\mathbf{1}\|_2$  can be evaluated explicitly. Firstly,  $\|\mathbf{1}\|_2^2 = \sigma(\mathbb{S}^2) = 4\pi$ . Secondly,

$$(\sigma * \sigma)(x) = 2\pi |x|^{-1} \chi_{|x| \leq 2}.$$

Indeed, it follows from trigonometry that  $\sigma * \sigma(x) = a|x|^{-1} \chi_{|x| \leq 2}$  for some  $a > 0$ , and  $a$  can be evaluated by

$$(4\pi)^2 = \sigma(\mathbb{S}^2)^2 = \int_{\mathbb{R}^3} (\sigma * \sigma)(x) dx = \int_0^2 ar^{-1} \cdot 4\pi r^2 dr = 8\pi a.$$

Therefore

$$\Psi(\mathbf{1}) = \int_{\mathbb{R}^3} (\sigma * \sigma(x))^2 dx = \int_{\mathbb{R}^3} 4\pi^2 |x|^{-2} dx = 4\pi^2 \int_0^2 r^{-2} \cdot 4\pi r^2 dr = 4\pi^2 \cdot 4\pi \cdot 2 = 32\pi^3.$$

Therefore it suffices to prove that

$$\sup_{\|g\|_2=1} \langle g\sigma * g\sigma, \sigma * \sigma \rangle < \frac{1}{3} \cdot 32\pi^3 \cdot (4\pi)^{-1} = \frac{8}{3}\pi^2,$$

where the supremum is taken over all real-valued, even  $g \in L^2(\mathbb{S}^2)$  satisfying  $\|g\|_2 = 1$  and  $\int g d\sigma = 0$ .

The following key bound will be established below.

**Lemma 16.1.** *For all real-valued even functions  $g \in L^2(\mathbb{S}^2)$  satisfying  $\int g d\sigma = 0$ ,*

$$\left| \iint_{\mathbb{S}^2 \times \mathbb{S}^2} g(x)g(y)|x-y|^{-1} d\sigma(x) d\sigma(y) \right| \leq \frac{4}{5}\pi \|g\|_{L^2(\mathbb{S}^2)}^2.$$

The factor  $\frac{4}{5}\pi$  is optimal, and is attained if and only if  $g$  is a spherical harmonic of degree 2.

Now for such  $g$  satisfying  $\|g\|_2 = 1$ ,

$$\begin{aligned} \langle g\sigma * g\sigma, \sigma * \sigma \rangle &= \langle g\sigma * (\sigma * \sigma), g \rangle \\ &= 2\pi \iint_{\mathbb{S}^2 \times \mathbb{S}^2} g(x)g(y)|x-y|^{-1} d\sigma(x) d\sigma(y) \leq 2\pi \cdot \frac{4}{5}\pi = \frac{8}{5}\pi^2 < \frac{8}{3}\pi^2, \end{aligned}$$

completing the proof of Theorem 1.8. □

*Proof of Lemma 16.1.* We first recall the Funk–Hecke formula in the theory of spherical harmonics, see e.g., [Müller 1998, page 29] or [Xu 2000, Theorem A].

**Theorem 16.2** (Funk–Hecke formula). *Let  $d \geq 2$  and  $k \geq 0$  be integers. Let  $f$  be a continuous function on  $[-1, 1]$  and  $Y_k$  be a spherical harmonic of degree  $k$ , on the sphere  $S^d$ . Then for any  $x \in S^d$ ,*

$$\int_{S^d} f(x \cdot y) Y_k(y) d\sigma(y) = \lambda_k Y_k(x),$$

where  $x \cdot y$  is the usual inner product in  $\mathbb{R}^{d+1}$ , and

$$\lambda_k = \frac{\omega_d \int_{-1}^1 f(t) C_k^{(d-1)/2}(t) (1-t^2)^{(d-2)/2} dt}{C_k^{(d-1)/2}(1) \int_{-1}^1 (1-t^2)^{(d-2)/2} dt},$$

where  $\omega_d := 2\pi^{(d+1)/2}/\Gamma((d+1)/2)$  denotes the surface area of the unit sphere  $S^d$  and  $C_k^\nu(t)$  is the Gegenbauer polynomial defined by the generating function

$$(1 - 2rt + r^2)^{-\nu} = \sum_{k=0}^{\infty} C_k^\nu r^k \tag{16-1}$$

for  $0 \leq r < 1$  and  $-1 \leq t \leq 1$  and  $\nu > 0$ .

For  $\nu = 1/2$  and  $t = 1$ , the generating formula becomes  $(1 - r)^{-2/2} = \sum_{k=0}^{\infty} C_k^{1/2} r^k$ , so

$$C_k^{1/2} = 1 \quad \text{for all } k \geq 0.$$

For  $d = 2$ ,  $(d - 2)/2 = 0$  and  $\omega_d = 4\pi$ , and the relevant index  $\nu$  is  $\nu = (d - 1)/2 = 1/2$ . Therefore for  $d = 2$ ,

$$\lambda_k = 2\pi \int_{-1}^1 f(t) C_k^{1/2}(t) dt. \tag{16-2}$$

Choosing  $\nu = 1/2$  and setting  $r = 1$  in the generating function (16-1), we obtain

$$(2 - 2t)^{-1/2} = \sum_{k=0}^{\infty} C_k^{1/2}(t).$$

This formula is not entirely valid, since (16-1) only applies for  $r < 1$ ; but all calculations below can be justified by writing the corresponding formulas for  $r < 1$  and then passing to the limit  $r = 1$ . We will omit these details, and work directly with  $r = 1$ .

We also recall the following fact in [Stein and Weiss 1971, Chapter 4, Corollary 2.16]: For  $\mathbb{S}^2$ , the polynomials  $C_k^{1/2}(t)$  for  $k = 0, 1, \dots$  are mutually orthogonal with respect to the inner product  $\langle f, g \rangle = \int_{-1}^1 f(t)g(t)dt$ . So for  $f = (2 - 2t)^{-1/2}$  in (16-2) and for any  $k \geq 0$ , by orthogonality,

$$\begin{aligned} \lambda_k &= 2\pi \int_{-1}^1 (2 - 2t)^{-1/2} C_k^{1/2}(t) dt = 2\pi \int_{-1}^1 \sum_{m=0}^{\infty} C_m^{1/2}(t) C_k^{1/2}(t) dt \\ &= 2\pi \int_{-1}^1 (C_k^{1/2}(t))^2 dt = \frac{4\pi}{2k+1}, \end{aligned}$$

where the last identity follows from the normalized value of  $C_k^{1/2}(t)$  over  $(-1, 1)$ , see e.g., [Andrews et al. 1999, page 461] or [Müller 1998, 10.15, page 54]. Hence for  $f(t) = (2 - 2t)^{-1/2}$ , for  $x \in \mathbb{S}^2$ ,

$$\int_{\mathbb{S}^2} f(x \cdot y) Y_k(y) d\sigma(y) = \frac{4\pi}{2k+1} Y_k(x) \quad \text{for all } k \geq 0.$$

Now return to  $\iint g(x)g(y)|x - y|^{-1} d\sigma(x) d\sigma(y)$ . Here  $|x - y|^{-1} = (2 - 2x \cdot y)^{-1/2} = f(x \cdot y)$ , where  $f(t) = (2 - 2t)^{-1/2}$ . Since all spherical harmonics of odd degrees are odd, and since  $g \perp \mathbf{1}$ ,  $g$  may be expanded as  $g = \sum_{k=1}^{\infty} Y_{2k}$ , where each  $Y_{2k}$  is a spherical harmonic of degree  $2k$ . These are of course

pairwise orthogonal in  $L^2(\mathbb{S}^2)$ . Therefore

$$\begin{aligned} \iint g(x)g(y)|x-y|^{-1}d\sigma(x)d\sigma(y) &= \sum_{k=1}^{\infty} \langle \lambda_{2k}Y_{2k}, Y_{2k} \rangle \\ &= \sum_{k=1}^{\infty} \left\langle \frac{4\pi}{2(2k)+1} Y_{2k}, Y_{2k} \right\rangle \leq \frac{4\pi}{5} \sum_{k=1}^{\infty} \|Y_{2k}\|_2^2 = \frac{4\pi}{5} \|g\|_2^2. \end{aligned}$$

This completes the proof of Lemma 16.1.  $\square$

**Remark 16.3.** Consider inequalities of the modified form

$$\int_{\mathbb{R}^3} |(f\sigma * f\sigma)(x)|^2 w(x) dx \leq C \|f\|_{L^4(\mathbb{S}^2)}^4, \quad (16-3)$$

where  $w \geq 0$  is any radial weight. The modification consists in placing the  $L^4$  norm on the right side of the inequality instead of the  $L^2$  norm.

If the inequality holds for some  $C < \infty$ , and if  $w$  satisfies  $|\lambda_k(w)| \leq \lambda_0(w)$ , where

$$\lambda_k(w) = 2\pi \int_{-1}^1 w((2+2t)^{1/2})(2+2t)^{-1/2} C_k^{1/2}(t) dt,$$

then constant functions are (global) extremals. This holds in particular for  $w \equiv 1$ .

This is proved as follows, in the spirit of Foschi [2007]. We may assume that  $f \geq 0$ .

$$\begin{aligned} \int_{\mathbb{R}^3} (f\sigma * f\sigma)(x)^2 w(x) dx &\leq \int_{\mathbb{R}^3} ((f^2\sigma * \sigma)(x))^2 w(x) dx \\ &= 2\pi \iint_{\mathbb{S}^2 \times \mathbb{S}^2} f^2(x)f^2(y)|x+y|^{-1}w(|x+y|)d\sigma(x)d\sigma(y). \end{aligned}$$

The first inequality follows from the Cauchy–Schwarz inequality, and is an equality if  $f$  is constant modulo null sets on almost every circle (that is, the intersection of  $\mathbb{S}^2$  with an affine plane) in  $\mathbb{S}^2$ ; thus if and only if  $f$  is constant modulo  $\sigma$ -null sets. Expand  $f^2 = \sum_{k=0}^{\infty} Y_k$  in spherical harmonics. Then

$$2\pi \iint_{\mathbb{S}^2 \times \mathbb{S}^2} f^2(x)f^2(y)|x+y|^{-1}w(|x+y|)d\sigma(x)d\sigma(y) = 2\pi \sum_{k=0}^{\infty} \lambda_k \|Y_k\|_2^2 \leq 2\pi \sup_k \lambda_k \|f\|_4^4,$$

for certain coefficients  $\lambda_k$  that depend only on  $w$ . If there is a valid inequality (16-3) with  $C < \infty$ , then  $\lambda_0 < \infty$ . Thus constant functions are extremizers. If  $\max_{k \neq 0} |\lambda_k(w)| < \lambda_0(w)$ , then  $f$  is an extremizer if and only if  $f^2$  has a spherical harmonic expansion with  $Y_k = 0$  for all  $k \geq 1$ , that is, if and only if  $f^2$  is constant. For  $f \geq 0$ , this forces  $f$  to be constant.  $\square$

## 17. A variational calculation

Recall the notation  $e_\xi(x) = e^{x \cdot \xi}$ . It is natural to study  $\|\widehat{f\sigma}\|_4 / \|f\|_2$  for  $f(x) = e_\xi(x)$ , for several reasons.

- (i) Extremizers for the paraboloid  $\mathbb{P}^2 = \{x : x_3 = \frac{1}{2}|x'|^2\}$ , where  $x' = (x_1, x_2)$ , are Gaussian functions of  $x'$ ; but these are simply restrictions to  $\mathbb{P}^2$  of simple exponentials  $e^{x \cdot \xi}$  for  $\xi \in \mathbb{C}^3$  satisfying  $\text{Re}(\xi_3) < 0$ .
- (ii)  $(f\sigma * f\sigma)(x)$  is expressed for each  $x$  as an integral of a product of two factors. When  $f = e_\xi$ , the integrand becomes a constant for each  $x$ , and hence the Cauchy–Schwarz inequality becomes an equality when applied to each such integral in an appropriate way. Such equalities are the key to one proof [Foschi 2007] that Gaussians are extremal for  $\mathbb{P}^2$ .
- (iii) The functional  $\|e_\xi\sigma * e_\xi\sigma\|_2 / \|e_\xi\|_2^2$  is susceptible to a perturbative analysis for large  $|\xi|$ .
- (iv) This analysis appears more likely to be generalizable to other manifolds than  $\mathbb{S}^2$  than does the calculation of Lemma 2.4 for  $f \equiv 1$ .

For these reasons, we carry out in this section a perturbative analysis of  $\|e_\xi\sigma * e_\xi\sigma\|_2 / \|e_\xi\|_2^2$ , thereby establishing Proposition 2.19.

We will work with functions concentrated principally in a very small neighborhood of the north pole  $(0, 0, 1)$ . A point  $z \approx (0, 0, 1)$  in  $\mathbb{S}^2$  can be written as

$$(y, (1 - |y|^2)^{1/2}) = (y, 1 - \frac{1}{2}|y|^2 - \frac{1}{8}|y|^4 + O(|y|^6)),$$

where  $y \in \mathbb{R}^2$  and  $|y| < 1$ . Let  $\sigma$  denote surface measure on  $\mathbb{S}^2$ ;

$$d\sigma = (1 + \frac{1}{2}|y|^2 + O(|y|^4)) dy.$$

For  $z \in \mathbb{S}^2$  and  $\varepsilon > 0$  define

$$f_\varepsilon(z) = \varepsilon^{-1/2} e^{(z_3-1)/\varepsilon} \chi_{|(z_1, z_2)| < \frac{1}{2}} \chi_{z_3 > 0}.$$

Within the domain of  $f_\varepsilon$ , the mapping  $(z_1, z_2, z_3) \leftrightarrow (z_1, z_2)$  is a one-to-one correspondence between  $\mathbb{S}^2$  and a ball in  $\mathbb{R}^2$ .

We observe that  $f_\varepsilon$  is essentially  $\varepsilon^{-1/2} e^{-1/\varepsilon} e_\xi$ , where  $\xi = (0, 0, \varepsilon^{-1})$ ; the two functions differ by  $O(e^{-c/\varepsilon})$  in  $L^2$  norm for some  $c > 0$ . The cutoff functions are inserted for convenience in the calculation.

For  $(t, x) \in \mathbb{R}^{1+2}$ , define

$$u_\varepsilon(t, x) = \int_{\mathbb{S}^2} f_\varepsilon(z) e^{-i(x,t) \cdot z} d\sigma(z),$$

where of course  $(x, t) \cdot z = x_1 z_1 + x_2 z_2 + t z_3$ . Then

$$\begin{aligned} u_\varepsilon(t, x) &= \varepsilon^{-1/2} \int_{\mathbb{S}^2} e^{(z_3-1)/\varepsilon} e^{-ix \cdot (z_1, z_2)} e^{-it z_3} \tilde{\chi}(z) d\sigma(z) \\ &= \varepsilon^{-1/2} e^{-it} \int_{\mathbb{R}^2} e^{(-|y|^2/2 - |y|^4/8 + O(|y|^6))\varepsilon^{-1}} \\ &\quad \cdot e^{-ix \cdot y} e^{-it(-|y|^2/2 - |y|^4/8 + O(|y|^6))} (1 + |y|^2/2 + O(|y|^4)) \chi(y) dy, \end{aligned}$$

where  $\tilde{\chi}$  and  $\chi$  denote disks centered respectively at  $(0, 0, 1) \in \mathbb{S}^2$  and  $0 \in \mathbb{R}^2$ , which are independent of  $\varepsilon$ . A change of variables gives

$$u_\varepsilon(t, x) = \varepsilon^{1/2} e^{-it} \int_{\mathbb{R}^2} e^{-i\varepsilon^{1/2}x \cdot y} e^{-(1-i\varepsilon t)(|y|^2/2 + \varepsilon|y|^4/8 + O(\varepsilon^{-1}|\varepsilon^{1/2}y|^6))} \cdot (1 + \varepsilon|y|^2/2 + O(|\varepsilon^{1/2}y|^4)) \chi(\varepsilon^{1/2}y) dy.$$

Setting

$$\begin{aligned} v_\varepsilon(t, x) &= e^{-it/\varepsilon} \varepsilon^{-1/2} u_\varepsilon(-\varepsilon^{-1}t, \varepsilon^{-1/2}x) \\ &= \int_{\mathbb{R}^2} e^{-ix \cdot y} e^{-(1+it)(|y|^2/2 + \varepsilon|y|^4/8 + O(\varepsilon^{-1}|\varepsilon^{1/2}y|^6))} (1 + \varepsilon|y|^2/2 + O(|\varepsilon^{1/2}y|^4)) \chi(\varepsilon^{1/2}y) dy, \end{aligned}$$

we have

$$\|v_\varepsilon\|_{L^4(\mathbb{R}^3)}^4 = \|u_\varepsilon\|_{L^4(\mathbb{R}^3)}^4. \quad (17-1)$$

Set

$$w_\varepsilon(t, x) = \int_{\mathbb{R}^2} e^{-ix \cdot y} e^{-(1+it)(|y|^2/2 + \varepsilon|y|^4/8)} (1 + \frac{1}{2}\varepsilon|y|^2) dy \quad \text{for } \varepsilon \geq 0.$$

Using the exact definition of  $f_\varepsilon$  rather than the approximate expressions above, it is routine to verify that

$$\|w_\varepsilon\|_4^4 = \|v_\varepsilon\|_4^4 + O(\varepsilon^2) \quad \text{as } \varepsilon \rightarrow 0^+.$$

Since we are interested in first variations with respect to  $\varepsilon$  of the  $L^4$  norm at  $\varepsilon = 0$ , it will suffice to analyze  $\|w_\varepsilon\|_4^4$ . Also introduce

$$g_\varepsilon(y) = e^{-|y|^2/2 - \varepsilon|y|^4/8} \quad \text{and} \quad d\sigma_\varepsilon(y) = (1 + \varepsilon|y|^2/2) dy.$$

Then

$$\|f_\varepsilon\|_{L^2(\sigma)}^2 = \|g_\varepsilon\|_{L^2(\sigma_\varepsilon)}^2 + O(\varepsilon^2).$$

Although  $f_\varepsilon$  is not well defined in the limit  $\varepsilon = 0$ , the limit  $\lim_{\varepsilon \rightarrow 0^+} \|f_\varepsilon\|_2^2 > 0$  does exist, and we will abuse notation by writing  $\|f_0\|_2^2$  to denote this quantity. We have

$$\|f_0\|_2^2 = \int_{\mathbb{R}^2} e^{-|y|^2} dy.$$

It is a routine exercise to verify that  $\varepsilon \mapsto \|v_\varepsilon\|_4^4$  is a  $C^\infty$  function on  $[0, \infty)$ ; hence the same goes for  $\|w_\varepsilon\|_4^4$ , and for  $\|u_\varepsilon\|_4^4$  by (17-1). Similarly,  $\varepsilon \mapsto \|f_\varepsilon\|_2^2$  is  $C^\infty$  on  $[0, \infty)$ .

Consider the functional

$$\Psi(\varepsilon) = \log \frac{\|u_\varepsilon\|_{L^4}^4}{\|f_\varepsilon\|_{L^2}^4},$$

which is initially defined for  $\varepsilon > 0$  but extends continuously and differentially to  $\varepsilon = 0$ . Its derivative is

$$\partial_\varepsilon|_{\varepsilon=0} \Psi(\varepsilon) = \frac{\partial_\varepsilon \|w_\varepsilon\|_4^4|_{\varepsilon=0}}{\|w_0\|_4^4} - 2 \frac{\partial_\varepsilon|_{\varepsilon=0} \|g_\varepsilon\|_2^2}{\|g_0\|_2^2}, \quad (17-2)$$

and of course  $\Psi(0) = \log(\mathcal{R}_{\mathbb{P}^2}^4)$ , where  $\mathcal{R}_{\mathbb{P}^2}$  from (1-3) is the optimal constant for the adjoint restriction inequality for the paraboloid.

**Lemma 17.1.** 
$$\left. \frac{\partial \Psi}{\partial \varepsilon} \right|_{\varepsilon=0} > 0.$$

Proposition 2.19 follows, since by radial symmetry,  $\|e_\xi \sigma * e_\xi \sigma\|_2 / \|e_\xi \sigma\|_2^2$  depends only on  $|\xi|$ .

The most involved calculation is that of the numerator in the first term of (17-2). To begin that calculation,

$$\begin{aligned} \partial_\varepsilon \Big|_{\varepsilon=0} w_\varepsilon(t, x) &= \int \left( -\frac{1}{8}(1+it)|y|^4 + \frac{1}{2}|y|^2 \right) e^{-ix \cdot y} e^{-(1+it)|y|^2/2} dy \\ &= \left( -\frac{1}{8}(1+it)(-i/2)^{-2} \partial_t^2 + \frac{1}{2}(-i/2)^{-1} \partial_t \right) \int e^{-ix \cdot y} e^{-(1+it)|y|^2/2} dy \\ &= \left( \frac{1}{2}(1+it) \partial_t^2 + i \partial_t \right) \int e^{-ix \cdot y} e^{-(1+it)|y|^2/2} dy \\ &= \left( \frac{1}{2}(1+it) \partial_t^2 + i \partial_t \right) w_0(t, x) \\ &= \left( \frac{1}{2}(1+it) \partial_t^2 + i \partial_t \right) c_0 (1+it)^{-1} e^{-|x|^2/2(1+it)}, \end{aligned}$$

where  $c_0$  is a positive constant whose precise value will play no role, since it will ultimately appear in both the numerator and denominator of a certain ratio.

Define

$$\phi(t, x) = -\frac{1}{2}|x|^2(1+it)^{-1} - \log(1+it),$$

so that  $w_0 = c_0 e^\phi$ . The last quantity above may be written as

$$\begin{aligned} c_0 \left( \frac{1}{2}(1+it) \partial_t^2 + i \partial_t \right) e^\phi &= \frac{1}{2} c_0 (1+it) (\phi_t^2 + \phi_{tt}) e^\phi + c_0 i \phi_t e^\phi \\ &= \left( \frac{1}{2}(1+it) (\phi_t^2 + \phi_{tt}) + i \phi_t \right) w_0, \end{aligned}$$

where  $\phi_t$  and  $\phi_{tt}$  denote respectively the first and second partial derivatives of  $\phi$  with respect to  $t$ .

Now

$$\begin{aligned} \phi_t &= \frac{i}{2}|x|^2(1+it)^{-2} - i(1+it)^{-1}, \\ \phi_{tt} &= \frac{i}{2}(-2i)|x|^2(1+it)^{-3} - i(-i)(1+it)^{-2} = |x|^2(1+it)^{-3} - (1+it)^{-2}, \\ \phi_t^2 &= -\frac{1}{4}|x|^4(1+it)^{-4} + |x|^2(1+it)^{-3} - (1+it)^{-2}, \end{aligned}$$

so

$$\phi_t^2 + \phi_{tt} = -\frac{1}{4}|x|^4(1+it)^{-4} + 2|x|^2(1+it)^{-3} - 2(1+it)^{-2}.$$

Consequently

$$\begin{aligned} &\frac{1}{2}(1+it)(\phi_t^2 + \phi_{tt}) + i \phi_t \\ &= -\frac{1}{8}|x|^4(1+it)^{-3} + |x|^2(1+it)^{-2} - (1+it)^{-1} - \frac{1}{2}|x|^2(1+it)^{-2} + (1+it)^{-1} \\ &= -\frac{1}{8}|x|^4(1+t^2)^{-3}(1-it)^3 + \frac{1}{2}|x|^2(1+t^2)^{-2}(1-it)^2, \end{aligned}$$

whose real part is

$$\operatorname{Re}\left(\frac{1}{2}(1+it)(\phi_t^2 + \phi_{tt}) + i\phi_t\right) = -\frac{1}{8}|x|^4(1+t^2)^{-3}(1-3t^2) + \frac{1}{2}|x|^2(1+t^2)^{-2}(1-t^2).$$

Now  $\partial_\varepsilon \|w_\varepsilon\|_4^4 = 4 \int |w_\varepsilon|^4 \operatorname{Re}(\partial_\varepsilon w_\varepsilon / w_\varepsilon)$ , and therefore

$$\begin{aligned} \partial_\varepsilon \|w_\varepsilon\|_4^4|_{\varepsilon=0} &= 4 \iint_{\mathbb{R}^2} \operatorname{Re}\left(\frac{1}{2}(1+it)(\phi_t^2 + \phi_{tt}) + i\phi_t\right) |w_0(t, x)|^4 dx dt \\ &= c_0^4 \iint_{\mathbb{R}^2} \left(-\frac{1}{2}|x|^4(1+t^2)^{-3}(1-3t^2) + 2|x|^2(1+t^2)^{-2}(1-t^2)\right) \\ &\quad \cdot (1+t^2)^{-2} |e^{-|x|^2/2(1+it)}|^4 dx dt \\ &= c_0^4 \iint_{\mathbb{R}^2} \left(-\frac{1}{2}|x|^4(1+t^2)^{-3}(1-3t^2) + 2|x|^2(1+t^2)^{-2}(1-t^2)\right) \\ &\quad \cdot (1+t^2)^{-2} e^{-2|x|^2/(1+t^2)} dx dt. \end{aligned}$$

Substituting  $x = (1+t^2)^{1/2}\tilde{x}$  and then replacing  $\tilde{x}$  by  $x$  gives

$$\partial_\varepsilon \|w_\varepsilon\|_4^4|_{\varepsilon=0} = c_0^4 \int_{\mathbb{R}} \int_{\mathbb{R}^2} \left(-\frac{1}{2}|x|^4(1-3t^2) + 2|x|^2(1-t^2)\right) (1+t^2)^{-2} e^{-2|x|^2} dx dt.$$

By substituting  $x = 2^{-1/2}y$  in  $\mathbb{R}^2$  and then  $r = s^{1/2}$  in  $(0, \infty)$ , we derive the identities

$$\begin{aligned} \int_{\mathbb{R}^2} e^{-2|x|^2} dx &= \frac{1}{2} \int_{\mathbb{R}^2} e^{-|y|^2} dy = \pi \int_0^\infty e^{-r^2} r dr = \frac{1}{2}\pi \int_0^\infty e^{-s} ds = \frac{\pi}{2}, \\ \int_{\mathbb{R}^2} |x|^2 e^{-2|x|^2} dx &= \frac{\pi}{4} \int_0^\infty s e^{-s} ds = \frac{\pi}{4}, \\ \int_{\mathbb{R}^2} |x|^4 e^{-2|x|^2} dx &= \frac{\pi}{8} \int_0^\infty s^2 e^{-s} ds = \frac{\pi}{4}. \end{aligned}$$

Recall also that

$$\int_{\mathbb{R}} (1+t^2)^{-1} dt = \pi \quad \text{and} \quad \int_{\mathbb{R}} (1+t^2)^{-2} dt = \frac{\pi}{2}.$$

Using these formulas we obtain

$$\begin{aligned} \partial_\varepsilon \|w_\varepsilon\|_4^4|_{\varepsilon=0} &= c_0^4 \int_{\mathbb{R}} \left(-\frac{1}{2}(1-3t^2)\frac{\pi}{4} + 2(1-t^2)\frac{\pi}{4}\right) (1+t^2)^{-2} dt \\ &= \frac{\pi}{4} c_0^4 \int_{\mathbb{R}} \left(-\frac{1}{2}t^2 + \frac{3}{2}\right) (1+t^2)^{-2} dt \\ &= \frac{\pi}{4} c_0^4 \int_{\mathbb{R}} \left(-\frac{1}{2}(1+t^2)^{-1} + 2(1+t^2)^{-2}\right) dt = \frac{\pi}{4} c_0^4 \left(-\frac{\pi}{2} + 2\frac{\pi}{2}\right) = c_0^4 \frac{\pi^2}{8}. \end{aligned}$$

On the other hand,

$$\|w_0\|_4^4 = c_0^4 \int_{\mathbb{R}} \int_{\mathbb{R}^2} (1+t^2)^{-2} e^{-2|x|^2/(1+t^2)} dx dt = c_0^4 \int_{\mathbb{R}} \int_{\mathbb{R}^2} (1+t^2)^{-1} e^{-2|y|^2} dy dt = c_0^4 \frac{1}{2} \pi^2.$$

Therefore

$$\frac{\partial_\varepsilon \|w_\varepsilon\|_4^4|_{\varepsilon=0}}{\|w_0\|_4^4} = \frac{\pi^2 c_0^4 / 8}{\pi^2 c_0^4 / 2} = \frac{1}{4}.$$

The variation of  $\|g_\varepsilon\|_2^2$  must also be taken into account:

$$\begin{aligned} \partial_\varepsilon \int_{\mathbb{R}^2} g_\varepsilon(y)^2 d\sigma_\varepsilon(y) \Big|_{\varepsilon=0} &= \partial_\varepsilon \int_{\mathbb{R}^2} e^{-|y|^2 - \varepsilon \frac{1}{4}|y|^4} (1 + \varepsilon \frac{1}{2}|y|^2) dy \Big|_{\varepsilon=0} \\ &= \int_{\mathbb{R}^2} (-\frac{1}{4}|y|^4 + \frac{1}{2}|y|^2) e^{-|y|^2} dy = -\frac{2\pi}{4} + \frac{\pi}{2} = 0. \end{aligned}$$

Therefore  $2\partial_\varepsilon \|g_\varepsilon\|_{L^2(\sigma_\varepsilon)}^2 \Big|_{\varepsilon=0} / \|g_0\|_2^2 = 0$ . Putting it all together,  $\partial_\varepsilon \Psi(\varepsilon) \Big|_{\varepsilon=0} = \frac{1}{4} - 0 > 0$ .

### 18. Proof of Lemma 6.1

*Proof of Lemma 6.1.* Suppose that  $f = \chi_E$  is the characteristic function of a set  $E$ . We will begin by showing that there exist  $C < \infty$  and exponents  $s, t > 0$  such that for any set  $E$  and any index  $k$ ,

$$\sum_j |\mathcal{C}_k^j|^2 \left( |\mathcal{C}_k^j|^{-1} \int_{\mathcal{C}_k^j} |\chi_E|^p \right)^{4/p} \leq C |E|^2 \cdot \min(2^{-2k} |E|^{-1}, 2^{2k} |E|)^t \cdot \max_i \left( \frac{|E \cap \mathcal{C}_k^i|}{|E| + |\mathcal{C}_k^i|} \right)^s. \quad (18-1)$$

Indeed,

$$\begin{aligned} \sum_j |\mathcal{C}_k^j|^2 \left( |\mathcal{C}_k^j|^{-1} \int_{\mathcal{C}_k^j} \chi_E^p \right)^{4/p} &= \sum_j |\mathcal{C}_k^j|^2 |E \cap \mathcal{C}_k^j|^{4/p} |\mathcal{C}_k^j|^{-4/p} \\ &\leq \sum_j |E \cap \mathcal{C}_k^j| \cdot \max_i (|E \cap \mathcal{C}_k^i|^{4/p-1} |\mathcal{C}_k^i|^{2-4/p}) \\ &= |E| \max_i (|E \cap \mathcal{C}_k^i|^{4/p-1} |\mathcal{C}_k^i|^{2-4/p}). \end{aligned}$$

The analysis now splits into two cases. Note that  $|\mathcal{C}_k^j| \sim 2^{-2k}$  uniformly for all indices  $j$  and  $k$ . If  $2^{-2k} \geq |E|$ , then

$$\begin{aligned} |E| \max_i (|E \cap \mathcal{C}_k^i|^{4/p-1} |\mathcal{C}_k^i|^{2-4/p}) &\leq |E|^2 \max_i \left( \frac{|E \cap \mathcal{C}_k^i|}{|\mathcal{C}_k^i|} \right)^{4/p-2} \\ &\leq |E|^2 (2^{2k} |E|)^{2/p-1} \max_i \left( \frac{|E \cap \mathcal{C}_k^i|}{|\mathcal{C}_k^i|} \right)^{2/p-1}. \end{aligned}$$

Since  $1 \leq p < 2$ , we have  $2/p - 1 > 0$  and hence this is a bound of the required form (18-1). If instead  $2^{-2k} < |E|$ , then since  $4/p - 1 > 1 \geq \frac{1}{2}$ ,

$$\begin{aligned} |E| \max_i (|E \cap \mathcal{C}_k^i|^{4/p-1} |\mathcal{C}_k^i|^{2-4/p}) &= |E|^2 (2^{2k} |E|)^{-1} \max_i \left( \frac{|E \cap \mathcal{C}_k^i|}{|\mathcal{C}_k^i|} \right)^{4/p-1} \\ &\leq |E|^2 (2^{2k} |E|)^{-1} \max_i \left( \frac{|E \cap \mathcal{C}_k^i|}{|\mathcal{C}_k^i|} \right)^{1/2} \\ &= |E|^2 (2^{2k} |E|)^{-1/2} \max_i \left( \frac{|E \cap \mathcal{C}_k^i|}{|E|} \right)^{1/2}, \end{aligned}$$

which again is a bound of the desired form. Thus (18-1) is proved.



Next consider a general function  $f \in L^2(\mathbb{S}^2)$ . By sacrificing a constant factor in the inequality, we may assume that  $f$  takes the form  $f = \sum_{\alpha=-\infty}^{\infty} 2^\alpha \chi_{E_\alpha}$ , where the sets  $E_\alpha$  are pairwise disjoint and  $|E_\alpha| < \infty$ . Invoking the preceding analysis for each summand together with the triangle inequality for the sum with respect to  $\alpha$  yields

$$\|f\|_{X_p}^4 \leq C \sum_k \left( \sum_\alpha 2^\alpha |E_\alpha|^{1/2} \cdot \min(2^{-2k} |E_\alpha|^{-1}, 2^{2k} |E_\alpha|)^{t/4} \cdot \max_i \left( \frac{|E_\alpha \cap \mathcal{C}_k^i|}{|E_\alpha| + |\mathcal{C}_k^i|} \right)^{s/4} \right)^4 \quad (18-2)$$

$$\leq C \left( \sum_\alpha 2^{4\alpha} |E_\alpha|^2 \max_{k,i} \left( \frac{|E_\alpha \cap \mathcal{C}_k^i|}{|E_\alpha| + |\mathcal{C}_k^i|} \right)^s \right)^{1/2} \|f\|_2^2. \quad (18-3)$$

The second inequality in (18-3) is deduced as follows. For each integer  $r$  define

$$a_r = \sum_{\beta: |E_\beta| \in [2^r, 2^{r+1})} 2^\beta |E_\beta|^{1/2} \max_{m,i} \left( \frac{|E_\beta \cap \mathcal{C}_m^i|}{|E_\beta| + |\mathcal{C}_m^i|} \right)^{s/4} \quad \text{and} \quad b_{k,r} = \min(2^{-(r+2k)t/4}, 2^{(r+2k)t/4}).$$

Then by (18-2),

$$\begin{aligned} \|f\|_{X_p} &\leq C \left( \sum_{k=0}^{\infty} \left( \sum_{r=-\infty}^{\infty} a_r b_{k,r} \right)^4 \right)^{1/4} \\ &\leq C \left( \sum_{k=0}^{\infty} \left( \sum_r a_r^4 b_{k,r} \right) \left( \sum_r b_{k,r} \right)^3 \right)^{1/4} \leq C \left( \sum_{k=0}^{\infty} \sum_r a_r^4 b_{k,r} \right)^{1/4} \leq C \left( \sum_r a_r^4 \right)^{1/4}. \end{aligned} \quad (18-4)$$

Finally for each  $r$ , an application of Hölder's inequality with exponents 8 and  $\frac{8}{7}$  gives

$$\begin{aligned} a_r &= \sum_{\beta: |E_\beta| \sim 2^r} 2^\beta |E_\beta|^{1/2} \max_{m,i} \left( \frac{|E_\beta \cap \mathcal{C}_m^i|}{|E_\beta| + |\mathcal{C}_m^i|} \right)^{s/4}, \\ &\leq C 2^{r/2} \left( \sum_{\beta: |E_\beta| \sim 2^r} 2^{4\beta} \max_{m,i} \left( \frac{|E_\beta \cap \mathcal{C}_m^i|}{|E_\beta| + |\mathcal{C}_m^i|} \right)^{2s} \right)^{1/8} \left( \sum_{\beta: |E_\beta| \sim 2^r} 2^{4\beta/7} \right)^{7/8} \\ &\leq C \left( \sum_{\beta: |E_\beta| \sim 2^r} 2^{4\beta} |E_\beta|^2 \max_{m,i} \left( \frac{|E_\beta \cap \mathcal{C}_m^i|}{|E_\beta| + |\mathcal{C}_m^i|} \right)^s \right)^{1/8} \|f\|_2^{1/2}, \end{aligned}$$

since the sum of the finite series  $\sum_{\beta: |E_\beta| \sim 2^r} 2^{4\beta/7}$  is comparable to its largest term.

Continuing now from (18-4), we have

$$\begin{aligned} \|f\|_{X_p}^8 \|f\|_2^{-4} &\leq C \sum_\alpha 2^{2\alpha} |E_\alpha| \cdot \sup_\alpha 2^{2\alpha} |E_\alpha| \max_{k,i} \left( \frac{|E_\alpha \cap \mathcal{C}_k^i|}{|E_\alpha| + |\mathcal{C}_k^i|} \right)^s \\ &= C \|f\|_2^4 \cdot \sup_\alpha \left( (2^{2\alpha} |E_\alpha| \|f\|_2^{-2}) \max_{k,i} \left( \frac{|E_\alpha \cap \mathcal{C}_k^i|}{|E_\alpha| + |\mathcal{C}_k^i|} \right)^s \right) \\ &\leq C \|f\|_2^4 \cdot \sup_\alpha \left( (2^{2\alpha} |E_\alpha| \|f\|_2^{-2})^s \max_{k,i} \left( \frac{|E_\alpha \cap \mathcal{C}_k^i|}{|E_\alpha| + |\mathcal{C}_k^i|} \right)^s \right) \end{aligned}$$

for some  $0 < s \leq 1$ .

It remains to show that

$$X := \sup_{\alpha} \left( (2^{2\alpha} |E_{\alpha}| \|f\|_2^{-2}) \max_{k,i} \left( \frac{|E_{\alpha} \cap \mathcal{C}_k^i|}{|E_{\alpha}| + |\mathcal{C}_k^i|} \right) \right) \leq C \sup_{m,j} \Lambda_{m,j}(f)^r$$

for some positive exponent  $r$ . Choose an index  $\alpha$  for which the supremum is attained up to a factor of at most 2. Then

$$\frac{1}{2} X \leq (2^{2\alpha} |E_{\alpha}| \cdot \|f\|_2^{-2}) \max_{k,i} \left( \frac{|E_{\alpha} \cap \mathcal{C}_k^i|}{|E_{\alpha}| + |\mathcal{C}_k^i|} \right).$$

The right side is a product of two nonnegative factors, neither of which can exceed 1, so

$$2^{2\alpha} |E_{\alpha}| \|f\|_2^2 \geq X/2 \quad \text{and there exist } k \text{ and } i \text{ such that } \frac{|E_{\alpha} \cap \mathcal{C}_k^i|}{|E_{\alpha}| + |\mathcal{C}_k^i|} \geq X/4.$$

Set  $\mathcal{C} = \mathcal{C}_k^i$ . We have  $|E_{\alpha}| \geq 2^{-2\alpha-1} X \|f\|_2^2$ , and since  $|E_{\alpha} \cap \mathcal{C}| \leq 2^{-\alpha} \int_{\mathcal{C}} |f|$ ,

$$|\mathcal{C}|^{-1} \int_{\mathcal{C}} |f| \geq 2^{\alpha} \frac{|E_{\alpha} \cap \mathcal{C}|}{|\mathcal{C}|} \geq 2^{\alpha} \frac{|E_{\alpha} \cap \mathcal{C}|}{|E_{\alpha}| + |\mathcal{C}|} \geq c 2^{\alpha} X.$$

Also

$$\begin{aligned} |\mathcal{C}|^{-1} \int_{\mathcal{C}} |f| &\geq 2^{\alpha} \frac{|E_{\alpha} \cap \mathcal{C}|}{|E_{\alpha}|} \cdot \frac{|E_{\alpha}|}{|\mathcal{C}|} \geq 2^{\alpha} \frac{|E_{\alpha} \cap \mathcal{C}|}{|E_{\alpha}| + |\mathcal{C}|} |\mathcal{C}|^{-1} |E_{\alpha}| \geq c 2^{\alpha} X |\mathcal{C}|^{-1} |E_{\alpha}| \\ &\geq c 2^{\alpha} X |\mathcal{C}|^{-1} \cdot 2^{-2\alpha} \|f\|_2^2 X = c 2^{-\alpha} \|f\|_2^2 X^2. \end{aligned}$$

Taking the geometric mean of these two bounds yields

$$\frac{|\mathcal{C}|^{-1} \int_{\mathcal{C}} |f|}{|\mathcal{C}|^{-1/2} \|f\|_2} \geq c X^{3/2},$$

which by the definitions of  $X$  and  $\Lambda_{k,i}(f)$  is a bound of the desired form.  $\square$

### Acknowledgement

We are indebted to Terence Tao for bringing the question to our attention, to Diogo Oliveira e Silva for useful comments on the exposition, and to an anonymous referee for exceptionally detailed and helpful comments.

### References

- [Andrews et al. 1999] G. E. Andrews, R. Askey, and R. Roy, *Special functions*, Encyclopedia of Mathematics and its Applications **71**, Cambridge University Press, 1999. MR 2000g:33001 Zbl 0920.33001
- [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 Zbl 1171.35109
- [Bennett et al. 2009] J. Bennett, N. Bez, A. Carbery, and D. Hundertmark, “Heat-flow monotonicity of Strichartz norms”, *Anal. PDE* **2**:2 (2009), 147–158. MR 2010j:35418 Zbl 1190.35043
- [Carneiro 2009] E. Carneiro, “A sharp inequality for the Strichartz norm”, *Int. Math. Res. Not.* **2009**:16 (2009), 3127–3145. MR 2010h:35328 Zbl 1178.35090

- [Christ 2011a] M. Christ, “On extremals for a Radon-like transform”, preprint, 2011. arXiv 1106.0728
- [Christ 2011b] M. Christ, “Quasixtremals for a Radon-like transform”, preprint, 2011. arXiv 1106.0722
- [Christ and Quilodrán 2010] M. Christ and R. Quilodrán, “Gaussians rarely extremize adjoint Fourier restriction inequalities for paraboloids”, preprint, 2010. To appear in *Trans. Amer. Math. Soc.* arXiv 1012.1346v1
- [Christ and Shao 2012] M. Christ and S. Shao, “On the extremizers for an adjoint Fourier restriction inequality”, *Adv. Math.* **230** (2012), 957–97.
- [Fanelli et al. 2011] L. Fanelli, L. Vega, and N. Visciglia, “On the existence of maximizers for a family of restriction theorems”, *Bull. Lond. Math. Soc.* **43**:4 (2011), 811–817. MR 2012g:42020 Zbl 1225.42012
- [Foschi 2007] D. Foschi, “Maximizers for the Strichartz inequality”, *J. Eur. Math. Soc.* **9**:4 (2007), 739–774. MR 2008k:35389 Zbl 1231.35028
- [Hundertmark and Zharnitsky 2006] D. Hundertmark and V. Zharnitsky, “On sharp Strichartz inequalities in low dimensions”, *Int. Math. Res. Not.* **2006** (2006), Art. ID 34080. MR 2007b:35277 Zbl 1131.35308
- [Kunze 2003] M. Kunze, “On the existence of a maximizer for the Strichartz inequality”, *Comm. Math. Phys.* **243**:1 (2003), 137–162. MR 2004i:35006 Zbl 1060.35133
- [Lions 1984a] P.-L. Lions, “The concentration-compactness principle in the calculus of variations, I: The locally compact case”, *Ann. Inst. H. Poincaré Anal. Non Linéaire* **1**:2 (1984), 109–145. MR 87e:49035a Zbl 0541.49009
- [Lions 1984b] P.-L. Lions, “The concentration-compactness principle in the calculus of variations, II: The locally compact case”, *Ann. Inst. H. Poincaré Anal. Non Linéaire* **1**:4 (1984), 223–283. MR 87e:49035b Zbl 0704.49004
- [Lions 1985a] P.-L. Lions, “The concentration-compactness principle in the calculus of variations, I: The limit case”, *Rev. Mat. Iberoamericana* **1**:1 (1985), 145–201. MR 87c:49007 Zbl 0704.49005
- [Lions 1985b] P.-L. Lions, “The concentration-compactness principle in the calculus of variations, II: The limit case”, *Rev. Mat. Iberoamericana* **1**:2 (1985), 45–121. MR 87j:49012 Zbl 0704.49006
- [Moyua et al. 1999] A. Moyua, A. Vargas, and L. Vega, “Restriction theorems and maximal operators related to oscillatory integrals in  $\mathbb{R}^3$ ”, *Duke Math. J.* **96**:3 (1999), 547–574. MR 2000b:42017 Zbl 0946.42011
- [Müller 1998] C. Müller, *Analysis of spherical symmetries in Euclidean spaces*, Applied Mathematical Sciences **129**, Springer, New York, 1998. MR 2001f:33004 Zbl 0884.33001
- [Shao 2009] S. Shao, “Maximizers for the Strichartz and the Sobolev–Strichartz inequalities for the Schrödinger equation”, *Electron. J. Differential Equations* **13**:3 (2009), 1072–6691. MR 2010a:35032 Zbl 1173.35692
- [Sogge 1993] C. D. Sogge, *Fourier integrals in classical analysis*, Cambridge Tracts in Mathematics **105**, Cambridge University Press, 1993. MR 94c:35178 Zbl 0783.35001
- [Stein and Weiss 1971] E. M. Stein and G. Weiss, *Introduction to Fourier analysis on Euclidean spaces*, Princeton Mathematical Series **32**, Princeton University Press, 1971. MR 46 #4102 Zbl 0232.42007
- [Xu 2000] Y. Xu, “Funk–Hecke formula for orthogonal polynomials on spheres and on balls”, *Bull. London Math. Soc.* **32**:4 (2000), 447–457. MR 2001g:33024 Zbl 1032.33005

Received 21 Jun 2010. Revised 11 Nov 2010. Accepted 22 Dec 2010.

MICHAEL CHRIST: [mchrist@math.berkeley.edu](mailto:mchrist@math.berkeley.edu)

*University of California, Berkeley, Department of Mathematics, Berkeley, CA 94720-3840, United States*

SHUANGLIN SHAO: [slshao@ima.umn.edu](mailto:slshao@ima.umn.edu)

*Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN 55455, United States*

and

*School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540, United States*

## DISPERSION AND CONTROLLABILITY FOR THE SCHRÖDINGER EQUATION ON NEGATIVELY CURVED MANIFOLDS

NALINI ANANTHARAMAN AND GABRIEL RIVIÈRE

We study the time-dependent Schrödinger equation  $i \frac{\partial u}{\partial t} = -\frac{1}{2} \Delta u$ , on a compact Riemannian manifold on which the geodesic flow has the Anosov property. Using the notion of semiclassical measures, we prove various results related to the dispersive properties of the Schrödinger propagator, and to controllability.

### 1. Introduction

Let  $M$  be a smooth compact Riemannian manifold of dimension  $d$  (without boundary). We denote by  $\Delta$  the Laplacian on  $M$ . We are interested in understanding the regularizing properties of the Schrödinger equation

$$i \frac{\partial u}{\partial t} = -\frac{1}{2} \Delta u, \quad \text{where } u|_{t=0} \in L^2(M).$$

More precisely, given a sequence of initial conditions  $u_n \in L^2(M)$ , we investigate the asymptotic behavior of the family

$$v_n(dx) = \left( \int_0^T |e^{it\Delta/2} u_n(x)|^2 dt \right) d\text{Vol}(x) \tag{1}$$

of measures (where  $\text{Vol}$  denotes the Riemannian volume measure on  $M$ ).

We want to relate this question to the behavior of the geodesic flow, using results on propagation of singularities. For that purpose, we reformulate the question using the semiclassical formalism, and more specifically the notion of semiclassical measures. We consider a sequence of states  $(u_\hbar)_{\hbar \rightarrow 0^+}$  normalized in  $L^2(M)$  (indexed by a parameter  $\hbar > 0$  going to 0, which plays the role of Planck's constant in quantum mechanics), and for every  $t \in \mathbb{R}$  we define the following family of distributions on the cotangent bundle  $T^*M$ :

$$\mu_\hbar(t)(a) = \int_{T^*M} a(x, \xi) d\mu_\hbar(x, \xi) := \langle e^{it\Delta/2} u_\hbar | \text{Op}_\hbar(a) | e^{it\Delta/2} u_\hbar \rangle_{L^2(M)} \quad \text{for all } a \in \mathcal{C}_0^\infty(T^*M), \tag{2}$$

where  $\text{Op}_\hbar(a)$  is a  $\hbar$ -pseudodifferential operator of principal symbol  $a$  (see [Dimassi and Sjöstrand 1999], or Appendix A for a brief reminder). This construction gives a description of a state in terms of position and impulsion variables. Throughout the paper, we will denote by  $U^t := e^{it\Delta/2}$  the quantum propagator.

---

N. Anantharaman wishes to acknowledge the support of Agence Nationale de la Recherche, under the grant ANR-09-JCJC-0099-01.

MSC2000: 35B37.

Keywords: Schrödinger equation, semiclassical analysis, control theory.

By standard estimates on the norm of  $\text{Op}_{\hbar}(a)$  (the Calderón–Vaillancourt theorem), the map  $t \mapsto \mu_{\hbar}(t)$  belongs to  $L^{\infty}(\mathbb{R}; \mathcal{D}'(T^*M))$ , and is uniformly bounded in that space as  $\hbar \rightarrow 0^+$ . Thus, one can extract subsequences that converge in the weak-\* topology of  $L^{\infty}(\mathbb{R}; \mathcal{D}'(T^*M))$ . In other words, after possibly extracting a subsequence, we have

$$\mu_{\hbar}(\theta \otimes a) := \int_{\mathbb{R}} \theta(t) a(x, \xi) \mu_{\hbar}(t)(dx, d\xi) dt \xrightarrow{\hbar \rightarrow 0} \int_{\mathbb{R}} \theta(t) a(x, \xi) \mu(t)(dx, d\xi) dt \quad (3)$$

for all  $\theta \in L^1(\mathbb{R})$  and  $a \in \mathcal{C}_0^{\infty}(T^*M)$ . The main example to keep in mind is the case when  $\theta$  is the characteristic function of some interval  $[0, T]$ . In that case we can write

$$\mu_{\hbar}(\theta \otimes a) = \int_0^T \langle e^{it\Delta/2} u_{\hbar} | \text{Op}_{\hbar}(a) | e^{it\Delta/2} u_{\hbar} \rangle dt = \hbar \int_0^{T/\hbar} \langle e^{i\tau\hbar\Delta/2} u_{\hbar} | \text{Op}_{\hbar}(a) | e^{i\tau\hbar\Delta/2} u_{\hbar} \rangle d\tau.$$

In the last term we used the change of variable  $t = \hbar\tau$  to express everything in terms of the flow  $e^{i\tau\hbar\Delta/2}$ , which solves the equation  $-\hbar^2\Delta v/2 = i\hbar\partial v/\partial\tau$  with the time-parametrization of quantum mechanics. Thus, in the time-scale of quantum mechanics, we are averaging over time intervals of order  $\hbar^{-1}$ .

It follows from standard properties of  $\text{Op}_{\hbar}(a)$  that the limit  $\mu$  has the following properties:

- For almost all  $t$ ,  $\mu(t)$  is a positive measure on  $T^*M$ .
- The unitarity of  $U^t$  implies that  $\int_{T^*M} \mu(t)(dx, d\xi)$  does not depend on  $t$ ; from the normalization of  $u_{\hbar}$ , we have  $\int_{T^*M} \mu(t)(dx, d\xi) \leq 1$ , the inequality coming from the fact that  $T^*M$  is not compact, and that there may be an escape of mass to infinity.
- Define the geodesic flow  $g^{\tau} : T^*M \rightarrow T^*M$  as the Hamiltonian flow associated with the energy  $p(x, \xi) = \|\xi\|_x^2/2$ . From the *Egorov theorem*, we have

$$e^{-i\tau\hbar\Delta/2} \text{Op}_{\hbar}(a) e^{i\tau\hbar\Delta/2} = \text{Op}_{\hbar}(a \circ g^{\tau}) + O_{\tau,a}(\hbar) \quad \text{for all } \tau \in \mathbb{R}$$

and for  $a \in \mathcal{C}_0^{\infty}(T^*M)$ . At the limit  $\hbar \rightarrow 0^+$ , this implies that  $\mu(t)$  is invariant under  $g^{\tau}$  for almost all  $t$  and all  $\tau$ .

These sequences of distributions were already studied by Macià [2009]; we refer to that paper for details about the facts mentioned above. Macià was mostly interested in describing the properties of the measures  $\mu(t)$  in the case where the geodesic flow on the manifold  $M$  was not chaotic (Zoll manifolds for instance, or the flat torus [Macià 2010; Anantharaman and Macià 2011]).

In this paper, we are interested in a completely different situation where the geodesic flow has the Anosov property (manifolds of negative curvature are the main example). In this setting, the case where the initial states  $u_{\hbar}$  are eigenfunctions of the Laplacian, satisfying  $-\hbar^2\Delta u_{\hbar} = u_{\hbar}$ , has been much studied; in this particular situation  $\mu_{\hbar}(t)$  does not depend on  $t$ . The Shnirelman theorem (also called quantum ergodicity theorem) says that for a “typical” sequence of eigenfunctions  $u_{\hbar}$ , the limit  $\mu$  is the Liouville measure on the unit cotangent bundle  $S^*M$ ; see [Shnirelman 1974; Zelditch 1987; Colin de Verdière 1985] for the precise statement. It is also known, by the work of Anantharaman and Nonnenmacher, that for any sequence of eigenfunctions the limit  $\mu$  has positive entropy [Anantharaman 2008; Anantharaman and Nonnenmacher 2007; Anantharaman et al. 2009]. The aim of this paper is twofold: *extend the*

*Shnirelman theorem to the setting of the time dependent equation and prove lower bounds on the metric entropy of the measures  $\mu(t)$ . We shall also show how these results apply to the controllability problem for the Schrödinger equation.*

**2. Statement of the results**

**2a. Semiclassical large deviations.** Our first result is a generalization (and a reinforcement in the case of Anosov geodesic flows) of the quantum ergodicity theorem. Recall that the Shnirelman theorem is originally a result on orthonormal bases of eigenfunctions of the Laplacian. In order to state an analogue for solutions of the time dependent Schrödinger equation, we introduce a notion of generalized orthonormal families.

**2a1. Generalized orthonormal family.** We fix  $\alpha > 0$  and a sequence  $I(\hbar) := [a(\hbar), b(\hbar)]$  of subintervals that are of length at least  $2\alpha\hbar$  for every  $\hbar > 0$ . We also assume that  $\lim_{\hbar \rightarrow 0^+} a(\hbar) = \lim_{\hbar \rightarrow 0^+} b(\hbar) = 1$ . We denote by  $N(I(\hbar))$  the number of eigenvalues  $\lambda_j^2$  of  $\Delta$  (counted with their multiplicities) satisfying  $\hbar^2\lambda_j^2 \in I(\hbar)$ . We assume that

$$N(I(\hbar)) = \frac{\text{Vol}(M)}{(2\pi\hbar)^d} \text{Vol}(B_d(0, 1))(b(\hbar) - a(\hbar))(1 + o(1)) \tag{4}$$

(where  $\text{Vol}(M)$  is the Riemannian volume of  $M$ , and  $\text{Vol}(B_d(0, 1))$  is the volume of the unit ball in  $\mathbb{R}^d$ ). According to [Duistermaat and Guillemin 1975], we know that the Weyl law (4) holds in the case where  $b(\hbar) - a(\hbar) = 2\alpha\hbar$  if we suppose that the set of closed geodesics is of zero Liouville measure on  $S^*M$  (this is the case for Anosov geodesic flows).

We introduce the notion of generalized orthonormal family localized in the “energy window”  $I(\hbar)$ :

**Definition 2.1.** For  $\hbar > 0$ , let  $(\Omega_\hbar, \mathbb{P}_\hbar)$  be a probability space and  $u_\hbar : \Omega_\hbar \rightarrow L^2(M)$  a measurable map. We say that  $(u_\hbar(\omega))_{\omega \in (\Omega_\hbar, \mathbb{P}_\hbar)}$  is a generalized orthonormal family (GOF) in the spectral window  $I(\hbar)$  if

- $\|u_\hbar(\omega)\|_{L^2(M)} = 1 + o(1)$  as  $\hbar$  tends to 0 (uniformly for  $\omega$  in  $\Omega_\hbar$ );
- $\|(\text{Id}_{L^2(M)} - \mathbb{1}_{I(\hbar)}(-\hbar^2\Delta))u_\hbar(\omega)\|_{L^2(M)} = o(1)$  as  $\hbar$  tends to 0 (uniformly for  $\omega$  in  $\Omega_\hbar$ );
- for every  $B$  in  $\mathcal{L}(L^2(M))$ ,

$$\int_{\Omega_\hbar} \langle u_\hbar(\omega) | B | u_\hbar(\omega) \rangle_{L^2(M)} d\mathbb{P}_\hbar(\omega) = \frac{1}{N(I(\hbar))} \text{Tr}(B \mathbb{1}_{I(\hbar)}(-\hbar^2\Delta)). \tag{5}$$

We stress the fact that if  $(u_\hbar(\omega))_{\omega \in (\Omega_\hbar, \mathbb{P}_\hbar)}$  is a GOF, then  $(U^t u_\hbar(\omega))_{\omega \in (\Omega_\hbar, \mathbb{P}_\hbar)}$  is also one for every  $t$ . This is a strong requirement which is crucial in the sequel. In Section 4, we will provide two examples of GOF.

We will denote by  $\mu_{\hbar,\omega}(t)$  the (time-dependent) distribution associated to  $u_\hbar(\omega)$  by formula (2).

**2a2. Semiclassical large deviations.** The quantum ergodicity theorem says that, for a given orthonormal basis of eigenvectors of  $\Delta$ , “most of” the associated distributions on  $T^*M$  converge to the Liouville measure on the unit cotangent bundle  $S^*M := \{p = 1/2\}$  (we recall that  $p(x, \xi) = \|\xi\|_x^2/2$  is the classical energy). This holds under the assumption that the geodesic flow acts ergodically on  $S^*M$  endowed with

the Liouville measure. Here we aim for a more precise statement, and will assume that the geodesic flow has the Anosov property. Our result will, in particular, imply a reinforced version of the usual Shnirelman theorem.

We recall that the Liouville measure on  $T^*M$  is the measure given by  $d\mathcal{L} = dx d\xi$  in local coordinates. In a region where the Hamiltonian  $p$  has no critical point, one can find local symplectic coordinates  $(x_1, \dots, x_d, \xi_1, \dots, \xi_d)$  such that  $x_1 = p$ , and the Liouville measure can be decomposed into  $d\mathcal{L} = dx_1 dL_{x_1}(x, \xi)$ , where  $L_{x_1}$  is a smooth positive measure carried by the energy layer  $\{p = x_1\}$ . We shall restrict our attention to the unit cotangent bundle,  $S^*M = \{p = \frac{1}{2}\}$ , and will denote  $L = L_{1/2}$ . This is the Liouville measure on  $S^*M$ .

Given a GOF  $(u_{\hbar}(\omega))_{\omega \in (\Omega_{\hbar}, \mathbb{P}_{\hbar})}$ , our result says that for “most”  $\omega$  (in the sense of  $\mathbb{P}_{\hbar}$ ) the distributions  $\mu_{\hbar, \omega}(t)$  are close to the Liouville measure  $L$ . We will use a large deviations result due to Kifer [1992] to give an estimate on the proportion of  $\omega$  for which  $\mu_{\hbar, \omega}(t)$  is far away from  $L$ . To state our result, we need to introduce two dynamical quantities. First, we define the maximal expansion rate of the geodesic flow on  $S^*M$  as

$$\chi_{\max} := \lim_{\tau \rightarrow \pm\infty} \frac{1}{\tau} \log \sup_{\rho \in S^*M} \|d_{\rho} g^{\tau}\|.$$

This quantity gives an upper bound on the Lyapunov exponents over  $S^*M$  and it is linked to the range of validity of the semiclassical approximation in the Egorov theorem [Bouzouina and Robert 2002]. We also introduce, for every  $\delta$  in  $\mathbb{R}$  and every  $a$  in  $\mathcal{C}_o^{\infty}(T^*M, \mathbb{R})$  such that  $L(a) = 0$ ,

$$H(\delta) := \inf_{s \in \mathbb{R}} \{-s\delta + P(sa + \varphi^u)\},$$

where  $f \mapsto P(f)$  is the topological pressure of the continuous map  $f$  and  $\varphi^u$  is the infinitesimal unstable Jacobian (see Section 3 for details). The map  $\delta \mapsto -H(\delta)$  is the Legendre transform of  $s \mapsto P(sa + \varphi^u)$ , which is a smooth and convex function on  $\mathbb{R}$ . In particular,  $-H$  is a convex map on  $\mathbb{R}$  and it satisfies  $H(0) = 0$  and  $H(\delta) < 0$  for all  $\delta \neq 0$  (see Section 3c).

**Theorem 2.2.** *Suppose  $(S^*M, (g^{\tau}))$  has the Anosov property. We fix a sequence of generalized orthonormal families  $(u_{\hbar}(\omega))_{\omega \in (\Omega_{\hbar}, \mathbb{P}_{\hbar})}$  (with  $\hbar \rightarrow 0^+$ ). We fix two observables,*

- *an element  $\theta$  in  $L^1(\mathbb{R}, \mathbb{R}_+)$  such that  $\int \theta(t) dt = 1$ , and*
- *an element  $a$  in  $\mathcal{C}_o^{\infty}(T^*M, \mathbb{R})$  such that  $\int_{S^*M} a dL = 0$ .*

*Then, we have, for any  $\delta > 0$ ,*

$$\limsup_{\hbar \rightarrow 0} \frac{\log \mathbb{P}_{\hbar}(\{\omega \in \Omega_{\hbar} : \mu_{\hbar, \omega}(\theta \otimes a) \geq \delta\})}{|\log \hbar|} \leq \frac{H(\delta)}{\chi_{\max}}.$$

From this theorem and the properties of  $H(\delta)$ , one can deduce the following corollary:

**Corollary 2.3.** *Suppose  $(S^*M, (g^{\tau}))$  has the Anosov property. Fix a sequence of GOF  $(u_{\hbar}(\omega))_{\omega \in (\Omega_{\hbar}, \mathbb{P}_{\hbar})}$  (with  $\hbar \rightarrow 0^+$ ). Then, for every  $\delta > 0$ , for every  $a \in \mathcal{C}_o^{\infty}(T^*M, \mathbb{C})$  and for every function  $\theta$  in  $L^1(\mathbb{R}, \mathbb{R}_+)$ , we have*

$$\mathbb{P}_{\hbar} \left( \left\{ \omega \in \Omega_{\hbar} : \left| \mu_{\hbar, \omega}(\theta \otimes a) - \int_{S^*M} a dL \int_{\mathbb{R}} \theta(t) dt \right| \geq \delta \right\} \right) = \mathcal{O}_{a, \delta, \theta}(\hbar^{\tilde{H}(\delta)}), \tag{6}$$

where  $\tilde{H}(\delta) > 0$  depends on  $a, \theta$  and  $\delta$ .

