LECTURE NOTES 5 FOR 254A

TERENCE TAO

1. Overview

In this set of notes we return to the phase plane. This time, we will study the *Walsh model* of the phase plane as opposed to the standard Fourier model, the difference being that the underlying group is $\mathbf{Z}_2[[x]]$ rather than $\mathbf{R}$. This model is less useful for applications (for instance, there is no good notion of a derivative, and hence there are no applications to PDE), however it is a good model for building intuition because it is "cleaner" than the Fourier model. For instance, in the Walsh model it is possible to localize both in space and frequency simultaneously.

To describe the Walsh model, we begin with the half-line $\mathbf{R}_+$. Except on a set of measure zero (which we shall always ignore), every point x on $\mathbf{R}_+$ can be identified with a doubly-infinite set of binary digits

$$x = \dots a_3 a_2 a_1 a_0 . a_{-1} a_{-2} a_{-3} \dots$$

where $a_n \in Z_2$ and a_n is eventually zero as $n \to +\infty$. We use $a_n(x)$ to denote the n^{th} binary digit of x. We give this half-line a norm $\|x\|$, defined to be the largest power of 2 which is less than or equal to x.

We give $\mathbf{R}^+$ the field structure of $\mathbf{Z}_2[[x]]$ (i.e. identifying binary strings with $\mathbf{Z}_2$-valued Laurent polynomials $\sum_n a_n x^n$; this space can also be thought of as $\mathbf{Z}_2^{\mathbf{Z}}$). By this, we mean that we define addition $\oplus$ on $\mathbf{R}^+$ by

$$a_n(x \oplus y) := a_n(x) + a_n(y),$$

and multiplication $\otimes$ by

$$a_n(x \otimes y) := \sum_{m \in \mathbf{Z}} a_m(x) a_{n-m}(y).$$

These operations are just like ordinary addition and multiplication except one never carries a digit. These operations turn $\mathbf{R}^+$ into a field[1] of characteristic two i.e. $x \oplus x = 0$. (This means in particular that $x \ominus y = x \oplus y$, which has the pleasant effect that it is impossible to make sign errors.) The field is more "discrete" than $\mathbf{R}$ and takes some getting used to (e.g. the interval $[0, 1]$ is closed under addition!). We give $\mathbf{R}^+$ the usual Lebesgue measure; this is invariant under Walsh translations

[1] To see the invertibility of $\otimes$, first observe that $(1 \oplus x)^{-1} = 1 \oplus x \oplus (x \otimes x) \oplus \dots$ when $|x| < 1$, and then generalize by scaling.

$x \mapsto x \oplus x_0$, while for Walsh dilations we have

$$\int_{\mathbf{R}^+} f(k \otimes x)\ dx = \frac{1}{\|k\|} \int_{\mathbf{R}^+} f(x)\ dx. \tag{1}$$

(see Exercises). All integrals will be on $\mathbf{R}^+$ unless otherwise specified.

Note that $\mathbf{R}^+$ is a metric space under the metric $d(x,y) := \|x \oplus y\|$. This metric is not quite equivalent to the usual metric $|x-y|$ (for instance, $1-\varepsilon$ and $1+\varepsilon$ are a distance 1 apart for any $0 < \varepsilon < 1$) but is very similar.

On $\mathbf{R}$ one uses Schwartz functions or C_0^∞ functions as the test functions to work on. The analogue in the Walsh model are *dyadic step functions*, defined as finite linear combinations of characteristic functions χ_I, where the I are dyadic intervals in $\mathbf{R}^+$.

The Fourier analysis of $\mathbf{R}$ is fundamentally driven by the plane waves $\exp(2\pi i x\xi)$. The Walsh analogue of this is the dyadic plane waves $e(x \otimes \xi)$, where $e : \mathbf{R}^+ \to \{-1,+1\}$ is defined by $e(x) := +1$ when $a_{-1}(x) = 0$ and $e(x) := -1$ when $a_{-1}(x) = 1$. Thus the graph of $e(x)$ is just a square wave with period 1 - it's the Walsh analogue of $\exp(2\pi i x)$. One minor pleasantness in the Walsh case is that e is always real, thus avoiding the need for complex numbers. Also, note that $e()$ is multiplicative in the sense that $e(x \oplus y) = e(x)e(y)$. We can then define the *Walsh-Fourier transform*, defined on dyadic step functions f by

$$\hat{f}(\xi) := \int e(x \otimes \xi) f(x)\ dx.$$

The Walsh-Fourier transform maps dyadic step functions to dyadic step functions, and its own inverse. Note that the function $\chi_{[0,1]}$ is its own Walsh-Fourier transform. (We delegate this to the exercises). This function plays the same role in Walsh-Fourier analysis that the Gaussian $e^{-\pi x^2}$ plays in Fourier analysis. Note that unlike the Gaussian, we have *perfect* localization in both space and frequency simultaneously (which we know to be impossible on $\mathbf{R}$).

By scaling and translating the Walsh Gaussian $\chi_{[0,1]}$ we obtain the handy formula

$$\hat{\chi}_I(\xi) := |I| \chi_{[0,|I|^{-1}]}(\xi) e(x_I \xi) \tag{2}$$

for all dyadic intervals I, where we use x_I to denote an arbitrary element of I.

The Walsh-Fourier transform has basically the same properties as the Fourier transform: translation in space corresponds to modulation by dyadic plane waves on the frequency side, and vice versa; scaling by k in physical space corresponds to scaling by k^{-1} in frequency space; multiplication goes to Walsh convolution and vice versa; and we have the Parseval and Plancherel theorems. The proofs are identical to the corresponding statements in the Fourier case.

Now we look at what frequency projections look like. Specifically, we want to understand the frequency projections π_ω for dyadic intervals[2] ω, defined by

$$\widehat{\pi_\omega f}(\xi) := \chi_\omega(\xi)\hat{f}(\xi).$$

Let's first understand this for the case $\omega = [0,1]$. In this case we have

$$\pi_{[0,1]} f(x) = f * \hat{\chi}_{[0,1]}(x) = f * \chi_{[0,1]}(x) = \int f(y)\chi_{[0,1]}(x \oplus y)\ dy.$$

If x lies in the interval $(n, n+1)$, we thus have

$$\pi_{[0,1]} f(x) = \int_n^{n+1} f(y)\ dy,$$

More generally, if $k \in \mathbf{Z}$ and x lies in the dyadic interval $(2^{-k}n, 2^{-k}(n+1))$, then

$$\pi_{[0,2^k]} f(x) = 2^k \int_{2^{-k}n}^{2^{-k}(n+1)} f(y)\ dy.$$

Thus $\pi_{[0,2^k]}$ is an averaging operator at spatial scale 2^{-k}. (Compare this with the Fourier situation). Shifting this by some frequency ξ, we see that

$$\pi_{[\xi, \xi\oplus 2^{k+1}]} f(x) = 2^k e(\xi \otimes x) \int_{2^{-k}n}^{2^{-k}(n+1)} e(\xi \otimes y) f(y)\ dy.$$

In particular, we have

$$\pi_{[2^k, 2^{k+1}]} f(x) = 2^k e(2^k \otimes x) \int_{2^{-k}n}^{2^{-k}(n+1)} e(2^k \otimes y) f(y)\ dy.$$

To understand this better, let's have some notation. For any dyadic interval I, define the *parent* $2I$ of this interval to be the unique dyadic interval containing I of twice the length. The interval $2I$ is thus the union of I and another interval I^*, which we call the *conjugate* of I. We then have

$$\pi_{[2^k, 2^{k+1}]} f(x) = \frac{1}{2}\Big(\frac{1}{|I|}\int_I f - \frac{1}{|I^*|}\int_{I^*} f\Big),$$

where I is the unique dyadic interval of length 2^{-k-1} which contains x. This operator is the analogue of the Littlewood-Paley projection P_k - note how it captures precisely the mean and averaging properties we were heuristically deducing in the Week 2 notes (averaging at scale 2^{-k-1}, but mean zero at scale 2^{-k}).

