Polynomial Szemerédi in finite fields

Talk by Ashwin Sah

Notes by Sanjana Das

October 13, 2023

§1 Introduction

We'll discuss a paper of Sarah Peluse on polynomial Szemerédi in finite fields.

Question 1.1. Given a sequence of m polynomials $\mathcal{P} = (P_1, \dots, P_m) \in (\mathbb{Z}[x])^m$ and a set G, what is the largest subset of G that does not contain a nontrivial copy of $\mathcal{P}$?

We often take G to be $[N]$ or $\mathbb{F}_p$. By a copy of $\mathcal{P}$, we mean a pattern $\{x, x + P_1(y), \dots, x + P_m(y)\}$.

Definition 1.2. We use $r_{\mathcal{P}}(G)$ to denote the size of the largest $S \subseteq G$ with no nontrivial copy of $\mathcal{P}$.

When the polynomials are all linear, the interesting cases correspond to Szemerédi's theorem:

Theorem 1.3 (Szemerédi 1975)
For all k, we have

$$\frac{r_{(y,2y,\dots,(k-1)y)}([N])}{N} = o_k(1).$$

There are some sequences of polynomials that don't have this property.

Example 1.4

- If we consider $(y, y+1)$ (corresponding to $\{x, x+y, x+y+1\}$), we can take S to be the set of even numbers — this is a dense set which avoids this pattern.
- If we consider (y^2+1) (corresponding to $\{x, x+y^2+1\}$), then since y^2+1 is never divisible by 3, we can take S to consist of all multiples of 3.

So there can be mod conditions in the way; but the idea is that if we *don't* have such conditions, then we expect something like Szemerédi should hold. One example of this is the following.

Theorem 1.5 (Bergelson–Leibman)
If $P_1(0) = \cdots = P_m(0) = 0$, then we have

$$\frac{r_{\mathcal{P}}([N])}{N} = o_{\mathcal{P}}(1).$$

We generally want good bounds for these types of statements. The first good bound for Szemerédi's theorem is due to Gowers, and gives a bound of $(\log\log N)^{-c_k}$.

The proof of Theorem 1.5 is through ergodic theory, so it's completely non-quantitative.

Question 1.6. Can we get a quantitative bound for Theorem 1.5 at all? Can we get a *reasonable* one?

Furstenberg (1977) and Sárközy (1978) proved such a result for the pattern $\{x, x+y^2\}$. The bounds for this have been improved to a growing power of log; most recently, Slijepčević (2003) showed such a result for $\{x, x+P_1(y)\}$ whenever P_1 has no constant term.

§1.1 History

From now, we'll purely focus on the case $G = \mathbb{F}_p$, where we think of p as a large prime.

First, why is the problem easier over $\mathbb{F}_p$? In $[N]$, if we have a pattern $\{x, x+y^3, x+y^4\}$ in S, then since S has width N, this automatically tells you that $|y| \le N^{1/4}$. But in $\mathbb{F}_p$, this doesn't place any restrictions on y, because the powers of y wrap around mod p. With linear patterns, the wrap-around is generally not an issue (you can embed from $\mathbb{Z}$ into $\mathbb{Z}/M\mathbb{Z}$ on the same scale); but with higher-degree polynomials the wrap-around is much more of a problem. So in $\mathbb{F}_p$ we have extra room to work with, and this will be helpful when dealing with certain types of sums.

Here's what was known before, beyond having a single polynomial difference:

- Work by Bourgain and Chang (2016) got the pattern $\{x, x+y, x+y^2\}$.
- Work by Peluse (2016) and further improved by Dong, Li, and Sawin (2017) got the pattern

$$\{x, x+P_1(y), x+P_2(y)\}$$

  for any two polynomials P_1 and P_2 which are linearly independent. (Linear independence is important because if for example you had $P_2 = 2P_1$, then not only would you need to solve Roth's theorem in your group to begin with, but you'd also need it to be true with a special type of difference.)