**2a3. Comments.** As already mentioned, this result reinforces the Shnirelman theorem in the case of Anosov geodesic flows. The Shnirelman theorem (suitably adapted to the time-dependent Schrödinger equation) would simply assert that for an ergodic geodesic flow, and for every  $\delta > 0$ ,

$$\mathbb{P}_\hbar \left( \left\{ \omega \in \Omega_\hbar : \left| \mu_{\hbar, \omega}(\theta \otimes a) - \int_{S^*M} a \, dL \int_{\mathbb{R}} \theta(t) \, dt \right| \geq \delta \right\} \right) = o_{a, \delta, \theta}(1).$$

The algebraic rate of Corollary 2.3 can be compared with a classical conjecture in quantum chaos, known as the quantum variance conjecture [Feingold and Peres 1986; Eckhardt et al. 1995]. This conjecture is usually formulated for eigenfunctions of the Laplacian and states that the quantum variance behaves (modulo a prefactor related to a classical variance) like  $1/T_H(\hbar)$ , where  $T_H(\hbar)$  is the Heisenberg time. Recall that the Heisenberg time is defined as  $\hbar \bar{\rho}(\hbar)$ , where  $\bar{\rho}(\hbar)$  is the mean density of states (which is proportional to  $\hbar^{-d}$  in our case). Translated in our context, it would predict that

$$\int_{\Omega_\hbar} \left| \mu_{\hbar, \omega}(\theta \otimes a) - \int_{S^*M} a \, dL \int_{\mathbb{R}} \theta(t) \, dt \right|^2 d\mathbb{P}_\hbar(\omega) \sim V(a, \theta) \hbar^{d-1},$$

where  $V(a, \theta)$  would be a classical dynamical variance. If this conjecture is true, it implies

$$\mathbb{P}_\hbar \left( \left\{ \omega \in \Omega_\hbar : \left| \mu_{\hbar, \omega}(\theta \otimes a) - \int_{S^*M} a \, dL \int_{\mathbb{R}} \theta(t) \, dt \right| \geq \delta \right\} \right) = \mathcal{O}(\hbar^{d-1}),$$

which is stronger than our result.

Related to this kind of question, Zelditch [1994] proved that

$$\int_{\Omega_\hbar} \left| \mu_{\hbar, \omega}(\theta \otimes a) - \int_{S^*M} a \, dL \int_{\mathbb{R}} \theta(t) \, dt \right|^p d\mathbb{P}_\hbar(\omega) = \mathcal{O}(|\log \hbar|^{-p/2})$$

for all  $p \geq 1$ ; see also [Schubert 2006]. Again, his proof is written for the eigenfunction problem, but could easily be transposed to the time-dependent Schrödinger equation (see [Rivière 2009] — and note that we have to make the extra assumption  $\|u_\hbar(\omega)\|_{L^2} = 1 + \mathcal{O}(|\log \hbar|^{-1})$  uniformly in  $\omega$ ). Using the Bienaymé–Chebyshev inequality, Zelditch’s result implies that

$$\mathbb{P}_\hbar \left( \left\{ \omega \in \Omega_\hbar : \left| \mu_{\hbar, \omega}(\theta \otimes a) - \int_{S^*M} a \, dL \int_{\mathbb{R}} \theta(t) \, dt \right| \geq \delta \right\} \right) = \mathcal{O}(|\log \hbar|^{-\infty}).$$

Our theorem — although it does not say anything about the quantum variance — improves this aspect of Zelditch’s result, as we can replace  $\mathcal{O}(|\log \hbar|^{-\infty})$  by  $\mathcal{O}(\hbar^{\tilde{H}(\delta)})$ .

**2b. Entropy of semiclassical measures.** Our second result is a lower bound on the Kolmogorov–Sinai entropy of the measures  $\mu(t)$ . We will consider a sequence of normalized states  $(u_\hbar)_{\hbar \rightarrow 0^+}$  in  $L^2(M)$ . We fix two energy levels  $0 \leq E_1 < E_2$  and we suppose that the family of states is localized in the energy window  $[E_1, E_2]$ . Precisely, we make the assumption that

$$\lim_{\hbar \rightarrow 0^+} \left\| (\text{Id}_{L^2(M)} - \mathbb{1}_{[E_1, E_2]}(-\hbar^2 \Delta)) u_\hbar \right\|_{L^2(M)} = 0. \tag{7}$$



This assumption implies that each  $\mu(t)$  is a probability measure carried by the set  $\{E_1 \leq \|\xi\|_x^2 \leq E_2\}$  (it prevents escape of mass in the fibers of  $T^*M$ ). In addition, we recall that  $\mu(t)$  is invariant under the geodesic flow. Using the invariance of the energy under the geodesic flow, we see that for Lebesgue almost every  $t$ ,  $\mu(t)(dx, d\xi)$  is of the form  $\int \mu_{t,E}(dx, d\xi)\nu(dE)$ , where  $\nu$  is a positive measure on the interval  $[E_1, E_2]$  and  $\mu_{t,E}$  is a probability measure on  $\{\|\xi\|_x^2 = E\}$  invariant under the geodesic flow.

**Remark 1.** The measure  $\nu$  is independent of  $t$ . It is the weak limit (after extraction of a subsequence) of the measures  $\nu_h$  defined on  $\mathbb{R}$  by  $\nu_h([E, E']) = \|\mathbb{1}_{[E, E']}(-\hbar^2 \Delta) u_h\|^2$ .

In the following theorem,  $h_{\text{KS}}(\mu, (g^\tau))$  denotes the entropy of the invariant probability measure  $\mu$  for the geodesic flow  $g^\tau$  (its definition is recalled in Section 3).

**Theorem 2.4.** *Let  $M$  be a compact Riemannian manifold of dimension  $d$  and constant curvature  $\equiv -1$ . We fix two energy levels  $0 \leq E_1 < E_2$  and we consider a sequence  $(u_h)_{h \rightarrow 0^+}$  in  $L^2(M)$  that satisfies*

- *the energy localization  $\lim_{h \rightarrow 0} \|\text{Id}_{L^2(M)} - \mathbb{1}_{[E_1, E_2]}(-\hbar^2 \Delta) u_h\|_{L^2(M)} = 0$  and*
- *$\lim_{h \rightarrow 0} \|u_h\|_{L^2(M)} = 1$ .*

*Consider  $\mu(t)(dx, d\xi) = \int \mu_{t,E}(dx, d\xi)\nu(dE)$  a weak- $*$  limit in  $L^\infty(\mathbb{R}; \mathcal{D}'(T^*M))$  of the sequence of distributions  $\mu_h(t)$  defined in (2). Then, one has,  $\text{Leb} \otimes \nu$  almost everywhere,*

$$h_{\text{KS}}(\mu_{t,E}, (g^\tau)) \geq \frac{d-1}{2} \sqrt{E},$$

where  $h_{\text{KS}}(\mu_{t,E}, (g^\tau))$  is the Kolmogorov–Sinai entropy of  $\mu_{t,E}$ .

**Remark 2.** For the sake of simplicity, we only state and prove the results in the case of constant curvature. In principle the methods from [Anantharaman and Nonnenmacher 2007; Anantharaman et al. 2009] for general Anosov manifolds or from [Rivière 2010] for Anosov surfaces could be adapted in this setting. However, one step requires a nontrivial adaptation: see Remark 8. Modulo this extra work, the result in variable curvature would read

$$h_{\text{KS}}(\mu_{t,E}, (g^\tau)) \geq \left( \int |\varphi^u| d\mu_{t,E} - \frac{d-1}{2} \chi_{\max}(E) \right)$$

where  $\varphi^u$  is the unstable Jacobian and  $\chi_{\max}(E)$  is the maximal expansion rate of the geodesic flow on the energy layer  $\{p = E/2\}$ ; see Section 3. This lower bound may be negative (and thus trivial) if  $\chi_{\max}$  is too large compared to the average of  $\varphi^u$ . For surfaces, the adaptation of the ideas of [Rivière 2010] would lead to the better result

$$h_{\text{KS}}(\mu_{t,E}, (g^\tau)) \geq \frac{1}{2} \int |\varphi^u| d\mu_{t,E} > 0.$$

**Remark 3.** We note that  $\sqrt{E}$  is the speed of trajectories of  $g^\tau$  on the energy layer  $\{p = E/2\}$ . It is also natural to consider the geodesic flow  $\phi^\tau = g^{\tau/\sqrt{E}}$  parametrized to have speed 1 on any energy layer, and our result then reads  $h_{\text{KS}}(\mu_{t,E}, (\phi^\tau)) \geq (d-1)/2$ .

If one wants, one can avoid assumption (7) and deal with the issue of escape of mass in a different manner: Consider the space  $\mathcal{S}_0$  of smooth functions  $a$  on  $T^*M$  that are 0-homogeneous outside a compact

set. The distributions  $\mu_{\hbar}(t)$  are bounded in  $L^\infty(\mathbb{R}, \mathcal{S}'_0)$ , and one can consider convergent subsequences in the corresponding weak- $*$  topology. The corresponding limits  $\mu \in L^\infty(\mathbb{R}, \mathcal{S}'_0)$  are actually positive for almost all  $t$ , and each  $\mu(t)$  defines a probability measure on  $\widehat{T^*M}$ , the cotangent bundle compactified by spheres at infinity. We note that the flow  $\phi^t$  can be extended to the spheres at infinity. We can then write  $\mu(t) = \int \mu_{t,E}(dx, d\xi)\nu(dE)$ , where now  $\nu$  is a probability measure on  $[0, +\infty]$ . Our result reads  $h_{\text{KS}}(\mu_{t,E}, (g^\tau)) \geq \sqrt{E}(d-1)/2$  for  $0 \leq E < +\infty$ , and  $h_{\text{KS}}(\mu_{t,E}, (\phi^\tau)) \geq (d-1)/2$  for  $0 < E \leq +\infty$ .

**Remark 4** ( $u_{\hbar}$  versus  $u_n$ ). Let  $(u_n)$  be a normalized sequence in  $L^2(M)$ , and suppose we want to study the sequence of probability measures (1). No scale  $\hbar_n$  is given *a priori*. We can always choose  $\hbar_n$  such that (7) is satisfied, and apply Theorem 2.4. However, the statement of the theorem is trivial for the part of the limit measure carried on  $\{\xi = 0\}$ : It just says that  $h_{\text{KS}}(\mu_{t,0}, g^\tau) \geq 0$ . Thus, it is preferable to choose  $\hbar_n$  such that none of the limit mass goes to  $\{\xi = 0\}$ . If  $u_n$  converges weakly to 0 in  $L^2$ , this is also possible but in general (7) will no longer be satisfied (some of the mass will escape to infinity) and one must in this case use the version of the theorem stated in Remark 3. If  $u_n$  converges weakly to 0 in  $L^2$  and if one is ready to have all the mass escape to infinity (thus losing some information about the rate of escape), one can even let  $\hbar_n = 1$ . This means that one considers the “distribution”

$$\mu_n(t)(b) := \langle u_n | e^{-it\Delta/2} \text{Op}_1(b) e^{it\Delta/2} u_n \rangle_{L^2(M)},$$

defined for all  $b \in \mathcal{S}_0$ . This is the analogue of (2) in the microlocal setting [Gérard 1991]. The map  $t \mapsto \mu_n(t)$  belongs to  $L^\infty(\mathbb{R}, \mathcal{S}'_0)$ . Thus, there exists a subsequence  $(u_{n_k})_k$  and  $\mu$  in  $L^\infty(\mathbb{R}, \mathcal{S}'_0)$  such that

$$\int_{\mathbb{R} \times \widehat{T^*M}} \theta(t)b(x, \xi)\mu_{n_k}(t)(dx, d\xi) dt \xrightarrow{k \rightarrow +\infty} \int_{\mathbb{R} \times \widehat{T^*M}} \theta(t)b(x, \xi)\mu(t)(dx, d\xi) dt$$

for all  $\theta \in L^1(\mathbb{R})$  and  $b \in \mathcal{S}_0$ . Besides, as above,  $\mu(t)$  is a probability measure on the compactified cotangent bundle  $\widehat{T^*M}$ , and is invariant under the normalized geodesic flow. As  $u_n(t) = e^{it\Delta/2}u_n$  converges weakly to 0 for every  $t$  in  $\mathbb{R}$ , each  $\mu(t)$  is actually supported at infinity, and may thus be identified with a probability measure on the unit sphere bundle  $S^*M$ , invariant under the geodesic flow.

Theorem 2.4 adapted to this setting says that  $h_{\text{KS}}(\mu(t), (g^\tau)) \geq (d-1)/2$  for every  $t$  in  $\mathbb{R}$ .

**2c. Application to controllability.** Theorem 2.4, in the form given in Remark 4, implies the following observability inequality:

**Theorem 2.5.** *Let  $M$  be a compact Riemannian manifold of dimension  $d$  and constant curvature identically equal to  $-1$ . Let  $a$  be a smooth function on  $M$ , and define a closed  $g^\tau$ -invariant subset of  $S^*M$  by*

$$K_a = \{\rho \in S^*M, a^2(g^\tau(\rho)) = 0 \text{ for all } \tau \in \mathbb{R}\}.$$

*Assume that the topological entropy of  $K_a$  is less than  $(d-1)/2$ . Then, for all  $T > 0$ , there exists  $C_{T,a} > 0$  such that, for all  $u$  in  $L^2(M)$ ,*

$$\|u\|_{L^2(M)}^2 \leq C_{T,a} \int_0^T \|ae^{it\Delta/2}u\|_{L^2(M)}^2 dt. \tag{8}$$

**Remark 5.** The topological entropy of a  $(g^\tau)$ -invariant compact subset  $K$  of  $S^*M$  is related to the Kolmogorov–Sinai entropy by the variational principle [Walters 1982]

$$h_{\text{top}}(K, (g^\tau)) := \sup_{\mu \in \mathcal{M}(S^*M, g^\tau)} \{h_{\text{KS}}(\mu, (g^\tau)) : \mu(K) = 1\},$$

where  $\mathcal{M}(S^*M, g^\tau)$  is the set of probability measures on  $S^*M$  invariant under the geodesic flow. Thanks to [Barreira and Wolf 2007, Corollary 4], our assumption on the topological entropy of  $K_a$  is satisfied when the Hausdorff dimension of  $K_a$  is less than  $d$ . The converse is also true if  $K_a$  is a locally maximal subset [Pesin and Sadovskaya 2001, Theorem 4.1], that is, there exists an open neighborhood  $\mathcal{U}$  of  $K_a$  such that  $K_a = \bigcap_{\tau \in \mathbb{R}} g^\tau \mathcal{U}$ .

The proof that Theorem 2.4 implies Theorem 2.5 is given in Section 7. This follows a classical argument due to Lebeau [1992], who used it to prove that if  $M$  is an arbitrary Riemannian manifold, and if  $K_a = \emptyset$  (the “geometric control condition”), then (8) holds.

We can give an example where our assumption on the topological entropy of  $K_a$  is satisfied. Consider a closed geodesic  $\gamma$  and a small tubular neighborhood of this geodesic in  $M$  that does not contain another complete geodesic. We take  $a$  to be nonzero on the complement of this neighborhood and 0 near the closed geodesic. In this case, one has  $K_a = \gamma$  so that our condition holds. Another example, in dimension  $d = 2$ , goes as follows: Take a decomposition of the hyperbolic surface  $M$  into “hyperbolic pairs of pants” (there are  $2g - 2$  pairs of pants if  $M$  has genus  $g$ ). The boundary of each pair of pants consists of 3 simple closed geodesics. Take a function  $a$  supported in a neighborhood of the union of these  $3g - 3$  simple closed geodesics, and assume that  $a$  does not vanish on the union of these curves. Thus, any geodesic that avoids the support of  $a$  must stay inside one of the pairs of pants. If the length of each of the  $3g - 3$  boundary components is large enough, this will imply that  $K_a$  has dimension less than  $d$ , and our condition will be satisfied. The existence of a hyperbolic pants decomposition with boundary components of arbitrary large lengths follows, for instance, from [Rees 1981, Proposition 2.2]. It would be interesting to find a larger variety of geometric situations in which our assumption on  $K_a$  holds.

Following the Hilbert uniqueness method, one knows that inequality (8) implies that for any  $u_0, u_T \in L^2(M)$  and any  $T > 0$ , there exists  $f(t, x) \in L^2([0, T] \times M)$  such that the solutions of

$$i \frac{\partial u}{\partial t} + \frac{\Delta}{2} u = a(x) f(t, x)$$

with initial condition  $u|_{t=0} = u_0$  satisfy  $u|_{t=T} = u_T$ . This is called the controllability problem.

**Remark 6.** As already mentioned, this application to the controllability problem relies on the entropic estimate of Theorem 2.4, which is proved for manifolds of constant negative curvature. In Remark 2, we indicated what should be (modulo extra work) the extension of Theorem 2.4 in the case of manifolds of variable negative curvature. Let us mention what would then be the consequences for controllability. In the case of manifolds of variable negative curvature, controllability would hold under the condition that

$$P_{\text{top}}(K_a, (g^\tau), \varphi^u) < -\frac{d-1}{2} \chi_{\text{max}},$$

where  $P_{\text{top}}(K_a, (g^\tau), \varphi^u)$  is the topological pressure of  $K_a$  with respect to  $\varphi^u$  [Pesin 1997, Appendix II]. If  $M$  is of variable curvature, there is no precise relation between such a condition and the Hausdorff dimension of  $K_a$ . In the case of *surfaces* of variable negative curvature, the entropic estimate of Remark 2 would imply that controllability holds under the more general condition

$$P_{\text{top}}(K_a, (g^\tau), \frac{1}{2}\varphi^u) < 0.$$

This condition is satisfied when the Hausdorff dimension of  $K_a$  is less than 2 [Barreira and Wolf 2007, Corollary 4].

**Organization of the paper.** In Section 3, we describe some background in dynamical systems that we will need at different points of the article. In Section 4, we give two examples of GOF and apply Theorem 2.2 to them. In Sections 5 and 6, we prove Theorems 2.2 and 2.4. Finally, in Section 7, we show how to derive an observability result from Theorem 2.4. In the appendix, we give a brief review of semiclassical calculus on a manifold.

### 3. Dynamical systems background

**3a. Anosov property.** In this paper, we suppose that  $M$  is a smooth, compact, Riemannian manifold of dimension  $d$  (without boundary). The geodesic flow  $(g^\tau)$  on  $T^*M$  is the Hamiltonian flow associated to the Hamiltonian  $p(x, \xi) = \|\xi\|_x^2/2$ . We also assume that, for any  $E > 0$ , the geodesic flow  $g^\tau$  is Anosov on the energy layer  $p^{-1}(\{E/2\}) \subset T^*M$ : For all  $\rho \in p^{-1}(\{E/2\})$ , we have a decomposition

$$T_\rho p^{-1}(\{E/2\}) = E^u(\rho) \oplus E^s(\rho) \oplus \mathbb{R}X_p(\rho),$$

with  $X_p$  is the Hamiltonian vector field associated to  $p$ ,  $E^u$  the unstable space and  $E^s$  the stable space [Katok and Hasselblatt 1995]. We can introduce the infinitesimal unstable Jacobian as follows [Bowen and Ruelle 1975]:

$$\varphi^u(\rho) := -\frac{d}{d\tau}(\det(d_\rho g^\tau|_{E^u(\rho)}))_{\tau=0}.$$

**3b. Kolmogorov–Sinai entropy.** Let us recall a few facts about Kolmogorov–Sinai (or metric) entropy, which can be found for example in [Walters 1982]. Let  $(X, \mathcal{B}, T, \mu)$  be a measurable dynamical system, and  $\mathcal{P} := (P_\alpha)_{\alpha \in I}$  a finite measurable partition of  $X$ , that is, a finite collection of measurable subsets that forms a partition. Each  $P_\alpha$  is called an atom of the partition. With the convention  $0 \log 0 = 0$ , one defines

$$H_n(\mu, T, \mathcal{P}) = -\sum_{|\alpha|=n} \mu(P_{\alpha_0} \cap \dots \cap T^{-(n-1)}P_{\alpha_{n-1}}) \log \mu(P_{\alpha_0} \cap \dots \cap T^{-(n-1)}P_{\alpha_{n-1}}). \tag{9}$$

This quantity satisfies a subadditivity property

$$H_{n+m}(\mu, T, \mathcal{P}) \leq H_n(\mu, T, \mathcal{P}) + H_m(\mu, T, T^{-n}\mathcal{P}) = H_n(\mu, T, \mathcal{P}) + H_m(\mu, T, \mathcal{P}). \tag{10}$$

The first inequality is true even if the probability measure  $\mu$  is not  $T$ -invariant, while the last equality holds for  $T$ -invariant measures. A classical argument for subadditive sequences allows to define the

quantity

$$h_{\text{KS}}(\mu, T, \mathcal{P}) := \lim_{n \rightarrow \infty} \frac{H_n(\mu, T, \mathcal{P})}{n}, \quad (11)$$

the *Kolmogorov–Sinai entropy* of  $(T, \mu)$  with respect to the partition  $\mathcal{P}$ . The Kolmogorov–Sinai entropy  $h_{\text{KS}}(\mu, T)$  of  $(\mu, T)$  is then defined as the supremum of  $h_{\text{KS}}(\mu, T, \mathcal{P})$  over all finite partitions  $\mathcal{P}$  of  $X$ . In the case of a flow (for instance the dynamical system  $(S^*M, g^\tau, \mu)$ ), we define the entropy  $h_{\text{KS}}(\mu, (g^\tau)) := h_{\text{KS}}(\mu, g^1)$ . Entropy can *a priori* be infinite. However, for a smooth flow on a compact finite dimensional manifold, entropy is bounded thanks to the Ruelle inequality [1978]. In the case of the geodesic flow on a negatively curved manifold, it reads

$$h_{\text{KS}}(\mu, (g^\tau)) \leq - \int_{S^*M} \varphi^u(\rho) d\mu(\rho),$$

and equality holds if and only if  $\mu$  is the disintegration  $L$  of the Liouville measure on  $S^*M$  (defined in Section 2a2) [Pesin 1977; Ledrappier and Young 1985].

**Notation.** In the rest of this paper, we will write  $h_{\text{KS}}(\mu)$  for  $h_{\text{KS}}(\mu, (g^\tau))$ , unless we want to consider a flow different from  $(g^\tau)$ .

**3c. Topological pressure.** To conclude this section, we introduce the topological pressure of the dynamical system  $(S^*M, g^\tau)$  as the Legendre transform of the Kolmogorov–Sinai entropy [Walters 1982; Parry and Pollicott 1990; Pesin 1997]: for all  $f \in \mathcal{C}^0(S^*M, \mathbb{R})$ ,

$$P(f) = P(S^*M, (g^\tau), f) := \sup \left\{ h_{\text{KS}}(\mu) + \int_{S^*M} f d\mu : \mu \in \mathcal{M}(S^*M, g^\tau) \right\},$$

where  $\mathcal{M}(S^*M, g^\tau)$  is the set of probability measures on  $S^*M$  invariant under the geodesic flow. This defines a continuous and convex function on  $\mathcal{C}^0(S^*M, \mathbb{R})$  [Walters 1982].

We shall be particularly interested in the behavior of  $P(f)$  near  $f = \varphi^u$ . By the Ruelle inequality, we have  $P(\varphi^u) = 0$  (the sup defining  $P(\varphi^u)$  is achieved at  $\mu = L$ ; see Section 3b). Moreover, it can be proved that for any real-valued Hölder function  $f$  on  $S^*M$ , the function  $s \mapsto P(\varphi^u + sf)$  is real analytic on  $\mathbb{R}$  [Bowen and Ruelle 1975; Ruelle 1976] and its derivatives of order 1 and 2 can be computed explicitly [Parry and Pollicott 1990].

We have  $\frac{d}{ds}(P(\varphi^u + sf))|_{s=0} = \int_{S^*M} f dL$ . If  $\int_{S^*M} f dL = 0$ , the convex function  $s \mapsto P(\varphi^u + sf)$  achieves its minimum at 0. Moreover, if  $\int_{S^*M} f dL = 0$ , then we have

$$\frac{d^2}{ds^2}(P(\varphi^u + sf))|_{s=0} = \sigma^2(f),$$

where

$$\sigma^2(f) := \lim_{T \rightarrow +\infty} \frac{1}{T} \int_{S^*M} \left( \int_0^T f \circ g^\tau(\rho) d\tau \right)^2 dL(\rho)$$

is called the dynamical variance of the function  $f$ . It is known that  $\sigma^2(f)$  vanishes if and only if  $f$  is of the form  $f = \frac{d}{d\tau}(h \circ g^\tau)|_{\tau=0}$  for some function  $h$ . In this case, one says that  $f$  is a coboundary.

**3d. Kifer’s large deviation upper bound.** We shall use the following result, due to Kifer [1992], and valid for more general Anosov flows:

$$\lim_{T \rightarrow \infty} \frac{1}{T} \log \int_{S^*M} \exp\left(\int_0^T a \circ g^\tau(\rho) d\tau\right) dL(\rho) = P(a + \varphi^u), \tag{12}$$

for all continuous  $a$ . In fact, we will only use that the lim sup is uniform for  $a$  running over compact sets in the  $\mathcal{C}^1$  topology (this property can be derived from the proof of [Kifer 1992, Theorem 3.2]).

**Remark 7.** This result implies the following strengthened version of the Birkhoff ergodic theorem. Fix  $a$  such that  $\int_{S^*M} a dL = 0$ , and fix  $\delta > 0$ . Then

$$\begin{aligned} \limsup \frac{1}{T} \log L\left(\left\{\rho \in S^*M : \frac{1}{T} \int_0^T a \circ g^\tau(\rho) d\tau > \delta\right\}\right) &\leq \inf_{s \geq 0} \{-s\delta + P(sa + \varphi^u)\} \\ &= \inf_{s \in \mathbb{R}} \{-s\delta + P(sa + \varphi^u)\} = H(\delta). \end{aligned}$$

Similarly, for  $\delta < 0$ , one has

$$\limsup \frac{1}{T} \log L(\{\rho \in S^*M : \frac{1}{T} \int_0^T a \circ g^\tau(\rho) d\tau < \delta\}) \leq H(\delta).$$

The function  $-H$ , which is the Legendre transform of  $s \mapsto P(\varphi^u + sa)$ , satisfies  $H(\delta) = 0$ , is convex and is positive for  $\delta \neq 0$  (it is infinite for  $\delta \neq 0$  if  $a$  is a coboundary).

### 4. Examples of generalized orthonormal families

In this section, we provide two examples of GOF and show how Theorem 2.2 applies to them. Our examples are of distinct types: basis of eigenvectors of  $\Delta$  and truncated Dirac distributions. In the first example, Theorem 2.2 provides a strengthened version of Shnirelman’s theorem for Anosov flows.

**4a. Orthonormal basis of eigenvectors.** Consider  $(\psi_n)_{n \in \mathbb{N}}$ , an orthonormal basis of  $L^2(M)$  made of eigenfunctions of  $\Delta$ , that is, there exists a sequence  $0 = \lambda_0 < \lambda_1 \leq \dots \leq \lambda_n \leq \dots$  such that for every  $n$  in  $\mathbb{N}$ ,

$$\Delta \psi_n = -\lambda_n^2 \psi_n.$$

For  $\hbar > 0$ , we take  $\Omega_\hbar := \{n \in \mathbb{N} : \hbar^2 \lambda_n^2 \in [1 - \alpha\hbar, 1 + \alpha\hbar]\}$ , where  $\alpha$  is some fixed positive number. In this case, the probability measure is given by  $\mathbb{P}_\hbar := \frac{1}{|\Omega_\hbar|} \sum_{n \in \Omega_\hbar} \delta_n$  and the measurable map is given by  $u_\hbar(n) := \psi_n$ . Applying Corollary 2.3 to this example, we find that for every  $a$  in  $\mathcal{C}_o^\infty(T^*M)$ , and for every  $\delta > 0$ , there exists  $\tilde{H}(\delta) > 0$  such that

$$\frac{1}{|\Omega_\hbar|} \left| \left\{ n \in \Omega_\hbar : \left| \mu_{\hbar,n}(a) - \int_{S^*M} a dL \right| \geq \delta \right\} \right| = \mathcal{O}_{a,\delta}(\hbar^{\tilde{H}(\delta)}).$$

Shnirelman’s theorem provides a  $o_{a,\delta}(1)$ , and using the results from [Zelditch 1994] on eigenfunctions of  $\Delta$ , one can obtain a  $\mathcal{O}_{a,\delta,p}(|\log \hbar|^{-p})$  for arbitrarily large  $p$ .

**4b. Truncated Dirac distributions.** The second class of examples we will consider is given by families of vectors constructed from the Dirac distributions. For  $y$  in  $M$ , we denote by  $\delta_y$  the Dirac distribution given by  $\langle \delta_y, f \rangle := f(y)$  (where  $f$  is in  $\mathcal{C}^\infty(M)$ ). To construct our GOF, we will project  $\delta_y$  on  $L^2(M)$ . To do this, recall that we have defined  $I(\hbar) := [a(\hbar), b(\hbar)]$ , where  $b(\hbar) - a(\hbar) \geq 2\alpha\hbar$ , and that we have defined  $N(I(\hbar)) := |\{n : \hbar^2\lambda_n^2 \in I(\hbar)\}|$ . Using this notation, we can introduce a truncated Dirac distribution by

$$\delta_y^\hbar := \left( \frac{\text{Vol}_M(M)}{N(I(\hbar))} \right)^{1/2} \mathbb{1}_{I(\hbar)}(-\hbar^2\Delta) \delta_y.$$

According to (global and local) Weyl laws from [Duistermaat and Guillemin 1975] and from [Sogge and Zelditch 2002, Theorem 1.2]), we know that in the Anosov case,

$$\left( M, \frac{\text{Vol}_M}{\text{Vol}_M(M)}, \delta_y^\hbar \right) \text{ is a GOF in the spectral window } I(\hbar).$$

Applying Corollary 2.3 to this example, we find that for every  $a$  in  $\mathcal{C}_0^\infty(T^*M, \mathbb{C})$ , for every  $\theta$  in  $L^1(\mathbb{R}, \mathbb{R}_+)$  and for every  $\delta > 0$ , there exists  $\tilde{H}(\delta) > 0$  such that

$$\text{Vol}_M \left( \left\{ y \in M : \left| \mu_{\hbar, y}(\theta \otimes a) - \int_{S^*M} a \, dL \int_{\mathbb{R}} \theta(t) \, dt \right| \geq \delta \right\} \right) := \mathbb{O}_{a, \theta, \delta}(\hbar^{\tilde{H}(\delta)}).$$

Thus, if we choose  $y$  randomly on  $M$  according to the volume measure, and consider the solution of the Schrödinger equation  $e^{it\Delta/2}\delta_y^\hbar$ , our result says that we have convergence of the associated semiclassical measure to the uniform measure, for most  $y$  (in the probability sense, and with an explicit bound) as  $\hbar$  tends to 0. Taking a subsequence  $(\hbar_n)_n$  that tends to 0 fast enough, we can apply the Borel–Cantelli lemma and derive convergence for almost every  $y$  [Rivière 2009]. An interesting question would be to understand more precisely for which subsequences  $(\hbar_n)$  we have convergence for almost every  $y$ .

**4c. Coherent states.** Similar results could, in principle, apply to bases of coherent states (e.g., gaussian states). Such bases can be constructed easily in euclidean situations; see [Rivière 2009] for an application of Theorem 2.2 to the “cat map” toy model. However, on an arbitrary manifold, it seems difficult to construct bases of coherent states meeting all the requirements of the definition of a GOF, which are actually quite strong.

## 5. Proof of Theorem 2.2

The proof has two steps. To begin with, we combine the Bienaymé–Chebyshev inequality and the Egorov theorem to obtain a first bound (Section 5b). Then we apply a large deviations estimate due to Kifer [1992] to obtain a bound in terms of the topological pressure. This proof follows the steps of Zelditch [1994], the new input being

- the use of the exponential function  $x \mapsto e^x$  in Section 5b instead of the power functions  $x \mapsto x^p$ ;
- the use of Kifer’s large deviation result for the geodesic flow instead of the central-limit theorem;<sup>1</sup>

<sup>1</sup>Rigorously speaking, one cannot say that the LDP is stronger than the CLT. When the large deviation principle holds with a rate function that is  $C^2$  and strictly convex, one usually expects to have a central limit theorem; the variance of the limiting

- a more careful treatment of the trace asymptotics (Lemma 5.3) to make sure that the remainder term is not larger than the leading term for the symbols we consider.

We fix  $\theta$  an element of  $L^1(\mathbb{R}, \mathbb{R}_+)$  such that  $\int \theta(t) dt = 1$ . Let  $a$  be an element in  $\mathcal{C}_0^\infty(T^*M, \mathbb{R})$  that satisfies  $\int_{S^*M} a dL = 0$ . Recall that we defined

$$\chi_{\max} := \lim_{\tau \rightarrow \pm\infty} \frac{1}{\tau} \log \sup_{\rho \in S^*M} \|d_\rho g^\tau\|.$$

Since the states  $u_\hbar(\omega)$  are uniformly microlocalized in a thin neighborhood of  $S^*M$ , we can assume that  $a$  is compactly supported in a tubular neighborhood  $p^{-1}([\frac{1}{2} - \eta, \frac{1}{2} + \eta])$  of  $S^*M$  (with  $\eta > 0$  arbitrarily small). Letting  $\chi_\eta = \chi_{\max} \sqrt{1 + 2\eta}$ , we have, for all  $\tau \in \mathbb{R}$ , for all  $\rho \in T^*M$  and for all  $\alpha$ ,

$$\|\partial^\alpha (a \circ g^\tau)(\rho)\| \leq C_{a,\alpha} e^{\chi_\eta |\alpha| |\tau|}.$$

**5a. Long-time Egorov theorem.** We fix  $c$  such that  $c\chi_\eta < \frac{1}{2}$ . The positive quantization  $\text{Op}_\hbar^+$  procedure described in Appendix A satisfies the following “long time Egorov property”:

$$\|U^{-\tau\hbar} \text{Op}_\hbar^+(a) U^{\tau\hbar} - \text{Op}_\hbar^+(a \circ g^\tau)\|_{L^2(M) \rightarrow L^2(M)} = \mathcal{O}_a(\hbar^{1/2-\nu}) \quad \text{for all } |\tau| \leq c|\log \hbar|, \quad (13)$$

where  $\nu := c\chi_\eta$ ; see [Anantharaman and Nonnenmacher 2007].

**Lemma 5.1.** For every  $\delta_0 > 0$ , there exists  $\hbar_0$  (depending on  $a, \theta$  and  $\delta_0$ ) such that for every  $\hbar < \hbar_0$ ,

$$\left\| \int \theta(t) U^{-t} \left( \text{Op}_\hbar^+(a) - \frac{1}{2T} \int_{-T}^T \text{Op}_\hbar^+(a \circ g^\tau) d\tau \right) U^t dt \right\|_{L^2(M) \rightarrow L^2(M)} \leq \delta_0 \quad \text{for every } |T| \leq c|\log \hbar|.$$

*Proof.* The proof of this lemma relies on the application of the Egorov property (13). For  $T$  a real number such that  $|T| \leq c|\log \hbar|$ , we have

$$\int \theta(t) U^{-t} \left( \frac{1}{2T} \int_{-T}^T \text{Op}_\hbar^+(a \circ g^\tau) d\tau \right) U^t dt = \frac{1}{2T} \int_{-T}^T \int \theta(t) U^{-t-\tau\hbar} \text{Op}_\hbar^+(a) U^{t+\tau\hbar} dt d\tau + \mathcal{O}_a(\hbar^{1/2-\nu}).$$

We make the change of variables  $t' = t + \tau\hbar$  and use the fact that  $\|\theta(\cdot) - \theta(\cdot - \tau)\|_{L^1} \xrightarrow{\tau \rightarrow 0} 0$  to conclude. □

**5b. Bienaymé–Chebyshev and Jensen’s inequality.** For simplicity of notation, we will denote the quantity we want to bound as follows:

$$\mathbb{P}_\hbar(\theta \otimes a, \delta) := \mathbb{P}_\hbar(\{\omega \in \Omega_\hbar : \mu_{\hbar,\omega}(\theta \otimes a) \geq \delta\}).$$

The first step is to combine the previous lemma with the Bienaymé–Chebyshev inequality to obtain a bound on  $\mathbb{P}_\hbar(\theta \otimes a, \delta)$ .

---

gaussian being the second derivative of the rate function at its minimum. Formally, one makes a Taylor expansion of order 2 of the LDP near the minimum of the rate function to derive a gaussian behavior. However, the implementation of this idea requires a very precise and strong version of the LDP, and in practice one prefers to prove the CLT independently.



**Lemma 5.2.** *Let  $\delta, \delta_0 > 0$  be arbitrary positive numbers. For  $s \in \mathbb{R}$ , let*

$$a_s(T(\hbar), \rho) := \exp\left(s \int_{-T(\hbar)}^{T(\hbar)} a \circ g^\tau(\rho) d\tau\right),$$

where  $T(\hbar) = c|\log \hbar|$  (and  $c$  is such that  $c\chi_\eta < 1/2$ ). Then, given  $s > 0$  and for  $\hbar$  small enough, one has

$$\mathbb{P}_\hbar(\theta \otimes a, \delta) \leq 2 \frac{e^{(-2\delta+4\delta_0)sT(\hbar)}}{N(I(\hbar))} \text{Tr}[\mathbb{1}_{I(\hbar)}(-\hbar^2\Delta) \text{Op}_\hbar^+(a_s(T(\hbar), \cdot))]. \quad (14)$$

*Proof.* Let  $s > 0$ . A direct application of the Bienaymé–Chebyshev inequality allows us to write

$$\mathbb{P}_\hbar(\theta \otimes a, \delta) := \mathbb{P}_\hbar(\{\omega \in \Omega_\hbar : \mu_{\hbar,\omega}(\theta \otimes a) \geq \delta\}) \leq e^{-2s\delta T(\hbar)} \int_{\Omega_\hbar} \exp(2sT(\hbar)\mu_{\hbar,\omega}(\theta \otimes a)) d\mathbb{P}_\hbar(\omega).$$

We can now use Lemma 5.1 and deduce that, for  $\hbar$  small enough,

$$\mathbb{P}_\hbar(\theta \otimes a, \delta) \leq e^{-2s\delta T(\hbar)} \int_{\Omega_\hbar} \exp\left(s\mu_{\hbar,\omega}\left(\theta \otimes \left(\int_{-T(\hbar)}^{T(\hbar)} a \circ g^\tau d\tau\right)\right) + 2s\delta_0 T(\hbar)\|u_\hbar(\omega)\|^2\right) d\mathbb{P}_\hbar(\omega).$$

Using the fact that  $\|u_\hbar(\omega)\| = 1 + o(1)$  uniformly for  $\omega$  in  $\Omega_\hbar$ , the quantity  $e^{2s\delta_0 T(\hbar)\|u_\hbar(\omega)\|^2}$  is uniformly bounded by  $e^{3s\delta_0 T(\hbar)}$  for  $\hbar$  small enough. The map  $x \mapsto e^{sx}$  is convex and we can use Jensen's inequality to write

$$\mathbb{P}_\hbar(\theta \otimes a, \delta) \leq e^{s(-2\delta+3\delta_0)T(\hbar)} \int_{\Omega_\hbar} \mu_{\hbar,\omega}\left(\exp\left(s\mu_{\hbar,\omega}(\theta \otimes 1)\left(\int_{-T(\hbar)}^{T(\hbar)} a \circ g^\tau d\tau\right)\right) \otimes \theta\right) \frac{d\mathbb{P}_\hbar(\omega)}{\mu_{\hbar,\omega}(\theta \otimes 1)}.$$

Using again that  $\|u_\hbar(\omega)\| = 1 + o(1)$  uniformly for  $\omega$  in  $\Omega_\hbar$  and that  $\theta$  is nonnegative and  $\int \theta(t) dt = 1$ , one has

$$\mu_{\hbar,\omega}(\theta \otimes 1) = 1 + o(1),$$

uniformly in  $\omega$  for  $\hbar$  small enough. All this can be summarized as follows:

$$\mathbb{P}_\hbar(\theta \otimes a, \delta) \leq 2e^{s(-2\delta+4\delta_0)T(\hbar)} \int_{\Omega_\hbar} \mu_{\hbar,\omega}(\theta \otimes a_s(T(\hbar), \cdot)) d\mathbb{P}_\hbar(\omega).$$

Note that the function  $a_s(T(\hbar), \cdot)$  belongs to the class of symbols  $S_v^{0,k_0}(T^*M)$ , where  $v := c\chi_\eta < 1/2$  and  $k_0 := 2cs\|a\|_\infty$  (Appendix A); moreover  $a_s(T(\hbar), \cdot)$  is constant in a neighborhood of infinity. The previous inequality can be rewritten as

$$\mathbb{P}_\hbar(\theta \otimes a, \delta) \leq 2e^{(-2\delta+4\delta_0)sT(\hbar)} \int \theta(t) \int_{\Omega_\hbar} \langle u_\hbar(\omega) | U^{-t} \text{Op}_\hbar^+(a_s(T(\hbar), \cdot)) U^t | u_\hbar(\omega) \rangle d\mathbb{P}_\hbar(\omega) dt.$$

We recall that if  $(u_\hbar(\omega))_{\omega \in (\Omega_\hbar, \mathbb{P}_\hbar)}$  is a GOF then for every  $t$  in  $\mathbb{R}$ ,  $(U^t u_\hbar(\omega))_{\omega \in (\Omega_\hbar, \mathbb{P}_\hbar)}$  is also a GOF. Using point 3 of the definition of a GOF, we get for  $\hbar$  small enough the bound

$$\mathbb{P}_\hbar(\theta \otimes a, \delta) \leq \frac{2e^{(-2\delta+4\delta_0)sT(\hbar)}}{N(I(\hbar))} \text{Tr}[\mathbb{1}_{I(\hbar)}(-\hbar^2\Delta) \text{Op}_\hbar^+(a_s(T(\hbar), \cdot))]. \quad \square$$

**5c. Trace asymptotics.** We now have to estimate (from above) the trace

$$\mathrm{Tr}[\mathbb{1}_{I(\hbar)}(-\hbar^2 \Delta) \mathrm{Op}_\hbar^+(a_s(T(\hbar), \cdot))]. \tag{15}$$

We first underline that, for every  $\hbar > 0$ , there exist energy levels  $E_1 < \dots < E_P$  (depending on  $\hbar$ ) such that

$$I(\hbar) = [a(\hbar), b(\hbar)] \subset \bigsqcup_{p=1}^P [E_p - \alpha\hbar, E_p + \alpha\hbar] \subset [a(\hbar) - \alpha\hbar, b(\hbar) + \alpha\hbar],$$

for some fixed positive  $\alpha$ . Note that  $P = \mathcal{O}((b(\hbar) - a(\hbar))/\hbar)$ . We decompose (15) into

$$\sum_{p=1}^P \mathrm{Tr}[\mathbb{1}_{[E_p - \alpha\hbar, E_p + \alpha\hbar]}(-\hbar^2 \Delta) \mathrm{Op}_\hbar^+(a_s(T(\hbar), \cdot))].$$

We shall bound each term of the previous sum (uniformly with respect to  $p$ ), using standard trace estimates, and then sum over  $p$ . We consider for instance the interval  $[1 - \alpha\hbar, 1 + \alpha\hbar]$ , and recall how to determine the asymptotic behavior of

$$\mathrm{Tr}[\mathbb{1}_{[1 - \alpha\hbar, 1 + \alpha\hbar]}(-\hbar^2 \Delta) \mathrm{Op}_\hbar^+(a_s(T(\hbar), \cdot))].$$

Introduce a function  $f$  that is  $\mathcal{C}^\infty$ , compactly supported in a small neighborhood of 1, equal to 1 in a neighborhood of 1 and taking values in  $[0, 1]$ . We shall also use a function  $\chi$  in  $\mathcal{S}(\mathbb{R}^d)$  whose Fourier transform is compactly supported in a small neighborhood of 0, containing no period of the closed geodesics of  $(g^\tau)$  on  $S^*M$ . We assume that  $\chi \geq 0$  and that it is greater than 1 on  $[-\alpha, \alpha]$ . Using the fact that the quantization is positive, we can bound the previous quantity as

$$\mathrm{Tr}[\mathbb{1}_{[1 - \alpha\hbar, 1 + \alpha\hbar]}(-\hbar^2 \Delta) \mathrm{Op}_\hbar^+(a_s(T(\hbar), \cdot))] \leq \mathrm{Tr}\left[f(-\hbar^2 \Delta) \chi\left(\frac{-\hbar^2 \Delta - 1}{\hbar}\right) \mathrm{Op}_\hbar^+(a_s(T(\hbar), \cdot))\right]. \tag{16}$$

The study of this last quantity now follows well-known lines. We use the Fourier inversion formula,

$$2\pi \chi\left(\frac{E-1}{\hbar}\right) = \int_{\mathbb{R}} e^{i(E-1)\tau/\hbar} \hat{\chi}(\tau) d\tau.$$

As a consequence, the right hand side of (16) can be written as

$$\frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\tau/\hbar} \mathrm{Tr}(\mathrm{Op}_\hbar^+(a_s(T(\hbar), \cdot)) U^{2\tau\hbar} f(-\hbar^2 \Delta)) \hat{\chi}(\tau) d\tau.$$

The asymptotic behavior of the trace comes from an asymptotic expansion of the kernel of the operator  $\mathrm{Op}_\hbar^+(a_s(T(\hbar), \cdot)) U^{2\tau\hbar} f(-\hbar^2 \Delta)$ . This expansion is given by the theory of Fourier integral operators [Dimassi and Sjöstrand 1999, Chapter 11; Zworski 2012, Chapter 10]. The trace is then expressed as the integral of the kernel over the diagonal, and the asymptotic behavior of this integral is determined by the method of stationary phase [Dimassi and Sjöstrand 1999, Chapter 11].

**Lemma 5.3.** *For every integer  $N \geq 1$ , we have*

$$\begin{aligned} \text{Tr} \left[ f(-\hbar^2 \Delta) \chi \left( \frac{-\hbar^2 \Delta - 1}{\hbar} \right) \text{Op}_\hbar^+ (a_s(T(\hbar), \cdot)) \right] \\ = \frac{1}{(2\pi \hbar)^{d-1}} \left( \sum_{n=0}^{N-1} \hbar^n \int_{S^*M} D^{2n} a_s(T(\hbar), \rho) dL(\rho) + \mathbb{O}_{a, \chi, \theta, N}(\hbar^{N(1-2\nu) - \beta\nu - k_0}) \right), \end{aligned}$$

where  $\beta > 0$  depends only on the dimension of  $M$ , and where  $D^{2n}$  is a differential operator of order  $2n$  on  $T^*M$  (depending on the cutoff functions and on the choice of the quantization  $\text{Op}_\hbar^+$ ).

There are many references for these kind of estimates. For instance, a very similar calculation is done by Schubert [2006, Proposition 1] (he stops at  $N = 1$  but the stationary phase method actually provides asymptotic expansions at any order).

Recall that  $\nu = c\chi_\eta < \frac{1}{2}$ . It is important here to note that  $a_s(T(\hbar), \cdot)$  belongs to the class  $S_v^{0, k_0}(T^*M)$ , and that the observable  $a_s(T(\hbar), x, \xi)$  satisfies the particular property that  $D^{2n} a_s(T(\hbar), \rho)$  is of the form  $a_s(T(\hbar), x, \xi) b_{2n}(x, \xi)$ , with  $\|b_{2n}\|_\infty = \mathbb{O}(|s|^{2n} \hbar^{-2n\nu})$  as  $\hbar \rightarrow 0$  and  $s \rightarrow \infty$ . If  $s$  stays in a bounded interval, and if we choose  $N$  large enough accordingly, this implies that

$$\text{Tr} \left[ f(-\hbar^2 \Delta) \chi \left( \frac{-\hbar^2 \Delta - 1}{\hbar} \right) \text{Op}_\hbar^+ (a_s(T(\hbar), \cdot)) \right] \leq \frac{1}{(2\pi \hbar)^{d-1}} \left( \int_{S^*M} a_s(T(\hbar), \rho) dL(\rho) \right) (1 + \mathbb{O}(\hbar^{1-2\nu})).$$

Combing this with Lemma 5.2 and using the Weyl law (4), we finally have, for every  $N \geq 1$  and  $\hbar$  small enough,

$$\mathbb{P}_\hbar(\theta \otimes a, \delta) \leq C e^{(-2\delta + 4\delta_0)sT(\hbar)} \left( \int_{S^*M} a_s(T(\hbar), \rho) dL(\rho) \right) (1 + \mathbb{O}(\hbar^{1-2\nu})), \tag{17}$$

for some constant  $C$  that does not depend on  $\hbar$ .

**5d. A large deviations bound.** To conclude, we use Kifer’s large deviations result (12). For our proof, we only need an upper bound on the quantity

$$\int_{S^*M} \exp \left( s \int_{-T}^T a \circ g^\tau(\rho) d\tau \right) dL(\rho).$$

Compared with (12), there is a parameter  $s$  in the exponential that stays in a bounded interval  $I$ . Following the proof of the upper bound (12) in [Kifer 1992, Section 3], one can say that for every  $\delta' > 0$  and any bounded interval  $I$  in  $\mathbb{R}_+$ , there exists  $c_{\delta'} > 0$  and  $n(\delta', I) \in \mathbb{N}$  such that for every  $T \geq n(\delta', I)$  and every  $s$  in  $I$ ,

$$\int_{S^*M} \exp \left( s \int_{-T}^T a \circ g^\tau(\rho) d\tau \right) dL(\rho) \leq c_{\delta'} e^{T\delta'} e^{2TP(sa + \varphi^u)}. \tag{18}$$

This last bound will allow us to conclude. In fact, combining this inequality to the bound (17) on  $\mathbb{P}_\hbar(\theta \otimes a, \delta)$ , we find that

$$\mathbb{P}_\hbar(\theta \otimes a, \delta) \leq C e^{(-2\delta + 4\delta_0)sT(\hbar)} e^{T(\hbar)\delta'} e^{2T(\hbar)P(sa + \varphi^u)},$$

where the constant  $C$  depends on the various parameters but not on  $\hbar$ . This implies

$$\limsup_{\hbar \rightarrow 0} \frac{\log (\mathbb{P}_{\hbar}(\theta \otimes a, \delta))}{c|\log \hbar|} \leq \delta' + (-2\delta + 4\delta_0)s + 2P(sa + \varphi^u).$$

This last inequality holds for any  $\delta_0 > 0$  and any  $\delta' > 0$ . It implies that for every  $s > 0$  in the interval  $I$ ,

$$\limsup_{\hbar \rightarrow 0} \frac{\log (\mathbb{P}_{\hbar}(\theta \otimes a, \delta))}{c|\log \hbar|} \leq -2s\delta + 2P(sa + \varphi^u) \quad \text{for all } c \in \left(0, \frac{1}{2\chi_{\max}}\right).$$

In particular, we find that

$$\limsup_{\hbar \rightarrow 0} \frac{\log (\mathbb{P}_{\hbar}(\theta \otimes a, \delta))}{|\log \hbar|/(2\chi_{\max})} \leq 2 \inf_{s \in \mathbb{R}_+} \{-s\delta + P(sa + \varphi^u)\} \quad \text{for all } \delta \in \mathbb{R}.$$

Since  $\delta > 0$ , we have  $\inf_{s \in \mathbb{R}_+} \{-s\delta + P(sa + \varphi^u)\} = \inf_{s \in \mathbb{R}} \{-s\delta + P(sa + \varphi^u)\}$ . This concludes the proof of Theorem 2.2.

### 6. Proof of Theorem 2.4

In this section, we assume that  $M$  has constant sectional curvature  $-1$ , and we fix two energy levels  $0 \leq E_1 < E_2$  and consider a sequence  $(u_{\hbar})_{\hbar \rightarrow 0^+}$  in  $L^2(M)$  that satisfies

$$\lim_{\hbar \rightarrow 0} \|(\text{Id}_{L^2(M)} - \mathbb{1}_{[E_1, E_2]}(-\hbar^2 \Delta))u_{\hbar}\|_{L^2(M)} = 0.$$

Moreover, we suppose that  $\|u_{\hbar}\|_{L^2(M)} = 1$ . The proof follows essentially the same lines as the one in [Anantharaman and Nonnenmacher 2007], and we refer the reader to that paper for a detailed account.

**6a. Quantum partitions.** As usual when computing the Kolmogorov–Sinai entropy, we start by decomposing the manifold  $M$  into finitely many pieces (of small diameter). Let  $(P_k)_{k=1, \dots, K}$  be a family of smooth real functions on  $M$  such that

$$\sum_{k=1}^K P_k^2(x) = 1 \quad \text{for all } x \in M. \tag{19}$$

Later on we will assume that the diameters of the supports of the  $P_k$  are small enough. We shall denote by  $\hat{P}_k$  the operator of multiplication by  $P_k(x)$  on the Hilbert space  $L^2(M)$ . We denote the Schrödinger flow by  $U^t = \exp(it\Delta/2)$ . With no loss of generality, we will assume that the injectivity radius of  $M$  is greater than 2, and work with this propagator at time  $\hbar$ , that is,  $U^{\hbar}$ . This unitary operator is a Fourier integral operator associated with the geodesic flow  $g^1$  taken at time  $\tau = 1$ . As one does to compute the Kolmogorov–Sinai entropy of an invariant measure, we define a new quantum partition of unity by evolving and refining the initial partition under the quantum evolution. For each time  $n \in \mathbb{N}$  and any sequence of symbols  $\alpha = (\alpha_0, \dots, \alpha_{n-1})$ , where  $\alpha_i \in [1, K]$  (we say that the sequence  $\alpha$  is of length  $|\alpha| = n$ ), we define the operators

$$\pi_{\alpha} = \hat{P}_{\alpha_{n-1}}(n-1)\hat{P}_{\alpha_{n-2}}(n-2) \cdots \hat{P}_{\alpha_0}. \tag{20}$$

Throughout the paper we use the notation  $\hat{A}(\tau) = U^{-\tau\hbar}\hat{A}U^{\tau\hbar}$  for the quantum evolution of an operator  $\hat{A}$ . From (19) and the unitarity of  $U^t$ , the family of operators  $\{\pi_\alpha : |\alpha| = n\}$  obviously satisfies the resolution of identity  $\sum_{|\alpha|=n} \pi_\alpha \pi_\alpha^* = \text{Id}_{L^2(M)}$ . We also have  $\sum_{|\alpha|=n} \pi_\alpha^* \pi_\alpha = \text{Id}_{L^2(M)}$ .

**6b. Quantum entropy, and entropic uncertainty principle.** For each time  $n$  and each normalized  $\phi$  in  $L^2(M)$ , we define two quantities that are noncommutative analogues of the entropy (9):

$$h_n^-(\phi) = - \sum_{|\alpha|=n} \|\pi_\alpha^* \phi\|^2 \log(\|\pi_\alpha^* \phi\|^2), \tag{21}$$

$$h_n^+(\phi) = - \sum_{|\alpha|=n} \|\pi_\alpha \phi\|^2 \log(\|\pi_\alpha \phi\|^2). \tag{22}$$

In all that follows, the integer  $n$  is of order  $\tilde{c}|\log \hbar|$  (with  $\tilde{c} > 0$  to be chosen later), and thus the number of terms in the sum  $\sum_{|\alpha|=n}$  is of order  $\hbar^{-K_0}$  for some  $K_0 > 0$ . The following is proved in [Anantharaman and Nonnenmacher 2007], using the entropic uncertainty principle of [Maassen and Uffink 1988].

**Proposition 6.1.** *Let  $\chi$  be real-valued, smooth, compactly supported function on  $\mathbb{R}$ . Define*

$$c(\chi, n) := \max_{|\alpha|=|\alpha'|=n} (\|\pi_{\alpha'}(n)\pi_\alpha \chi(-\hbar^2 \Delta)\|). \tag{23}$$

*Then for any  $\hbar > 0$  and  $L > 0$ , and for any normalized state  $\phi$  satisfying*

$$\sup_{|\alpha|=n} \|(I - \chi(-\hbar^2 \Delta))\pi_\alpha^* \phi\| \leq \hbar^L, \tag{24}$$

*we have*

$$h_n^+(U^{n\hbar}\phi) + h_n^-(\phi) \geq -2 \log(c(\chi, n) + \hbar^{L-K_0}).$$

Finally everything boils down to the main estimate:

**Theorem 6.2** [Anantharaman 2008; 2011; Anantharaman and Nonnenmacher 2007]. *If the diameters of the supports of the functions  $P_k$  are small enough (compared with the injectivity radius), the following holds.*

*For  $E > 0$  and  $0 < \varepsilon < E$ , choose  $\chi$  smooth, compactly supported in  $[E - \varepsilon, E + \varepsilon]$ , and such that  $\|\chi\|_\infty \leq 1$ . For any  $\tilde{c} > 0$ , there exists  $\hbar_{\tilde{c}} > 0$  such that, for all  $\hbar < \hbar_{\tilde{c}}$ , for  $n \leq \tilde{c}|\log \hbar|$ , and for any pair  $\alpha, \alpha'$  of sequences of length  $n$ ,*

$$\|\pi_{\alpha'}(n)\pi_\alpha \chi(-\hbar^2 \Delta)\| \leq C\hbar^{-(d-1)/2} e^{-n(d-1)\sqrt{E-\varepsilon}}. \tag{25}$$

*(The constant  $C$  is an absolute constant).*

**Remark 8.** This result is an improvement of the estimate of [Anantharaman 2008] (where the prefactor was only  $\hbar^{-d/2}$ ) and [Anantharaman and Nonnenmacher 2007] (where the support of  $\chi$  was assumed to shrink with  $\hbar$ ). Proving Theorem 2.4 using the weaker results of these papers turned out to be more painful than reproving Theorem 6.2 directly. This proof is provided in [Anantharaman 2011, Section 5]. Unfortunately, the arguments of there are specific to constant curvature, although we believe the result should also hold in variable negative curvature (parts of the proof rely on the fact that the stable and

unstable foliations of the geodesic flow are smooth). Thus, if we wanted to extend Theorem 2.4 so as to get the results claimed in Remark 2, we would have to use the hyperbolic dispersive estimate in the form used in [Anantharaman and Nonnenmacher 2007], which would need a rather different, and more technical, presentation.

In what follows, the integer  $n$  will always be taken equal to  $\lfloor \tilde{c}|\log \hbar| \rfloor$ , where  $\tilde{c}$  will be fixed in the next section. We assume that  $L$  is large enough so that  $\hbar^{L-K_0}$  is negligible in comparison with  $\hbar^{-(d-1)/2} e^{-n(d-1)\sqrt{E-\varepsilon}}$ . As a corollary of Theorem 6.2 and Proposition 6.1, we have this:

**Corollary 6.3.** *Let  $(\phi_{\hbar})_{\hbar \rightarrow 0}$  be a sequence of normalized states satisfying the assumptions of 6.1, with  $L$  large enough that  $\hbar^{L-K_0}$  is negligible in comparison with  $\hbar^{-(d-1)/2} e^{-n(d-1)\sqrt{E-\varepsilon}}$  for  $n = \lfloor \tilde{c}|\log \hbar| \rfloor$ . Then, in the semiclassical limit, the entropies of  $\phi_{\hbar}$  at time  $n = \lfloor \tilde{c}|\log \hbar| \rfloor$  satisfy*

$$\frac{h_n^+(U^{n\hbar}\phi_{\hbar}) + h_n^-(\phi_{\hbar})}{2n} \geq (d-1)\sqrt{E-\varepsilon} - \frac{(d-1)}{2\tilde{c}} + \mathcal{O}(n^{-1}). \tag{26}$$

**6c. Subadditivity until the Ehrenfest time.** In this section, we fix a sequence of normalized states  $(\phi_{\hbar})_{\hbar \rightarrow 0}$  satisfying (24) ( $\chi$  is always assumed to be supported in  $[E - \varepsilon, E + \varepsilon]$ ). We fix some arbitrary  $\delta > 0$ , and introduce the Ehrenfest time,

$$n_{\text{Ehr}}(\hbar, E, \varepsilon) := \left\lfloor \frac{(1-\delta)|\log \hbar|}{\sqrt{E+\varepsilon}} \right\rfloor. \tag{27}$$

**Remark 9.** The Ehrenfest time is the largest time on which the (noncommutative) dynamical system formed by the flow  $(U^{\tau\hbar})$  acting on pseudodifferential operators (supported in  $\{\|\xi\|^2 \in [E - \varepsilon, E + \varepsilon]\}$ ) is commutative, up to small errors going to 0 with  $\hbar$ .

