DECOUPLING ESTIMATES IN FOURIER ANALYSIS

LARRY GUTH

ABSTRACT

Decoupling is a recent development in Fourier analysis, which has applications in harmonic analysis, PDEs, and number theory. We survey some applications of decoupling and some of the ideas in the proof.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 42B15; Secondary 42B20, 11L15

KEYWORDS

Restriction theory, circle method

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 2, pp. 1054–1089
DOI 10.4171/ICM2022/172

1. INTRODUCTION

Decoupling is a recent development in Fourier analysis, which has applications in harmonic analysis, PDEs, and number theory. To put it in context, let us start by recalling some basic ideas of Fourier analysis. In Fourier analysis, we represent a function as a Fourier series or Fourier integral. For instance, if $f : \mathbb{R}^n \to \mathbb{C}$ is a reasonably nice function, then we can write it as a Fourier integral

$$f(x) = \int_{\mathbb{R}}^{n} \hat{f}(\omega) e^{2\pi i \omega \cdot x} dx. \tag{1.1}$$

Here $\omega \cdot x$ is the dot product of ω and x, which we will also abbreviate as just ωx.

Here are a couple reasons that it is useful to represent a function f using a Fourier series or integral. First, the functions $e^{2\pi i \omega x}$ are eigenfunctions for the partial derivative operators ∂x_j. This makes the Fourier representation interact well with partial derivatives, and it helps to study PDEs. Second, the functions $e^{2\pi i \omega \cdot x}$ are eigenfunctions of the translation operator T_v defined by $T_v f(x) = f(x + v)$. This makes the Fourier representation useful in problems involving the translation structure of $\mathbb{R}^n$, including problems in additive number theory.

But there is also a serious downside to representing a function f as a Fourier series/integral. To evaluate $f(x)$, we have to compute an integral or a sum with many terms. It often happens that the terms have various phases in the complex plane, and it is difficult to tell what happens when we add them all up. In general, given some information about $\hat{f}$, it can be difficult to determine what that information it has to say about f. We will see some longstanding open questions of this flavor below.

Decoupling is helpful for estimating $\| f \|_{L^p}$ in terms of information about $\hat{f}$. Now $\| f \|_{L^2}$ is directly related to $\hat{f}$ because of orthogonality: Plancherel's theorem states that

$$\| f \|_{L^2} = \| \hat{f} \|_{L^2}. \tag{1.2}$$

But for other values of p, it is much harder to connect $\| f \|_{L^p}$ with information about $\hat{f}$.

Estimates for $\| f \|_{L^p}$ for $p \neq 2$ occur often in harmonic analysis, PDEs, and analytic number theory. You may wonder, if we have a good understanding of $\| f \|_{L^2}$, what more do we learn by understanding $\| f \|_{L^p}$ for other values of p. I like to think of this question in terms of superlevel sets. Define the superlevel set $U_\lambda(f)$ by

$$U_\lambda(f) := \{ x : |f(x)| > \lambda \}. \tag{1.3}$$

We denote the volume of a set U by $|U|$. If we know $\| f \|_{L^p}$ for every p, we typically get accurate estimates for $|U_\lambda(f)|$ for every λ, which gives us basically all the possible information about how "big" the function f is. But if we only know $\| f \|_{L^2}$, we get only limited information about $|U_\lambda(f)|$.

Other motivations for studying $\| f \|_{L^p}$ come from applications in PDE and analytic number theory. In nonlinear PDEs, bounds involving $\| f \|_{L^p}$ are important for understanding how close a solution to a nonlinear PDE is to a solution of a corresponding linear PDE. In analytic number theory, the number of solutions to certain diophantine systems is equal to

$\int |f|^p$ for a well-chosen function f and exponent p. These are just a couple samples among many applications for estimating $\|f\|_{L^p}$.

Decoupling is a new tool for estimating $\|f\|_{L^p}$ in terms of Fourier-analytic information about f. It was first formulated by Wolff in [56], where he was able to prove sharp estimates for large values of p. In [14], Bourgain and Demeter proved sharp decoupling estimates for all p. This breakthrough has led to solutions to problems in harmonic analysis that had seemed far out of reach a decade ago.

In the next two subsections, we will introduce two main areas where decoupling has had an impact. We will give examples of hard open problems and also examples of problems that were solved using decoupling.

1.1. Restriction theory

The Fourier representation of a function $f : \mathbb{R}^n \to \mathbb{C}$ is

$$f(x) = \int_{\mathbb{R}^n} \hat{f}(\omega) e^{2\pi i \omega \cdot x} d\omega.$$

There are two basic estimates connecting the L^p-norms of f and the L^p-norms of $\hat{f}$:

- (Orthogonality) $\|f\|_{L^2} = \|\hat{f}\|_{L^2}$;
- (Triangle inequality) $\|f\|_{L^\infty} \leq \|\hat{f}\|_{L^1}$.

Interpolating between these gives the Hausdorff–Young inequality

$$\|f\|_{L^p} \leq \|\hat{f}\|_{L^q} \quad \text{if } 1 \leq p \leq 2 \text{ and } \frac{1}{q} = 1 - \frac{1}{p}. \tag{1.4}$$

These are all of the L^p-type estimates for the Fourier transform operator.

If $\hat{f}$ is supported in a subset $\Omega \subset \mathbb{R}^n$, we can write

$$f(x) = \int_{\Omega} \hat{f}(\omega) e^{2\pi i \omega \cdot x} d\omega.$$

Restriction theory studies how the geometry of Ω relates to properties of f such as $\|f\|_{L^p}$. One of the most interesting cases is when $\hat{f}$ is supported in a compact submanifold $S \subset \mathbb{R}^n$. In this case, the Fourier representation of f has the form

$$f(x) = \int_S a(\omega) e^{2\pi i \omega x} d\mu_S(\omega), \tag{1.5}$$

where $d\mu_S$ is the surface area measure of S.

Stein proposed studying L^p-estimates of the form

$$\|f\|_{L^p(\mathbb{R}^n)} \leq C \|a\|_{L^q(S)} \tag{1.6}$$

and made the remarkable discovery that the estimates for the operator E_S depend on the geometry of S. If S is a flat disk, then the only estimate of the form (1.6) is the triangle inequality $\|f\|_{L^\infty} \leq \|a\|_{L^1(S)}$. But if S is a curved surface, then there are more inequalities.

One central problem in the field is to understand all the L^p-inequalities of form (1.6) when S is a curved hypersurface, like a paraboloid. Let us write P for the truncated paraboloid

$$P := \left\{ \omega \in \mathbb{R}^n \mid \omega_n = \sum_{j=1}^{n-1} \omega_j^2 \text{ and } \sum_{j=1}^{n-1} \omega_j^2 \leq 1 \right\}. \tag{1.7}$$

In this case, the Fourier representation of f is

$$f(x) = \int_P a(\omega) e^{2\pi i \omega x} d\mu_P(\omega). \tag{1.8}$$

Example 1.1. Suppose $a(\omega) = 1$ on P, and f is given by (1.8).

First note that $f(0) = \int_P d\mu_P$ is equal to the area of P, which is ~ 1. When x is large, there is a lot of cancellation in the integral (1.8) coming from rapid oscillation of the function $e^{2\pi i \omega x}$ as ω varies over P. This effect can be estimated accurately using stationary phase, and one finds that

$$\left| f(x) \right| \lesssim |x|^{-\frac{n-1}{2}}.$$

This bound is sharp for most x. Therefore $\| f \|_{L^p(\mathbb{R}^n)} < \infty$ if and only if $p > \frac{2n}{n-1}$.

Stein conjectured that the same L^p-bounds hold whenever $|a(\omega)| \leq 1$ for all ω.

Conjecture 1.2 (Restriction conjecture [49]). *Suppose that f has the form (1.8) and that $|a(\omega)| \leq 1$ for all $\omega \in P$. If $p > \frac{2n}{n-1}$, then*

$$\| f \|_{L^p(\mathbb{R}^n)} \leq C(p, n).$$

Notice that the hypothesis that f has the form (1.8) with $|a(\omega)| \leq 1$ for all ω is a hypothesis about $\hat{f}$. The restriction conjecture asks what this information about $\hat{f}$ tells us about $\| f \|_{L^p}$. The 2-dimensional case of Conjecture 1.2 was proven by Fefferman in [26]. But for dimension $n \geq 3$, the conjecture remains open after intensive work by many people. In Section 5, we will some reasons the problem is so difficult.

In Conjecture 1.2, we considered the bound $\|a(\omega)\|_{L^\infty} \leq 1$. Bounds of the form $\|a\|_{L^q(P)}$ are also interesting for other q. The case $q = 2$ is the most important, and it was completely worked out by Strichartz [51] following work by Tomas and Stein. It has turned out to be important in PDEs. It reads as follows.

Theorem 1.3 (Strichartz inequality [51]). *Suppose that f has the form (1.8). If $p \geq \frac{2(n+1)}{n-1}$, then*

$$\| f \|_{L^p(\mathbb{R}^n)} \leq C(n) \left\| a(\omega) \right\|_{L^2(P)}.$$

This theorem plays an important role in the study of the Schrödinger equation. Recall that the linear Schödinger equation for a function $u(x, t)$ with $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$ is

$$\partial_t u = i \sum_{j=1}^{d} \partial_{x_j}^2 u. \tag{1.9}$$

If u obeys the linear Schrödinger equation, then $\hat{u}$ is a distribution supported on the paraboloid, and so the Strichartz estimate can be used to understand $\|u\|_{L^p}$. Theorem 1.3 tells us that for any solution of the linear Schrödinger equation (1.9) with initial data $u(x,0) = u_0(x)$,

$$\|u\|_{L^{\frac{2(d+2)}{d}}(\mathbb{R}^d \times \mathbb{R})} \leq C \|u_0\|_{L^2(\mathbb{R}^d)}. \tag{1.10}$$

This theorem has played a central role in PDEs, especially in nonlinear PDEs. The L^2-norm on the right-hand side is important in PDEs because $\|u_0\|_{L^2} = \|\hat{u}_0\|_{L^2}$ and also $\|u_0\|_{L^2} = \|u(y,t)\|_{L^2_y}$ for every t. In nonlinear PDEs, it leads to sharp estimates about when the solution to a nonlinear PDE is close to the solution of the corresponding linear PDE.

The Strichartz estimate describes a spreading-out effect. To get a sense of it, first suppose that u_0 is a smooth bump concentrated on a ball in spacetime. As t increases, the function $u(x,t)$ spreads out and gets smaller. As it does so, $\int_{\mathbb{R}^d} |u(x,t)|^2 dx$ remains constant, and $\int_{\mathbb{R}^d} |u(x,t)|^p dx$ gets smaller for any $p > 2$. Because of this spreading out effect, $\int_{\mathbb{R}^d \times \mathbb{R}} |u(x,t)|^{\frac{2(d+2)}{d}} dx dt$ is finite.

The exponent $\frac{2(d+2)}{d}$ is the only exponent for which (1.10) holds. To see what is special about this exponent, it helps me to translate the Strichartz estimate into an estimate for superlevel sets. Let $U_\lambda(u) := \{(x,t) \in \mathbb{R}^d \times \mathbb{R} : |u(x,t)| > \lambda\}$. The Strichartz inequality implies that if $\|u_0\|_{L^2(\mathbb{R}^d)} = 1$, then

$$\big|U_\lambda(u)\big| \leq C\lambda^{-\frac{2(d+2)}{d}}.$$

This estimate is sharp: for any choice of λ, we can find initial data u_0 with $\|u_0\|_{L^2(\mathbb{R}^d)} = 1$ so that the solution of the Schrödinger equation has $|U_\lambda(u)| \geq c\lambda^{-\frac{2(d+2)}{d}}$.

It is also worth mentioning that the choice of the paraboloid in this discussion is just one interesting example. There are similar theorems and conjectures for other surfaces, such as the sphere and the cone, and these help to study other PDEs, such as the Laplace eigenfunction equation $\triangle u = \lambda u$ and the wave equation.

One striking application of decoupling involves Strichartz estimates on flat tori. The Schrödinger equation makes sense on any Riemannian manifold, and for each manifold we can ask for the best inequality in the spirit of (1.10). Understanding the Strichartz estimates on closed manifolds is extremely difficult. It is known that different closed manifolds behave quite differently from each other—for example, round spheres behave differently from flat tori. But very few examples are understood. Before decoupling, sharp Strichartz estimates were only known for S^1 and $S^1 \times S^1$ (by Bourgain in the 1990s [8]) and S^3 (by Burq–Gerard–Tzvetkov [19]). In all these examples, the value of the exponent p is an even integer, and we will discuss in Section 1.2 why this is important.

The simplest flat torus in the unit cube torus $\mathbb{R}^d/\mathbb{Z}^d$. A solution to the Schrödinger equation on the unit cube torus is just a solution $u(x,t)$ on $\mathbb{R}^d \times \mathbb{R}$ which is $\mathbb{Z}^d$-periodic in the x variable. Any such solution can be written in the form

$$u(x,t) = \sum_{n \in \mathbb{Z}^d} a_n e^{2\pi i (n \cdot x + |n|^2 t)}. \tag{1.11}$$

Notice that this Fourier representation is analogous to (1.8), except that the integral in (1.8) is replaced by a sum. We say that u has "frequency at most N" if the coefficients a_n are supported in the cube $Q_N := \{(n_1, \dots, n_d) \in \mathbb{Z}^d : |n_j| \le N \text{ for all } j\}$.