These results build off a method introduced by Bourgain and Chang regarding nonlinear convolutions. A lot of these also use certain algebraic geometric inputs — for example, you need to constrain what varieties show up.

§1.2 The result

Definition 1.7. We define $\Lambda_{\mathcal{P}}(f_0, \dots, f_m) = \mathbb{E}_{x,y} f_0(x) f_1(x+P_1(y)) \cdots f_m(x+P_m(y))$.

Theorem 1.8 (Peluse 2019)

If $P_j(0) = 0$ for all j and the polynomials P_j are linearly independent, then for all $f_0, \dots, f_m : \mathbb{F}_p \to \mathbb{C}$ which are 1-bounded, we have

$$\Lambda_{\mathcal{P}}(f_0, \dots, f_m) = \prod_{j=0}^{m} \mathbb{E} f_j + O(p^{-c})$$

(where the constants only depend on the pattern $\mathcal{P}$).

You can think of the functions f_i as indicator functions of sets, but we allow this more general setting where they can be arbitrary $\mathbb{C}$-valued functions. For example, if the functions f_i are indicator functions of a set

with density α, then the right-hand side is the number of copies of the pattern that we'd expect in a *random* set with density α — Theorem 1.8 says we see the same number of copies in *any* set of that density.

Dong, Li, and Sawin improved the error term $O(p^{-c})$; for us, we'll have so many iterations that we won't be able to say much about this error term. (In Peluse's result, c depends on the pattern; in theirs it doesn't.)

In particular, Theorem 1.8 immediately implies a Szemerédi-type result regarding $r_{\mathcal{P}}(\mathbb{F}_p)$ — if we have a set S, then even if it has some *polynomial* density $p^{-\alpha}$, we can plug it into Theorem 1.8 and get that the normalized count of patterns is

$$p^{-\alpha(m+1)} + O(p^{-c}).$$

As long as α is small enough that the error term doesn't matter, this tells us exactly how many patterns are in our set, and we can subtract out the trivial ones and still be left with patterns. So this actually gives a *power-saving* bound for $r_{\mathcal{P}}(\mathbb{F}_p)$.

§2 Some motivation

First, why can something like Theorem 1.8 even possibly be true? To see this, we need a notion of complexity of patterns.

Example 2.1

For Roth's theorem (where our pattern is a 3-AP, i.e., $\{x, x+y, x+2y\}$), we consider

$$\Lambda_{\text{Roth}}(f_0, f_1, f_2) = \mathbb{E}_{x,y} f_0(x) f_1(x+y) f_2(x+2y).$$

The key idea in the proof of Roth's theorem is that if we have a set A with density α, there's a *randomness vs. structure* dichotomy. In the random case, we have

$$\Lambda_{\text{Roth}}(\mathbf{1}_A, \mathbf{1}_A, \mathbf{1}_A) \approx \alpha^3,$$

and then there's lots of 3-APs in A. If not, then we have lots of structure. Specifically, some U^2 norm will be large, and then we'll have some large correlation — $\mathbf{1}_A - \alpha$ will be correlated with a function $e^{i\theta x}$. And in that case, we can pass to a subprogression.

One way to see why these functions $e^{i\theta x}$ show up is to imagine plugging in $f_0 = e^{i\theta x}$, $f_1 = e^{i(-2\theta)x}$, and $f_2 = e^{i\theta x}$. If θ is pretty nonzero, then each f_i should roughly average out to 0 in the long term; but we have

$$f_0(x) f_1(x+y) f_2(x+2y) = 1$$

(both the x's and the y's cancel out). So in some sense, these Fourier phases allow us to construct sets with strange 3-AP densities.

It's not just linear phases that come up:

Example 2.2

For the pattern $\{x, x+y, x+2y, x+3y\}$, you can take $e^{-i\theta x^2}$, $e^{i(3\theta)x^2}$, $e^{-i(3\theta)x^2}$, and $e^{i\theta x^2}$.