Let's see how the frequency cutoffs π_ω interact with spatial cutoffs χ_I. From our formula for $\pi_{[0,1]}$ we see that $\pi_{[0,1]}$ commutes with χ_I whenever the dyadic interval I has length at least 1. More generally (by scaling and modulation) we see that π_ω commutes with χ_I whenever $I \times \omega$ has area at least 1. This fact is extremely pleasant, and is not quite present in the Fourier case (any projection in frequency space must cause leakage in physical space, and vice versa, so there is no way physical and Fourier cutoffs can commute). We can thus define *phase space projections*

$$\pi_{I\times\omega} := \pi_\omega \chi_I = \chi_I \pi_\omega$$

whenever $I \times \omega$ respects the Heisenberg uncertainty principle (i.e. has area at least 1). Observe that this really is a projection in that it is self-adjoint and idempotent

[2] Conventionally, intervals in physical space are given names like I, while intervals in frequency space are given names like ω.

(as opposed to the Fourier case, in which all smoothed out projections are only projections in a moral sense).

Consider the projection $\pi_{[0,1]\times[0,1]}$. From our formula for $\pi_{[0,1]}$ we see that

$$\pi_{[0,1]\times[0,1]} f := (\int_{[0,1]} f)\chi_{[0,1]},$$

i.e. $\pi_{[0,1]\times[0,1]}$ is the orthogonal rank 1 projection (in $L^2(\mathbf{R}^+)$ to the line spanned by the dyadic Gaussian $\chi_{[0,1]}$. More generally, if $P = I \times \omega$ is a *Heisenberg tile*, i.e. a dyadic rectangle of area exactly 1, then π_P is the orthogonal rank 1 projection to the line spanned by the *Walsh wave packet* (or *Walsh function*)

$$\phi_P(x) := |I|^{-1/2} e(\xi_\omega x)\chi_I(x).$$

(In principle, the choice of ξ_ω could affect the definition of ϕ_P, but one can check that this only affects ϕ_P by a sign, which turns out to always be irrelevant). In other words, since we have defined ϕ_P to be L^2-normalized, we have

$$\pi_P f = \langle f, \phi_P \rangle \phi_P.$$

(This result can be deduced from the $P = [0,1] \times [0,1]$ case by translation, modulation, and scaling).

Since $\pi_{I\times\omega}$ is a rank 1 projection, we see that there is only a one-dimensional family of functions which have spatial support in I and frequency support in ω, namely the constant multiples of ϕ_P.

It is instructive to compare some examples of wave packets ϕ_P, when P is some integer shift of the unit square $[0,1]^2$, with their Fourier counterparts, namely the Gaussian translated and modulated by integer amounts. One is made of square waves and the other of sine waves, but otherwise they are very similar.

If $P = I \times \omega$ is a tile, we write I_P for the spatial interval I and ω_P for the frequency interval ω. Tiles have the nice property that given any two tiles which intersect, one of them is wider than the other and the spatial interval of the narrower one is contained in that of the wider one. Conversely, the frequency interval of the wider one is contained in that of the narrower one (so the narrower interval is taller). More succinctly, if we define $P < P'$ as the statement that P and P' intersect and that $I_P \subset I_{P'}$, then $<$ is a partial order and one has $P < P'$ or $P' < P$ if and only if $P \cap P' \neq \emptyset$.

Observe from (2) that the Walsh-Fourier transform of a Walsh wave packet is another Walsh wave packet, up to a sign:

$$\hat{\phi}_{I\times\omega} = \pm\phi_{\omega\times I}.$$

Thus the Walsh-Fourier transform intertwines $\pi_{I\times\omega}$ and $\pi_{\omega\times I}$. As a consequence of this and Plancherel we see that $\phi_{I\times\omega}$ and $\phi_{I'\times\omega'}$ are orthogonal whenever I and I' are disjoint, or ω and ω' are disjoint. In other words,

$$\langle \phi_P, \phi_{P'} \rangle = 0 \text{ whenever } P \cap P' = \emptyset.$$

In other words, the projections π_P and $\pi_{P'}$ are orthogonal to each other whenever $P \cap P' = \emptyset$.

This suggests putting together several disjoint tiles (which correspond to rank 1 projections) to create higher rank projections. To make this formal, we first observe from our formulae for π_P that

$$\pi_P f(x) = |I|^{-1}\chi_I(x)e(\xi_\omega x)\int f(y)\chi_I(y)e(\xi_\omega y)\ dy.$$

We use Plancherel:

$$\pi_P f(x) = |I|^{-1}\chi_I(x)e(\xi_\omega x)\int \hat{f}(\xi)\hat{\chi}_I(\xi \oplus \xi_\omega)\ d\xi.$$

Now we use (2):

$$\pi_P f(x) = \chi_I(x)e(\xi_\omega x)\int \hat{f}(\xi)\chi_\omega(\xi)e((\xi \oplus \xi_\omega) \otimes x_I)\ d\xi.$$

Observe that

$$(x \oplus x_I) \otimes (\xi \oplus xi_\omega) = 0$$

for all $(x, \xi) \in I \times \omega$. We can thus rewrite this formula as

$$\pi_P f(x) = \int \chi_P(x, \xi)e(x, \xi)\hat{f}(\xi)\ d\xi.$$

Because of this, it is tempting to define the phase space projection

$$\pi_\Omega f(x) := \int \chi_\Omega(x, \xi)e(x, \xi)\hat{f}(\xi)\ d\xi$$

for arbitrary Ω. It turns out this is only useful for Ω which *respect the uncertainty principle*, in the sense that they can be written as the union of disjoint tiles. Clearly, for such Ω we have

$$\pi_\Omega = \sum_{P \in \mathbf{P}} \pi_P$$

whenever $\mathbf{P}$ is a collection of disjoint tiles which tile Ω; since each tile has area 1, the cardinality of $\mathbf{P}$ must equal the area $|\Omega|$ of Ω. From this and orthogonality of the π_P we see that π_Ω is a rank $|\Omega|$ projection. It is also clear that any function of the form π_Ω has spatial support in $\{x : (x, \xi) \in \Omega \text{ for some } \xi\}$ and frequency support in $\{\xi : (x, \xi) \in \Omega \text{ for some } x\}$. (Virtually none of these properties are retained if Ω does not respect the uncertainty principle).

For future reference we shall need

Lemma 1.1. *If Ω respects the uncertainty principle, and P is a tile which is strictly contained inside Ω, then $\Omega - P$ also respects the uncertainty principle.*

Proof By definition, Ω is the union of a disjoint collection of tiles $Q_1, \ldots, Q_N$. If one of these Q_i were equal to P, we would be done. Otherwise, the idea is to shuffle the Q around until one of them does equal P.

We may throw away all the Q_j which don't intersect P (and shrink Ω accordingly), since they can always be added back later. Now observe it is impossible for one of the Q_j to be wider than P and another to be taller than P, otherwise we would contradict disjointness. By symmetry we may thus assume that all the Q_j are wider than P.