Example 1.4. Suppose that u is given by (1.11) where $a_n = 1$ if $n \in Q_N$ and $a_n = 0$ otherwise. In other words,

$$u(x,t) = \sum_{n \in Q_N} e^{2\pi i (n\cdot x + |n|^2 t)}.$$

First note that $u(0,0) = |Q_N| \sim N^d$. We have $|u(x,t)| \sim N^d$ when $|x| \le \frac{1}{10dN}$ and $|t| \le \frac{1}{10dN^2}$, because then each term in the sum is almost 1. As x and t increase, we get cancellation in the sum coming from oscillations in $e^{2\pi i(n\cdot x + |n|^2 t)}$. So far this behavior is similar to Example 1.1.

However, in the torus case, $|u(x,t)|$ is also large when (x,t) lies near to a rational point of the form $(\frac{p_1}{q}, \dots, \frac{p_d}{q}, \frac{p_t}{q})$. Taking account of all these peaks near rational points, it turns out that $U_\lambda(u) \cap [0,1]^{d+1}$ has volume $\sim N^{d+2}\lambda^{-\frac{2(d+2)}{d}}$ for all λ in the range $N^{d/2} \le \lambda \le N^d$. (This range includes all interesting values of λ.)

A natural analogue of the restriction conjecture in the periodic setting would say

Conjecture 1.5. *Suppose that u is given by (1.11) and that $|a_n| \le 1$ for all $n \in Q_N$ and $a_n = 0$ for $n \notin Q_N$. Then $|U_\lambda(u) \cap [0,1]^{d+1}| \le C(d,\varepsilon) N^{d+2+\varepsilon} \lambda^{-\frac{2(d+2)}{d}}$ for all λ in the range $N^{d/2} \le \lambda \le N^d$.*

This conjecture used to sound to me just as hard as the restriction conjecture or maybe harder. The setup is similar. And Example 1.4 in this periodic setting is more intricate and complex than Example 1.1 in the setting of the original restriction conjecture. However, Bourgain and Demeter proved this conjecture as a corollary of their sharp Strichartz estimate on tori. This theorem is one of the first applications of decoupling.

Theorem 1.6 (Bourgain and Demeter [14]). *Suppose that u is given by (1.11) and that a_n is supported in Q_N. Then*

$$\|u\|_{L^{\frac{2(d+2)}{d}}([0,1]^{d+1})} \le C(d,\varepsilon) N^\varepsilon \|a_n\|_{\ell^2}. \tag{1.12}$$

Notice that if $u_0(x) = u(x,0)$, then $\|a_n\|_{\ell^2} = \|u_0\|_{L^2([0,1]^d)}$, so this inequality is very similar to the Strichartz inequality for the Schrödinger equation on $\mathbb{R}^d$ recorded in (1.10).

To finish this section, let us try to roughly indicate why the Strichartz inequality on the torus is much harder than the Strichartz inequality on $\mathbb{R}^d$. Recall that the Strichartz inequality encodes a spreading-out effect. First, imagine a solution $u(x,t)$ on a Euclidean space, and suppose that the initial data u_0 is concentrated in a very small ball. As time increases, the solution $u(x,t)$ spreads out. At a small time t_0, the solution is spread over a unit ball. In the Euclidean space, it can continue to spread out in all directions indefinitely. The proof of Strichartz estimates this effect in a quantitative way.

Now let u_P be the solution on the torus with the same initial data u_0. The function u_P is given by periodizing u:

$$u_P(x,t) = \sum_{z \in \mathbb{Z}^d} u(x+z,t). \tag{1.13}$$

For times up to t_0, $u(x,t)$ is supported on a unit ball in the x variable, and so $u_P(x,t) = u(x,t)$. But beyond this time, $u(x,t)$ is spread over a much bigger ball, and there are many nonzero terms in the sum (1.13). If we visualize $u_P(x,t)$, the solution starts to wrap around the torus. Different pieces of the solution, which have traveled around the torus in different ways, get added up, and we have to prove that there is a lot of cancellation in that sum.

Before decoupling, Theorem 1.6 was known for $d = 1, 2$ only because of a connection with number theory. In the next section, we describe some connections between Fourier analysis and number theory, and we will flesh this out.

1.2. Analytic number theory

When p is an even integer, L^p-estimates have a special interpretation which connects them with problems in additive number theory.

Suppose that $A \subset \mathbb{Z}^d$ is a finite set. We define $E_s(A)$ (the additive s-energy of A) by

$$E_s(A) := \#\big\{(a_1,\dots,a_s,b_1,\dots,b_s) \in A^{2s} : a_1 + \cdots + a_s = b_1 + \cdots + b_s\big\}. \tag{1.14}$$

For each A, we can also define a function $f_A(x)$ with Fourier series

$$f_A(x) = \sum_{a \in A} e^{2\pi i a \cdot x}. \tag{1.15}$$

The function $f_A : \mathbb{R}^d \to \mathbb{C}$ is $\mathbb{Z}^d$-periodic because $A \subset \mathbb{Z}^d$, and so each function $e^{2\pi i a \cdot x}$ is $\mathbb{Z}^d$-periodic.

Lemma 1.7. *For any finite set $A \subset \mathbb{Z}^d$,*

$$\int_{[0,1]^d} \big|f_A(x)\big|^{2s} dx = E_s(A).$$

Proof sketch. We expand the integral on the left-hand side:

$$\begin{aligned}\int_{[0,1]^d} \big|f_A(x)\big|^{2s} dx &= \int_{[0,1]^d} f_A^s \bar{f}_A^s dx \\ &= \int_{[0,1]^d} \sum_{a_1,\dots,a_s,b_1,\dots,b_s \in A} e^{2\pi i (a_1+\cdots+a_s-b_1-\cdots-b_s)x} dx.\end{aligned}$$

Now if $m \in \mathbb{Z}^d$, then $\int_{[0,1]^d} e^{2\pi i m \cdot x} dx$ is 1 if $m = 0$ and 0 otherwise. And so the only terms that contribute to the integral above are terms where $a_1 + \cdots + a_s - b_1 - \cdots - b_s = 0$. So the last integral is $E_s(A)$. ∎

For instance, if $A_{k,N} := \{1^k, 2^k, \dots, N^k\} \subset \mathbb{Z}$ then

$$E_s(A_{k,N}) = \#\text{ of solutions to } a_1^k + \cdots + a_s^k = b_1^k + \cdots + b_s^k, \\ \text{with } a_j, b_j \in \mathbb{Z},\ 1 \le a_j, b_j \le N. \tag{1.16}$$

In this case, the relevant function f is

$$f_{k,N}(x) = \sum_{a=1}^{N} e^{2\pi i a^k x}, \tag{1.17}$$

and Lemma 1.7 tells us that

$$\int_0^1 \left| f_{k,N}(x) \right|^{2s} dx = E_s(A_{k,N}). \tag{1.18}$$

Lemma 1.7 tells us that a certain L^p-norm is equal to the number of solutions to a certain diophantine equation. The lemma is useful in both directions. If we know something about the number of solutions to the diophantine equation, then we can get information about the L^p-norm. If we know something about the L^p-norm, then we can get information about the number of solutions to the diophantine equation.

For instance, consider the diophantine equation $a_1^2 + a_2^2 = b_1^2 + b_2^2$, with a_i, b_i between 1 and N. First let us estimate the number of solution directly. Rearranging we get $a_1^2 - b_1^2 = b_2^2 - a_2^2$, and factoring one side we see that

$$(a_1 + b_1)(a_1 - b_1) = b_2^2 - a_2^2.$$

If we fix a_2, b_2, then the number of (a_1, b_1) solving this equation depends on the number of factors of $b_2^2 - a_2^2$. Because of unique factorization, the number of different factors of an integer M is fairly small, at most $C_\varepsilon M^\varepsilon$ for any $\varepsilon > 0$. Using this, we see that the number of integer solutions to $a_1^2 + a_2^2 = b_1^2 + b_2^2$ with $1 \le a_j, b_j \le N$ is at most $C_\varepsilon N^{2+\varepsilon}$. Lemma 1.7 tells us that the number of solutions is equal to $\int_0^1 |f_{2,N}(x)|^4 dx$, and so we conclude that this integral is bounded by $C_\varepsilon N^{2+\varepsilon}$.

On the other hand, Weyl used the differencing method to give pointwise estimates for the function $f_{2,N}$. These estimates imply that $\int_0^1 |f_{2,N}(x)|^4 dx \le C_\varepsilon N^{2+\varepsilon}$ which then gives an analytic proof that the number of integer solutions to $a_1^2 + a_2^2 = b_1^2 + b_2^2$ with $1 \le a_j, b_j \le N$ is at most $C_\varepsilon N^{2+\varepsilon}$.

Hardy and Littlewood made a conjecture that generalizes these estimates from squares to higher powers.

Conjecture 1.8 (Hardy and Littlewood). *For any $k \ge 2$, $E_k(A_{k,N}) \le C_\varepsilon N^{k+\varepsilon}$. Equivalently,*

$$\int_0^1 \left| f_{k,N}(x) \right|^{2k} dx \le C_\varepsilon N^{k+\varepsilon}.$$

This conjecture is open for all $k \ge 3$. The Fourier series of $f_{3,N}$ is fairly simple to write down. But it is very difficult to determine good bounds for the L^p-norms of $f_{3,N}$, or for the size of superlevel sets $U_\lambda(f_{3,N})$. This is a classical and striking example of how difficult it is to read off information about $f(x)$ from information about its Fourier series.

On the other hand, there are cases when we can use Fourier analysis to estimate an L^p-norm and then use Lemma 1.7 to get a new estimate for the number of solutions to a diophantine equation. One of the most interesting examples of this kind concerns Vinogradov's mean value theorem, which is a multivariable generalization of the functions we just considered.

Define

$$F_{k,N}(x_1,\dots,x_k) = \sum_{a=1}^{N} e^{2\pi i(ax_1+a^2x_2+\cdots+a^kx_k)}.$$

By Lemma 1.7, $\int_{[0,1]^k} |F_{k,N}(x)|^{2s}dx$ is equal to the number of solutions to the following diophantine system of equations:

$$a_1^j+\cdots+a_s^j = b_1^j+\cdots+b_s^j \quad \text{for all } 1\le j\le k, \text{ with } a_i,b_i\in\mathbb{Z}, 1\le a_i,b_i\le N.$$

Vinogradov [52] studied the L^p-norms of $F_{k,N}$ in the 1930s. He was able to prove sharp estimates for $\|F_{k,N}\|_{L^p}$ for sufficiently large p. He used these bounds to greatly improve the estimates for Weyl sums and Waring's problem in large degree, and also to improve the bounds on the zero-free region of the Riemann zeta function. Vinogradov's argument cleverly exploited both sides of equation (1.7): some parts of the argument directly count the number of solutions to some diophantine systems in the variables a_i, b_i, and other parts of the argument estimate integrals in the x variable. Some important ideas in the proof of decoupling are related to Vinogradov's argument, and we will discuss this more in Section 4.5.

In the last decade, mathematicians have proven estimates for $\|F_{k,N}\|_{L^p}$ that are sharp up to factors of $C(k,\varepsilon)N^\varepsilon$ for every k and p. As a corollary, we get estimates for the number of solutions to the Vinogradov system that are sharp up to a factor $C(k,\varepsilon)N^\varepsilon$.

Theorem 1.9 ([16, 58, 59]).

$$\|F_{k,N}\|_{L^p([0,1]^k)} \le C(k,\varepsilon)N^\varepsilon\big(N^{1/2}+N^{1-\frac{k(k+1)}{2p}}\big).$$

The proof in [16] uses decoupling and the proof in [59] uses the method of efficient congruencing. (Historically, Wooley developed efficient congruencing starting in the 1990s, cf. [57]. He improved Vinogradov's estimates and gave sharp estimates for $k=3$ in [58]. Then [16] used decoupling to prove Theorem 1.9 and immediately afterwards, [59] used efficient congruencing to give a different proof of Theorem 1.9.)

Both [16] and [59] are quite technical. Recently, Guo–Li–Yung–Zorin-Kranich [28] gave a dramatically simpler proof of Theorem 1.9, combining some of the features of [16] and [59] with some new clarifying ideas. Their paper is ten pages long and essentially self-contained.

Lemma 1.7 is a special trick for understanding L^p-norms when p is an even integer. This even integer trick also plays an important role in the problems we discussed in Section 1.1. In [26], Fefferman used a version of the even integer trick to prove Conjecture 1.2 in dimension $n=2$. The L^p-exponent in Conjecture 1.2 is $p=\frac{2n}{n-1}$, which is an even integer when $n=2$ but not for any $n\ge 3$. In the early 1990s, in [8], Bourgain used

the even integer trick to prove sharp periodic Strichartz estimates when $d = 1, 2$ (the cases $d = 1, 2$ in Theorem 1.6). The exponent in the Strichartz estimate is $\frac{2(d+2)}{d}$, which is an even integer when $d = 1, 2$, but not for any $d > 2$. Another important problem in this circle is Montgomery's conjecture about the L^p-norms of Dirichlet polynomials. When p is an even integer, Montgomery gave sharp estimates for the relevant L^p-norms in just a page (cf. [42]). But giving a sharp estimate for any other value of p is a major open problem. This might help explain why, even though Theorem 1.6 was already known in dimensions $d = 1, 2$, it still seemed far out of reach to prove it for any other d.

Before decoupling, the situation concerning periodic Strichartz estimates in dimensions $1, 2$ was rather curious. The periodic Strichartz estimate can be considered as a result in PDEs, resolving a problem of mathematical physics. But the proof depended on number theory facts, such as unique factorization. The decoupling proof of Theorem 1.6 is purely analytic—with no input from number theory. The argument can then recover some of the number theory that went into the original proof. The relevant number theory estimates are not that difficult, but proving them by analysis is still interesting. Building on this, Bourgain and Demeter began to work on Vinogradov's mean value theorem in [15], eventually leading to Theorem 1.9 and new results in number theory.