Morally, every pattern has this collection of possible obstructions that you can get; and those should provide the key things to look for in the structure vs. randomness dichotomy. (This leads to things like higher uniformity norms and higher-order Fourier analysis.)

But for us (in the setting of Theorem 1.8), let's look at what happens with linear phases (higher-order stuff won't happen here) — let's consider

$$\Lambda_{\mathcal{P}} = \mathbb{E}_{x,y} f_0(x) f_1(x + P_1(y)) \cdots f_m(x + P_m(y)),$$

and look at what happens if each f_i is some exponential. For convenience, let $e_p(t) = e^{2\pi it/p}$ (these are the Fourier characters over $\mathbb{F}_p$). If we suppose that

$$f_j(x) = e_p(\alpha_j x)$$

for all j and plug this into our counting operator, then we get

$$\Lambda_{\mathcal{P}} = \mathbb{E}_{x,y} e_p\left(\sum_{j=0}^{m} \alpha_j x + \sum_{j=1}^{m} \alpha_j P_j(y)\right)$$

(adding together the exponents, and combining the x-terms into one sum and the y-terms into another).

Let's assume that $(\alpha_0, \dots, \alpha_m)$ is nonzero (if it were the zero vector, then we'd just be plugging in a bunch of constant functions, which isn't interesting). For the statement in Theorem 1.8 to be true (which essentially says that the random case holds), we want the right-hand side to be small. It suffices to consider the case where $\sum_{j=0}^{m} \alpha_j = 0$, since otherwise averaging over x *already* cancels out everything. Then the above expression becomes

$$\Lambda_{\mathcal{P}} = \mathbb{E}_y e_p\left(\sum_{j=1}^{m} \alpha_j P_j(y)\right).$$

Now this is where the property that the polynomials are linearly independent comes in — it's essentially to avoid Roth-like configurations. Specifically, linear independence means that $\sum_{j=1}^{m} \alpha_j P_j(y)$ is some nonzero polynomial $Q(y)$. (We don't know anything about its coefficients, but we do know it's some nonzero polynomial mod p.)

So now we've reduced the question of evaluating our operator in this example to evaluating an exponential sum of some polynomial Q. For example, if $Q(y) = y^2$, then this becomes the traditional Gauss sum; and the Gauss sum is $p^{1/2}$, so the corresponding average is $p^{-1/2}$, which goes into the error term in Theorem 1.8. In general, each sum like this will be uniformly small. (Trying to encode this is where the algebraic geometry comes in — for example, Peluse uses the Weil bound for curves to show that there's cancellation.)

So the point is that since we have linear independence, there are *no* obstructions at all — there's no 'structure' case. Now the question is, how do we make this formal? This argument worked well for pure phases, but how do we show *every* function is represented by this behavior?

> **Remark 2.3.** These patterns might be called as having *true complexity* 0; 3-APs (as in Roth's theorem) have true complexity 1; and 4-APs have true complexity 2. This has to do with the degree of the polynomials involved — the definition of true complexity is sort of as the smallest k such that the operator will be small if all the Gowers U^k norms of your functions are. But there are also many other definitions of complexity.

§3 Outline

We won't look at all the steps of the proof in detail; we'll talk mostly about a key step called the *degree-lowering phase*. We'll also focus on the explicit pattern $\{x, x+y, x+y^2\}$. (The result for this pattern was already known, but we'll prove it using the degree-lowering method rather than the original proof.) In this case, we have

$$\Lambda(f_0, f_1, f_2) = \mathbb{E}_{x,y} f_0(x) f_1(x+y) f_2(x+y^2).$$

(We'll always assume our functions are 1-bounded, and we won't worry about complex conjugates unless we need to.)

The classic thing to do in order to evaluate the right-hand side is to write each function as a combination of its mean value and a mean-0 portion (i.e., $f = \mathbb{E}f + (f - \mathbb{E}f)$) and expand out the multilinear form, and show that everything except the dominant contribution (the one coming from the three means) is small. So the key proposition we'll prove is the following.