We take  $n = n_{\text{Ehr}}(\hbar, E, \varepsilon)$  (in other words, we take  $\tilde{c} = (1-\delta)/\sqrt{E+\varepsilon}$ ), and we use a subadditivity property of the entropies  $h_n^+$  and  $h_n^-$  to go from (26) for  $n = n_{\text{Ehr}}(\hbar, E, \varepsilon)$  to a fixed, arbitrary, integer  $n_0$ . The proof of the next proposition is given in [Anantharaman and Nonnenmacher 2007] in the case when  $\phi_{\hbar}$  is an eigenfunction of  $\Delta$ . It can easily be adapted to the case of an arbitrary  $\phi_{\hbar}$  and yields this:

**Proposition 6.4** (subadditivity). *Let  $E \geq 0$  and  $\varepsilon > 0$ . For  $\delta > 0$  arbitrary, define the Ehrenfest time  $n_{\text{Ehr}}(\hbar, E, \varepsilon)$  as in (27). Let  $(\phi_{\hbar})_{\hbar \rightarrow 0}$  be a normalized family satisfying (24), where  $\chi$  is supported in  $[E - \varepsilon, E + \varepsilon]$ , and  $L$  is chosen large enough.*

*For any  $n_0 \in \mathbb{N}$ , there exists a positive  $R_{n_0}(\hbar)$ , with  $R_{n_0}(\hbar) \rightarrow 0$  as  $\hbar \rightarrow 0$ , such that for any  $\hbar \in (0, 1]$  and any  $n_0, m \in \mathbb{N}$  with  $n_0 + m \leq n_{\text{Ehr}}(\hbar)$ , we have*

$$\begin{aligned} h_{n_0+m}^+(\phi_{\hbar}) &\leq h_m^+(\phi_{\hbar}) + h_{n_0}^+(U^{m\hbar}\phi_{\hbar}) + R_{n_0}(\hbar), \\ h_{n_0+m}^-(\phi_{\hbar}) &\leq h_{n_0}^-(\phi_{\hbar}) + h_m^-(U^{n_0\hbar}\phi_{\hbar}) + R_{n_0}(\hbar). \end{aligned}$$

Let  $n_0 \in \mathbb{N}$  be fixed and  $n = n_{\text{Ehr}}(\hbar, E, \varepsilon)$ . Using the Euclidean division  $n = qn_0 + r$ , with  $r < n_0$ , Proposition 6.4 implies that for  $\hbar$  small enough,

$$\frac{h_n^+(\phi_{\hbar})}{n} \leq \frac{\sum_{k=0}^{q-1} h_{n_0}^+(U^{kn_0\hbar}\phi_{\hbar})}{qn_0} + \frac{h_r^+(U^{qn_0\hbar}\phi_{\hbar})}{n} + \frac{R_{n_0}(\hbar)}{n_0}$$

and

$$\frac{h_n^-(\phi_{\hbar})}{n} \leq \frac{\sum_{k=0}^{q-1} h_{n_0}^-(U^{(r+kn_0)\hbar} \phi_{\hbar})}{qn_0} + \frac{h_r^-(U^{r\hbar} \phi_{\hbar})}{n} + \frac{R_{n_0}(\hbar)}{n_0}.$$

Note that  $h_r^+(U^{qn_0\hbar} \phi_{\hbar}) + h_r^-(U^{r\hbar} \phi_{\hbar})$  stays uniformly bounded (by  $\log n_0$ ) when  $\hbar \rightarrow 0$ . Combining the subadditivity property with Corollary 6.3, we find that

$$\frac{\sum_{k=0}^{q-1} (h_{n_0}^+(U^{kn_0\hbar} U^{n\hbar} \phi_{\hbar}) + h_{n_0}^-(U^{(r+kn_0)\hbar} \phi_{\hbar}))}{2qn_0} \geq (d-1)\sqrt{E-\varepsilon} - \frac{(d-1)\sqrt{E+\varepsilon}}{2(1-\delta)} - \frac{R_{n_0}(\hbar)}{n_0} + \mathbb{O}_{n_0}(1/n) \quad (28)$$

for  $n = n_{\text{Ehr}}(\hbar, E, \varepsilon)$ .

**6d. The conclusion.** The interval  $[E_1, E_2]$  is fixed. Consider  $E$  in  $[E_1, E_2]$  and a sequence of normalized states  $(u_{\hbar})_{\hbar \rightarrow 0}$  that satisfies (7). We may assume without loss of generality that  $\mathbb{1}_{[E_1, E_2]}(-\hbar^2 \Delta)u_{\hbar} = u_{\hbar}$  (since the semiclassical limits associated with  $u_{\hbar}$  and  $\mathbb{1}_{[E_1, E_2]}(-\hbar^2 \Delta)u_{\hbar}$  will be the same). We fix a function  $\chi \in \mathcal{C}_0^\infty(\mathbb{R})$ , supported in  $[-1, 1]$  such that  $\sum_{k \in \mathbb{Z}} \chi^2(x-k) \equiv 1$ . For  $N \in \mathbb{N}$ , we write

$$\varepsilon = \frac{E_2 - E_1}{N} \quad \text{and} \quad \chi_j(x) = \chi\left(\frac{x - E_1 - j\varepsilon}{\varepsilon}\right) \quad \text{for } j = 0, \dots, N.$$

We have  $u_{\hbar} = \sum_{j=0}^N \chi_j^2(-\hbar^2 \Delta)u_{\hbar}$  and thus  $\|u_{\hbar}\|^2 = \sum_{j=0}^N \|\chi_j(-\hbar^2 \Delta)u_{\hbar}\|^2$ . We will write  $u_j = \chi_j(-\hbar^2 \Delta)u_{\hbar}$  and  $\tilde{u}_j = u_j/\|u_j\|$ . For  $t \in \mathbb{R}$ , we apply (28) to  $\phi_{\hbar} = U^t \tilde{u}_j$  and obtain

$$\frac{\sum_{k=0}^{q-1} (h_{n_0}^+(U^{kn_0\hbar} U^{n\hbar} U^t \tilde{u}_j) + h_{n_0}^-(U^{(r+kn_0)\hbar} U^t \tilde{u}_j))}{2qn_0} \geq (d-1)\sqrt{E_1 + (j-1)\varepsilon} - \frac{(d-1)}{2(1-\delta)}\sqrt{E_1 + (j+1)\varepsilon} - \frac{R_{n_0}(\hbar)}{n_0} + \mathbb{O}_{n_0}(1/|\log \hbar|). \quad (29)$$

If we multiply by  $\theta(t)$  (satisfying  $\theta \in L^1(\mathbb{R}, \mathbb{R}_+)$  and  $\int \theta = 1$ ), integrate with respect to  $t$ , and take into account the fact that  $(kn_0 + r)\hbar \rightarrow 0$  and  $n\hbar \rightarrow 0$ , we find that

$$\int \theta(t) \frac{h_{n_0}^+(U^t \tilde{u}_j) + h_{n_0}^-(U^t \tilde{u}_j)}{2n_0} dt \geq (d-1)\sqrt{E_1 + (j-1)\varepsilon} - \frac{(d-1)}{2(1-\delta)}\sqrt{E_1 + (j+1)\varepsilon} + o_{n_0}(1). \quad (30)$$

This yields that

$$\sum_{j=0}^N \|u_j\|^2 \int \theta(t) \frac{h_{n_0}^+(U^t \tilde{u}_j) + h_{n_0}^-(U^t \tilde{u}_j)}{2n_0} dt \geq \sum_{j=0}^N \|u_j\|^2 \left[ (d-1)\sqrt{E_1 + (j-1)\varepsilon} - \frac{(d-1)}{2(1-\delta)}\sqrt{E_1 + (j+1)\varepsilon} \right] + o_{n_0}(1). \quad (31)$$

We define the averaged entropy

$$h_n^-(\phi, \theta) = - \sum_{|\alpha|=n} \left( \int \theta(t) \|\pi_\alpha^* U^t \phi\|^2 dt \right) \log \left( \int \theta(t) \|\pi_\alpha^* U^t \phi\|^2 dt \right), \tag{32}$$

$$h_n^+(\phi, \theta) = - \sum_{|\alpha|=n} \left( \int \theta(t) \|\pi_\alpha U^t \phi\|^2 dt \right) \log \left( \int \theta(t) \|\pi_\alpha U^t \phi\|^2 dt \right). \tag{33}$$

Using the concavity of  $x \mapsto -x \log x$ , (31) implies

$$\begin{aligned} \sum_{j=0}^N \|u_j\|^2 \frac{h_{n_0}^+(\tilde{u}_j, \theta) + h_{n_0}^-(\tilde{u}_j, \theta)}{2n_0} \\ \geq \sum_{j=0}^N \|u_j\|^2 \left[ (d-1)\sqrt{E_1 + (j-1)\varepsilon} - \frac{(d-1)}{2(1-\delta)}\sqrt{E_1 + (j+1)\varepsilon} \right] + o_{n_0}(1). \end{aligned} \tag{34}$$

We can now take the limit  $\hbar \rightarrow 0$ . If the semiclassical measure associated with the family  $(U^t u_\hbar)$  decomposes as  $\mu_t = \int \mu_{t,E} d\nu(E)$ , then  $\|u_j\|^2$  converges to  $\int \chi_j^2(E) d\nu(E)$ . On the left side of (34),  $h_{n_0}^+(\tilde{u}_j, \theta)$  and  $h_{n_0}^-(\tilde{u}_j, \theta)$  both converge to

$$\sum_{|\alpha|=n_0} \eta \left( \frac{1}{\int \chi_j^2(E) d\nu(E)} \int \theta(t) \chi_j^2(E) \mu_{t,E} ((P_{\alpha_{n-1}}^2 \circ g^{n-1}) \cdots (P_{\alpha_1}^2 \circ g^1) P_{\alpha_0}^2) d\nu(E) dt \right),$$

where  $\eta(x) = -x \log x$ .

Then, we let  $n_0 \rightarrow +\infty$ , which allows to go from the previous quantity to the Kolmogorov–Sinai entropy  $h_{KS}$ ; for this step, details can be found in [Anantharaman and Nonnenmacher 2007, Section 2.2.8]. This gives us the inequality

$$\begin{aligned} \sum_{j=0}^N \int \chi_j^2(E) d\nu(E) h_{KS} \left( \frac{1}{\int \chi_j^2(E) d\nu(E)} \int \theta(t) \chi_j^2(E) \mu_{t,E} d\nu(E) dt \right) \\ \geq \sum_{j=0}^N \left[ (d-1)\sqrt{E_1 + (j-1)\varepsilon} - \frac{(d-1)}{2(1-\delta)}\sqrt{E_1 + (j+1)\varepsilon} \right] \int \chi_j^2(E) d\nu(E). \end{aligned} \tag{35}$$

At this stage, we use the fact that  $h_{KS}$  is affine and derive that

$$\int \theta(t) \left( h_{KS}(\mu_{t,E}) - \sum_{j=0}^N \chi_j^2(E) \left[ (d-1)\sqrt{E_1 + (j-1)\varepsilon} - \frac{(d-1)}{2(1-\delta)}\sqrt{E_1 + (j+1)\varepsilon} \right] \right) d\nu(E) dt \geq 0.$$

Finally, we can take the limit  $N \rightarrow +\infty$ , to obtain

$$\int \theta(t) \left( h_{KS}(\mu_{t,E}) - \frac{d-1}{2} \sqrt{E} \right) d\nu(E) dt \geq 0.$$



If we use the same argument, replacing  $u_{\hbar}$  by  $f(-\hbar^2 \Delta)u_{\hbar}$  (where  $f$  is a smooth function on  $[E_1, E_2]$  such that  $\int f^2(E) dv(E) = 1$ ), we obtain by the same argument

$$\int \theta(t) f^2(E) \left( h_{\text{KS}}(\mu_{t,E}) - \frac{d-1}{2} \sqrt{E} \right) dv(E) dt \geq 0;$$

this holds for all  $\theta$  in  $L^1(\mathbb{R}, \mathbb{R}_+)$  such that  $\int \theta = 1$  and  $f$  in  $\mathcal{C}_0^\infty(\mathbb{R}_+, \mathbb{R})$  such that  $\int f^2(E) dv(E) = 1$ . As a consequence, one has for Lebesgue  $\otimes \nu$ -almost every  $(t, E)$ ,

$$h_{\text{KS}}(\mu_{t,E}) \geq \frac{d-1}{2} \sqrt{E}. \quad \square$$

**Remark 10.** If one wants to consider the microlocal setting (see Remark 4) where one uses  $\text{Op}_1$  instead of  $\text{Op}_{\hbar}$ , one introduces a partition of unity based on the Paley–Littlewood decomposition. For a fixed  $\epsilon > 0$ , arbitrarily small, one introduces a smooth function  $\psi_\epsilon$  on  $\mathbb{R}_+$  satisfying  $\psi_\epsilon(E) = 1$  for  $0 \leq E \leq 2^{-\epsilon}$  and  $\psi_\epsilon(E) = 0$  for  $E \geq 1$ . Then, one can define  $\varphi_\epsilon(E) = \psi_\epsilon(E/2^\epsilon) - \psi_\epsilon(E)$  and verify that

$$1 = \psi_\epsilon(E) + \sum_{j \geq 0} \varphi_\epsilon(2^{-j\epsilon} E).$$

We stress that for every  $j \geq 0$ , the cutoff function  $\varphi_\epsilon(2^{-j\epsilon} E)$  is compactly supported in  $[2^{\epsilon(j-1)}, 2^{\epsilon(j+1)}]$ . On the energy window  $E \in [2^{\epsilon(j-1)}, 2^{\epsilon(j+1)}]$ , one can adapt the proof above, doing the change of variable  $\xi \rightsquigarrow 2^{-\epsilon j} \xi$ , and using the relation  $\text{Op}_1(a(x, 2^{-\epsilon j} \xi)) = \text{Op}_{2^{-\epsilon j}}(a(x, \xi))$ . One then copies the steps of Section 6, using  $\hbar_j = 2^{-\epsilon j}$  as the effective Planck constant, and taking  $\chi_j(E) = \varphi_\epsilon^{1/2}(2^{-j\epsilon} E)$  in Section 6d.

### 7. From entropy estimates to observability

In this section, we explain how we can go from the entropy estimates of Theorem 2.4 to the observability estimate of Theorem 2.5. According to Lebeau [1992], it suffices to prove the following weak observability result to deduce Theorem 2.5:

**Theorem 7.1.** *Under the assumptions of Theorem 2.5, for all  $T > 0$ , there exists  $C_{T,a} > 0$  such that*

$$\|u\|_{L^2(M)}^2 \leq C_{T,a} \left( \int_0^T \|ae^{it\Delta/2}u\|_{L^2(M)}^2 dt + \|u\|_{H^{-1}(M)}^2 \right) \quad \text{for all } u \text{ in } L^2(M). \quad (36)$$

For the sake of completeness, we briefly recall the argument of Lebeau to deduce observability from a weak observability estimate at time  $T$ . First, for  $T' > T$ , we introduce the subspace

$$N(T') := \{\varphi \in L^2(M) : a(x)(e^{it\Delta}\varphi)(x) = 0 \text{ for all } 0 \leq t \leq T'\}.$$

From weak observability and the compactness of the injection  $L^2 \subset H^{-1}$ , we can deduce that for  $T' > T$ , this subspace is finite-dimensional. One can also verify that  $\Delta\varphi$  belongs to  $N(T'')$  for every  $T < T'' < T'$  and every  $\varphi$  in  $N(T')$  (by taking the limit of the sequence  $(e^{t\epsilon\Delta}\varphi - \varphi)/\epsilon$ , which belongs to  $N(T'')$  for  $\epsilon$  small enough, and is bounded in  $H^{-2}(M)$ ).

This implies that  $\Delta$  is an operator from the finite-dimensional subspace  $N(T')$  into itself. As  $a$  is nontrivial, one can deduce the existence of an eigenfunction of the Laplacian that is equal to 0 on a

nonempty open set. By the Aronszajn–Cordes theorem [Hörmander 1985, Section 17.2], this eigenfunction is necessarily 0 and the subspace  $N(T')$  is reduced to  $\{0\}$ . By contradiction, we can finally deduce that observability holds for  $T' > T$ .

To prove Theorem 7.1, we proceed by contradiction and make the assumption that there exists a sequence of normalized vectors  $(u_n)_{n \in \mathbb{N}}$  in  $L^2(M)$  and  $T > 0$  such that

$$\lim_{n \rightarrow +\infty} \left( \int_0^T \|ae^{it\Delta/2}u_n\|_{L^2(M)}^2 dt + \|u_n\|_{H^{-1}(M)}^2 \right) = 0. \tag{37}$$

This implies that  $u_n$  converges to 0, weakly in  $L^2$ . For every  $t$  in  $\mathbb{R}$ , we introduce the “distribution”

$$\mu_n(t)(b) := \langle u_n | e^{-it\Delta/2} \text{Op}_1(b)e^{it\Delta/2}u_n \rangle_{L^2(M)},$$

defined for all  $b \in \mathcal{S}_0$ . The map  $t \mapsto \mu_n(t)$  belongs to  $L^\infty(\mathbb{R}, \mathcal{S}'_0)$ . Thus, there exists a subsequence  $(u_{n_k})_k$  and  $\mu$  in  $L^\infty(\mathbb{R}, \mathcal{S}'_0)$  such that

$$\int_{\mathbb{R} \times \widehat{T^*M}} \theta(t)b(x, \xi)\mu_{n_k}(t)(dx, d\xi) dt \xrightarrow{k \rightarrow +\infty} \int_{\mathbb{R} \times \widehat{T^*M}} \theta(t)b(x, \xi)\mu(t)(dx, d\xi) dt$$

for all  $\theta \in L^1(\mathbb{R})$  and  $b \in \mathcal{S}_0$ . As  $u_n$  converges weakly to 0, each  $\mu(t)$  is actually supported at infinity, and may thus be identified with a probability measure on the unit sphere bundle  $S^*M$ , invariant under the geodesic flow (see Remark 4).

From Theorem 2.4 and Remark 4, we know that  $h_{\text{KS}}(\mu(t)) \geq \frac{1}{2}(d - 1)$  for almost every  $t$  in  $\mathbb{R}$ . We will now use the fact that the topological entropy of  $K_a$  is less than  $\frac{1}{2}(d - 1)$ , that is,

$$h_{\text{top}}(K_a, (g^\tau)) := \sup_{\mu \in \mathcal{M}(S^*M, g^\tau)} \{h_{\text{KS}}(\mu) : \mu(K_a) = 1\} < \frac{1}{2}(d - 1).$$

Using property (37), we know that  $\int_{S^*M \times [0, T]} a^2(x, \xi)\mu(t)(dx, d\xi) dt = 0$ . In particular, this implies that  $\mu(t)(S^*M \setminus K_a) = 0$  for almost every  $t$  in  $[0, T]$  (as  $\mu(t)$  is  $g^\tau$ -invariant), leading to a contradiction.  $\square$

### Appendix A. Pseudodifferential calculus on a manifold

In this section, we recall some facts of pseudodifferential calculus; details can be found in [Zworski 2012]. We define on  $\mathbb{R}^{2d}$  the following class of (semiclassical) symbols:

$$S^{m,k}(\mathbb{R}^{2d}) := \{a = a_h \in C^\infty(\mathbb{R}^{2d}) : |\partial_x^\alpha \partial_\xi^\beta a| \leq C_{\alpha,\beta} h^{-k} \langle \xi \rangle^{m-|\beta|}$$

for all  $K \subset \mathbb{R}^d$  compact,  $\alpha, \beta$ , some  $C_{\alpha,\beta}$ , and all  $(x, \xi) \in K \times \mathbb{R}^d\}$ .

Let  $M$  be a smooth compact Riemannian  $d$ -manifold without boundary. Consider a finite smooth atlas  $(f_l, V_l)$  of  $M$ , where each  $f_l$  is a smooth diffeomorphism from the open subset  $V_l \subset M$  to a bounded open set  $W_l \subset \mathbb{R}^d$ . To each  $f_l$  corresponds a pull-back  $f_l^* : C^\infty(W_l) \rightarrow C^\infty(V_l)$  and a canonical map  $\tilde{f}_l$  from  $T^*V_l$  to  $T^*W_l$ :

$$\tilde{f}_l : (x, \xi) \mapsto (f_l(x), (Df_l(x))^{-1})^T \xi).$$

Consider now a smooth locally finite partition of identity  $(\phi_l)$  adapted to the previous atlas  $(f_l, V_l)$ . That means  $\sum_l \phi_l = 1$  and  $\phi_l \in \mathcal{C}_o^\infty(V_l)$ . Then, any observable  $a$  in  $C^\infty(T^*M)$  can be decomposed as  $a = \sum_l a_l$ , where  $a_l = a\phi_l$ . Each  $a_l$  belongs to  $C^\infty(T^*V_l)$  and can be pushed to a function  $\tilde{a}_l = (f_l^{-1})^* a_l \in C^\infty(T^*W_l)$ . As in [Zworski 2012], define a class of symbols of order  $m$  and index  $k$  by

$$S^{m,k}(T^*M) := \{a = a_\hbar \in C^\infty(T^*M) : |\partial_x^\alpha \partial_\xi^\beta a| \leq C_{\alpha,\beta} \hbar^{-k} \langle \xi \rangle^{m-|\beta|} \text{ for all } \alpha, \beta \text{ and some } C_{\alpha,\beta}\}. \quad (38)$$

Then, for  $a \in S^{m,k}(T^*M)$  and for each  $l$ , one can associate to the symbol  $\tilde{a}_l \in S^{m,k}(\mathbb{R}^{2d})$  the standard Weyl quantization:

$$\text{Op}_\hbar^w(\tilde{a}_l)u(x) := \frac{1}{(2\pi\hbar)^d} \int_{\mathbb{R}^{2d}} e^{i(\hbar)^{-1}(x-y, \xi)} \tilde{a}_l\left(\frac{x+y}{2}, \xi; \hbar\right) u(y) dy d\xi,$$

where  $u \in \mathcal{C}_o^\infty(\mathbb{R}^d)$ . Consider now a smooth cutoff  $\psi_l \in \mathcal{C}_c^\infty(V_l)$  such that  $\psi_l = 1$  close to the support of  $\phi_l$ . A quantization of  $a \in S^{m,k}(T^*M)$  is then defined by

$$\text{Op}_\hbar(a)(u) := \sum_l \psi_l \times (f_l^* \text{Op}_\hbar^w(\tilde{a}_l)(f_l^{-1})^*)(\psi_l \times u), \quad (39)$$

where  $u \in \mathcal{C}^\infty(M)$ . According to the appendix of [Zworski 2012], the quantization procedure  $\text{Op}_\hbar$  sends  $S^{m,k}(T^*M)$  onto the space of pseudodifferential operators of order  $m$  and of index  $k$ , denoted  $\Psi^{m,k}(M)$ . It can be shown that the dependence in the cutoffs  $\phi_l$  and  $\psi_l$  only appears at order 2 in  $\hbar$  and the principal symbol map  $\sigma_0 : \Psi^{m,k}(M) \rightarrow S^{m,k}/S^{m,k-1}(T^*M)$  is then intrinsically defined.

At various places in this paper, a larger class of symbols should be considered, as in [Dimassi and Sjöstrand 1999] or [Zworski 2012]. For  $0 \leq \nu < 1/2$ ,

$$S_\nu^{m,k}(T^*M) = \{a = a_\hbar \in \mathcal{C}^\infty(T^*M) : |\partial_x^\alpha \partial_\xi^\beta a| \leq C_{\alpha,\beta} \hbar^{-k-\nu|\alpha+\beta|} \langle \xi \rangle^{m-|\beta|} \text{ for all } \alpha, \beta \text{ and some } C_{\alpha,\beta}\}.$$

Results of [Dimassi and Sjöstrand 1999] can be applied to this new class of symbols. For example, if  $M$  is compact, a symbol of  $S_\nu^{0,0}$  gives a bounded operator on  $L^2(M)$  (with norm independent of  $\hbar \leq 1$ ).

Even if the Weyl procedure is a natural choice to quantize an observable  $a$  on  $\mathbb{R}^{2d}$ , it is sometimes preferable to use a quantization that also satisfies the property that  $\text{Op}_\hbar(a) \geq 0$  if  $a \geq 0$  (such a quantization procedure is said to be *positive*). This can be achieved using to the anti-Wick procedure; see [Helffer et al. 1987]. For  $a$  in  $S_\nu^{0,0}(\mathbb{R}^{2d})$  that coincides with a function on  $\mathbb{R}^d$  outside a compact subset of  $T^*\mathbb{R}^d$ , one has

$$\|\text{Op}_\hbar^w(a) - \text{Op}_\hbar^{AW}(a)\|_{L^2} \leq C \sum_{|\alpha| \leq D} \hbar^{(|\alpha|+1)/2} \|\partial^\alpha a\|_\infty, \quad (40)$$

where  $C$  and  $D$  are some positive constants that depend only on the dimension  $d$ . To get a positive procedure of quantization on a manifold, one replaces in definition (39) the Weyl quantization by the anti-Wick one. We will denote by  $\text{Op}_\hbar^+(a)$  this new choice of quantization, which is well defined for every element in  $S_\nu^{0,0}(T^*M)$  of the form  $b(x) + c(x, \xi)$ , where  $b$  belongs to  $S_\nu^{0,0}(T^*M)$  and  $c$  belongs to  $\mathcal{C}_o^\infty(T^*M) \cap S_\nu^{0,0}(T^*M)$ . We underline the fact that  $\text{Op}_\hbar^+(1) = \text{Id}_{L^2(M)}$ .

## References

- [Anantharaman 2008] N. Anantharaman, “Entropy and the localization of eigenfunctions”, *Ann. of Math. (2)* **168**:2 (2008), 435–475. MR 2011g:35076 Zbl 1175.35036
- [Anantharaman 2011] N. Anantharaman, “Exponential decay for products of Fourier integral operators”, *Methods Appl. Anal.* **18**:2 (2011), 165–181. MR 2847483
- [Anantharaman and Macià 2011] N. Anantharaman and F. Macià, “Semiclassical measures for the Schrödinger equation on the torus”, preprint, 2011. arXiv 1005.0296
- [Anantharaman and Nonnenmacher 2007] N. Anantharaman and S. Nonnenmacher, “Half-delocalization of eigenfunctions for the Laplacian on an Anosov manifold”, *Ann. Inst. Fourier (Grenoble)* **57**:7 (2007), 2465–2523. MR 2009m:81076 Zbl 1145.81033
- [Anantharaman et al. 2009] N. Anantharaman, H. Koch, and S. Nonnenmacher, “Entropy of eigenfunctions”, pp. 1–22 in *New trends in mathematical physics* (Rio de Janeiro, 2006), edited by V. Sidoravičius, Springer, Dordrecht, 2009. Zbl 1175.81118 arXiv 0704.1564
- [Barreira and Wolf 2007] L. Barreira and C. Wolf, “Dimension and ergodic decompositions for hyperbolic flows”, *Discrete Contin. Dyn. Syst.* **17**:1 (2007), 201–212. MR 2007m:37050 Zbl 1144.37012
- [Bouzouina and Robert 2002] A. Bouzouina and D. Robert, “Uniform semiclassical estimates for the propagation of quantum observables”, *Duke Math. J.* **111**:2 (2002), 223–252. MR 2003b:81049 Zbl 1069.35061
- [Bowen and Ruelle 1975] R. Bowen and D. Ruelle, “The ergodic theory of Axiom A flows”, *Invent. Math.* **29**:3 (1975), 181–202. MR 52 #1786 Zbl 0311.58010
- [Colin de Verdière 1985] Y. Colin de Verdière, “Ergodicité et fonctions propres du Laplacien”, *Comm. Math. Phys.* **102**:3 (1985), 497–502. MR 87d:58145 Zbl 0592.58050
- [Dimassi and Sjöstrand 1999] M. Dimassi and J. Sjöstrand, *Spectral asymptotics in the semi-classical limit*, London Mathematical Society Lecture Note Series **268**, Cambridge University Press, 1999. MR 2001b:35237 Zbl 0926.35002
- [Duistermaat and Guillemin 1975] J. J. Duistermaat and V. W. Guillemin, “The spectrum of positive elliptic operators and periodic bicharacteristics”, *Invent. Math.* **29**:1 (1975), 39–79. MR 53 #9307 Zbl 0307.35071
- [Eckhardt et al. 1995] B. Eckhardt, S. Fishman, J. P. Keating, O. Agam, J. Main, and K. Müller, “Approach to ergodicity in quantum wave functions”, *Phys. Rev. E* **52**:6 (1995), 5893–5903.
- [Feingold and Peres 1986] M. Feingold and A. Peres, “Distribution of matrix elements of chaotic systems”, *Phys. Rev. A* (3) **34**:1 (1986), 591–595. MR 87i:81055
- [Gérard 1991] P. Gérard, “Microlocal defect measures”, *Comm. Partial Differential Equations* **16**:11 (1991), 1761–1794. MR 92k:35027 Zbl 0770.35001
- [Helffer et al. 1987] B. Helffer, A. Martinez, and D. Robert, “Ergodicité et limite semi-classique”, *Comm. Math. Phys.* **109**:2 (1987), 313–326. MR 88e:81029 Zbl 0624.58039
- [Hörmander 1985] L. Hörmander, *The analysis of linear partial differential operators, III: Pseudo-differential operators*, Grundlehren der Mathematischen Wissenschaften **274**, Springer, Berlin, 1985. MR 87d:35002a Zbl 0601.35001
- [Katok and Hasselblatt 1995] A. Katok and B. Hasselblatt, *Introduction to the modern theory of dynamical systems*, Encyclopedia of Mathematics and its Applications **54**, Cambridge University Press, 1995. MR 96c:58055 Zbl 0878.58020
- [Kifer 1992] Y. Kifer, “Averaging in dynamical systems and large deviations”, *Invent. Math.* **110**:2 (1992), 337–370. MR 93m:60118 Zbl 0791.58072
- [Lebeau 1992] G. Lebeau, “Contrôle de l’équation de Schrödinger”, *J. Math. Pures Appl. (9)* **71**:3 (1992), 267–291. MR 93i:35018 Zbl 0838.35013
- [Ledrappier and Young 1985] F. Ledrappier and L.-S. Young, “The metric entropy of diffeomorphisms, I: Characterization of measures satisfying Pesin’s entropy formula”, *Ann. of Math. (2)* **122**:3 (1985), 509–539. MR 87i:58101a Zbl 0605.58028
- [Maassen and Uffink 1988] H. Maassen and J. B. M. Uffink, “Generalized entropic uncertainty relations”, *Phys. Rev. Lett.* **60**:12 (1988), 1103–1106. MR 89f:81027

- [Macià 2009] F. Macià, “Semiclassical measures and the Schrödinger flow on Riemannian manifolds”, *Nonlinearity* **22**:5 (2009), 1003–1020. MR 2010i:58032 Zbl 1166.81020
- [Macià 2010] F. Macià, “High-frequency propagation for the Schrödinger equation on the torus”, *J. Funct. Anal.* **258**:3 (2010), 933–955. MR 2011b:81143 Zbl 1180.35438
- [Parry and Pollicott 1990] W. Parry and M. Pollicott, *Zeta functions and the periodic orbit structure of hyperbolic dynamics*, Astérisque **187-188**, Société Mathématique de France, Montrouge, 1990. MR 92f:58141 Zbl 0726.58003
- [Pesin 1977] Y. B. Pesin, “Characteristic Lyapunov exponents, and smooth ergodic theory”, *Uspehi Mat. Nauk* **32**:4 (196) (1977), 55–112. In Russian; translated in *Russ. Math. Surv.* **32**:4 (1977), 55–114. MR 57 #6667 Zbl 0359.58010
- [Pesin 1997] Y. B. Pesin, *Dimension theory in dynamical systems: contemporary views and applications*, University of Chicago Press, Chicago, IL, 1997. MR 99b:58003 Zbl 0895.58033
- [Pesin and Sadovskaya 2001] Y. B. Pesin and V. Sadovskaya, “Multifractal analysis of conformal Axiom A flows”, *Comm. Math. Phys.* **216**:2 (2001), 277–312. MR 2002g:37035 Zbl 0992.37023
- [Rees 1981] M. Rees, “An alternative approach to the ergodic theory of measured foliations on surfaces”, *Ergodic Theory Dynamical Systems* **1**:4 (1981), 461–488. MR 84c:58065 Zbl 0539.58018
- [Rivière 2009] G. Rivière, *Délocalisation des mesures semi-classiques pour des systèmes dynamiques chaotiques*, thesis, Centre de Mathématiques Laurent Schwartz, Palaiseau, 2009, available at <http://tel.archives-ouvertes.fr/tel-00437912>.
- [Rivière 2010] G. Rivière, “Entropy of semiclassical measures in dimension 2”, *Duke Math. J.* **155**:2 (2010), 271–336. MR 2012a:58056 Zbl 1230.37048
- [Ruelle 1976] D. Ruelle, “Generalized zeta-functions for Axiom A basic sets”, *Bull. Amer. Math. Soc.* **82**:1 (1976), 153–156. MR 53 #4146 Zbl 0316.58016
- [Ruelle 1978] D. Ruelle, “An inequality for the entropy of differentiable maps”, *Bol. Soc. Brasil. Mat.* **9**:1 (1978), 83–87. MR 80f:58026 Zbl 0432.58013
- [Schubert 2006] R. Schubert, “Upper bounds on the rate of quantum ergodicity”, *Ann. Henri Poincaré* **7**:6 (2006), 1085–1098. MR 2008a:81075 Zbl 1099.81032
- [Shnirelman 1974] A. I. Shnirelman, “Ergodic properties of eigenfunctions”, *Uspehi Mat. Nauk* **29**:6 (180) (1974), 181–182. In Russian. MR 53 #6648 Zbl 0324.58020
- [Sogge and Zelditch 2002] C. D. Sogge and S. Zelditch, “Riemannian manifolds with maximal eigenfunction growth”, *Duke Math. J.* **114**:3 (2002), 387–437. MR 2004b:58053 Zbl 1018.58010
- [Walters 1982] P. Walters, *An introduction to ergodic theory*, Graduate Texts in Mathematics **79**, Springer, New York, 1982. MR 84e:28017 Zbl 0475.28009
- [Zelditch 1987] S. Zelditch, “Uniform distribution of eigenfunctions on compact hyperbolic surfaces”, *Duke Math. J.* **55**:4 (1987), 919–941. MR 89d:58129 Zbl 0643.58029
- [Zelditch 1994] S. Zelditch, “On the rate of quantum ergodicity, I: Upper bounds”, *Comm. Math. Phys.* **160**:1 (1994), 81–92. MR 95f:58084 Zbl 0788.58043
- [Zworski 2012] M. Zworski, *Semiclassical analysis*, American Mathematical Society, Providence, RI, 2012.

Received 28 Jul 2010. Revised 27 Aug 2011. Accepted 31 Aug 2011.

NALINI ANANTHARAMAN: [nalini.anantharaman@math.u-psud.fr](mailto:nalini.anantharaman@math.u-psud.fr)

Laboratoire de Mathématiques (UMR 8628), Université Paris-Sud XI, Bâtiment 425, 91405 Orsay Cedex, France

GABRIEL RIVIÈRE: [gabriel.riviere@polytechnique.edu](mailto:gabriel.riviere@polytechnique.edu)

Laboratoire Paul Painlevé (UMR 8524), Université Lille 1, Bâtiment M3, 59655 Villeneuve d’Ascq Cedex, France

# A BILINEAR OSCILLATORY INTEGRAL ESTIMATE AND BILINEAR REFINEMENTS TO STRICHARTZ ESTIMATES ON CLOSED MANIFOLDS

ZAHER HANI

We prove a bilinear  $L^2(\mathbb{R}^d) \times L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^{d+1})$  estimate for a pair of oscillatory integral operators with different asymptotic parameters and phase functions satisfying a transversality condition. This is then used to prove a bilinear refinement to Strichartz estimates on closed manifolds, similar to that derived by Bourgain on  $\mathbb{R}^d$ , but at a relevant semiclassical scale. These estimates will be employed elsewhere to prove global well-posedness below  $H^1$  for the cubic nonlinear Schrödinger equation on closed surfaces.

## 1. Introduction

We consider oscillatory integrals defined by

$$T_\lambda f(t, x) = \int_{\mathbb{R}^d} e^{i\lambda\phi(t, x, \xi)} a(t, x, \xi) f(\xi) d\xi, \quad (1-1)$$

where  $t \in \mathbb{R}$ ,  $x, \xi \in \mathbb{R}^d$ ,  $a \in C_0^\infty(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d)$ . The phase function  $\phi$  is a real-valued smooth function on the support of  $a$ . We shall assume that it satisfies a usual nondegeneracy condition, namely that the  $(d + 1) \times d$  matrix

$$\frac{\partial^2 \phi}{\partial \xi \partial(x, t)}(t_0, x_0, \xi_0) \text{ has maximal rank } d \text{ for every } (t_0, x_0, \xi_0) \in \text{supp } a. \quad (1-2)$$

This implies that for each fixed  $(t_0, x_0) \in \mathbb{R}^{d+1}$ , the map given by

$$\xi \mapsto \nabla_{(t, x)} \phi(t_0, x_0, \xi)$$

defines a smooth immersion from  $\mathbb{R}^d$  into  $\mathbb{R}^{d+1}$ . The image of this map is a hypersurface which we denote by  $S_\phi(t_0, x_0)$ , or just  $S_\phi$  when no confusion arises. Our objective is to prove bilinear estimates for such operators and use them to get bilinear refinements to Strichartz estimates on compact manifolds without boundary.

Operators as in (1-1) can be thought of as variable coefficient generalizations of usual dual restriction (extension) operators where  $\phi(t, x, \xi) = x \cdot \xi + t\psi(\xi)$  and (1-1) becomes the dual of the operator given by restricting the Fourier transform to the hypersurface  $S_\psi = \{(\tau, \xi) \in \mathbb{R}^{d+1} : \tau = \psi(\xi)\}$ . As in the case of restriction operators, one is interested in obtaining asymptotic decay estimates for  $\|T_\lambda\|_{L^p(\mathbb{R}^d) \rightarrow L^q(\mathbb{R}^{d+1})}$  in terms of  $\lambda$ . It is well known that in order to obtain nontrivial decay estimates (the optimal one being

*MSC2000:* primary 35B45, 42B20, 58J40; secondary 35A17, 35S30.

*Keywords:* bilinear oscillatory integrals, bilinear Strichartz estimates, transversality, semiclassical time scale, nonlinear Schrödinger equation on compact manifolds.

$\lambda^{-(d+1)/q}$ ), one has to impose some curvature condition on the hypersurfaces  $S_\phi$ , namely that the Gaussian curvature does not vanish anywhere. The pairs of exponents  $(p, q)$  for which this decay is possible were specified by Hörmander [1973] when  $d = 1$  and posed as a question for higher dimensions. Since then, there has been a tremendous amount of research in proving such bounds. (See [Stein 1993] and references therein for an introduction and [Tao 2004] for a more current survey).

We will be interested in bilinear versions of such estimates. In this case, one considers the product  $T_\lambda f \tilde{T}_\mu g$ , where  $\tilde{T}_\mu g$  is an operator similar to (1-1):

$$\tilde{T}_\mu g(t, x) = \int_{\mathbb{R}^d} e^{i\mu\psi(t, x, \xi)} b(t, x, \xi) g(\xi) d\xi, \quad (1-3)$$

where  $b \in C_0^\infty(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d)$  and  $\psi$  is smooth on the support of  $b$  and satisfies the same nondegeneracy assumption (1-2). The initial motivation behind such estimates was proving and refining the linear estimates in the case when the exponent  $q$  is an even number. However, such an improvement is only possible when the surfaces  $S_\phi$  and  $S_\psi$  satisfy a certain transversality assumption. This transversality turns out to be more important than any curvature assumption in certain instances. To be precise, the type of estimates one is often interested in are of the form

$$\|T_\lambda f \tilde{T}_\mu g\|_{L^q(\mathbb{R} \times \mathbb{R}^d)} \lesssim \Lambda(\lambda, \mu) \|f\|_{L^2(\mathbb{R}^d)} \|g\|_{L^2(\mathbb{R}^d)}. \quad (1-4)$$

(For us, the case when  $q = 2$  and  $\lambda \neq \mu$  will be of particular interest.) Great progress has been achieved in proving estimates like (1-4) especially in the case  $\lambda = \mu$  and when the surfaces  $S_\phi$  and  $S_\psi$  satisfy some nonvanishing curvature assumption. In the constant coefficient (restriction) case, Wolff was able to prove (1-4) in the cone restriction case for all  $q > 1 + 2/(d + 1)$  with  $\Lambda(\lambda, \lambda) \lesssim \lambda^{-(d+1)/q}$  [Wolff 2001]. This estimate was later extended to the endpoint in [Tao 2001]. The same estimate was then proven for transverse subsets of the paraboloid [Tao 2003]. In the variable coefficient case, Lee proved a similar estimate when  $\lambda = \mu$ ,  $q \geq 1 + 2/(d + 1)$ , and  $\Lambda(\lambda, \lambda) \lesssim \lambda^{-(d+1)/q+\epsilon}$  under certain curvature assumptions on the surfaces  $S_\phi(t_0, x_0)$  and  $S_\psi(t_0, x_0)$  [Lee 2006].

In this paper, we prove an  $L^2$  estimate when  $\lambda \neq \mu$  and the only assumption we impose on the hypersurfaces  $S_\phi$  and  $S_\psi$  is transversality. In particular, no curvature assumptions are taken.

**Theorem 1.1.** *Suppose that  $T_\lambda$  and  $\tilde{T}_\mu$  are two oscillatory integral operators of the form given in (1-1) with  $\mu \leq \lambda$  and assume that the canonical hypersurfaces associated with the phase functions  $\phi$  and  $\psi$  satisfy the standard transversality condition (1-6), then*

$$\|T_\lambda f \tilde{T}_\mu g\|_{L^2(\mathbb{R} \times \mathbb{R}^d)} \lesssim \frac{1}{\lambda^{d/2} \mu^{1/2}} \|f\|_{L^2(\mathbb{R}^d)} \|g\|_{L^2(\mathbb{R}^d)}. \quad (1-5)$$

*The implicit constants are allowed to depend on  $\delta$ ,  $d$ , and uniform bounds on a fixed number of derivatives of  $\phi$ ,  $\psi$ ,  $a$ , and  $b$ .*

A couple of remarks are in order. First, we mention that (1-5) is sharp (see the remark at the end of Section 2). Second, we note that without curvature assumptions on the surfaces, the linear estimate is easily seen to fail (consider the restriction to hyperplanes). However, the  $L^2$  bilinear estimate is true as

long as the surfaces are transverse.<sup>1</sup> Even when the linear estimate is true (which requires as mentioned a nonvanishing curvature assumption on the surfaces), (1-5) is an improvement on applying Hölder and the linear estimates available especially in the case when  $\mu \ll \lambda$  (for example, when  $d = 2$  linear estimates give the bound  $(\lambda\mu)^{-3/4}$ ). This improvement is often of great importance in applications (see [Bourgain 1999; 1998; Hani 2012]).

We now specify the transversality condition needed. The canonical hypersurfaces  $S_\phi(t_0, x_0)$  and  $S_\psi(t_0, x_0)$ , given by the maps  $\xi \mapsto \nabla_{(t,x)}\phi(t_0, x_0, \xi)$  and  $\xi \mapsto \nabla_{(t,x)}\psi(t_0, x_0, \xi)$  respectively, live in the cotangent space  $T_{(t_0,x_0)}^*\mathbb{R}^{d+1}$  to  $\mathbb{R}^{d+1}$  at  $(t_0, x_0)$ . The nondegeneracy condition defined in (1-2) for  $\phi$  (and defined similarly for  $\psi$ ), implies that for every  $\xi_0 \in \text{supp}_\xi a(t_0, x_0, \cdot)$ , there exists a locally defined unit normal vector field  $\nu_1(t_0, x_0, \xi_0) = \nu_1(\xi_0)$  to this surface at the point  $\nabla_{(t,x)}\phi(t_0, x_0, \xi_0) \in T_{(t_0,x_0)}^*\mathbb{R}^{d+1}$ . In other words, the map

$$\xi \mapsto \langle \nu_1(\xi_0), \nabla_{(t,x)}\phi(t_0, x_0, \xi) \rangle$$

has a critical point at  $\xi = \xi_0$  (in linear algebra terms,  $\nu(\xi_0)$  is the unit vector spanning the one dimensional orthogonal complement of the image of the matrix appearing in (1-2)). Similarly, we define the associated unit normal vector  $\nu_2(\xi_0)$  to  $S_\psi(t_0, x_0)$  at the point  $\nabla_{(t,x)}\psi(t_0, x_0, \xi_0)$  satisfying

$$\xi \mapsto \langle \nu_2(\xi_0), \nabla_{(t,x)}\psi(t_0, x_0, \xi) \rangle$$

has a critical point at  $\xi = \xi_0$ .

The transversality condition we impose on the phase functions  $\phi$  and  $\psi$  is that *the two surfaces  $S_\phi(t_0, x_0)$  with  $S_\psi(t_0, x_0)$  are uniformly transverse for every  $(t_0, x_0)$* : by which we mean that there exists a  $\delta > 0$  such that for each  $(t_0, x_0, \xi_1) \in \text{supp } a$ ,  $(t_0, x_0, \xi_2) \in \text{supp } b$ , we have

$$|\langle \nu_1(\xi_1), \nu_2(\xi_2) \rangle| \leq 1 - \delta. \tag{1-6}$$

This transversality condition is standard in all bilinear oscillatory integral estimates. We remark that there is a slight difference between this definition of transversality and that used in most differential topology textbooks in which the definition of transversality includes manifolds that do not intersect. Here we say that two hypersurfaces are transverse if the intersection of all their translates is transverse in the sense of differential topology.

**Remark.** The phase functions  $\phi$  and  $\psi$  can depend on  $\lambda$  and  $\mu$  as long as the quantitative estimates needed in the proof (namely (1-6) and the derivative bounds mentioned in Equation (1-5)) are satisfied uniformly in  $\lambda$  and  $\mu$  on the support of  $a$  and  $b$ .

The proof of Theorem 1.1 is based on a  $TT^*$  argument and delicate analysis of a cumulative phase function.

**Bilinear Strichartz estimates.** Our main application of the bilinear estimate in Theorem 1.1 is to derive short-range or semiclassical bilinear Strichartz estimates for the Schrodinger equation on closed (compact

---

<sup>1</sup>This is well known in the constant coefficient case; see [Tao 2004].



without boundary)  $d$ -manifolds  $M^d$ . We will also be able to prove mixed bilinear estimates of Schrödinger-wave type as well (see Section 4). Bilinear estimates are of great importance in PDE as they offer refinements to linear Strichartz estimates. The latter are given on  $\mathbb{R}^d$  with its Euclidean Laplacian by

$$\|e^{it\Delta}u_0\|_{L_t^q L_x^r(\mathbb{R}\times\mathbb{R}^d)} \lesssim \|u_0\|_{L^2(\mathbb{R}^d)}, \tag{1-7}$$

where  $(q, r)$  is any *Schrödinger admissible* pair, i.e.,  $2 \leq q, r \leq \infty$ ,  $2/q + d/r = d/2$ , and  $(q, r, d) \neq (2, \infty, 2)$ . The implicit constants depend on  $(q, r, d)$ . These estimates are of fundamental importance in proving both local and global results for nonlinear Schrödinger equations. (See [Tao 2006; Keel and Tao 1998].)

In the case of compact manifolds, the first Strichartz estimates were proved by Bourgain [1993] in the case of the torus. The case of general compact Riemannian manifolds  $(M, g)$  without boundary was dealt with by Burq, Gerard, and Tzvetkov in [Burq et al. 2004] and [Staffilani and Tataru 2002]. In [Burq et al. 2004], the authors prove the estimates

$$\|e^{it\Delta_g}u_0\|_{L_t^q L_x^r([0,1]\times M)} \lesssim_{q,r,M} \|u_0\|_{H^{1/q}(M)} \tag{1-8}$$

for any admissible pair  $(q, r)$ . The proof relies on a construction of an approximate parametrix to the semiclassical operator  $e^{ih\Delta_g}\varphi(h\sqrt{-\Delta_g})$  (where  $\varphi$  is Schwartz) which is used to prove the *semiclassical linear Strichartz estimate*

$$\|e^{it\Delta_g}u_0\|_{L_t^q L_x^r([0,\alpha/N]\times M)} \lesssim_{q,r,M} \|u_0\|_{L^2(M)} \tag{1-9}$$

whenever  $u_0$  is frequency (spectrally) localized at the dyadic scale  $N$  and  $\alpha \ll 1$ . This estimate conforms with the heuristic that Schrödinger evolution moves wavepackets localized at frequency  $\sim N$  at speeds  $\sim N$ , which means that in the time interval  $[0, \alpha/N]$ , one expects the wave packet to remain in a coordinate patch and hence satisfy the same estimates like those on  $\mathbb{R}^d$ . This heuristic will be very useful in predicting the right bilinear estimate later on as well. Notice that (1-8) follows directly from (1-9) by splitting the time interval  $[0, 1]$  into  $N$  subintervals of lengths  $N^{-1}$  and using the conservation of mass and a square function estimate (see [Burq et al. 2004]).

Turning to bilinear estimates, we will start by mentioning the relevant estimate on  $\mathbb{R}^d$  for which we wish to find an analogue on compact manifolds. This estimate first appeared as a refinement to linear Strichartz estimates in [Bourgain 1998]: assuming that  $u_0$  is frequency localized at frequencies  $\{\xi \in \mathbb{R}^d : |\xi| \sim N_1\}$  and  $v_0$  is frequency localized at frequencies  $\{\xi \in \mathbb{R}^d : |\xi| \lesssim N_2\}$  with  $N_2 \leq N_1$ , then

$$\|e^{it\Delta}u_0e^{it\Delta}v_0\|_{L^2(\mathbb{R}\times\mathbb{R}^d)} \lesssim_d \frac{N_2^{(d-1)/2}}{N_1^{1/2}} \|u\|_{L^2(\mathbb{R}^d)} \|v\|_{L^2(\mathbb{R}^d)}. \tag{1-10}$$

We first notice that this estimate is an improvement on applying Hölder’s inequality and the linear Strichartz estimates. In fact, applying the linear estimates only, one would get instead of the  $N_2^{(d-1)/2}/N_1^{1/2}$  constant on the left side of (1-10): 1 for  $d = 2$  (here one uses the  $L_x^2 \rightarrow L_{t,x}^4$  Strichartz estimate) and  $N_2^{d/2-1}$  for  $d \geq 3$  (here one should use Hölder, the  $L_x^2 \rightarrow L_{t,x}^{2(d+2)/d}$  estimate for  $e^{it\Delta}u_0$ , and Bernstein combined

with the

$$L_x^2 \rightarrow L_t^{d+2} L_x^{\frac{2d(d+2)}{d(d+2)-4}}$$

estimate for  $e^{it\Delta} v_0$ ). Bourgain used this improvement (when  $N_2 \ll N_1$ ) to prove, among other things, global well-posedness below energy norm for certain mass (and  $\dot{H}^{1/2}$ )-critical equations (which incidentally is also an application that will be considered in the context of closed manifolds in [Hani 2012]). Since then, this improvement and variants of it proved to be of essential use in studying nonlinear Schrödinger equations.

In the context of compact manifolds, some bilinear estimates on the torus were already implicit in [Bourgain 1993] (see also [Burq et al. 2005a]), and other variants were proved in [De Silva et al. 2007]. In [Burq et al. 2005a; 2005b], the authors prove bilinear Strichartz estimates on spheres  $S^2$  and  $S^3$  (and on the bit wider class of Zoll manifolds) using bilinear eigenfunction cluster estimates. These bilinear Strichartz estimates take the form

$$\|e^{it\Delta_g} u_0 e^{it\Delta_g} v_0\|_{L_{t,x}^2([0,1] \times S^d)} \lesssim_d N_2^{\alpha_d} \|u_0\|_{L^2(S^d)} \|v_0\|_{L^2(S^d)}$$

whenever  $u_0$  is spectrally localized in the dyadic region  $\sqrt{-\Delta_g} \in [N_1, 2N_1)$ ,  $v_0$  in the region  $\sqrt{-\Delta_g} \in [N_2, 2N_2)$ ,  $N_2 \leq N_1$ , with  $\alpha = \frac{1}{4} + \epsilon$  when  $d = 2$  and  $\alpha = \frac{1}{2} + \epsilon$  when  $d = 3$ .

Using Theorem 1.1, we will be able to prove the following bilinear estimate for any closed manifold  $(M, g)$ :

**Theorem 1.2.** *Suppose  $u_0, v_0 \in L^2(M^d)$  are spectrally localized at dyadic scales  $N_1$  and  $N_2$  as above with  $N_2 \leq N_1$ . Then the estimate*

$$\|e^{it\Delta_g} u_0 e^{it\Delta_g} v_0\|_{L_{t,x}^2([-1/N_1, 1/N_1] \times M)} \lesssim_M \frac{N_2^{(d-1)/2}}{N_1^{1/2}} \|u_0\|_{L^2(M)} \|v_0\|_{L^2(M)}. \tag{1-11}$$

holds. More generally,

$$\|e^{it\Delta_g} u_0 e^{it\Delta_g} v_0\|_{L^2([-T, T] \times M)} \leq \Lambda(T, N_1, N_2) \|u_0\|_{L^2(M)} \|v_0\|_{L^2(M)}, \tag{1-12}$$

where

$$\Lambda(T, N_1, N_2) \lesssim_M \begin{cases} N_2^{(d-1)/2} / N_1^{1/2} & \text{if } T \ll N_1^{-1}, \\ T^{1/2} N_2^{(d-1)/2} & \text{if } T \gtrsim N_1^{-1}. \end{cases} \tag{1-13}$$

In particular, for  $T = 1$  we have

$$\|e^{it\Delta_g} u_0 e^{it\Delta_g} v_0\|_{L^2([-1, 1] \times M)} \lesssim N_2^{(d-1)/2} \|u_0\|_{L^2(M)} \|v_0\|_{L^2(M)}. \tag{1-14}$$

Some notes are in order: First we notice that in the semiclassical/ short-range case (1-11), the coefficient  $N_2^{(d-1)/2} / N_1^{1/2}$  is the same as that on  $\mathbb{R}^d$ . This conforms with the heuristic that in the time interval  $[0, 1/N_1]$ , the two waves  $e^{it\Delta_g} v_0$  (which is moving with speed  $\sim N_1$ ) and  $e^{it\Delta_g} u_0$  (moving at speed  $\sim N_2 \leq N_1$ ) do not leave a coordinate patch and hence their product satisfies the same estimate as that on  $\mathbb{R}^d$ . Second, the estimates in (1-12) and (1-14) are essentially obtained from (1-11) by splitting the time interval into pieces of length  $N_1^{-1}$ . It should be emphasized though that the exact dependence of

$\Lambda(T, N_1, N_2)$  on its all parameters is often of great importance in applications (see [Hani 2012]). In fact, it is easy to see that bilinear estimates on the interval  $[0, T]$  translate by scaling into bilinear estimates on the interval  $[0, 1]$  for the rescaled manifold  $\lambda M$ .<sup>2</sup> The  $\lambda$ -dependence of those estimates is dictated by dependence of  $\Lambda(T, N_1, N_2)$  on all its parameters. The bilinear Strichartz estimates on  $\lambda M$  take the following form (see [Hani 2012] for relevant calculations):

**Corollary 1.3** (Time  $T$  estimate on  $M$  implies time 1 estimate on  $\lambda M$ ). *Let  $M$  be a 2D closed manifold and suppose that  $N_1, N_2 \in 2^{\mathbb{Z}}$  and suppose  $u_0, v_0 \in L^2(\lambda M)$  are spectrally localized around  $N_1$  and  $N_2$  respectively, with  $N_2 \leq N_1$ . Then*

$$\|e^{it\Delta_\lambda} u_0 e^{it\Delta_\lambda} v_0\|_{L^2([0,1] \times \lambda M)} \lesssim_M \Lambda(\lambda^{-2}, \lambda N_1, \lambda N_2) \|u_0\|_{L^2(\lambda M)} \|v_0\|_{L^2(\lambda M)} \tag{1-15}$$

$$\lesssim_M \begin{cases} (N_2/N_1)^{1/2} \|u_0\|_{L^2(\lambda M)} \|v_0\|_{L^2(\lambda M)} & \text{if } \lambda \gg N_1, \\ (N_2/\lambda)^{1/2} \|u_0\|_{L^2(\lambda M)} \|v_0\|_{L^2(\lambda M)} & \text{if } \lambda \lesssim N_1, \end{cases} \tag{1-16}$$

where we have denoted by  $\Delta_\lambda$  the Laplace–Beltrami operator on the rescaled manifold  $\lambda M$ .

Having favorable bounds (in terms of  $\lambda$  and  $N_2$ ) on the right hand side of (1-16) is crucial to obtaining global well-posedness of some nonlinear equations on  $M$  below energy norm. In fact, in [Hani 2012] it is proven that the cubic nonlinear Schrödinger equation is globally well-posed in  $H^s(M)$  for any closed 2D surface  $M^2$  and all  $s > \frac{2}{3}$ , a result which matches the current (to the best of our knowledge) minimum regularity needed for global well-posedness on the 2-torus.

Finally, we note that as in the case of bilinear estimates on  $\mathbb{R}^d$ , the bilinear estimates in (1-11) and (1-12) offer a refinement to those obtained by using linear estimates alone. However, this refinement is only visible when one looks at estimates over time intervals  $[0, T]$  for  $T \ll N_2^{-1}$  (or alternatively, estimates on rescaled manifolds). For example, for  $d \geq 3$ , applying Hölder’s inequality, the  $L_t^\infty L_x^2$  bound on  $e^{it\Delta} u_0$ , Bernstein and the  $L_t^2 L_x^{2d/(d-2)}$  for  $e^{it\Delta} v_0$ , one gets

$$\|e^{it\Delta} u_0 e^{it\Delta} v_0\|_{L_{t,x}^2([0,T] \times M)} \lesssim C(T, N_2) \|u_0\|_{L^2(M)} \|v_0\|_{L^2(M)},$$

where  $C(T, N_2) = N_2^{(d-2)/2} = N_2^{(d-1)/2} / N_2^{1/2}$  for  $T \lesssim N_2^{-1}$  and  $C(T, N_2) = T^{1/2} N_2^{(d-1)/2}$  for  $T \geq N_2^{-1}$ . This shows the improvement offered by (1-12) in the range  $T \ll N_2^{-1}$  (especially when dealing with low-high frequency interaction  $N_2 \ll N_1$ ). This improvement is due to the cancellation happening when we multiply the high frequency wave with the low frequency one. This cancellation is completely ignored by linear estimates. In the case,  $d = 2$ , one would need to prove an estimate for the inadmissible pair  $(q, r) = (2, \infty)$ . This is possible with an  $N^\epsilon$  loss. See [Jiang 2011]. In this case, the bilinear estimate (1-12) not only offers a refinement to linear estimates at time scales  $T \ll 1$  and in the range  $N_2 \ll N_1$ , but also yields better estimates in the time scale  $T = 1$  (no  $N_2^\epsilon$  loss in (1-14)). See [Hani 2012] for details.

The paper is organized as follows. In Section 2 we provide the proof of Theorem 1.1. In Section 3, we review the needed facts about the parametrix construction in [Burq et al. 2004] and prove Theorem 1.2.

---

<sup>2</sup>Here  $\lambda M$  can either be viewed as the Riemannian manifold  $(M, (1/\lambda^2)g)$  or by embedding  $M$  into some ambient space  $\mathbb{R}^N$  and then applying a dilation by  $\lambda$  to get  $\lambda M$ .

Finally in Section 4 we prove inhomogeneous versions of the bilinear Strichartz estimates stated above in addition to mixed type bilinear estimates for products of the Schrödinger propagator  $e^{it\Delta}u_0$  and the half wave propagators  $e^{\pm it|\nabla|v_0}$ . These estimates can also be deduced from Theorem 1.1 and have potential applications (to be investigated elsewhere) in studying Zakharov type systems on closed manifolds. We use the notation  $A \lesssim B$  to denote  $A \leq CB$  for some  $C > 0$  and  $A \sim B$  to denote  $A \lesssim B \lesssim A$ .

**2. Proof of Theorem 1.1**

All implicit constants are allowed to depend on  $d, \delta$  and uniform bounds on a finite number of derivatives of  $\phi, \psi, a$  and  $b$ . We have

$$T_\lambda f(t, x) \tilde{T}_\mu g(t, x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{i(\lambda\phi(t,x,\xi_1) + \mu\psi(t,x,\xi_2))} a(t, x, \xi_1) b(t, x, \xi_2) f(\xi_1) g(\xi_2) d\xi_1 d\xi_2. \quad (2-1)$$

Since the supports of  $a$  and  $b$  are compact, one can use a finite partition of unity to split  $a$  and  $b$  into finitely many pieces so that on the support of each piece there exists  $t_0, x_0, \xi_0, \xi_{2,0}$  such that

$$|t - t_0|, |x - x_0|, |\xi_1 - \xi_0|, |\xi_2 - \xi_{2,0}| \leq \frac{1}{C},$$

where  $C$  is some large constant depending only on  $\delta$  and the uniform norms of  $\phi$  and  $\psi$  and their derivatives on the compact supports of  $a$  and  $b$ .

Also notice that by applying a rotation  $L$  of the domain  $\mathbb{R} \times \mathbb{R}^d: (t, x) = L^T(s, y)$ , the left hand side of (1-5) is unaffected, whereas the hypersurfaces  $S_\phi$  and  $S_\psi$  are both rotated by  $L$ . In fact, since

$$\nabla_{(s,y)}(\phi(L^T(s, y), x, \xi)) = L(\nabla\phi)(L^T(s, y), \xi),$$

where  $\nabla$  is taken in the first  $d + 1$  variables of  $\phi$ . Consequently, if we apply the change of variable  $(t, x) = L^T(s, y)$ , the canonical hypersurfaces  $S_\phi$  and  $S_\psi$  are both rotated by  $L$ . Using this symmetry, one can assume that

$$\left| \det \left( \frac{\partial^2 \phi}{\partial \xi \partial x}(t_0, x_0, \xi_0) \right) \right| \gtrsim 1 \quad \text{and} \quad \left| \det \left( \frac{\partial^2 \psi}{\partial \xi \partial x}(t_0, x_0, \xi_{2,0}) \right) \right| \gtrsim 1 \quad (2-2)$$

on the support of  $a$  and of  $b$ , respectively. This means that the surfaces  $S_\phi$  and  $S_\psi$  can be regarded as graphs of functions of the form  $(\xi, \tau_1(\xi))$  and  $(\xi, \tau_2(\xi)) \subset T_{(t_0,x_0)}^* \mathbb{R}^{d+1}$  respectively.