Let Q_{wide} be a tile Q_j with maximal width. Consider the tile Q^*_{wide}, defined as the tile with the same spatial interval as Q_{wide} but the conjugate frequency interval. This tile also intersects P, and thus must encounter some Q_j. Q_j cannot be taller than Q_{wide}, else it would intersect Q_{wide}, and it cannot be wider, by definition of Q_{wide}. Thus it must be Q^*_{wide}. In other words, Q_{wide} and Q^*_{wide} are both in the collection of Q_j. We can then replace $Q_{wide} \cup Q^*_{wide}$ by the union of two tiles of half the width and twice the height. One of these does not intersect P and can thus be thrown away. Continuing in this way we may remove more and more tiles from the Q_j until eventually one of them is equal to P, at which point we are done. ∎

Observe that $\pi_{I\times\mathbf{R}^+}$ corresponds to the spatial projection χ_I. To see this, first see that it is true for any Walsh wave packet on a tile P in $I \times \mathbf{R}^+$ (since one can use P to begin a tiling of $I \times \mathbf{R}^+$). From this it is clear that it is true for any function on I, and outside I it is clear that the projection vanishes. Similarly $\pi_{\mathbf{R}^+\times\omega}$ equals π_ω.

Sometimes it is convenient to pass from the Walsh continuum to a discrete analogue. Suppose that $0 < a < A$ are integers, and a dyadic step function f is supported on an interval $[0, 2^A]$, but is constant on dyadic intervals of length 2^a. Thus 2^a is the smallest interesting scale and 2^A is the largest interesting scale. Then f is a convolution of a function with $\chi_{[0,2^a]}$, and so its Fourier support is contained in that of $\chi_{[0,2^a]}$, i.e. $\hat{f}$ is supported on $[0, 2^{-a}]$. Conversely, f is a product of a function with $\chi_{[0,2^A]}$, hence $\hat{f}$ is a convolution of a function with $2^A\chi_{[0,2^{-A}]}$. Thus $\hat{f}$ is constant on dyadic intervals of length 2^{-A}. Thus on the frequency side, 2^{-A} is the smallest interesting scale and 2^{-a} is the largest.

We can then discretize physical space to the discrete interval (which is a subgroup of $\mathbf{R}^+$)

$$I_{a,A} := 2^a\mathbf{Z} \cap [0, 2^A)$$

and frequency space to the dual interval $I_{-A,-a}$, so that phase space is discretized to $I_{a,A} \times I_{-A,-a}$. One can give these intervals normalized counting measure

$$\mu_{a,A}(E) := 2^{-a}\#(E)$$

so that phase space has measure $\mu_{a,A} \otimes \mu_{-A,-a}$. One can then define the Walsh-Fourier transform on these discrete spaces as usual, and one has analogues of all the previous statements.

For instance, when $a = -1$, $A = 0$, then functions only have two values, $f(0)$ and $f(.1)$ (here 0.1 is 1/2, not 1/10!), and the Fourier transform only has two values, $\hat{f}(0) = (f(0) + f(.1))/2$ and $\hat{f}(1) = (f(0) - f(.1))/2$. Phase space consists of four points, and there are four tiles one can play with.

We won't use discretized models much in this course, but they are handy for building up intuition.

Let's return to the phase space projections π_Ω, defined for sets Ω which respect the uncertainty principle. Observe that $\pi_\Omega f$ and $\pi_{\Omega'} f$ are orthogonal whenever Ω, Ω'

respect the uncertainty principle and are disjoint. In particular we have

$$\|\pi_{\Omega\cup\Omega'} f\|_2^2 = \|\pi_\Omega f\|_2^2 + \|\pi_{\Omega'} f\|_2^2. \tag{3}$$

Also, we clearly have

$$\|\|\pi_\Omega f\|_2^2 \geq 0.$$

Thus the quantity $\|\|\pi_\Omega f\|_2^2$ is behaving like some sort of non-negative measure on the phase plane. It's almost as if there was some function $W(f,f)(x,\xi)$ on phase space for which

$$W(f,f)(x,\xi) \geq 0 \text{ for all } x,\xi \tag{4}$$

and

$$\|\|\pi_\Omega f\|_2^2 = \int_\Omega W(f,f) \text{ for all } \Omega \text{ which respects the uncertainty principle.} \tag{5}$$

We are writing $W(f,f)$ because the expression is clearly quadratic in f. To make this rigorous, it would be nice if $W(f,f)$ was just a special case of some $W(f,g)$ such that

$$W(f,g) \text{ is linear with respect to } f \text{ or } g. \tag{6}$$

Also, by taking $\Omega = I \times \mathbf{R}$ in (5) and letting I shrink to a point we see that

$$|f(x)|^2 = \int W(f,f)(x,\xi)\ d\xi \tag{7}$$

and similarly

$$|\hat{f}(\xi)|^2 = \int W(f,f)(x,\xi)\ dx \tag{8}$$

thus $|\hat{f}|^2$ and $|f|^2$ are the marginal distributions of $W(f,f)$. Combining this with (4), we see that

$$f \text{ has physical support in } I \implies W(f,f) \text{ has support in } I \times \mathbf{R}^+ \tag{9}$$

and

$$f \text{ has frequency support in } \omega \implies W(f,f) \text{ has support in } \mathbf{R}^+ \times \omega \tag{10}$$

It would also be nice if translating f in physical space corresponded to translating $W(f,f)$ in the x variable, and similarly for frequency space and the ξ variable, that scaling f by k in physical space in an L^2-preserving manner should just replace $W(f,f)(x,\xi)$ by $W(f,f)(k^{-1}\otimes x, k\otimes\xi)$. If all these properties held, then $W(f,f)$ would be a realization of the ideal we discussed in the Week 1 notes, namely a *phase space portrait* of f. $W(f,f)$ would somehow be the probability density of f at (x,ξ), measuring how much of f had a position at x and frequency at ξ. (The functions $|f(x)|^2$ and $|f(\xi)|^2$ only do this for one of the two variables).

It turns out that such W do not actually exist, but one can find candidates which almost obey all of the above properties.

For instance, the function

$$W(f,g)(x,\xi) := f(x)g(\xi)$$

obeys all the properties desired except for the non-negativity property (4). The only one which is remotely tricky is (5). By (3) it suffices to do this when Ω is a tile; by translation, modulation, and scaling we can set the tile to be $[0,1]^2$. Then we have

$$\int_{[0,1]^2} W(f,f) = \int_0^1 \int_0^1 f(x)\hat f(\xi)\, dx d\xi = \langle f, \chi_{[0,1]}\rangle \langle \hat f, \chi_{[0,1]}\rangle = |\int_{[0,1]} f|^2 = \|\pi_{[0,1]^2} f\|_2^2$$

as desired.

Example: In the discrete model $a = -1$, $A = +1$ mentioned earlier, we have $W(0,0) = f(0)(f(0)+f(.1))/2, W(.1,0) = f(1)(f(0)+f(.1))/2, W(0,1) = f(0)(f(0)-f(.1))/2$, and $W(.1,1) = f(1)(f(0)-f(.1))/2$. Note that W is not always positive, but its average on any set which respects the uncertainty principle is positive. This is a good example of things being weird at the quantum level but seemingly okay at the classical level.