Proposition 3.1

If f_0, f_1, and f_2 are 1-bounded and $|\Lambda(f_0, f_1, f_2)| \geq \delta$, then

$$\min_{j \in \{0,1,2\}} |\mathbb{E}f_j| \gtrsim \delta^C.$$

(Here and in most lemmas, we need $p \geq \delta^{-\Omega(1)}$, i.e., p needs to be polynomially large in δ.)

So this states that if $|\Lambda|$ is large, then *each* of the expectations has to be large.

We can't prove this immediately. Instead, we first show something weaker — that all the Gowers U^k norms are large for some k.

Lemma 3.2 (PET induction)

If $|\Lambda(f_0, f_1, f_2)| \geq \delta$, then

$$\min_j \|f_j\|_{U^s}^{2^s} \gtrsim \delta^C,$$

where s is some explicit constant depending on $\mathcal{P}$.

For the specific pattern we're considering, s is 3 or 4; but it blows up very quickly. But the point is that once you prove Lemma 3.2, it's some fixed number (e.g., something like 10^9).

To prove Lemma 3.2, you do something similar to the proof of Roth's theorem, with iterated Gowers–Cauchy–Schwarz. It's a bit more complicated because we have polynomials involved, which causes multilinear products of differences and basepoints to show up. But it turns out that over $\mathbb{F}_p$, since multiplication is an isomorphism over the group, we can get away with doing the things we want enough times. Over the integers, this causes significant problems, and there's a concatenation step in papers that involve this situation (which we won't talk about).

First, here's the definition of the Gowers norms.

Definition 3.3. For a function $f : \mathbb{F}_p \to \mathbb{C}$ and some difference $h \in \mathbb{F}_p$, we define

$$\Delta_h f(x) = f(x)\overline{f(x+h)}.$$

We also define $\Delta_{h_1,\dots,h_k}$ as the composition $\Delta_{h_1} \cdots \Delta_{h_k}$.

If f is the exponential of a polynomial, then Δ_h takes the discrete derivative of the exponent with respect to h — so this reduces the degree of the polynomial in the exponent. This means the way to detect whether some function is a phase of a kth power is that if we take a bunch of derivatives of this form, we should end up with the 1 function.

To turn this into a norm, we look at what happens on average.

Definition 3.4. The Gowers U^k norm is defined as

$$\|f\|_{U^k}^{2^k} = \mathbb{E}_{x,h_1,\dots,h_k} \Delta_{h_1,\dots,h_k} f(x).$$

This is well-defined and is actually a norm — if $k \geq 1$, this quantity will always be a nonnegative real. A key property is that we can extract out a couple of differences and put them somewhere else:

Fact 3.5 — For any t, we have $\|f\|_{U^k}^{2^k} = \mathbb{E}_{h_1,\dots,h_t} \|\Delta_{h_1,\dots,h_t} f\|_{U^{k-t}}^{2^{k-t}}$.

Note that when $k = 1$, we get

$$\|f\|_{U^1}^2 = \mathbb{E}_{x,h} f(x)\overline{f(x+h)} = |\mathbb{E} f|^2 .$$

(This one is a seminorm, but all the others are genuinely norms.)

So the point is that Lemma 3.2 gives a statement for some large constant s, e.g., $s = 10^9$; and what we *want* is the same statement for $s = 1$. So the key idea is to somehow go from $s = 10^9$ to $s = 1$. This is what the degree-lowering step does, which we'll talk about for the rest of the time.

§4 Degree lowering

First, here's the naive hope of degree lowering. To start with, we know that $\|f_j\|_{U^s}^{2^s}$ is large. The dream would be to somehow show that then $\|f_j\|_{U^{s-1}}^{2^{s-1}}$ is also large, and then iterate to $s-2$, and so on.