Define

$$A := \frac{\partial^2 \phi}{\partial \xi \partial x}(t_0, x_0, \xi_0) \quad \text{and} \quad B := \frac{\partial^2 \psi}{\partial \xi \partial x}(t_0, x_0, \xi_{2,0}).$$

By the above, we have that  $A$  and  $B$  are invertible. It will be convenient later on to do the following change of variables in the  $\xi_1$  integral and define  $\xi = \xi_1 + (\mu/\lambda)A^{-1}B\xi_2$ .<sup>3</sup> This gives

$$\begin{aligned} & T_\lambda f(t, x) \tilde{T}_\mu g(t, x) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{i\lambda(\phi(t,x,\xi - (\mu/\lambda)A^{-1}B\xi_2) + (\mu/\lambda)\psi(t,x,\xi_2))} c(t, x, \xi, \xi_2) f\left(\xi - \frac{\mu}{\lambda}A^{-1}B\xi_2\right) g(\xi_2) d\xi d\xi_2, \end{aligned} \quad (2-3)$$

<sup>3</sup>The justification for this change of variables will be obvious later on. However, at a heuristic level this corresponds to adding the momenta of the two waves.

where we set  $c(t, x, \xi, \xi_2) = a(t, x, \xi - (\mu/\lambda)A^{-1}B\xi_2)b(t, x, \xi_2)$  and all we have to remember about  $c$  is that it is uniformly bounded along with all its derivatives (since  $\mu/\lambda \leq 1$ ) and is supported in a small neighborhood of  $(t_0, x_0, \xi_0 + (\mu/\lambda)A^{-1}B\xi_{2,0}, \xi_{2,0})$  of diameter  $\lesssim 1/C$ . In particular, we have

$$\left| \xi - \frac{\mu}{\lambda}A^{-1}B\xi_2 - \xi_0 \right| \leq \frac{1}{C} \quad (2-4)$$

for every  $\xi, \xi_2$  in the support of  $c$ .

We now fix a particular coordinate direction  $e_j$  (to be specified later), and write  $\xi_2 = pe_j + \xi'_2$ . Roughly speaking, the direction will be chosen using the transversality assumption of the two surfaces  $S_\phi$  and  $S_\psi$  so that

$$\left\langle v_1(\xi_0), \frac{\partial^2 \psi(t_0, x_0, \xi_{2,0})}{\partial \xi \partial(t, x)} e_j \right\rangle \gtrsim_\delta 1. \quad (2-5)$$

(The inner product is in  $\mathbb{R}^{d+1}$ , the second entry being the product of a  $(d+1) \times d$  matrix with a vector in  $\mathbb{R}^d$ .) This will be possible because  $v_2$  is the unique direction for which

$$\left\langle v_2, \frac{\partial^2 \psi(t_0, x_0, \xi_{2,0})}{\partial \xi \partial(t, x)} \right\rangle = \vec{0}_{\mathbb{R}^d};$$

since  $v_1$  is not a multiple of  $v_2$ , the vector

$$\left\langle v_1, \frac{\partial^2 \psi(t_0, x_0, \xi_{2,0})}{\partial \xi \partial(t, x)} \right\rangle$$

is also nonzero, so there exists a coordinate direction  $e_j$  onto which the projection of this nonzero vector does not vanish. In other words, the inner product in (2-5) can be thought of as the projection of  $v_1$  onto the curve in  $S_\phi(t_0, x_0)$  given by  $t \mapsto \nabla_{(t,x)} \psi(t_0, x_0, \xi_{2,0} + te_j)$ .

For convenience of notation, when confusion does not arise, we will assume that  $j = 1$  and write  $\xi_2 = (p, \xi'_2)$  where  $p \in \mathbb{R}$  and  $\xi'_2 \in \mathbb{R}^{d-1}$ . As a result, we have

$$\begin{aligned} & \|T_\lambda f(t, x) \tilde{T}_\mu g(t, x)\|_2 \\ &= \left\| \int_{\mathbb{R}_{\xi'_2}^{d-1}} \int_{\mathbb{R}_\xi^d} \int_{\mathbb{R}_p} e^{i\lambda(\phi(t,x,\xi - (\mu/\lambda)A^{-1}B\xi_2) + (\mu/\lambda)\psi(t,x,\xi_2))} c(t, x, \xi, \xi_2) f\left(\xi - \frac{\mu}{\lambda}\xi_2\right) g(\xi_2) d\xi dp d\xi'_2 \right\|_2 \\ &\leq \int_{\mathbb{R}_{\xi'_2}^{d-1}} \left\| \int_{\mathbb{R}_\xi^d} \int_{\mathbb{R}_p} e^{i\lambda(\phi(t,x,\xi - (\mu/\lambda)A^{-1}B\xi_2) + (\mu/\lambda)\psi(t,x,\xi_2))} c(t, x, \xi, \xi_2) f\left(\xi - \frac{\mu}{\lambda}\xi_2\right) g(\xi_2) d\xi dp \right\|_{L_{t,x}^2} d\xi'_2. \end{aligned}$$

Freezing  $\xi'_2$ , we define the operator  $S = S_{\xi'_2} : L^2(\mathbb{R}^{d+1}) \rightarrow L^2(\mathbb{R}^{d+1})$  given by

$$SF(t, x) = \int_{\mathbb{R}_\xi^d} \int_{\mathbb{R}_p} e^{i\lambda(\phi(t,x,\xi - (\mu/\lambda)A^{-1}B\xi_2) + (\mu/\lambda)\psi(t,x,\xi_2))} c(t, x, \xi, \xi_2) F(\xi, p) d\xi dp, \quad (2-6)$$

where  $\xi_2 = (p, \xi'_2)$ . As a result of this definition, our estimate is reduced to proving that for each  $\xi'_2$ , the estimate

$$\|SF\|_{L_{t,x}^2(\mathbb{R}^{d+1})} \lesssim \frac{1}{\lambda^{d/2} \mu^{1/2}} \|F\|_{L_{p,\xi}^2(\mathbb{R}^{d+1})} \quad (2-7)$$

holds for  $S$ . In fact, with such an estimate and by Cauchy–Schwarz in the  $\xi'_2$  integral (keeping in mind that  $c$  is compactly supported), we get

$$\begin{aligned} \|T_\lambda f(t, x) \tilde{T}_\mu g(t, x)\|_2 &\lesssim \frac{1}{\lambda^{d/2} \mu^{1/2}} \int_{|\xi'_2| \lesssim 1} \left\| f\left(\xi - \frac{\mu}{\lambda}(p, \xi'_2)\right) g(p, \xi'_2) \right\|_{L^2_{p, \xi}} d\xi'_2 \\ &\lesssim \frac{1}{\lambda^{d/2} \mu^{1/2}} \|f\|_{L^2} \|g\|_{L^2}. \end{aligned}$$

The bound on  $S$  is proved using a  $T^*T$  argument. For convenience of notation, let us define

$$\Phi(t, x, \xi, p) = \phi\left(t, x, \xi - \frac{\mu}{\lambda} A^{-1} B \xi_2\right) + \frac{\mu}{\lambda} \psi(t, x, \xi_2), \tag{2-8}$$

where  $\xi_2 = (p, \xi'_2)$ . With this notation,  $S$  takes the form

$$SF(t, x) = \int_{\mathbb{R}_\xi^d} \int_{\mathbb{R}_p} e^{i\lambda\Phi(t, x, \xi, p)} c(t, x, \xi, p) F(\xi, p) d\xi dp.$$

The adjoint of  $S$  is given by the operator

$$S^*G(\xi, p) = \int_{\mathbb{R}_x^d} \int_{\mathbb{R}_t} e^{-i\lambda\Phi(t, x, \xi, p)} \bar{c}(t, x, \xi, p) G(x, t) dx dt.$$

As a result, we get

$$S^*SF(\zeta, q) = \int_{\mathbb{R}_\xi^d} \int_{\mathbb{R}_p} K(\zeta, q, \xi, p) F(\xi, p) d\xi dp, \tag{2-9}$$

where

$$K(\zeta, q, \xi, p) = \int_{\mathbb{R}_t} \int_{\mathbb{R}_x^d} e^{i\lambda[\Phi(t, x, \xi, p) - \Phi(t, x, \zeta, q)]} c(t, x, \xi, p) \bar{c}(t, x, \zeta, q) dx dt. \tag{2-10}$$

Our aim will be to show that  $K$  satisfies the bound

$$K(\zeta, q, \xi, p) \lesssim_N \frac{1}{(1 + \lambda|\xi - \zeta| + \mu|q - p|)^N} \tag{2-11}$$

for a sufficiently large  $N$  (any  $N > d + 1$  would do).

In fact, with such an estimate, one can easily see (using Schur’s test for example) that  $\|S^*S\|_{L^2 \rightarrow L^2} \lesssim 1/(\lambda^d \mu)$ . Since  $\|S\|_{L^2 \rightarrow L^2} = \|S^*S\|_{L^2 \rightarrow L^2}^{1/2}$  one gets that  $\|S\|_{L^2 \rightarrow L^2}$  is bounded by  $O(1/(\lambda^{d/2} \mu^{1/2}))$ .

The bound on  $K$  is based on nonstationary-phase-type estimates and integration by parts. These are based on the following estimates on the phase function  $\Phi$  and its derivatives.

**Lemma 2.1.** *There exists  $\Omega \in S^d$  such that*

$$\left| \langle \nabla_{t,x} \Phi(t, x, \xi, p) - \nabla_{t,x} \Phi(t, x, \zeta, q), \Omega \rangle \right| \gtrsim |\xi - \zeta| + \frac{\mu}{\lambda} |p - q| \tag{2-12}$$

and

$$\left| \frac{\partial}{\partial x^\alpha \partial t^\beta} (\Phi(t, x, \xi, p) - \Phi(t, x, \zeta, q)) \right| \lesssim_{\alpha, \beta} |\xi - \zeta| + \frac{\mu}{\lambda} |p - q|. \tag{2-13}$$

*Proof.* The second estimate (2-13) is a direct consequence of the definition (2-8), the Taylor expansion, and the uniform boundedness of all the  $t, x$  derivatives of  $\phi$  and  $\psi$ . We now turn to the proof of (2-12).

Here we split the analysis into two cases:

**Case I:**  $|\xi - \zeta| \geq \frac{1}{100}(\mu/\lambda)|p - q|$ . The change of variables we have made in (2-3) will allow us to prove (2-12) in this case using only the  $x$  derivative part of  $\nabla_{t,x}\Phi$ . In fact, using (2-8), we have

$$\nabla_x \Phi(t, x, \xi, p) - \nabla_x \Phi(t, x, \zeta, q) = \nabla_x \phi\left(t, x, \xi - \frac{\mu}{\lambda} A^{-1} B \xi_2\right) - \nabla_x \phi\left(t, x, \zeta - \frac{\mu}{\lambda} A^{-1} B \zeta_2\right) \quad (2-14)$$

$$+ \frac{\mu}{\lambda} (\nabla_x \psi(t, x, \xi_2) - \nabla_x \psi(t, x, \zeta_2)), \quad (2-15)$$

where  $\zeta_2 = (q, \xi_2')$ . We estimate (2-14) in the following manner:

$$\begin{aligned} \nabla_x \phi\left(t, x, \xi - \frac{\mu}{\lambda} A^{-1} B \xi_2\right) - \nabla_x \phi\left(t, x, \zeta - \frac{\mu}{\lambda} A^{-1} B \zeta_2\right) &= \left\langle \frac{\partial^2 \phi}{\partial \xi \partial x}\left(t, x, \xi - \frac{\mu}{\lambda} A^{-1} B \xi_2\right), \xi - \zeta - \frac{\mu}{\lambda} A^{-1} B (\xi_2 - \zeta_2) \right\rangle + O(|\xi - \zeta|^2) \\ &= \left\langle \frac{\partial^2 \phi}{\partial \xi \partial x}(t_0, x_0, \xi_0), \xi - \zeta - \frac{\mu}{\lambda} A^{-1} B (\xi_2 - \zeta_2) \right\rangle + \text{Error}_1 \\ &= A(\xi - \zeta) - \frac{\mu}{\lambda} B(\xi_2 - \zeta_2) + \text{Error}_1, \end{aligned}$$

where we used the fact that  $A = (\partial^2 \phi / \partial \xi \partial x)(t_0, x_0, \xi_0)$ . Here

$$\begin{aligned} \text{Error}_1 &= \left\langle \frac{\partial^2 \phi}{\partial \xi \partial x}\left(t, x, \xi - \frac{\mu}{\lambda} A^{-1} B \xi_2\right), \xi - \zeta - \frac{\mu}{\lambda} A^{-1} B (\xi_2 - \zeta_2) \right\rangle \\ &\quad - \left\langle \frac{\partial^2 \phi}{\partial \xi \partial x}(t_0, x_0, \xi_0), \xi - \zeta - \frac{\mu}{\lambda} A^{-1} B (\xi_2 - \zeta_2) \right\rangle + O(|\xi - \zeta|^2). \end{aligned}$$

By our assumption of smallness of the support of  $c$  (cf. (2-4)), the error can be estimated (if  $C$  is chosen large enough depending on the uniform norms of derivatives of  $\phi$ ) by

$$|\text{Error}_1| \lesssim_\phi \frac{1}{C} |\xi - \zeta - \frac{\mu}{\lambda} A^{-1} B (\xi_2 - \zeta_2)| + O(|\xi - \zeta|^2) \leq \frac{1}{10} \gamma_1 |\xi - \zeta|,$$

where  $\gamma_1$  is chosen to be the smallest singular value of  $A$  (or equivalently  $\gamma_1 = \min_{z \in S^{d-1}} |Az|$ ).

Next we estimate (2-15) by

$$\begin{aligned} \frac{\mu}{\lambda} (\nabla_x \psi(t, x, \xi_2) - \nabla_x \psi(t, x, \zeta_2)) &= \frac{\mu}{\lambda} \left\langle \frac{\partial^2 \psi}{\partial \xi \partial x}(t, x, \xi_2), \xi_2 - \zeta_2 \right\rangle + O\left(\frac{\mu}{\lambda} |\xi_2 - \zeta_2|^2\right) \\ &= \frac{\mu}{\lambda} \left\langle \frac{\partial^2 \psi}{\partial \xi \partial x}(t_0, x_0, \xi_{2,0}), \xi_2 - \zeta_2 \right\rangle + \text{Error}_2 \\ &= \frac{\mu}{\lambda} B(\xi_2 - \zeta_2) + \text{Error}_2, \end{aligned}$$

where

$$\text{Error}_2 = \frac{\mu}{\lambda} \left( \left\langle \frac{\partial^2 \psi}{\partial \xi \partial x}(t, x, \xi_2), \xi_2 - \zeta_2 \right\rangle - \left\langle \frac{\partial^2 \psi}{\partial \xi \partial x}(t_0, x_0, \xi_{2,0}), \xi_2 - \zeta_2 \right\rangle \right) + O\left(\frac{\mu}{\lambda} |\xi_2 - \zeta_2|^2\right),$$

which, as before, can be bounded (using the bounds  $|\xi_2 - \zeta_2|$ ,  $|\xi_2 - \xi_{2,0}| \lesssim \frac{1}{C}$  and  $\frac{\mu}{\lambda} |\xi_2 - \zeta_2| \leq 100|\xi - \zeta|$ ) by

$$|\text{Error}_2| \leq \frac{1}{10} \gamma_1 |\xi - \zeta|.$$

Collecting these estimates we get

$$\nabla_x \Phi(t, x, \xi, p) - \nabla_x \Phi(t, x, \zeta, q) = A(\zeta - \xi) + \text{Error}_1 + \text{Error}_2, \quad (2-16)$$

where  $\text{Error}_1 + \text{Error}_2$  is bounded by  $\frac{1}{5} \gamma_1 |\zeta - \xi|$ . We now let  $\omega \in S^{d-1}$  be equal to  $A(\zeta - \xi)/|A(\zeta - \xi)|$ . Since

$$|\langle A(\zeta - \xi), \omega \rangle| = |A(\xi - \zeta)| \geq \gamma_1 |\xi - \zeta|$$

by the definition of  $\gamma_1$ , we get

$$|\langle \nabla_x \Phi(t, x, \xi, p) - \nabla_x \Phi(t, x, \zeta, q), \omega \rangle| \gtrsim |\xi - \zeta|.$$

As a result, by taking  $\Omega \in S^d$  equal to  $(\omega, 0)$  we get

$$|\langle \nabla_{t,x} \Phi(t, x, \xi, p) - \nabla_{t,x} \Phi(t, x, \zeta, q), \Omega \rangle| \gtrsim |\xi - \zeta| \gtrsim |\xi - \zeta| + \frac{\mu}{\lambda} |p - q|, \quad (2-17)$$

which is (2-12) in Case 1.

**Case 2:**  $|\xi - \zeta| \leq \frac{1}{100} (\mu/\lambda) |p - q|$ . The analysis in this case is a bit more delicate, and it is here that the transversality assumption is used. In this case, we will take  $\Omega = \nu_1(\xi_0)$ , the normal to the surface  $\xi \mapsto \nabla_{t,x} \phi(t_0, x_0, \xi)$  at  $\xi_0$ . With this choice we have

$$\begin{aligned} & \langle \nabla_{t,x} \Phi(t, x, \xi, p) - \nabla_{t,x} \Phi(t, x, \zeta, q), \Omega \rangle \\ &= \left\langle \nabla_{t,x} \phi \left( t, x, \xi - \frac{\mu}{\lambda} A^{-1} B \xi_2 \right) - \nabla_{t,x} \phi \left( t, x, \zeta - \frac{\mu}{\lambda} A^{-1} B \zeta_2 \right), \nu_1(\xi_0) \right\rangle \end{aligned} \quad (2-18)$$

$$+ \frac{\mu}{\lambda} \langle \nabla_{t,x} \psi(t, x, \xi_2) - \nabla_{t,x} \psi(t, x, \zeta_2), \nu_1(\xi_0) \rangle. \quad (2-19)$$

The main term in this expression comes from (2-19), whereas (2-18) will be treated as an error. We start by lower bounding (2-19).

We have

$$\begin{aligned} \nabla_{t,x} \psi(t, x, \xi_2) - \nabla_{t,x} \psi(t, x, \zeta_2) &= \left\langle \frac{\partial^2 \psi}{\partial \xi \partial(x, t)}(t, x, \xi_2), \xi_2 - \zeta_2 \right\rangle + O(|\xi_2 - \zeta_2|^2) \\ &= \left\langle \frac{\partial^2 \psi}{\partial \xi \partial(x, t)}(t_0, x_0, \xi_{2,0}), \xi_2 - \zeta_2 \right\rangle + \text{Error}_1 \\ &= (p - q) \left\langle \frac{\partial^2 \psi}{\partial \xi \partial(x, t)}(t_0, x_0, \xi_{2,0}), e_j \right\rangle + \text{Error}_1, \end{aligned}$$

where

$$\text{Error}_1 = \left\langle \frac{\partial^2 \psi}{\partial \xi \partial(x, t)}(t, x, \xi_2), \xi_2 - \zeta_2 \right\rangle - \left\langle \frac{\partial^2 \psi}{\partial \xi \partial(x, t)}(t_0, x_0, \xi_{2,0}), \xi_2 - \zeta_2 \right\rangle + O(|\xi_2 - \zeta_2|^2).$$



This is estimated as before using the small support assumption to get

$$|\text{Error}_1| \lesssim_\psi \frac{1}{C} |\xi_2 - \zeta_2| \leq \frac{1}{C} |p - q|, \tag{2-20}$$

where we have used in the last inequality the fact that  $\xi_2 = (p, \xi'_2)$  and  $\zeta_2 = (q, \xi'_2)$ . We remark that

$$N := \frac{\partial^2 \psi}{\partial \xi \partial (x, t)}(t_0, x_0, \xi_{2,0})$$

is a  $(d + 1) \times d$  matrix, so  $\langle N, e_j \rangle$  is a vector in  $\mathbb{R}^{d+1}$ . From a geometric point of view, this vector lies in the tangent space to  $S_\psi(t_0, x_0)$  at  $\xi_{2,0}$ .

Recall that by definition,  $v_2 := v_2(\xi_{2,0})$  is the unique vector (up to sign) in  $S^d$  such that  $v_2^T N = 0$  where  $v_2^T$  is the row vector corresponding to  $v_2$ . In particular, the map from the  $d$ -dimensional subspace  $v_2^\perp \subset \mathbb{R}^{d+1}$  into  $\mathbb{R}^d$  given by

$$v \in v_2^\perp \mapsto v^T N \in \mathbb{R}^d$$

is an isomorphism. Let  $\gamma_2 > 0$  denote its smallest singular value (or equivalently  $\gamma_2$  is the *positive* infimum of the above map when  $v \in v_2^\perp$  satisfies  $\|v\| = 1$ ).

Writing  $v_1(\xi_0) = \alpha v_2 + \beta v_3$  with  $v_3 \in v_2^\perp$ ,  $\|v_3\| = 1$ , and  $|\alpha|, |\beta| \leq 1$ , we notice that since  $1 - \delta > |\langle v_1, v_2 \rangle| = |\alpha|$  we have that  $|\beta| = \sqrt{1 - \alpha^2} \geq \sqrt{\delta}$ .

As a result, we have

$$\langle v_1, \nabla_{t,x} \psi(t, x, \xi_2) - \nabla_{t,x} \psi(t, x, \zeta_2) \rangle = (p - q) v_1^T N e_j + \text{Error}_1 = \beta(p - q) v_3^T N e_j + \text{Error}_1.$$

Since  $\|v_3^T N\| \geq \gamma_2$ , one can choose  $e_j$  so that  $|v_3^T N e_j| \geq \gamma_2 / \sqrt{d} =: c_1$ . Combining this to the estimate on  $\text{Error}_1$  in (2-20) above we get that if  $C$  is large enough, then

$$\left| \langle v_1, \nabla_{t,x} \psi(t, x, \xi_2) - \nabla_{t,x} \psi(t, x, \zeta_2) \rangle \right| \geq c_1 \sqrt{\delta} |p - q| - \frac{c_1 \sqrt{\delta}}{100} |p - q| \geq \frac{99}{100} c_1 \sqrt{\delta} |p - q|. \tag{2-21}$$

As mentioned before, we will treat (2-18) as an error. Indeed,

$$\begin{aligned} & \left\langle \nabla_{t,x} \phi \left( t, x, \xi - \frac{\mu}{\lambda} A^{-1} B \xi_2 \right) - \nabla_{t,x} \phi \left( t, x, \zeta - \frac{\mu}{\lambda} A^{-1} B \zeta_2 \right), v_1(\xi_0) \right\rangle \\ &= v_1(\xi_0)^T D_{(d+1) \times d} \left( t, x, \xi - \frac{\mu}{\lambda} A^{-1} B \xi_2 \right) \left[ \xi - \zeta - \frac{\mu}{\lambda} A^{-1} B (\xi_2 - \zeta_2) \right] + O \left( \left| \frac{\mu}{\lambda} (p - q) \right|^2 \right), \end{aligned}$$

where we have defined

$$D_{(d+1) \times d}(t, x, \eta) = \frac{\partial^2 \phi}{\partial \xi \partial (x, t)}(t, x, \eta)$$

and also used that  $|\xi - \zeta| \leq (\mu/\lambda)|p - q|$  in this case. Since the derivatives of  $D$  are uniformly bounded and because of the small support assumption (2-4), we have

$$\left\| D_{(d+1) \times d} \left( t, x, \xi - \frac{\mu}{\lambda} A^{-1} B \xi_2 \right) - D_{(d+1) \times d}(t_0, x_0, \xi_0) \right\| \lesssim \frac{1}{C} \leq \frac{c_1 \sqrt{\delta}}{100(\|A^{-1} B\| + 1)}$$

if  $C$  is large enough.

Using the fact that  $v_1^T D_{(d+1) \times d}(t_0, x_0, \xi_0) = 0$ , we get that

$$\left| \left\langle \nabla_{t,x} \phi \left( t, x, \xi - \frac{\mu}{\lambda} A^{-1} B \xi_2 \right) - \nabla_{t,x} \phi \left( t, x, \zeta - \frac{\mu}{\lambda} A^{-1} B \zeta_2 \right), v_1(\xi_0) \right\rangle \right| \leq \frac{c_1 \sqrt{\delta} \mu}{50 \lambda} |p - q| \quad (2-22)$$

again using the small support assumption.

Combining (2-22) and (2-21), we get (2-12) for Case 2.

Now we are ready to perform the integration by parts needed to prove the estimate (2-11). Recall that

$$K(\zeta, q, \xi, p) = \int_{\mathbb{R}_t} \int_{\mathbb{R}_x^d} e^{i\lambda[\Phi(t,x,\xi,p) - \Phi(t,x,\zeta,q)]} c(t, x, \xi, p) \bar{c}(t, x, \zeta, q) dx dt.$$

Let  $D_\Omega$  be the operator given by

$$D_\Omega := \frac{1}{i\lambda \langle \nabla_{t,x} \Phi(t, x, \xi, p) - \nabla_{t,x} \Phi(t, x, \zeta, q), \Omega \rangle} \langle \nabla_{(x,t)}, \Omega \rangle. \quad (2-23)$$

Then

$$D_\Omega(e^{i\lambda(\Phi(t,x,\xi,\xi_2) - \Phi(t,x,\zeta,\zeta_2))}) = e^{i\lambda(\Phi(t,x,\xi,\xi_2) - \Phi(t,x,\zeta,\zeta_2))}.$$

Noticing that the formal adjoint of  $D_\Omega$  acting on  $L^2$  is

$$D_\Omega^T = \langle \nabla_{(x,t)}, \Omega \rangle \frac{1}{(i\lambda \langle \nabla_{t,x} \bar{\Phi}(t, x, \xi, p) - \nabla_{t,x} \bar{\Phi}(t, x, \zeta, q), \Omega \rangle)},$$

we get

$$\begin{aligned} K(\zeta, q, \xi, p) &= \int_{\mathbb{R}_t} \int_{\mathbb{R}_x^d} e^{i\lambda[\Phi(t,x,\xi,p) - \Phi(t,x,\zeta,q)]} c(t, x, \xi, p) \bar{c}(t, x, \zeta, q) dx dt \\ &= \int_{\mathbb{R}_t} \int_{\mathbb{R}_x^d} e^{i\lambda[\Phi(t,x,\xi,p) - \Phi(t,x,\zeta,q)]} \overline{(D_\Omega^T)^N \bar{c}(t, x, \xi, p) c(t, x, \zeta, q)} dx dt. \end{aligned}$$

Using the estimates in Lemma 2.1, it is easy to see that

$$(D_\Omega^T)^N \bar{c}(t, x, \xi, p) c(t, x, \zeta, q) \lesssim_N \frac{1}{(\lambda|\xi - \zeta| + \mu|p - q|)^N}.$$

When  $\lambda|\xi - \zeta| + \mu|p - q| \leq 1$ , we do not perform any integration by parts and estimate the  $K$  integrand by  $O(1)$  and hence  $K$  by  $O(1)$  as well. Otherwise we use the above decay. As a result, we get

$$K(\xi, \xi_2, \zeta, \zeta_2) \lesssim_N \frac{1}{(1 + \lambda|\xi - \zeta| + \mu|p - q|)^N},$$

which finishes the proof. □

**Remark.** It is not hard to see that the estimate (1-5) is sharp. In fact, by considering the restriction case and taking  $\phi(t, x, \xi) = \psi(t, x, \xi) = x \cdot \xi + t|\xi|^2$  with  $a$  having its  $\xi$  support in the region  $|\xi| \geq 100$  and  $b$  having its  $\xi$  support near  $|\xi| \leq 1$ , one can reduce the sharpness of (1-5) to that of (1-10) which is known to be sharp. In fact, this can be seen by first reducing to the case when  $N_2 = 1$  (again using scaling) and taking  $\widehat{u}_0$  to be the characteristic function of  $[N_1, N_1 + N_1^{-1}] \times [-1, 1]^{d-1}$  (hence  $\|u_0\|_{L_x^2} \sim N_1^{-1/2}$ ); and  $\widehat{v}_0$  to be the characteristic function of  $[-1, 1]^d$  (hence  $\|v_0\|_{L_x^2} \sim 1$ ). By Plancherel's

theorem in space and time, we get that the left side of (1-10) is  $\gtrsim \|\chi_{R_1} * \chi_{R_2}\|_{L^2(\mathbb{R}^{d+1})}$  where  $R_1 = [N_1, N_1 + N_1^{-1}] \times [0, 1]^d$  and  $R_2 = [-1, 1]^{d+1}$ . A direct calculation now shows that  $\chi_{R_1} * \chi_{R_2} \gtrsim (1/N_1)\chi_{R_3}$  where  $R_3 = [N_1 + \frac{1}{4}, N_1 + \frac{3}{4}] \times [-\frac{1}{2}, \frac{1}{2}]^d$  and hence  $\|\chi_{R_1} * \chi_{R_2}\|_{L^2(\mathbb{R}^{d+1})} \sim 1/N_1$ , which shows that the left side of (1-10) is  $\gtrsim (1/N_1^{1/2})\|u_0\|_{L_x^2}\|v_0\|_{L_x^2}$ .

### 3. Bilinear Strichartz estimates

We will apply the result of the previous section to get bilinear Strichartz estimates for the free Schrödinger evolution on compact manifolds without boundary. These will be analogues in the variable coefficient case to the estimate (1-10) on  $\mathbb{R}^d$  with the Euclidean Laplacian which we recall here for convenience

$$\|e^{it\Delta}u_0e^{it\Delta}v_0\|_{L^2(\mathbb{R}\times\mathbb{R}^d)} \lesssim \frac{N_2^{(d-1)/2}}{N_1^{1/2}}\|u\|_{L^2(\mathbb{R}^d)}\|v\|_{L^2(\mathbb{R}^d)},$$

where  $u, v \in L^2(\mathbb{R}^d)$  are frequency localized on the dyadic annuli  $\{\xi \in \mathbb{R}^d : |\xi| \in [N_1, 2N_1]\}$  and  $\{\xi \in \mathbb{R}^d : |\xi| \in [N_2, 2N_2]\}$  respectively.

By scaling time and space, one can easily see that this estimate is equivalent to the same one on the time interval  $[0, 1/N_1]$ . On this time scale, the numerology in (1-10) can be understood (heuristically at least) by a simple back-of-the-envelope calculation. Thinking of  $e^{it\Delta}u_0$  as a “bump function” localized in frequency at scale  $N_1$  and *initially* (at  $t = 0$ ) localized in space at scale  $1/N_1$ . The evolution moves this bump function at a speed  $N_1$  thus expanding its support at this rate while keeping the  $L^2$  norm conserved. Similarly,  $e^{it\Delta}v_0$  could be thought of as a “bump function” that is initially concentrated in space at scale  $\sim 1/N_2$  and moving (expanding) at speed  $N_2$ . A simple schematic diagram allows to estimate the space-time overlap of the two expanding “bump functions” thus giving the estimate  $N_2^{(d-1)/2}/N_1^{1/2}$  for the  $L^2_{t,x}([0, N_1^{-1}] \times \mathbb{R}^d)$  of the product.

The goal of this section is to prove the analogue of (1-10) for the linear evolution of the Schrödinger equation on a  $C^\infty$  compact manifold  $M$  without boundary. This was stated in Theorem 1.2. All implicit constants are allowed to depend on  $M$  and the uniform bounds of its metric functions (they are all finite since  $M$  is compact). To fix notation, we consider two functions  $u_0, v_0 \in C^\infty(M)^4$  such that  $u_0 = \varphi(\sqrt{-\Delta}/N_1)u_0$  and  $v_0 = \varphi(\sqrt{-\Delta}/N_2)v_0$  where  $\varphi \in C_0^\infty(\mathbb{R})$ , and we would like to estimate the  $L^2_{t,x}$  norm of the product  $e^{it\Delta}u_0e^{it\Delta}v_0$ . We assume further that  $\varphi$  vanishes in a small neighborhood of the origin.

**Remark.** The same analysis allows to consider different frequency localizations for  $u_0$  and  $v_0$  like  $u_0 = \varphi(\sqrt{-\Delta}/N_1)u_0$  and  $v_0 = \psi(\sqrt{-\Delta}/N_2)v_0$  with  $\varphi, \psi \in C_0^\infty$  as long as  $\varphi$  vanishes in a neighborhood of the origin and  $N_1$  is sufficiently larger than  $N_2$ . In particular,  $\psi$  does not need to vanish near the origin.

To simplify notation, we use  $\Delta$  to denote the Laplace–Beltrami operator  $\Delta_g$  on  $M$ , and  $|\xi|_{g(x)}$  to denote  $\sqrt{g(x)^{ij}\xi_i\xi_j}$ .

*Proof of Theorem 1.2.* The proof is organized as follows. We will first review some important facts about microlocalizing  $\varphi(h\sqrt{-\Delta})$  and constructing the Schrödinger parametrix (as in [Burq et al. 2004]) that will

<sup>4</sup>The full result for  $u_0, v_0 \in L^2(M)$  can be obtained in the end by a standard limiting argument.

be used to approximate the linear evolutions. The case when  $N_2 \sim N_1$ , will then follow directly from the semiclassical linear Strichartz estimates already proven in [Burq et al. 2004, Proposition 2.9]. As a result, we will only need to consider the case when  $N_2 \ll N_1$ . This will ensure that the canonical hypersurfaces associated to the phase functions of the parametrices are transversal as defined in the previous section, a fact which will allow us to apply Theorem 1.1.  $\square$

**Microlocalizing  $\varphi(h\sqrt{-\Delta})$**  [Burq et al. 2004; Sogge 1993; Hörmander 1994a; 1994b]. In this section, we will briefly review how spectrally localizing a function  $f \in C^\infty(M)$  using the spectral multiplier  $\varphi(h\sqrt{-\Delta})$  is expressed in local coordinates. Essentially, up to smooth remainder terms,  $\varphi(h\sqrt{-\Delta})f$  is given in local coordinates as a pseudodifferential operator whose symbol  $a(x, \xi)$  has a support that reflects the spectral localization dictated by  $\varphi$ :

**Proposition 3.1.** *Let  $\varphi \in C_0^\infty(\mathbb{R})$  and  $\kappa : U \subset \mathbb{R}^d \rightarrow V \subset M$  be a coordinate parametrization of  $M$ . Also let  $\chi_1, \chi_2 \in C_0^\infty(V)$  be such that  $\chi_2 = 1$  near the support of  $\chi_1$ . Then for every  $N \in \mathbb{N}$ , every  $h \in (0, 1)$ , and every  $\sigma \in [0, N]$ , there exists  $a_N(x, \xi)$  supported in  $\{(x, \xi) \in U \times \mathbb{R}^d : \kappa(x) \in \text{supp}(\chi_1), |\xi|_{g(x)} \in \text{supp}(\varphi)\}$  such that*

$$\|\kappa^*(\chi_1\varphi(h\sqrt{-\Delta})f) - a(x, hD)\kappa^*(\chi_2f)\|_{H^\sigma(\mathbb{R}^d)} \lesssim_N h^{N-\sigma} \|f\|_{L^2(M)} \tag{3-1}$$

for every  $f \in C^\infty(M)$ . In particular, if  $\varphi$  is supported away from the origin, then so is the  $\xi$  support of  $a(x, \xi)$ . Here  $\kappa^*$  is used to denote the pull-back map given by  $\kappa^*f = f \circ \kappa$ .

*Proof.* See Proposition 2.1 of [Burq et al. 2004] (alternatively, one can use the parametrix expression of the half-wave operator  $e^{it\sqrt{-\Delta}}$  (see [Sogge 1993] for example), along with the expression of  $\varphi$  in terms of its Fourier transform).

A consequence of this proposition and a finite partition of unity in  $M$ , one can split  $u_0 = \varphi(h\sqrt{-\Delta})u_0$  into pieces of the form  $\chi_1\varphi(h\sqrt{-\Delta})u_0$  and replace each of those pieces (incurring an error that is  $O(h^N \|u_0\|_{L^2})$ ) by  $a(x, hD)\kappa^*(\chi_2u_0)$  which is a compactly supported function in space and is pseudolocalized in frequency in the following sense:

There exists a function  $\psi \in C_0^\infty(\mathbb{R}^d)$  such that for all  $h \in (0, 1)$ ,  $\sigma > 0$ , and  $N > 0$ ,

$$\kappa^*(\chi_1\varphi(h\sqrt{-\Delta})f) = \psi(hD)\kappa^*(\chi_1\varphi(h\sqrt{-\Delta})f) + r_1, \tag{3-2}$$

with  $\|r_1\|_{H^\sigma(\mathbb{R}^d)} \lesssim_{\sigma, N} h^N \|f\|_{L^2}$ . If  $\varphi$  is supported away from 0, one can also take  $\psi$  to be supported at a positive distance from the origin in  $\mathbb{R}^d$ . This follows easily from Proposition 3.1 and standard pseudodifferential calculus (See [Stein 1993], for example). We will denote  $w_0(x) = a(x, hD)\kappa^*(\chi_2u_0)$ . In brief,  $w_0$  is compactly supported in space and can be replaced by  $\psi(hD)w_0$  at the cost of an error that is  $O(h^N \|u_0\|_{L^2(M)})$ .

**The parametrix** [Burq et al. 2004]. With this microlocalization setup, Burq, Gerard, and Tzvetkov constructed an approximate solution in local coordinates to the semiclassical equation

$$ih\partial_t w + h^2\Delta_g w = 0, \tag{3-3}$$

$$w(0) = \varphi(h\sqrt{-\Delta})v_0. \tag{3-4}$$

More precisely, using the usual WKB construction (see for example [Hörmander 1994a; 1994b; Burq et al. 2004], or the lecture notes [Evans and Zworski 2003]), they show that there exists  $\alpha > 0$ , such that on the time interval  $[-\alpha, \alpha]$

$$w(s) = \tilde{w}(s) + r_2(s),$$

where  $r_2(s)$  satisfies  $\|r_2(t)\|_{L_t^\infty([-\alpha, \alpha] \times H^\sigma(M))} \lesssim h^N \|w_0\|_{L^2(M)}$  (with  $N$  sufficiently large) and  $\tilde{w}(t)$  is supported in a compact subset of  $V \subset M$  and is given in local coordinates by the oscillatory integral

$$\tilde{w}(s, x) = \frac{1}{(2\pi h)^d} \int_{\mathbb{R}^d} e^{(i/h)\tilde{\phi}(s, x, \xi)} a(s, x, \xi, h) \widehat{w}_0\left(\frac{\xi}{h}\right) d\xi. \tag{3-5}$$

Here  $a(s, x, \xi, h) = \sum_{j=0}^N h^j a_j(s, x, \xi)$ , and  $a_j \in C_0^\infty([-\alpha, \alpha] \times U \times U' \subseteq \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d)$ , while  $w_0$  is the microlocalization of  $\varphi(h\sqrt{-\Delta})v_0$  described above. Since  $w_0$  can be replaced by  $\psi(hD)w_0$  at the cost of an error that is  $O(h^N \|w_0\|_{L^2(\mathbb{R}^d)})$  one can assume without loss of generality that  $a(s, x, \xi, h)$  has its  $\xi$  support at a positive distance from the origin in frequency space if  $\varphi$  is supported away from 0 itself.

The phase function  $\tilde{\phi}$  appearing in the integral (3-5) satisfies the eikonal equation

$$\partial_s \tilde{\phi} + \sum_{ij} g^{ij} \partial_i \tilde{\phi} \partial_j \tilde{\phi} = 0, \tag{3-6}$$

$$\tilde{\phi}(0, x, \xi) = x \cdot \xi. \tag{3-7}$$

**Semiclassical linear Strichartz estimates and the case  $N_1 \sim N_2$ .** Using this representation, one can easily use stationary phase (see [Burq et al. 2004] for details) to get the semiclassical dispersion estimate

$$\|e^{it\Delta} \varphi^2(h\sqrt{-\Delta})v_0\|_{L^\infty(M)} \lesssim_M \frac{1}{t^{d/2}} \|v_0\|_{L^1(M)} \tag{3-8}$$

for every  $t \in [-\alpha h, \alpha h]$  with  $0 < \alpha \ll 1$ . Combining this with the Keel–Tao machinery [1998] one immediately gets the semiclassical Strichartz estimate

$$\|e^{it\Delta} \varphi(h\sqrt{-\Delta})u_0\|_{L_t^q L_x^r([-\alpha h, \alpha h] \times M)} \lesssim_M \|u_0\|_{L^2(M)} \tag{3-9}$$

whenever  $2 \leq q, r \leq \infty$  satisfy  $2/q + d/r = d/2$  and  $(q, r, d) \neq (2, \infty, 2)$ .

This estimate is enough to prove (1-11) in the case when  $h = 1/N_1 \sim m = 1/N_2$ . In fact, for  $d = 2$ , one can use the  $L_{t,x}^4$  Strichartz estimate to get

$$\|e^{it\Delta} u_0 e^{it\Delta} v_0\|_{L_{t,x}^2([-\alpha h, \alpha h] \times M^2)} \leq \|e^{it\Delta} \varphi(h\sqrt{-\Delta})u_0\|_{L_{t,x}^4} \|e^{it\Delta} \varphi(h\sqrt{-\Delta})v\|_{L_{t,x}^4} \lesssim \|u_0\|_{L^2(M^2)} \|v_0\|_{L^2(M^2)}.$$

Whereas for  $d \geq 3$ , one can apply Hölder’s inequality, the  $L_t^\infty L_x^2$  bound on  $e^{it\Delta} u_0$ , Bernstein<sup>5</sup> and the  $L_t^2 L_x^{2d/(d-2)}$  for  $e^{it\Delta} v_0$  to get

$$\|e^{it\Delta} u_0 e^{it\Delta} v_0\|_{L_{t,x}^2([0, \alpha h] \times M)} \lesssim N_2^{(d-2)/2} \|u_0\|_{L^2(M)} \|v_0\|_{L^2(M)}$$

as desired.

<sup>5</sup>One can verify Bernstein’s inequality in the setting of compact manifolds by using Proposition 3.2 and the fact that the kernel  $K(x, y)$  of  $a(x, hD)$  satisfies the bound  $\|K(x, y)\|_{L_x^r L_y^p(\mathbb{R}^d \times \mathbb{R}^d)} \lesssim_a h^{-d(1-1/r-1/p)}$ .

**The case  $N_1 \gg N_2$ .** In this section, we will reduce the case  $N_1 \gg N_2$  to a verification of the conditions of (1-5). By rescaling time, we have

$$\begin{aligned} \|e^{it\Delta}u_0e^{it\Delta}v_0\|_{L^2_{t,x}([- \alpha h, \alpha h] \times M)} &= h^{1/2} \|e^{iht\Delta}u_0e^{iht\Delta}v_0\|_{L^2_{t,x}([- \alpha, \alpha] \times M)} \\ &= h^{1/2} \|e^{iht\Delta}u_0e^{im(h/m)t\Delta}v_0\|_{L^2_{t,x}([- \alpha, \alpha] \times M)}. \end{aligned} \tag{3-10}$$

As a result it is enough to show

$$\|e^{iht\Delta}u_0e^{im(\frac{h}{m}t)\Delta}v_0\|_{L^2_{t,x}([- \alpha, \alpha] \times M)} \lesssim \frac{1}{m^{(d-1)/2}} \|u_0\|_{L^2(M)} \|v_0\|_{L^2(M)}. \tag{3-11}$$

The advantage of writing the estimate in this way is that we can now use the parametrices for  $e^{ith\Delta}u_0$  and  $e^{itm\Delta}v_0$  constructed above to write<sup>6</sup>

$$e^{ith\Delta}u_0(x) = \tilde{T}_h u_0(t, x) + R_h u_0(t, x)$$

and

$$e^{im(ht/m)\Delta}v_0(x) = \tilde{S}_m v_0(t, x) + R_m v_0(t, x),$$

where  $\tilde{T}_h$  and  $\tilde{S}_m$  are defined according to (3-5) by

$$\tilde{T}_h u_0(t, x) = \frac{1}{(2\pi h)^d} \int_{\mathbb{R}^d} e^{(i/h)\tilde{\phi}(t,x,\xi)} a_1(t, x, \xi, h) \widehat{u}_0\left(\frac{\xi}{h}\right) d\xi \tag{3-12}$$

and

$$\tilde{S}_m v_0(t, x) = \frac{1}{(2\pi m)^d} \int_{\mathbb{R}^d} e^{(i/m)\tilde{\phi}(ht/m,x,\xi_2)} a_2\left(\frac{h}{m}t, x, \xi_2, m\right) \widehat{v}_0\left(\frac{\xi_2}{m}\right) d\xi_2, \tag{3-13}$$

where  $\tilde{u}_0$  and  $\tilde{v}_0$  are the respective microlocalizations of  $u_0$  and  $v_0$  in the considered coordinate patch (in particular  $\|\tilde{u}_0\|_{L^2(\mathbb{R}^d)} \lesssim \|u_0\|_{L^2(M)}$  and  $\|\tilde{v}_0\|_{L^2(M)} \lesssim \|v_0\|_{L^2(M)}$ ). Also we have

$$\|R_h u_0\|_{L^\infty H^\sigma([- \alpha, \alpha] \times M)} \lesssim h^N \|u_0\|_{L^2(M)} \quad \text{and} \quad \|R_m v_0\|_{L^\infty H^\sigma([- \alpha, \alpha] \times M)} \lesssim m^N \|v_0\|_{L^2(M)}. \tag{3-14}$$

The main contribution comes of course from the product  $\tilde{T}_h u_0 \tilde{S}_m v_0$ . For example the cross terms  $\tilde{T}_h u_0 R_m v_0$  and  $R_h u_0 \tilde{S}_m v_0$  can be bounded as follows:

$$\|\tilde{T}_h u_0 R_m v_0\|_{L^2_{t,x}} \leq \|\tilde{T}_h u_0\|_{L^\infty L^2_x} \|R_m v_0\|_{L^2_t L^\infty_x} \lesssim \|u_0\|_{L^2} \|v_0\|_{L^2},$$

where in the last step we used (3-14) and a crude Sobolev embedding to bound  $\|R_m v_0\|_{L^2_t L^\infty_x}$  by  $\|R_m\|_{L^2_t H^\sigma_x}$  for some  $\sigma > d/2$ . The  $L^\infty L^2_x$  bound on  $\tilde{T}_h u_0$  follows from the  $L^\infty L^2_x$  boundedness of  $e^{ith\Delta}u_0$ . Similarly, one bounds the contributions of  $R_h u_0 \tilde{S}_m v_0$  and  $R_h u_0 R_m v_0$ .

To bound the contribution of  $\tilde{T}_h u_0 \tilde{S}_m v_0$ , we now apply Theorem 1.1 with  $\phi(t, x, \xi) = \tilde{\phi}(t, x, \xi)$  and  $\psi(t, x, \xi_2) = \tilde{\phi}((h/m)t, x, \xi_2)$ ,  $f(\xi) := \tilde{u}(\xi/h)$ , and  $g(\xi) = \tilde{v}_0(\xi/m)$ , to get

$$\|\tilde{T}_h u_0 \tilde{S}_m v_0\|_{L^2_{t,x}([- \alpha, \alpha] \times \mathbb{R}^d)} \lesssim \frac{1}{(hm)^d} (h^d m)^{1/2} \|f\|_{L^2(\mathbb{R}^d)} \|g\|_{L^2(\mathbb{R}^d)} \lesssim \frac{1}{m^{(d-1)/2}} \|\tilde{u}_0\|_{L^2(\mathbb{R}^d)} \|\tilde{v}_0\|_{L^2(\mathbb{R}^d)},$$

---

<sup>6</sup>Strictly speaking this representation only holds in an open neighborhood of  $x_0 \in M$ . Since  $M$  is compact, we can cover it by finitely many of such neighborhood, and hence we only need to prove the estimate on each one of them.

which clearly gives (3-11) and hence (1-11). As a result, all we need to do is to verify that the requirements of Theorem 1.1 are satisfied.

Obviously all derivatives of  $\phi$  and  $\psi$  are uniformly bounded on the compact supports of  $a_1$  and  $a_2$  ( $h/m \leq 1$ ). Moreover, since  $\tilde{\phi}(0, x, \xi) = x \cdot \xi$ , we have that  $(\partial^2 \phi / \partial \xi \partial x)(0, x, \xi) = \text{Id}$  (invertible), the nondegeneracy condition (1-2) is satisfied at  $t = 0$  and hence for all  $t \in [-\alpha, \alpha]$  if  $\alpha$  is small enough.

Now we consider the canonical surfaces  $S_\phi$  and  $S_\psi$ :

Recall that  $S_\phi$  and  $S_\psi$  are the images of the maps

$$\begin{aligned} \xi_1 &\mapsto \nabla_{t,x} \phi(t, x, \xi_1) = (\nabla_x \tilde{\phi}(t, x, \xi_1), \partial_t \tilde{\phi}(t, x, \xi_1)), \\ \xi_2 &\mapsto \nabla_{t,x} \psi(t, x, \xi_2) = \left( \nabla_x \tilde{\phi}\left(\frac{h}{m}t, x, \xi_2\right), \frac{h}{m} \partial_t \tilde{\phi}\left(\frac{h}{m}t, x, \xi_2\right) \right), \end{aligned}$$

respectively. By the nondegeneracy condition above,  $S_\phi$  and  $S_\psi$  are smooth embedded hypersurfaces in  $T_{(t,x)}^* \mathbb{R}^{d+1}$ . We need to show that if  $\nu_1(\xi_1)$  is the normal to  $S_\phi$  at  $\nabla_{t,x} \phi(t, x, \xi_1)$  and  $\nu(\xi_2)$  is the normal to  $S_\psi$  at  $\nabla_{t,x} \psi(t, x, \xi_2)$ , then there is a  $\delta > 0$  (uniform in  $\xi_1$  and  $\xi_2$ ) such that

$$|\langle \nu_1, \nu_2 \rangle| \leq 1 - \delta. \tag{3-15}$$

By continuity, we only need to verify (3-15) at  $t = 0$  for all  $x, \xi_1, \xi_2$ . This will imply that the same holds for all  $t \in [-\alpha, \alpha]$  if  $\alpha$  is small enough. We now fix  $(0, x_0) \in \mathbb{R}^{d+1}$  and consider the surfaces  $S_\phi$  and  $S_\psi$  in  $T_{(0,x_0)}^* \mathbb{R}^{d+1}$ . From the eikonal equation (3-6),  $\tilde{\phi}(0, x, \xi) = x \cdot \xi$  and  $\partial_t \tilde{\phi}(0, x, \xi) = g^{ij}(x) \xi_i \xi_j$ . A straightforward computation gives

$$\nu_1(\xi) = \frac{(2g^{1j} \xi_j, 2g^{2j} \xi_j, \dots, 2g^{dj} \xi_j, -1)}{\sqrt{1+4|\xi|_{g(x)}^2}}$$

and

$$\nu_2(\xi) = \frac{(2(h/m)g^{1j} \xi_j, 2(h/m)g^{2j} \xi_j, \dots, 2(h/m)g^{dj} \xi_j, -1)}{\sqrt{1+4|(h/m)\xi|_{g(x)}^2}},$$

where we recall our notation that  $|\xi|_{g(x)} = \sqrt{g(x)^{ij} \xi_i \xi_j}$ . As a result,

$$\langle \nu_1(\xi_1), \nu_2(\xi_2) \rangle = \frac{1}{\sqrt{1+4|\xi_1|_{g(x)}^2} \sqrt{1+4|(h/m)\xi_2|_{g(x)}^2}} + O\left(\frac{h}{m}\right).$$

Since  $|\xi_1| \gtrsim 1$  and  $|\xi_2| \lesssim 1$ ,<sup>7</sup> we get that (3-15) holds true if  $h/m$  is small enough.

The proof of (1-12) follows by splitting the time interval  $[0, T]$  into pieces of length  $N_1^{-1}$ . That of (1-14) follows by setting  $T = 1$  in (1-14) when  $N_1 \geq 1$  and by using the  $L_t^\infty L_x^2$  estimates and Sobolev's inequality if  $N_1 \leq 1$ . □

**Remark.** If  $P(D)$  is a differential operator on  $M$  of degree  $n$ , then  $P(D)e^{iht\Delta}u_0$  has the expression

$$P(D)e^{iht\Delta}u_0(x) = h^{-n} \tilde{T}'_h u_0(t, x) + R'_h u_0(t, x),$$

<sup>7</sup>Without loss of generality, we can assume that  $\|g^{ij} - \delta^{ij}\| \leq \text{frac}1C$  for some large enough  $C$  on the coordinate patch considered. This is enough to have  $|\xi|_{g(x)} \sim |\xi|$ .

where  $\tilde{T}'_h$  and  $R'_h$  are operators of the same form as  $T_h$  and  $R_h$ . In particular,  $T'_h$  has an expression as in (3-12) (just with different  $a$ ) and  $R'_h$  obeys similar estimates to (3-14) (by choosing  $h$  small enough). Similar expressions for  $e^{imt\Delta}v_0$  allow us, using the exact same analysis performed above, to get:

**Corollary 3.2.** *Suppose the  $u_0, v_0 \in L^2(M)$  are spectrally localized around  $N_1, N_2 \in 2^{\mathbb{Z}}$  respectively as in Corollary 1.3. Let  $P(D)$  and  $Q(D)$  be differential operators on  $M$  of orders  $n$  and  $m$  respectively:*

$$\|P(D)e^{it\Delta}u_0Q(D)e^{it\Delta}v_0\|_{L^2([0,T] \times M)} \leq N_1^n N_2^m \Lambda(T, N_1, N_2) \|u_0\|_{L^2(M)} \|v_0\|_{L^2(M)}, \tag{3-16}$$

where  $\Lambda(T, N_1, N_2)$  is given in (1-13).

This variant will be useful in some applications of the bilinear Strichartz estimates proved here (see [Hani 2012] for example).

#### 4. Further results and remarks

**Bilinear inhomogeneous estimates.** Here we will present some inhomogeneous versions of the bilinear estimates proved in the previous section. We will assume that  $u(t)$  and  $v(t)$  solve the inhomogeneous Schrödinger equation with forcing terms  $F$  and  $G$  respectively. More precisely,

$$i\partial_t u + \Delta u = F, \tag{4-1}$$

$$i\partial_t v + \Delta v = G. \tag{4-2}$$

$F$  and  $G$  can be assumed to be a priori in  $C^\infty$ .<sup>8</sup> The question now is to determine estimates for  $\|uv\|_{L^2_{t,x}}$  in terms of the initial data  $u(0) = u_0, v(0) = v_0$  and the forcing terms  $F$  and  $G$ .

We will prove two types of inhomogeneous estimates: one corresponding to spectrally localized functions generalizing (1-11) and another is a time  $T = 1$  estimate generalizing (1-14).

**Theorem 4.1.** *Suppose  $u(t)$  and  $v(t)$  solve the inhomogeneous Schrödinger equations (4-1) and (4-2) with initial data  $u(0) = u_0$  and  $v(0) = v_0$  respectively. Also suppose that  $(q, r)$  and  $(\tilde{q}, \tilde{r})$  are two Schrödinger admissible exponents.*

(i) *If  $u(t) = \varphi(\sqrt{-\Delta}/N_1)u(t)$  and  $v(t) = \varphi(\sqrt{-\Delta}/N_2)v(t)$  for all  $t$ , then*

$$\|uv\|_{L^2_{t,x}([0,1/N_1] \times M)} \lesssim \frac{N_2^{(d-1)/2}}{N_1^{1/2}} (\|u_0\|_{L^2(M)} + \|F\|_{L^q_t L^r_x}) (\|v_0\|_{L^2(M)} + \|G\|_{L^{\tilde{q}}_t L^{\tilde{r}}_x}), \tag{4-3}$$

where for any  $p \in [1, \infty]$ ,  $p'$  denotes its conjugate exponent  $1/p + 1/p' = 1$ .

(ii) *In general, for any  $\delta > 0$  we have*

$$\|uv\|_{L^2_{t,x}([0,1] \times M)} \lesssim (\|u_0\|_{H^\delta(M)} + \|(\sqrt{1-\Delta})^{\delta+1/q} F\|_{L^q_t L^r_x}) (\|v_0\|_{H^{1/2-\delta}(M)} + \|(\sqrt{1-\Delta})^{1/2-\delta+1/\tilde{q}} G\|_{L^{\tilde{q}}_t L^{\tilde{r}}_x}). \tag{4-4}$$

For the proof, we will need the Christ–Kiselev lemma [2001], which we state following [Smith and Sogge 2000]:

<sup>8</sup>This assumption can be removed a posteriori using standard density arguments.



**Lemma 4.2.** *Let  $X$  and  $Y$  be Banach spaces and  $K(t, x)$  a continuous function taking values in  $B(X, Y)$ , the space of bounded linear mappings from  $X$  to  $Y$ . Suppose that  $-\infty \leq a < b \leq \infty$  and let*

$$Tf(t) = \int_a^b K(t, s) f(s) ds.$$

Suppose that

$$\|Tf\|_{L^q([a,b];Y)} \leq C \|f\|_{L^p([a,b];X)},$$

and define the lower triangular operator

$$Wf(t) = \int_a^t K(t, s) f(s) ds.$$

Then, if  $1 \leq p < q \leq \infty$ ,

$$\|Wf\|_{L^q([a,b];Y)} \lesssim C \|f\|_{L^p([a,b];X)}.$$

*Proof of Theorem 4.1.* We start by proving the spectrally localized version in (4-3). The integral equations satisfied by  $u(t)$  and  $v(t)$  are given by Duhamel’s formula:

$$u(t) = e^{it\Delta} u_0 - i \int_0^t e^{i(t-s)\Delta} F(s) ds \quad \text{and} \quad v(t) = e^{it\Delta} v_0 - i \int_0^t e^{i(t-s)\Delta} G(s) ds.$$

As a result,

$$\begin{aligned} u(t)v(t) &= e^{it\Delta} u_0 e^{it\Delta} v_0 - i e^{it\Delta} u_0 \int_0^t e^{i(t-s)\Delta} G(s) ds \\ &\quad - i e^{it\Delta} v_0 \int_0^t e^{i(t-s)\Delta} F(s) ds - \int_0^t e^{i(t-s)\Delta} F(s) ds \int_0^t e^{i(t-r)\Delta} G(r) dr. \end{aligned} \quad (4-5)$$

Recall that  $u_0, u(t), F(t)$  are all spectrally localized at dyadic scale  $N_1$  and  $v_0, v(t), G(t)$  localized at scale  $N_2$ . The estimate for the first term on the right in (4-5) is the bilinear Strichartz estimate proved in the previous section. We turn to the second term. Applying the Christ–Kiselev lemma (with  $Y = L_t^{\tilde{q}'} L_x^{\tilde{r}'}$ ,  $X = L_{t,x}^2([0, 1/N_1] \times M)$ , and  $C \sim N_2^{(d-1)/2} / N_1^{1/2} \|u_0\|_{L^2(M)}$ ), it is enough to show

$$\left\| e^{it\Delta} u_0 \int_0^{1/N_1} e^{i(t-s)\Delta} G(s) ds \right\|_{L_{t,x}^2([0, 1/N_1] \times M)} \lesssim \frac{N_2^{(d-1)/2}}{N_1^{1/2}} \|u_0\|_{L^2(M)} \|G\|_{L_t^{\tilde{q}'} L_x^{\tilde{r}'}}.$$

But this follows from the bilinear estimate (1-11) and

$$\left\| \int_0^{1/N_1} e^{-is\Delta} \varphi\left(\frac{\sqrt{-\Delta}}{N_1}\right) G(s) ds \right\|_{L_x^2(M)} \lesssim \|G\|_{L_t^{\tilde{q}'} L_x^{\tilde{r}'}} ,$$

which is the dual estimate to (1-9).

The third term on the right in (4-5) is estimated similarly. For the fourth term, we first apply the Christ–Kiselev lemma to reduce the estimate to

$$\begin{aligned} & \left\| \int_0^{1/N_1} e^{i(t-s)\Delta} F(s) ds \int_0^t e^{i(t-r)\Delta} G(r) dr \right\|_{L^2_{t,x}([0,1/N_1] \times M)} \\ &= \left\| e^{it\Delta} \left( \int_0^{1/N_1} e^{-is\Delta} F(s) ds \right) \int_0^t e^{i(t-r)\Delta} G(r) dr \right\|_{L^2_{t,x}} \\ &\lesssim \frac{N_2^{(d-1)/2}}{N_1} \left\| \int_0^{N_1^{-1}} e^{-is\Delta} F(s) ds \right\|_{L^2(M)} \|G\|_{L_t^{\tilde{q}'} L_x^{\tilde{r}'}} \lesssim \|F\|_{L_t^{q'} L_x^{r'}} \|G\|_{L_t^{\tilde{q}'} L_x^{\tilde{r}'}} \end{aligned}$$

where in the first inequality we apply the same analysis as that used to estimate the second and third term on the right in (4-5) (or apply Christ–Kiselev lemma again) while in the second we use the dual homogeneous Strichartz estimate. This finishes the proof of (4-3).

We now turn to the time 1 estimate (4-4). We start by mentioning that the first term on the right in (4-5) satisfies the needed estimate

$$\|e^{it\Delta} u_0 e^{it\Delta} v_0\|_{L^2([0,1] \times M)} \lesssim \|u_0\|_{H^\delta} \|v_0\|_{H^{1/2-\delta}}.$$

This follows directly by splitting into Littlewood–Paley pieces  $u = \sum_{\substack{N_1 \geq 1 \\ \text{(dyadic)}}} u_{N_1}$  and  $v = \sum_{\substack{N_2 \geq 1 \\ \text{(dyadic)}}} v_{N_2}$  and estimating by

$$\begin{aligned} & \|e^{it\Delta} u_0 e^{it\Delta} v_0\|_{L^2_{t,x}([0,1] \times M)} \\ &\leq \sum_{N_1 \leq N_2} \|e^{it\Delta} u_{N_1} e^{it\Delta} v_{N_2}\|_{L^2_{t,x}} + \sum_{N_1 > N_2} \|e^{it\Delta} u_{N_1} e^{it\Delta} v_{N_2}\|_{L^2_{t,x}} \\ &\lesssim \sum_{N_1 \leq N_2} N_1^{(d-1)/2} \|u_{N_1}\|_{L^2} \|v_{N_2}\|_{L^2} + \sum_{N_2 < N_1} N_2^{(d-1)/2} \|u_{N_1}\|_{L^2} \|v_{N_2}\|_{L^2} \\ &\lesssim \sum_{N_1 \leq N_2} \frac{N_1^{(d-1)/2-\delta}}{N_2^{(d-1)/2-\delta}} \|u_{N_1}\|_{H^\delta} \|v_{N_2}\|_{H^{(d-1)/2-\delta}} + \sum_{N_2 < N_1} \frac{N_2^\delta}{N_1^\delta} \|u_{N_1}\|_{H^\delta} \|v_{N_2}\|_{H^{(d-1)/2-\delta}} \\ &\lesssim \|u\|_{H^\delta} \|v\|_{H^{(d-1)/2-\delta}}, \end{aligned}$$

where we have used Schur’s test to sum in the last step. The rest of the proof of (4-4) follows as that of (4-3) above except that here we use the estimate dual to (1-8) given by

$$\left\| \int_0^1 e^{i(t-s)\Delta} F(s) ds \right\|_{L^2(M)} \lesssim \|(\sqrt{1-\Delta})^{1/q} F\|_{L_t^{q'} L_x^{r'}([0,1] \times M)}. \quad \square$$

**Bilinear estimates of mixed type.** Here we present an instance of a mixed-type bilinear estimate of Schrödinger-wave type that can be proved using Theorem 1.1. Constant coefficient versions of such estimates are often useful when studying coupled Schrödinger-wave systems such as the Zakharov system (see [Bejenaru et al. 2009] for instance). Theorem 4.3 below serves as an example of a variable coefficient

Schrödinger-wave bilinear estimates and has potential applications in studying Zakharov systems (or other Schrödinger-wave systems) on manifolds.