There are many other choices of $W(f,g)$ which would work. For instance, any expression of the form

$$W(f,g)(x,\xi) := \int f(x \oplus (a \otimes t)) g(x \oplus (b \otimes t)) e(\xi \otimes t)\ dt$$

will work, if a, b are numbers such that $a \oplus b = 1$. (Note that $a = 0, b = 1$ corresponds to the form $f(x)g(\xi)$ used earlier). Bilinear expressions of this type are known as *Walsh-Wigner transforms*. The Fourier analogue is

$$W(f,g)(x,\xi) := \int f(x+at)g(x-bt)e^{-2\pi i \xi t}\ dt$$

where $a + b = 1$; when $a = b = 1/2$, this is known as the *Wigner transform* of f and g.

Now let's see why one cannot find such a W with all the properties required (including (4)). Since $W(f,g)(0,0)$ is bilinear, it must have the form

$$W(f,g)(0,0) = \int\int f(x)g(y)K(x,y)\ dxdy$$

for some kernel K, possibly distributional. From (9) and polarization we see that $W(f,g)$ vanishes whenever f, g are supported on a dyadic interval I not containing the origin. In particular K vanishes on $I \times I$ in this case, which implies that $K(x,x) = 0$ for all $x \neq 0$. On the other hand, the positivity condition (4) enforces (morally at least) the condition $K(x,y)^2 \leq K(x,x)K(y,y)$ (i.e. every two-by-two minor of K must be non-negative). This implies that K vanishes everywhere except at the origin, which means that $W(f,g)(0,0) = cf(0)g(0)$ for some constant c. But this is easily seen to disobey (10).

On the other hand, if one gives up the linearity condition (6), then one can grab (4):

Proposition 1.2. *Let f be a dyadic step function. Then there exists a non-negative function $W(f,f)$ on the phase plane which obeys (5).*

One can also make this function obey all the other properties required (other than linearity), but we shall not try to do so here.

Proof We are looking for a non-negative function $W(f, f)$ such that

$$\langle W(f, f), \chi_P \rangle = \|\pi_P f\|_2^2$$

for all tiles P (the claim (5) then follows by (3)).

Let's first do some linear programming[3]. Consider the set $\Sigma \subset L^2(\mathbf{R}^+)$ as the affine span of the functions $-\chi_P / \|\pi_P f\|_2^2$; this has the form $v_0 + V$ for some vector space $V \subset L^2(\mathbf{R}^+)$ and $v_0 in L^2(\mathbf{R}^+)$. Also, consider the positive cone Γ defined by

$$\Gamma := \{f \in L^2(\mathbf{R}^+) : f(x) \geq 0 \text{ for all } x \in \mathbf{R}^+.\}$$

Clearly we must have $\langle W(f, f), F \rangle = -1$ for all F in the affine space Σ. Thus if Σ intersects Γ, no such $W(f, f)$ is possible. On the other hand, if Σ is disjoint from Γ, then by repeating the standard proof of the Hahn-Banach theorem we can find an affine hyperspace containing Σ which is disjoint from Γ. This hyperspace does not go through the origin, and so has the form $\{F \in L^2 : \langle W(f, f), F \rangle = -1\}$ for some $W(f, f) \in L^2$. This function obeys the desired properties.

This it remains to show that Σ is disjoint from Γ. Suppose for contradiction that this were not the case, then some non-zero affine combination of $-\chi_P / \|\pi_P f\|_2^2$ gives a non-negative function. Since f is a dyadic step function we can discretize, and assume that this convex combination is a finite one. This is the same as saying that we can find a finite collection $P_1, \ldots, P_N$ of tiles and real numbers $a_1, \ldots, a_N$ such that $\sum_{i=1}^N a_i \chi_{P_i}$ is non-negative, while $\sum_{i=1}^N a_i \|\pi_{P_i} f\|_2^2$ is negative. By adding a positive multiple of some χ_P we may assume that $\sum_{i=1}^N a_i \chi_{P_i}$ is not identically zero and $\sum_{i=1}^N a_i \|\pi_{P_i} f\|_2^2 = 0$. Equivalently - separating a_i into positive and negative numbers, and discarding any zeroes - we may find tiles $P_1, \ldots, P_N, Q_1, \ldots, Q_M$ and positive constants $a_1, \ldots, a_N$ and $b_1, \ldots, b_M$ such that we have the pointwise estimate

$$\sum_{i=1}^N a_i \chi_{P_i} \geq \sum_{j=1}^M b_j \chi_{Q_j} \tag{11}$$

as well as the bound

$$\sum_{i=1}^N a_i \|\pi_{P_i} f\|_2^2 = -\sum_{i=1}^M b_i \|\pi_{Q_j} f\|_2^2.$$

Integrating (11) we have (since tiles have area 1)

$$\sum_{i=1}^N a_i > \sum_{j=1}^M b_j.$$

[3] Specifically, we are deriving *Farkas's lemma*, which states that the only obstructions to solving a linear programming problem occur when one can derive $0 > 0$ from non-negative linear combinations of the constraints.

Since neither side can be zero, we may thus normalize

$$\sum_{i=1}^{N} a_i > \sum_{j=1}^{M} b_j + 1.$$

Suppose that of all possible such P_i, a_i, Q_j, b_j, we pick an example which minimizes $\sum_{j=1}^{M} b_j$. This means in particular that $P_i \neq Q_j$ for all i, j, since one could otherwise decrement a_i and b_j by the same factor.

More generally, by Lemma 1.1 we can assume that $Q_j \not\subset P_{i_1} \cup \ldots \cup P_{i_k}$ for any disjoint tiles $P_{i_1}, \ldots, P_{i_k}$. For if this were the case, we could use Lemma 1.1 to find disjoint tiles $P'_1, \ldots, P'_{k-1}, Q_j$ which cover the same region of phase space as $P_{i_1}, \ldots, P_{i_k}$. By transferring an epsilon of the a_{i_s} to these new tiles, one can then cancel out some of the Q_j, again contradicting minimality.

Let Q_{wide} be a tile Q_j which maximizes the length of the spatial interval I_{Q_j}. Pick any x in $I_{Q_{wide}}$, and let ξ vary in $\omega_{Q_{wide}}$. By (11), (x, ξ) must be contained in some P_{i_ξ}, which is distinct from Q_{wide}. There are two cases: either there exists a ξ such that P_{i_ξ} is taller than Q_{wide}, or that P_{i_ξ} is wider than Q_{wide} for all ξ.

Consider the latter case. Then we can cover Q_{wide} by a collection $\mathbf{P}$ of the tiles in $\{P_1, \ldots, P_n\}$, each of which are wider than Q_{wide}. By passing to the tiles in $\mathbf{P}$ whose frequency intervals are maximal with respect to set inclusion, we may assume that the tiles in $\mathbf{P}$ are disjoint. But this contradicts the conclusion of two paragraphs ago.

In the former case, we see that for every $x \in I_{Q_{wide}}$ the interval $x \times \omega_{Q_{wide}}$ is contained in some P_i. Now we may repeat the previous argument (but for height instead of width) to again get a contradiction. ■

2. A Walsh version of Carleson's theorem

Recall the Fourier inversion formula

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi) e^{-2\pi i x \xi}\, d\xi$$

(where $\hat{f}$ temporarily denotes the Fourier transform once again). In particular, we have (formally at least)

$$f(x) = \lim_{N \to \infty} \int_{-\infty}^{N} \hat{f}(\xi) e^{-2\pi i x \xi}\, d\xi.$$

Thus, if we let P_N denote the Fourier multiplier corresponding to $\chi_{(-\infty, A]}$, we have

$$f(x) = \lim_{N \to \infty} P_N f(x).$$

If f is Schwartz, then this formula is valid pointwise in x (indeed the convergence is uniform). However, it is not as clear what is going on for rough functions f.