Theorem 1.9 leads to improved bounds for Waring's problem on the number of ways to write an integer as a sum of kth powers and the related problem of Weyl sums. Other applications of decoupling have led to incremental improvements in very classical problems of analytic number theory such as the Lindelof hypothesis [13] and the Gauss circle problem and [18]. Guo–Zhang [29] and Guo–Zorin-Kranich [30] have extended Theorem 1.9 to more complex systems of diophantine equations, introduced in number theory by Arkhipov–Chubarikov–Karatsuba [1].

1.3. Influence of the proof

Besides the new results, the method of proof of decoupling has had a big influence on the field. There is a classical toolbox in harmonic analysis with tools like orthogonality, integration by parts, and Hölder's inequality. For hard problems in this area, such as the restriction conjecture, people who have worked a lot on them generally feel that this set of classical tools is not sufficient to understand the problem. Over the last 25 years, mathematicians have brought into play ideas from other areas in order to attack some of these hard problems. For instance, Wolff ([54] and [56]) brought in ideas from combinatorial geometry and topology, Bourgain [9] brought in ideas from combinatorial number theory, and Dvir [24] brought in ideas from error-correcting codes and algebraic geometry. In contrast to these developments, the proof of decoupling is based on the classical toolbox. The most important idea in the proof is to take advantage of estimates at many different scales. Using many different scales is also a classical idea in harmonic analysis. But it is really striking how powerful it turns out to be in the context of decoupling. I personally was shocked that it is possible to prove Theorem 1.6 using only these tools. The main goal of the article is to explore how combining information at many scales helps to prove theorems like Theorems 1.6 and 1.9.

1.4. Outline of the rest of the article

In Section 2, we will introduce the statement of decoupling. In Section 3, we will begin to discuss multiscale arguments, and we will see how the statement of decoupling was carefully crafted to work well in such arguments. In Section 4, we will discuss some ideas of the proof of decoupling.

In Section 5, we will discuss the connection between the restriction problem and the Kakeya problem, and try to explain why the restriction problem seems to be so difficult. Then we will discuss why decoupling turns out to be easier than restriction.

In Section 6, we will survey some other applications of decoupling in harmonic analysis.

In Section 7, we will discuss some limitations of the method, some frustrating aspects of the proof, and some open problems.

2. THE STATEMENT OF DECOUPLING

Now that we have seen some applications of decoupling, we turn to the actual statement of decoupling. The statement of decoupling was crafted carefully, and after we state it we will spend two sections digesting it and discussing some of the choices involved in the statement.

Suppose that $\Omega \subset \mathbb{R}^n$ and that Ω is a disjoint union of subsets θ, $\Omega = \sqcup \theta$. If $f : \mathbb{R}^n \to \mathbb{C}$ is a function, and $\hat{f}$ is supported in Ω, then we can decompose $f = \sum_\theta f_\theta$ where f_θ is defined by

$$f_\theta = \int_\theta \hat{f}(\omega) e^{2\pi i \omega x} d\omega.$$

Decoupling has to do with the relationship between L^p-norm of f and the L^p-norms of f_θ for the different θ in the decomposition $\Omega = \sqcup \theta$.

Definition 2.1. Suppose that $\Omega \subset \mathbb{R}^n$ and Ω is a disjoint union of subsets θ, $\Omega = \sqcup \theta$. For each exponent p, we define the decoupling constant $D_p(\Omega = \sqcup \theta)$ to be the smallest constant so that for every function f with $\hat{f}$ supported in Ω,

$$\| f \|^2_{L^p(\mathbb{R}^n)} \leq D_p(\Omega = \sqcup \theta)^2 \sum_\theta \| f_\theta \|^2_{L^p(\mathbb{R}^n)}. \tag{2.1}$$

If $p = 2$, then orthogonality gives $\| f \|^2_{L^2} = \sum_\theta \| f_\theta \|^2_{L^2}$, and so $D_2(\Omega = \sqcup \theta) = 1$ for any decomposition $\Omega = \sqcup \theta$. Decoupling theorems for higher p are a kind of strengthening of orthogonality. For $p > 2$, the value of $D_p(\Omega = \sqcup \theta)$ depends on the geometry of the decomposition.

As an example of a decomposition, first let P denote the truncated parabola:

$$P = \{(\omega_1, \omega_2) \in \mathbb{R}^2 : \omega_2 = \omega_1^2, -1 \leq \omega_1 \leq 1\}.$$

Definition 2.2. For a large parameter N, we let Ω be the N^{-2}-neighborhood of P. For $j = -N, \ldots, N$, we define

$$\theta_j := \Omega \cap \left\{ \frac{j}{N} - \frac{1}{2N} \leq \omega_1 \leq \frac{j}{N} + \frac{1}{2N} \right\}.$$

Each θ_j is approximately a rectangular box of dimensions $N^{-2} \times N^{-1}$.

We have $\Omega = \bigsqcup_{j=1}^{N} \theta_j$, and we abbreviate this whole decomposition as P_N.

We can now state our first decoupling theorem.

Theorem 2.3 ([14]). *For each $\varepsilon > 0$, for each $2 \le p \le 6$, $D_p(P_N) \le C_\varepsilon N^\varepsilon$.*

In other words, if $2 \le p \le 6$, and if $\hat{f}$ is supported in the N^{-2}-neighborhood of P, then

$$\| f \|^2_{L^p(\mathbb{R}^2)} \le C_\varepsilon N^\varepsilon \sum_{j=1}^{N} \| f_{\theta_j} \|^2_{L^p(\mathbb{R}^2)}. \tag{2.2}$$

This decoupling theorem can be applied to exponential sums, and it implies Theorem 1.6 in the case $d = 1$ and Theorem 1.9 in the case $k = 2$. Theorem 1.6 for a d-dimensional torus follows from a decoupling theorem for the paraboloid in $\mathbb{R}^{d+1}$, and Theorem 1.9 for higher k follows from a decoupling theorem for the moment curve in $\mathbb{R}^k$.

Let us see how this decoupling theorem leads to L^p-estimates for exponential sums. This will help a little to digest the definition of D_p. Suppose we start with an exponential sum using frequencies on the truncated parabola. For $j = -N, \dots, N$, we define the frequency $\omega_j = (\frac{j}{N}, \frac{j^2}{N^2}) \in P$, and we let f be the exponential sum

$$f(x) = \sum_{j=-N}^{N} a_j e^{2\pi i \omega_j x}.$$

If the a_j were chosen randomly, then with high probability, we would have $|f(x)| \sim (\sum_j |a_j|^2)^{1/2}$ for most x. In this random case, we would have $\| f \|_{L^p(B_R)} \sim (\sum_j |a_j|^2)^{1/2} |B_R|^{1/p}$. So the best possible bound we could hope for has the form

$$\| f \|_{L^p(B_R)} \sim \left(\sum_j |a_j|^2 \right)^{1/2} |B_R|^{1/p}.$$

Decoupling achieves such a bound up to a factor of N^ε when $2 \le p \le 6$ and R is large enough. This bound in turn implies Theorem 1.6 for $d = 1$ and Theorem 1.9 for $k = 2$.

Here is how to apply decoupling. Note that the frequency ω_j lies in θ_j. In fact, if we write $f = \sum_j f_{\theta_j}$, then $f_{\theta_j} = a_j e^{2\pi i \omega_j x}$. Directly applying Theorem 2.3 does not tell us anything because $\| f_{\theta_j} \|_{L^p(\mathbb{R}^2)}$ is infinite. But with a little technical work, one can prove that a similar estimate holds with L^p-norms on large balls instead of L^p-norms on the whole plane. In particular, if $R \ge N^2$, then

$$\| f \|^2_{L^p(B_R)} \le 100 D_p(P_N)^2 \sum_{j=-N}^{N} \left\| a_j e^{2\pi i \omega_j x} \right\|^2_{L^p(B_R)}.$$

(The extra factor 100 comes from the technical work of passing from $\mathbb{R}^2$ to B_R.) If $p = 6$, then we can plug in $D_6(P_N) \le C_\varepsilon N^\varepsilon$ and simplify everything to get

$$\| f \|_{L^6(B_R)} \le C_\varepsilon N^\varepsilon \left(\sum_{j=1}^{N} |a_j|^2 \right)^{1/2} |B_R|^{1/6}.$$

This bound matches the random example above up to the factor $C_\varepsilon N^\varepsilon$, and so in particular it is tight up to this factor. This estimate is the periodic Strichartz estimate for $d = 1$ and the Vinogradov mean value theorem for $k = 2$.

The definition of the decoupling constant D_p was crafted partly to make this computation work. This explains the squares in Definition 2.1.

3. INDUCTION ON SCALES

The definition of decoupling was crafted by Thomas Wolff in his work on local smoothing [56]. He noticed that this definition is well suited for combining information from many scales. The whole field of decoupling leans on this observation. The first example of combining scales is the following lemma, which essentially appears in [56].

Lemma 3.1. $D_p(P_{N_1 N_2}) \leq D_p(P_{N_1}) D_p(P_{N_2})$.

Let us first discuss why this is significant, and then we will sketch the proof. If we iterate this lemma k times, we get

$$D_p(P_{N_1^k}) \leq D_p(P_{N_1})^k. \tag{3.1}$$

Suppose that we are able to find a single number N_1 for which we can prove $D_p(P_{N_1}) \leq N_1^{\frac{1}{1000}}$. Then equation (3.1) implies that $D_p(P_N) \leq N^{\frac{1}{1000}}$ when N is any power of N_1. This implies the decoupling theorem, Theorem 2.3, with $\varepsilon = \frac{1}{1000}$. For any particular N_1, the decoupling constant $D_p(N_1)$ can be approximated to a given accuracy by a finite computation. This is not immediately obvious from the definition, but it is not that difficult to show. So, in principle, there exists a brute force proof of Theorem 2.3 with $p = 6$ (the most interesting p) and $\varepsilon = \frac{1}{1000}$, where the proof is a giant finite computation to check that $D_6(P_{N_1}) \leq N_1^{\frac{1}{1000}}$ for a particular N_1 together with Lemma 3.1.

This situation is very different from the periodic Strichartz estimate, Theorem 1.6, or Vinogradov's mean value theorem, Theorem 1.9. For instance, suppose we somehow knew that Theorem 1.6 holds when $d = 3$ and $N = 10^{10}$. Recall that Theorem 1.6 is an L^p-estimate for periodic solutions to the Schrödinger equation with frequencies at most N. If we somehow knew optimal bounds for periodic solutions with frequency at most 10^{10}, I do not see how we could use that information to say anything about solutions with much larger frequencies, like 10^{1000}.

By switching our point of view from the original problem of periodic Strichartz estimates to the decoupling problem, we make it easier to combine information from different scales. The real proof of the decoupling theorem does not involve a giant brute force computation like we described above. It combines the multiscale idea from Lemma 3.1 with other ideas from the field, and we will discuss it more in the next section.

Next let us talk about the proof of Lemma 3.1. The proof is very short, and it illustrates how the statement of decoupling was crafted to combine information from different scales.

The first observation is that decoupling behaves in a nice way under translations and under linear changes of variable. Suppose that $L : \mathbb{R}^n \to \mathbb{R}^n$ is a linear change of variables, or a translation, or a composition of those. If we start with a decomposition $\Omega = \sqcup \theta$, then we get a new decomposition $L\Omega = \sqcup L\theta$. The first observation is that the new decomposition has the same decoupling constant as the original one:

$$D_p(L\Omega = \sqcup L\theta) = D_p(\Omega = \sqcup \theta). \tag{3.2}$$

If g has Fourier support in Ω and $g = \sum g_\theta$, then we can perform a change of a variables to get a new function $\tilde{g}$ with Fourier support in $L\theta$. Since the Fourier transform behaves in a nice way with respect to linear changes of variables and to translations, it is easy to track how the decoupling constant behaves and check (3.2).

Now we start the proof sketch of Lemma 3.1. Suppose that $\hat{f}$ is supported in Ω, the $(N_1 N_2)^{-2}$-neighborhood of P. This neighborhood is divided into blocks θ of length $(N_1 N_2)^{-1}$, and we need to prove that

$$\|f\|_{L^p}^2 \leq D_p(P_{N_1})^2 D_p(P_{N_2})^2 \sum_\theta \|f_\theta\|_{L^p}^2.$$

We prove this bound in two steps. Note that Ω is contained in the N_1^{-2}-neighborhood of P, which we can divide into blocks τ of length N_1^{-1}. By definition of $D_p(P_{N_1})$, we have

$$\|f\|_{L^p}^2 \leq D_p(P_{N_1})^2 \sum_\tau \|f_\tau\|_{L^p}^2. \tag{Step 1}$$

The support of $\hat{f_\tau}$ is contained in $\Omega \cap \tau$, which we can decompose as $\Omega \cap \tau = \bigsqcup_{\theta \subset \tau} \theta$.

By the definition of D_p,

$$\|f_\tau\|_{L^p}^2 \leq D_p\left(\Omega \cap \tau = \bigsqcup_{\theta \subset \tau} \theta\right)^2 \sum_{\theta \subset \tau} \|f_\theta\|_{L^p}^2.$$

Notice that there are N_2 different θ in each τ. In fact, there is a linear change of variables that takes $\Omega \cap \tau$ to the N_2^{-1}-neighborhood of P and takes each θ to a block of length N_2^{-1}. Therefore, $D_p(\Omega \cap \tau = \bigsqcup_{\theta \subset \tau} \theta) = D_p(P_{N_2})$. Plugging in to the last indented equation, we get

$$\|f_\tau\|_{L^p}^2 \leq D_p(N_2)^2 \sum_{\theta \subset \tau} \|f_\theta\|_{L^p}^2. \tag{Step 2}$$

Now if we combine Steps 1 and 2, we get the desired inequality:

$$\|f\|_{L^p}^2 \leq D_p(P_{N_1})^2 \sum_\tau \|f_\tau\|_{L^p}^2 \leq D_p(N_1)^2 D_p(N_2)^2 \sum_\theta \|f_\theta\|_{L^p}^2.$$

4. IDEAS OF THE PROOF

In this section, we discuss some of the ideas in the proof of the decoupling theorem for the parabola, Theorem 2.3. By now there are actually several proofs of Theorem 2.3 (cf.