**Theorem 4.3.** *Suppose  $u_0, v_0 \in L^2(M^d)$  are spectrally localized at dyadic scales  $N_1$  and  $N_2$  as above with  $1 \ll N_1$ . Then the estimate*

$$\|e^{it\Delta} u_0 e^{\pm it|\nabla|} v_0\|_{L^2_{t,x}([-1/N_1, 1/N_1] \times M)} \lesssim_M \frac{\min(N_1, N_2)^{(d-1)/2}}{N_1^{1/2}} \|u_0\|_{L^2(M)} \|v\|_{L^2(M)} \quad (4-6)$$

holds. Of course, an estimate over the time interval  $[0, T]$  follows as well by splitting into pieces of length  $1/N_1$ .

*Proof.* We present the proof in the case of the forward half wave operator, the proof for the backwards operator being similar. As before, we use the parametrix for  $e^{it|\nabla|} v_0$  which is given, up to a smoothing remainder  $R_m v_0$ , by the oscillatory integral

$$S_m^W v_0 = \frac{1}{(2\pi m)^d} \int_{\mathbb{R}^d} e^{(i/m)\psi(t,x,\xi_2)} a(t, x, \xi_2) \widehat{v}_0\left(\frac{\xi_2}{m}\right) d\xi_2,$$

where  $\psi$  is a nondegenerate phase function (in particular  $\det(\partial^2/\partial\xi \partial x) \tilde{\psi} \neq 0$ ) and homogeneous in  $\xi_2$  of degree 1 and  $\tilde{v}_0$  is a microlocalization of  $v_0$  as explained in Section 3 (cf. [Hörmander 1994b, Chapter XXIX]). As before, we used the convention that  $h = 1/N_1$  and  $m = 1/N_2$ . As a result, we have

$$\|e^{it\Delta} u_0 e^{it|\nabla|} v_0\|_{L^2_{t,x}([- \alpha/N_1, \alpha/N_1] \times M)} = h^{1/2} \|e^{iht\Delta} u_0 e^{iht|\nabla|} v_0\|_{L^2_{t,x}([- \alpha, \alpha] \times M)}.$$

Ignoring the smooth remainder terms  $R_h$  and  $R_m$  (as they are inconsequential as in Section 3) we get that (4-6) follows from the estimate

$$\begin{aligned} \|\tilde{T}_h u_0(t, x) \tilde{S}_m^W v_0(ht, x)\|_{L^2_{t,x}([- \alpha, \alpha] \times \mathbb{R}^d)} &\lesssim \frac{1}{(hm)^{d/2}} \min(m, h)^{d/2} \max(m, h)^{1/2} \|\tilde{u}_0\|_{L^2(\mathbb{R}^d)} \|\tilde{v}_0\|_{L^2(\mathbb{R}^d)} \\ &= C \max(m, h)^{-(d-1)/2} \|\tilde{u}_0\|_{L^2(\mathbb{R}^d)} \|\tilde{v}_0\|_{L^2(\mathbb{R}^d)}. \end{aligned}$$

This inequality follows by applying Equation (1-5) with the nondegenerate phase functions  $\phi(t, x, \xi_1) = \tilde{\phi}(t, x, \xi_1)$  and  $\psi(t, x, \xi_2) = \tilde{\psi}(ht, x, \xi_2)$ . The transversality condition is directly verified as follows. The normal vectors to the two surfaces

$$\begin{aligned} S_\phi : \xi_1 &\mapsto \nabla_{t,x} \phi(t, x, \xi_1) = (\nabla_x \tilde{\phi}(t, x, \xi_1), \partial_t \tilde{\phi}(t, x, \xi_1)), \\ S_\psi : \xi_2 &\mapsto \nabla_{t,x} \psi(t, x, \xi_2) = (\nabla_x \tilde{\psi}(ht, x, \xi_2), h \partial_t \tilde{\psi}(ht, x, \xi_2)) \end{aligned}$$

can be written as  $v_1 = (\eta_1, \tau_1)$  and  $v_2 = (\eta_2, \tau_2)$  with  $\eta_1, \eta_2 \in \mathbb{R}^n$  and  $\tau_1, \tau_2 \in \mathbb{R}$ . The fact that  $\langle v_2, (\partial^2/\partial\xi \partial x) \psi \rangle = \vec{0}$  implies that  $\langle \eta_2, (\partial^2/\partial\xi \partial x) \tilde{\psi}(ht, x, \xi_2) \rangle + h \tau_2 \partial_t \partial_\xi \tilde{\psi}(ht, x, \xi_2) = \vec{0}$ , which implies that

$$\eta_2 = -h \tau_2 \left\langle \partial_t \partial_\xi \tilde{\psi}, \left[ \frac{\partial^2}{\partial \xi \partial x} \tilde{\psi} \right]^{-1} \right\rangle = O(h).$$

This gives that

$$\langle v_1, v_2 \rangle \leq |\tau_1 \tau_2| + O(h) \leq |\tau_1| + O(h).$$

As a result, the transversality condition (1-6) holds if  $h \ll 1$  (i.e.,  $N_1 \gg 1$ ) and  $|\tau_1| < 1$ , which is the case since  $\tau_1 = -1/\sqrt{1 + 4|\xi|_{g(x)}^2}$  and  $|\xi_1| \gtrsim 1$  (see end of the proof of Theorem 1.2).  $\square$

**Applications in PDE.** The bilinear estimate (1-14) directly implies local well-posedness for 2-dimensional cubic NLS

$$\begin{aligned} i\partial_t u + \Delta u &= |u|^2 u, \\ u(t = 0) &= u_0 \in H^s(M^2) \end{aligned} \tag{4-7}$$

in  $X^{s,b} \subset C_t H_x^s$  spaces for all  $s > 1/2$  and some  $b > \frac{1}{2}$ . It should be noted that local well-posedness of (4-7) in  $C_t H^s$  for  $s > \frac{1}{2}$  has already been proven in [Burq et al. 2004] using linear Strichartz estimates. Here  $X^{s,b}$  is the closure of  $C_0^\infty(\mathbb{R} \times M)$  in the norm

$$\|u\|_{X^{s,b}} = \left( \int_{\mathbb{R}} \sum_{\nu} \langle \tau + \nu \rangle^{2b} \langle \nu \rangle^s \|\widehat{\pi_{\nu} u}(\tau)\|_{L^2(M)}^2 d\tau \right)^{1/2},$$

where the sum runs over the distinct eigenvalues of the Laplacian and  $\pi_{\nu}$  is the projection onto the eigenspace corresponding to the eigenvalue  $\nu$ . It is worth remarking that (1-11) translates into the following estimate for functions  $u, v \in C_0^\infty(\mathbb{R} \times M)$  satisfying  $u(t) = \mathbf{1}_{[N_1, 2N_1]}(\sqrt{-\Delta})u(t)$  and  $v(t) = \mathbf{1}_{[N_2, 2N_2]}(\sqrt{-\Delta})v(t)$ :

$$\|uv\|_{L^2(\mathbb{R} \times M)} \lesssim \min(N_2, N_1)^{1/2} \|u\|_{X^{0,b}} \|v\|_{X^{0,b}} \tag{4-8}$$

for any  $b > \frac{1}{2}$  (cf. [Burq et al. 2005a; Hani 2012]). Using this and a standard dyadic decomposition one can prove the crucial cubic estimate that yields local well-posedness via Picard iteration (see [Burq et al. 2005a] for example).

One interesting application of Theorem 1.2 is that of proving global well-posedness of (4-7) for  $s < 1$ . As mentioned in the introduction, the bilinear Strichartz estimate (1-12) on the time interval  $[0, T]$  translates into a bilinear Strichartz estimate on the rescaled manifold  $\lambda M$  over the time interval  $[0, 1]$ . Here  $\lambda M$  can either be viewed as the Riemannian manifold  $(M, (1/\lambda^2)g)$  or by embedding  $M$  into some ambient space  $R^N$  and then applying a dilation by  $\lambda$  to get  $\lambda M$ . The relevant result was cited in the introduction in Corollary 1.3: if  $u_0, v_0 \in L^2(\lambda M)$  are spectrally localized around  $N_1$  and  $N_2$  respectively, with  $N_2 \leq N_1$ . Then

$$\begin{aligned} \|e^{it\Delta_{\lambda}} u_0 e^{it\Delta_{\lambda}} v_0\|_{L^2([0,1] \times \lambda M)} &\lesssim \Lambda(\lambda^{-2}, \lambda N_1, \lambda N_2) \|u_0\|_{L^2(\lambda M)} \|v_0\|_{L^2(\lambda M)} \\ &\lesssim \begin{cases} (N_2/N_1)^{1/2} \|u_0\|_{L^2(\lambda M)} \|v_0\|_{L^2(\lambda M)} & \text{if } \lambda \gg N_1, \\ (N_2/\lambda)^{1/2} \|u_0\|_{L^2(\lambda M)} \|v_0\|_{L^2(\lambda M)} & \text{if } \lambda \lesssim N_1. \end{cases} \end{aligned}$$

This estimate turns out to be crucial in [Hani 2012] where it is proved that (4-7) is globally well-posed for all  $s > \frac{2}{3}$ . This generalizes, *without any loss in regularity*, a similar result from [Bourgain 2004] (see also [De Silva et al. 2007]), where global well-posedness for  $s > \frac{2}{3}$  is proved for the torus  $\mathbb{T}^2$ . Global well-posedness for  $s \geq 1$  follows using conservation of energy and standard arguments. To go below the energy regularity  $s = 1$ , the I-method of Colliander, Keel, Staffilani, Takaoka, and Tao should be used

and most of the analysis is done on  $\lambda M$  rather than  $M$ . As a result, the factor of  $1/\lambda^{1/2}$  on the right side of (1-16) in the range  $\lambda \lesssim N_1$  becomes crucial to get the full regularity range of  $s > \frac{2}{3}$  (see [Hani 2012]).

### Acknowledgements

The author is deeply grateful to his advisor, Prof. Terence Tao, for his invaluable support, encouragement, and guidance. He also wishes to extend his immense gratitude to the referee for his careful review of the manuscript and his helpful comments and suggestions that considerably improved and clarified the exposition.

### References

- [Bejenaru et al. 2009] I. Bejenaru, S. Herr, J. Holmer, and D. Tataru, “On the 2D Zakharov system with  $L^2$ -Schrödinger data”, *Nonlinearity* **22**:5 (2009), 1063–1089. MR 2010f:35383 Zbl 1173.35651
- [Bourgain 1993] J. Bourgain, “Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations, I: Schrödinger equations”, *Geom. Funct. Anal.* **3**:2 (1993), 107–156. MR 95d:35160a Zbl 0787.35097
- [Bourgain 1998] J. Bourgain, “Refinements of Strichartz’ inequality and applications to 2D-NLS with critical nonlinearity”, *Internat. Math. Res. Notices* **1998**:5 (1998), 253–283. MR 99f:35184 Zbl 0917.35126
- [Bourgain 1999] J. Bourgain, *Global solutions of nonlinear Schrödinger equations*, American Mathematical Society Colloquium Publications **46**, American Mathematical Society, Providence, RI, 1999. MR 2000h:35147 Zbl 0933.35178
- [Bourgain 2004] J. Bourgain, “A remark on normal forms and the “ $I$ -method” for periodic NLS”, *J. Anal. Math.* **94** (2004), 125–157. MR 2006b:37155 Zbl 1084.35085
- [Burq et al. 2004] N. Burq, P. Gérard, and N. Tzvetkov, “Strichartz inequalities and the nonlinear Schrödinger equation on compact manifolds”, *Amer. J. Math.* **126**:3 (2004), 569–605. MR 2005h:58036 Zbl 1067.58027
- [Burq et al. 2005a] N. Burq, P. Gérard, and N. Tzvetkov, “Bilinear eigenfunction estimates and the nonlinear Schrödinger equation on surfaces”, *Invent. Math.* **159**:1 (2005), 187–223. MR 2005m:35275 Zbl 1092.35099
- [Burq et al. 2005b] N. Burq, P. Gérard, and N. Tzvetkov, “Multilinear eigenfunction estimates and global existence for the three dimensional nonlinear Schrödinger equations”, *Ann. Sci. École Norm. Sup. (4)* **38**:2 (2005), 255–301. MR 2006m:35337 Zbl 1116.35109
- [Christ and Kiselev 2001] M. Christ and A. Kiselev, “Maximal functions associated to filtrations”, *J. Funct. Anal.* **179**:2 (2001), 409–425. MR 2001i:47054 Zbl 0974.47025
- [De Silva et al. 2007] D. De Silva, N. Pavlović, G. Staffilani, and N. Tzirakis, “Global well-posedness for a periodic nonlinear Schrödinger equation in 1D and 2D”, *Discrete Contin. Dyn. Syst.* **19**:1 (2007), 37–65. MR 2008g:35191 Zbl 05236234
- [Evans and Zworski 2003] L. C. E. Evans and M. Zworski, “Lectures on semiclassical analysis”, University of California, Berkeley, CA, 2003, Available at <http://math.berkeley.edu/~evans/semiclassical.pdf>.
- [Hani 2012] Z. Hani, “Global well-posedness of the 2D–cubic nonlinear Schrödinger equation on compact manifolds without boundary”, *Comm. Partial Differential Equations* **37**:7 (2012), 1186–1236. arXiv 1008.2826
- [Hörmander 1973] L. Hörmander, “Oscillatory integrals and multipliers on  $FL^p$ ”, *Ark. Mat.* **11** (1973), 1–11. MR 49 #5674 Zbl 0254.42010
- [Hörmander 1994a] L. Hörmander, *The analysis of linear partial differential operators, III: Pseudo-differential operators*, Grundlehren der Mathematischen Wissenschaften **274**, Springer, Berlin, 1994. MR 95h:35255 Zbl 0601.35001
- [Hörmander 1994b] L. Hörmander, *The analysis of linear partial differential operators, IV: Fourier integral operators*, Grundlehren der Mathematischen Wissenschaften **275**, Springer, Berlin, 1994. MR 98f:35002 Zbl 0612.35001
- [Jiang 2011] J.-C. Jiang, “Bilinear Strichartz estimates for Schrödinger operators in two-dimensional compact manifolds with boundary and cubic NLS”, *Differential Integral Equations* **24**:1-2 (2011), 83–108. MR 2012a:35332 Zbl 05944786

- [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
- [Lee 2006] S. Lee, “Linear and bilinear estimates for oscillatory integral operators related to restriction to hypersurfaces”, *J. Funct. Anal.* **241**:1 (2006), 56–98. MR 2007g:42024 Zbl 1121.35151
- [Smith and Sogge 2000] H. F. Smith and C. D. Sogge, “Global Strichartz estimates for nontrapping perturbations of the Laplacian”, *Comm. Partial Differential Equations* **25**:11-12 (2000), 2171–2183. MR 2001j:35180 Zbl 0972.35014
- [Sogge 1993] C. D. Sogge, *Fourier integrals in classical analysis*, Cambridge Tracts in Mathematics **105**, Cambridge University Press, Cambridge, 1993. MR 94c:35178 Zbl 0783.35001
- [Staffilani and Tataru 2002] G. Staffilani and D. Tataru, “Strichartz estimates for a Schrödinger operator with nonsmooth coefficients”, *Comm. Partial Differential Equations* **27**:7-8 (2002), 1337–1372. MR 2003f:35248 Zbl 1010.35015
- [Stein 1993] E. M. Stein, *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals*, Princeton Mathematical Series **43**, Princeton University Press, Princeton, NJ, 1993. MR 95c:42002 Zbl 0821.42001
- [Tao 2001] T. Tao, “Endpoint bilinear restriction theorems for the cone, and some sharp null form estimates”, *Math. Z.* **238**:2 (2001), 215–268. MR 2003a:42010 Zbl 0992.42004
- [Tao 2003] T. Tao, “A sharp bilinear restriction estimate for paraboloids”, *Geom. Funct. Anal.* **13**:6 (2003), 1359–1384. MR 2004m:47111 Zbl 1068.42011
- [Tao 2004] T. Tao, “Some recent progress on the restriction conjecture”, pp. 217–243 in *Fourier analysis and convexity*, edited by L. Brandolini et al., Birkhäuser, Boston, 2004. MR 2005i:42015 Zbl 1083.42008 arXiv math/0311181
- [Tao 2006] T. Tao, *Nonlinear dispersive equations: local and global analysis*, CBMS Regional Conference Series in Mathematics **106**, American Mathematical Society, Providence, RI, 2006. MR 2008i:35211 Zbl 1106.35001
- [Wolff 2001] T. Wolff, “A sharp bilinear cone restriction estimate”, *Ann. of Math. (2)* **153**:3 (2001), 661–698. MR 2002j:42019 Zbl 1125.42302

Received 16 Aug 2010. Revised 13 Jan 2011. Accepted 13 Feb 2011.

ZAHER HANI: [zhani@math.ucla.edu](mailto:zhani@math.ucla.edu)

Mathematics Department, University of California, Los Angeles, 520 Portola Plaza, Math Sciences Building,  
Los Angeles, CA 90095, United States

<http://www.math.ucla.edu/~zhani>



## THE CAUCHY PROBLEM FOR THE BENJAMIN–ONO EQUATION IN $L^2$ REVISITED

LUC MOLINET AND DIDIER PILOD

Ionescu and Kenig proved that the Cauchy problem associated with the Benjamin–Ono equation is globally well posed in  $L^2(\mathbb{R})$ . In this paper we give a simpler proof of Ionescu and Kenig’s result, which moreover provides stronger uniqueness results. In particular, we prove unconditional well-posedness in  $H^s(\mathbb{R})$  for  $s > \frac{1}{4}$ . Note that our approach also permits us to simplify the proof of the global well-posedness in  $L^2(\mathbb{T})$  and yields unconditional well-posedness in  $H^{\frac{1}{2}}(\mathbb{T})$ .

### 1. Introduction

The Benjamin–Ono equation is one of the fundamental equations describing the evolution of weakly nonlinear internal long waves. It has been derived by Benjamin [1967] as an approximate model for long-crested unidirectional waves at the interface of a two-layer system of incompressible inviscid fluids, one being infinitely deep. In nondimensional variables, the initial value problem (IVP) associated with the Benjamin–Ono equation (BO) is

$$\begin{cases} \partial_t u + \mathcal{H} \partial_x^2 u = u \partial_x u, \\ u(x, 0) = u_0(x), \end{cases} \quad (1-1)$$

where  $x \in \mathbb{R}$  or  $\mathbb{T}$ ,  $t \in \mathbb{R}$ ,  $u$  is a real-valued function, and  $\mathcal{H}$  is the Hilbert transform, defined on the line by

$$\mathcal{H}f(x) = \text{p.v.} \frac{1}{\pi} \int_{\mathbb{R}} \frac{f(y)}{x-y} dy. \quad (1-2)$$

The Benjamin–Ono equation is, at least formally, completely integrable [Fokas and Ablowitz 1983] and thus possesses an infinite number of conservation laws. For example, the momentum and the energy, respectively given by

$$M(u) = \int u^2 dx \quad \text{and} \quad E(u) = \frac{1}{2} \int |D_x^{\frac{1}{2}} u|^2 dx + \frac{1}{6} \int u^3 dx, \quad (1-3)$$

are conserved by the flow of (1-1).

The IVP associated with the Benjamin–Ono equation presents interesting mathematical difficulties and has been extensively studied in recent years. In the continuous case, well-posedness in  $H^s(\mathbb{R})$  for  $s > \frac{3}{2}$  was proved by Iório [1986] by using purely hyperbolic energy methods (see also [Abdelouhab et al. 1989] for global well-posedness in the same range of  $s$ ). Then Ponce [1991] derived a local smoothing effect associated with the dispersive part of the equation, which, combined with compactness methods, enables

---

MSC2010: primary 35A07, 35Q53; secondary 76B55.

Keywords: Benjamin–Ono equation, initial value problem, gauge transformation.



us to reach  $s = \frac{3}{2}$ . This technique was refined by Koch and Tzvetkov [2003] and Kenig and Koenig [2003], who reached  $s > \frac{5}{4}$  and  $s > \frac{9}{8}$ , respectively. On the other hand, Molinet, Saut, and Tzvetkov [Molinet et al. 2001] proved that the flow map associated with BO, when it exists, fails to be  $C^2$  in any Sobolev space  $H^s(\mathbb{R})$ ,  $s \in \mathbb{R}$ . This result is based on the fact that the dispersive smoothing effects of the linear part of BO are not strong enough to control the low-high frequency interactions appearing in the nonlinearity of (1-1). It was improved by Koch and Tzvetkov [2005], who showed that the flow map fails even to be uniformly continuous in  $H^s(\mathbb{R})$  for  $s > 0$  (see [Biagioni and Linares 2001] for the same result in the case  $s < -\frac{1}{2}$ ). As the consequence of those results, one cannot solve the Cauchy problem for the Benjamin–Ono equation by a Picard iterative method implemented on the integral equation associated with (1-1) for initial data in the Sobolev space  $H^s(\mathbb{R})$ ,  $s \in \mathbb{R}$ . In particular, the methods introduced by Bourgain [1993b] and Kenig, Ponce, and Vega [Kenig et al. 1993; 1996] for the Korteweg–de Vries equation do not apply directly to the Benjamin–Ono equation.

Therefore, the problem of obtaining well-posedness in less regular Sobolev spaces turns out to be far from trivial. Due to the conservations laws (1-3),  $L^2(\mathbb{R})$  and  $H^{\frac{1}{2}}(\mathbb{R})$  are two natural spaces where well-posedness is expected. In this direction, a decisive breakthrough was achieved by Tao [2004]. By combining a complex variant of the Cole–Hopf transform (which linearizes the Burgers equation) with Strichartz estimates, he proved well-posedness in  $H^1(\mathbb{R})$ . More precisely, to obtain estimates at the  $H^1$ -level, he introduced the new unknown

$$w = \partial_x P_{+hi}(e^{-\frac{i}{2}F}), \quad (1-4)$$

where  $F$  is some spatial primitive of  $u$  and  $P_{+hi}$  denotes the projection on high positive frequencies. Then  $w$  satisfies an equation of the form

$$\partial_t w - i \partial_x^2 w = -\partial_x P_{+hi}(\partial_x^{-1} w P_- \partial_x u) + \text{negligible terms}. \quad (1-5)$$

Observe that, thanks to the frequency projections, the nonlinear term appearing in the right-hand side of (1-5) does not exhibit any low-high frequency interaction terms. Finally, to invert this gauge transformation, one gets an equation of the form

$$u = 2ie^{\frac{i}{2}F} w + \text{negligible terms}. \quad (1-6)$$

Very recently, Burq and Planchon [2008] and Ionescu and Kenig [2007] were able to use Tao’s ideas in the context of Bourgain’s spaces to prove well-posedness for the Benjamin–Ono equation in  $H^s(\mathbb{R})$  for  $s > \frac{1}{4}$  and  $s \geq 0$ , respectively. The main difficulty arising here is that Bourgain’s spaces do not enjoy an algebra property so that one is losing regularity when estimating  $u$  in terms of  $w$  via Equation (1-6). Burq and Planchon first parilinearized the equation and then used a localized version of the gauge transformation on the worst nonlinear term. On the other hand, Ionescu and Kenig decomposed the solution in two parts: the first one is the smooth solution of BO evolving from the low-frequency part of the initial data while the second one solves a dispersive system renormalized by a gauge transformation involving the first part. The authors were then able to solve the system via a fixed-point argument in a dyadic version of Bourgain’s spaces (already used in the context of wave maps [Tataru 1998]) with a

special structure in low frequencies. We observe that their result only ensures the uniqueness in the class of limits of smooth solutions while Burq and Planchon obtained a stronger uniqueness result. Indeed, by applying their approach to the equation satisfied by the difference of two solutions, they succeed in proving that the flow map associated with BO is Lipschitz in a weaker topology when the initial data belongs to  $H^s(\mathbb{R})$ ,  $s > \frac{1}{4}$ .

In the periodic setting, Molinet [2007; 2008] proved well-posedness in  $H^s(\mathbb{T})$  for  $s \geq \frac{1}{2}$  and  $s \geq 0$ , successively. (This last result is proven to be sharp in [Molinet 2009].) Once again, these works combined Tao’s gauge transformation with estimates in Bourgain’s spaces. It should be pointed out that in the periodic case, one can assume that  $u$  has mean value zero to define a primitive. Then it is easy to check by the mean-value theorem that the gauge transformation in (1-4) is Lipschitz from  $L^2$  into  $L^\infty$ . This property, which is not true on the real line, is crucial to prove the uniqueness and the Lipschitz property of the flow map.

The aim of this paper is to give a simpler proof of Ionescu and Kenig’s result, which also provides a stronger uniqueness result for the solutions at the  $L^2$  level. It is worth noticing that to reach  $L^2$  in [Ionescu and Kenig 2007] and [Molinet 2008], the authors replaced  $u$  in (1-4) by the formula given in (1-6). The benefit of this substitution is that then  $u$  no longer appears in (1-4). On the other hand, it introduces new technical difficulties in handling the multiplication by  $e^{\mp i F/2}$  in Bourgain spaces. Here we are able to avoid this substitution, which will simplify the proof. Our main result is the following:

**Theorem 1.1.** *Let  $s \geq 0$  be given.*

Existence: *For all  $u_0 \in H^s(\mathbb{R})$  and all  $T > 0$ , there exists a solution*

$$u \in C([0, T]; H^s(\mathbb{R})) \cap X_T^{s-1,1} \cap L_T^4 W_x^{s,4} \tag{1-7}$$

of (1-1) such that

$$w = \partial_x P_{+hi}(e^{-\frac{i}{2} F[u]}) \in Y_T^s, \tag{1-8}$$

where  $F[u]$  is some primitive of  $u$  defined in (3-2).

Uniqueness: *This solution is unique in the following classes:<sup>1</sup>*

- (i)  $u \in L^\infty([0, T]; L^2(\mathbb{R})) \cap L^4([0, T] \times \mathbb{R})$  and  $w \in X_T^{0, \frac{1}{2}}$ ,
- (ii)  $u \in L^\infty([0, T]; H^s(\mathbb{R})) \cap L_T^4 W_x^{s,4}$  whenever  $s > 0$ ,
- (iii)  $u \in L^\infty([0, T]; H^s(\mathbb{R}))$  whenever  $s > \frac{1}{4}$ .

Moreover,  $u \in C_b(\mathbb{R}; L^2(\mathbb{R}))$ , and the flow-map data solution  $u_0 \mapsto u$  is continuous from  $H^s(\mathbb{R})$  into  $C([0, T]; H^s(\mathbb{R}))$ .

Note that  $H^s(\mathbb{R})$  above denotes the space of all real-valued functions with the usual norm, and  $X_T^{s,b}$  and  $Y_T^s$  are Bourgain spaces defined in Section 2B while the primitive  $F[u]$  of  $u$  is defined in Section 3A.

**Remark 1.2.** Since the function spaces in the uniqueness class (i) are reflexive and since  $\partial_x P_{+hi}(e^{-\frac{i}{2} F[u_n]})$  converges to  $\partial_x P_{+hi}(e^{-\frac{i}{2} F[u]})$  in  $L^\infty([-T, T]; L^2(\mathbb{R}))$  when  $u_n$  converges to  $u$  in  $L^\infty([-T, T]; L^2(\mathbb{R}))$ , our result clearly implies the uniqueness in the class of  $L^\infty([-T, T]; L^2(\mathbb{R}))$ -limits of smooth solutions.

<sup>1</sup>Note that according to the equation, the time derivative of a solution in these classes belongs to  $L^\infty(-T, T; H^{-2})$ , and thus such a solution has to belong to  $C(-T, T; H^{-2})$ .

**Remark 1.3.** For  $s > 0$  we get a uniqueness class without any conditions on  $w$  (see [Burq and Planchon 2008] for the case  $s > \frac{1}{4}$ ).

**Remark 1.4.** According to (iii) we get unconditional well-posedness in  $H^s(\mathbb{R})$  for  $s > \frac{1}{4}$ . Such a result was first proven, in a much less direct way, in [Burq and Planchon 2006] for  $s \geq \frac{1}{2}$ . It implies in particular the uniqueness of the (energy) weak solutions that belong to  $L^\infty(\mathbb{R}; H^{\frac{1}{2}}(\mathbb{R}))$ . These solutions are constructed by regularizing the equation and passing to the limit as the regularizing coefficient goes to 0 (taking into account some energy estimate for the regularizing equation related to the energy conservation of (1-1)).

Our proof also combines Tao's ideas with the use of Bourgain's spaces. Actually, it closely follows the strategy introduced by the first author in [Molinet 2007]. The main new ingredient is a bilinear estimate for the nonlinear term appearing in (1-5), which allows us to recover one derivative at the  $L^2$  level. It is interesting to note that, at the  $H^s$  level with  $s > 0$ , this estimate follows from the Cauchy–Schwarz method introduced by Kenig, Ponce, and Vega [Kenig et al. 1996] (see the Appendix for the use of this method in some region of integration). To reach  $L^2$ , one of the main difficulties is that we cannot substitute the Fourier transform of  $u$  by its modulus in the bilinear estimate since we are not able to prove that  $\mathcal{F}^{-1}(|\hat{u}|)$  belongs to  $L^4_{x,t}$  but only that  $u$  belongs to  $L^4_{x,t}$ . To overcome this difficulty we use a Littlewood–Paley decomposition of the functions and carefully divide the domain of integration into suitable disjoint subdomains.

To obtain our uniqueness result, following the same method as in the periodic setting, we derive a Lipschitz bound for the gauge transformation from some affine subspaces of  $L^2(\mathbb{R})$  into  $L^\infty(\mathbb{R})$ . Recall that this is clearly not possible for general initial data since it would imply the uniform continuity of the flow map. The main idea is to notice that such a Lipschitz bound holds for solutions emanating from initial data having the same low frequency part, and this is sufficient for our purpose.

Let us point out some applications. First our uniqueness result allows us to simplify the proof of the continuity of the flow map associated with the Benjamin–Ono equation for the weak topology of  $L^2(\mathbb{R})$ . This result was recently proved by Cui and Kenig [2010].

It is also interesting to observe that the method of proof used here still works in the periodic setting, and thus, we reobtain the well-posedness result [Molinet 2008] in a simpler way. Moreover, as in the continuous case, we prove new uniqueness results (see Theorem 7.1). In particular, we get unconditional well-posedness in  $H^s(\mathbb{T})$  as soon as  $s \geq \frac{1}{2}$ .

Finally, we believe that this technique may be useful for other nonlinear dispersive equations presenting the same kind of difficulties as the Benjamin–Ono equation. For example, consider the higher-order Benjamin–Ono equation

$$\partial_t v - b\mathcal{H}\partial_x^2 v + a\partial_x^3 v = cv\partial_x v - d\partial_x(v\mathcal{H}\partial_x v + \mathcal{H}(v\partial_x v)), \quad (1-9)$$

where  $x, t \in \mathbb{R}$ ,  $v$  is a real-valued function,  $a \in \mathbb{R}$ , and  $b, c$ , and  $d$  are positive constants. The equation above corresponds to a second-order approximation model of the same phenomena described by the Benjamin–Ono equation. It was derived by Craig, Guyenne, and Kalisch [2005] using a Hamiltonian perturbation theory and possesses an energy at the  $H^1$  level. As for the Benjamin–Ono equation, the flow map associated with (1-9) fails to be smooth in any Sobolev space  $H^s(\mathbb{R})$ ,  $s \in \mathbb{R}$  [Pilod 2008]. Recently,

the Cauchy problem associated with (1-9) was proved to be well posed in  $H^2(\mathbb{R})$  [Linares et al. 2011]. In a forthcoming paper, the authors will show that it is actually well posed in the energy space  $H^1(\mathbb{R})$ .

This paper is organized as follows: in the next section, we introduce the notations, define the function spaces, and recall some classical linear estimates. Section 3 is devoted to the key nonlinear estimates, which are used in Section 4 to prove the main part of Theorem 1.1 while the assertions (ii) and (iii) are proved in Section 5. In Section 6, we give a simple proof of the continuity of the flow map for the weak  $L^2(\mathbb{R})$  topology whereas Section 7 is devoted to some comments and new results in the periodic case. Finally, in the Appendix we prove the bilinear estimate used in Section 5.

## 2. Notation, function spaces, and preliminary estimates

**2A. Notation.** For any positive numbers  $a$  and  $b$ , the notation  $a \lesssim b$  means that there exists a positive constant  $c$  such that  $a \leq cb$ . We also write  $a \sim b$  when  $a \lesssim b$  and  $b \lesssim a$ . Moreover, if  $\alpha \in \mathbb{R}$ ,  $\alpha_+$  and  $\alpha_-$  will denote a number slightly greater and lesser than  $\alpha$ , respectively.

For  $u = u(x, t) \in \mathcal{S}(\mathbb{R}^2)$ ,  $\mathcal{F}u = \widehat{u}$  will denote its space-time Fourier transform whereas  $\mathcal{F}_x u = (u)^{\wedge_x}$  and  $\mathcal{F}_t u = (u)^{\wedge_t}$  will denote its Fourier transform in space and time, respectively. For  $s \in \mathbb{R}$ , we define the Bessel and Riesz potentials of order  $-s$ ,  $J_x^s$  and  $D_x^s$ , by

$$J_x^s u = \mathcal{F}_x^{-1} (1 + |\xi|^2)^{\frac{s}{2}} \mathcal{F}_x u \quad \text{and} \quad D_x^s u = \mathcal{F}_x^{-1} (|\xi|^s \mathcal{F}_x u).$$

Throughout the paper, we fix a cutoff function  $\eta$  such that

$$\eta \in C_0^\infty(\mathbb{R}), \quad 0 \leq \eta \leq 1, \quad \eta|_{[-1,1]} = 1, \quad \text{supp}(\eta) \subset [-2, 2].$$

We define

$$\phi(\xi) := \eta(\xi) - \eta(2\xi) \quad \text{and} \quad \phi_{2^l}(\xi) := \phi(2^{-l}\xi).$$

Summations over capitalized variables such as  $N$  are presumed to be dyadic with  $N \geq 1$ ; i.e., these variables range over numbers of the form  $2^n$ ,  $n \in \mathbb{Z}_+$ . Then we have

$$\sum_N \phi_N(\xi) = 1 - \eta(2\xi) \quad \forall \xi \neq 0 \quad \text{and} \quad \text{supp}(\phi_N) \subset \{\frac{1}{2}N \leq |\xi| \leq 2N\}.$$

Let us define the Littlewood–Paley multipliers by

$$P_N u = \mathcal{F}_x^{-1} (\phi_N \mathcal{F}_x u) \quad \text{and} \quad P_{\geq N} := \sum_{K \geq N} P_K.$$

We also define the operators  $P_{hi}$ ,  $P_{HI}$ ,  $P_{lo}$ , and  $P_{LO}$  by

$$P_{hi} = \sum_{N \geq 2} P_N, \quad P_{HI} = \sum_{N \geq 8} P_N, \quad P_{lo} = 1 - P_{hi}, \quad \text{and} \quad P_{LO} = 1 - P_{HI}.$$

Let  $P_+$  and  $P_-$  denote the projections on the positive and the negative Fourier frequencies, respectively. Then

$$P_{\pm} u = \mathcal{F}_x^{-1} (\chi_{\mathbb{R}_{\pm}} \mathcal{F}_x u),$$

and we also define  $P_{\pm hi} = P_{\pm} P_{hi}$ ,  $P_{\pm HI} = P_{\pm} P_{HI}$ ,  $P_{\pm lo} = P_{\pm} P_{lo}$ , and  $P_{\pm LO} = P_{\pm} P_{LO}$ . Observe that  $P_{hi}$ ,  $P_{HI}$ ,  $P_{lo}$ , and  $P_{LO}$  are bounded operators on  $L^p(\mathbb{R})$  for  $1 \leq p \leq \infty$  while  $P_{\pm}$  is only bounded on  $L^p(\mathbb{R})$  for  $1 < p < \infty$ . We also note that

$$\mathcal{H} = -iP_+ + iP_-.$$

Finally, we denote by  $U(\cdot)$  the free group associated with the linearized Benjamin–Ono equation, which is to say,

$$\mathcal{F}_x(U(t)f)(\xi) = e^{-it|\xi|\xi} \mathcal{F}_x f(\xi).$$

**2B. Function spaces.** For  $1 \leq p \leq \infty$ ,  $L^p(\mathbb{R})$  is the usual Lebesgue space with the norm  $\|\cdot\|_{L^p}$ , and for  $s \in \mathbb{R}$ , the real-valued Sobolev spaces  $H^s(\mathbb{R})$  and  $W^{s,p}(\mathbb{R})$  denote the spaces of all real-valued functions with the usual norms

$$\|f\|_{H^s} = \|J^s u\|_{L^2} \quad \text{and} \quad \|f\|_{W^{s,p}} = \|J_x^s f\|_{L^p}.$$

For  $1 < p < \infty$ , we define the space  $\tilde{L}^p$  as

$$\|f\|_{\tilde{L}^p} = \|P_{lo} f\|_{L^p} + \left( \sum_N \|P_N f\|_{L^p}^2 \right)^{\frac{1}{2}}.$$

Observe that when  $p \geq 2$ , the Littlewood–Paley theorem on the square function and Minkowski’s inequality imply that the injection  $\tilde{L}^p \hookrightarrow L^p$  is continuous. Moreover, if  $u = u(x, t)$  is a real-valued function defined for  $x \in \mathbb{R}$  and  $t$  in the time interval  $[0, T]$  with  $T > 0$ ,  $B$  is one of the spaces defined above, and  $1 \leq p \leq \infty$ , we will define the mixed space-time spaces  $L_T^p B_x$  and  $L_t^p B_x$  by the norms

$$\|u\|_{L_T^p B_x} = \left( \int_0^T \|u(\cdot, t)\|_B^p dt \right)^{\frac{1}{p}} \quad \text{and} \quad \|u\|_{L_t^p B_x} = \left( \int_{\mathbb{R}} \|u(\cdot, t)\|_B^p dt \right)^{\frac{1}{p}},$$

respectively.

For  $s, b \in \mathbb{R}$ , we introduce the Bourgain spaces  $X^{s,b}$  and  $Z^{s,b}$  related to the Benjamin–Ono equation as the completion of the Schwartz space  $\mathcal{S}(\mathbb{R}^2)$  under the norms

$$\|u\|_{X^{s,b}} = \left( \int_{\mathbb{R}^2} \langle \tau + |\xi|\xi \rangle^{2b} \langle \xi \rangle^{2s} |\widehat{u}(\xi, \tau)|^2 d\xi d\tau \right)^{\frac{1}{2}}, \tag{2-1}$$

$$\|u\|_{Z^{s,b}} = \left( \int_{\mathbb{R}} \left( \int_{\mathbb{R}} \langle \tau + |\xi|\xi \rangle^b \langle \xi \rangle^s |\widehat{u}(\xi, \tau)| d\tau \right)^2 d\xi \right)^{\frac{1}{2}}, \tag{2-2}$$

$$\|u\|_{\tilde{Z}^{s,b}} = \|P_{lo} u\|_{Z^{s,b}} + \left( \sum_N \|P_N u\|_{Z^{s,b}}^2 \right)^{\frac{1}{2}}, \tag{2-3}$$

$$\|u\|_{Y^s} = \|u\|_{X^{s, \frac{1}{2}}} + \|u\|_{\tilde{Z}^{s,0}}, \tag{2-4}$$

where  $\langle x \rangle := 1 + |x|$ . We will also use the localized (in time) version of these spaces. Let  $T > 0$  be a positive time and  $\|\cdot\|_B = \|\cdot\|_{X^{s,b}}$ ,  $\|\cdot\|_{\tilde{Z}^{s,b}}$ , or  $\|\cdot\|_{Y^s}$ . If  $u : \mathbb{R} \times [0, T] \rightarrow \mathbb{C}$ , then

$$\|u\|_{B_T} := \inf \{ \|\tilde{u}\|_B \mid \tilde{u} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{C}, \tilde{u}|_{\mathbb{R} \times [0, T]} = u \}.$$

We recall that

$$Y_T^s \hookrightarrow Z_T^{s,0} \hookrightarrow C([0, T]; H^s(\mathbb{R})).$$

**2C. Linear estimates.** In this subsection, we recall some linear estimates in Bourgain’s spaces that will be needed later. The first ones are well known (see [Ginibre et al. 1997], for example).

**Lemma 2.1** (homogeneous linear estimate). *Let  $s \in \mathbb{R}$ . Then*

$$\|\eta(t) U(t) f\|_{Y^s} \lesssim \|f\|_{H^s}. \tag{2-5}$$

**Lemma 2.2** (nonhomogeneous linear estimate). *Let  $s \in \mathbb{R}$ . Then, for any  $0 < \delta < \frac{1}{2}$ ,*

$$\left\| \eta(t) \int_0^t U(t-t') g(t') dt' \right\|_{X^{s, \frac{1}{2}+\delta}} \lesssim \|g\|_{X^{s, -\frac{1}{2}+\delta}} \tag{2-6}$$

and

$$\left\| \eta(t) \int_0^t U(t-t') g(t') dt' \right\|_{Y^s} \lesssim \|g\|_{X^{s, -\frac{1}{2}}} + \|g\|_{\tilde{Z}^{s, -1}}. \tag{2-7}$$

*Proof.* Lemmas 2.1 and 2.2 follow directly from the classical linear estimates for  $X^{s,b}$  and  $Z^{s,b}$  together with the fact that

$$\|u\|_{X^{s,b}} \sim \|P_{I_0} u\|_{X^{s,b}} + \left( \sum_N \|P_N u\|_{X^{s,b}}^2 \right)^{\frac{1}{2}}. \quad \square$$

**Lemma 2.3.** *For any  $T > 0$ ,  $s \in \mathbb{R}$  and for all  $-\frac{1}{2} < b' \leq b < \frac{1}{2}$ ,*

$$\|u\|_{X_T^{s,b'}} \lesssim T^{b-b'} \|u\|_{X_T^{s,b}}. \tag{2-8}$$

The following Bourgain–Strichartz estimates will also be useful:

**Lemma 2.4.** *It holds that*

$$\|u\|_{L_{x,t}^4} \lesssim \|u\|_{\tilde{L}_{x,t}^4} \lesssim \|u\|_{X^{0, \frac{3}{8}}}, \tag{2-9}$$

and for any  $T > 0$  and  $\frac{3}{8} \leq b \leq \frac{1}{2}$ ,

$$\|u\|_{L_{x,T}^4} \lesssim T^{b-\frac{3}{8}} \|u\|_{X_T^{0,b}}. \tag{2-10}$$

*Proof.* Estimate (2-9) follows directly by applying the estimate

$$\|u\|_{L_{x,t}^4} \lesssim \|u\|_{X^{0, \frac{3}{8}}},$$

proved in the appendix of [Molinet 2007], to each dyadic block on the left-hand side of (2-9).

To prove (2-10), we choose an extension  $\tilde{u} \in X^{0,b}$  of  $u$  such that  $\|\tilde{u}\|_{X^{0,b}} \leq 2\|u\|_{X_T^{0,b}}$ . Therefore, it follows from (2-8) and (2-9) that

$$\|u\|_{L_{x,T}^4} \leq \|\tilde{u}\|_{L_{x,t}^4} \lesssim \|\tilde{u}\|_{X^{0, \frac{3}{8}}} \lesssim T^{b-\frac{3}{8}} \|u\|_{X_T^{0,b}}. \quad \square$$

**2D. Fractional Leibniz rules.** First we state the classical fractional Leibniz rule estimate derived by Kenig, Ponce, and Vega (see Theorems A.8 and A.12 in [Kenig et al. 1993]).

**Proposition 2.5.** *Let  $0 < \alpha < 1$ ,  $p, p_1, p_2 \in (1, +\infty)$  with  $\frac{1}{p_1} + \frac{1}{p_2} = \frac{1}{p}$ , and  $\alpha_1, \alpha_2 \in [0, \alpha]$  with  $\alpha = \alpha_1 + \alpha_2$ . Then*

$$\|D_x^\alpha(fg) - fD_x^\alpha g - gD_x^\alpha f\|_{L^p} \lesssim \|D_x^{\alpha_1} g\|_{L^{p_1}} \|D_x^{\alpha_2} f\|_{L^{p_2}}. \quad (2-11)$$

Moreover, for  $\alpha_1 = 0$ , the value  $p_1 = +\infty$  is allowed.

The next estimate is a frequency-localized version of estimate (2-11) in the same spirit as Lemma 3.2 in [Tao 2004]. It allows sharing most of the fractional derivative in the first term on the right-hand side of (2-12).

**Lemma 2.6.** *Let  $\alpha \geq 0$  and  $1 < q < \infty$ . Then*

$$\|D_x^\alpha P_+(fP_- \partial_x g)\|_{L^q} \lesssim \|D_x^{\alpha_1} f\|_{L^{q_1}} \|D_x^{\alpha_2} g\|_{L^{q_2}} \quad (2-12)$$

with  $1 < q_i < \infty$ ,  $\frac{1}{q_1} + \frac{1}{q_2} = \frac{1}{q}$ , and  $\alpha_1 \geq \alpha$ ,  $\alpha_2 \geq 0$ , and  $\alpha_1 + \alpha_2 = 1 + \alpha$ .

*Proof.* See Lemma 3.2 in [Molinet 2007]. □

Finally, we derive an estimate to handle the multiplication by a term of the form  $e^{\pm \frac{i}{2}F}$ , where  $F$  is a real-valued function, in fractional Sobolev spaces.

**Lemma 2.7.** *Let  $2 \leq q < \infty$  and  $0 \leq \alpha \leq \frac{1}{q}$ . Consider  $F_1$  and  $F_2$ , two real-valued functions such that  $u_j = \partial_x F_j$  belongs to  $L^2(\mathbb{R})$  for  $j = 1, 2$ . Then*

$$\|J_x^\alpha(e^{\pm \frac{i}{2}F_1} g)\|_{L^q} \lesssim (1 + \|u_1\|_{L^2}) \|J_x^\alpha g\|_{L^q}, \quad (2-13)$$

and

$$\|J_x^\alpha((e^{\pm \frac{i}{2}F_1} - e^{\pm \frac{i}{2}F_2})g)\|_{L^q} \lesssim (\|u_1 - u_2\|_{L^2} + \|e^{\pm \frac{i}{2}F_1} - e^{\pm \frac{i}{2}F_2}\|_{L^\infty}(1 + \|u_1\|_{L^2})) \|J_x^\alpha g\|_{L^q}. \quad (2-14)$$

*Proof.* In the case  $\alpha = 0$ , we deduce from Hölder's inequality that

$$\|e^{\pm \frac{i}{2}F_1} g\|_{L^q} \leq \|g\|_{L^q} \quad (2-15)$$

since  $F_1$  is real-valued. Therefore, we assume that  $0 < \alpha \leq \frac{1}{q}$ , and it is enough to bound  $\|D_x^\alpha(e^{\pm \frac{i}{2}F_1} g)\|_{L^q}$ . First we observe that

$$\|D_x^\alpha(e^{\pm \frac{i}{2}F_1} g)\|_{L^q} \leq \|D_x^\alpha(P_{lo}e^{\pm \frac{i}{2}F_1} g)\|_{L^q} + \|D_x^\alpha(P_{hi}e^{\pm \frac{i}{2}F_1} g)\|_{L^q}. \quad (2-16)$$

Estimate (2-11) and Bernstein's inequality imply that

$$\|D_x^\alpha(P_{lo}e^{\pm \frac{i}{2}F_1} g)\|_{L^q} \lesssim \|P_{lo}e^{\pm \frac{i}{2}F_1}\|_{L^\infty} \|D_x^\alpha g\|_{L^q} + \|D_x^\alpha P_{lo}e^{\pm \frac{i}{2}F_1}\|_{L^\infty} \|g\|_{L^q} \lesssim \|J_x^\alpha g\|_{L^q}. \quad (2-17)$$

On the other hand, by using estimate (2-11) again, we get that

$$\|D_x^\alpha(P_{hi}e^{\pm \frac{i}{2}F_1} g)\|_{L^q} \lesssim \|P_{hi}e^{\pm \frac{i}{2}F_1}\|_{L^\infty} \|D_x^\alpha g\|_{L^q} + \|g\|_{L^{q_1}} \|D_x^\alpha P_{hi}e^{\pm \frac{i}{2}F_1}\|_{L^{q_2}}$$

with  $\frac{1}{q_1} = \frac{1}{q} - \alpha$  and  $\frac{1}{q_2} = \alpha$ , so  $\frac{1}{q_1} + \frac{1}{q_2} = \frac{1}{q}$ . Then it follows from the real-valuedness of  $F_1$ , the equality

$\partial_x F_1 = u_1$ , and the Sobolev embedding that

$$\begin{aligned} \|D_x^\alpha (P_{hi} e^{\pm \frac{i}{2} F_1} g)\|_{L^q} &\lesssim \|D_x^\alpha g\|_{L^q} + \|J_x^\alpha g\|_{L^q} \|D_x^{\alpha + \frac{1}{2}} P_{hi} e^{\pm \frac{i}{2} F_1}\|_{L^2} \\ &\lesssim \|J_x^\alpha g\|_{L^q} (1 + \|\partial_x e^{\pm \frac{i}{2} F_1}\|_{L^2}) \\ &\lesssim \|J_x^\alpha g\|_{L^q} (1 + \|u_1\|_{L^2}). \end{aligned} \tag{2-18}$$

The proof of estimate (2-13) is concluded gathering (2-15)–(2-18).

Estimate (2-14) can be obtained exactly in the same way, using that

$$\|\partial_x (e^{\pm \frac{i}{2} F_1} - e^{\pm \frac{i}{2} F_2})\|_{L^2} \lesssim \|u_1 - u_2\|_{L^2} + \|e^{\pm \frac{i}{2} F_1} - e^{\pm \frac{i}{2} F_2}\|_{L^\infty} \|u_1\|_{L^2}. \tag{2-19}$$

This completes the proof. □

### 3. A priori estimates in $H^s(\mathbb{R})$ for $s \geq 0$

In this section we will derive *a priori* estimates on a solution  $u$  to (1-1) at the  $H^s$ -level for  $s \geq 0$ . First, following Tao [2004], we perform a nonlinear transformation on the equation to weaken the high-low frequency interaction in the nonlinearity. Furthermore, since we want to reach  $L^2$ , we will need to use Bourgain spaces. This requires a new bilinear estimate, which we derive in Section 3B.

**3A. The gauge transformation.** Let  $u$  be a solution to the equation in (1-1). First we construct a spatial primitive  $F = F[u]$  of  $u$  (i.e.,  $\partial_x F = u$ ) that satisfies the equation

$$\partial_t F = -\mathcal{H} \partial_x^2 F + \frac{1}{2} (\partial_x F)^2. \tag{3-1}$$

Note that these two properties defined  $F$  up to a constant. In order to construct  $F$  for  $u$  with low regularity, we use the construction of Burq and Planchon [2008]. Consider  $\psi \in C_0^\infty(\mathbb{R})$  such that  $\int_{\mathbb{R}} \psi(y) dy = 1$  and define

$$F(x, t) = \int_{\mathbb{R}} \psi(y) \left( \int_y^x u(z, t) dz \right) dy + G(t) \tag{3-2}$$

as a mean of antiderivatives of  $u$ . Obviously,  $\partial_x F = u$  and

$$\begin{aligned} \partial_t F(x, t) &= \int_{\mathbb{R}} \psi(y) \left( \int_y^x \partial_t u(z, t) dz \right) dy + G'(t) \\ &= \int_{\mathbb{R}} \psi(y) \left( \int_y^x (-\mathcal{H} \partial_z^2 u(z, t) + \frac{1}{2} \partial_z (u(z, t)^2)) dz \right) dy + G'(t) \\ &= -\mathcal{H} \partial_x u(x, t) + \frac{1}{2} u(x, t)^2 + \int_{\mathbb{R}} (\mathcal{H} \psi'(y) u(y, t) - \psi(y) \frac{1}{2} u(y, t)^2) dy + G'(t). \end{aligned}$$

Therefore, we choose  $G$  as

$$G(t) = \int_0^t \int_{\mathbb{R}} (-\mathcal{H} \psi'(y) u(y, s) + \psi(y) \frac{1}{2} u(y, s)^2) dy ds$$



to ensure that (3-1) is satisfied. Observe that this construction makes sense for  $u \in L_{\text{loc}}^2(\mathbb{R}^2)$ . Next, we introduce the new unknown

$$W = P_{+hi}(e^{-\frac{i}{2}F}) \quad \text{and} \quad w = \partial_x W = -\frac{1}{2}i P_{+hi}(e^{-\frac{i}{2}F} u). \quad (3-3)$$

Then it follows from (3-1) and the identity  $\mathcal{H} = -i(P_+ - P_-)$  that

$$\begin{aligned} \partial_t W + \mathcal{H} \partial_x^2 W &= \partial_t W - i \partial_x^2 W = -\frac{1}{2}i P_{+hi}(e^{-\frac{i}{2}F} (\partial_t F - i \partial_x^2 F - \frac{1}{2}(\partial_x F)^2)) \\ &= -P_{+hi}(W P_- \partial_x u) - P_{+hi}(P_{lo} e^{-\frac{i}{2}F} P_- \partial_x u) \end{aligned}$$

since the term  $-P_{+hi}(P_{-hi} e^{-\frac{i}{2}F} P_- \partial_x u)$  cancels due to the frequency localization. Thus, it follows from differentiating that

$$\partial_t w - i \partial_x^2 w = -\partial_x P_{+hi}(W P_- \partial_x u) - \partial_x P_{+hi}(P_{lo} e^{-\frac{i}{2}F} P_- \partial_x u). \quad (3-4)$$

On the other hand, one can write  $u$  as

$$u = F_x = e^{\frac{i}{2}F} e^{-\frac{i}{2}F} F_x = 2i e^{\frac{i}{2}F} \partial_x (e^{-\frac{i}{2}F}) = 2i e^{\frac{i}{2}F} w - e^{\frac{i}{2}F} P_{lo}(e^{-\frac{i}{2}F} u) - e^{\frac{i}{2}F} P_{-hi}(e^{-\frac{i}{2}F} u) \quad (3-5)$$

so that it follows from the frequency localization

$$P_{+HI} u = 2i P_{+HI}(e^{\frac{i}{2}F} w) - P_{+HI}(P_{+hi} e^{\frac{i}{2}F} P_{lo}(e^{-\frac{i}{2}F} u)) + 2i P_{+HI}(P_{+HI} e^{\frac{i}{2}F} \partial_x P_{-hi} e^{-\frac{i}{2}F} u). \quad (3-6)$$

**Remark 3.1.** Note that the use of  $P_{+HI}$  allows us to replace  $e^{\frac{i}{2}F}$  by  $P_{+hi} e^{\frac{i}{2}F}$  in the second term on the right-hand side of (3-6). This fact will be useful to obtain at least a quadratic term in  $\|u\|_{L_T^\infty L_x^2}$  on the right-hand side of estimate (3-8) in Proposition 3.2.

Then we have the following *a priori* estimates for  $u$  in terms of  $w$ :

**Proposition 3.2.** *Let  $0 \leq s \leq 1$ ,  $0 < T \leq 1$ ,  $0 \leq \theta \leq 1$ , and  $u$  be a solution to (1-1) in the time interval  $[0, T]$ . Then*

$$\|u\|_{X_T^{s-\theta, \theta}} \lesssim \|u\|_{L_T^\infty H_x^s} + \|u\|_{L_{T,x}^4} \|J_x^s u\|_{L_{T,x}^4}. \quad (3-7)$$

Moreover, if  $0 \leq s \leq \frac{1}{4}$ , we have

$$\|J_x^s u\|_{L_T^p L_x^q} \lesssim \|u_0\|_{L^2} + (1 + \|u\|_{L_T^\infty L_x^2})(\|w\|_{Y_T^s} + \|u\|_{L_T^\infty L_x^2}^2) \quad (3-8)$$

for  $(p, q) = (\infty, 2)$  or  $(4, 4)$ .

**Remark 3.3.** One can rewrite (3-8) in a convenient form for  $s \geq \frac{1}{4}$ ; see [Molinet 2007].

*Proof.* We begin with the proof of estimate (3-7) and construct a suitable extension in time  $\tilde{u}$  of  $u$ . First we consider  $v(t) = U(-t)u(t)$  on the time interval  $[0, T]$  and extend  $v$  on  $[-2, 2]$  by setting  $\partial_t v = 0$  on  $[-2, 2] \setminus [0, T]$ . Then it is pretty clear that

$$\|\partial_t v\|_{L_{[-2,2]}^2 H_x^r} = \|\partial_t v\|_{L_T^2 H_x^r} \quad \text{and} \quad \|v\|_{L_{[-2,2]}^2 H_x^r} \lesssim \|v\|_{L_T^\infty H_x^r}$$

for all  $r \in \mathbb{R}$ . Now we define  $\tilde{u}(x, t) = \eta(t) U(t) v(t)$ . Obviously,

$$\|\tilde{u}\|_{X^{s-1,1}} \lesssim \|\partial_t v\|_{L_{[-2,2]}^2 H_x^{s-1}} + \|v\|_{L_{[-2,2]}^2 H_x^{s-1}} \lesssim \|\partial_t v\|_{L_T^2 H_x^{s-1}} + \|v\|_{L_T^\infty H_x^{s-1}} \quad (3-9)$$

and

$$\|\tilde{u}\|_{X^{s,0}} \lesssim \|v\|_{L^2_{[-2,2]}H^s_x} \lesssim \|v\|_{L^{\infty}_T H^s_x} = \|u\|_{L^{\infty}_T H^s_x}. \tag{3-10}$$

Interpolating between (3-9) and (3-10) and using the identity

$$\partial_t v = \mathcal{H} \partial_x^2 U(-t) u + U(-t) \partial_t u = U(-t) [\mathcal{H} \partial_x^2 u + \partial_t u],$$

we then deduce that

$$\|\tilde{u}\|_{X^{s-\theta,\theta}} \lesssim \|\partial_t u + \mathcal{H} \partial_x^2 u\|_{L^2_T H^{s-1}_x} + \|u\|_{L^{\infty}_T H^s_x} \tag{3-11}$$

for all  $0 \leq \theta \leq 1$ . Therefore, the fact that  $u$  is a solution to (1-1) and the fractional Leibniz rule [Kenig et al. 1993] yield

$$\|\tilde{u}\|_{X^{s-\theta,\theta}} \lesssim \|u\|_{L^{\infty}_T H^s_x} + \|u\|_{L^4_{x,T}} \|J_x^s u\|_{L^4_{x,T}},$$

which concludes the proof of (3-7) since  $\tilde{u}$  extends  $u$  outside of  $[0, T]$ .