Plancherel's theorem combined with monotone convergence tells us that $P_N f$ converges to f in L^2 norm, but this does not imply pointwise convergence. A possible difficulty is indicated by *Gibbs phenomenon*: if f has a jump discontinuity (for instance), then $P_N f$ will oscillate around that discontinuity by a sizeable amount, and thus does not converge uniformly (although it still converges pointwise a.e.)

Indeed, examples due to Kolmogorov and others show that if f is merely in L^1, then $P_N f$ can diverge a.e. On the other hand, in 1966 Lennart Carleson (currently here at UCLA!) proved (in a famously difficult paper):

Theorem 2.1. *If f is in L^2, then $P_N f$ converges a.e. to f as $N \to \infty$.*

As is usual in this business, pointwise convergence is obtained by showing a bound on the *maximal function*

$$\mathcal{C}f(x) := \sup_N |P_N f(x)|. \tag{12}$$

Suppose in particular that one was able to show the weak-type bound

$$\|\mathcal{C}f\|_{L^{2,\infty}} \lesssim \|f\|_2. \tag{13}$$

Then, for any f in L^2 and any $\varepsilon > 0$, we could approximate f by a smooth function g and an error of L^2 norm $O(\varepsilon)$. From the above estimate we see that $P_N(f)$ will not fluctuate very much from $P_N(g)$ outside of a set of small measure. By the usual methods one can then demonstrate pointwise convergence. (In fact this is reversible: if the above weak-type estimate failed, then one could not obtain pointwise convergence. This is known as *Stein's maximal principle*, which we will not discuss here).

We will not quite prove Carleson's theorem here, as it is a little technical. However, we will prove the Walsh analogue of Carleson's theorem, which already contains all the major ideas of the proof. We redefine P_N for $N > 0$ to be the *Walsh* multiplier corresponding to $[0, N]$ (acting on functions on $\mathbf{R}^+$), and define the Walsh-Carleson operator $\mathcal{C}f(x)$ by (12). We shall show the bound (13) for this operator.

We follow the arguments of Lacey and Thiele in a short paper last year; this was inspired by a brilliant 1973 paper of Fefferman, and also by their earlier work on the bilinear Hilbert transform in 1998. The techniques here have since been applied to a number of very delicate multi-linear operators and maximal operators, and is still an active area of research (especially at UCLA!)

We begin with a couple standard reductions. The first is *linearization* - we replace the sub-linear operator $\mathcal{C}$ by the linear variant

$$P_{N(x)}f(x),$$

where $N : \mathbf{R}^+ \to \mathbf{R}^+$ is an arbitrary measurable function. Clearly $P_{N(x)}f(x)$ is dominated by $\mathcal{C}$; conversely, one can choose an $N(x)$ so that $P_{N(x)}f(x)$ is comparable to $\mathcal{C}f(x)$ in magnitude everywhere. Thus it suffices to prove that

$$\|P_{N(x)}f(x)\|_{L^{2,\infty}} \lesssim \|f\|_2$$

for some arbitrary function N, which we now fix. (Of course, N is likely to be extremely rough). For technical reasons we shall assume that N is never a terminating binary.

We fix f to be some dyadic step function, which we normalize to have L^2 norm 1. We have to show that

$$|\{x : |P_{N(x)} f(x)| \geq \lambda\}| \lesssim \lambda^{-2}.$$

It suffices to show that

$$\int_E |P_{N(x)} f(x)| \lesssim |E|^{1/2}$$

for all sets $E \subset \mathbf{R}^+$ of finite measure, since the claim then follows by setting $E := \{x : |P_{N(x)} f(x)| \geq \lambda\}$. Clearly, it suffices to show that

$$|\int_E P_{N(x)} f(x)| \lesssim |E|^{1/2}$$

for all E (split E into sets where $P_{N(x)} f(x)$ is positive or negative). We now fix E.

So far we have only taken very cautious, reversible steps. Now let's do more interesting operations. We would like to understand the projections P_N in terms of the Walsh projections of the previous section. If we can decompose the interval $[0, N]$ as the disjoint union of dyadic intervals ω then we would have $P_N = \sum_\omega \pi_\omega$. Of course, there are several ways to do this, but we would like to be as efficient as possible and use as few intervals as we can.

Call an interval ω *left-handed* if it is the left-half of its parent interval 2ω. Let $\mathcal{L}$ denote the collection of left-handed intervals. Observe that $[0, N]$ is the disjoint union of those intervals $\omega \in \mathcal{L}$ for which $N \in \omega^*$. (This is the partition obtained by the greedy algorithm, and is the maximal one). Thus

$$P_{N(x)} f(x) = \sum_{\omega \in \mathcal{L}} \chi_{\omega^*}(N(x)) \pi_\omega f(x).$$

The projection π_ω can be written as $\pi_{\mathbf{R}^+ \times \omega}$, which can be split as $\sum_{P : \omega_P = \omega} \pi_P$, the sum of all the tile projections of tiles P whose frequency interval is ω. However $\pi_P f(x)$ vanishes unless $x \in I_P$. Thus we can write

$$P_{N(x)} f(x) = \sum_{\omega \in \mathcal{L}} \sum_{P : \omega_P = \omega} \chi_{\omega^*}(N(x)) \chi_{I_P}(x) \pi_P f(x).$$

Call a tile *left-handed* if its frequency interval is, and call $\mathbf{P}_0$ the set of left-handed tiles. If $P = I \times \omega$ is a tile, define the conjugate tile P^* by $P^* := I \times \omega^*$. (Thus if P is left-handed, then P^* lies directly atop of P). We thus have

$$P_{N(x)} f(x) = \sum_{P \in \mathbf{P}_0} \chi_{P^*}(x, N(x)) \pi_P f(x) = \sum_{P \in \mathbf{P}_0} \chi_{P^*}(x, N(x)) \langle f, \phi_P \rangle \phi_P(x).$$

Thus each tile P can potentially contribute something to $P_{N(x)} f(x)$, but only if the point $(x, N(x))$ is captured by the conjugate tile P^*. Integrating on E, we thus reduce to showing that

$$|\sum_{P \in \mathbf{P}_0} a_P b_P| \lesssim |E|^{1/2} \tag{14}$$

where a_P, b_P are the numbers

$$a_P := \int_{E_P} \phi_P(x)\, dx$$

$$b_P := \langle f, \phi_P \rangle$$

and E_P is the set

$$E_P := \{x \in E : (x, N(x)) \in P^*\} = \{x \in E \cap I_P : N(x) \in \omega_P^*\}.$$

Notice how we have managed to decouple the function f from the set E and the function N. Initially, f, E, and N were all interacting with each other in a rather non-linear manner. Now, f only interacts with P, while E and N^* only interact with P^*, while P and P^* of course are highly correlated with each other. One can think of this as a kind of "factorization" of the operator.

As we shall see, the relationship between P and P^* will be crucial - and more specifically, the fact that these two tiles are close together, but not identical. A key observation is the following: if P and Q are two distinct left-handed tiles which overlap, then P^* and Q^* will *not* overlap. Equivalently, if P^* and Q^* overlap, then P and Q cannot. So we have disjointness in one or the other - but to begin with, we don't know which type of disjointness we have. The strategy will be to somehow separate the two types of disjointness.

To estimate $\sum_{P \in \mathbf{P}_0} a_P b_P$ we must obtain good bounds on the a_P and b_P separately. However, the problem is that even after we do this there are many, many tiles P (a three-parameter set) to sum, with much overlap (both in the Ps and in the P^*s). We will need to sort these tiles into more manageable clumps. In some clumps the Ps will be disjoint, while in others the P^*s will be disjoint.