[14, 32, 38]). Each proof has some advantages. We will focus on the original proof in [14], but as we go we will try to highlight certain ideas that appear in all of the proofs.

Recall that P is the truncated parabola in $\mathbb{R}^2$. We let Ω be the N^{-2}-neighborhood of P, and we decompose Ω into N pieces θ, which are each approximately rectangles of dimensions $N^{-2} \times N^{-1}$. Suppose $\hat{f}$ is supported on Ω and decompose $f = \sum_\theta f_\theta$. To help illustrate the ideas, we focus on the following corollary of Theorem 2.3.

Corollary 4.1. *If $f = \sum_\theta f_\theta$ as in the last paragraph, and $\|f_\theta\|_{L^\infty(\mathbb{R}^2)} \leq 1$ for every θ, then*

$$\left|U_{N/10}(f) \cap B_{N^2}\right| \leq C_\varepsilon N^{1+\varepsilon}.$$

First, let us give a little context for the numbers that appear in this bound. By the triangle inequality, $|f(x)| \leq \sum_\theta |f_\theta(x)| \leq N$. So $U_{N/10}(f)$ is the region where $|f(x)|$ is biggest. The bound in Corollary 4.1 is sharp, as we can see from the following example.

Example 4.2. Let $f(x)$ be the exponential sum

$$f(x) = \sum_{n=1}^{N} e^{2\pi i(\frac{n}{N}x_1 + \frac{n^2}{N^2}x_2)}.$$

Each f_θ is a single term in the sum, and so $\|f_\theta\|_{L^\infty} = 1$.

We can check directly that $f(mN, 0) = N$ for any integer m because each term in the sum is 1. Also if x lies in a ball of radius $1/100$ around $(mN, 0)$, then each term in the sum has real part more than $1/2$, and so $|f(x)| \geq N/2$. Therefore, $U_{N/10}(f) \cap B_{N^2}$ contains $\sim N$ balls of measure ~ 1 and so itself has measure $\gtrsim N$.

We will give a rough sketch of the proof of Corollary 4.1. The proof of Corollary 4.1 is simpler than the whole proof of Theorem 2.3, but it shows most of the main ideas.

4.1. Orthogonality

Under the hypotheses of Corollary 4.1, it may well happen that $|f_\theta(x)| \sim 1$ for every x and θ. To prove Corollary 4.1, we need to show that for most points $x \in B_{N^2}$, there is a lot of cancellation in the sum $f(x) = \sum_\theta f_\theta(x)$. The most fundamental tool for proving cancellation in Fourier analysis is orthogonality. Since the sets θ are disjoint, the functions f_θ are orthogonal, and so

$$\int_{\mathbb{R}^2} |f|^2 = \sum_\theta \int_{\mathbb{R}^2} |f_\theta|^2.$$

The functions f_θ are exactly orthogonal on $\mathbb{R}^2$. They are also approximately orthogonal over any sufficiently large set. Since the distance between any two (nonadjacent) θ's is at least $1/N$, the functions f_θ are morally orthogonal on any ball of radius N. The rough reason for this approximate orthogonality is the following. Suppose $\omega_1 \in \theta_1$ and $\omega_2 \in \theta_2$. We have to check that the functions $e^{2\pi i\omega_1 x}$ and $e^{2\pi i\omega_2 x}$ are approximately orthogonal on a ball $B_N(x_0)$. The inner product of $e^{2\pi i\omega_1 x}$ and $e^{2\pi i\omega_2 x}$ on $B_N(x_0)$ is

$$\int_{B_N(x_0)} e^{2\pi i\omega_1 x}\overline{e^{2\pi i\omega_2 x}}dx = \int_{B_N(x_0)} e^{2\pi i(\omega_1-\omega_2)x}dx.$$

Since $|\omega_1 - \omega_2| \geq 1/N$, the function $e^{2\pi i(\omega_1-\omega_2)x}$ oscillates significantly on $B_N(x_0)$, which causes some cancellation in that integral. This approximate argument suggests the following heuristic.

Heuristic 4.3 (Approximate orthogonality). If B is a square box of side length at least N, then

$$\int_B |f|^2 dx \approx \sum_\theta \int_B |f_\theta|^2 dx.$$

As written, this heuristic is not quite true, but there are more technical substitutes for it. It is morally true, and it helps to imagine it in our proof sketch.

By approximate orthogonality,

$$\int_{B_{N^2}} |f|^2 dx \approx \sum_\theta \int_{B_{N^2}} |f_\theta|^2 dx \leq CN|B_{N^2}| = CN^5.$$

This gives an upper bound

$$\left|U_{N/10}(f) \cap B_{N^2}\right| \leq CN^3. \tag{4.1}$$

To prove Corollary 4.1, we will have to improve the bound N^3 to $N^{1+\varepsilon}$.

So far, we have only used that the rectangles θ are disjoint (and separated by at least $1/N$). We will have to use more information about the θ in order to do better. In fact, if the rectangles θ were laid out along a straight line, then bound (4.1) would be best possible. (We can see that by considering the exponential sum $f(x) = \sum_{n=1}^N e^{2\pi i \frac{n}{N} x_1}$.) To do better, we will have to take advantage of the way the rectangles θ follow the curve of the parabola. In the next two subsections we set up some basic tools that will allow us to take advantage of the curvature of the parabola.

4.2. Multiple scales

We want to study $f = \sum_\theta f_\theta$. We can divide this sum into pieces in various ways. If $M < N$, then we can cover Ω by M rectangles τ of dimensions $M^{-1} \times M^{-2}$. Imagine that M divides N so that each θ is contained in exactly one τ. Then we can write

$$f_\tau = \sum_{\theta \subset \tau} f_\theta \quad \text{and} \quad f = \sum_\tau f_\tau.$$

In order to get better bounds for f, we will consider the functions f_τ at many different intermediate scales (many different choices of M).

The number of $\theta \subset \tau$ is $\frac{N}{M}$, and so $|f_\tau(x)| \leq \frac{N}{M}$. We define $N_\tau = \frac{N}{M}$, which is the number of θ in τ.

Since $|f(x)| \leq \sum_\tau |f_\tau(x)|$, we see that if $|f(x)| \geq N/10$, then $|f_\tau(x)| \geq \frac{N_\tau}{20}$ for at least $M/20$ different τ. This suggests studying $U_{N_\tau/20}(f_\tau)$ for each τ.

We can use orthogonality (Lemma 4.3) to bound $|U_\lambda(f_\tau)|$. By itself, this will not lead to any new bounds. In addition to that, we will study the shape of $U_\lambda(f_\tau)$. Because of their shapes, the sets $U_{N_\tau/20}(f_\tau)$ cannot overlap too much. This geometric input will lead to improvement on the bound for $U_{N/10}(f)$.

4.3. Wave packets

Suppose that $\tau \subset \mathbb{R}^2$ is a rectangle. Suppose that $\hat{f_\tau}$ is supported in τ. Then f_τ itself has a special geometric structure, which is called a wave packet decomposition.

This wave packet decomposition is based on a tiling of $\mathbb{R}^2$ which is in some sense dual to τ. First, let τ^* be the dual rectangle. If τ has dimensions $M^{-1} \times M^{-2}$, then τ^* would have dimensions $M \times M^2$. The axis of τ^* with length M^2 corresponds to the axis of τ with length M^{-2}. Next, let $\mathbb{T}_\tau$ be a tiling of $\mathbb{R}^2$ by rectangles congruent to τ^*.

Heuristic 4.4 (Locally constant heuristic). If $\hat{f_\tau}$ is supported on a rectangle τ with center ω_τ, then for each rectangle $T \in \mathbb{T}_\tau$,

$$f_\tau(x) \approx a_T e^{2\pi i \omega_\tau x},$$

where $a_T \in \mathbb{C}$ is a constant. In particular,

$$\big|f_\tau(x)\big| \quad \text{is approximately constant on each rectangle } T \in \mathbb{T}_\tau.$$

According to this heuristic, we can describe f_τ on all of $\mathbb{R}^2$ in the form

$$f_\tau(x) \approx \sum_{T \in \mathbb{T}_\tau} a_T e^{2\pi i \omega_\tau x} \chi_T. \tag{4.2}$$

Here χ_T is the characteristic function of T (or a smoothed out version of it). Each term on the right-hand side is called a wave packet, and equation (4.2) is called the wave packet decomposition of f_τ.

This heuristic is again not quite literally true, but it can be replaced by more technical statements that are true. It is morally true.

One origin of wave packet decompositions is particle–wave duality in quantum mechanics. If $\hat{f}$ is supported in the parabola P, then f satisfies the Schrödinger equation, which describes a quantum mechanical particle moving in a vacuum. (Here we have $f(x_1, x_2)$, and we think of x_2 as the time variable t.) Quantum mechanical particles can behave almost like classical particles for significant time periods. A classical particle in a vacuum moves with constant velocity, tracing out a straight line in space time. A single wave packet describes a quantum mechanical particle behaving almost classically.

Let us try to give some idea why the locally constant heuristic makes sense. The Fourier transformation behaves in a nice way with respect to linear changes of variables and translations. Because of this, it actually suffices to understand the wave packet decomposition when τ is the square $[-1, 1]^2$. Also the wave packet decomposition makes sense in any dimension, and the proofs are basically the same. For simplicity, let us consider dimension 1. Now we have a function $f : \mathbb{R} \to \mathbb{C}$ with $\hat{f}$ supported in $[-1, 1]$. The wave packet decomposition, equation (4.2), says that $f(x)$ is roughly constant on each unit interval. This vague statement is closely related to the the Whittaker–Shannon–Nyquist interpolation theorem, which says that if $\hat{f}$ is supported in $[-1, 1]$, then the whole function $f(x)$ can be recovered from the values $f(n/2)$, with $n \in \mathbb{Z}$. Informally, this suggests that "nothing significant is happening on length scales smaller than 1/2." Here is another way to think about it. Since $\hat{f}$

is supported in $[-1, 1]$,

$$f(x) = \int_{-1}^{1} \hat{f}(\omega) e^{2\pi i \omega x} dx.$$

For $|\omega| \leq 1$, each function $e^{2\pi i \omega x}$ varies slowly, and looks roughly constant on any scale significantly smaller than 1. The function f itself is a linear combination of these slowly varying functions, and so we may hope that f also looks roughly constant at scales smaller than 1.

4.4. Transversality

We are now ready to return to the proof sketch of Corollary 4.1. By bringing into play the wave packet structure of f_τ, we will see how to improve on the bound from Section 4.1, which only used orthogonality. At this point, the curvature of the parabola will come into play.

Recall that each θ is an $N^{-2} \times N^{-1}$ rectangle in the N^{-2}-neighborhood of the truncated parabola P. By the hypotheses of Corollary 4.1, we know that $\|f_\theta\|_{L^\infty(\mathbb{R}^2)} \leq 1$. We want to bound $U_{N/10}(f) \cap B_{N^2}$.

As in Section 4.2, set $M = N^{1/2}$, and cover the parabola with M rectangles τ of dimensions $M^{-1} \times M^{-2}$. Set $N_\tau = N/M = N^{1/2}$, and let us try to understand $U_{N_\tau/20}(f_\tau)$ for each τ. We know that $|f_\tau|$ is locally constant on translates of τ^*, which have dimensions $M \times M^2 = N^{1/2} \times N$. We can also use orthogonality to estimate $\int_{B_N} |f_\tau|^2 dx$ for each ball B_N of radius N. Putting together this information, we conclude that for each B_N, $U_{N_\tau/10}(f_\tau) \cap B_N$ is contained in $\lesssim 1$ translates of τ^*. In other words, on each B_N, each f_τ has only around 1 wave packet of amplitude $\sim N_\tau$.

Now we are ready to take advantage of the curvature of the parabola. Because of the curvature of P, the rectangles τ are oriented in different directions, and so the dual rectangles τ^* point in different directions. On each B_N, $U_{N_\tau/10}(f_\tau)$ is essentially one translate of τ^*. Because all these rectangles point in different directions, they do not overlap very much. The set $U_{N/10}(f)$ should lie in $U_{N_\tau/20}(f_\tau)$ for most τ, and so $U_{N/10}(f) \cap B_N$ has to lie in a constant number of balls of radius $N^{1/2}$. This geometric observation allows us to improve the bound for $U_{N/10}(f)$ beyond what we got from orthogonality alone.

The most effective way to study $U_{N/10}(f)$ on each of these balls of radius $N^{1/2}$ is to repeat the same method, using larger τ's with $M = N^{1/4}$. Continuing in this way through many scales, we eventually see that $U_{N/10}(f) \cap B_N$ has to lie in at most N^ε balls of radius 1. This gives an upper bound

$$\left|U_{N/10}(f) \cap B_{N^2}\right| \leq C_\varepsilon N^{2+\varepsilon}. \tag{4.3}$$

We will call the argument in this section the orthogonality/transversality method, because those are the two main tools that go into it. This argument is essentially due to Bennett–Carbery–Tao [3]. We will discuss their work more in Section 5.1 below. The orthogonality/transversality method improves on just orthogonality, but to prove Corollary 4.1, we will have to improve the bound $N^{2+\varepsilon}$ to $N^{1+\varepsilon}$.