Next, we turn to the proof of (3-8). Let  $0 \leq T \leq 1, 0 \leq s \leq \frac{1}{4}, (p, q) = (\infty, 2)$  or  $(4, 4)$ , and  $u$  be a smooth solution to the equation in (1-1). Since  $u$  is real-valued, it that holds  $P_- u = \overline{P_+ u}$  so that

$$\|J_x^s u\|_{L^p_T L^q_x} \lesssim \|P_{Lo} u\|_{L^p_T L^q_x} + \|D_x^s P_{+HI} u\|_{L^p_T L^q_x}. \tag{3-12}$$

To estimate the second term on the right-hand side of (3-12), we use (3-6) to deduce that

$$\begin{aligned} \|D_x^s P_{+HI} u\|_{L^p_T L^q_x} &\lesssim \|D_x^s P_{+HI}(e^{\frac{i}{2}F} w)\|_{L^p_T L^q_x} + \|D_x^s P_{+HI}(P_{+hi} e^{\frac{i}{2}F} P_{lo}(e^{-\frac{i}{2}F} u))\|_{L^p_T L^q_x} \\ &\quad + \|D_x^s P_{+HI}(P_{+HI} e^{\frac{i}{2}F} \partial_x P_{-hi} e^{-\frac{i}{2}F})\|_{L^p_T L^q_x} \\ &=: I + II + III. \end{aligned}$$

Estimates (2-10) and (2-13) yield

$$I \lesssim (1 + \|u\|_{L^{\infty}_T L^2_x}) \|J_x^s w\|_{L^p_T L^q_x} \lesssim (1 + \|u\|_{L^{\infty}_T L^2_x}) \|w\|_{Y^s_T}. \tag{3-13}$$

On the other hand, the fractional Leibniz rule (Proposition 2.5), Hölder’s inequality in time, and the Sobolev embedding imply that

$$\begin{aligned} II &\lesssim \|D_x^s P_{+hi} e^{\frac{i}{2}F}\|_{L^p_T L^q_x} \|P_{+lo}(ue^{-\frac{i}{2}F})\|_{L^{\infty}_{T,x}} + \|P_{+hi} e^{\frac{i}{2}F}\|_{L^{\infty}_{T,x}} \|D_x^s P_{+lo}(ue^{-\frac{i}{2}F})\|_{L^p_T L^q_x} \\ &\lesssim \|\partial_x P_{+hi} e^{\frac{i}{2}F}\|_{L^p_T L^2_x} \|P_{+lo}(ue^{-\frac{i}{2}F})\|_{L^{\infty}_T L^2_x} \lesssim T^{\frac{1}{p}} \|u\|_{L^{\infty}_T L^2_x}^2. \end{aligned} \tag{3-14}$$

Finally, estimate (2-12) with  $\alpha_1 = \alpha_2 = (1 + s)/2$  and  $q_1 = q_2 = q$ , Hölder’s inequality in time, and the Sobolev embedding lead to

$$\begin{aligned} III &\lesssim \|D_x^{(1+s)/2} P_{+HI} e^{\frac{i}{2}F}\|_{L^{2p}_T L^{2q}_x} \|D_x^{(1+s)/2} P_{-hi} e^{-\frac{i}{2}F}\|_{L^{2p}_T L^{2q}_x} \\ &\lesssim T^{\frac{1}{p}} \|D_x^{1+\frac{s}{2}-\frac{1}{2q}} P_{+HI} e^{\frac{i}{2}F}\|_{L^{\infty}_T L^2_x} \|D_x^{1+\frac{s}{2}-\frac{1}{2q}} P_{-hi} e^{-\frac{i}{2}F}\|_{L^{\infty}_T L^2_x} \\ &\lesssim T^{\frac{1}{p}} \|\partial_x P_{+HI} e^{\frac{i}{2}F}\|_{L^{\infty}_T L^2_x} \|\partial_x P_{-hi} e^{-\frac{i}{2}F}\|_{L^{\infty}_T L^2_x} \lesssim T^{\frac{1}{p}} \|u\|_{L^{\infty}_T L^2_x}^2, \end{aligned} \tag{3-15}$$

since  $0 \leq s \leq \frac{1}{q}$ . Therefore, we deduce by gathering (3-13)–(3-15) that

$$\|D_x^s P_{+HI} u\|_{L_T^p L_x^q} \lesssim (1 + \|u\|_{L_T^\infty L_x^2}) (\|w\|_{Y_T^s} + T^{\frac{1}{p}} \|u\|_{L_T^\infty L_x^2}^2). \quad (3-16)$$

Next we turn to the first term on the right-hand side of (3-12) and consider the integral equation satisfied by  $P_{LO} u$ ,

$$P_{LO} u = U(t) P_{LO} u_0 + \int_0^t U(t-\tau) P_{LO} \partial_x(u^2)(\tau) d\tau. \quad (3-17)$$

First observe that

$$\|P_{LO} u\|_{L_T^p L_x^q} \lesssim T^{\frac{1}{p}} \|P_{LO} u\|_{L_T^\infty L_x^2}.$$

Then we deduce from (3-17), using the fact that  $U$  is a unitary group in  $L^2$  and Bernstein's inequality, that

$$\begin{aligned} \|P_{LO} u\|_{L_T^p L_x^q} &\lesssim T^{\frac{1}{p}} \|u_0\|_{L_x^2} + T^{1+\frac{1}{p}} \|\partial_x P_{LO}(u^2)\|_{L_T^\infty L_x^2} \\ &\lesssim T^{\frac{1}{p}} \|u_0\|_{L_x^2} + T^{1+\frac{1}{p}} \|P_{LO}(u^2)\|_{L_T^\infty L_x^1} \\ &\lesssim \|u_0\|_{L_x^2} + \|u\|_{L_T^\infty L_x^2}^2, \end{aligned} \quad (3-18)$$

since  $0 \leq T \leq 1$ .

Thus, estimate (3-8) follows combining (3-12), (3-16), and (3-18). This concludes the proof of Proposition 3.2.  $\square$

**3B. Bilinear estimates.** The aim of this subsection is to derive the following estimate of  $\|w\|_{Y_T^s}$ :

**Proposition 3.4.** *Let  $0 < T \leq 1$ ,  $0 \leq s \leq \frac{1}{2}$ , and  $u$  be a solution to (1-1) on the time interval  $[0, T]$ . Then*

$$\|w\|_{Y_T^s} \lesssim (1 + \|u_0\|_{L^2}) \|u_0\|_{H^s} + \|u\|_{L_{x,T}^4}^2 + \|w\|_{X_T^{s,\frac{1}{2}}} (\|u\|_{L_T^\infty L_x^2} + \|u\|_{L_{x,T}^4} + \|u\|_{X_T^{-1,1}}). \quad (3-19)$$

The main tools to prove Proposition 3.4 are the following crucial bilinear estimates:

**Proposition 3.5.** *For any  $s \geq 0$ , we have*

$$\|\partial_x P_{+hi}(\partial_x^{-1} w P_- \partial_x u)\|_{X^{s,-\frac{1}{2}}} \lesssim \|w\|_{X^{s,\frac{1}{2}}} (\|u\|_{L_{x,t}^2} + \|u\|_{L_{x,t}^4} + \|u\|_{X^{-1,1}}) \quad (3-20)$$

and

$$\|\partial_x P_{+hi}(\partial_x^{-1} w P_- \partial_x u)\|_{\tilde{Z}^{s,-1}} \lesssim \|w\|_{X^{s,\frac{1}{2}}} (\|u\|_{L_{x,t}^2} + \|u\|_{L_{x,t}^4} + \|u\|_{X^{-1,1}}). \quad (3-21)$$

**Remark 3.6.** Note that  $\partial_x^{-1} w$  is well defined since  $w$  is localized in high frequencies.

*Proof.* We will only give the proof in the case of  $s = 0$  since the case  $s > 0$  can be deduced by using similar arguments. By duality, to prove (3-20) is equivalent to prove that

$$|I| \lesssim \|h\|_{L_{x,t}^2} \|w\|_{X^{0,\frac{1}{2}}} (\|u\|_{L_{x,t}^2} + \|u\|_{L_{x,t}^4} + \|u\|_{X^{-1,1}}), \quad (3-22)$$

where

$$I = \int_{\mathbb{Q}} \frac{\xi}{\langle \sigma \rangle^{1/2}} \widehat{h}(\xi, \tau) \xi_1^{-1} \widehat{w}(\xi_1, \tau_1) \xi_2 \widehat{u}(\xi_2, \tau_2) d\nu, \quad (3-23)$$

$$d\nu = d\xi d\xi_1 d\tau d\tau_1, \quad \xi_2 = \xi - \xi_1, \quad \tau_2 = \tau - \tau_1, \quad \sigma_i = \tau_i + \xi_i |\xi_i|, \quad i = 1, 2, \quad (3-24)$$

and

$$\mathcal{D} = \{ (\xi, \xi_1, \tau, \tau_1) \in \mathbb{R}^4 \mid \xi \geq 1, \xi_1 \geq 1, \xi_2 \leq 0 \}. \tag{3-25}$$

Observe that we always have in  $\mathcal{D}$  that

$$\xi_1 \geq \xi \geq 1 \quad \text{and} \quad \xi_1 \geq |\xi_2|. \tag{3-26}$$

In the case where  $|\xi_2| \leq 1$ , we have by using Hölder’s inequality and estimate (2-9) that

$$|I| \lesssim \int_{\mathbb{R}^4} \frac{|\widehat{h}|}{\langle \sigma \rangle^{1/2}} |\widehat{w}(\xi_1, \tau_1)| |\widehat{u}(\xi_2, \tau_2)| \, d\nu \lesssim \left\| \left( \frac{|\widehat{h}|}{\langle \sigma \rangle^{1/2}} \right)^\vee \right\|_{L^4_{x,t}} \|(|\widehat{w}|)^\vee\|_{L^4_{x,t}} \|u\|_{L^2_{x,t}} \lesssim \|h\|_{L^2_{x,t}} \|w\|_{X^{\frac{3}{8}}} \|u\|_{L^2_{x,t}}.$$

From now on we will assume that  $|\xi_2| \geq 1$  in  $\mathcal{D}$ .

By using a dyadic decomposition in space-frequency for the functions  $h, w$ , and  $u$ , one can rewrite  $I$  as

$$I = \sum_{N, N_1, N_2} I_{N, N_1, N_2} \tag{3-27}$$

with

$$I_{N, N_1, N_2} := \int_{\mathcal{D}} \frac{\xi}{\langle \sigma \rangle^{1/2}} \widehat{P_N h}(\xi, \tau) \xi_1^{-1} \widehat{P_{N_1} w}(\xi_1, \tau_1) \xi_2 \widehat{P_{N_2} u}(\xi_2, \tau_2) \, d\nu$$

and the dyadic numbers  $N, N_1$ , and  $N_2$  ranging from 1 to  $+\infty$ . Moreover, the resonance identity

$$\sigma_1 + \sigma_2 - \sigma = \xi_1^2 + (\xi - \xi_1)|\xi - \xi_1| - \xi^2 = -2\xi\xi_2 \tag{3-28}$$

holds in  $\mathcal{D}$ . Therefore, to calculate  $I_{N, N_1, N_2}$ , we split the integration domain  $\mathcal{D}$  into the disjoint regions

$$\begin{aligned} \mathcal{A}_{N, N_2} &= \{ (\xi, \xi_1, \tau, \tau_1) \in \mathcal{D} \mid |\sigma| \geq \frac{1}{6} N N_2 \}, \\ \mathcal{B}_{N, N_2} &= \{ (\xi, \xi_1, \tau, \tau_1) \in \mathcal{D} \mid |\sigma| < \frac{1}{6} N N_2, |\sigma_1| \geq \frac{1}{6} N N_2 \}, \\ \mathcal{C}_{N, N_2} &= \{ (\xi, \xi_1, \tau, \tau_1) \in \mathcal{D} \mid |\sigma| < \frac{1}{6} N N_2, |\sigma_1| < \frac{1}{6} N N_2, |\sigma_2| \geq \frac{1}{6} N N_2 \}, \end{aligned}$$

and denote by  $I_{N, N_1, N_2}^{\mathcal{A}_{N, N_2}}$ ,  $I_{N, N_1, N_2}^{\mathcal{B}_{N, N_2}}$ , and  $I_{N, N_1, N_2}^{\mathcal{C}_{N, N_2}}$  the restriction of  $I_{N, N_1, N_2}$  to each of these regions. Then it follows that

$$I_{N, N_1, N_2} = I_{N, N_1, N_2}^{\mathcal{A}_{N, N_2}} + I_{N, N_1, N_2}^{\mathcal{B}_{N, N_2}} + I_{N, N_1, N_2}^{\mathcal{C}_{N, N_2}},$$

and thus

$$|I| \leq |I_{\mathcal{A}}| + |I_{\mathcal{B}}| + |I_{\mathcal{C}}|, \tag{3-29}$$

where

$$I_{\mathcal{A}} := \sum_{N, N_1, N_2} I_{N, N_1, N_2}^{\mathcal{A}_{N, N_2}}, \quad I_{\mathcal{B}} := \sum_{N, N_1, N_2} I_{N, N_1, N_2}^{\mathcal{B}_{N, N_2}}, \quad I_{\mathcal{C}} := \sum_{N, N_1, N_2} I_{N, N_1, N_2}^{\mathcal{C}_{N, N_2}}.$$

Therefore, it suffices to bound  $|I_{\mathcal{A}}|$ ,  $|I_{\mathcal{B}}|$ , and  $|I_{\mathcal{C}}|$ . Note that one of the two following cases holds:

- (1) high-low interaction:  $N_1 \sim N$  and  $N_2 \leq N_1$ ,
- (2) high-high interaction:  $N_1 \sim N_2$  and  $N \leq N_1$ .

*Estimate for  $|I_{\mathcal{A}}|$ .* In the first case, we observe from the Cauchy–Schwarz inequality that

$$\begin{aligned} |I_{\mathcal{A}}| &\sim \left| \int_{\mathbb{R}^2} \widehat{h} \sum_{N_1} \sum_{j=0}^{\frac{\ln N_1}{\ln 2}} \phi_{N_1} \xi \langle \sigma \rangle^{-\frac{1}{2}} \chi_{\{|\sigma| \geq \frac{1}{6} N_1 2^{2-j}\}} \overline{\mathcal{F}}(P_+(\partial_x^{-1} P_{N_1} w P_- \partial_x P_{2^{-j} N_1} u)) d\xi d\tau \right| \\ &\lesssim \|\widehat{h}\|_{L_{\xi, \tau}^2} \left\| \sum_{N_1} \sum_{j \geq 0} N_1 (N_1^2 2^{-j})^{-\frac{1}{2}} \phi_{N_1} |\overline{\mathcal{F}}(P_+(\partial_x^{-1} P_{N_1} w P_- \partial_x P_{2^{-j} N_1} u))| \right\|_{L_{\xi, \tau}^2}. \end{aligned}$$

Then the Plancherel identity and the triangular inequality imply that

$$|I_{\mathcal{A}}| \lesssim \|h\|_{L_{x,t}^2} \sum_{j \geq 0} \left( \sum_{N_1} 2^j \|P_{N_1}(\partial_x^{-1} P_{N_1} w P_- \partial_x P_{2^{-j} N_1} u)\|_{L_{x,t}^2}^2 \right)^{\frac{1}{2}}.$$

By using the Hölder and Bernstein inequalities, we deduce that

$$\begin{aligned} |I_{\mathcal{A}}| &\lesssim \|h\|_{L_{x,t}^2} \sum_{j \geq 0} \left( \sum_{N_1} 2^{-j} \|P_{N_1} w\|_{L_{x,t}^4}^2 \|P_{2^{-j} N_1} u\|_{L_{x,t}^4}^2 \right)^{\frac{1}{2}} \\ &\lesssim \|h\|_{L_{x,t}^2} \left( \sum_N \|P_N w\|_{L_{x,t}^4}^2 \right)^{\frac{1}{2}} \|u\|_{L_{x,t}^4}. \end{aligned} \quad (3-30)$$

In the second case, it follows using the same strategy as in the first case that

$$|I_{\mathcal{A}}| \lesssim \|h\|_{L_{x,t}^2} \sum_{j \geq 0} \left( \sum_{N_1} (2^{-j} N_1)^2 (2^{-j} N_1 N_1)^{-1} \|P_{2^{-j} N_1}(\partial_x^{-1} P_{N_1} w P_- \partial_x P_{N_1} u)\|_{L_{x,t}^2}^2 \right)^{\frac{1}{2}},$$

which implies using the Hölder and Bernstein inequalities that

$$\begin{aligned} |I_{\mathcal{A}}| &\lesssim \|h\|_{L_{x,t}^2} \sum_{j \geq 0} \left( \sum_{N_1} 2^{-j} \|P_{N_1} w\|_{L_{x,t}^4}^2 \|P_{N_1} u\|_{L_{x,t}^4}^2 \right)^{\frac{1}{2}} \\ &\lesssim \|h\|_{L_{x,t}^2} \left( \sum_{N_1} \|P_{N_1} w\|_{L_{x,t}^4}^2 \right)^{\frac{1}{2}} \|u\|_{L_{x,t}^4}. \end{aligned} \quad (3-31)$$

Therefore, we deduce by gathering (3-30)–(3-31) and using estimate (2-9) that

$$|I_{\mathcal{A}}| \leq \|h\|_{L_{x,t}^2} \|w\|_{X^{0, \frac{3}{8}}} \|u\|_{L_{x,t}^4}. \quad (3-32)$$

*Estimate for  $|I_{\mathcal{B}}|$ .* By again using the triangular and the Cauchy–Schwarz inequalities, we have in the first case that

$$|I_{\mathcal{B}}| \leq \|w\|_{X^{0, \frac{1}{2}}} \sum_{j \geq 0} \left( \sum_{N_1} N_1^{-2} (N_1 2^{-j} N_1)^{-1} \left\| P_{N_1} \left( \partial_x P_{+hi} P_{N_1} \left( \frac{\widehat{h}}{\langle \sigma \rangle^{1/2}} \right)^\vee P_+ \partial_x P_{2^{-j} N_1} \tilde{u} \right) \right\|_{L_{x,t}^2}^2 \right)^{\frac{1}{2}},$$

where  $\tilde{u}(x, t) = u(-x, -t)$ . Thus, it follows from the Bernstein and Hölder inequalities that

$$\begin{aligned} |I_{\mathfrak{B}}| &\lesssim \|w\|_{X^{0, \frac{1}{2}}} \sum_{j \geq 0} \left( \sum_{N_1} 2^{-j} \left\| P_{N_1} \left( \frac{\widehat{h}}{\langle \sigma \rangle^{1/2}} \right)^\vee \right\|_{L^4_{x,t}}^2 \|P_{2^{-j}N_1} u\|_{L^4_{x,t}}^2 \right)^{\frac{1}{2}} \\ &\lesssim \|w\|_{X^{0, \frac{1}{2}}} \left( \sum_{N_1} \left\| P_{N_1} \left( \frac{\widehat{h}}{\langle \sigma \rangle^{1/2}} \right)^\vee \right\|_{L^4_{x,t}}^2 \right)^{\frac{1}{2}} \|u\|_{L^4_{x,t}}. \end{aligned} \tag{3-33}$$

In the second case, we bound  $|I_{\mathfrak{B}}|$  by

$$|I_{\mathfrak{B}}| \leq \|w\|_{X^{0, \frac{1}{2}}} \sum_{j \geq 0} \left( \sum_{N_1} N_1^{-2} (2^{-j} N_1 N_1)^{-1} \left\| P_{N_1} \left( \partial_x P_{+hi} P_{2^{-j}N_1} \left( \frac{\widehat{h}}{\langle \sigma \rangle^{1/2}} \right)^\vee P_+ \partial_x P_{N_1} \tilde{u} \right) \right\|_{L^2_{x,t}}^2 \right)^{\frac{1}{2}}$$

so that

$$\begin{aligned} |I_{\mathfrak{B}}| &\lesssim \|w\|_{X^{0, \frac{1}{2}}} \sum_{j \geq 0} \left( \sum_{N_1} 2^{-j} \left\| P_{2^{-j}N_1} P_{+hi} \left( \frac{\widehat{h}}{\langle \sigma \rangle^{1/2}} \right)^\vee \right\|_{L^4_{x,t}}^2 \|P_{N_1} u\|_{L^4_{x,t}}^2 \right)^{\frac{1}{2}} \\ &\lesssim \|w\|_{X^{0, \frac{1}{2}}} \sum_{j \geq 0} 2^{-\frac{j}{2}} \left( \sum_{N_1} \left\| P_{2^{-j}N_1} P_{+hi} \left( \frac{\widehat{h}}{\langle \sigma \rangle^{1/2}} \right)^\vee \right\|_{L^4_{x,t}}^2 \right)^{\frac{1}{2}} \|u\|_{L^4_{x,t}} \\ &\lesssim \|w\|_{X^{0, \frac{1}{2}}} \left( \sum_{N_1} \left\| P_{N_1} \left( \frac{\widehat{h}}{\langle \sigma \rangle^{1/2}} \right)^\vee \right\|_{L^4_{x,t}}^2 \right)^{\frac{1}{2}} \|u\|_{L^4_{x,t}}. \end{aligned} \tag{3-34}$$

In conclusion, we obtain by gathering (3-33)–(3-34) and using estimate (2-9) that

$$|I_{\mathfrak{B}}| \leq \|h\|_{L^2_{x,t}} \|w\|_{X^{0, \frac{1}{2}}} \|u\|_{L^4_{x,t}}. \tag{3-35}$$

*Estimate for  $|I_{\mathfrak{C}}|$ .* First observe that

$$|I_{\mathfrak{C}}| \lesssim \int_{\tilde{\mathfrak{C}}} \frac{|\xi|}{\langle \sigma \rangle^{1/2}} |\widehat{h}(\xi, \tau)| |\xi_1|^{-1} |\widehat{w}(\xi_1, \tau_1)| \frac{|\xi_2|^2 \langle \sigma_2 \rangle}{\langle \sigma_2 \rangle |\xi_2|} |\widehat{u}(\xi_2, \tau_2)| d\nu, \tag{3-36}$$

where

$$\tilde{\mathfrak{C}} = \left\{ (\xi, \xi_1, \tau, \tau_1) \in \mathcal{D} \mid (\xi, \xi_1, \tau, \tau_1) \in \bigcup_{N, N_2} \mathcal{C}_{N, N_2} \right\}.$$

Since  $|\sigma_2| > |\sigma|$  and  $|\sigma_2| > |\sigma_1|$  in  $\tilde{\mathfrak{C}}$ , it follows from (3-28) that  $|\sigma_2| \gtrsim |\xi \xi_2|$ . Then

$$|\xi \xi_1^{-1} \xi_2^2 \langle \sigma_2 \rangle^{-1}| \lesssim 1 \tag{3-37}$$

holds in  $\tilde{\mathfrak{C}}$  so that, using Hölder’s inequality and estimate (2-9), we deduce

$$\begin{aligned} |I_{\mathfrak{C}}| &\lesssim \int_{\tilde{\mathfrak{C}}} \frac{|\widehat{h}(\xi, \tau)|}{\langle \sigma \rangle^{1/2}} |\widehat{w}(\xi_1, \tau_1)| \frac{\langle \sigma_2 \rangle}{|\xi_2|} |\widehat{u}(\xi_2, \tau_2)| d\nu \\ &\lesssim \left\| \left( \frac{\widehat{h}}{\langle \sigma \rangle^{1/2}} \right)^\vee \right\|_{L^4_{x,t}} \|(|\widehat{w}|)^\vee\|_{L^4_{x,t}} \|u\|_{X^{-1,1}} \lesssim \|h\|_{L^2_{x,t}} \|w\|_{X^{0, \frac{3}{8}}} \|u\|_{X^{-1,1}}. \end{aligned} \tag{3-38}$$

Therefore, estimates (3-29), (3-32), (3-35), and (3-38) imply estimate (3-22), which concludes the proof of estimate (3-20).

To prove estimate (3-21), we also proceed by duality. Then it is sufficient to show that

$$|J| \lesssim \left( \sum_N \|g_N\|_{L_\xi^2 L_\tau^\infty}^2 \right)^{\frac{1}{2}} \|w\|_{X^{0,\frac{1}{2}}} (\|u\|_{L_{x,t}^2} + \|u\|_{L_{x,t}^4} + \|u\|_{X^{-1,1}}), \tag{3-39}$$

where

$$J = \sum_N \int_{\mathcal{D}} \frac{\xi}{\langle \sigma \rangle} g_N(\xi, \tau) \phi_N(\xi) \xi_1^{-1} \widehat{w}(\xi_1, \tau_1) \xi_2 \widehat{u}(\xi_2, \tau_2) d\nu,$$

and  $d\nu$  and  $\mathcal{D}$  are defined in (3-24) and (3-25). As in the case of  $I$ , we can also assume that  $|\xi_2| \geq 1$ . By using dyadic decompositions as in (3-27),  $J$  can be rewritten as

$$J = \sum_{N, N_1, N_2} J_{N, N_1, N_2},$$

where

$$J_{N, N_1, N_2} := \int_{\mathcal{D}} \frac{\xi}{\langle \sigma \rangle} \phi_N(\xi) g_N(\xi, \tau) \xi_1^{-1} \widehat{P_{N_1} w}(\xi_1, \tau_1) \xi_2 \widehat{P_{N_2} u}(\xi_2, \tau_2) d\nu,$$

and the dyadic numbers  $N, N_1$ , and  $N_2$  range from 1 to  $+\infty$ . Moreover, we will denote by  $J_{N, N_1, N_2}^{\mathcal{A}_{N, N_2}}$ ,  $J_{N, N_1, N_2}^{\mathcal{B}_{N, N_2}}$ , and  $J_{N, N_1, N_2}^{\mathcal{C}_{N, N_2}}$  the restriction of  $J_{N, N_1, N_2}$  to the regions  $\mathcal{A}_{N, N_2}$ ,  $\mathcal{B}_{N, N_2}$ , and  $\mathcal{C}_{N, N_2}$  defined in (3-28). Then it follows that

$$|J| \leq |J_{\mathcal{A}}| + |J_{\mathcal{B}}| + |J_{\mathcal{C}}|, \tag{3-40}$$

where

$$J_{\mathcal{A}} := \sum_{N, N_1, N_2} J_{N, N_1, N_2}^{\mathcal{A}_{N, N_2}}, \quad J_{\mathcal{B}} := \sum_{N, N_1, N_2} J_{N, N_1, N_2}^{\mathcal{B}_{N, N_2}}, \quad J_{\mathcal{C}} := \sum_{N, N_1, N_2} J_{N, N_1, N_2}^{\mathcal{C}_{N, N_2}}$$

so that it suffices to estimate  $|J_{\mathcal{A}}|$ ,  $|J_{\mathcal{B}}|$ , and  $|J_{\mathcal{C}}|$ .

*Estimate for  $|J_{\mathcal{A}}|$ .* To estimate  $|J_{\mathcal{A}}|$ , we divide each region  $\mathcal{A}_{N, N_2}$  into disjoint subregions

$$\mathcal{A}_{N, N_2}^q = \{ (\xi, \xi_1, \tau, \tau_1) \in \mathcal{A}_{N, N_2} \mid 2^{q-3} N N_2 \leq |\sigma| < 2^{q-2} N N_2 \}$$

for  $q \in \mathbb{Z}_+$ . Thus, if  $J_{N, N_1, N_2}^{\mathcal{A}_{N, N_2}^q}$  denotes the restriction of  $J_{N, N_1, N_2}^{\mathcal{A}_{N, N_2}}$  to each of these regions, we have

$$J_{\mathcal{A}} = \sum_{q \geq 0} \sum_{N, N_1, N_2} J_{N, N_1, N_2}^{\mathcal{A}_{N, N_2}^q}.$$

In the case of high-low interactions, we deduce by using the Plancherel identity and the Cauchy–Schwarz and Minkowski inequalities that

$$|J_{\mathcal{A}}| \leq \sum_{q \geq 0} \sum_{N_1} \sum_{N_2 \leq N_1} \|g_{N_1} \chi_{\{|\sigma| \sim 2^q N_1 N_2\}}\|_{L_{\xi, \tau}^2} \times (2^q N_1 N_2)^{-1} N_1 \|\partial_x^{-1} P_{N_1} w P_- \partial_x P_{N_2} u\|_{L_{x,t}^2}.$$

Moreover, we get from Hölder’s inequality

$$\|g_{N_1} \chi_{\{|\sigma| \sim 2^q N_1 N_2\}}\|_{L_{\xi, \tau}^2} \lesssim (2^q N N_2)^{\frac{1}{2}} \|g_{N_1}\|_{L_\xi^2 L_\tau^\infty}$$

so that the Cauchy–Schwarz inequality yields

$$\begin{aligned} |J_{\mathcal{A}}| &\lesssim \sum_{N_1} \sum_{N_2 \leq N_1} (N_2 N_1^{-1})^{\frac{1}{2}} \|g_{N_1}\|_{L_{\xi}^2 L_{\tau}^{\infty}} \|P_{N_1} w\|_{L_{x,t}^4} \|P_{N_2} u\|_{L_{x,t}^4} \\ &\lesssim \|u\|_{L_{x,t}^4} \sum_{N_1} \|g_{N_1}\|_{L_{\xi}^2 L_{\tau}^{\infty}} \|P_{N_1} w\|_{L_{x,t}^4} \lesssim \left( \sum_{N_1} \|g_{N_1}\|_{L_{\xi}^2 L_{\tau}^{\infty}}^2 \right)^{\frac{1}{2}} \|w\|_{\tilde{L}_{x,t}^4} \|u\|_{L_{x,t}^4}. \end{aligned} \quad (3-41)$$

In the high-high interaction case, it follows from the Minkowski and Cauchy–Schwarz inequalities that

$$|J_{\mathcal{A}}| \leq \sum_{q \geq 0} \sum_{N_1} \sum_{N \leq N_1} \|g_N \chi_{\{|\sigma| \sim 2^q N N_1\}}\|_{L_{\xi,\tau}^2} \times (2^q N N_1)^{-1} N \|\partial_x^{-1} P_{N_1} w P_{-} \partial_x P_{N_1} u\|_{L_{x,t}^2}.$$

Moreover, we deduce from Hölder’s inequality that

$$\|g_N \chi_{\{|\sigma| \sim 2^q N N_1\}}\|_{L_{\xi,\tau}^2} \lesssim (2^q N N_1)^{\frac{1}{2}} \|g_N\|_{L_{\xi}^2 L_{\tau}^{\infty}}.$$

Then the Cauchy–Schwarz inequality implies that

$$\begin{aligned} |J_{\mathcal{A}}| &\lesssim \sum_{j \geq 0} \sum_{N_1} (N_1^{-1} 2^{-j} N_1)^{\frac{1}{2}} \|g_{2^{-j} N_1}\|_{L_{\xi}^2 L_{\tau}^{\infty}} \|P_{N_1} w\|_{L_{x,t}^4} \|P_{N_1} u\|_{L_{x,t}^4} \\ &\lesssim \sum_{j \geq 0} 2^{-\frac{j}{2}} \left( \sum_{N_1} \|g_{2^{-j} N_1}\|_{L_{\xi}^2 L_{\tau}^{\infty}}^2 \right)^{\frac{1}{2}} \left( \sum_{N_1} \|P_{N_1} w\|_{L_{x,t}^4}^2 \right)^{\frac{1}{2}} \|u\|_{L_{x,t}^4} \\ &\lesssim \left( \sum_{N_1} \|g_{N_1}\|_{L_{\xi}^2 L_{\tau}^{\infty}}^2 \right)^{\frac{1}{2}} \|w\|_{\tilde{L}_{x,t}^4} \|u\|_{L_{x,t}^4}. \end{aligned} \quad (3-42)$$

Then estimates (2-9), (3-41), and (3-42) yield

$$|J_{\mathcal{A}}| \lesssim \left( \sum_N \|g_N\|_{L_{\xi}^2 L_{\tau}^{\infty}}^2 \right)^{\frac{1}{2}} \|w\|_{X^{0, \frac{3}{8}}} \|u\|_{L_{x,t}^4}. \quad (3-43)$$

*Estimate for  $|J_{\mathcal{B}}|$  and  $|J_{\mathcal{C}}|$ .* Arguing as in the proof of (3-20), it is deduced that

$$|J_{\mathcal{B}}| + |J_{\mathcal{C}}| \lesssim \left( \left\| \left( \frac{g}{\langle \sigma \rangle} \right)^{\vee} \right\|_{\tilde{L}_{x,t}^4} + \left\| \left( \frac{|g|}{\langle \sigma \rangle} \right)^{\vee} \right\|_{\tilde{L}_{x,t}^4} \right) \|w\|_{X^{0, \frac{1}{2}}} (\|u\|_{L_{x,t}^4} + \|u\|_{X^{-1,1}}),$$

where  $g = \sum_N \phi_N g_N$ . Moreover, estimate (2-9) and Hölder’s inequality imply

$$\left\| \left( \frac{g}{\langle \sigma \rangle} \right)^{\vee} \right\|_{\tilde{L}_{x,t}^4} + \left\| \left( \frac{|g|}{\langle \sigma \rangle} \right)^{\vee} \right\|_{\tilde{L}_{x,t}^4} \lesssim \left\| \langle \sigma \rangle^{-\frac{5}{8}} \sum_N \phi_N g_N \right\|_{L_{\xi,\tau}^2} \lesssim \left( \sum_N \|\langle \sigma \rangle^{-\frac{5}{8}} g_N\|_{L_{\xi,\tau}^2}^2 \right)^{\frac{1}{2}} \lesssim \left( \sum_N \|g_N\|_{L_{\xi}^2 L_{\tau}^{\infty}}^2 \right)^{\frac{1}{2}}$$

so that

$$|J_{\mathcal{B}}| + |J_{\mathcal{C}}| \lesssim \left( \sum_N \|g_N\|_{L_{\xi}^2 L_{\tau}^{\infty}}^2 \right)^{\frac{1}{2}} \|w\|_{X^{0, \frac{1}{2}}} (\|u\|_{L_{x,t}^4} + \|u\|_{X^{-1,1}}). \quad (3-44)$$

Finally (3-40), (3-43), and (3-44) imply (3-39), which concludes the proof of estimate (3-21).  $\square$



**Lemma 3.7.** *Let  $0 < T \leq 1$ ,  $s \geq 0$ ,  $u_1, u_2 \in L^\infty(\mathbb{R}; L^2(\mathbb{R})) \cap L^4(\mathbb{R}^2)$  be supported in the time interval  $[-2T, 2T]$ , and  $F_1, F_2$  be some spatial primitives of  $u_1$  and  $u_2$ , respectively. Then*

$$\|\partial_x P_{+hi}(P_{lo}e^{-\frac{i}{2}F_1}P_- \partial_x u_1)\|_{\tilde{Z}^{s,-1}} + \|\partial_x P_{+hi}(P_{lo}e^{-\frac{i}{2}F_1}P_- \partial_x u_1)\|_{X^{s,-\frac{1}{2}}} \lesssim \|u_1\|_{L^4_{x,t}}^2, \tag{3-45}$$

and

$$\begin{aligned} \|\partial_x P_{+hi}(P_{lo}(e^{-\frac{i}{2}F_1} - e^{-\frac{i}{2}F_2})P_- \partial_x u_2)\|_{\tilde{Z}^{s,-1}} + \|\partial_x P_{+hi}(P_{lo}(e^{-\frac{i}{2}F_1} - e^{-\frac{i}{2}F_2})P_- \partial_x u_2)\|_{X^{s,-\frac{1}{2}}} \\ \lesssim (\|u_1 - u_2\|_{L_t^\infty L_x^2} + \|e^{-\frac{i}{2}F_1} - e^{-\frac{i}{2}F_2}\|_{L_{x,t}^\infty} \|u_2\|_{L_t^\infty L_x^2}) \|u_2\|_{L^4_{x,t}}. \end{aligned} \tag{3-46}$$

*Proof.* We deduce from the Cauchy–Schwarz inequality, the Sobolev embedding  $\|f\|_{H_t^{-1/2+\varepsilon}} \lesssim \|f\|_{L_t^{1+\varepsilon}}$  with  $1 + \varepsilon' = 1/(1 - \varepsilon)$ , and the Minkowski inequality that

$$\begin{aligned} \|f\|_{\tilde{Z}^{s,-1}} + \|f\|_{X^{s,-\frac{1}{2}}} &\lesssim \|f\|_{X^{s,-\frac{1}{2}+\varepsilon}} = \left\| \left( J_x^s U(-t) f \right)^{\wedge_x}(\xi) \right\|_{H_t^{-\frac{1}{2}+\varepsilon}} \Big|_{L_\xi^2} \\ &\lesssim \left\| \left( J_x^s U(-t) f \right)^{\wedge_x}(\xi) \right\|_{L_t^{1+\varepsilon'}} \Big|_{L_\xi^2} \lesssim \|f\|_{L_t^{1+\varepsilon'} H_x^s}. \end{aligned} \tag{3-47}$$

On the other hand, it follows from the frequency localization that

$$\partial_x P_{+hi}(P_{lo}e^{-\frac{i}{2}F}P_- \partial_x u) = \partial_x P_{+LO}(P_{lo}e^{-\frac{i}{2}F}P_{-LO} \partial_x u).$$

Therefore, by using (3-47), Bernstein’s inequalities, and estimate (2-12), we can bound the left-hand side of (3-45) by

$$\|P_{+LO}(P_{lo}e^{-\frac{i}{2}F}P_{-LO} \partial_x u)\|_{L_t^{1+\varepsilon'} L_x^2} \lesssim T^\gamma \|\partial_x e^{-\frac{i}{2}F}\|_{L_{x,t}^4} \|u\|_{L_{x,t}^4} \tag{3-48}$$

with  $\frac{1}{\gamma} = \frac{1}{2} - \varepsilon'$ , which concludes the proof of estimate (3-45), recalling that  $\partial_x F = u$  and  $0 < T \leq 1$ . Estimate (3-46) can be proved exactly as above by recalling (2-19).  $\square$

*Proof of Proposition 3.4.* Let  $0 \leq s \leq \frac{1}{2}$ ,  $0 < T \leq 1$ , and  $\tilde{u}$  and  $\tilde{w}$  be extensions of  $u$  and  $w$  such that  $\|\tilde{u}\|_{X^{-1,1}} \leq 2\|u\|_{X_T^{-1,1}}$  and  $\|\tilde{w}\|_{X^{s,1/2}} \leq 2\|w\|_{X_T^{s,1/2}}$ . By the Duhamel principle, the integral formulation associated with (3-4) reads

$$\begin{aligned} w(t) = \eta(t) U(t) w(0) - \eta(t) \int_0^t U(t-t') \partial_x P_{+hi}(\eta_T \partial_x^{-1} \tilde{w} P_-(\eta_T \partial_x u))(t') dt' \\ - \eta(t) \int_0^t U(t-t') \partial_x P_{+hi}(P_{lo}(\eta_T e^{-\frac{i}{2}\tilde{F}})P_-(\eta_T \partial_x \tilde{u}))(t') dt' \end{aligned}$$

for  $0 < t \leq T \leq 1$ . Therefore, we deduce gathering estimates (2-5), (2-7), (3-20), (3-21), and (3-45) that

$$\|w\|_{Y_T^s} \lesssim \|w(0)\|_{H^s} + \|u\|_{L_{x,T}^4}^2 + \|w\|_{X_T^{s,\frac{1}{2}}} (\|u\|_{L_T^\infty L_x^2} + \|u\|_{L_{x,T}^4} + \|u\|_{X_T^{-1,1}}).$$

This concludes the proof of estimate (3-19) since

$$\|w(0)\|_{H^s} \lesssim \|J_x^s(e^{-\frac{i}{2}F(\cdot,0)}u_0)\|_{L^2} \lesssim (1 + \|u_0\|_{L^2})\|u_0\|_{H^s} \tag{3-49}$$

follows from estimate (2-13) and the fact that  $0 \leq s \leq \frac{1}{2}$ .  $\square$

**4. Proof of Theorem 1.1**

First, we can always assume that we deal with data having small  $L^2(\mathbb{R})$ -norm. Indeed, if  $u$  is a solution to the IVP (1-1) on the time interval  $[0, T]$ , then for every  $0 < \lambda < \infty$ ,  $u_\lambda(x, t) = \lambda u(\lambda x, \lambda^2 t)$  is also a solution to the equation in (1-1) on the time interval  $[0, \lambda^{-2}T]$  with initial data  $u_{0,\lambda} = \lambda u_0(\lambda, \cdot)$ . For  $\varepsilon > 0$  let us denote by  $B_\varepsilon$  the ball of  $L^2(\mathbb{R})$  centered at the origin with radius  $\varepsilon$ . Since  $\|u_\lambda(\cdot, 0)\|_{L^2} = \lambda^{1/2} \|u_0\|_{L^2}$ , we see that we can force  $u_{0,\lambda}$  to belong to  $B_\varepsilon$  by choosing  $\lambda \sim \min(\varepsilon^2 \|u_0\|_{L^2}^{-2}, 1)$ . Therefore, the existence and uniqueness of a solution of (1-1) on the time interval  $[0, 1]$  for small  $L^2(\mathbb{R})$  initial data will ensure the existence of a unique solution  $u$  to (1-1) for arbitrary large  $L^2(\mathbb{R})$  initial data on the time interval  $T \sim \lambda^2 \sim \min(\|u_0\|_{L^2}^{-4}, 1)$ . Using the conservation of the  $L^2(\mathbb{R})$ -norm, this will lead to global well-posedness in  $L^2(\mathbb{R})$ .

**4A. Uniform bound for small initial data.** First we begin by deriving *a priori* estimates on smooth solutions associated with initial data  $u_0 \in H^\infty(\mathbb{R})$  that are small in  $L^2(\mathbb{R})$ . It is known from the classical well-posedness theory [Iório 1986] that such initial data gives rise to a global solution  $u \in C(\mathbb{R}; H^\infty(\mathbb{R}))$  to the Cauchy problem (1-1). Setting  $0 < T \leq 1$ ,

$$N_T^s(u) := \max\left(\|u\|_{L_T^\infty H^s}, \|J_x^s u\|_{L_{x,T}^4}, \|w\|_{X_T^{s, \frac{1}{2}}}\right), \tag{4-1}$$

and it follows from the smoothness of  $u$  that  $T \mapsto N_T^s(u)$  is continuous and nondecreasing on  $\mathbb{R}_+^*$ . Moreover, from (3-4), the linear estimate (2-7), (3-49), and (3-7), we infer that  $\lim_{T \rightarrow 0+} N_T^s(u) \lesssim (1 + \|u_0\|_{L^2}) \|u_0\|_{H^s}$ . On the other hand, combining (3-7)–(3-8) and (3-19) and the conservation of the  $L^2$ -norm, we infer that

$$N_T^0(u) \lesssim (1 + \|u_0\|_{L^2}) \|u_0\|_{L^2} + (N_T^0(u))^2 + (N_T^0(u))^3.$$

By continuity, this ensures there exist  $\varepsilon_0 > 0$  and  $C_0 > 0$  such that  $N_1^0(u) \leq C_0 \varepsilon$  given  $\|u_0\|_{L^2} \leq \varepsilon \leq \varepsilon_0$ . Finally, again using (3-7)–(3-8) and (3-19), this leads to  $N_1^s(u) \lesssim \|u_0\|_{H^s}$  given  $\|u_0\|_{L^2} \leq \varepsilon \leq \varepsilon_0$ .

**4B. Lipschitz bound for initial data having the same low-frequency part.** To prove the uniqueness as well as the continuity of the solution, we will derive a Lipschitz bound on the solution map on some affine subspaces of  $H^s(\mathbb{R})$  with values in  $L_T^\infty H^s(\mathbb{R})$ . We know from [Koch and Tzvetkov 2003] that such a Lipschitz bound does not exist in general in  $H^s(\mathbb{R})$ . Here we will restrict ourselves to solutions emanating from initial data having the same low-frequency part. This is clearly sufficient to get uniqueness, and it will turn out to be sufficient to get the continuity of the solution as well as the continuity of the flow map. Let  $\varphi_1, \varphi_2 \in B_\varepsilon \cap H^s(\mathbb{R})$ ,  $s \geq 0$ , such that  $P_{LO}\varphi_1 = P_{LO}\varphi_2$ , and let  $u_1, u_2$  be two solutions to (1-1) emanating from  $\varphi_1$  and  $\varphi_2$ , respectively, that satisfy (7-1) on the time interval  $[0, T]$ ,  $0 < T < 1$ . We also assume that the primitives  $F_1 := F[u_1]$  and  $F_2 := F[u_2]$  of  $u_1$  and  $u_2$ , respectively, are such that the associated gauge functions  $W_1, w_1$  and  $W_2, w_2$ , respectively, constructed in Section 3A, satisfy (7-2). Finally, we assume that

$$N_T^0(u_i) \leq C_0 \varepsilon \leq C_0 \varepsilon_0. \tag{4-2}$$

First, by construction, we observe that since  $F(x) - F(y) = \int_x^y u(z) dz$ ,

$$P_{LO} \int_y^x u dz = P_{LO}(F(x) - F(y)) = P_{LO}F(x) - F(y)$$

holds. On the other hand, since  $P_{LO}$  and  $\partial_x$  do commute, we have  $\partial_x P_{LO}F = P_{LO}u$  and, by integrating,  $\int_y^x P_{LO}u dz = P_{LO}F(x) - P_{LO}F(y)$ . Gathering these two identities, we get

$$\int_y^x P_{LO}u dz - P_{LO} \int_y^x u dz = F(y) - P_{LO}F(y) = P_{HI}F(y),$$

which leads to

$$P_{lo} \int_y^x u dz = P_{lo} \int_y^x P_{LO}u dz.$$

We thus infer that

$$\begin{aligned} P_{lo}(F_1 - F_2)(x, 0) &= \int_{\mathbb{R}} \psi(y) P_{lo} \int_y^x (u_1 - u_2)(z, 0) dz dy \\ &= \int_{\mathbb{R}} \psi(y) P_{lo} \int_y^x P_{LO}(\varphi_1(z) - \varphi_2(z)) dz dy = 0. \end{aligned} \quad (4-3)$$

Then we set  $v = u_1 - u_2$ ,  $Z = W_1 - W_2$ , and  $z = w_1 - w_2$ . Obviously,  $z$  satisfies

$$\begin{aligned} \partial_t z - i \partial_x^2 z &= -\partial_x P_{+hi}(W_1 P_- \partial_x v) - \partial_x P_{+hi}(Z P_- \partial_x u_2) \\ &\quad - \partial_x P_{+hi}(P_{lo} e^{-\frac{i}{2}F_1} P_- \partial_x v) - \partial_x P_{+hi}(P_{lo}(e^{-\frac{i}{2}F_1} - e^{-\frac{i}{2}F_2}) P_- \partial_x u_2). \end{aligned} \quad (4-4)$$

Thus, by gathering estimates (2-7), (3-20), (3-21), (3-45), and (3-46), we deduce that

$$\begin{aligned} \|z\|_{Y_1^s} &\lesssim \|z(0)\|_{H^s} + \|w_1\|_{X_1^{s, \frac{1}{2}}} (\|v\|_{X_1^{-1,1}} + \|v\|_{L_{x,1}^4} + \|v\|_{L_1^\infty L_x^2}) + \|v\|_{L_{x,1}^4}^2 \\ &\quad + \|z\|_{X_1^{s, \frac{1}{2}}} (\|u_2\|_{X_1^{-1,1}} + \|u_2\|_{L_{x,1}^4} + \|u_2\|_{L_1^\infty L_x^2}) + (\|v\|_{L_1^\infty L_x^2} + \|e^{-\frac{i}{2}F_1} - e^{-\frac{i}{2}F_2}\|_{L_{x,1}^\infty}) \|u_2\|_{L_{x,1}^4}, \end{aligned}$$

which, recalling (4-1) and (4-2), implies that

$$\|z\|_{Y_1^s} \lesssim \|z(0)\|_{H^s} + \varepsilon (\|v\|_{X_1^{-1,1}} + \|v\|_{L_{x,1}^4} + \|v\|_{L_1^\infty L_x^2}) + \varepsilon \|e^{-\frac{i}{2}F_1} - e^{-\frac{i}{2}F_2}\|_{L_{x,1}^\infty}, \quad (4-5)$$

where, by the mean-value theorem,

$$\begin{aligned} \|z(0)\|_{H^s} &\lesssim \|\varphi_1 - \varphi_2\|_{H^s} (1 + \|\varphi_1\|_{H^s} + \|\varphi_2\|_{L^2}) + \|e^{-\frac{i}{2}F_1(0)} - e^{-\frac{i}{2}F_2(0)}\|_{L^\infty} \|\varphi_1\|_{H^s} (1 + \|\varphi_1\|_{L^2}) \\ &\lesssim \|\varphi_1 - \varphi_2\|_{H^s} + \|F_1(0) - F_2(0)\|_{L^\infty}. \end{aligned} \quad (4-6)$$

On the other hand, the equation for  $v = u_1 - u_2$  reads

$$\partial_t v + \mathcal{H} \partial_x^2 v = \frac{1}{2} \partial_x ((u_1 + u_2)v)$$

so that it is deduced from (3-11), (4-1), and the fractional Leibniz rule that

$$\|v\|_{X_1^{-1,1}} \lesssim \|\partial_t v + \mathcal{H} \partial_x^2 v\|_{L_1^2 H_x^{-1}} + \|v\|_{L_T^\infty L_x^2} \lesssim \varepsilon \|v\|_{L_{x,1}^4} + \|v\|_{L_1^\infty L_x^2}. \quad (4-7)$$

Next, proceeding as in (3-6), we infer that

$$\begin{aligned} P_{+HI}v &= 2i P_{+HI}(e^{\frac{i}{2}F_1}z) + 2i P_{+HI}((e^{\frac{i}{2}F_1} - e^{\frac{i}{2}F_2})w_2) \\ &\quad + 2i P_{+HI}(P_{+hi}e^{\frac{i}{2}F_1}\partial_x P_{+lo}(e^{-\frac{i}{2}F_1} - e^{-\frac{i}{2}F_2})) \\ &\quad + 2i P_{+HI}(P_{+hi}(e^{\frac{i}{2}F_1} - e^{\frac{i}{2}F_2})\partial_x P_{+lo}e^{-\frac{i}{2}F_2}) \\ &\quad + 2i P_{+HI}(P_{+HI}e^{\frac{i}{2}F_1}\partial_x P_{-}(e^{-\frac{i}{2}F_1} - e^{-\frac{i}{2}F_2})) \\ &\quad + 2i P_{+HI}(P_{+HI}(e^{\frac{i}{2}F_1} - e^{\frac{i}{2}F_2})\partial_x P_{-}e^{-\frac{i}{2}F_2}). \end{aligned}$$

Thus, we deduce using estimates (2-14) and (2-19) and arguing as in the proof of Proposition 3.2 that

$$\begin{aligned} \|J_x^s v\|_{L_1^p L_x^q} &\lesssim (\|u_1\|_{L_1^\infty L_x^2} + \|u_2\|_{L_1^\infty L_x^2})\|v\|_{L_1^\infty L_x^2} + (1 + \|u_1\|_{L_1^\infty L_x^2})\|z\|_{Y_1^s} \\ &\quad + (\|v\|_{L_1^\infty L_x^2} + \|e^{\frac{i}{2}F_1} - e^{\frac{i}{2}F_2}\|_{L_{x,1}^\infty} (1 + \|u_1\|_{L_1^\infty L_x^2}))\|w_2\|_{Y_1^s} \\ &\quad + \|u_1\|_{L_1^\infty L_x^2} (\|v\|_{L_1^\infty L_x^2} + \|e^{-\frac{i}{2}F_1} - e^{-\frac{i}{2}F_2}\|_{L_{x,1}^\infty} \|u_1\|_{L_1^\infty L_x^2}) \\ &\quad + \|u_2\|_{L_1^\infty L_x^2} (\|v\|_{L_1^\infty L_x^2} + \|e^{\frac{i}{2}F_1} - e^{\frac{i}{2}F_2}\|_{L_{x,1}^\infty} \|u_1\|_{L_1^\infty L_x^2}) \end{aligned}$$

for  $(p, q) = (\infty, 2)$  or  $(p, q) = (4, 4)$ , which, recalling (4-2), implies that

$$\|J_x^s v\|_{L_1^\infty L_x^2} + \|J_x^s v\|_{L_{x,1}^4} \lesssim \|z\|_{Y_1^s} + \varepsilon \|e^{-\frac{i}{2}F_1} - e^{-\frac{i}{2}F_2}\|_{L_{x,1}^\infty} + \varepsilon \|e^{\frac{i}{2}F_1} - e^{\frac{i}{2}F_2}\|_{L_{x,1}^\infty}. \tag{4-8}$$

Finally, we use the mean-value theorem to get the bound

$$\|e^{\pm\frac{i}{2}F_1} - e^{\pm\frac{i}{2}F_2}\|_{L_{x,1}^\infty} \lesssim \|F_1 - F_2\|_{L_{x,1}^\infty}. \tag{4-9}$$

The following crucial lemma gives an estimate for the right-hand side of (4-9):

**Lemma 4.1.** *It holds that*

$$\|F_1(0) - F_2(0)\|_{L^\infty} \lesssim \|\varphi_1 - \varphi_2\|_{L^2} \tag{4-10}$$

and

$$\|F_1 - F_2\|_{L_{x,1}^\infty} \lesssim \|v\|_{L_1^\infty L_x^2}. \tag{4-11}$$

*Proof.* Equation (4-10) clearly follows from (4-3) together with Bernstein's inequality. To prove (4-11), we set  $G = F_1 - F_2$ ,  $G_{lo} = P_{lo}G$ , and  $G_{hi} = P_{hi}G$ . Then

$$\|G\|_{L_{x,1}^\infty} \leq \|G_{lo}\|_{L_{x,1}^\infty} + \|G_{hi}\|_{L_{x,1}^\infty}. \tag{4-12}$$

Observe, from the Duhamel principle and (4-3), that  $G_{lo}$  satisfies

$$G_{lo} = \frac{1}{2} \int_0^t U(t - \tau) P_{lo}((u_1 + u_2)v)(\tau) d\tau.$$

Therefore, using Bernstein and Hölder's inequalities, it follows that

$$\|G_{lo}\|_{L_{x,1}^\infty} \lesssim \|(u_1 + u_2)v\|_{L_1^\infty L_x^1} \lesssim (\|u_1\|_{L_1^\infty L_x^2} + \|u_2\|_{L_1^\infty L_x^2})\|v\|_{L_1^\infty L_x^2}. \tag{4-13}$$

On the other hand, Bernstein's inequality ensures that

$$\|G_{hi}\|_{L_{x,1}^\infty} \lesssim \|\partial_x G_{hi}\|_{L_1^\infty L_x^2} \lesssim \|v\|_{L_1^\infty L_x^2} \quad (4-14)$$

since  $\partial_x G = v$ . The proof of Lemma 4.1 is concluded gathering (4-2) and (4-12)–(4-14).  $\square$

Finally, estimates (4-5)–(4-11) lead to

$$\|z\|_{Y_1^s} + \|v\|_{X_1^{s-1,1}} + \|v\|_{L_1^\infty H_x^s} + \|J_x^s v\|_{L_{x,1}^4} \lesssim \|\varphi_1 - \varphi_2\|_{H^s} + \varepsilon \left( \|z\|_{Y_1^s} + \|v\|_{X_1^{s-1,1}} + \|v\|_{L_1^\infty H_x^s} + \|J_x^s v\|_{L_{x,1}^4} \right).$$

Therefore, we conclude that there exists  $0 < \varepsilon_1 \leq \varepsilon_0$  such that

$$\|z\|_{Y_1^s} + \|v\|_{X_1^{s-1,1}} + \|v\|_{L_1^\infty H_x^s} + \|J_x^s v\|_{L_{x,1}^4} \lesssim \|\varphi_1 - \varphi_2\|_{H^s}, \quad (4-15)$$

provided  $u_1$  and  $u_2$  satisfy (4-2) with  $0 < \varepsilon \leq \varepsilon_1$ .

**4C. Well-posedness.** Let  $u_0 \in B_{\varepsilon_1} \cap H^s(\mathbb{R})$ , and consider the sequence of initial data  $\{u_0^j\} \subset H^\infty(\mathbb{R})$ , defined by

$$u_0^j = \mathcal{F}_x^{-1}(\chi_{|[-j,j]} \mathcal{F}_x u_0) \quad \text{for all } j \geq 20. \quad (4-16)$$

Clearly  $\{u_0^j\}$  converges to  $u_0$  in  $H^s(\mathbb{R})$ . By the classical well-posedness theory, the associated sequence of solutions  $\{u^j\}$  is a subset of  $C([0, 1]; H^\infty(\mathbb{R}))$ , and according to Section 4A, it satisfies  $N_1^s(u^j) \leq C_0 \varepsilon_1$ . Moreover, since  $P_{LO} u_0^j = P_{LO} u_0$  for all  $j \geq 20$ , it follows from the preceding subsection that

$$\|u^j - u^{j'}\|_{L_1^\infty H_x^s} + \|u^j - u^{j'}\|_{L_1^4 W_x^{s,4}} + \|w^j - w^{j'}\|_{X_1^{0, \frac{1}{2}}} \lesssim \|u_0^j - u_0^{j'}\|_{H_x^s}. \quad (4-17)$$

Therefore, the sequence  $\{u^j\}$  converges strongly in  $L_1^\infty H^s(\mathbb{R}) \cap L_1^4 W_x^{s,4}$  to some function

$$u \in C([0, 1]; H^s(\mathbb{R})),$$

and  $\{w_j\}_{j \geq 4}$  converges strongly to some function  $w$  in  $X_1^{s, 1/2}$ . Thanks to these strong convergences, it is easy to check that  $u$  is a solution to (1-1) emanating from  $u_0$  and that  $w = \partial_x P_{+hi}(e^{-\frac{i}{2}F[u]})$ . Moreover, from the conservation of the  $L^2(\mathbb{R})$ -norm,  $u \in C_b(\mathbb{R}; L^2(\mathbb{R})) \cap C(\mathbb{R}; H^s(\mathbb{R}))$ .

Now let  $\tilde{u}$  be another solution of (1-1) on  $[0, T]$  emanating from  $u_0$  belonging to the same class of regularity as  $u$ . By again using the scaling argument we can always assume that  $\|\tilde{u}\|_{L_T^\infty L_x^2} + \|\tilde{u}\|_{L_{x,T}^4} \leq C_0 \varepsilon_1$ . Moreover, setting  $\tilde{w} := P_{+hi}(e^{-iF[\tilde{u}]})$ , by the Lebesgue monotone convergence theorem, there exists  $N > 0$  such that  $\|P_{\geq N} \tilde{w}\|_{X_T^{0, 1/2}} \leq C_0 \varepsilon_1 / 2$ . On the other hand, using Lemmas 2.1–2.2, it is easy to check that

$$\begin{aligned} \|(1 - P_{\geq N}) \tilde{w}\|_{X_T^{0, \frac{1}{2}}} &\lesssim \|u_0\|_{L^2} + NT^{\frac{1}{4}} \|\tilde{u}\|_{L_{x,T}^4} \|\tilde{w}\|_{L_{x,T}^4} + \|\tilde{u}\|_{L_{x,T}^4}^2 \\ &\lesssim \|u_0\|_{L^2} + NT^{\frac{1}{4}} \|\tilde{w}\|_{X_T^{0, \frac{1}{2}}} \|\tilde{u}\|_{L_{x,T}^4} + \|\tilde{u}\|_{L_{x,T}^4}^2. \end{aligned}$$

Therefore, for  $T > 0$  small enough, we can require that  $\tilde{u}$  satisfies the smallness condition (4-2) with  $\varepsilon_1$ , and thus by (4-15),  $\tilde{u} \equiv u$  on  $[0, T]$ . This proves the uniqueness result for initial data belonging to  $B_{\varepsilon_1}$ .

Next we turn to the continuity of the flow map. Fix  $u_0 \in B_{\varepsilon_1}$  and  $\lambda > 0$  and consider the emanating solution  $u \in C([0, 1]; H^s(\mathbb{R}))$ . We will prove that if  $v_0 \in B_{\varepsilon_1}$  satisfies  $\|u_0 - v_0\|_{H^s} \leq \delta$ , where  $\delta$  will be fixed later, then the solution  $v$  emanating from  $v_0$  satisfies

$$\|u - v\|_{L_1^\infty H_x^s} \leq \lambda. \tag{4-18}$$

For  $j \geq 1$ , let  $u_0^j$  and  $v_0^j$  be constructed as in (4-16), and denote by  $u^j$  and  $v^j$  the solutions emanating from  $u_0^j$  and  $v_0^j$ . Then it follows from the triangular inequality that

$$\|u - v\|_{L_1^\infty H_x^s} \leq \|u - u^j\|_{L_1^\infty H_x^s} + \|u^j - v^j\|_{L_1^\infty H_x^s} + \|v - v^j\|_{L_1^\infty H_x^s}. \tag{4-19}$$

First, according to (4-17), we can choose  $j_0$  large enough so that

$$\|u - u^{j_0}\|_{L_1^\infty H_x^s} + \|v - v^{j_0}\|_{L_1^\infty H_x^s} \leq \frac{2}{3}\lambda.$$

Second, from the definition of  $u_0^j$  and  $v_0^j$  in (4-16), we infer that

$$\|u_0^j - v_0^j\|_{H^3} \leq j^{3-s} \|u_0 - v_0\|_{H^s} \leq j^{3-s} \delta.$$

Therefore, by using the continuity of the flow map for smooth initial data, we can choose  $\delta > 0$  such that

$$\|u^{j_0} - v^{j_0}\|_{L_1^\infty H_x^s} \leq \frac{\lambda}{3}.$$

This concludes the proof of Theorem 1.1.

### 5. Improvement of the uniqueness result for $s > 0$

Now we prove that uniqueness holds for initial data  $u_0 \in H^s(\mathbb{R})$ ,  $s > 0$ , in the class  $u \in L_T^\infty H_x^s \cap L_T^4 W_x^{s,4}$ . The great interest of this result is that we no longer assume any condition on the gauge transform of  $u$ . Moreover, when  $s > \frac{1}{4}$ , the Sobolev embedding  $L_T^\infty H_x^s \hookrightarrow L_T^4 W_x^{0+,4}$  ensures that uniqueness holds in  $L_T^\infty H_x^s$ , and thus, the Benjamin–Ono equation is unconditionally well posed in  $H^s(\mathbb{R})$  for  $s > \frac{1}{4}$ .

According to the uniqueness result (i) of Theorem 1.1, it suffices to prove that for any solution  $u$  to (1-1) that belongs to  $L_T^\infty H_x^s \cap L_T^4 W_x^{s,4}$ , the associated gauge function  $w = \partial_x P_{hi}(e^{-\frac{i}{2}F[u]})$  belongs to  $X_T^{0, \frac{1}{2}}$ . The proof is based on the following bilinear estimate that is shown in the Appendix:

**Proposition 5.1.** *Let  $s > 0$ . Then there exist  $0 < \delta < s/10$  and  $\theta \in (\frac{1}{2}, 1)$ , let us say  $\theta = \frac{1}{2} + \delta$ , such that*

$$\|P_{+hi}(WP_- \partial_x u)\|_{X^{\frac{1}{2}, -\frac{1}{2}+2\delta}} \lesssim \|W\|_{X^{\frac{1}{2}, \frac{1}{2}+\delta}} (\|J^s u\|_{L_{x,t}^2} + \|J^s u\|_{L_{x,t}^4} + \|u\|_{X^{s-\theta, \theta}}). \tag{5-1}$$

First note that by the same scaling argument as in Section 4C, for any given  $\varepsilon > 0$ , we can always assume that  $\|J^s u\|_{L_T^\infty L_x^2} + \|J^s u\|_{L_{Tx}^4} \leq \varepsilon$ , and by (3-7) it follows that  $\|u\|_{X_T^{s-\theta, \theta}} \lesssim \varepsilon$  for  $0 \leq \theta \leq 1$ .