Since f is normalized in L^2, and the ϕ_P are normalized, we see that $|b_P| \leq 1$ for all P. Actually, we can do a lot better: from Bessel's inequality and orthogonality we have

$$(\sum_{P \in \mathbf{D}} |b_P|^2)^{1/2} \leq 1 \tag{15}$$

whenever $\mathbf{D}$ is a disjoint collection of tiles in $\mathbf{P}_0$. This is the only bound we will use on the b_P.

What about the a_P? Let's first consider estimating a single a_P. Since ϕ_P has size $|I_P|^{-1/2}$ and is supported on I_P, we have the crude bound

$$|a_P| \leq |I_P|^{-1/2} |E_P|. \tag{16}$$

Since $|E_P| \leq |E|, |I_P|$ we thus have

$$|a_P| \leq \min(|I_P|^{-1/2}|E|, |I_P|^{1/2}) \leq |E|^{1/2}. \tag{17}$$

Thus we can control a single factor $a_P b_P$ by $O(|E|^{1/2})$ easily.

How could we boost this to sums of $a_P b_P$? The estimate (15) helps, but is clearly not enough by itself. To build some intuition, let's first try summing $\sum_{P \in \mathbf{D}} a_P b_P$, where $\mathbf{D}$ ranges over a collection of disjoint tiles. Expanding this out we have

$$\sum_{P \in \mathbf{D}} \int_{E_P} b_P \phi_P.$$

Since E_P is smaller than E, we expect this expression to morally be smaller than

$$\sum_{P \in \mathbf{D}} \int_E b_P \phi_P,$$

though this is not rigorous because of the possibility of cancellation. Interchanging the sum and integral and applying Cauchy-Schwartz, we bound this by

$$|E|^{1/2} \| \sum_{P \in \mathbf{D}} b_P \phi_P \|_2,$$

which is OK by Bessel's inequality and (15). Of course, we have to go back and deal with the E_P vs E issue, but morally at least we can cope with this type of sum.

Now suppose we have a sum $\sum_{P \in \mathbf{D}} a_P b_P$ in which the P are not necessarily disjoint, but their *conjugates* P^* are. Now we cannot exploit (15) so easily, and we must find other tricks. To simplify matters let us consider a model type of collection $\mathbf{D}$, namely the set

$$\mathbf{D} = \{P \in \mathbf{P}_0 : P \leq P_T\}$$

where the "top tile" P_T is some arbitrary tile. This type of collection is known as a *tree* (from its shape in the phase plane) and will play an important role in our arguments. (Actually this is an example of a *complete tree*; in our arguments we shall actually deal with much more "pruned" versions of this object).

To be consistent with (15), let us assume that

$$|b_P| \lesssim \left(\frac{|I_P|}{|I_{P_T}|}\right)^{1/2}. \tag{18}$$

(It turns out that this is not such an unreasonable assumption - we will eventually choose trees so that something like (18) holds for each tree). Thus we can bound

$$|\sum_{P \in \mathbf{D}} a_P b_P| \lesssim |I_{P_T}|^{-1/2} \sum_{P \in \mathbf{D}} |I_P|^{1/2} |a_P|.$$

Since ϕ_P has size $|I_P|^{-1/2}$, we can bound

$$|I_P|^{1/2} |a_P| \lesssim |E_P|.$$

Now the key observation is that because the P^* are disjoint, the E_P are also disjoint, and are contained in $E \cap I_{P_T}$. Thus we have

$$|\sum_{P \in \mathbf{D}} a_P b_P| \lesssim |I_{P_T}|^{-1/2} |E \cap I_{P_T}|.$$

Now we argue as in (17) to obtain

$$|\sum_{P \in \mathbf{D}} a_P b_P| \lesssim |I_{P_T}|^{-1/2} |E|^{1/2} |I_{P_T}|^{1/2} = |E|^{1/2}$$

as desired.

To summarize, we have managed to (morally) handle summations over certain types of tile sets $\mathbf{D}$ which are either disjoint, or are conjugate-disjoint. Of course, the entire tile set $\mathbf{P}$ has too much self-overlap to fall into either one of these categories. However, it turns out that one can split $\mathbf{P}$ into the countable union of "trees", each of which fall into one or the other of the above categories. At first this doesn't seem like enough, because one still gets a divergence when summing. However, as we shall see, in this sum, the trees slowly get "lighter" (in the sense that the ratio $|E_P|/|I_P|$ is no longer 1, but gets smaller and smaller), and less "energetic" (in the sense that estimates like (18) improve), and eventually we will be able to sum things in a geometric series.

Now for the technical stuff. From here onwards all tiles will be assumed to be left-handed.

Definition 2.2. Let P_T be a tile. A *1-tree with top* P_T is a collection T of tiles in $\mathbf{P}_0$ containing P_T such that $P \leq P_T$ for all $P \in T$. A *2-tree with top* P_T is a collection T of tiles containing P_T in $\mathbf{P}_0$ such that $P^* \leq P_T^*$ for all $P \in T$. We define a *tree with top* P_T to be any collection T of tiles in $\mathbf{P}_0$ containing P_T such that $P \leq P_T$ or $P^* \leq P_T^*$ for all $P \in T$.

Note that in a 1-tree, the conjugate tiles P^* are all disjoint, while in a 2-tree the P tiles are all disjoint. A tree is composed of a 1-tree and 2-tree, thus we have a chance of estimating an individual tree.

Definition 2.3. A collection $\mathbf{P}$ of tiles in $\mathbf{P}_0$ is *convex* if, whenever $P_1 < P_2 < P_3$ are tiles in $\mathbf{P}_0$ such that $P_1, P_3 \in \mathbf{P}$, then P_2 must also be in $\mathbf{P}$.

Convexity is a technical thing needed to make the "we can replace E_P by E" step rigorous in the above heuristic argument. Note that $\mathbf{P}_0$ is trivially convex.

Definition 2.4. For any collection $\mathbf{P}$ of tiles in $\mathbf{P}_0$, we define the *mass* $\mathrm{mass}(\mathbf{P})$ by

$$\mathrm{mass}(\mathbf{P}) := \sup_{P \in \mathbf{P}} \frac{|\tilde{E}_P|}{|I_P|}$$

where $\tilde{E}_P$ is a slight enlargement of E_P, defined by

$$\tilde{E}_P := \{x \in E : (x, N(x)) \in P \cup P^*\} = \{x \in E \cap I_P : N(x) \in \omega_P \cup \omega_P^*\}.$$

Clearly we have $\mathrm{mass}(\mathbf{P}) \leq 1$ for any $\mathbf{P}$, however for many of our trees the mass will be much smaller. The mass will control the size of the a_P. The corresponding quantity controlling the b_P is called the *energy*:

Definition 2.5. For any collection $\mathbf{P}$ of tiles in $\mathbf{P}_0$, we define the *energy* $\mathrm{energy}(\mathbf{P})$ by

$$\mathrm{energy}(\mathbf{P}) := \sup_{T \subset \mathbf{P}} \Big(\frac{1}{|I_{P_T}|} \sum_{P \in T} |b_P|^2\Big)^{1/2},$$

where T ranges over all 2-trees contained in $\mathbf{P}$.

Since the tiles in a 2-tree are disjoint, we expect to use (15) to get some control on *energy*$(\mathbf{P})$ (at least, we should be able to control the contribution of the big trees). Since every singleton tile is itself a 2-tree with itself as its top, we obtain the estimate

$$|b_P| \leq |I_P|^{1/2}\text{energy}(\mathbf{P}) \tag{19}$$

which is the analogue of (18).