4.5. Induction on scales and transversality together

To get the sharp bound in Corollary 4.1, Bourgain and Demeter combined the ideas from the last subsection with induction on scales (as in Section 3). As in the last subsection, we set $M = N^{1/2}$ and cover the parabola with M rectangles τ of dimensions $M^{-1} \times M^{-2}$. In the orthogonality/transversality argument, we had to understand $U_{N_\tau/20}(f_\tau)$, and we controlled it with the following observation:

(1) Local orthogonality gives an upper bound on $|U_{N_\tau/20}(f_\tau) \cap B_N|$ for each box B_N of side length N.

We can also bring into play induction on scales. After a change of variables, estimating $|U_{N_\tau}(f_\tau)|$ is equivalent to our original problem, Corollary 4.1, but with $N^{1/2}$ rectangular tiles instead of N tiles. So we can also use induction on scales to bound $|U_{N_\tau}(f_\tau)|$.

(2) Induction on scales gives an upper bound on $|U_{N_\tau}(f_\tau) \cap B_{N^2}|$.

The proof of decoupling in [14] uses (1) and (2) together. Combining them leads to the sharp bound in Corollary 4.1

When I was first reading the proof of decoupling, I was surprised and even troubled that combining induction on scales with orthogonality/transversality is so powerful. The orthogonality/transversality method gives an interesting but suboptimal bound. Induction on scales by itself does not give any bound. Why do these ingredients become so much stronger when we mix them together?

Initially, the argument even felt fishy to me. Let us look back at points (1) and (2) above. Why should we combine them? If (2) is stronger than (1), then why not just use (2)? If (1) is stronger than (2), then why not just use (1)? I gradually realized that (1) and (2) give different types of information about $U_{N_\tau/20}(f_\tau)$. Neither one is stronger than the other. They are different and give complementary information.

Induction on scales gives information about the total measure of $U_{N_\tau/20}(f_\tau)$ in B_{N^2}. Local orthogonality also implies a bound on the total measure of $U_{N_\tau/20}(f_\tau)$ in B_{N^2}. The bound on the total measure coming from induction is stronger than the bound coming from orthogonality. But (1) is a local bound: it bounds $|U_{N_\tau/20}(f_\tau) \cap B_N|$ for each box of side N. For a small box B_N of side length N, the bound on $|U_{N_\tau/20}(f_\tau) \cap B_N|$ coming from (1) is stronger than the bound coming from (2). Induction on scales controls the total measure of $U_{N_\tau/20}(f_\tau)$, and local orthogonality forces $U_{N_\tau/20}(f_\tau)$ to be rather spread out.

To summarize, the bound (2) from induction gives the best information about the measure of $U_{N_\tau/20}(f_\tau)$. But the bound (1) from local orthogonality gives us additional information about the shape of $U_{N_\tau/20}(f_\tau)$: in particular, for any box B_N of side length N, $U_{N_\tau/20}(f_\tau) \cap B_N$ consists of at most a constant number of $N^{1/2} \times N$ rectangles.

Now we have digested the information that (1) and (2) give us about f_τ, for each τ. The reader may wonder why information about the shape of $U_{N_\tau/20}(f_\tau)$ helps bound the measure of $U_{N/10}(f)$. The point is that it is difficult for different functions f_{τ_1} and f_{τ_2} to be large in the same place. Notice that if $|f(x)| = |\sum_\tau f_\tau(x)|$ is large, then we must have

$|f_\tau(x)|$ large for many different τ at the same point x. If we knew (2) but not (1), it would be possible for $U_{N_\tau/20}(f_{\tau_1})$ and $U_{N_\tau/20}(f_{\tau_2})$ to be equal to each other. But if we use (1) and (2) together, then we get a much stronger estimate for the measure of the intersection $U_{N_\tau/20}(f_{\tau_1}) \cap U_{N_\tau/20}(f_{\tau_2})$.

Here is another way to think about the leverage we get by adding induction on scales to the transversality/orthogonality argument from Section 4.4. Recall that we covered our original tiles θ with M rectangles τ with dimensions $M^{-1} \times M^{-2}$, and we considered f_τ. In the argument from Section 4.4, we started by picking $M = N^{1/2}$. Continuing through the argument, we then used $M = N^{1/4}$, then $M = N^{1/8}$, and so on. At each of these scales, we used the wave packet structure of the f_τ and we took advantage of transversality between the wave packets of the different f_τ's.

When we add in induction on scales, we are implicitly considering many different scales. We started as before by using the scale $M = N^{1/2}$. When we apply induction to a given f_τ, and we unwind the induction, then we are really applying the same argument to f_τ. When we apply the argument to f_τ, it gets decomposed as $f_\tau = \sum_\gamma f_\gamma$, where each γ contains $N^{1/4}$ of the θ in τ. The total number of γ covering all the different τ's is $M = N^{3/4}$, a scale that we never used in Section 4.4. If we fully unwind the inductive argument, it brings into play wave packets at every scale. And it takes advantage of transversality between wave packets at every scale. In some sense, the extra power comes from using transversality at every scale instead of just the special scales $M = N^{1/2}, N^{1/4}, N^{1/8}, \ldots$, which were used in Section 4.4.

4.6. Final comments

As we mentioned earlier, there are a number of different proofs of decoupling. In [38], Zane Li gave a new proof of Theorem 2.3 based on Wooley's method of efficient congruencing (cf. [57–59]). In [32], Maldague, Wang, and I gave a new proof of Theorem 2.3 based on ideas from projection theory in geometric measure theory such as Orponen's work [43]. One common feature of all these proofs is to bring into play f_τ with τ at every scale, and to take advantage of some type of transversality at every scale.

Vinogradov's work on the mean value conjecture [52] already has this key feature: it uses f_τ for τ of every scale (after unwinding the induction) and takes advantage of some type of transversality at every scale. Vinogradov's work [52] is the first work I am aware of to take advantage of many scales of τ in estimating an exponential sum. Within harmonic analysis, Wolff's work on local smoothing [56] used this key feature. Bourgain's work on the restriction problem [6] took advantage of the transversality of wave packets of f_τ for a single scale of τ. Wolff's work [56] introduced a version of induction on scales which allowed him to take advantage of transversality of wave packets at every scale. Using this method, he proved a decoupling theorem (for the cone) for large exponents p.

The papers [52] and [56] prove estimates for $|U_\lambda(f)|$, which are sharp when λ takes the largest possible value, but not sharp for smaller λ. For instance, the methods of [52] or [56] could prove Corollary 4.1. The advantage of Theorem 2.3 is to give sharp estimates for $|U_\lambda(f)|$ for every λ. For simplicity, we illustrated the method with $\lambda = N/10$, the largest

possible value. The same general method works for every value of λ, although there are some extra wrinkles in the argument.

5. THE KAKEYA CONJECTURE

In this section, we discuss why the restriction conjecture, Conjecture 1.2, remains out of reach in dimension $n \geq 3$. As we saw in the last section, Fourier-analytic estimates in restriction theory are related to understanding how much rectangles pointing in different directions can overlap each other. The Kakeya conjecture is a precise question about how much rectangles pointing in different directions can overlap each other. (Actually, there are several related conjectures.)

Let us formulate the Kakeya conjecture in a way that connects with our discussion of wave packets. Recall that $P \subset \mathbb{R}^n$ denotes the truncated paraboloid:

$$P = \left\{ \omega \in \mathbb{R}^n : \omega_n = \sum_{j=1}^{n-1} \omega_j^2 \text{ and } 0 \leq \omega_n \leq 1 \right\}.$$

Cover P with N^{n-1} rectangular boxes θ of dimensions $\frac{1}{N} \times \cdots \times \frac{1}{N} \times \frac{1}{N^2}$. For each θ, let θ^* denote the dual box with dimensions $N \times \cdots \times N \times N^2$. The long direction of θ^* is equal to the short direction of θ. For each θ, let T_θ denote a translate of θ^*.

The tubes T_θ are related to wave packets that occur in the restriction problem. In the restriction problem, we consider a function f of the form

$$f(x) = \int_P a(\omega) e^{2\pi i \omega x} d\mu_P(\omega). \tag{5.1}$$

The restriction problem asks to estimate $\|f\|_{L^p(\mathbb{R}^n)}$ assuming that $|a(\omega)| \leq 1$ for every ω. We can decompose f as $f = \sum_\theta f_\theta$ where

$$f_\theta(x) = \int_{P \cap \theta} a(\omega) e^{2\pi i \omega x} d\mu_P(\omega). \tag{5.2}$$

Heuristically, each function f_θ is organized into wave packets, and in particular $|f_\theta|$ is locally constant on translates of θ^*. So the tubes T_θ correspond to wave packets of f. Understanding how much the wave packets overlap helps estimate $\|f\|_{L^p}$.

Now we are ready to formulate one version of the Kakeya conjecture.

Conjecture 5.1 (Kakeya conjecture for volume). *Suppose $n \geq 2$. For each θ in the covering of $P \subset \mathbb{R}^n$, let T_θ be a translate of θ^*. Then for each $\varepsilon > 0$,*

$$\left| \bigcup_\theta T_\theta \right| \geq C(n, \varepsilon) N^{-\varepsilon} \sum_\theta |T_\theta|.$$

An argument of Fefferman [25] shows that the restriction conjecture implies the Kakeya conjecture. If a set of tubes $\{T_\theta\}$ is a counterexample to the Kakeya conjecture, we could build a counterexample to the restriction conjecture by choosing f_θ to concentrate on a single wave packet supported on T_θ.

Around 1920, Besicovitch constructed a remarkable example in 2 dimensions where $|\bigcup_\theta T_\theta| \sim \frac{1}{\log N} \sum_\theta |T_\theta|$. Fefferman used this construction in [25] to give a counterexample to a cousin of the restriction conjecture called the ball multiplier problem.

When $n = 2$, Besicovitch's construction turns out to be tight: Davies proved that $|\bigcup_\theta T_\theta| \geq \frac{c}{\log N} \sum_\theta |T_\theta|$. If $n \geq 3$, Besicovitch's construction still works, but we do not know good bounds in the other direction. For example, if $n = 3$, then Davies's method gives only

$$\left|\bigcup_\theta T_\theta\right| \geq \frac{c}{N} \sum_\theta |T_\theta|.$$

Bourgain [6] improved the $\frac{c}{N}$ to $\frac{c}{N^{2/3}}$ and Wolff [53] improved it further to $\frac{c}{N^{1/2}}$. At this point, it becomes very difficult to go further. The best current bound is

$$\left|\bigcup_\theta T_\theta\right| \geq \frac{c}{N^{1/2-\varepsilon_0}} \sum_\theta |T_\theta|,$$

where ε_0 is a small positive constant. The proofs do not make ε_0 explicit, but the best value given by current techniques is probably around $1/1000$. This estimate was proven under an extra assumption by Katz–Laba–Tao [35] and then proven in full generality by Katz–Zahl [36]. The arguments of [6] and [53] are fairly short, about five pages each, but the arguments of [35] and [36] are much more complex, about 50 pages each.

The reason that it is very difficult to improve on $\frac{c}{N^{1/2}}$ has to do with an "almost counterexample" which takes place in $\mathbb{C}^3$. This almost counterexample was first described in [35]. Consider the set

$$H = \{(z_1, z_2, z_3) \in \mathbb{C}^3 : |z_1|^2 + |z_2|^2 - |z_3|^2 = 1\}.$$

This set is a 5-dimensional real manifold in $\mathbb{C}^3$. Its key feature is that it contains many complex lines. Each point of H lies in infinitely many complex lines contained in H. Using this set H as a guide, [35] constructed a set of "complex tubes" T_j with "dimensions" $N \times N \times N^2$, where $|\bigcup_j T_j| = \frac{c}{N^{1/2}} \sum_j |T_j|$. These tubes overlap each other in a very intricate way. They are complex tubes instead of real tubes, and they do not actually all point in different directions, but Wolff's argument from [53] does apply to them. To beat the Kakeya estimate from [53], one has to introduce into the argument some tool that rules out this "almost counterexample." The papers [35] and [36] succeed in doing this, but the tools are much more complex and the quantitative bounds are rather weak. It would be major progress in the field to give a good quantitative improvement to the Kakeya bound in [53], let alone proving the Kakeya conjecture in full.

There is also a stronger version of the Kakeya conjecture which involves L^p-norms. This version is important for the coming subsection.

Conjecture 5.2 (Kakeya conjecture for L^p-norms). *Suppose $n \geq 2$. For each θ in the covering of $P \subset \mathbb{R}^n$, let T_θ be the characteristic function of translate of θ^*, and let $T_{\theta,0}$ be the characteristic function of θ^* itself. The difference is that θ^* is centered at 0, but T_θ could*

have any center. Then for any $\varepsilon > 0$ and any p,

$$\left\| \sum_{\theta} T_\theta \right\|_{L^p(\mathbb{R}^n)} \leq C(n,\varepsilon) N^{\varepsilon} \left\| \sum_{\theta} T_{\theta,0} \right\|_{L^p(\mathbb{R}^n)}.$$

To digest this formula, notice that $\sum_\theta T_\theta(x)$ is the number of tubes through x. The pth power of the left-hand side is $\int_{\mathbb{R}}^n |\sum_\theta T_\theta(x)|^p dx$. This is large if many points x lie in many tubes from our set of tubes. So the L^p Kakeya conjecture says that not too many points x can lie in many different tubes.

The restriction conjecture implies this stronger version of the Kakeya conjecture, which in turn implies the Kakeya conjecture for volumes, Conjecture 5.1.