Since  $u \in L^\infty([0, T]; H^s(\mathbb{R})) \cap L_T^4 W_x^{s,4}$  and satisfies (1-1), it follows that  $u_t \in L^\infty([0, T]; H^{s-2}(\mathbb{R}))$ . Therefore,  $F := F[u] \in L^\infty([0, T]; H_{loc}^{s+1})$ , and  $\partial_t F \in L^\infty([0, T]; H_{loc}^{s-1})$ . It ensures that

$$W := P_{hi}(e^{-\frac{i}{2}F}) \in L^\infty([0, T]; H^{s+1}(\mathbb{R})) \cap L_T^4 W_x^{s+1,4} \hookrightarrow X_T^{1,0}, \tag{5-2}$$

$e^{-\frac{i}{2}F} \in L^\infty([0, T]; H_{\text{loc}}^{s+1})$ , and the following calculations are thus justified:

$$\begin{aligned} \partial_t W &= \partial_t P_+(e^{-\frac{i}{2}F}) = -\frac{1}{2}i P_{hi}(F_t e^{-\frac{i}{2}F}) \\ &= -\frac{1}{2}i P_{hi}(e^{-\frac{i}{2}F}(-\mathcal{H}F_{xx} + \frac{1}{2}F_x^2)), \\ \partial_{xx} W &= \partial_{xx} P_{hi}(e^{-\frac{i}{2}F}) = P_{hi}(e^{-\frac{i}{2}F}(-\frac{1}{4}F_x^2 - \frac{i}{2}F_{xx})). \end{aligned}$$

It follows that  $W$  satisfies, at least in a distributional sense,

$$\begin{cases} \partial_t W - i \partial_x^2 W = -P_{+hi}(W P_- \partial_x u) - P_{+hi}(P_{lo} e^{-\frac{i}{2}F} P_- \partial_x u) \\ W(\cdot, 0) = P_{+hi}(e^{-\frac{i}{2}F} u_0). \end{cases} \tag{5-3}$$

From (5-2) and Lemma 2.6, we thus deduce that  $W \in X_T^{s,1}$  so that, by interpolation with (5-2),  $W \in X_T^{\frac{1}{2}, \frac{1}{2}+}$ . But since  $u$  is given in  $L_T^\infty H_x^s \cap L_T^4 W_x^{s,4} \cap X_T^{s-\theta, \theta}$ , considering (2-6), the bilinear estimate (5-1), and (3-48), we infer that there exists only one solution to (5-3) in  $X_T^{\frac{1}{2}, \frac{1}{2}+}$ . Hence,  $w = \partial_x W$  belongs to  $X_T^{-\frac{1}{2}, \frac{1}{2}+}$  and is the unique solution to (3-4) in  $X_T^{-\frac{1}{2}, \frac{1}{2}+}$  emanating from the initial data  $w_0 = \partial_x P_{hi}(e^{-\frac{i}{2}F} u_0) \in L^2(\mathbb{R})$ . On the other hand, according to Proposition 3.4, one can construct a solution to (3-4) emanating from  $w_0$  and belonging to  $Y_T^s$  by using a Picard iterative scheme. Moreover, using (1-1) and Lemma 2.6 we can easily check that this solution belongs to  $X_T^{-1,1}$  and thus by interpolation to  $X_T^{s-\frac{1}{2}+, \frac{1}{2}+} \hookrightarrow X_T^{-\frac{1}{2}, \frac{1}{2}+}$ . This ensures that  $w = \partial_x P_{hi}(e^{-iF/2})$  belongs to  $Y_T^s \hookrightarrow X_T^{0, \frac{1}{2}}$ , which concludes the proof.

### 6. Continuity of the flow map for the weak $L^2$ -topology

In [Cui and Kenig 2010] it is proven that, for any  $t \geq 0$ , the flow map  $u_0 \mapsto u(t)$  associated with the Benjamin–Ono equation is continuous from  $L^2(\mathbb{R})$  equipped with the weak topology into itself. In this section, we explain how the uniqueness part of Theorem 1.1 enables us to simplify the proof of this result by following the approach developed in [Goubet and Molinet 2009].

Let  $\{u_{0,n}\}_n \subset L^2(\mathbb{R})$  be a sequence of initial data that converges weakly to  $u_0$  in  $L^2(\mathbb{R})$ , and let  $u$  be the solution emanating from  $u_0$  given by Theorem 1.1. From the Banach–Steinhaus theorem, we know that  $\{u_{0,n}\}_n$  is bounded in  $L^2(\mathbb{R})$ , and from Theorem 1.1 we know that  $\{u_{0,n}\}_n$  gives rise to a sequence  $\{u_n\}_n$  of solutions to (1-1) bounded in  $C([0, 1]; L^2(\mathbb{R})) \cap L^4([0, 1] \times \mathbb{R})$  with an associated sequence of gauge functions  $\{w_n\}_n$  bounded in  $X_1^{0, \frac{1}{2}}$ . Therefore, there exist  $v \in L^\infty([0, 1]; L^2(\mathbb{R})) \cap X_1^{-1,1} \cap L^4([0, 1] \times \mathbb{R})$  and  $z \in X_1^{0, \frac{1}{2}}$  such that, up to the extraction of a subsequence,  $\{u_n\}_n$  converges to  $v$  weakly in  $L^4([0, 1] \times \mathbb{R})$  and weakly star in  $L^\infty([0, 1] \times \mathbb{R})$ , and  $\{w_n\}_n$  converges to  $z$  weakly in  $X_1^{0, \frac{1}{2}}$ . We now need some compactness on  $\{u_n\}_n$  to ensure that  $z$  is the gauge transform of  $v$ . In this direction, we first notice, since  $\{w_n\}_n$  is bounded in  $X_1^{0, \frac{1}{2}}$  and by using the Kato’s smoothing effect injected in Bourgain’s spaces framework, that  $\{D_x^{1/4} w_n\}_n$  is bounded in  $L_x^4 L_1^2$ . Let  $\eta_R(\cdot) := \eta(\cdot/R)$ . Using (3-6) and Lemma 2.6 we

infer that

$$\begin{aligned} \|D_x^{\frac{1}{4}} P_{+HI} u_n\|_{L^2([0, 1[\times]-R, R])} &\lesssim \|D_x^{\frac{1}{4}} P_{+HI}(e^{\frac{i}{2}F[u_n]} w_n \eta_R)\|_{L^2_{1,x}} + \|D_x^{\frac{1}{4}} P_{+HI}(P_{+hi} e^{\frac{i}{2}F[u_n]} \partial_x P_{lo} e^{-\frac{i}{2}F[u_n]})\|_{L^2_{1,x}} \\ &\quad + \|D_x^{\frac{1}{4}} P_{+HI}(P_{+HI} e^{\frac{i}{2}F[u_n]} \partial_x P_{-hi} e^{-\frac{i}{2}F[u_n]})\|_{L^2_{1,x}} \\ &\lesssim \|D_x^{\frac{1}{4}}(w_n \eta_R)\|_{L^2_x L^2_1} + \|D_x^{\frac{1}{4}} e^{iF[u_n]}\|_{L^8_{1,x}} \|w_n\|_{L^{\frac{8}{3}}_{1,x}} + \|u_n\|_{L^4_{1,x}}^2. \end{aligned}$$

But clearly

$$\|D_x^{\frac{1}{4}}(w_n \eta_R)\|_{L^2_x L^2_1} \lesssim C(R) \left( \|D_x^{\frac{1}{4}} w_n\|_{L^4_x L^2_1} + \|w_n\|_{L^2_{1,x}} \right),$$

and by interpolation  $\|D_x^{1/4} e^{iF[u_n]}\|_{L^8_{1,x}} \lesssim \|u_n\|_{L^2_{1,x}}^{3/4}$ . Therefore, recalling that the  $u_n$  are real-valued functions, it follows that  $\{u_n\}_n$  is bounded in  $L^2_1 H^{1/4}([ -R, R])$ .

Since, according to Equation (1-1),  $\{\partial_t u_n\}_n$  is bounded in  $L^2_1 H_x^{-2}$ , Aubin–Lions compactness theorem and standard diagonal extraction arguments ensure that there exists an increasing sequence of integers  $\{n_k\}_k$  such that  $u_{n_k} \rightarrow v$  a.e. in  $]0, 1[ \times \mathbb{R}$  and  $u_{n_k}^2 \rightharpoonup v^2$  in  $L^2([0, 1[ \times \mathbb{R})$ . In view of our construction of the primitive  $F[u_n]$  of  $u_n$  (see Section 3A), it is then easy to check that  $F[u_{n_k}]$  converges to the primitive  $F[v]$  of  $v$  a.e. in  $]0, 1[ \times \mathbb{R}$ . This ensures that  $P_{+hi}(e^{-\frac{i}{2}F[u_{n_k}]})$  converges weakly to  $P_{+hi}(e^{-\frac{i}{2}F[v]})$  in  $L^2([0, 1[ \times \mathbb{R})$ , and thus,  $z$  is the gauge transform of  $v$ . Passing to the limit in the equation, we conclude that  $v$  satisfies (1-1) and belongs to the class of uniqueness of Theorem 1.1.

Moreover, setting  $(\cdot, \cdot)$  for the  $L^2_x$  scalar product, by (1-1) and the bounds above, it is easy to check that for any smooth space function  $\phi$  with compact support, the family  $\{t \mapsto (u_{n_k}(t), \phi)\}$  is uniformly equicontinuous on  $[0, 1]$ . Ascoli’s theorem then ensures that  $(u_{n_k}(\cdot), \phi)$  converges to  $(v(\cdot), \phi)$  uniformly on  $[0, 1]$ , and thus,  $v(0) = u_0$ . By uniqueness, it follows that  $v \equiv u$ , which ensures that the whole sequence  $\{u_n\}$  converges to  $v$  in the sense above and not only a subsequence. Finally, from the above convergence result, it follows that  $u_n(t) \rightharpoonup u(t)$  in  $L^2_x$  for all  $t \in [0, 1]$ .  $\square$

### 7. The periodic case

In this section, we explain how the bilinear estimate proved in Proposition 3.5 can lead to a great simplification of the global well-posedness result in  $L^2(\mathbb{T})$  derived in [Molinet 2008] and to new uniqueness results in  $H^s(\mathbb{T})$ , where  $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ . With the notations of [Molinet 2007], these new results lead to the following global well-posedness theorem:

**Theorem 7.1.** *Let  $s \geq 0$  be given. For all  $u_0 \in H^s(\mathbb{T})$  and all  $T > 0$ , there exists a solution*

$$u \in C([0, T]; H^s(\mathbb{T})) \cap X_T^{s-1,1} \cap L_T^4 W^{s,4}(\mathbb{T}) \tag{7-1}$$

of (1-1) such that

$$w = \partial_x P_{+hi}(e^{-\frac{i}{2}\partial_x^{-1}\tilde{u}}) \in Y_T^s, \tag{7-2}$$

where

$$\tilde{u} := u\left(t, x - t \int u_0\right) - \int u_0 \quad \text{and} \quad \widehat{\partial_x^{-1}} := \frac{1}{i\xi}, \quad \xi \in \mathbb{Z}^*.$$



This solution is unique in the following classes:

- (i)  $u \in L^\infty([0, T]; L^2(\mathbb{T})) \cap L^4([0, T] \times \mathbb{T})$  and  $w \in X_T^{0, \frac{1}{2}}$ ,
- (ii)  $u \in L^\infty([0, T]; H^{\frac{1}{4}}(\mathbb{T})) \cap L_T^4 W^{\frac{1}{4}, 4}(\mathbb{T})$  whenever  $s \geq \frac{1}{4}$ ,
- (iii)  $u \in L^\infty([0, T]; H^{\frac{1}{2}}(\mathbb{T}))$  whenever  $s \geq \frac{1}{2}$ .

Moreover,  $u \in C_b(\mathbb{R}; L^2(\mathbb{T}))$ , and the flow map data-solution  $u_0 \mapsto u$  is continuous from  $H^s(\mathbb{T})$  into  $C([0, T]; H^s(\mathbb{T}))$ .

*Sketch of the proof.* In the periodic case, following [Molinet 2007], the gauge transform is defined as follows: let  $u$  be a smooth  $2\pi$ -periodic solution of BO with initial data  $u_0$ . In the sequel, we will assume that  $u(t)$  has mean value zero for all time. Otherwise, we perform the change of unknowns

$$\tilde{u}(t, x) := u\left(t, x - t \int u_0\right) - \int u_0, \tag{7-3}$$

where  $\int u_0 := \frac{1}{2\pi} \int_{\mathbb{T}} u_0$  is the mean value of  $u_0$ . It is easy to see that  $\tilde{u}$  satisfies BO with  $u_0 - \int u_0$  as initial data, and since  $\int \tilde{u}$  is preserved by the flow of BO,  $\tilde{u}(t)$  has mean value zero for all time. We take for the primitive of  $u$  the unique periodic, zero mean value primitive of  $u$  defined by

$$\hat{F}(0) = 0 \quad \text{and} \quad \hat{F}(\xi) = \frac{1}{i\xi} \hat{u}(\xi), \quad \xi \in \mathbb{Z}^*.$$

The gauge transform is then defined by

$$W := P_+(e^{-iF/2}). \tag{7-4}$$

Since  $F$  satisfies

$$F_t + \mathcal{H}F_{xx} = \frac{1}{2}F_x^2 - \frac{1}{2} \int F_x^2 = \frac{1}{2}F_x^2 - \frac{1}{2}P_0(F_x^2),$$

we finally obtain that  $w := W_x = -\frac{1}{2}iP_{+hi}(e^{-iF/2}F_x) = -\frac{1}{2}iP_+(e^{-iF/2}u)$  satisfies

$$\begin{aligned} w_t - iw_{xx} &= -\partial_x P_{hi}\left[e^{-iF/2}\left(P_-(F_{xx}) - \frac{i}{4}P_0(F_x^2)\right)\right] \\ &= -\partial_x P_{+hi}(WP_-(u_x)) + \frac{i}{4}P_0(F_x^2)w. \end{aligned} \tag{7-5}$$

Clearly the second term is harmless, and the first one has exactly the same structure as the one that we estimated in Proposition 3.5. Carefully following the proof of this proposition, it is not too hard to check that it also holds in the periodic case independent of the period  $\lambda \geq 1$ . Note in particular that (2-9) also holds with  $L_{x,t}^4$  and  $X^{0, \frac{3}{8}}$  respectively replaced by  $L_{t,\lambda}^4$  and  $X_{\lambda}^{0, \frac{3}{8}}$ ,  $\lambda \geq 1$ , where the subscript  $\lambda$  denotes spaces of functions with space variable on the torus  $\mathbb{R}/2\pi\lambda\mathbb{Z}$  (see [Bourgain 1993a] and also [Molinet 2007]). This leads to a great simplification of the proof the global well-posedness in  $L^2(\mathbb{T})$  proved in [Molinet 2008].

Now to derive the new uniqueness result we proceed exactly as in Section 5 except that Proposition 5.1 does not hold on the torus. Actually, on the torus it should be replaced by the following:

**Proposition 7.2.** *For  $s \geq \frac{1}{4}$  and all  $\lambda \geq 1$ , we have*

$$\|P_{+hi}(WP_{-} \partial_x u)\|_{X_{\lambda}^{s+\frac{1}{2},-\frac{1}{2}}} \lesssim \|W\|_{X_{\lambda}^{s+\frac{1}{2},\frac{1}{2}}} \left( \|J_x^s u\|_{L_{T,\lambda}^2} + \|J_x^s u\|_{L_{T,\lambda}^4} + \|u\|_{X_{\lambda}^{s-1,1}} \right). \tag{7-6}$$

Going back to the proof of the bilinear estimate, it is easy to be convinced that the above estimate works at the level  $s = 0_+$  in the regions  $\mathcal{A}$  and  $\mathcal{B}$  (see the proof of Proposition 5.1), whereas in the region  $\mathcal{C}$  we are clearly in trouble. Indeed, when  $s = 0$ , (3-37) must then be replaced by

$$|k^{\frac{1}{2}} k_1^{-\frac{1}{2}} k_2^2 \langle \sigma_2 \rangle^{-1}| \sim |k^{-\frac{1}{2}} k_1^{-\frac{1}{2}} k_2|,$$

which cannot be bounded when  $|k_2| \gg k$ . On the other hand, at the level  $s = \frac{1}{4}$  it becomes

$$|k^{\frac{3}{4}} k_1^{-\frac{3}{4}} k_2^{\frac{7}{4}} \langle \sigma_2 \rangle^{-1}| \sim |k^{-\frac{1}{4}} k_1^{-\frac{3}{4}} k_2^{\frac{3}{4}}| \lesssim k^{-\frac{1}{4}} \lesssim 1,$$

which yields the result.

With Proposition 7.2 in hand, exactly the same procedure as in Section 5 leads to the uniqueness result in the class  $u \in L_T^\infty H^{\frac{1}{4}}(\mathbb{T}) \cap L_T^4 W^{\frac{1}{4},4}(\mathbb{T})$  and by Sobolev embedding to the uniqueness in the class  $u \in L_T^\infty H^{\frac{1}{2}}(\mathbb{T})$ , i.e., unconditional uniqueness in  $H^{\frac{1}{2}}(\mathbb{T})$ . As in the real line case, it proves the uniqueness of the (energy) weak solutions that belong to  $L^\infty(\mathbb{R}; H^{\frac{1}{2}}(\mathbb{T}))$ .

### Appendix

*Proof of Proposition 5.1.* We will need the following calculus lemma stated in [Ginibre et al. 1997].

**Lemma A.3.** *Let  $0 < a_- \leq a_+$  such that  $a_- + a_+ > \frac{1}{2}$ . Then for all  $\mu \in \mathbb{R}$ ,*

$$\int_{\mathbb{R}} \langle y \rangle^{-2a_-} \langle y - \mu \rangle^{-2a_+} dy \lesssim \langle \mu \rangle^{-s}, \tag{A-1}$$

where  $s = 2a_-$  if  $a_+ > \frac{1}{2}$ ,  $s = 2a_- - \varepsilon$  if  $a_+ = \frac{1}{2}$ , and  $s = 2(a_+ + a_-) - 1$  if  $a_+ < \frac{1}{2}$ , and let  $\varepsilon$  denote any small positive number.

The proof of Proposition 5.1 closely follows the one of Proposition 3.5 except in the region  $\sigma_2$ -dominant where we use the approach developed in [Kenig et al. 1996]. Recalling the notation used in (3-24)–(3-25), we need to prove that

$$|K| \lesssim \|h\|_{L_{x,t}^2} \|f\|_{L_{x,t}^2} \left( \|u\|_{L_{x,t}^2} + \|u\|_{L_{x,t}^4} + \|u\|_{X^{-\theta,\theta}} \right), \tag{A-2}$$

where

$$K = \int_{\mathcal{D}} \frac{\langle \xi \rangle^{\frac{1}{2}}}{\langle \sigma \rangle^{\frac{1}{2}-2\delta}} \widehat{h}(\xi, \tau) \frac{\langle \xi_1 \rangle^{-\frac{1}{2}}}{\langle \sigma_1 \rangle^{\frac{1}{2}+\delta}} \widehat{f}(\xi_1, \tau_1) \xi_2 \langle \xi_2 \rangle^{-s} \widehat{u}(\xi_2, \tau_2) d\nu. \tag{A-3}$$

For the same reason as in the proof of Proposition 3.5, we can assume that  $|\xi_2| \leq 1$ . By using a Littlewood–Paley decomposition on  $h$ ,  $f$ , and  $u$ ,  $K$  can be rewritten as

$$K = \sum_{N, N_1, N_2} K_{N, N_1, N_2}, \tag{A-4}$$

with

$$K_{N,N_1,N_2} := \int_{\mathcal{D}} \frac{\langle \xi \rangle^{\frac{1}{2}}}{\langle \sigma \rangle^{\frac{1}{2}-2\delta}} \widehat{P_N h}(\xi, \tau) \frac{\langle \xi_1 \rangle^{-\frac{1}{2}}}{\langle \sigma_1 \rangle^{\frac{1}{2}+\delta}} \widehat{P_{N_1} f}(\xi_1, \tau_1) \xi_2 \langle \xi_2 \rangle^{-s} \widehat{P_{N_2} u}(\xi_2, \tau_2) dv$$

and the dyadic numbers  $N, N_1,$  and  $N_2$  ranging from 1 to  $+\infty$ . Moreover, we will denote by  $K_{N,N_1,N_2}^{\mathcal{A}_{N,N_2}}, K_{N,N_1,N_2}^{\mathcal{B}_{N,N_2}}$ , and  $K_{N,N_1,N_2}^{\mathcal{C}_{N,N_2}}$  the restriction of  $K_{N,N_1,N_2}$  to the regions  $\mathcal{A}_{N,N_2}, \mathcal{B}_{N,N_2},$  and  $\mathcal{C}_{N,N_2}$  defined in (3-28). Then it follows that

$$|K| \leq |K_{\mathcal{A}}| + |K_{\mathcal{B}}| + |K_{\mathcal{C}}|, \tag{A-5}$$

where

$$K_{\mathcal{A}} := \sum_{N,N_1,N_2} J_{N,N_1,N_2}^{\mathcal{A}_{N,N_2}}, \quad K_{\mathcal{B}} := \sum_{N,N_1,N_2} K_{N,N_1,N_2}^{\mathcal{B}_{N,N_2}}, \quad \text{and} \quad K_{\mathcal{C}} := \sum_{N,N_1,N_2} J_{N,N_1,N_2}^{\mathcal{C}_{N,N_2}}$$

so that it suffices to estimate  $|K_{\mathcal{A}}|, |K_{\mathcal{B}}|,$  and  $|K_{\mathcal{C}}|$ . Recall that due to the structure of  $\mathcal{D}$ , one of the following case must hold:

- (1) high-low interaction:  $N_1 \sim N$  and  $N_2 \leq N_1,$
- (2) high-high interaction:  $N_1 \sim N_2$  and  $N \leq N_1.$

*Estimate for  $|K_{\mathcal{A}}|$ .* In the first case, it follows from the triangular inequality, Plancherel’s identity, and Hölder’s inequality that

$$\begin{aligned} |K_{\mathcal{A}}| &\lesssim \|h\|_{L_{x,t}^2} \sum_{N_1} \sum_{N_2 \leq N_1} \frac{N_1^{\frac{1}{2}}}{(N_1 N_2)^{\frac{1}{2}-2\delta}} \left\| P_{N_1} \left( J_x^{-\frac{1}{2}} P_{N_1} \left( \frac{\widehat{f}}{\langle \sigma_1 \rangle^{\frac{1}{2}+\delta}} \right)^\vee P_- \partial_x J_x^{-s} P_{N_2} u \right) \right\|_{L_{x,t}^2} \\ &\lesssim \|h\|_{L_{x,t}^2} \sum_{N_1} \sum_{N_2 \leq N_1} \frac{N_2^{\frac{1}{2}-s+2\delta}}{(N_1)^{\frac{1}{2}-2\delta}} \left\| P_{N_1} \left( \frac{\widehat{f}}{\langle \sigma_1 \rangle^{\frac{1}{2}+\delta}} \right)^\vee \right\|_{L_{x,t}^4} \|P_{N_2} u\|_{L_{x,t}^4} \\ &\lesssim \|h\|_{L_{x,t}^2} \|u\|_{L_{x,t}^4} \sum_{N_1} N_1^{4\delta-s} \left\| P_{N_1} \left( \frac{\widehat{f}}{\langle \sigma_1 \rangle^{\frac{1}{2}+\delta}} \right)^\vee \right\|_{L_{x,t}^4}. \end{aligned}$$

Then it is deduced from the Cauchy–Schwarz inequality in  $N_1$  that

$$|K_{\mathcal{A}}| \lesssim \|h\|_{L_{x,t}^2} \left( \sum_{N_1} \left\| P_{N_1} \left( \frac{\widehat{f}}{\langle \sigma_1 \rangle^{\frac{1}{2}+\delta}} \right)^\vee \right\|_{L_{x,t}^4}^2 \right)^{\frac{1}{2}} \|u\|_{L_{x,t}^4} \tag{A-6}$$

since  $s > 10\delta$ . On the other, estimate (A-6) also holds in the case of high-high interaction by arguing exactly as in (3-31) so that estimate (2-9) yields

$$|K_{\mathcal{A}}| \lesssim \|h\|_{L_{x,t}^2} \|f\|_{L_{x,t}^2} \|u\|_{L_{x,t}^4}. \tag{A-7}$$

*Estimate for  $|K_{\mathcal{B}}|$ .* The estimate

$$|K_{\mathcal{B}}| \lesssim \|h\|_{L_{x,t}^2} \|f\|_{L_{x,t}^2} \|u\|_{L_{x,t}^4} \tag{A-8}$$

follows by the same argument as in (A-6).

Estimate for  $|K_{\mathcal{C}}|$ . First observe that

$$|K_{\mathcal{C}}| \lesssim \int_{\tilde{\mathcal{C}}} \frac{|\xi|^{\frac{1}{2}}}{\langle \sigma \rangle^{\frac{1}{2}-2\delta}} |\widehat{h}(\xi, \tau)| \frac{|\xi_1|^{-\frac{1}{2}}}{\langle \sigma_1 \rangle^{\frac{1}{2}+\delta}} |\widehat{f}(\xi_1, \tau_1)| \frac{|\xi_2|^{(1+\theta-s)}}{\langle \sigma_2 \rangle^\theta} \frac{\langle \sigma_2 \rangle^\theta}{|\xi_2|^\theta} |\widehat{u}(\xi_2, \tau_2)| dv, \tag{A-9}$$

where

$$\tilde{\mathcal{C}} = \left\{ (\xi, \xi_1, \tau, \tau_1) \in \mathcal{D} \mid (\xi, \xi_1, \tau, \tau_1) \in \bigcup_{N, N_2} \mathcal{C}_{N, N_2} \right\}.$$

Since  $|\sigma_2| > |\sigma|$  and  $|\sigma_2| > |\sigma_1|$  in  $\tilde{\mathcal{C}}$ , (3-28) implies that  $|\sigma_2| \gtrsim |\xi \xi_2|$ . Applying the Cauchy-Schwarz inequality twice, we deduce that

$$|K_{\mathcal{C}}| \lesssim \sup_{\xi_2, \tau_2} (L_{\tilde{\mathcal{C}}}(\xi_2, \tau_2))^{\frac{1}{2}} \|f\|_{L^2_{\xi, \tau}} \|g\|_{L^2_{\xi, \tau}} \|h\|_{L^2_{\xi, \tau}},$$

where

$$L_{\tilde{\mathcal{C}}}(\xi_2, \tau_2) = \frac{|\xi_2|^{2+2(\theta-s)}}{\langle \sigma_2 \rangle^{2\theta}} \int_{\mathcal{C}(\xi_2, \tau_2)} \frac{|\xi| |\xi_1|^{-1}}{\langle \sigma \rangle^{1-4\delta} \langle \sigma_1 \rangle^{1+2\delta}} d\xi_1 d\tau_1$$

and

$$\tilde{\mathcal{C}}(\xi_2, \tau_2) = \{ (\xi_1, \tau_1) \in \mathbb{R}^2 \mid (\xi, \xi_1, \tau, \tau_1) \in \mathcal{C} \}.$$

Thus, to prove that

$$|K_{\mathcal{C}}| \lesssim \|h\|_{L^2_{x,t}} \|f\|_{L^2_{x,t}} \|u\|_{X^{-\theta, \theta}}, \tag{A-10}$$

it is enough to prove that  $L_{\tilde{\mathcal{C}}}(\xi_2, \tau_2) \lesssim 1$  for all  $(\xi_2, \tau_2) \in \mathbb{R}^2$ . We deduce from (A-1) and (3-28) that

$$L_{\tilde{\mathcal{C}}}(\xi_2, \tau_2) \lesssim \frac{|\xi_2|^{2+2(\theta-s)}}{\langle \sigma_2 \rangle^{1+2\delta}} \int_{\xi_1} \frac{|\xi| |\xi_1|^{-1}}{\langle \sigma_2 + 2\xi \xi_2 \rangle^{1-4\delta}} d\xi_1$$

since  $\theta = 1 + \delta$ . To integrate with respect to  $\xi_1$ , we change variables

$$\mu_2 = \sigma_2 + 2\xi \xi_2 \quad \text{so that} \quad d\mu_2 = 2\xi_2 d\xi_1 \quad \text{and} \quad |\mu_2| \leq 4|\sigma_2|.$$

Moreover, (3-26) and (3-28) imply that

$$\frac{|\xi| |\xi_1|^{-1} |\xi_2|^{1+2(\theta-s)}}{|\xi_1|^2} \leq |\xi \xi_2|^{\frac{1}{2}+\theta-s} \lesssim |\sigma_1|^{\frac{1}{2}+\theta-s}$$

in  $\tilde{\mathcal{C}}$ . Then

$$L_{\tilde{\mathcal{C}}}(\xi_2, \tau_2) \lesssim \frac{|\xi_2|^{1+2(\theta-s)}}{\langle \sigma_2 \rangle^{1+2\delta}} \int_0^{4|\sigma_2|} \frac{|\xi| |\xi_1|^{-1}}{\langle \mu_2 \rangle^{1-4\delta}} d\mu_2 \lesssim \frac{\langle \sigma_2 \rangle^{\frac{1}{2}+\theta-s+4\delta}}{\langle \sigma_2 \rangle^{1+2\delta}} \lesssim \langle \sigma_2 \rangle^{3\delta-s} \lesssim 1$$

since  $s - 3\delta > 0$ .

Finally, we conclude the proof of Proposition 5.1 by gathering (A-2), (A-5), (A-7), (A-8), and (A-10).

### Acknowledgments

This work was initiated during a visit of the second author at the L.M.P.T., Université François Rabelais, Tours. He would like to thank the L.M.P.T. for the kind hospitality. The first author was partially supported by the ANR project Equa-disp.

### References

- [Abdelouhab et al. 1989] L. Abdelouhab, J. L. Bona, M. Felland, and J.-C. Saut, “Nonlocal models for nonlinear, dispersive waves”, *Phys. D* **40**:3 (1989), 360–392. MR 91d:58033 Zbl 0699.35227
- [Benjamin 1967] T. B. Benjamin, “Internal waves of permanent form in fluid of great depth”, *J. Fluid Mech.* **29** (1967), 559–592.
- [Biagioni and Linares 2001] H. A. Biagioni and F. Linares, “Ill-posedness for the derivative Schrödinger and generalized Benjamin–Ono equations”, *Trans. Amer. Math. Soc.* **353**:9 (2001), 3649–3659. MR 2002e:35215 Zbl 0970.35154
- [Bourgain 1993a] J. Bourgain, “Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations, I: Schrödinger equations”, *Geom. Funct. Anal.* **3**:2 (1993), 107–156.
- [Bourgain 1993b] J. Bourgain, “Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations, II: The KdV-equation”, *Geom. Funct. Anal.* **3**:3 (1993), 209–262. MR 95d:35160b
- [Burq and Planchon 2006] N. Burq and F. Planchon, “The Benjamin–Ono equation in energy space”, pp. 55–62 in *Phase space analysis of partial differential equations*, edited by A. Bove et al., Progr. Nonlinear Differential Equations Appl. **69**, Birkhäuser, Boston, MA, 2006. MR 2007h:35276 Zbl 1127.35302
- [Burq and Planchon 2008] N. Burq and F. Planchon, “On well-posedness for the Benjamin–Ono equation”, *Math. Ann.* **340**:3 (2008), 497–542. MR 2009a:35205 Zbl 1148.35074
- [Craig et al. 2005] W. Craig, P. Guyenne, and H. Kalisch, “Hamiltonian long-wave expansions for free surfaces and interfaces”, *Comm. Pure Appl. Math.* **58**:12 (2005), 1587–1641. MR 2006i:76012 Zbl 1151.76385
- [Cui and Kenig 2010] S. Cui and C. E. Kenig, “Weak continuity of dynamical systems for the KdV and mKdV equations”, *Differential Integral Equations* **23**:11–12 (2010), 1001–1022. MR 2012b:35299 Zbl 05944721
- [Fokas and Ablowitz 1983] A. S. Fokas and M. J. Ablowitz, “The inverse scattering transform for the Benjamin–Ono equation: a pivot to multidimensional problems”, *Stud. Appl. Math.* **68**:1 (1983), 1–10. MR 84f:35139 Zbl 0505.76031
- [Ginibre et al. 1997] J. Ginibre, Y. Tsutsumi, and G. Velo, “On the Cauchy problem for the Zakharov system”, *J. Funct. Anal.* **151**:2 (1997), 384–436. MR 2000c:35220 Zbl 0894.35108
- [Goubet and Molinet 2009] O. Goubet and L. Molinet, “Global attractor for weakly damped nonlinear Schrödinger equations in  $L^2(\mathbb{R})$ ”, *Nonlinear Anal.* **71**:1–2 (2009), 317–320. MR 2010h:35033 Zbl 1170.35534
- [Ionescu and Kenig 2007] A. D. Ionescu and C. E. Kenig, “Global well-posedness of the Benjamin–Ono equation in low-regularity spaces”, *J. Amer. Math. Soc.* **20**:3 (2007), 753–798. MR 2008f:35350 Zbl 1123.35055
- [Iório 1986] R. J. Iório, Jr., “On the Cauchy problem for the Benjamin–Ono equation”, *Comm. Partial Differential Equations* **11**:10 (1986), 1031–1081. MR 88b:35034 Zbl 0608.35030
- [Kenig and Koenig 2003] C. E. Kenig and K. D. Koenig, “On the local well-posedness of the Benjamin–Ono and modified Benjamin–Ono equations”, *Math. Res. Lett.* **10**:5–6 (2003), 879–895. MR 2004j:35249 Zbl 1044.35072
- [Kenig et al. 1993] C. E. Kenig, G. Ponce, and L. Vega, “Well-posedness and scattering results for the generalized Korteweg–de Vries equation via the contraction principle”, *Comm. Pure Appl. Math.* **46**:4 (1993), 527–620. MR 94h:35229 Zbl 0808.35128
- [Kenig et al. 1996] C. E. Kenig, G. Ponce, and L. Vega, “A bilinear estimate with applications to the KdV equation”, *J. Amer. Math. Soc.* **9**:2 (1996), 573–603. MR 96k:35159 Zbl 0848.35114
- [Koch and Tzvetkov 2003] H. Koch and N. Tzvetkov, “On the local well-posedness of the Benjamin–Ono equation in  $H^s(\mathbb{R})$ ”, *Int. Math. Res. Not.* **2003**:26 (2003), 1449–1464. MR 2004b:35284 Zbl 1039.35106
- [Koch and Tzvetkov 2005] H. Koch and N. Tzvetkov, “Nonlinear wave interactions for the Benjamin–Ono equation”, *Int. Math. Res. Not.* **2005**:30 (2005), 1833–1847. MR 2006f:35245 Zbl 1156.35460

- [Linares et al. 2011] F. Linares, D. Pilod, and G. Ponce, “Well-posedness for a higher-order Benjamin–Ono equation”, *J. Differential Equations* **250**:1 (2011), 450–475. MR 2011j:35210 Zbl 1205.35261
- [Molinet 2007] L. Molinet, “Global well-posedness in the energy space for the Benjamin–Ono equation on the circle”, *Math. Ann.* **337**:2 (2007), 353–383. MR 2007i:35205 Zbl 1140.35001
- [Molinet 2008] L. Molinet, “Global well-posedness in  $L^2$  for the periodic Benjamin–Ono equation”, *Amer. J. Math.* **130**:3 (2008), 635–683. MR 2009f:35300 Zbl 1157.35001
- [Molinet 2009] L. Molinet, “Sharp ill-posedness result for the periodic Benjamin–Ono equation”, *J. Funct. Anal.* **257**:11 (2009), 3488–3516. MR 2011c:35517 Zbl 1181.35246
- [Molinet et al. 2001] L. Molinet, J. C. Saut, and N. Tzvetkov, “Ill-posedness issues for the Benjamin–Ono and related equations”, *SIAM J. Math. Anal.* **33**:4 (2001), 982–988. MR 2002k:35281 Zbl 0999.35085
- [Pilod 2008] D. Pilod, “On the Cauchy problem for higher-order nonlinear dispersive equations”, *J. Differential Equations* **245**:8 (2008), 2055–2077. MR 2009f:35302 Zbl 1152.35017
- [Ponce 1991] G. Ponce, “On the global well-posedness of the Benjamin–Ono equation”, *Differential Integral Equations* **4**:3 (1991), 527–542. MR 92e:35137 Zbl 0732.35038
- [Tao 2004] T. Tao, “Global well-posedness of the Benjamin–Ono equation in  $H^1(\mathbf{R})$ ”, *J. Hyperbolic Differ. Equ.* **1**:1 (2004), 27–49. MR 2005f:35273 Zbl 1055.35104
- [Tataru 1998] D. Tataru, “Local and global results for wave maps. I”, *Comm. Partial Differential Equations* **23**:9-10 (1998), 1781–1793. MR 99j:58209 Zbl 0914.35083

Received 7 Sep 2010. Accepted 15 Jan 2011.

LUC MOLINET: luc.molinet@lmpt.univ-tours.fr

Laboratoire de Mathématiques et Physique Théorique, Université de Tours, Parc Grandmont, 37200 Tours, France

DIDIER PILOD: didier@im.ufrj.br

Instituto de Matemática, Universidade Federal do Rio de Janeiro, Caixa Postal 68530, 21945-970 Rio de Janeiro, Brazil



# ON TRIANGLES DETERMINED BY SUBSETS OF THE EUCLIDEAN PLANE, THE ASSOCIATED BILINEAR OPERATORS AND APPLICATIONS TO DISCRETE GEOMETRY

ALLAN GREENLEAF AND ALEX IOSEVICH

We prove that if the Hausdorff dimension of a compact set  $E \subset \mathbb{R}^2$  is greater than  $\frac{7}{4}$ , then the set of three-point configurations determined by  $E$  has positive three-dimensional measure. We establish this by showing that a natural measure on the set of such configurations has Radon–Nikodym derivative in  $L^\infty$  if  $\dim_{\mathcal{H}}(E) > \frac{7}{4}$ , and the index  $\frac{7}{4}$  in this last result cannot, in general, be improved. This problem naturally leads to the study of a bilinear convolution operator,

$$B(f, g)(x) = \iint f(x - u) g(x - v) dK(u, v),$$

where  $K$  is surface measure on the set  $\{(u, v) \in \mathbb{R}^2 \times \mathbb{R}^2 : |u| = |v| = |u - v| = 1\}$ , and we prove a scale of estimates that includes  $B : L^2_{-1/2}(\mathbb{R}^2) \times L^2(\mathbb{R}^2) \rightarrow L^1(\mathbb{R}^2)$  on positive functions.

As an application of our main result, it follows that for finite sets of cardinality  $n$  and belonging to a natural class of discrete sets in the plane, the maximum number of times a given three-point configuration arises is  $O(n^{\frac{9}{7}+\epsilon})$  (up to congruence), improving upon the known bound of  $O(n^{\frac{4}{3}})$  in this context.

## 1. Introduction

The classical Falconer distance conjecture says that if a compact set  $E \subset \mathbb{R}^d$ ,  $d \geq 2$ , has Hausdorff dimension  $\dim_{\mathcal{H}}(E) > \frac{d}{2}$ , then the one-dimensional Lebesgue measure  $\mathcal{L}^1(\Delta(E))$  of its *distance set*,

$$\Delta(E) := \{|x - y| \in \mathbb{R} : x, y \in E\},$$

is positive. Here and throughout,  $|\cdot|$  denotes the Euclidean distance. A beautiful example due to Falconer, based on the integer lattice, shows that the exponent  $d/2$  is best possible. The best results currently known, culminating almost three decades of efforts by Falconer [1985b], Mattila [1987], Bourgain [1994] and others, are due to Wolff [1999] for  $d = 2$  and Erdoğlan [2005] for  $d \geq 3$ . They prove that  $\mathcal{L}^1(\Delta(E)) > 0$  if

$$\dim_{\mathcal{H}}(E) > \frac{d}{2} + \frac{1}{3}.$$

Since two-point configurations are equivalent, up to Euclidean motions of  $\mathbb{R}^d$ , precisely if the corresponding distances are the same, one may think of the Falconer conjecture as stating that the set of

---

The authors were supported by NSF grants DMS-0853892 and DMS-1045404.  
MSC2010: 42B15, 52C10.

Keywords: Falconer–Erdős distance problem, distance set, geometric combinatorics, multilinear operators, triangles.



two-point configurations determined by a compact  $E$  of sufficiently high Hausdorff dimension has positive measure. A natural extension of the Falconer problem is this:

**Question.** *For  $N \geq 3$ , how great does the Hausdorff dimension of a compact set need to be in order to ensure that the set of  $N$ -point configurations it determines is of positive measure?*

To make this more precise, define the *space of  $(k + 1)$ -point configurations in  $E$*  or the *quotient space of (possibly degenerate)  $k$ -simplices with vertices in  $E$* , modulo Euclidean motions, as

$$T_k(E) := E^{k+1} / \sim,$$

where  $E^{k+1} = E \times E \times \dots \times E$  ( $k + 1$  times) and the congruence relation

$$(x^1, x^2, \dots, x^{k+1}) \sim (y^1, y^2, \dots, y^{k+1})$$

holds if and only if there exists an element  $R$  of the orthogonal group  $O(d)$  and a translation  $\tau \in \mathbb{R}^d$  such that

$$y^j = \tau + R(x^j), \quad 1 \leq j \leq k + 1.$$

Observe that we may identify  $T_k(E)$  as a subset of  $\mathbb{R}^{\binom{k+1}{2}}$  since rigid motions may be encoded by fixing distances, and this induces  $\binom{k+1}{2}$ -dimensional Lebesgue measure on  $T_k(E)$ . The problem under consideration was first taken up in [Erdoğan et al. 2011], where it was shown that

$$\text{if } \dim_{\mathcal{H}}(E) > \frac{d+k+1}{2}, \text{ then } \mathcal{L}^{\binom{k+1}{2}}(T_k(E)) > 0.$$

Unfortunately, these results do not give a nontrivial exponent for what are arguably the most natural cases, namely three-point configurations in  $\mathbb{R}^2$ , four-point configurations in  $\mathbb{R}^3$  and, more generally,  $(d + 1)$ -point configurations (generically spanning  $d$ -dimensional simplices) in  $\mathbb{R}^d$ . (Nor does it yield results for  $(d - 1)$ -simplices.) Here, we partially fill this gap by establishing a nontrivial exponent for three-point configurations in the plane.

As for counterexamples, it is easy to see that  $\mathcal{L}^{\binom{k+1}{2}}(T_k(E)) > 0$  does not hold if the Hausdorff dimension of  $E$  is less than or equal to  $d - 1$ ; to see this, just take  $E$  to be a subset of a  $(d - 1)$ -dimensional plane. We do not currently know if more restrictive conditions exist in this context. However, more restrictive counterexamples do exist if we consider the following related question. For any symmetric matrix  $t = \{t_{ij}\}$  with zeros on the diagonal, let

$$\mathcal{S}_t^k(E) = \{(x^1, \dots, x^{k+1}) \in E^{k+1} : |x^i - x^j| = t_{ij}, \forall i, j\}.$$

Conditions under which

$$\dim_{\mathcal{M}}(\mathcal{S}_t^k(E)) \leq (k + 1) \dim_{\mathcal{H}}(E) - \binom{k+1}{2} = (k + 1) \left( \dim_{\mathcal{H}}(E) - \frac{k}{2} \right), \tag{1-1}$$

where  $\dim_{\mathcal{M}}$  denotes the Minkowski dimension, are analyzed in [Eswarathasan et al. 2011] in the case  $k = 1$  in a rather general setting and in [Erdoğan et al.  $\geq 2012$ ] in the case  $k > 1$ . (See [Falconer 1985a; Mattila 1995] for background on  $\dim_{\mathcal{H}}$ ,  $\dim_{\mathcal{M}}$  and connections with harmonic analysis.)

The estimate (1-1) follows easily if one can show that

$$(\mu \times \mu \times \cdots \times \mu) \{ (x^1, \dots, x^{k+1}) : t_{ij} \leq |x^i - x^j| \leq t_{ij} + \epsilon, \forall i, j \} \lesssim \epsilon^{\binom{k+1}{2}} \text{ as } \epsilon \searrow 0, \tag{1-2}$$

where  $\mu$  is a Frostman measure (defined in (2-1) below) on  $E$ , under the assumption that  $\dim_{\mathcal{H}}(E) > s_0$  for some threshold  $s_0 < d$ . This is shown in [Erdoğan et al. ≥ 2012] under the assumption that the Hausdorff dimension of  $E$  is greater than  $(k/k + 1)d + k/2$ , but observe that this only yields a nontrivial exponent (less than  $d$ ) if  $\binom{k}{2} < d$  and, in particular, does not cover the important case of  $k = d$ .

Our main result is the following:

**Theorem 1.1.** *Let  $E \subset [0, 1]^2$  be compact and  $\mu$  a Frostman measure on  $E$ .*

- (i) *If  $\dim_{\mathcal{H}}(E) > \frac{7}{4}$ , then estimate (1-2) holds with  $d = k = 2$ .*
- (ii) *If  $\dim_{\mathcal{H}}(E) > \frac{7}{4}$ , then  $\mathcal{L}^3(T_2(E)) > 0$ .*

The proof that part (i) of Theorem 1.1 implies part (ii) is presented in Section 2 below; part (i) is then proved by analysis of a bilinear operator (or trilinear form) in Sections 2, 3, and 4.

We observe that the result in part (i) is sharp in the following sense. Define a measure  $\nu$  on  $T_2(E)$  by the relation

$$\int f(t_{12}, t_{13}, t_{23}) d\nu(t_{12}, t_{13}, t_{23}) = \iiint f(|x^1 - x^2|, |x^1 - x^3|, |x^2 - x^3|) d\mu(x^1) d\mu(x^2) d\mu(x^3), \tag{1-3}$$

where  $\mu$  is any Frostman measure on  $E$ . We shall prove that the Radon–Nikodym derivative  $d\nu/dt \in L^\infty$ , which is just a rephrasing of the statement that (1-2) holds for  $d = k = 2$ , if the Hausdorff dimension of  $E$  is greater than  $\frac{7}{4}$ . On the other hand, we also use a variant of an example from [Mattila 1987] to show that if  $s < \frac{7}{4}$ , then  $d\nu/dt$  need not be, in general, in  $L^\infty$ , in the sense that for every  $s < \frac{7}{4}$  there exists a set  $E$  of Hausdorff dimension  $s$  and a Frostman measure  $\mu$  supported on  $E$  such that  $d\nu/dt \notin L^\infty$ . (This issue is taken up in Section 5.) Thus, in order to try to improve part (ii) of the theorem, i.e., to prove that  $\mathcal{L}^3(T_2(E)) > 0$  if  $\dim_{\mathcal{H}}(E) = s_0$  for some  $s_0 \leq \frac{7}{4}$ , it would be reasonable to try to obtain an  $L^p$ , rather than an  $L^\infty$  bound on the measure  $\nu$  defined by (1-3). We hope to address this in a subsequent paper.

Theorem 1.1 may be viewed as a local version of the following theorem due to Furstenberg, Katznelson and Weiss; see also [Bourgain 1986; Ziegler 2006] for subsequent results along these lines.

**Theorem 1.2** [Furstenberg et al. 1990]. *Let  $E \subset \mathbb{R}^2$  be of positive upper Lebesgue density in the sense that*

$$\limsup_{R \rightarrow \infty} \frac{\mathcal{L}^d \{ E \cap [-R, R]^2 \}}{(2R)^2} > 0,$$

where  $\mathcal{L}^2$  denotes 2-dimensional Lebesgue measure. For  $\delta > 0$ , let  $E_\delta$  denote the  $\delta$ -neighborhood of  $E$ . Then, given vectors  $u, v$  in  $\mathbb{R}^2$ , there exists  $l_0$  such that for any  $l > l_0$  and  $\delta > 0$ , there exist  $x, y, z \in E_\delta$  forming a triangle congruent to  $\{\mathbf{0}, lu, lv\}$ , where  $\mathbf{0}$  denotes the origin in  $\mathbb{R}^2$ .

We note in passing that it is generally believed that the conclusion of Theorem 1.2 still holds if the  $\delta$ -neighborhood of  $E$  is replaced by  $E$  under an additional assumption that the triangles under consideration are nondegenerate. For degenerate triangles, i.e., allowing line segments, the necessity of considering the  $\delta$ -neighborhood of  $E$  was established by Bourgain (see [Furstenberg et al. 1990]).

In contrast to Theorem 1.2, we are able in the local version to go beyond subsets of the plane of positive Lebesgue measure, and we do not need to allow for dilations of the triangles. On the other hand, we only obtain a positive Lebesgue measure’s worth of the possible three-point configurations, not all of them.

It is also not difficult to show (see Section 2) that if the estimate (1-2) holds under the assumption that  $\dim_{\mathcal{H}}(E) > s_0$ , then  $\mathcal{L}^{\binom{k+1}{2}}(T_k(E)) > 0$  for these sets. In [Erdoğan et al. ≥ 2012], a number of estimates of the type (1-2) are proved but, as we note above, do not cover the cases  $k = d$  or  $k = d - 1$ .

**A combinatorial perspective.** Finite configuration problems have their roots in geometric combinatorics. For example, the Falconer distance problem is a continuous analog of the celebrated Erdős distance problem; see [Solymosi and Tóth 2001; Katz and Tardos 2004; Brass et al. 2005; Székely 1997] and the references therein. The discrete precursor of the problem discussed in this paper is the following question posed in [Erdős and Purdy 1971] (see also [Brass et al. 2005; Erdős and Purdy 1975; 1976; 1977; 1978; 1995]):

**Question.** *What is the maximum number of mutually congruent  $k$ -simplices with vertices from among a set of  $n$  points in  $\mathbb{R}^d$ ?*

In Section 6 we shall see that Theorem 1.1 (ii) implies that for a large class of finite sets  $P$  of cardinality  $n$  in  $\mathbb{R}^2$ , namely those that are  $s$ -adaptable, the maximum number of mutually congruent triangles determined by points of  $P$  is  $O(n^{\frac{9}{7}+\epsilon})$ .

For explicit quantitative connections between discrete and continuous finite configuration problems in other contexts, see, for example, [Hofmann and Iosevich 2005; Iosevich and Łaba 2005; Iosevich et al. 2007].

**Notation.** Throughout the paper,  $X \lesssim Y$  means that there exists  $C > 0$  such that  $X \leq CY$ , and  $X \approx Y$  means that  $X \lesssim Y$  and  $Y \lesssim X$ . We also define  $X \lesssim\lesssim Y$  as follows. If  $X$  and  $Y$  are quantities that depend on a large parameter  $N$ , then  $X \lesssim\lesssim Y$  means that for every  $\epsilon > 0$  there exists  $C_\epsilon > 0$  such that  $X \leq C_\epsilon N^\epsilon Y$ , while if  $X$  and  $Y$  depend on a small parameter  $\delta$ , then  $X \lesssim\lesssim Y$  means that for every  $\epsilon > 0$  there exists  $C_\epsilon > 0$  such that  $X \leq C_\epsilon \delta^{-\epsilon} Y$  as  $\delta$  tends to 0.

## 2. Reduction of the proof to the estimation of a trilinear form

We shall work exclusively with Frostman measures. Recall that a probability measure  $\mu$  on a compact set  $E \subset \mathbb{R}^d$  is a *Frostman measure* if, for any ball  $B_\delta$  of radius  $\delta$ ,

$$\mu(B_\delta) \lesssim\lesssim \delta^s, \tag{2-1}$$

where  $s = \dim_{\mathcal{H}}(E)$ . For discussion and proof of the existence of such measures see, e.g., [Mattila 1995].

Let  $\mu$  be a Frostman measure on  $E$ . Cover  $T_2(E)$  by cubes of the form

$$(t_{12}^l - \epsilon_l, t_{12}^l + \epsilon_l) \times (t_{13}^l - \epsilon_l, t_{13}^l + \epsilon_l) \times (t_{23}^l - \epsilon_l, t_{23}^l + \epsilon_l).$$

It follows that

$$1 = (\mu \times \mu \times \mu)\{E \times E \times E\} \leq \sum_l (\mu \times \mu \times \mu)\{(x^1, x^2, x^3) : t_{ij}^l - \epsilon_l \leq |x^i - x^j| \leq t_{ij}^l + \epsilon_l, \forall i, j\}. \tag{2-2}$$

Suppose that we could show that this expression is  $\lesssim \sum \epsilon_j^3$ . It would then follow, by definition of sets of measure 0, that the three dimensional Lebesgue measure of  $T_2(E)$  is positive.

In light of (2-1), to establish the positive measure of  $T_2(E)$  we may assume that  $t_{ij} \geq c > 0$ . To see this, observe that if each  $t_{ij}$  is  $\leq r$ , then fixing  $x^1$  results in  $x^2$  and  $x^3$  being contained in a ball of radius  $r$  centered at  $x^1$ . It follows that

$$(\mu \times \mu \times \mu)\{E \times E \times E\} \leq \sum_l (\mu \times \mu \times \mu)\{(x^1, x^2, x^3) : t_{ij}^l - \epsilon_l \leq |x^i - x^j| \leq t_{ij}^l + \epsilon_l, \forall i, j\} \leq Cr^{2s},$$

and taking  $r$  to be small enough, this expression is  $\leq \frac{1}{10}$ . This means that in place of equality on the left-hand side of (2-2), we have an inequality with 1 replaced by  $\frac{9}{10}$ , and the rest of the argument goes through as before.

Therefore, the proof of Theorem 1.1 (i) is reduced to proving the trilinear estimate

$$\Lambda_t^\epsilon(\mu, \mu, \mu) := \iiint \sigma_{t_{12}}^\epsilon(x^1 - x^2) \sigma_{t_{13}}^\epsilon(x^1 - x^3) \sigma_{t_{23}}^\epsilon(x^2 - x^3) d\mu(x^1) d\mu(x^2) d\mu(x^3) \lesssim 1. \quad (2-3)$$

Here,  $t = (t_{12}, t_{13}, t_{23})$ ,  $\sigma_r$  is arc length measure on the circle of radius  $r$  in  $\mathbb{R}^2$  and  $\sigma_r^\epsilon = \sigma_r * \rho_\epsilon$ , where  $\rho_\epsilon(x) = \epsilon^{-2} \rho(x/\epsilon)$  is an approximate identity with  $\rho \in C_0^\infty(\{|x| \leq 1\})$ ,  $\rho \geq 0$ ,  $\int \rho(x) dx = 1$ . Note that the right-hand side is 1 instead of  $\epsilon^3$  because the characteristic function of the annulus of radius  $t_{ij}$  and thickness  $\epsilon$ , divided by  $\epsilon$ , is dominated by  $\sigma_{t_{ij}}^\epsilon$ . We now turn to the proof of (2-3).

### 3. Reducing the trilinear form estimate to a bilinear operator estimate

Define trilinear forms

$$\Lambda_t^\epsilon(f_1, f_2, f_3) := \iiint \sigma_{t_{12}}^\epsilon(x^1 - x^2) \sigma_{t_{13}}^\epsilon(x^1 - x^3) \sigma_{t_{23}}^\epsilon(x^2 - x^3) f_1(x^1) f_2(x^2) f_3(x^3) dx^1 dx^2 dx^3, \quad (3-1)$$

and consider  $\Lambda_t^\epsilon(\mu_{-\alpha}^\delta, \mu, \mu_\alpha^\delta)$ , where

$$\mu_\alpha(x) := \frac{2^{(2-\alpha)/2}}{\Gamma(\alpha/2)} (\mu * |\cdot|^{-2+\alpha})(x), \quad (3-2)$$

initially defined for  $\text{Re } \alpha > 0$ , is extended to the complex plane by analytic continuation, and

$$\mu^\delta(x) := \mu * \rho_\delta(x),$$

and  $\rho_\delta(x) = \delta^{-2} \rho(x/\delta)$  is an approximate identity as above. Observe that  $\widehat{\mu}_\alpha^\delta(\xi) = C_\alpha \widehat{\mu}(\xi) \widehat{\rho}(\delta\xi) |\xi|^{-\alpha}$ , where

$$C_\alpha = \frac{2\pi \cdot 2^{\alpha/2}}{\Gamma((2-\alpha)/2)}. \quad (3-3)$$

(See [Gelfand and Shilov 1958] for relevant calculations.) This shows, in view of Plancherel, that  $\mu_\alpha^\delta$  is an  $L^2(\mathbb{R}^2)$  function with bounds depending on  $\delta$ . Moreover, since we have compact support, this shows that one has a trivial finite upper bound on the trilinear form with constants depending on  $\delta$ . Taking the

modulus in (3-2), we see that

$$|\mu_\alpha^\delta(x)| \leq \left| \frac{2^{(2-\alpha)/2}}{\Gamma(\alpha/2)} \right| (\mu^\delta * |\cdot|^{-2+\operatorname{Re}\alpha})(x) = 2^{(2-\operatorname{Re}\alpha)/2} \frac{\Gamma(\operatorname{Re}\alpha/2)}{|\Gamma(\alpha/2)|} \mu_{\operatorname{Re}\alpha}^\delta(x)$$

and note that the right-hand side is nonnegative.

Now define

$$F(\alpha) := \Lambda_t^\epsilon(\mu_{-\alpha}^\delta, \mu, \mu_\alpha^\delta) = \langle B(\mu_{-\alpha}^\delta, \mu^\delta), \mu_\alpha^\delta \rangle, \tag{3-4}$$

where  $\langle \cdot, \cdot \rangle$  is the  $L^2(\mathbb{R}^2)$  inner product and  $B$  is the bilinear operator given by the relation

$$B^\epsilon(f, g)(x) := \iint f(x-u) g(x-v) \sigma_a^\epsilon(u) \sigma_b^\epsilon(v) \sigma^\epsilon(u-v) du dv. \tag{3-5}$$

Here, for simplicity we have rescaled one side of the triangle to have unit length; the other two,  $a, b \lesssim 1$ , are bounded away from 0.

Our main bilinear estimate is the following, which is proved in Section 4.

**Theorem 3.1.** *Let  $B^\epsilon$  be defined as above and suppose that  $f, g \geq 0$ . Then*

$$\|B^\epsilon(f, g)\|_{L^1(\mathbb{R}^2)} \lesssim \|f\|_{L^2_{-\beta_1}(\mathbb{R}^2)} \cdot \|g\|_{L^2_{-\beta_2}(\mathbb{R}^2)} \quad \text{if } \beta_1 + \beta_2 = \frac{1}{2}, \beta_1, \beta_2 \geq 0, \tag{3-6}$$

with constants independent of  $\epsilon$ .

Using (3-6), we see that, with  $F(\alpha)$  defined as in (3-4), we have

$$|F(\alpha)| \lesssim \langle B^\epsilon(\mu_{-\operatorname{Re}\alpha}^\delta, \mu^\delta), \mu_{\operatorname{Re}\alpha}^\delta \rangle \leq \|B^\epsilon(\mu_{-\operatorname{Re}\alpha}^\delta)\|_{L^1(\mathbb{R}^2)} \cdot \|\mu_{\operatorname{Re}\alpha}^\delta\|_{L^\infty(\mathbb{R}^2)}, \tag{3-7}$$

where the  $\lesssim$  symbol includes factors of the gamma functions.

**Lemma 3.2.** Suppose that  $\mu$  is a Frostman measure on a set of Hausdorff dimension  $> \frac{7}{4}$ . Then

$$\|\mu_\alpha^\delta\|_\infty \lesssim 1 \quad \text{if } \operatorname{Re}\alpha = \frac{1}{4}.$$

To prove the lemma, observe that if  $\operatorname{Re}\alpha = \frac{1}{4}$ ,

$$\mu_\alpha^\delta(x) \leq \int |x-y|^{-2+\frac{1}{4}} d\mu^\delta(y) \approx \sum_m 2^{m(2-\frac{1}{4})} \int_{|x-y|\approx 2^{-m}} d\mu^\delta(y) \lesssim \sum_m 2^{m(2-\frac{1}{4})} 2^{-ms},$$

and this is  $\lesssim 1$  since  $\mu$  is a Frostman measure on a set of Hausdorff dimension  $> \frac{7}{4}$ . Substituting this into (3-7) and applying (3-6) with  $\beta_1 = \frac{3}{8}, \beta_2 = \frac{1}{8}$ , we see that if  $\operatorname{Re}\alpha = \frac{1}{4}$ ,

$$|F(\alpha)| \leq \|B^\epsilon(\mu_{-1/4}^\delta, \mu^\delta)\|_{L^1(\mathbb{R}^2)} \lesssim \|\mu_{-1/4}^\delta\|_{L^2_{-3/8}(\mathbb{R}^2)} \cdot \|\mu^\delta\|_{L^2_{-1/8}(\mathbb{R}^2)}. \tag{3-8}$$

A straightforward calculation using the definition of  $\mu_\alpha^\delta$  from above shows that the square of either of the terms in (3-8) is bounded by

$$\iint |x-y|^{-7/4} d\mu(x) d\mu(y),$$

which is the energy integral of  $\mu$  of order  $\frac{7}{4}$ . This integral is bounded since the Hausdorff dimension of  $E$  is greater than  $\frac{7}{4}$  and  $\mu$  is a Frostman measure; see, e.g., [Falconer 1985a; Mattila 1995].

By symmetry, the same bound holds when  $\operatorname{Re} \alpha = -\frac{1}{4}$  because we can reverse the roles of  $d\mu(x^1)$  and  $d\mu(x^3)$ . When  $-\frac{1}{4} < \operatorname{Re} \alpha < \frac{1}{4}$ , we use the fact that  $|F(\alpha)|$  is bounded from above with constants depending on  $\delta$  as we noted in the beginning of this section. By the three lines lemma (see the version in Hirschman [1953]), for example) we conclude that,  $\Lambda_t^\epsilon(\mu, \mu, \mu) \lesssim 1$ , which completes the proof of Theorem 1.1, conditional on Theorem 3.1, which we now prove.

#### 4. Estimating the bilinear operator

Since we are assuming  $f, g \geq 0$ , we have

$$\|B^\epsilon(f, g)\|_{L^1(\mathbb{R}^2)} = \iiint f(x-u) g(x-v) K^\epsilon(u, v) du dv dx, \tag{4-1}$$

where

$$K^\epsilon(u, v) = \sigma_a^\epsilon(u) \sigma_b^\epsilon(v) \sigma^\epsilon(u-v);$$

recall that we scaled one of the sigmas to the unit radius. Let  $\psi \in C_0^\infty(\{|x| \leq 2\})$ ,  $\psi \geq 0$ ,  $\psi \equiv 1$  on  $\{|x| \leq 1\}$ . Then it suffices to estimate

$$\iiint f(x-u) g(x-v) K^\epsilon(u, v) du dv \psi(x/R) dx$$

with bounds independent of  $R \geq 1$ . Using Fourier inversion, the expression (4-1) equals

$$R^2 \iint \hat{f}(\xi) \hat{g}(\eta) \widehat{K}^\epsilon(\xi, \eta) \widehat{\psi}(R(\xi + \eta)) d\xi d\eta. \tag{4-2}$$

**Lemma 4.1.** Let  $K(u, v) = K^0(u, v)$ , interpreted in the sense of distributions. We have

$$\widehat{K}(\xi, \eta) = \sum_{\pm} \widehat{\sigma}(U_{a,b}^\pm(\xi, \eta)), \tag{4-3}$$

where

$$U_{a,b}^\pm : \mathbb{R}^4 \rightarrow \mathbb{R}^2$$

are defined by

$$U_{a,b}^\pm(\xi, \eta) = \left( a\xi_1 + b\eta_1\gamma_{a,b} \pm b\eta_2\sqrt{1-\gamma_{a,b}^2}, a\xi_2 - b\eta_1\sqrt{1-\gamma_{a,b}^2} \mp b\gamma_{a,b}\eta_2 \right) \tag{4-4}$$

with  $\gamma_{a,b} = (a^2 + b^2 - 1)/(2ab)$ . Consequently, using stationary phase, we see that

$$|\widehat{K}^\epsilon(\xi, \eta) \widehat{\psi}(R(\xi + \eta))| \lesssim (1 + |\xi| + |\eta|)^{-\frac{1}{2}} \tag{4-5}$$

uniformly for  $R \geq 1$ .