Clearly, mass and energy are monotone in the sense that adding more tiles to a collection can only serve to increase the mass and energy.

There will be three key lemmas in the proof of (14). The first one shows how to bound the contribution of a single tree.

Lemma 2.6 (Tree estimate)**.** *If T is a convex tree, then*

$$|\sum_{P \in T} a_P b_P| \lesssim |I_{P_T}|\text{mass}(T)\text{energy}(T).$$

To use this lemma we need to split $\mathbf{P}_0$ up into trees, each of which has a relatively small mass and energy. This is the purpose of the other two lemmas:

Lemma 2.7 (Mass decomposition lemma)**.** *Let $\mathbf{P}$ be a convex collection of tiles. Then we can partition $\mathbf{P} = \bigcup_{T \in \mathbf{T}} T \cup \mathbf{P}'$ where $\mathbf{P}'$ is a convex collection of tiles with*

$$\text{mass}(\mathbf{P}') \leq \frac{1}{2}\text{mass}(\mathbf{P}),$$

and $\mathbf{T}$ is a collection of convex trees with

$$\sum_{T \in \mathbf{T}} |I_{P_T}| \lesssim |E|/\text{mass}(\mathbf{P}). \tag{20}$$

Lemma 2.8 (Energy decomposition lemma)**.** *Let $\mathbf{P}$ be a convex collection of tiles. Then we can partition $\mathbf{P} = \bigcup_{T \in \mathbf{T}} T \cup \mathbf{P}'$ where $\mathbf{P}'$ is a convex collection of tiles with*

$$\text{energy}(\mathbf{P}') \leq \frac{1}{2}\text{energy}(\mathbf{P}),$$

and $\mathbf{T}$ is a collection of convex trees with

$$\sum_{T \in \mathbf{T}} |I_{P_T}| \lesssim \text{energy}(\mathbf{P})^{-2}. \tag{21}$$

We will prove these lemmas in the next few sections. For now, let us see how they imply (14). Define the auxiliary quantity width$(\mathbf{P})$ by

$$\text{width}(\mathbf{P}) = \min(|E|/\text{mass}(\mathbf{P}), \text{energy}(\mathbf{P})^{-2}).$$

From the mass and energy decomposition lemmas we have

Corollary 2.9. *Let $\mathbf{P}$ be a convex collection of tiles. Then we can partition $\mathbf{P} = \bigcup_{T \in \mathbf{T}} T \cup \mathbf{P}'$ where $\mathbf{P}'$ is a convex collection of tiles with*

$$\text{width}(\mathbf{P}') \geq 2\text{width}(\mathbf{P}),$$

and $\mathbf{T}$ *is a collection of convex trees with*

$$\sum_{T \in \mathbf{T}} |I_{P_T}| \lesssim \mathrm{width}(\mathbf{P}). \tag{22}$$

Now we do the following algorithm. Initialize $\mathbf{P} = \mathbf{P}_0$. If $\mathbf{P}$ has zero mass or zero energy then clearly $\sum_P a_P b_P$ vanishes, and we are done. Otherwise, $\mathbf{P}$ has some finite width. We apply the above Corollary, and split $\sum_{P \in \mathbf{P}} a_P b_P$ into

$$\sum_{T \in \mathbf{T}} \sum_{P \in T} a_P b_P$$

and

$$\sum_{P \in \mathbf{P}'} a_P b_P.$$

By the tree estimate, the first expression is bounded by

$$\sum_{T \in \mathbf{T}} |I_{P_T}| \mathrm{mass}(\mathbf{P}) \mathrm{energy}(\mathbf{P}).$$

Using the bounds

$$\mathrm{energy}(\mathbf{P}) \lesssim \mathrm{width}(\mathbf{P})^{-1/2}; \quad \mathrm{mass}(\mathbf{P}) \lesssim \min(1, |E|/\mathrm{width}(\mathbf{P})$$

and (22), we can bound this by

$$\mathrm{width}(\mathbf{P}) \min(1, |E|/\mathrm{width}(\mathbf{P})) \mathrm{width}(\mathbf{P})^{-1/2}.$$

This quantity is bounded by $|E|^{1/2}$ but decays geometrically as $\mathrm{width}(\mathbf{P})$ gets much larger or much narrower than $|E|$. To deal with $\mathbf{P}'$, we replace $\mathbf{P}$ by $\mathbf{P}'$ and iterate the above algorithm (but now with a width which is at least twice as large). Summing this geometric series we get $|E|^{1/2}$ as desired.

3. Proof of the mass decomposition lemma

Fix $\mathbf{P}$, and call a tile $P \in \mathbf{P}$ *heavy* if $|\tilde{E}_P| \geq \frac{1}{2}\mathrm{mass}(\mathbf{P})|I_P|$. We need to eliminate all the heavy tiles from $\mathbf{P}$ in order to create $\mathbf{P}'$. The idea is to organize the heavy tiles into trees and use some sort of disjointness of the $\tilde{E}_P$ to obtain (20).

The algorithm is as follows. Initialize $\mathbf{T}$ to be empty, and $\mathbf{P}'$ to equal $\mathbf{P}$. If $\mathbf{P}'$ contains no heavy tiles, terminate the algorithm. Otherwise, pick a heavy tile P_T in $\mathbf{P}'$ with a maximal width $|I_{P_T}|$. Form the tree $\{P \in \mathbf{P}' : P^* \leq P_T^* \text{ or } P \leq P_T\}$. Remove this tree from $\mathbf{P}'$ (adding it to $\mathbf{T}$); note that this does not disturb the convexity of $\mathbf{P}'$. Continue the algorithm until $\mathbf{P}$ no longer contains heavy tiles.

Note that since $\mathbf{P}'$ is always convex, so are the trees T. It remains to show (20). First observe that the conjugate tops P_T^* of the trees we remove are disjoint; if one P_T^* intersects another, then the wider of the trees was selected first, and when that tree was removed from $\mathbf{P}'$, the smaller top should also have been removed at that time, a contradiction. Similarly the tops P_T are disjoint. The $\tilde{E}_{P_T}$ are now essentially disjoint subsets of E, and (20) easily follows.

4. Proof of the energy decomposition lemma

This is a similar argument to the mass lemma. Fix $\mathbf{P}$, and call a tree $T \in \mathbf{P}$ *energized* if

$$(\frac{1}{|I_{P_T}|} \sum_{P \in T : P^* \leq P_T^*} |b_P|^2)^{1/2} \geq \frac{1}{2}\text{energy}(\mathbf{P}).$$

Thus the 2-tree component of an energized tree has a large energy. We shall therefore removed all energized trees from $\mathbf{P}$.

The algorithm is as follows. Initialize $\mathbf{T}$ to be empty, and $\mathbf{P}'$ to equal $\mathbf{P}$. If $\mathbf{P}'$ contains no energized trees, terminate the algorithm. Otherwise, pick an energized tree T in $\mathbf{P}'$, which we can assume to be maximal with respect to set inclusion (so in particular it is convex, and removing T from $\mathbf{P}'$ will not disturb convexity). We can also assume that T is as low as possible, or in other words that ξ_{P_T}, the midpoint of ω_{P_T}, is minimal. (As we shall see, it is important that we remove trees from the bottom before we work our way to the top). We then remove T from $\mathbf{P}'$ and repeat the iteration.