Bourgain and Demeter proved a sharp decoupling theorem for the paraboloid $P \subset \mathbb{R}^n$ for all n, which they used to give a sharp Strichartz estimate for tori in all dimensions, Theorem 1.6. One reason this result came as a big surprise has to do with the Kakeya conjecture. The proof of decoupling for the paraboloid involves estimating how much tubes pointing in different directions overlap. When $n = 2$, we know a great deal about how rectangles in different directions overlap, including the Kakeya conjecture for $n = 2$. But when $n \geq 3$, we do not know the Kakeya conjecture. Although there was no formal connection between Kakeya and decoupling for the paraboloid, the Kakeya conjecture still made a sharp decoupling theorem in high dimensions seem out of reach, especially for an approach which is heavily based on estimating the overlaps of tubes pointing in different directions.

5.1. Multilinear Kakeya

The Kakeya-type input into the proof of decoupling is called multilinear Kakeya. It was formulated and proven by Bennett–Carbery–Tao [3]. Multilinear Kakeya is a cousin of Kakeya. The setup is a little different, and we will explain it below, but it still gets at the idea that tubes pointing in different directions cannot overlap too much. Remarkably, Bennett–Carbery–Tao proved sharp multilinear Kakeya estimates in all dimensions. Their proof was simplified in [31] down to a few pages.

The multilinear Kakeya estimate in $\mathbb{R}^n$ is an L^p-type estimate. Suppose that $\ell_{j,a} \subset \mathbb{R}^n$ is a line that makes a small angle with the x_j axis (an angle at most $\frac{1}{100n}$ will do). Let $T_{j,a}$ be the characteristic function of the unit neighborhood of $\ell_{j,a}$—the characteristic function of a tube. Let $B_R \subset \mathbb{R}^n$ denote a cube of side length R.

Theorem 5.3 (Multilinear Kakeya [3]).

$$\int_{B_R} \prod_{j=1}^{n} \left(\sum_{a=1}^{A_j} T_{j,a}(x) \right)^{\frac{1}{n-1}} dx \leq C(n,\varepsilon) R^{\varepsilon} \prod_{j=1}^{n} A_j^{\frac{1}{n-1}}.$$

Let us take a moment to digest this estimate. For a fixed j, think of the tubes $\{T_{j,a}\}_{a=1}^{A_j}$ as tubes "in direction j." Now $\sum_{a=1}^{A_j} T_{j,a}(x)$ is the number of tubes in direction j going through x. The integrand is $\prod_{j=1}^{n} (\sum_{a=1}^{A_j} T_{j,a}(x))^{\frac{1}{n-1}}$, which is big if x lies in many tubes from each direction. So the integral on the left-hand side measures how many points x lie in many tubes from each direction. The multilinear Kakeya inequality says that there cannot be too many points which lie in many tubes from each direction.

The exponent $\frac{1}{n-1}$ makes the inequality sharp in two natural examples: the example when all the tubes go through the origin and an example when the tubes are arranged in a rectangular grid. The exponent $\frac{1}{n-1}$ is the most important, and this bound implies sharp estimates with any other exponent.

It makes sense to compare Theorem 5.3 with the L^p Kakeya conjecture, Conjecture 5.2. The main difference between them is that in the multilinear Kakeya theorem, the integrand is a product of n factors, and we assume that the n factors are transverse to each other in a strong sense. The word "multilinear" refers to this product structure.

Theorem 5.3 is also proven by induction on scales. In the case that the tubes $T_{j,a}$ are exactly parallel to the x_j axis (for all j and a), Theorem 5.3 reduces to the Loomis–Whitney inequality [39], which we will recall a moment. The general case of multilinear Kakeya is proven by applying Loomis–Whitney at many scales (cf. [31]). The multilinear Kakeya inequality grew out of work by Bennett–Carbery–Wright on nonlinear versions of the Loomis–Whitney inequality [4].

For completeness let us recall the statement of the Loomis–Whitney inequality. One version is an inequality for integrals that looks reminiscent of Hölder's inequality. Suppose that $\pi_j : \mathbb{R}^n \to \mathbb{R}^{n-1}$ are projections onto the coordinate hyperplanes. Then the Loomis–Whitney inequality says

$$\int_{\mathbb{R}^n} \prod_{j=1}^{n} f_j\big(\pi_j(x)\big)^{\frac{1}{n-1}}\, dx \le \prod_{j=1}^{n} \|f_j\|_{L^1(\mathbb{R}^{n-1})}^{\frac{1}{n-1}}.$$

There is a geometric corollary of this inequality which may feel more intuitive. Suppose that $U \subset \mathbb{R}^n$ is an open set, and that the projection of U onto every coordinate hyperplane has $(n-1)$-volume at most A. Then U has n-volume at most $A^{\frac{n}{n-1}}$. The case $n = 2$ is straightforward, but the case $n = 3$ is quite subtle. It is one of my favorite problems to think through with students studying analysis.

When multilinear Kakeya was first proven, it seemed natural and remarkable, but it was not clear just how much impact it would have in restriction theory. In [3], Bennett, Carbery, and Tao [3] formulated and proved an interesting multilinear restriction conjecture. They proved multilinear restriction by using multilinear Kakeya at many scales. But it was not clear whether these multilinear estimates would lead to bounds on problems that were not multilinear, such as the original restriction conjecture.

The paper [17] used these multilinear estimates to prove new partial results about the restriction problem. It introduced a technique called the broad/narrow method which can sometimes reduce linear estimates to multilinear estimates.

Remarkably, sharp decoupling theorems follow from multilinear Kakeya, even though there is nothing obviously multilinear about the statement of decoupling. This was one of the big surprises in the development of the field. The original Kakeya problem is much harder than multilinear Kakeya. The original restriction problem is much harder than multilinear restriction. There is also a multilinear version of decoupling. A key fact that makes decoupling accessible is that the original decoupling problem is EQUIVALENT to multilinear decoupling. This equivalence was noticed implicitly by Bourgain in [10], and

explicitly by Bourgain and Demeter in [14]. Because of this connection between decoupling and multilinear decoupling, we can prove sharp estimates for the original decoupling problem using multilinear Kakeya, even though we do not know sharp estimates for the original Kakeya problem.

The connection between decoupling and multilinear decoupling is another important application of induction on scales. It is based on the broad/narrow method. Because of considerations of space, we do not give a detailed description here.

When multilinear Kakeya first appeared, it seemed like it might not have very many applications in harmonic analysis compared with the original Kakeya conjecture. But now the situation has reversed: multilinear Kakeya currently has more applications in harmonic analysis than the original Kakeya conjecture would have even if we knew it.

6. APPLICATIONS OF DECOUPLING IN HARMONIC ANALYSIS

Decoupling theory has led to the solutions of several longstanding problems in harmonic analysis. We give three examples here. Each of these problems seemed out of reach a decade ago.

6.1. The helical maximal function

Hardy and Littlewood introduced their maximal function in the early 20th century. The Hardy–Littlewood maximal function is based on averages over balls. If $f : \mathbb{R}^n \to \mathbb{R}$, then the average value of f on the ball of radius r around x can be written as

$$\frac{1}{|B_r|}\int_{B_r} f(x+y)dy.$$

The Hardy–Littlewood maximal function is defined by taking the supremum over r,

$$Mf(x) = \sup_r \frac{1}{|B_r|}\int_{B_r} \big|f(x+y)\big|dy.$$

Hardy and Littlewood proved that $\|Mf\|_{L^p} \le C(p,n)\|f\|_{L^p}$ for all $p > 1$ but not for $p = 1$.

In the 1960s, Stein introduced a spherical maximal function [48]. Suppose $f : \mathbb{R}^n \to \mathbb{R}$. The average value of f on the sphere of radius r around x can be written as

$$\frac{1}{|S^{n-1}|}\int_{S^{n-1}} f(x+r\theta)d\theta.$$

The spherical maximal function is defined by taking the supremum over r,

$$M_S f(x) := \sup_{r>0} \frac{1}{|S^{n-1}|}\int_{S^{n-1}} \big|f(x+r\theta)\big|d\theta. \tag{6.1}$$

For $n \ge 3$, Stein proved that in $\mathbb{R}^n$, $\|M_S f\|_{L^p} \le C(n,p)\|f\|_{L^p}$ for all $p > \frac{n}{n-1}$, but not for $p \le \frac{n}{n-1}$. He conjectured that the same was true for $n = 2$. The case $n = 2$ was proven by Bourgain in [5].

Stein's result was striking for the following reason. A function $f \in L^p$ need only be defined almost everywhere. It may be undefined or infinite on a lower-dimensional submanifold like a sphere. So for a particular x and r, the integral on the right-hand side of

(6.1) may be infinite or undefined. Nevertheless, if $f \in L^p$ for $p > \frac{n}{n-1}$, Stein showed that the spherical maximal function is actually defined for almost every x. The curvature of the sphere is crucial in this estimate. The spherical maximal function and the restriction conjecture were two fundamental connections between curvature and harmonic analysis that Stein investigated.

The spherical maximal function can be generalized by replacing the sphere by other curved submanifolds. Many of the corresponding problems are still open. After the sphere and circle, the next most fundamental case to look at is the case of the moment curve in $\mathbb{R}^n$. Here is the definition. Consider the moment curve parametrized by $\gamma(t) = (t, t^2, t^3, \dots, t^n)$. We can build an averaging operator based on the moment curve as follows. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and define

$$Af(x) = \int_0^1 f\big(x + \gamma(t)\big)dt.$$

Geometrically, $Af(x)$ is the average value of f on the translate of the moment curve starting at x. Next we can consider different scalings of the moment curve. Define

$$A_r f(x) = \int_0^1 f\big(x + r\gamma(t)\big)dt.$$

Geometrically, $A_r f(x)$ is the average value of f on a moment curve which has been scaled by a factor of r and then translated to start at x. Finally, we can define the helical maximal function by taking the maximum of these averages over different choices of r,

$$M_{\text{hel}} f(x) := \sup_{r>0} A_r f(x).$$

In analogy with the work of Stein and Bourgain on the circular maximal function, it is natural to ask when $\|M_{\text{hel}} f\|_{L^p(\mathbb{R}^n)} \lesssim \|f\|_{L^p(\mathbb{R}^n)}$. In [45], Pramanik and Seeger connected this problem (when $n = 3$) to the decoupling problem for the cone, which Wolff had recently introduced in [56]. In [14], Bourgain and Demeter gave sharp estimates for the decoupling for the cone, but that by itself is not enough to give sharp estimates for the helical maximal function. Recently, Ko–Lee–Oh [37] and Beltran–Guo–Hickman–Seeger [2] independently proved the sharp L^p-estimate for the helical maximal function when $n = 3$.

Theorem 6.1 ([37] and [2]). *For $p > 3$ $\|M_{\text{hel}} f\|_{L^p(\mathbb{R}^3)} \le C(p)\|f\|_{L^p(\mathbb{R}^3)}$.*
If $p \le 3$, this estimate does not hold.

The case of higher dimensions remains open, although both groups have proven interesting estimates on helical averages in other dimensions as well.

6.2. Pointwise convergence for the Schrödinger equation

Consider the initial value problem for the linear Schrödinger equation in $\mathbb{R}^d \times \mathbb{R}$,

$$\partial_t u(x,t) = i\triangle u(x,t), u(x,0) = u_0(x).$$

We can write down the solution u with the help of the Fourier transform. If the initial data u_0 is rough, then the solution $u(x,t)$ will be rough also. In this situation, $u(x,t)$ will solve the differential equation in a distributional sense, even if $u(x,t)$ is discontinuous.

Carleson [20] raised the following problem.

Question 6.2. What is the smallest s so that whenever $u_0 \in H^s(\mathbb{R}^d)$ and $u(x,t)$ is a distributional solution to the Schrödinger equation on $\mathbb{R}^d \times \mathbb{R}$ with initial data $u_0(x)$, then $\lim_{t\to 0} u(x,t) = u_0(x)$ for almost every $x \in \mathbb{R}^d$?

This question helps describe how regular distributional solutions to the Schrodinger equation are. This question is actually a cousin of the restriction problem and the Strichartz estimate, although we will have to rewrite it a little bit to see how they are connected.

Because u solves the Schrödinger equation, the spacetime Fourier transform $\hat{u}$ is supported on the infinite paraboloid. One has to prove some estimates about how badly $u(x,t)$ oscillates for small t. After some standard arguments (scaling and Littlewood–Paley), one can reduce these estimates to the case that $\hat{u}$ is supported on the truncated paraboloid P and normalize so that $\|u_0\|_{L^2(\mathbb{R}^d)} = 1$. Now consider $U_\lambda(u) \subset \mathbb{R}^d \times \mathbb{R}$. The Strichartz estimates give sharp bounds for $|U_\lambda(u)|$ in terms of λ. A small variation gives sharp estimates for $|U_\lambda(u) \cap [0,R]^{d+1}|$ in terms of λ and R. Now let $\Pi_{\mathbb{R}^d}(x,t) = x$ be the projection from spacetime to space. Carleson's pointwise convergence problem is related to the following question about the size of $\Pi_{\mathbb{R}^d}(U_\lambda(u))$:

Question 6.3. Suppose that $\hat{u}$ is supported on the truncated paraboloid P. Let $u_0(x) = u(x,0)$, and suppose that $\|u_0\|_{L^2(\mathbb{R}^d)} = 1$. For any given λ, R, estimate the maximum possible size of $|\Pi_{\mathbb{R}^d}(U_\lambda u \cap [0,R]^{d+1})|$.