Obviously this won't work for general functions. Why not? Well, if f_j is the exponential of a degree-$(s-1)$ polynomial, then $\|f_j\|_{U^s}^{2^s}$ is large but $\|f_j\|_{U^{s-1}}^{2^{s-1}}$ is small. And the functions f_j we're given in Proposition 3.1 could be completely general; so this isn't going to work.

The key idea is that instead of naively applying this dream, we use an intermediate concept of *dual functions*. In this paper by Peluse over finite fields, she uses hyperplane separation to look at a decomposition of functions into one part that's large and another part that's large in the dual norm. We won't do this; instead we'll see something more concrete based on more recent papers.

§4.1 Dual functions

The idea is to use *dual functions*, and there's a nice way to package these insights through *stashing*.

First, we know that

$$\delta \le |\Lambda(f_0, f_1, f_2)| = \left|\mathbb{E}_x f_0(x) \mathbb{E}_y f_1(x+y) f_2(x+y^2)\right|$$

(pulling out the term $f_0(x)$ that only contains x). The second function $\mathbb{E}_y f_1(x+y) f_2(x+y^2)$ is another function of x; we write it as $D_0(f_1, f_2)(x)$, and call it the *dual* function to f_0 — it's a dual function in the sense that we have

$$\Lambda(f_0, f_1, f_2) = \langle f_0, D_0(f_1, f_2)\rangle. \tag{4.1}$$

Now we have a dot product of two functions, so we can use Cauchy–Schwarz; this gives

$$|\langle f_0, D_0(f_1, f_2)\rangle| \le \|f_0\|_2 \|D_0(f_1, f_2)\|_2 \le \|D_0(f_1, f_2)\|_2$$

(since f_0 is 1-bounded). And since the left-hand side was at least δ, this means

$$\delta^2 \le \|D_0(f_1, f_2)\|_2^2 = \langle D_0(f_1, f_2), D_0(f_1, f_2)\rangle.$$

And recall that the dual function $D_0(f_1, f_2)$ had the property (4.1) that if we take *any* function f_0 and dot it with this dual, we get the corresponding 3-fold operator $\Lambda(f_0, f_1, f_2)$. And 'any function' includes itself, so we get

$$\delta^2 \le \Lambda(D_0(f_1, f_2), f_1, f_2).$$

So what we've done is that one application of Cauchy–Schwarz replaces (f_0, f_1, f_2) with $(D_0(f_1, f_2), f_1, f_2)$. (This — i.e., replacing f_0 with a copy of $D_0(f_1, f_2)$ — is called *stashing*.)

Now here's the hope: We said before that we couldn't have our dream (of going directly from the U^s norm being large to the U^{s-1} norm being large) because f_0, f_1, and f_2 could be completely generic. But the dual functions that come from these operators aren't completely generic — they should be more regular than the original functions. So the hope is that you can somehow do something like the dream, but starting with the *dual* function instead.

So instead of applying Lemma 3.2 to our original operator $\Lambda(f_0, f_1, f_2)$ (which we were given is large), we apply Lemma 3.2 to this new operator $\Lambda(D_0(f_1, f_2), f_1, f_2)$ (which we just showed is also large by Cauchy–Schwarz). Then Lemma 3.2 tells us that

$$\|D_0(f_1, f_2)\|_{U^s}$$

is also large. And this function $D_0(f_1, f_2)$ has some structure we can potentially work with; so our goal is to show that this implies $\|f_1\|_{U^{s-1}}$ and $\|f_2\|_{U^{s-1}}$ are large.

> **Remark 4.1.** Note that we don't apply Lemma 3.2 to the functions f_j directly — we don't start with the hypothesis that $\|f_j\|_{U^s}$ is large. Instead, we use the fact that Lemma 3.2 works for *any* triple of functions. Lemma 3.2 gives U^s control, and our goal is to prove U^{s-1} control. And the idea is that to prove this new lemma with U^{s-1} control, we start with the given assumption $\delta \le |\Lambda(f_0, f_1, f_2)|$ and use the dual function and stashing trick to conclude that $\Lambda(D_0(f_1, f_2), f_1, f_2)$ is large as well, and we then apply Lemma 3.2 to *these* functions.