Recalling that, by the method of stationary phase (see, e.g., [Sogge 1993; Stein 1993]),

$$|\widehat{\sigma}(\xi)| \lesssim (1 + |\xi|)^{-\frac{1}{2}},$$

one sees that (4-5) will immediately follow from (4-3) and (4-4).

To prove the lemma, parametrize the Cartesian product of two circles as

$$\{(a \cos \theta, a \sin \theta, b \cos \phi, b \sin \phi)\}.$$

The restriction imposed by  $\sigma(u - v)$  says that

$$\text{dist}((a \cos \theta, a \sin \theta), (b \cos \phi, b \sin \phi)) = 1,$$

which implies via standard trigonometric identities that

$$\cos(\theta - \phi) = \frac{a^2 + b^2 - 1}{2ab} =: \gamma_{a,b}$$

and thus  $\theta - \phi = \pm \theta_{a,b} := \cos^{-1}(\gamma_{a,b})$ . It follows that

$$\begin{aligned} \widehat{K}(\xi, \eta) &= \int_0^{2\pi} \exp(2\pi i(a \cos(\theta)\xi_1 + a \sin(\theta)\xi_2 + b \cos(\theta + \theta_{a,b})\eta_1 + b \sin(\theta + \theta_{a,b})\eta_2)) d\theta \\ &= \widehat{\sigma}(U_{a,b}(\xi, \eta)), \end{aligned}$$

as claimed. This proves (4-3). The estimate (4-5) follows in the same way since  $\sigma_a^\epsilon(x) = \sigma_a * \rho_\epsilon(x)$ .

Using Lemma 4.1, Cauchy–Schwarz, and the assumption  $\beta_1 + \beta_2 = \frac{1}{2}$ ,  $\beta_1, \beta_2 \geq 0$ , we estimate the square of (4-2) by

$$\begin{aligned} \int |\widehat{f}(\xi)|^2 \left\{ R^2 \int |\widehat{K}^\epsilon(\xi, \eta)|^{4\beta_1} |\widehat{\psi}(R(\xi+\eta))| d\eta \right\} d\xi & \int |\widehat{g}(\eta)|^2 \left\{ R^2 \int |\widehat{K}^\epsilon(\xi, \eta)|^{4\beta_2} |\widehat{\psi}(R(\xi+\eta))| d\xi \right\} d\eta \\ & \lesssim \int |\widehat{f}(\xi)|^2 (1 + |\xi|)^{-2\beta_1} d\xi \int |\widehat{g}(\eta)|^2 (1 + |\eta|)^{-2\beta_2} d\eta \\ & = \|f\|_{L_{-\beta_1}^2(\mathbb{R}^2)}^2 \cdot \|g\|_{L_{-\beta_2}^2(\mathbb{R}^2)}^2, \end{aligned}$$

as desired, completing the proof of Theorem 3.1 and thus the proof of Theorem 1.1.

### 5. Sharpness of the trilinear estimate (2-3)

To understand the extent to which this result is sharp, we use a variant of the construction obtained for the case  $k = 1$ ,  $d = 2$  in [Mattila 1987]. See [Iosevich and Senger 2010; Erdoğan et al. ≥ 2012], where this issue is studied comprehensively. Let  $C_\alpha$  denote the standard  $\alpha$ -dimensional Cantor set contained in the interval  $[0, 1]$ . Let

$$F_\alpha = (C_\alpha - 1) \cup (C_\alpha + 1),$$

and let  $\mu$  denote the natural measure on this set. Let  $E = F_\alpha \times F_\beta$ . Observe that we can fit a  $\sqrt{\epsilon}$  by  $\epsilon$  rectangle in the annulus  $\{x : 1 \leq |x| \leq 1 + \epsilon\}$  near  $(0, \pm 1)$  and also near  $(\pm 1, 0)$ .

Fix  $x$  and observe that

$$(\mu \times \mu) \left\{ (y, z) : 1 \leq |x - z| \leq 1 + \epsilon; 1 \leq |x - y| \leq 1 + \epsilon; \sqrt{2} \leq |y - z| \leq \sqrt{2} + \epsilon \right\} \approx \epsilon^{\alpha/2 + \beta} \cdot \epsilon^{\alpha + \beta} = \epsilon^{\frac{3}{2}\alpha + 2\beta}.$$

Integrating in  $x$ , we see that

$$(\mu \times \mu \times \mu) \{ (x, y, z) : 1 \leq |x - z| \leq 1 + \epsilon; 1 \leq |x - y| \leq 1 + \epsilon; \sqrt{2} \leq |y - z| \leq \sqrt{2} + \epsilon \} \gtrsim \epsilon^{\frac{3}{2}\alpha + 2\beta}.$$

We need this quantity to be  $\lesssim \epsilon^3$ , which leads to the equation

$$\frac{3}{2}\alpha + 2\beta \geq 3.$$

Choosing  $\alpha = 1$  and  $\beta = \frac{3}{4}$  shows that the inequality (2-3) does not in general hold if  $s < \frac{7}{4}$ . It is important to note that this does not prove that  $\mathcal{L}^3(T_2(E)) > 0$  does not in general hold if  $s < \frac{7}{4}$ .

We stress that the calculation above pertains to the trilinear expression (2-3). We do not know of any example that shows that  $\mathcal{L}^3(T_2(E))$  is not in general positive if the Hausdorff dimension of  $E$  is greater than one. The discrepancy here is not particularly surprising because it already takes place in the study of distance sets. For example, as we point out in the introduction, it is known that if the Hausdorff dimension of  $E \subset \mathbb{R}^2$  is  $\leq 1$ , then it is not in general true that  $\mathcal{L}^1(\Delta(E)) > 0$ . A result due to Wolff [1999] says that if the Hausdorff dimension of  $E$  is greater than  $\frac{4}{3}$ , then  $\mathcal{L}^1(\Delta(E)) > 0$ . On the other hand, an example due to Mattila [1987] shows that if the Hausdorff dimension of  $E$  is less than  $\frac{3}{2}$  and  $\mu$  is a Frostman measure on  $E$ , then

$$\limsup_{\epsilon \rightarrow 0} \epsilon^{-1} (\mu \times \mu) \{ (x, y) \in E \times E : 1 \leq |x - y| \leq 1 + \epsilon \} = \infty. \tag{5-1}$$

We note that (5-1) is the analogue of (1-3). It says that the distance measure, defined by

$$\int f(t) dv(t) = \iint f(|x - y|) d\mu(x) d\mu(y),$$

has Radon–Nikodym derivative which is not in  $L^\infty$ .

### 6. Application to discrete geometry

**Definition 6.1.** Let  $P$  be a set of  $n$  points contained in  $[0, 1]^2$ . Define the measure

$$d\mu_P^s(x) = n^{-1} \cdot n^{d/s} \cdot \sum_{p \in P} \chi_{B_{n^{-1/s}}}(x) dx, \tag{6-1}$$

where  $\chi_{B_{n^{-1/s}}}(x)$  is the characteristic function of the ball of radius  $n^{-1/s}$  centered at  $p$ . We say that  $P$  is *s-adaptable* [Iosevich et al. 2007] if

$$I_s(\mu_P) = \iint |x - y|^{-s} d\mu_P^s(x) d\mu_P^s(y) < \infty.$$

This is equivalent to the statement

$$n^{-2} \sum_{p \neq p' \in P} |p - p'|^{-s} \lesssim 1. \tag{6-2}$$

To understand this condition in clearer geometric terms, suppose that  $P$  comes from a 1-separated set



$A$ , scaled down by its diameter. Then the condition (6-2) takes the form

$$n^{-2} \sum_{a \neq a' \in A} |a - a'|^{-s} \lesssim (\text{diameter}(A))^{-s}. \quad (6-3)$$

This says  $P$  is  $s$ -adaptable if it is a scaled 1-separated set where the expected value of the distance between two points raised to the power  $-s$  is comparable to the value of the diameter raised to the power of  $-s$ . This basically means that for the set to be  $s$ -adaptable, clustering is not allowed to be too severe.

To put it in more technical terms,  $s$ -adaptability means that a discrete point set  $P$  can be thickened into a set which is uniformly  $s$ -dimensional in the sense that its energy integral of order  $s$  is finite. Unfortunately, it is shown in [Iosevich et al. 2007] that there exist finite point sets which are not  $s$ -adaptable for certain ranges of the parameter  $s$ . The point is that the notion of Hausdorff dimension is much more subtle than the simple “size” estimate. However, many natural classes of sets are  $s$ -adaptable. For example, homogeneous sets studied by Solymosi and Vu [2004] and others are  $s$ -adaptable for all  $0 < s < d$ . See also [Iosevich et al. 2009], where  $s$ -adaptability of homogeneous sets is used to extract discrete incidence theorems from Fourier-type bounds.

Before we state the discrete result that follows from Theorem 1.1, let us briefly review what is known. If  $P$  is set of  $n$  points in  $[0, 1]^2$ , let  $u_{2,2}(n)$  denote the number of times a fixed triangle can arise among points of  $P$ . It is not hard to see that

$$u_{2,2}(n) = O(n^{\frac{4}{3}}). \quad (6-4)$$

This follows easily from the fact that a single distance cannot arise more than  $O(n^{\frac{4}{3}})$  times, which, in turn, follows from the celebrated Szemerédi–Trotter incidence theorem. See [Brass et al. 2005] and the references therein. By the pigeonhole principle, one can conclude that

$$\#T_2(P) \gtrsim \frac{n^3}{n^{4/3}} = n^{\frac{5}{3}}. \quad (6-5)$$

However, it is not difficult to see that one can do quite a bit better as far as the lower bound on  $\#T_2(P)$  is concerned. It is shown in [Brass et al. 2005, p. 263] that

$$\#T_2(P) \gtrsim n \cdot \#\{|x - y| : x, y \in P\}.$$

Guth and Katz [2011] have recently settled the Erdős distance conjecture, proving that

$$\#\{|x - y| : x, y \in P\} \gtrsim \frac{n}{\log n},$$

and it follows that

$$\#T_2(P) \gtrsim \frac{n^2}{\log n},$$

which, up to logarithmic factors, is the optimal bound. However, Theorem 1.1 does allow us to obtain an upper bound on  $u_{2,2}$  for  $s$ -adaptable sets that is better than the one in (6-4). Before we state the main result of this section, we need the following definition.

**Definition 6.2.** Let  $P$  be a subset of  $[0, 1]^2$  consisting of  $n$  points. Let  $\delta > 0$  and define

$$u_{2,2}^\delta(n) = \#\{(x^1, x^2, x^3) \in P \times P \times P : t_{ij} - \delta \leq |x^i - x^j| \leq t_{ij} + \delta\},$$

where the dependence on  $t = \{t_{ij}\}$  is suppressed.

Observe that obtaining an upper bound for  $u_{2,2}^\delta(n)$  with arbitrary  $t_{ij}$  immediately implies the same upper bound on  $u_{2,2}(n)$  defined above. The main result of this section is the following.

**Corollary 6.3.** Suppose  $P \subset [0, 1]^2$  is  $s$ -adaptable for  $s = \frac{7}{4} + a$  for every sufficiently small  $a > 0$ . Then for every  $b > 0$ , there exists  $C_b > 0$  such that

$$u_{2,2}^{n^{-\frac{4}{7}-b}}(n) \leq C_b n^{\frac{9}{7}+b}. \quad (6-6)$$

The proof follows from Theorem 1.1 in the following way. Let  $E$  denote the support of  $d\mu_P^s$ , defined as in (6-1) above. We know that if  $s > \frac{7}{4}$ , then

$$(\mu_P^s \times \mu_P^s \times \mu_P^s) \{(x^1, x^2, x^3) : t_{ij} \leq |x^i - x^j| \leq t_{ij} + \epsilon\} \lesssim \epsilon^3. \quad (6-7)$$

Taking  $\epsilon = n^{-1/s}$ , we see that the left-hand side is

$$\approx n^{-3} \cdot u_{2,2}^{n^{-1/s}}(n),$$

and we conclude that

$$u_{2,2}^{n^{-1/s}}(n) \lesssim n^{3-3/s},$$

which yields the desired result since  $s = \frac{7}{4} + a$ .

As we note above, this result is stronger than the previously known  $u_{2,2}(n) \lesssim n^{\frac{4}{3}}$ . However, our result holds under an additional restriction that  $P$  is  $s$ -adaptable. We hope to address this issue in a subsequent paper.

## References

- [Bourgain 1986] J. Bourgain, “A Szemerédi type theorem for sets of positive density in  $\mathbb{R}^k$ ”, *Israel J. Math.* **54**:3 (1986), 307–316. MR 87j:11012 Zbl 0609.10043
- [Bourgain 1994] J. Bourgain, “Hausdorff dimension and distance sets”, *Israel J. Math.* **87**:1-3 (1994), 193–201. MR 95h:28008 Zbl 0807.28004
- [Brass et al. 2005] P. Brass, W. Moser, and J. Pach, *Research problems in discrete geometry*, Springer, New York, 2005. MR 2006i:52001 Zbl 1086.52001
- [Erdős and Purdy 1971] P. Erdős and G. Purdy, “Some extremal problems in geometry”, *J. Combin. Theory Ser. A* **10** (1971), 246–252. MR 43 #1045 Zbl 0219.05006
- [Erdős and Purdy 1975] P. Erdős and G. Purdy, “Some extremal problems in geometry, III”, pp. 291–308 in *Proceedings of the Sixth Southeastern Conference on Combinatorics, Graph Theory and Computing* (Boca Raton, FL, 1975), edited by F. Hoffman et al., Congressus Numerantium **14**, Utilitas Math., Winnipeg, MB, 1975. MR 52 #13650 Zbl 0328.05018
- [Erdős and Purdy 1976] P. Erdős and G. Purdy, “Some extremal problems in geometry, IV”, pp. 307–322 in *Proceedings of the Seventh Southeastern Conference on Combinatorics, Graph Theory and Computing* (Baton Rouge, LA, 1976), edited by F. Hoffman et al., Congressus Numerantium **17**, Utilitas Math., Winnipeg, MB, 1976. MR 55 #10292 Zbl 0345.52007

- [Erdős and Purdy 1977] P. Erdős and G. Purdy, “Some extremal problems in geometry, V”, pp. 569–578 in *Proceedings of the Eighth Southeastern Conference on Combinatorics, Graph Theory and Computing* (Baton Rouge, LA, 1977), edited by F. Hoffman et al., *Congressus Numerantium* **19**, Utilitas Math., Winnipeg, MB, 1977. MR 57 #16104 Zbl 0403.52006
- [Erdős and Purdy 1978] P. Erdős and G. Purdy, “Some combinatorial problems in the plane”, *J. Combin. Theory Ser. A* **25:2** (1978), 205–210. MR 58 #21645 Zbl 0422.05023
- [Erdős and Purdy 1995] P. Erdős and G. Purdy, “Extremal problems in combinatorial geometry”, pp. 809–874 in *Handbook of combinatorics, 1–2*, edited by R. L. Graham et al., Elsevier, Amsterdam, 1995. MR 96m:52025 Zbl 0852.52009
- [Erdogan 2005] M. B. Erdoğan, “A bilinear Fourier extension theorem and applications to the distance set problem”, *Internat. Math. Res. Notices* **2005:23** (2005), 1411–1425. MR 2006h:42020 Zbl 1129.42353
- [Erdogan et al. 2011] B. Erdoğan, D. Hart, and A. Iosevich, “Multi-parameter projection theorems with applications to sums-products and finite point configurations in the Euclidean setting”, preprint, 2011. arXiv 1106.5544
- [Erdogan et al.  $\geq$  2012] B. Erdoğan, A. Iosevich, and K. Taylor, “Finite point configurations and dimensional inequalities in Euclidean space”, To appear in the volume in honor of Kostya Oskolkov’s 65th birthday, edited by D. Bilyk and A. Stokolos, Springer.
- [Eswarathasan et al. 2011] S. Eswarathasan, A. Iosevich, and K. Taylor, “Fourier integral operators, fractal sets, and the regular value theorem”, *Adv. Math.* **228:4** (2011), 2385–2402. MR 2836125 Zbl 05965623
- [Falconer 1985a] K. J. Falconer, *The geometry of fractal sets*, Cambridge Tracts in Mathematics **85**, Cambridge University Press, Cambridge, 1985. MR 88d:28001 Zbl 0587.28004
- [Falconer 1985b] K. J. Falconer, “On the Hausdorff dimensions of distance sets”, *Mathematika* **32:2** (1985), 206–212. MR 87j:28008 Zbl 0605.28005
- [Furstenberg et al. 1990] H. Furstenberg, Y. Katznelson, and B. Weiss, “Ergodic theory and configurations in sets of positive density”, pp. 184–198 in *Mathematics of Ramsey theory*, edited by J. Nešetřil and V. Rödl, Algorithms Combin. **5**, Springer, Berlin, 1990. MR 1083601 Zbl 0738.28013
- [Gelfand and Shilov 1958] I. M. Gelfand and G. E. Shilov, *Обобщенные функции и действия над ними, Обобщенные функции 1*, Gosudarstv. Izdat. Fiz.-Mat. Lit., Moscow, 1958. Translated in *Generalized functions, 1: Properties and operations*, Academic Press, New York, 1964. MR 20 #4182 Zbl 0091.11103
- [Guth and Katz 2011] L. Guth and N. Katz, “On the Erdős distinct distance problem in the plane”, preprint, 2011. arXiv 1011.4105
- [Hirschman 1953] I. I. Hirschman, Jr., “A convexity theorem for certain groups of transformations”, *J. Anal. Math.* **2** (1953), 209–218. MR 15,295b Zbl 0052.06302
- [Hofmann and Iosevich 2005] S. Hofmann and A. Iosevich, “Circular averages and Falconer/Erdős distance conjecture in the plane for random metrics”, *Proc. Amer. Math. Soc.* **133:1** (2005), 133–143. MR 2005k:42031 Zbl 1096.28004
- [Iosevich and Łaba 2005] A. Iosevich and I. Łaba, “ $K$ -distance sets, Falconer conjecture, and discrete analogs”, *Integers* **5:2** (2005), A8. MR 2006i:42033 Zbl 1139.28002
- [Iosevich and Senger 2010] A. Iosevich and S. Senger, “Sharpness of Falconer’s estimate in continuous and arithmetic settings, geometric incidence theorems and distribution of lattice points in convex domains”, preprint, 2010. arXiv 1006.1397
- [Iosevich et al. 2007] A. Iosevich, M. Rudnev, and I. Uriarte-Tuero, “Theory of dimension for large discrete sets and applications”, preprint, 2007. arXiv 0707.1322
- [Iosevich et al. 2009] A. Iosevich, H. Jorati, and I. Łaba, “Geometric incidence theorems via Fourier analysis”, *Trans. Amer. Math. Soc.* **361:12** (2009), 6595–6611. MR 2011b:42074 Zbl 1180.42014
- [Katz and Tardos 2004] N. H. Katz and G. Tardos, “A new entropy inequality for the Erdős distance problem”, pp. 119–126 in *Towards a theory of geometric graphs*, edited by J. Pach, Contemp. Math. **342**, Amer. Math. Soc., Providence, RI, 2004. MR 2005f:52033 Zbl 1069.52017
- [Mattila 1987] P. Mattila, “Spherical averages of Fourier transforms of measures with finite energy; dimension of intersections and distance sets”, *Mathematika* **34:2** (1987), 207–228. MR 90a:42009 Zbl 0645.28004
- [Mattila 1995] P. Mattila, *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*, Cambridge Studies in Advanced Mathematics **44**, Cambridge University Press, Cambridge, 1995. MR 96h:28006 Zbl 0819.28004

- [Sogge 1993] C. D. Sogge, *Fourier integrals in classical analysis*, Cambridge Tracts in Mathematics **105**, Cambridge University Press, Cambridge, 1993. MR 94c:35178 Zbl 0783.35001
- [Solymosi and Tóth 2001] J. Solymosi and C. D. Tóth, “Distinct distances in the plane”, *Discrete Comput. Geom.* **25**:4 (2001), 629–634. MR 2002c:52020 Zbl 0988.52027
- [Solymosi and Vu 2004] J. Solymosi and V. Vu, “Distinct distances in high dimensional homogeneous sets”, pp. 259–268 in *Towards a theory of geometric graphs*, edited by J. Pach, Contemp. Math. **342**, Amer. Math. Soc., Providence, RI, 2004. MR 2005m:52026 Zbl 1064.52011
- [Stein 1993] E. M. Stein, *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals*, Princeton Mathematical Series **43**, Princeton University Press, Princeton, NJ, 1993. MR 95c:42002 Zbl 0821.42001
- [Székely 1997] L. A. Székely, “Crossing numbers and hard Erdős problems in discrete geometry”, *Combin. Probab. Comput.* **6**:3 (1997), 353–358. MR 98h:52030 Zbl 0882.52007
- [Wolff 1999] T. Wolff, “Decay of circular means of Fourier transforms of measures”, *Internat. Math. Res. Notices* **1999**:10 (1999), 547–567. MR 2000k:42016 Zbl 0930.42006
- [Ziegler 2006] T. Ziegler, “Nilfactors of  $\mathbb{R}^m$ -actions and configurations in sets of positive upper density in  $\mathbb{R}^m$ ”, *J. Anal. Math.* **99** (2006), 249–266. MR 2008k:37008 Zbl 1145.37005

Received 15 Sep 2010. Revised 1 Jul 2011. Accepted 9 Oct 2011.

ALLAN GREENLEAF: [allan@math.rochester.edu](mailto:allan@math.rochester.edu)

*Department of Mathematics, University of Rochester, Rochester, NY 14627, United States*

ALEX IOSEVICH: [iosevich@math.rochester.edu](mailto:iosevich@math.rochester.edu)

*Department of Mathematics, University of Rochester, Rochester, NY 14627, United States*



## ASYMPTOTIC DECAY FOR A ONE-DIMENSIONAL NONLINEAR WAVE EQUATION

HANS LINDBLAD AND TERENCE TAO

We consider the asymptotic behaviour of finite energy solutions to the one-dimensional defocusing nonlinear wave equation  $-u_{tt} + u_{xx} = |u|^{p-1}u$ , where  $p > 1$ . Standard energy methods guarantee global existence, but do not directly say much about the behaviour of  $u(t)$  as  $t \rightarrow \infty$ . Note that in contrast to higher-dimensional settings, solutions to the linear equation  $-u_{tt} + u_{xx} = 0$  do not exhibit decay, thus apparently ruling out perturbative methods for understanding such solutions. Nevertheless, we will show that solutions for the nonlinear equation behave differently from the linear equation, and more specifically that we have the average  $L^\infty$  decay  $\lim_{T \rightarrow +\infty} \frac{1}{T} \int_0^T \|u(t)\|_{L_x^\infty(\mathbb{R})} dt = 0$ , in sharp contrast to the linear case. An unusual ingredient in our arguments is the classical Rademacher differentiation theorem that asserts that Lipschitz functions are almost everywhere differentiable.

### 1. Introduction

Fix  $p > 1$ . We consider solutions  $u : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  to the one-dimensional defocusing nonlinear wave equation

$$-u_{tt} + u_{xx} = |u|^{p-1}u, \tag{1}$$

with the finite energy initial condition

$$\|u(0)\|_{H_x^1(\mathbb{R})} + \|u_t(0)\|_{L_x^2(\mathbb{R})} < \infty.$$

Standard energy methods (using the Sobolev embedding  $H_x^1 \subset L_x^\infty$ ) show that the initial value problem is locally well-posed in this energy class. Furthermore, by using the conservation of energy<sup>1</sup>

$$E[u] = E[u(t)] := \int_{\mathbb{R}} \mathbb{T}_{00}(t, x) dx, \tag{2}$$

where  $\mathbb{T}_{00}$  is the *energy density*

$$\mathbb{T}_{00} := \frac{1}{2}u_t^2 + \frac{1}{2}u_x^2 + \frac{1}{p+1}|u|^{p+1},$$

---

Lindblad is supported by NSF grant DMS-0801120. Tao is supported by NSF Research Award DMS-0649473, the NSF Waterman award and a grant from the MacArthur Foundation.

MSC2010: 35L05.

Keywords: nonlinear wave equation.

<sup>1</sup>In order to justify energy conservation for solutions which are in the energy class, one can use standard local well-posedness theory to approximate such solutions by classical (i.e., smooth and compactly supported) solutions (regularising the nonlinearity  $|u|^{p-1}u$  if necessary), derive energy conservation for the classical solutions, and then take strong limits. We omit the standard details. More generally, we shall perform manipulations such as integration by parts on finite energy solutions as if they were classical without any further comment.

it is easy to show that the  $H_x^1 \times L_x^2$  norm of  $u(t)$  does not blow up in finite time, and that the solution to (1) can be continued globally in time.

In this paper we study the asymptotic behaviour of finite energy solutions  $u$  to (1) as  $t \rightarrow \pm\infty$ . Of course, from the conservation of energy (2) we know that  $u(t)$  stays bounded in  $\dot{H}_x^1(\mathbb{R}) \cap L_x^{p+1}(\mathbb{R})$ , and thus (by the Gagliardo–Nirenberg inequality) bounded in  $L_x^\infty(\mathbb{R})$  for all time, but this does not settle the question of whether  $\|u(t)\|_{L_x^\infty(\mathbb{R})}$  exhibits any *decay* as  $t \rightarrow \pm\infty$ .

For the linear equation  $-u_{tt} + u_{xx} = 0$ , the solutions are of course travelling waves  $u(t, x) = f(x+t) + g(x-t)$ , which do not decay along light rays  $x = x_0 \pm t$ . In particular, for any nontrivial linear solution,  $\|u(t)\|_{L_x^\infty(\mathbb{R})}$  stays bounded away from zero. It is thus natural to ask whether the same behaviour occurs for solutions to the nonlinear Equation (1). However, an easy energy argument shows that the behaviour must be slightly different. Indeed, if we introduce the *momentum density* (or *energy current*)

$$\mathbb{T}_{01} = \mathbb{T}_{10} := u_t u_x$$

and the *momentum current*

$$\mathbb{T}_{11} := \frac{1}{2}u_t^2 + \frac{1}{2}u_x^2 - \frac{1}{p+1}|u|^{p+1}$$

we observe the conservation laws

$$\partial_t \mathbb{T}_{00} = \partial_x \mathbb{T}_{01}, \tag{3}$$

$$\partial_t \mathbb{T}_{01} = \partial_x \mathbb{T}_{11}. \tag{4}$$

From (3) and the fundamental theorem of calculus we have

$$\partial_t \int_{x < x_0+t} \mathbb{T}_{00}(t, x) dx = \mathbb{T}_{00}(t, x_0+t) + \mathbb{T}_{01}(t, x_0+t)$$

for all  $x_0, t \in \mathbb{R}$ . On the other hand, from the nonnegativity of  $\mathbb{T}_{00}$  we clearly have

$$0 \leq \int_{x < x_0+t} \mathbb{T}_{00}(t, x) dx \leq E[u].$$

From the fundamental theorem of calculus (and the monotone convergence theorem), we thus obtain

$$\int_{-\infty}^{\infty} \mathbb{T}_{00}(t, x_0+t) + \mathbb{T}_{01}(t, x_0+t) dt \leq E[u]$$

for all  $x_0 \in \mathbb{R}$ . From the pointwise inequality  $\mathbb{T}_{00} + \mathbb{T}_{01} \geq \frac{1}{p+1}|u|^{p+1}$  we conclude in particular the nonlinear decay estimate

$$\int_{-\infty}^{\infty} |u|^{p+1}(t, x_0+t) dt \leq (p+1)E[u] \tag{5}$$

for any  $x_0 \in \mathbb{R}$ . From reflection symmetry we also have

$$\int_{-\infty}^{\infty} |u|^{p+1}(t, x_0-t) dt \leq (p+1)E[u] \tag{6}$$

for any  $x_0 \in \mathbb{R}$ . We thus see that solutions to the nonlinear equation  $u$  must decay (on average, at least) along any light ray  $x = x_0 \pm t$ , in sharp contrast to solutions to the linear equation. This simple calculation

already reveals that the nonlinear equation has somewhat different asymptotic behaviour from the linear equation, and in particular that it is highly unlikely that one can asymptotically analyse the former as a perturbation of the latter. This is in contrast with the one-dimensional nonlinear Klein–Gordon equation, for which the decay can be leveraged to obtain asymptotic results; see for instance [Lindblad and Soffer 2005]. Another contrast is with the local theory, which asserts that singularities for the nonlinear wave equation propagate along the same light rays as for the linear one; see [Reed 1978].

The estimates (5), (6) imply that finite energy solutions  $u$  cannot concentrate on light rays  $\{(t, x_0 \pm t) : t \in \mathbb{R}\}$ . However, it is *a priori* conceivable that such solutions might still concentrate on other worldlines  $\{(t, x(t)) : t \in \mathbb{R}\}$ . Concentration on spacelike worldlines (in which  $|x'(t)| > 1$ ) are easily ruled out by finite speed of propagation (or by a modification of the arguments used to derive (5), (6)), but concentration on timelike worldlines (in which  $|x'(t)| < 1$ ) are not so obviously ruled out. Nevertheless, we are able to rule out this scenario by the following theorem, which is the main result of this paper.

**Theorem 1.1** (Average  $L_x^\infty$  decay). *Let  $u$  be a finite energy solution to (1), with an upper bound  $E[u] \leq E$  on the energy. Then*

$$\frac{1}{2T} \int_{t_0-T}^{t_0+T} \|u(t)\|_{L_x^\infty(\mathbb{R})} dt \leq c_{E,p}(T)$$

for all  $t_0 \in \mathbb{R}$  and  $T > 0$ , where  $c_{E,p} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  is a function depending only on the energy bound  $E$  and the exponent  $p$  such that  $c_{E,p}(t) \rightarrow 0$  as  $t \rightarrow \infty$ . In particular, we have

$$\lim_{T \rightarrow +\infty} \sup_{t_0 \in \mathbb{R}} \frac{1}{2T} \int_{t_0-T}^{t_0+T} \|u(t)\|_{L_x^\infty(\mathbb{R})} dt = 0.$$

The proof of this theorem will use energy estimates combined with a version of the Rademacher differentiation theorem (or Lebesgue differentiation theorem), that Lipschitz functions are almost everywhere differentiable. The basic idea is to observe that if  $u$  concentrates on a timelike worldline  $\{(t, x(t)) : t \in \mathbb{R}\}$ , then  $x$  should be Lipschitz, and thus mostly differentiable. This implies that  $u$  concentrates on certain parallelograms in spacetime; we will then use energy estimates to rule out such concentration.

In principle, the decaying bound  $c_{E,p}(T)$  could be made explicit, but this would require a quantitative version of the Rademacher differentiation theorem. Such results exist (see [Tao 2009] or [Tao 2008, Section 2.4]), but they are fairly weak (involving the inverse tower exponential function  $\log_{**}$ ). Presumably a more refined argument than the one given in this paper would give better bounds. For instance, it is plausible to conjecture that  $\|u(t)\|_{L_x^\infty(\mathbb{R})}$  should decay at a polynomial rate in  $t$ , at least in the perturbative regime when  $u$  is small.

We remark that our methods do not seem to give any precise asymptotics for the solution. Of course Theorem 1.1 indicates that the solution will not scatter to a linear solution, but it is not clear what the solution scatters to instead, even in the perturbative regime. It may be that techniques from nonlinear geometric optics could be useful to settle this question, but the extremely weak decay of the solution means that it would be very difficult for these methods to be made rigorous, at least until one can improve the results of Theorem 1.1 significantly.



## 2. Energy estimates

In this section we derive the basic energy estimates needed to establish Theorem 1.1. Henceforth we fix  $p$  and the finite energy solution  $u$ . We adopt the notation  $X \lesssim Y$  or  $X = O(Y)$  to denote the estimate  $|X| \leq CY$ , where  $C$  can depend on  $p$  and the energy bound  $E$ . Thus from energy conservation we obtain the bounds

$$\int_{\mathbb{R}} |u_t|^2(t, x) + |u_x|^2(t, x) + |u|^{p+1}(t, x) dx \lesssim 1 \quad (7)$$

for all  $t$ .

**Lemma 2.1** (Hölder continuity). *For all  $t, x, t', x' \in \mathbb{R}$  we have the pointwise bound*

$$u(t, x) = O(1) \quad (8)$$

and the Hölder continuity property

$$u(t, x) - u(t', x') = O(|t - t'|^{1/2} + |x - x'|^{1/2}). \quad (9)$$

*Proof.* The bound (8) follows immediately from (7) and the Gagliardo–Nirenberg inequality. Using the bound on  $|u_x|^2$  in (7) together with the fundamental theorem of calculus and the Cauchy–Schwarz inequality, we also have the spatial Hölder continuity bound

$$u(t, x) - u(t, x') = O(|x - x'|^{1/2}).$$

Thus to prove (9) it will suffice to show that

$$u(t_1, x_0) - u(t_2, x_0) = O((t_2 - t_1)^{1/2}) \quad (10)$$

for all  $t_2 > t_1$ . In view of (8) we may also assume  $t_2 = t_1 + O(1)$ .

Fix  $t_1, t_2$ . From (4) and the fundamental theorem of calculus we have

$$\partial_t \int_{x < x_0} \mathbb{T}_{01}(t, x) dx = \mathbb{T}_{11}(t, x_0);$$

integrating this in time and using (7) we obtain the bounds

$$\int_{t_1}^{t_2} \mathbb{T}_{11}(t, x_0) dt = O(1).$$

Combining this with (8) we conclude

$$\int_{t_1}^{t_2} u_t(t, x_0)^2 dt = O(1)$$

and (10) follows from the fundamental theorem of calculus and Cauchy–Schwarz.  $\square$

Now we prove a more advanced energy estimate.

**Proposition 2.2** (nonlinear energy decay in a parallelogram). *Let  $T \geq R \geq 1$ , let  $x_0, t_0 \in \mathbb{R}$ , and let  $v \in \mathbb{R}$  be a velocity. Then we have*

$$\int_{t_0-T}^{t_0+T} \int_{x_0+vt-R}^{x_0+vt+R} |u(t, x)|^{p+1} dx dt \lesssim R^{1/2} T^{1/2} + \frac{T}{R}. \quad (11)$$

**Remark 2.3.** Energy conservation (7) only gives the bound of  $O(T)$  for this integral, thus this proposition is nontrivial when  $T$  is much larger than  $R$ . A key point here is that the bounds do not blow up in the neighbourhood of the speed of light  $v = 1$ . It may be possible to improve the right-hand side of (11), and to also control other components of the energy, but the above bound will suffice for our purposes.

*Proof.* By translation invariance we can set  $x_0 = t_0 = 0$ . By reflection symmetry we may assume that  $v \geq 0$ .

Let  $\chi : \mathbb{R} \rightarrow \mathbb{R}$  be a nonnegative bump function supported on  $[-2, 2]$  which equals 1 on  $[-1, 1]$ , and let  $\psi(x) := \int_{y < x} \chi(y) dy$  be the antiderivative of  $\chi$ . From (4) and integration by parts we have

$$\partial_t \int_{\mathbb{R}} \psi\left(\frac{x-vt}{R}\right) \mathbb{T}_{01}(t, x) dx = -\frac{1}{R} \int_{\mathbb{R}} \chi\left(\frac{x-vt}{R}\right) (\mathbb{T}_{11}(t, x) + v\mathbb{T}_{01}(t, x)) dx;$$

integrating this against  $\chi(t/T)$  using (7) we conclude that

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \chi\left(\frac{t}{T}\right) \chi\left(\frac{x-vt}{R}\right) (\mathbb{T}_{11}(t, x) + v\mathbb{T}_{01}(t, x)) dx dt = O(R). \quad (12)$$

A similar argument using (3) instead of (4) yields

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \chi\left(\frac{t}{T}\right) \chi\left(\frac{x-vt}{R}\right) (\mathbb{T}_{01}(t, x) + v\mathbb{T}_{00}(t, x)) dx dt = O(R). \quad (13)$$

On the other hand, if we define the nonlinear null form

$$Q := (-\partial_{tt} + \partial_{xx})u^2 = -2u_t^2 + 2u_x^2 + 2|u|^{p+1}$$

then from integration by parts and (8) we have

$$\begin{aligned} \left| \int_{\mathbb{R}} \int_{\mathbb{R}} \chi\left(\frac{t}{T}\right) \chi\left(\frac{x-vt}{R}\right) Q(t, x) dx dt \right| &= \left| \int_{\mathbb{R}} \int_{\mathbb{R}} u^2(t, x) (-\partial_{tt} + \partial_{xx}) \left( \chi\left(\frac{t}{T}\right) \chi\left(\frac{x-vt}{R}\right) \right) dx dt \right| \\ &\lesssim \int_{-2T}^{2T} \int_{v-2R}^{v+2R} \frac{1}{T^2} + \frac{1}{R^2} dx dt \\ &\lesssim \frac{R}{T} + \frac{T}{R} \lesssim \frac{T}{R}. \end{aligned} \quad (14)$$

Let us compare  $|u|^{p+1}$  against the quantities

$$\begin{aligned} \mathbb{T}_{11} + v\mathbb{T}_{01} &= \frac{1}{2}u_t^2 + vu_tu_x + \frac{1}{2}u_x^2 - \frac{1}{p+1}|u|^{p+1}, \\ \mathbb{T}_{01} + v\mathbb{T}_{00} &= \frac{1}{2}vu_t^2 + u_tu_x + \frac{1}{2}vu_x^2 + \frac{v}{p+1}|u|^{p+1}, \\ Q &= -2u_t^2 + 2u_x^2 + 2|u|^{p+1}. \end{aligned}$$

We divide into three cases.

Case 1 (spacelike):  $v \geq 1$ . In this case, we can verify the pointwise bound

$$\frac{1}{p+1} |u|^{p+1} \leq \mathbb{T}_{01} + v \mathbb{T}_{00}$$

and so (11) follows immediately from (13) (note that  $R = O(R^{1/2}T^{1/2})$ ).

Case 2 (lightlike):  $1 - \frac{R^{1/2}}{2T^{1/2}} < v < 1$ . In this case we have the bound

$$\frac{v}{p+1} |u|^{p+1} \leq (\mathbb{T}_{01} + v \mathbb{T}_{00}) + O\left(\frac{R^{1/2}}{T^{1/2}} \mathbb{T}_{00}\right)$$

and so from (13) and (7) we have

$$\frac{v}{p+1} \int_{\mathbb{R}} \int_{\mathbb{R}} \chi\left(\frac{t}{T}\right) \chi\left(\frac{x-vt}{R}\right) |u(t, x)|^{p+1} dt dx \lesssim R + R^{1/2}T^{1/2}$$

and (11) follows.

Case 3 (timelike):  $0 \leq v \leq 1 - \frac{R^{1/2}}{2T^{1/2}}$ . Here we use the identity

$$(\mathbb{T}_{11} + v \mathbb{T}_{01}) - v(\mathbb{T}_{01} + v \mathbb{T}_{00}) + \frac{1-v^2}{4} Q = (1-v^2)u_x^2 + \frac{(p-1)(1-v^2)}{2(p+1)} |u|^{p+1}.$$

Taking the indicated linear combination of (12), (13), (14) and discarding  $(1-v^2)u_x^2$ , which is non-negative, we conclude that

$$\frac{(p-1)(1-v^2)}{2(p+1)} \int_{\mathbb{R}} \int_{\mathbb{R}} \chi\left(\frac{t}{T}\right) \chi\left(\frac{x-vt}{R}\right) |u(t, x)|^{p+1} dt dx \lesssim R + \frac{1-v^2}{4} \frac{T}{R}$$

and thus (noting that  $1-v^2 = (1-v)(1+v)$  is comparable to  $1-v$ )

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \chi\left(\frac{t}{T}\right) \chi\left(\frac{x-vt}{R}\right) |u(t, x)|^{p+1} dt dx \lesssim \frac{R}{1-v} + \frac{T}{R}.$$

Since  $1-v \gtrsim R^{1/2}/T^{1/2}$  by hypothesis, the claim follows.  $\square$

### 3. Proof of Theorem 1.1

We are now ready to prove Theorem 1.1. Suppose that this claim failed for some  $E, p$ . Carefully negating the quantifiers, we may thus find a sequence of times  $T_n \rightarrow \infty$  and  $t_n \in \mathbb{R}$ , a  $\delta > 0$  independent of  $n$ , and a family of solutions  $u_n$  which uniformly obey the energy bound  $E[u_n] \leq E$  such that

$$\frac{1}{2T_n} \int_{t_n-T_n}^{t_n+T_n} \|u_n(t)\|_{L_x^\infty(\mathbb{R})} dt \geq \delta.$$

By translating each  $u_n$  by  $t_n$ , we may normalise  $t_n = 0$ .

Let  $n$  be large. We will now allow our implied constants in the  $\lesssim$  notation to depend on  $\delta$ , thus

$$\int_{-T_n}^{T_n} \|u_n(t)\|_{L^\infty(\mathbb{R})} dt \gtrsim T_n.$$

From this bound and (8), we now conclude that the set

$$\{t \in [-T_n, T_n] : \|u_n(t)\|_{L^\infty(\mathbb{R})} \gtrsim 1\}$$

has Lebesgue measure  $\gtrsim T_n$  (for suitable choices of implied constants). In particular, we can find a finite set  $\Delta_n \subset [-T_n, T_n]$  of times which are 1-separated and of cardinality

$$\#\Delta_n \gtrsim T_n$$

such that

$$\|u_n(t)\|_{L^\infty(\mathbb{R})} \gtrsim 1 \tag{15}$$

for all  $t \in \Delta_n$ .

For each  $t \in \Delta_n$ , let  $x_n(t) \in \mathbb{R}$  be a point such that  $|u_n(t, x_n(t))| \geq \frac{1}{2} \|u_n(t)\|_{L^\infty(\mathbb{R})}$ . From (15), one has

$$|u_n(t, x_n(t))| \gtrsim 1 \tag{16}$$

for all  $t \in \Delta_n$ .

Let us say that two times  $t, t' \in \Delta_n$  are *spacelike* if we have

$$|x_n(t') - x_n(t)| \geq |t - t'| + 1.$$

There is a limit as to how many spacelike pairs of times can exist:

**Lemma 3.1** (finite speed of propagation). *Let  $n$  be sufficiently large, and let  $t_1, \dots, t_m \in \Delta_n$  be times which are pairwise spacelike. Then we have  $m = O(1)$ .*

*Proof.* Without loss of generality we may assume that  $t_1 < \dots < t_m$ . Consider the spacetime region

$$\Omega := \mathbb{R} \times \mathbb{R} \setminus \bigcup_{1 \leq j \leq m} \left\{ (t, x) : t \geq t_j \text{ and } |x - x_n(t_j)| \leq t - t_j + \frac{1}{2} \right\}.$$

Standard energy estimates reveal that

$$\int_{(t_j, x) \in \Omega} \mathbb{T}_{00}(t_j, x) dx + \int_{|x - x_n(t_j)| \leq \frac{1}{2}} \mathbb{T}_{00}(t_j, x) dx \leq \int_{(t_{j-1}, x) \in \Omega} \mathbb{T}_{00}(t_{j-1}, x) dx$$

for all  $1 < j \leq m$ , where  $\mathbb{T}_{00} = \mathbb{T}_{00,n}$  is the energy density of  $u_n$ . Iterating this and then using (7), we conclude that

$$\sum_{1 < j \leq m} \int_{|x - x_n(t_j)| \leq \frac{1}{2}} \mathbb{T}_{00}(t_j, x) dx \lesssim 1$$

and in particular that

$$\sum_{1 < j \leq m} \int_{|x - x_n(t_j)| \leq \frac{1}{2}} |u_n(t_j, x)|^{p+1} dx \lesssim 1.$$

But from (16), (9) we see that

$$\int_{|x-x_n(t_j)| \leq \frac{1}{2}} |u_n(t_j, x)|^{p+1} dx \gtrsim 1.$$

for each  $j$ , and the claim follows.  $\square$

We now use this lemma and some combinatorial arguments to extract a Lipschitz worldline.

**Corollary 3.2** (existence of Lipschitz worldline). *Let  $\varepsilon_0 : (0, 1] \rightarrow (0, 1]$  be an arbitrary function. Then there exists a constant  $0 < c_0 = c_0(\varepsilon_0) \leq 1$  with the following property: for all sufficiently large  $n$ , there exists  $c_0 < c < 1$  (depending on  $n$ ) and a subset  $\Delta'_n$  of  $\Delta_n$  with*

$$\#\Delta'_n \geq cT_n$$

such that we have the Lipschitz property

$$|x_n(t') - x_n(t)| \leq |t - t'| + \varepsilon_0(c)T_n \quad (17)$$

for all  $t, t' \in \Delta'_n$ .

*Proof.* Fix  $\varepsilon$ , and let  $n$  be sufficiently large. Define the *particle number* of a set  $\Delta$  to be the largest integer  $m$  for which one can find pairwise spacelike times  $t_1, \dots, t_m$  in  $\Delta$ . By the previous lemma, we see that  $\Delta_n$  has particle number  $O(1)$ . The key lemma is the following:

**Lemma 3.3** (dichotomy). *Let  $\Delta' \subset \Delta_n$ ,  $m = O(1)$  and  $c > 0$  be such that*

$$\#\Delta' \geq 2cT_n$$

and  $\Delta'$  has particle number at most  $m$ . Suppose  $n$  is sufficiently large depending on  $c$ . Then at least one of the following is true:

- (i) *There exists a subset  $\Delta'' \subset \Delta'$  of cardinality at least  $cT_n$  such that (17) holds for all  $t, t' \in \Delta''$ .*
- (ii) *There exists a subset  $\Delta''' \subset \Delta'$  of cardinality at least  $c\varepsilon_0(c)T_n/16$  with particle number at most  $m - 1$ .*

Iterating this lemma at most  $O(1)$  times we obtain the claim.

It remains to prove the lemma. We subdivide the interval  $[-T_n, T_n]$  into intervals  $I$  of length between  $\varepsilon_0(c)T_n/4$  and  $\varepsilon_0(c)T_n/8$ . Call an interval *sparse* if  $\#(\Delta' \cap I) \leq c\varepsilon_0(c)T_n/8$ , and *dense* otherwise. Observe that at most  $cT_n$  elements of  $\Delta'$  lie in sparse intervals. Thus if we let  $\Delta''$  denote the intersection of  $\Delta'$  with the union of all the dense intervals, then  $\#\Delta'' \geq cT_n$ .

If  $\Delta''$  obeys (17) then we are done. Otherwise, we can find  $t_1, t_2 \in \Delta''$  such that

$$|x_n(t_1) - x_n(t_2)| > |t_1 - t_2| + \varepsilon_0(c)T_n.$$

The time  $t_1$  must lie in some dense interval  $I$ . We split  $\Delta'' \cap I = \Delta'''_1 \cup \Delta'''_2$ , where  $\Delta'''_1$  consists of all  $t \in \Delta'' \cap I$  with  $|x_n(t) - x_n(t_1)| \leq \varepsilon_0(c)T_n/2$ , and  $\Delta'''_2$  consists of the remainder of  $\Delta'' \cap I$ . Observe from the triangle inequality (if  $n$  is sufficiently large depending on  $c$ ) that all times in  $\Delta'''_1$  are spacelike with respect to  $t_2$ , and similarly all times in  $\Delta'''_2$  are spacelike with respect to  $t_1$ . Thus each of  $\Delta'''_1$  and  $\Delta'''_2$  can

have particle number at most  $m - 1$ . On the other hand, by the pigeonhole principle, one of  $\Delta_1'''$  and  $\Delta_2'''$  must have cardinality at least  $\frac{1}{2}\#(\Delta'' \cap I)$ , which is at least  $c\varepsilon_0(c)T_n/16$  since  $I$  is dense. The lemma, and hence the corollary, follows.  $\square$

Let  $\varepsilon_0 : (0, 1] \rightarrow (0, 1]$  to be a function to be chosen later (one should think of  $\varepsilon_0(c)$  as going to zero very rapidly as  $c \rightarrow 0$ ). For any sufficiently large  $n$ , let  $c_0, c$  and  $\Delta'_n$  be as in Corollary 3.2.

Define the function  $x'_n : [-T_n, T_n] \rightarrow \mathbb{R}$  by

$$x'_n(t) := \inf_{t' \in \Delta'_n} (x_n(t') - |t - t'|).$$

One easily verifies that  $x'_n$  is Lipschitz with constant at most 1. From (17) we also see that

$$|x_n(t) - x'_n(t)| \leq \varepsilon_0(c)T_n \quad (18)$$

for all  $t \in \Delta'_n$ .

We now apply a quantitative version of the Rademacher (or Lebesgue) differentiation theorem to ensure that  $x'_n(t)$  is approximately differentiable on a large interval.

**Proposition 3.4** (quantitative Rademacher differentiation theorem). *Let  $\varepsilon_1 : (0, 1] \rightarrow (0, 1]$  be a function, and let  $\delta > 0$ . Then there exists  $r_1 = r_1(\varepsilon_1, \delta) > 0$  with the following property: given any Lipschitz function  $f : [-1, 1] \rightarrow \mathbb{R}$  with Lipschitz constant at most 1, there exists  $r_1 \leq r \leq 1$  such that the set*

$$\left\{ x \in [-1, 1] : \text{there exists } L \in \mathbb{R} \text{ such that } \left| \frac{f(y) - f(x)}{y - x} - L \right| \leq \delta \right. \\ \left. \text{whenever } y \in [-1, 1] \text{ is such that } \varepsilon_1(r) \leq |y - x| \leq r \right\}$$

(which, intuitively, is the set where  $f$  is approximately differentiable) has Lebesgue measure at least  $2 - \delta$ .

*Proof.* We give an indirect ‘‘compactness and contradiction’’ proof. Suppose for contradiction that the claim failed. Negating the quantifiers carefully, this means that there exists a function  $\varepsilon_1 : (0, 1] \rightarrow (0, 1]$ , a  $\delta > 0$ , a sequence  $r_n \rightarrow 0$ , and a sequence  $f_n : [0, 1] \rightarrow \mathbb{R}$  of Lipschitz functions with constant at most 1, such that the sets

$$\left\{ x \in [-1, 1] : \text{there exists } L \in \mathbb{R} \text{ such that } \left| \frac{f_n(y) - f_n(x)}{y - x} - L \right| \leq \delta \right. \\ \left. \text{whenever } y \in [-1, 1] \text{ is such that } \varepsilon_1(r) \leq |y - x| \leq r \right\}$$

have Lebesgue measure at most  $2 - \delta$  for all  $n$  and all  $r_n \leq r \leq 1$ .

By translating each  $f_n$  by a constant if necessary, we may assume that  $f_n(0) = 0$ . The Lipschitz functions then form a bounded equicontinuous family on the compact domain  $[-1, 1]$ , and so by the Arzelà–Ascoli theorem we may (after passing to a subsequence if necessary) assume that the  $f_n$  converge

uniformly to a limit  $f$ . We conclude that the set

$$\left\{ x \in [-1, 1] : \text{there exists } L \in \mathbb{R} \text{ such that } \left| \frac{f(y) - f(x)}{y - x} - L \right| \leq \delta/2 \right. \\ \left. \text{whenever } y \in [-1, 1] \text{ is such that } \varepsilon_1(r) \leq |y - x| \leq r \right\}$$

has Lebesgue measure at most  $2 - \delta$  for all  $0 < r \leq 1$ . On the other hand,  $f$  is clearly Lipschitz with constant at most 1, and so by the Lipschitz differentiation theorem,  $f$  is differentiable almost everywhere. In particular, the set

$$\bigcup_{m=1}^{\infty} \left\{ x \in [-1, 1] : \text{there exists } L \in \mathbb{R} \text{ such that } \left| \frac{f(y) - f(x)}{y - x} - L \right| \leq \delta/2 \right. \\ \left. \text{whenever } y \in [-1, 1] \text{ is such that } 0 \leq |y - x| \leq 2^{-m} \right\}$$

has full measure in  $[-1, 1]$ . By the monotone convergence theorem, this implies that one of the sets in this union has measure greater than  $2 - \delta$ . But this contradicts the previous claim.  $\square$

**Remark 3.5.** It is also possible to give a more direct “martingale”<sup>2</sup> or “multiscale analysis” proof of this proposition, which we sketch as follows. For each  $n \geq 1$ , let  $f_n$  be the piecewise linear continuous function which agrees with  $f$  on multiples of  $2^{-n}$ , and is linear between such intervals. One easily verifies that the functions  $f_{n+1} - f_n$  are pairwise orthogonal in the Hilbert space  $\dot{H}^1([-1, 1])$ , and thus by Bessel’s inequality we have

$$\sum_{n=1}^{\infty} \|f_{n+1} - f_n\|_{\dot{H}^1([-1, 1])}^2 \leq 2.$$

Now let  $F : \mathbb{N} \rightarrow \mathbb{N}$  be a function to be chosen later, and let  $\sigma > 0$  be a small quantity to be chosen later. From the pigeonhole principle, one can find  $1 \leq n_0 \leq C(F, \sigma)$  such that

$$\sum_{n=n_0}^{F(n_0)} \|f_{n+1} - f_n\|_{\dot{H}^1([-1, 1])}^2 \leq \sigma.$$

If one then sets  $r := \sigma 2^{-n_0}$ , one can verify all the required claims if  $\sigma$  is chosen sufficiently small depending on  $\delta$ , and  $F$  is sufficiently rapidly growing depending on  $\delta$ ,  $\sigma$ , and  $\varepsilon_0$ ; the quantity  $L$  can basically be taken to be  $f'_n(x)$ . We omit the details, but see [Tao 2009] for some similar arguments in this spirit.

Let  $\delta > 0$  be a small quantity (depending on  $c$ ) to be chosen later, and let  $\varepsilon_1 : (0, 1] \rightarrow (0, 1]$  be the function  $\varepsilon_1(r) := \delta r$ . We let  $n$  be sufficiently large, and apply the above proposition to the Lipschitz function  $f = f_n : [-1, 1] \rightarrow \mathbb{R}$  defined by  $f(y) := 1/T_n x'_n(T_n y)$ . We conclude that there exists  $r_1 = r_1(\delta)$

<sup>2</sup>Indeed, the arguments here are closely related to some classical martingale inequalities of Doob [1953] and Lépingle [1976].

and  $r_1 < r < 1$  (depending on  $\delta$  and  $n$ ) such that the set

$$\left\{ x \in [-T_n, T_n] : \text{there exists } L \in \mathbb{R} \text{ such that } \left| \frac{x'_n(t') - x'_n(t)}{t' - t} - L \right| \leq \delta \right. \\ \left. \text{whenever } y \in [-T_n, T_n] \text{ is such that } \delta r T_n \leq |t - t'| \leq r T_n \right\}$$

has measure at least  $(2 - \delta)T_n$ .

On the other hand, the set  $\Delta'_n$  has cardinality at least  $cT_n$ . As in the proof of Lemma 3.3, we partition  $[-T_n, T_n]$  into intervals  $I$  of length between  $rT_n/4$  and  $rT_n/8$ , and let  $\Delta''_n$  be the portion of  $\Delta'_n$  which are contained inside those intervals  $I$  which are *dense* in the sense that they contain at least  $crT_n/16$  elements of  $\Delta'_n$ . It is easy to see that  $\Delta''_n$  has cardinality at least  $cT_n/2$ . Also,  $\Delta''_n$  is 1-separated.

Thus, if we let  $\delta = \delta(c)$  be sufficiently small compared to  $c$ , we can find  $t_* \in [-T_n, T_n]$  within a distance 1 of  $\Delta''_n$  and  $v \in \mathbb{R}$  such that

$$\left| \frac{x'_n(t') - x'_n(t_*)}{t' - t_*} - v \right| \leq \delta \quad \text{whenever } t' \in [-T_n, T_n] \text{ is such that } \delta r T_n \leq |t_* - t'| \leq r T_n.$$

Let  $t_0$  be an element of  $\Delta''_n$  within 1 of  $t_*$ . Applying (18), the triangle inequality, and the Lipschitz nature of  $x'_n$ , we conclude that

$$x_n(t_1) = x_n(t_0) + v(t_1 - t_0) + O(\delta|t_1 - t_0|) + O(\varepsilon_0(c)T_n) + O(1)$$

whenever  $t_1 \in \Delta''_n$  is such that  $\delta T_n + 1 \leq |t_1 - t_0| \leq rT_n - 1$ . Applying the Lipschitz property again, we conclude that

$$x_n(t_1) = x_n(t_0) + v(t_1 - t_0) + O(\delta r T_n) + O(\varepsilon_0(c)T_n) + O(1)$$

for all  $t_1 \in \Delta''_n$  with  $|t_1 - t_0| \leq rT_n - 1$ . If we set  $\varepsilon_0(c) := \delta(c)r_1(\delta(c))$ , and assume  $n$  is sufficiently large depending on all other parameters, we thus have

$$x_n(t_1) = x_n(t_0) + v(t_1 - t_0) + O(\delta r T_n)$$

whenever  $t_1 \in \Delta''_n$  and  $|t_1 - t_0| \leq rT_n/4$ . One should view this as an assertion that  $x_n$  is approximately differentiable near  $t_0$ .

By definition of  $\Delta''_n$ , we know that  $t_0$  is contained in an interval  $I$  of length at most  $rT_n/4$  which contains  $\gtrsim crT_n$  elements of  $\Delta_n$ . We thus see that the parallelogram

$$P := \{(t, x) : t \in I, |x - x_n(t_0) - v(t - t_0)| \leq R/2\}$$

contains at least  $\gtrsim crT_n$  points of the form  $(t, x_n(t))$  with  $t \in \Delta_n$ , where  $R$  is a quantity of size  $\sim \delta r T_n$ . On the other hand, by definition of  $\Delta_n$ , we have  $|u_n(t, x(t))| \gtrsim 1$  for all  $t \in \Delta_n$ . Applying (9), we conclude that

$$\int_P |u_n(t, x)|^{p+1} dt dx \gtrsim crT_n.$$



On the other hand, from Proposition 2.2 we have

$$\int_P |u_n(t, x)|^{p+1} dt dx \lesssim R^{1/2} (rT_n)^{1/2} + \frac{rT_n}{R} \lesssim \delta^{1/2} rT_n + \delta^{-1}.$$

If we set  $\delta$  to be sufficiently small depending on  $c$ , and let  $n$  be sufficiently large depending on all other parameters, we obtain a contradiction as desired. This completes the proof of Theorem 1.1.

### Acknowledgement

We thank Jason Murphy, Qing Tian Zhang and the anonymous referee for corrections.

### References

- [Doob 1953] J. L. Doob, *Stochastic processes*, Wiley, New York, 1953. MR 15,445b Zbl 0053.26802
- [Lépingle 1976] D. Lépingle, “La variation d’ordre  $p$  des semi-martingales”, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **36**:4 (1976), 295–316. MR 54 #8849 Zbl 0325.60047
- [Lindblad and Soffer 2005] H. Lindblad and A. Soffer, “A remark on asymptotic completeness for the critical nonlinear Klein–Gordon equation”, *Lett. Math. Phys.* **73**:3 (2005), 249–258. MR 2006i:35249 Zbl 1106.35072
- [Reed 1978] M. C. Reed, “Propagation of singularities for non-linear wave equations in one dimension”, *Comm. Partial Differential Equations* **3**:2 (1978), 153–199. MR 80d:35092 Zbl 0377.35047
- [Tao 2008] T. Tao, *Structure and randomness: pages from year one of a mathematical blog*, American Mathematical Society, Providence, RI, 2008. MR 2010h:00002 Zbl 05380664
- [Tao 2009] T. Tao, “A quantitative version of the Besicovitch projection theorem via multiscale analysis”, *Proc. Lond. Math. Soc.* (3) **98**:3 (2009), 559–584. MR 2010f:28009 Zbl 1173.28001

Received 3 Nov 2010. Revised 12 Jan 2011. Accepted 7 Feb 2011.

HANS LINDBLAD: lindblad@math.ucsd.edu

Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, United States

TERENCE TAO: tao@math.ucla.edu

Department of Mathematics, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095-1555, United States

## Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at [msp.berkeley.edu/apde](http://msp.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  $\text{\LaTeX}$  but submissions in other varieties of  $\text{\TeX}$ , and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

**References.** Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of  $\text{\BibTeX}$  is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures.** Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to [graphics@msp.org](mailto:graphics@msp.org) with details about how your graphics were generated.

**White space.** Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

**Proofs.** Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

# ANALYSIS & PDE

Volume 5 No. 2 2012

---

The geodesic X-ray transform with fold caustics PLAMEN STEFANOV and GUNTHER UHLMANN	219
Existence of extremals for a Fourier restriction inequality MICHAEL CHRIST and SHUANGLIN SHAO	261
Dispersion and controllability for the Schrödinger equation on negatively curved manifolds NALINI ANANTHARAMAN and GABRIEL RIVIÈRE	313
A bilinear oscillatory integral estimate and bilinear refinements to Strichartz estimates on closed manifolds ZAHER HANI	339
The Cauchy problem for the Benjamin–Ono equation in $L^2$ revisited LUC MOLINET and DIDIER PILOD	365
On triangles determined by subsets of the Euclidean plane, the associated bilinear operators and applications to discrete geometry ALLAN GREENLEAF and ALEX IOSEVICH	397
Asymptotic decay for a one-dimensional nonlinear wave equation HANS LINDBLAD and TERENCE TAO	411



2157-5045(2012)5:2;1-I