The key difference between this problem and the Strichartz inequality is we have to estimate the d-volume of the projection of $U_\lambda(u)$ instead of the $(d+1)$-volume of $U_\lambda(u)$ itself. This general problem is still open. However, we do understand a special case, which is sufficient to resolve the pointwise convergence problem. Here is the special case:

Question 6.4. Suppose that $\hat{u}$ is supported on the truncated paraboloid P. Let $u_0(x) = u(x,0)$, and suppose that $\|u_0\|_{L^2(\mathbb{R}^d)} = 1$. Suppose that $|\Pi_{\mathbb{R}^d}(U_\lambda u \cap [0,R]^{d+1})| \geq cR^d$. How big can λ be?

As a first example, suppose that u_0 is a smooth bump function approximating a constant function on $[0,R]^d$. Because $\|u_0\|_{L^2} = 1$, we have $|u_0(x)| \sim R^{-d/2}$ on most of $[0,R]^d$. In this case, $u(x,t)$ is roughly constant on $[0,R]^{d+1}$, and so λ is also $\sim R^{-d/2}$.

This first example is not the worst case. In case $d = 1$, the worst case example was found by Dahlberg–Kenig [21]. It is given when $u(x,t)$ is a single wave packet, essentially supported on a tilted rectangle with dimensions $R^{1/2} \times R$.

In this case, $u_0(x)$ is essentially supported on an interval of length $R^{1/2}$, and so $|u_0(x)| \sim R^{-1/4}$ on this interval. Then $|u(x,t)| \sim R^{-1/4}$ on the whole wave packet, and we get $\lambda \sim R^{-1/4}$. Carleson [20] had showed previously that this value of λ is optimal. This settles Carleson's problem in the case $d = 1$, but the case of higher dimensions was open for 30+ years.

In higher dimensions, we can adapt the Dahlberg–Kenig example by taking many parallel wave packets with disjoint projections onto $\mathbb{R}^d$. This gives $\lambda = R^{-\frac{d}{2}+\frac{1}{4}}$. For a long

time, it seemed plausible that this construction was sharp in any dimension. In the last decade, mathematicians found other much more intricate examples. The first was given by Bourgain [11] and there were several improvements leading up to [12] (cf. also [40]). The last example gives $\lambda = R^{-\frac{d}{2}+\frac{d}{2d+2}}$.

This last example turns out to be sharp. The case $d = 2$ was proven in [22] and the case of all d was proven in [23]. Even for $d = 2$, the proof in [23] is simpler. The key ingredient in these proofs is decoupling. Decoupling is applied in a somewhat indirect way. In particular, the proofs use decoupling many times at different scales.

Theorem 6.5 ([12, 23]). *Suppose that $s > \frac{d}{2d+2}$. If $u_0 \in H^s(\mathbb{R}^d)$, and $u(x,t)$ is a (distributional) solution to the linear Schrödinger equation with initial data u_0. Then $\lim_{t\to 0} u(x,t) = u_0(x)$ for almost every x.*

Suppose that $s < \frac{d}{2d+2}$. There exists a function $u_0 \in H^s(\mathbb{R}^d)$ with the following bad behavior. Let $u(x,t)$ be the (distributional) solution to the linear Schrödinger equation with initial data u_0. For this function, $\limsup_{t\to 0} |u(x,t)| = +\infty$ for almost every $x \in \mathbb{R}^d$.

6.3. The local smoothing problem

Wolff introduced decoupling in his work on the local smoothing problem [56]. This problem is an estimate about solutions to the wave equation.

Suppose that $u(x,t)$ solves the wave equation $\partial_t^2 u = \triangle u$, with $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$, and with initial data $u(x,0) = u_0(x)$ and $\partial_t u(x,0) = u_1(x)$. The local smoothing problem concerns Sobolev-type bounds for the wave equation: Given bounds on some Sobolev norms of u_0 and u_1, what bounds can we prove on the Sobolev norms of u?

To make things simple and concrete, let us suppose that $u_1 = 0$ and that $\hat{u}_0$ is supported in a ball of radius N in frequency space. Then we would like to find all bounds of the form

$$\|u(x,t)\|_{L^p(\mathbb{R}^d\times[0,1])} \le C N^\alpha \|u_0(x)\|_{L^p(\mathbb{R}^d)}.$$

The word "local" in "local smoothing" refers to the time interval $[0,1]$. A global estimate would give a bound on $\mathbb{R}^d \times \mathbb{R}$, whereas a fixed-time estimate would give a bound for $\mathbb{R}^d \times \{t_0\}$ for some fixed t_0 (such as $t_0 = 1$). Global in time estimates, local in time estimates, and fixed time estimates are all interesting. Sharp fixed time estimates were established by Peral [44] and Miyachi [41] around 1980. The word "smoothing" in "local smoothing" is because the power of α in the local in time estimates is smaller than the power in a fixed time estimate.

In [47], Sogge formulated the local smoothing conjecture, and he proved the first local smoothing estimates improving upon the α given by the fixed time estimates.

Conjecture 6.6 ([47]). *Suppose $d \ge 2$. Suppose that $u(x,t)$ solves the wave equation in $\mathbb{R}^d \times \mathbb{R}$, with initial data $u(x,0) = u_0(x)$ and $\partial_t u(x,0) = 0$. Suppose that $\hat{u}_0$ is supported in the ball of radius N. Then, if $2 \le p \le \frac{2d}{d-1}$, then*

$$\|u(x,t)\|_{L^p(\mathbb{R}^d\times[0,1])} \le C(d,\varepsilon) N^\varepsilon \|u_0\|_{L^p(\mathbb{R}^d)}.$$

If $p > \frac{2d}{d-1}$, *then*

$$\|u(x,t)\|_{L^p(\mathbb{R}^d \times [0,1])} \leq C(d,\varepsilon) N^{\frac{d-1}{2} - \frac{d}{p} + \varepsilon} \|u_0\|_{L^p(\mathbb{R}^d)}.$$

The case $p = \frac{2d}{d-1}$ is the critical exponent, and it implies all the other estimates for a given dimension d. In [56], Wolff introduced decoupling and used it to show that Conjecture 6.6 holds when $d = 2$ and $p > 74$. Wolff also observed that the local smoothing conjecture in dimension d implies the Kakeya conjecture in dimension d, by adapting Fefferman's argument from [25]. Therefore, the full conjecture remains out of reach for all $d \geq 3$.

In [14], Bourgain and Demeter proved a complete decoupling theorem for the cone. This implies that Conjecture 6.6 holds in $\mathbb{R}^d$ for all $p > \frac{2(d+1)}{d-1}$. In particular, when $d = 2$, local smoothing holds for all $p > 6$. When $d = 2$, the critical exponent for local smoothing is $p = 4$.

In [33], Wang, Zhang, and I proved the local smoothing conjecture when $d = 2$ for $p = 4$ (and hence for all p). The proof of local smoothing does not use decoupling per se, but it is strongly influenced by the ideas in the proof of decoupling, including induction on scales.

7. FRUSTRATIONS, LIMITATIONS, AND OPEN PROBLEMS

Decoupling and the ideas in the proof of decoupling have led to solutions of many problems that seemed out of reach a decade ago. The proof is elegant in some ways. In some ways, it feels like a proof "from the book." It is essentially self-contained and it is not that long. But in other ways the proof is frustrating. (Actually, there are now several proofs, and they have various advantages and disadvantages. The community is actively trying to understand decoupling from different angles, and in five or ten years, we may have a different sense of the essential ingredients.)

In this section, I discuss some of my frustrations with the proof of decoupling, some limitations of the method, and some open problems.

7.1. Too much induction

On the one hand, induction on scales is the central idea in the proof of decoupling. On the other hand, the heavy reliance on induction makes the proof difficult to read. A lot of important stuff is happening inside the induction.

For example, as we discussed in Section 4.5, I think that the leverage in the proof of decoupling comes from taking advantage of the transversality of wave packets of every scale, not just at a few scales. For instance, suppose we cut the parabola P into M rectangles τ with $M = N^{5/16}$. The proof of decoupling takes advantage of the transversality between the wave packets at this scale, but it is not easy to locate the place in the argument where this transversality is used because it is a little bit buried in the induction. Even though I have thought through the proof many times, it took me a good while to locate where wave packets at this particular scale are used.

Reading through the full proof of decoupling for the paraboloid, we see many different tricks for taking advantage of induction on scales. Loomis–Whitney is used at many scales to prove multilinear Kakeya. Multilinear Kakeya is used at many scales in the argument in Section 4.4. The key induction on scales is described in Section 4.5. Induction on scales is also used in a different way to go back and forth between multilinear estimates and the original linear estimates, as we discussed in Section 5.1. Finally, many applications of decoupling actually use decoupling many times at different scales, as in Section 6.2.

We might look at this and feel that using multiple scales is a craft with many aspects. But we might also start to get the feeling that this is too many different tricks, and that we should try to take advantage of many scales in a more systematic way.

7.2. What does decoupling say about the shapes of superlevel sets?

Decoupling gives an estimate for $\|f\|_{L^p}$ or for the measure of the superlevel sets $U_\lambda(f)$. Besides the measure of the sets $U_\lambda(f)$, decoupling also seems to be connected to the shape of the superlevel sets $U_\lambda(f)$. Looking back through our discussion in Sections 4.4 and 4.5, the shape of $U_\lambda(f)$ plays an important role, even though the final estimate only concerns the measure of $U_\lambda(f)$. In particular, during the argument, we make use of some information about $|U_\lambda(f_\tau) \cap B|$ for various balls B and for various τ. This information roughly describes how much the set $U_\lambda(f)$ can concentrate in balls. The shape of $U_\lambda(f)$ is also connected to some applications of decoupling, such as the work on Carleson's pointwise convergence problem discussed in Section 6.2.

Perhaps the shape of $U_\lambda(f)$ should be a more central character in decoupling. What is the full information about the shape of $U_\lambda(f)$ which the proof method of decoupling gives? Unfortunately this question is quite vague. There are many possible ways we could describe the shape of $U_\lambda(f)$, and it is not clear which language to use. But it is possible that discussing the shape of $U_\lambda(f)$ systematically throughout the whole story might make the arguments clearer or even stronger…

Here is one question from the harmonic analysis literature that has to do with the shape of $U_\lambda(f)$. We consider a measure μ supported on a large ball $B_R \subset \mathbb{R}^n$ which obeys the Frostman condition

$$\mu\big(B_r(x)\big) \le r^\alpha. \tag{7.1}$$

Here $0 < \alpha < n$ is fixed.

Question 7.1. As in the restriction problem or the Strichartz inequality, suppose that $f : \mathbb{R}^n \to \mathbb{C}$ is given by

$$f(x) = \int_P a(\omega) e^{2\pi i \omega x} d\mu_P(\omega).$$

For a given n and α, what is the best exponent γ in the inequality

$$\|f\|_{L^2(d\mu)} \le C R^\gamma \|a\|_{L^2(P)},$$

among all functions f as above and all measures μ obeying the Frostman condition (7.1) with exponent α.

In dimension $n = 2$, this question is well understood for all α by work of Mattila and Wolff, cf. [55]. But for $n \geq 3$, the problem is far from understood. In [23], Du and Zhang gave a sharp answer for $\alpha = n - 1$. No other cases are fully understood. The Du–Zhang estimate for $\alpha = n - 1$ is closely related to the solution of Carleson's problem on pointwise convergence for solutions of the Schrödinger equation. Decoupling and multilinear restriction are the essential tools in their approach, and they use decoupling at many different scales.

How much can the method of decoupling tell us about other values of α? Is there anything fundamentally special about $\alpha = n - 1$? Also the Frostman condition (7.1) can be replaced by other conditions, by replacing the function r^α by other functions of r. This would lead to other kinds of estimates about the shape of $U_\lambda(f)$.

7.3. Limitations of the information used in the proof

In the statement of decoupling, we assume that $\hat{f}$ is supported in Ω, and we try to bound $\|f\|_{L^p}$ in terms of some information about $\|f_\theta\|_{L^p}$ for all the θ in the decomposition of Ω. If we look through the proof and check where the hypothesis $\operatorname{supp}(\hat{f}) \subset \Omega$ is used, we find that it is used only in fairly simple ways.

In the course of the proof, we consider f_τ for many different rectangles τ. The proof relies crucially on two facts. The first is the locally constant heuristic:

$$\text{For each } \tau, \ |f_\tau| \text{ is approximately constant on each translate of } \tau*. \tag{7.2}$$

The second is the local orthogonality heuristic. If τ is a rectangle, and γ are smaller rectangles contained in τ, and if nonadjacent γ are separated by at least s, then

$$\int_B |f_\tau|^2 \approx \sum_{\gamma \subset \tau} \int_B |f_\gamma|^2, \tag{7.3}$$

whenever B is a cube whose side length is longer than s^{-1}.

The Fourier support properties of the different functions f, f_τ, f_θ are only really used to justify these two heuristics. These two heuristics are consequences of the Fourier support hypotheses, but they do not encode all the information given by the Fourier support hypotheses.

This raises the question: Which theorems of restriction theory can we prove only using the locally constant heuristic and local orthogonality? Which theorems require us to use the Fourier support hypothesis in some other way?

The proofs of the different decoupling theorems essentially only use these two properties. (I say essentially because some of the proofs also involve some pigeonholing of wave packets.) Also, the strongest current work on the restriction conjecture only uses these two properties. It is possible that the full restriction conjecture might follow only using these two properties.

In restriction theory there are currently very few examples of techniques for exploiting the Fourier support of f that use Fourier support information in some other way. (One example is to use the even integer trick, Lemma 1.7, together with number-theoretic input. An interesting recent example of this approach is the work on Strichartz-type estimates for the periodic Airy equation by Hughes–Wooley [34].)