If we can do this, then we'll have gone from Lemma 3.2 with s to the same statement with $s-1$. (This argument only got that the norms of f_1 and f_2 were large, but you can do the same argument replacing one of the other functions to show that the norm of f_0 is also large.) Then we can do the same thing to go from $s-1$ to $s-2$, and so on (this would give us a way to go from U^s control to U^{s-1} control for any $s \ge 2$).

So our goal is now to prove the statement

$$\|D_0(f_1, f_2)\|_{U^s} \text{ is large } \implies \|f_1\|_{U^{s-1}} \text{ and } \|f_2\|_{U^{s-1}} \text{ are large.} \tag{$*$}$$

(If we can prove this for all s, then that's enough.)

§4.2 Degree lowering from $s = 2$ to $s = 1$

First let's consider the final step of degree lowering, where we go from U^2 to U^1 (i.e., we're considering $(*)$ for $s = 2$). Here we start with the assumption that

$$\|D_0(f_1, f_2)\|_{U^2}^4 \gtrsim \delta^C,$$

and we want to say something about the U^1 norms (i.e., means) of f_1 and f_2. First, recall that

$$\|g\|_{U^2}^4 = \mathbb{E}_{x,a,b} g(x)\overline{g(x+a)g(x+b)}g(x+a+b).$$

The key property is that if we apply Fourier inversion to this, we can rewrite it as

$$\|g\|_{U^2}^4 = \sum_{\alpha \in \mathbb{F}_p} |\widehat{g}(\alpha)|^4.$$

This is used in the proof of Roth's theorem. There, the key step is bounding this by

$$\|g\|_{U^2}^4 \le \|g\|_\infty^2 \cdot \|g\|_2^2 \le \|g\|_\infty^2$$

(we have $\|g\|_2 \le 1$ because g is 1-bounded). This means that if the U^2 norm of our function is large, then its Fourier transform has large L^∞ norm, which means the function has a large correlation with $e_p(\alpha x)$ for some α (corresponding to the 'structure' case from the proof of Roth's theorem).

Here, we'll use the same fact:

Theorem 4.2 (U^2 inverse theorem)
If $\|D_0(f_1, f_2)\|_{U^2}^4 \gtrsim \delta^C$, then there exists $\alpha \in \mathbb{F}_p$ such that

$$|\mathbb{E}_x D_0(f_1, f_2)(x) e_p(\alpha x)| \gtrsim \delta^C.$$

And now we're again in the situation where we have some function dotted against a dual function; so we can rewrite the left-hand side as

$$\mathbb{E}_{x,y} e_p(\alpha x) f_1(x+y) f_2(x+y^2). \tag{4.2}$$

So at the beginning of the argument we had a completely generic function f_0; then we used stashing to replace it with the dual function $D_0(f_1, f_2)$; and then we used the U^2 inverse theorem to say that this dual function correlates with a linear phase. So somehow we've magically replaced the original arbitrary function f_0 with a linear phase.

This should stand out because if we knew that *all* the functions were linear phases, then we could run the argument we gave at the very beginning (in Section 2). We've only replaced *one* with a linear phase, and we're deep inside some iteration of proving some lemma, so we can't just simply do this again to replace another. But still, you can see why this is useful — we no longer have the complaint from earlier about our functions f_j being completely generic.

In this case, we're going to make an argument that only works if the pattern has 3 terms; we'll later explain how you might modify it to deal with larger patterns.