In each individual tree T, we extract out the 2-tree

$$T' := \{P \in T : P^* \leq P_T^*\}.$$

In this tree the tiles P are all disjoint. We in fact have the stronger statement that the tiles in $\bigcup_{T \in \mathbf{T}} T'$ are all disjoint. To see this, suppose for contradiction that we had two trees T_1, T_2 with tiles P_1, P_2 such that $P_1 < P_2$. Then P_{T_1} is higher than P_{T_2}, so T_2 must have been selected first. But since $P_1 < P_2$, we also have $P_1 < P_{T_2}$, which implies that P_1 should have been incorporated into T_2 (as it was chosen to be maximal), a contradiction.

From disjointness and (15) we have

$$\sum_{T \in \mathbf{T}} \sum_{P \in T'} |b_P|^2 \leq 1;$$

since each tree is energized, (21) follows.

5. Proof of the tree estimate for 1-trees

To prove Lemma 2.6 we shall prove it for 1-trees and 2-trees separately.

Suppose T is a 1-tree. We now use (16) and (19) to estimate $\sum_P a_P b_P$ by $\sum_P |E_P|\text{energy}(T)$, so it suffices to show that

$$\sum_P |E_P| \lesssim \text{mass}(T)|I_{P_T}|.$$

To do this we shall split I_{P_T} into disjoint intervals J, such that

$$\sum_P |E_P \cap J| \lesssim \text{mass}(T)|J|.$$

We introduce the intervals J which are not of the form I_P for any $P \in \mathbf{T}$, but such that $2J$ is of the form I_{P_J} for some $P_J \in \mathbf{T}$. Clearly these intervals partition I_{P_T}. The tiles P which give a non-zero contribution to $|E_P \cap J|$ all obey $I_P \supseteq J$, and since we have a 1-tree, the sets $E_P \cap J$ are all disjoint and live inside $\tilde{E}_{P_J}$. The claim follows.

6. Proof of the tree estimate for 2-trees

Now suppose T is a 2-tree. Now the P are all disjoint, which suggests repeating the first argument from the heuristic discussion. Again, we rewrite $\sum_P a_P b_P$ as

$$\sum_{P \in T} \int_{E_P} b_P \phi_P.$$

Now we run into the issue of what to do with the E_P. We can rewrite the above as

$$\int_{E \cap I_{P_T}} \sum_{P \in T : N(x) \in \omega_P^*} b_P \phi_P. \tag{23}$$

Let us first forget about E by estimating this crudely by

$$\int_{I_{P_T}} \Big| \sum_{P \in T : N(x) \in \omega_P^*} b_P \phi_P \Big|. \tag{24}$$

If we could remove the constraint $N(x) \in \omega_P^*$, then we could use Cauchy-Schwarz to bound (24) by

$$|I_{P_T}|^{1/2} \| \sum_{P \in T} b_P \phi_P \|_2,$$

which by the definition of energy is $O(|I_{P_T}| \text{energy}(T))$. This doesn't gain us the mass factor, but we threw that away before and we'll return to it. So the issue is how to deal with the $N(x) \in \omega_P^*$ constraint. For each x, let $W(x)$ denote the largest dyadic interval containing ω_{P_T} which does not contain $N(x)$ (so for instance $W(x) = \mathbf{R}$ if $N(x) \in \omega_{P_T}$). Note that if $N(x) \in \omega_P^*$, then ω_P is disjoint from $W(x)$, while if $N(x) \notin \omega_P^*$, then ω_P is contained in $W(x)$. Thus we can write

$$\sum_{P \in T : N(x) \in \omega_P^*} b_P \phi_P = (1 - \pi_{W(x)}) \sum_{P \in T} b_P \phi_P.$$

The operator $\pi_{W(x)}$ is just Walsh convolution with some function supported on $[0, |W(x)|^{-1}]$ and has magnitude $|W(x)|^{-1}$, hence $\pi_{W(x)} f$ is pointwise dominated by the Hardy-Littlewood maximal function of f. Thus we have

$$\Big| \sum_{P \in T : N(x) \in \omega_P^*} b_P \phi_P \Big| \lesssim M \sum_{P \in T} b_P \phi_P. \tag{25}$$

So now we can use Cauchy-Schwarz to bound (24) by

$$|I_{P_T}|^{1/2} \| M(\sum_{P \in T} b_P \phi_P) \|_2,$$

which is OK by the previous arguments and the Hardy-Littlewood maximal inequality. So we can at least get the right bound but for the mass factor.

To put this mass factor back in it will suffice if we can control (23) by

$$\operatorname{mass}(\mathbf{T}) \int_{I_{P_T}} M \sum_{P \in T} b_P \phi_P$$

since we can then run the previous argument.

We again partition I_{P_T} into the intervals J which are of the form I_P for any $P \in \mathbf{T}$, but such that $2J$ is not of the form I_{P_J} for some $P_J \in \mathbf{T}$. It thus suffices to show that

$$|\int_{E \cap J} \sum_{P \in T : N(x) \in \omega_P^*} b_P \phi_P| \lesssim \operatorname{mass}(T) \int_J M| \sum_{P \in T} b_P \phi_P|$$

for each such interval J.

Fix J. The only P which contribute are those such that $P \geq P_J$. Thus the integrand vanishes unless $N(x) \in \omega_{P_J}^*$. Applying (25) we can therefore bound the left-hand side by

$$\int_{E \cap J; N(x) \in \omega_{P_J}^*} M \sum_{P \in T} b_P \phi_P.$$

The integrand is essentially constant on J, while the set integrated on has measure at most $O(\operatorname{mass}(T)|J|)$. The claim follows.

7. Homework for week 6 - due Friday, March 2

- Prove that for any dyadic interval I, the set $k \otimes I$ is another dyadic interval of side length $\|k\||I|$ (ignoring sets of measure zero). As a consequence, prove (1), for dyadic step functions f.
- Show that the function $\chi_{[0,1]}$ is its own Walsh-Fourier transform. Conclude from this the Walsh-Fourier transform maps dyadic step functions to dyadic step functions, and its own inverse.
- A *dyadic Calderón-Zygmund operator* is an operator T defined on (possibly vector-valued) functions on $\mathbf{R}^+$ by the formula

  $$Tf(x) := \int_{R^+} K(x,y) f(y)\, dy$$

  where T is bounded on L^2, and K obeys the bounds

  $$|K(x,y)| \lesssim \frac{1}{\|x \oplus y\|}$$

  and

  $$|K(x,y) - K(x,z)|, |K(y,x) - K(z,x)| \lesssim \frac{\|y \oplus z\|}{\|y \oplus x\|^2}$$

  whenever $\|y \oplus z\| < \|y \oplus x\|$. Prove that if T is a dyadic Calderón-Zygmund operator, then T is weak-type $(1,1)$, and is bounded on L^p for all $1 < p < \infty$. (Note: the proof should be a little simpler than the corresponding statement on the field $\mathbf{R}$. This is a general fact in harmonic analysis: the Walsh case is cleaner than the Fourier.)

Note that by repeating the arguments from the Week 2 notes, we thus obtain as a consequence the *dyadic Littlewood-Paley inequality*

$$\|f\|_p \sim \| \sum_{k \in \mathbf{Z}} (\pi_{[2^k, 2^{k+1}]} f|^2)^{1/2} \|_p$$

for $1 < p < \infty$.

Note: Thanks to Dmitry Ryabogin for comments and misprints.

Department of Mathematics, UCLA, Los Angeles CA 90095-1555

E-mail address: tao@math.ucla.edu