However, there are a number of problems in restriction theory where I strongly doubt that these two properties are sufficient to give full answers. One example is the the problem of estimating the L^p-norms of the functions

$$f_{k,N}(x) = \sum_{a=1}^{N} e^{2\pi i a^k x}.$$

As we discussed in Section 1.2, the L^p-norms of $f_{k,N}$ are well understood for $k = 2$ and wide open for $k \geq 3$. When $k = 2$, the different proofs all use some information besides the locally constant heuristic and local orthogonality. I believe the sharp estimates for $k = 2$ cannot be proven by an argument using only those two properties.

There is an interesting generalization of this L^p-problem which I think is a good test case for going beyond the locally constant property and local orthogonality. As we mentioned in Section 1.2, $\| f_{2,N} \|_{L^4([0,1])} \leq C_\varepsilon N^{1/2+\varepsilon}$.

Question 7.2. We consider a sequence of frequencies ω_a, with $a = 1, \ldots, N$, which behave approximately like the squares a^2, in the sense that

$$\omega_{a+1} - \omega_a \sim a \quad \text{and} \quad (\omega_{a+1} - \omega_a) - (\omega_a - \omega_{a-1}) \sim 1.$$

For such a choice of frequencies ω_a, define

$$f(x) = \sum_{a=1}^{N} e^{2\pi i \omega_a x}.$$

Estimate $\| f \|_{L^4([0,1])}$. Is it true that $\| f \|_{L^4([0,1])} \leq C_\varepsilon N^{1/2+\varepsilon}$?

As far as I know, it is possible that $\| f \|_{L^4([0,1])} \leq C_\varepsilon N^{1/2+\varepsilon}$ in this much more general setting. However, the proofs that work for $f_{2,N}$ do not generalize to this setting. And the method of decoupling can prove only limited things. In [27], Fu, Maldague, and I explored how much we can say about this question using ideas of decoupling theory. As part of that investigation, we explain the version of the locally constant property which appears in this setting, which goes back to Bourgain's work [7] on Montgomery's conjecture. The main theorems of [27] give sharp L^p-estimates for much shorter sums, namely sums of length $\sim N^{1/2}$. For these shorter sums, the locally constant property and the methods of decoupling are effective. But for longer sums, they seem much less effective, and I believe that some different tools are needed.

Question 7.2 is also related to a question of Erdős about sumsets of convex sets. A sequence ω_a is called convex if $(\omega_{a+1} - \omega_a) - (\omega_a - \omega_{a-1}) > 0$ for all a. Notice that the set of frequencies in Question 7.2 is a convex sequence. If A is a convex sequence, then Erdős conjectured that $|A + A| \geq c_\varepsilon |A|^{2-\varepsilon}$. Here $A + A$ denotes all sums of two elements of A. This conjecture is open. There is interesting recent work on it by Schoen and Shkredov [46], who proved that $|A + A| \geq c_\varepsilon |A|^{1.6-\varepsilon}$. This beats the previous best estimate $|A|^{1.5}$, which had stood for a long time. If A denotes the frequencies in Question 7.2, and if indeed $\| f \|_{L^4([0,1])} \leq C_\varepsilon N^{1/2+\varepsilon}$, then it would follow that $|A + A| \geq c_\varepsilon |A|^{2-\varepsilon}$. The best bound I

could prove using the methods of decoupling gives $|A + A| \geq c|A|^{1.5}$. Work in combinatorics such as [46] may give clues on how to go further in problems like Question 7.2.

ACKNOWLEDGMENTS

The author is supported by a Simons Investigator award.

REFERENCES

[1] G. I. Arkhipov, V. N. Chubarikov, and A. A. Karatsuba, *Trigonometric sums in number theory and analysis*. de Gruyter Exp. Math. 39, Walter de Gruyter GmbH & Co. KG, Berlin, 2004. Translated from the 1987 Russian original.

[2] D. Beltran, J. Hickman, S. Guo, and A. Seeger, Sharp L^p bounds for the helical maximal function. 2021, arXiv:2102.08272.

[3] J. Bennett, A. Carbery, and T. Tao, On the multilinear restriction and Kakeya conjectures. *Acta Math.* **196** (2006), no. 2, 261–302.

[4] J. Bennett, A. Carbery, and J. Wright, A non-linear generalisation of the Loomis–Whitney inequality and applications. *Math. Res. Lett.* **12** (2005), no. 4, 443–457.

[5] J. Bourgain, Averages in the plane over convex curves and maximal operators. *J. Anal. Math.* **47** (1986), 69–85.

[6] J. Bourgain, Besicovitch type maximal operators and applications to Fourier analysis. *Geom. Funct. Anal.* **1** (1991), no. 2, 147–187.

[7] J. Bourgain, Remarks on Montgomery's conjectures on Dirichlet sums. In *Geometric aspects of functional analysis (1989–1990)*, pp. 153–165, Lecture Notes in Math. 1469, Springer, Berlin, 1991.

[8] J. Bourgain, Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations. I. Schrödinger equations. *Geom. Funct. Anal.* **3** (1993), no. 2, 107–156.

[9] J. Bourgain, On the dimension of Kakeya sets and related maximal inequalities. *Geom. Funct. Anal.* **9** (1999), no. 2, 256–282.

[10] J. Bourgain, Moment inequalities for trigonometric polynomials with spectrum in curved hypersurfaces. *Israel J. Math.* **193** (2013), no. 1, 441–458.

[11] J. Bourgain, On the Schrödinger maximal function in higher dimension. *Proc. Steklov Inst. Math.* **280** (2013), no. 1, 46–60.

[12] J. Bourgain, A note on the Schrödinger maximal function. *J. Anal. Math.* **130** (2016), 393–396.

[13] J. Bourgain, Decoupling, exponential sums and the Riemann zeta function. *J. Amer. Math. Soc.* **30** (2017), no. 1, 205–224.

[14] J. Bourgain and C. Demeter, The proof of the l^2 decoupling conjecture. *Ann. of Math. (2)* **182** (2015), no. 1, 351–389.

[15] J. Bourgain and C. Demeter, Decouplings for curves and hypersurfaces with nonzero Gaussian curvature. *J. Anal. Math.* **133** (2017), 279–311.

[16] J. Bourgain, C. Demeter, and L. Guth, Proof of the main conjecture in Vinogradov's mean value theorem for degrees higher than three. *Ann. of Math. (2)* **184** (2016), no. 2, 633–682.

[17] J. Bourgain and L. Guth, Bounds on oscillatory integral operators based on multilinear estimates. *Geom. Funct. Anal.* **21** (2011), no. 6, 1239–1295.

[18] J. Bourgain and N. Watt, Mean square of zeta function, circle problem and divisor problem revisited. 2017, arXiv:1709.04340.

[19] N. Burq, P. Gérard, and N. Tzvetkov, Strichartz inequalities and the nonlinear Schrödinger equation on compact manifolds. *Amer. J. Math.* **126** (2004), no. 3, 569–605.

[20] L. Carleson, Some analytic problems related to statistical mechanics. In *Euclidean harmonic analysis (Proc. Sem., Univ. Maryland, College Park, Md, 1979)*, pp. 5–45, Lecture Notes in Math. 779, 1979.

[21] B. Dahlberg and C. Kenig, A note on the almost everywhere behavior of solutions to the Schrödinger equation. In *Harmonic analysis (Minneapolis, MN, 1981)*, pp. 205–209, Lecture Notes in Math. 908, Springer, Berlin–New York, 1982.

[22] X. Du, L. Guth, and X. Li, A sharp Schrödinger maximal estimate in $\mathbb{R}^2$. *Ann. of Math. (2)* **186** (2017), no. 2, 607–640.

[23] X. Du and R. Zhang, Sharp L^2 estimates of the Schrödinger maximal function in higher dimensions. *Ann. of Math. (2)* **189** (2019), no. 3, 837–861.

[24] Z. Dvir, On the size of Kakeya sets in finite fields. *J. Amer. Math. Soc.* **22** (2009), no. 4, 1093–1097.

[25] C. Fefferman, The multiplier problem for the ball. *Ann. of Math. (2)* **94** (1971), 330–336.

[26] C. Fefferman, A note on spherical summation multipliers. *Israel J. Math.* **15** (1973), 44–52.

[27] Y. Fu, L. Guth, and D. Maldague, Decoupling inequalities for short generalized Dirichlet sequences. 2021, arXiv:2104.00856.

[28] S. Guo, Z. Li, P.-L. Yung, and P. Zorin-Kranich, A short proof of ℓ^2 decoupling for the moment curve. *Amer. J. Math.* **143** (2021), no. 6, 1983–1998.

[29] S. Guo and R. Zhang, On integer solutions of Parsell–Vinogradov systems. *Invent. Math.* **218** (2019), no. 1, 1–81.

[30] S. Guo and P. Zorin-Kranich, Decoupling for moment manifolds associated to Arkhipov–Chubarikov–Karatsuba systems. *Adv. Math.* **360** (2020).

[31] L. Guth, A short proof of the multilinear Kakeya inequality. *Math. Proc. Cambridge Philos. Soc.* **158** (2015), no. 1, 147–153.

[32] L. Guth, D. Maldague, and H. Wang, Improved decoupling for the parabola. 2020, arXiv:2009.07953.

[33] L. Guth, H. Wang, and R. Zhang, A sharp square function estimate for the cone in $\mathbb{R}^3$. *Ann. of Math. (2)* **192** (2020), no. 2, 551–581.

[34] K. Hughes and T. Wooley, Discrete restriction for (x, x^3) and related topics. 2019, arXiv:1911.12262.

[35] N. Katz, I. Laba, and T. Tao, An improved bound on the Minkowski dimension of Besicovitch sets in $\mathbb{R}^3$. *Ann. of Math. (2)* **152** (2000), no. 2, 383–446.

[36] N. Katz and J. Zahl, An improved bound on the Hausdorff dimension of Besicovitch sets in $\mathbb{R}^3$. *J. Amer. Math. Soc.* **32** (2019), no. 1, 195–259.

[37] H. Ko, S. Lee, and S. Oh, Maximal estimates for averages over space curves. 2021, arXiv:2102.07175.

[38] Z. Li, An ℓ^2 decoupling interpretation of efficient congruencing: the parabola. *Rev. Mat. Iberoam.* **37** (2021), no. 5, 1761–1802.

[39] L. Loomis and H. Whitney, An inequality related to the isoperimetric inequality. *Bull. Amer. Math. Soc.* **55** (1949), 961–962.

[40] R. Luca and K. Rogers, A note on pointwise convergence for the Schrödinger equation. *Math. Proc. Cambridge Philos. Soc.* **166** (2019), no. 2, 209–218.

[41] A. Miyachi, On some estimates for the wave equation in L^p and H^p. *J. Fac. Sci., Univ. Tokyo, Sect. 1A, Math.* **27** (1980), no. 2, 331–354.

[42] H. Montgomery, Mean and large values of Dirichlet polynomials. *Invent. Math.* **8** (1969), 334–345.

[43] T. Orponen, On the dimension and smoothness of radial projections. *Anal. PDE* **12** (2019), no. 5, 1273–1294.

[44] J. C. Peral, L^p estimates for the wave equation. *J. Funct. Anal.* **36** (1980), no. 1, 114–145.

[45] M. Pramanik and A. Seeger, L^p regularity of averages over curves and bounds for associated maximal operators. *Amer. J. Math.* **129** (2007), no. 1, 61–103.

[46] T. Schoen and I. Shkredov, On sumsets of convex sets. *Combin. Probab. Comput.* **20** (2011), no. 5, 793–798.

[47] C. Sogge, Propagation of singularities and maximal functions in the plane. *Invent. Math.* **104** (1991), no. 2, 349–376.

[48] E. Stein, Maximal functions. I. Spherical means. *Proc. Natl. Acad. Sci. USA* **73** (1976), no. 7, 2174–2175.

[49] E. Stein, Some problems in harmonic analysis. In *Harmonic analysis in Euclidean spaces (Proc. Sympos. Pure Math., Williams Coll., Williamstown, MA, 1978), Part 1*, pp. 3–20, Proc. Sympos. Pure Math. XXXV, Amer. Math. Soc., Providence, RI 1978.

[50] E. Stein, *Oscillatory integrals in Fourier analysis, Beijing lectures in harmonic analysis*. Ann. Math. Stat. 112, Princeton University Press, 1986.

[51] R. Strichartz, A priori estimates for the wave equation and some applications. *J. Funct. Anal.* **5** (1970), 218–235.

[52] I. Vinogradov, The method of trigonometrical sums in the theory of numbers (in Russian). *Trav. Inst. Math. Stekloff* **23** (1947), 109 pp.

[53] T. Wolff, An improved bound for Kakeya type maximal functions. *Rev. Mat. Iberoam.* **11** (1995), no. 3, 651–674.

[54] T. Wolff, A Kakeya-type problem for circles. *Amer. J. Math.* **119** (1997), no. 5, 985–1026.

[55] T. Wolff, Decay of circular means of Fourier transforms of measures. *Int. Math. Res. Not.* **10** (1999), 547–567.

[56] T. Wolff, Local smoothing type estimates on L^p for large p. *Geom. Funct. Anal.* **10** (2000), no. 5, 1237–1288.

[57] T. Wooley, Large improvements in Waring's problem. *Ann. of Math. (2)* **135** (1992), no. 1, 131–164.

[58] T. Wooley, The cubic case of the main conjecture in Vinogradov's mean value theorem. *Adv. Math.* **294** (2016), 532–561.

[59] T. Wooley, Nested efficient congruencing and relatives of Vinogradov's mean value theorem. *Proc. Lond. Math. Soc. (3)* **118** (2019), no. 4, 942–1016.

LARRY GUTH

Department of Mathematics, MIT, 77 Massachusetts Ave, Cambridge, MA 02139, USA, larryg@mit.edu