In (4.2), $e_p(\alpha x)$ is just a linear phase; so we just have the dot product of two functions with some weights, and it's natural to apply Fourier inversion. So let's write f_1 and f_2 as their Fourier series, expand out, and see what happens: this gives

$$\sum_{\theta_1, \theta_2 \in \mathbb{F}_p} \widehat{f_1}(\theta_1) \widehat{f_2}(\theta_2) \mathbb{E}_{x,y} e_p(\alpha x + \theta_1(x+y) + \theta_2(x+y^2)).$$

(Here we're using θ_1 for the Fourier expansion of f_1 and θ_2 for f_2.)

Again, if $\alpha + \theta_1 + \theta_2 \neq 0$, then just averaging over x will completely kill this — i.e., the inner expectation will vanish (even if we just take an expectation over x). So we only need to consider terms where $(\theta_1, \theta_2) = (\theta, -\alpha - \theta)$. Then we can remove the expectation over x, and we're left with

$$\sum_{\theta \in \mathbb{F}_p} \widehat{f_1}(\theta) \widehat{f_2}(-\alpha - \theta) \mathbb{E}_y e_p(\theta y - (\alpha + \theta) y^2).$$

Now we're in potentially good shape. Why? If we consider the inner exponential sum

$$\mathbb{E}_y e_p(\theta y - (\alpha + \theta) y^2),$$

the argument we gave at the beginning (in Section 2) applies — unless both the coefficients θ and $-(\alpha + \theta)$ are 0, this inner expectation is uniformly bounded by something like $p^{-1/2}$. (In this case it's actually a Gauss sum, so it's genuinely $p^{-1/2}$; but we really just need some cancellation.)

Case 1 ($\alpha \neq 0$). Then for any θ, one of θ or $-(\alpha + \theta)$ is 0, which means we get this cancellation; so then the above expression is at most

$$p^{-1/2} \sum_{\theta} |\widehat{f_1}(\theta)| |\widehat{f_2}(-\alpha - \theta)|.$$

And we assumed it was at least δ^C, so we get that

$$\delta^C \lesssim p^{-1/2} \sum_{\theta} |\widehat{f_1}(\theta)| |\widehat{f_2}(-\alpha - \theta)|.$$

Now since we have only two functions here, we can use Cauchy–Schwarz (this is why the argument only works if we start with 3-term patterns); this gives

$$\delta^C \lesssim p^{-1/2} \left(\sum |\widehat{f_1}(\theta)|^2\right)^{1/2} \left(\sum |\widehat{f_2}(\theta)|^2\right)^{1/2} = p^{-1/2} \|f_1\|_2 \|f_2\|_2 .$$

But $\|f_1\|_2$ and $\|f_2\|_2$ are both at most 1, so this is a contradiction.

Case 2 ($\alpha = 0$). In this case, plugging $\alpha = 0$ into (4.2) gives that

$$|\mathbb{E}_x D_0(f_1, f_2)(x)| = \left|\mathbb{E}_{x,y} f_1(x+y) f_2(x+y^2)\right| \gtrsim \delta^C.$$

And we can rewrite this as

$$\mathbb{E}_{x,y} f_1(x) f_2(x + y^2 - y).$$

This is a simpler pattern, so we can use induction — citing the final result for this simpler pattern — to get what we want. (For this special case, we could also just cite Sárközy.)

So that concludes the degree-lowering phase from $s = 2$ to $s = 1$. The key idea is that we use stashing to reduce to dealing with two functions, and use the U^2 inverse theorem to replace the most complicated thing with a single phase. Then in the end, it reduces to the fact that the exponential sum we checked at the beginning works out.

Remark 4.3. What differs for a larger pattern (with more than 3 terms)? Roughly, once we get to something of the form

$$\mathbb{E}_{x,y} e_p(\alpha x) f_1(x+y) f_2(x+y^2),$$

we treat this as a mixture of functions and explicitly given phases. And we prove the proposition in greater generality for such objects, doing an extra outer induction on how many functions we've replaced with phases. So we have an extra induction with mixed function-phase types of things, but we still do replacement one at a time.

§4.3 Larger values of s

So far, we've only proved the degree-lowering step ($*$) for $s = 2$; it remains to show it for larger s. (That, along with Lemma 3.2, completely finishes the proof.) We'll now give a flavor of how this works.

The final step is 'lifting via dual-difference exchange.' The vague idea is as follows: We're starting with U^s control, which means we're taking s derivatives of our dual function. We want to somehow remove $s-2$ of those derivatives and place them onto the *functions*, so that we can almost work with the U^2 situation as a black box (averaged over the choices of differences). So we want to take some of the difference operators that come with the U^s norm and exchange them with our operator.

Here's the key lemma for this, which is somewhat general:

Lemma 4.4

Let $g : \mathbb{F}_p^2 \to \mathbb{C}$ be 1-bounded. Let

$$A = \mathbb{E}_{x,h_1,\dots,h_k} \Delta^{(x)}_{h_1,\dots,h_k} [\mathbb{E}_y g(x,y)],$$
$$B = \mathbb{E}_{h_k} \mathbb{E}_{x,h_1,\dots,h_{k-1}} \Delta^{(x)}_{h_1,\dots,h_{k-1}} [\mathbb{E}_y \Delta^{(x)}_{h_k} g(x,y)].$$

Then we have $|A|^2 \le B$.

(The (x) denotes that we're taking discrete derivatives with respect to x.)

In our situation, what does this do? Recall that we're starting with

$$\|D_0(f_1, f_2)\|_{U^s}^{2^s} = \mathbb{E}_{x,h_1,\dots,h_s} \Delta_{h_1,\dots,h_s} D_0(f_1, f_2).$$

And by definition, we have

$$D_0(f_1, f_2)(x) = \mathbb{E}_y f_1(x+y) f_2(x+y^2).$$

This inner part is what we're calling g — i.e., $g(x, y) = f_1(x+y) f_2(x+y^2)$.

So with $\|D_0(f_1, f_2)\|_{U^s}^{2^s}$, we have an inner average over some dual function variable y, and then we're taking a s-fold Gowers norm; this means we have an expression of the form A. And Lemma 4.4 says that if A is large, then B is also large.

And what is B? You can think of

$$\mathbb{E}_{x,h_1,\dots,h_{k-1}} \Delta^{(x)}_{h_1,\dots,h_{k-1}} [\mathbb{E}_y \Delta^{(x)}_{h_k} g(x, y)]$$

as a $(k-1)$-fold Gowers norm where we've put the last difference on the inside; then in B we're averaging over this last difference. So this allows us to push some of the partials inside.

> **Remark 4.5.** In terms of how to prove Lemma 4.4, you expand both things out and do Cauchy–Schwarz once (cleverly) on h_k. Note that when you expand out, you'll have 2^k different copies of a y variable corresponding to each thing, so this takes some work.

With that in mind, how do we reduce from the situation with general s to the specific situation with 2? We're starting with the assumption that

$$\|D_0(f_1, f_2)\|_{U^s}^{2^s} \gtrsim \delta^C.$$

If we apply one copy of dual-difference exchange, it tells us that

$$\mathbb{E}_{h_s} \|D_0(\Delta_{h_s} f_1, \Delta_{h_s} f_2)\|_{U^{s-1}}^{2^{s-1}} \gtrsim \delta^{2C}.$$

We do this another $s-3$ times, to eventually get that

$$\mathbb{E}_{h_3,\dots,h_s} \|D_0(\Delta_{h_3,\dots,h_s} f_1, \Delta_{h_3,\dots,h_s} f_2)\|_{U^2}^4 \gtrsim \delta^{2^{s-2}C}.$$

Then we can apply Markov to say that $\|D_0(\Delta_{h_3,\dots,h_s} f_1, \Delta_{h_3,\dots,h_s} f_2)\|_{U^2}^4$ is large for a positive fraction of h_3, $\dots$, h_s; then we can apply the U^2 case; and then we can re-average to conclude.