The sum-difference conjecture

Talk by Manik Dhar

Notes by Sanjana Das

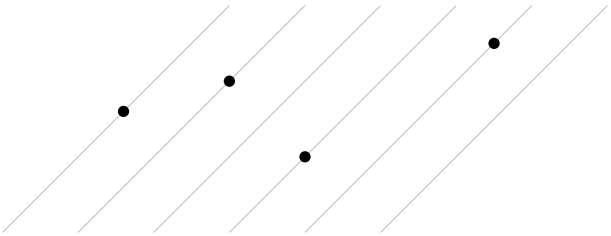
November 17, 2023

§1 The problem

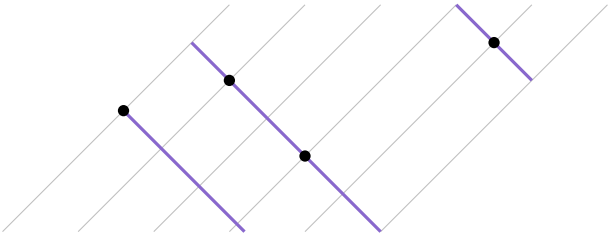
§1.1 Some motivation

Imagine we have a finite set of points $G \subseteq \mathbb{R}^2$. We'll also consider lines in $\mathbb{R}^2$; the equation of a line in $\mathbb{R}^2$ looks like $y + mx = c$, and we'll refer to m as the *slope* of this line. (This is nonstandard — usually the slope is defined as $-m$ — but it'll be a bit more convenient for our purposes.)

Suppose that we know every line with slope -1 intersects G at most once.

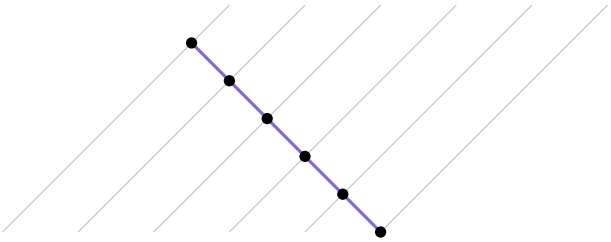


And suppose we also have information about G in one more direction — specifically, we'll look at lines with slope r, and consider the number of lines of slope r needed to cover G, which we define as $L_r(G)$.



Question 1.1. If we know that $L_r(G) \leq n$, then can we say anything about $|G|$?

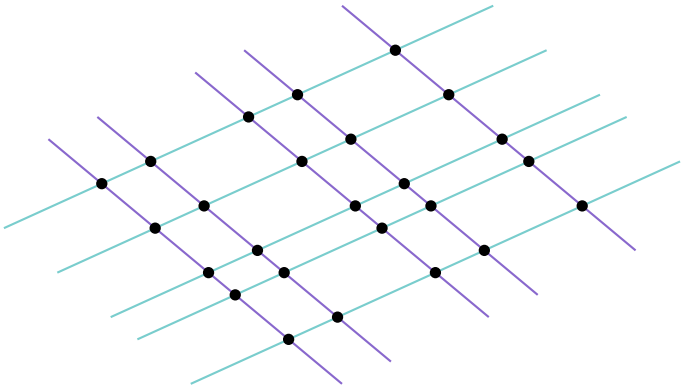
Unfortunately, the answer is no — imagine that we take a bunch of lines with slope 1 and a single line with slope r, and take G to consist of all the corresponding intersections. Then $L_r(G) = 1$, but $|G|$ can be arbitrarily large.



So this situation isn't so nice — having just one slope r isn't enough to tell us anything about $|G|$. So let's imagine we have one more slope to work with.

> **Question 1.2.** If we know that $L_{r_1}(G) \leq n$ and $L_{r_2}(G) \leq n$ (for distinct slopes r_1 and r_2), then can we say anything about G?

Now we're in business — we have lines in two different directions, which give us a sort of coordinate system. A point is completely determined by its coordinates, and there's at most n coordinates in both directions, so there's at most n^2 points — so we get $|G| \leq n^2$.



And this is the best bound we can get — we can achieve equality by taking a $n \times n$ grid (with directions chosen such that every line with slope -1 intersects G at most once).

> **Question 1.3.** What if we have information in another direction (i.e., that $L_{r_3}(G) \leq n$) — can we keep pushing the bound on $|G|$ downwards?

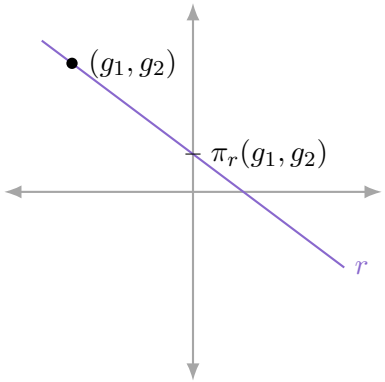
And that's essentially what the sum-difference conjecture is about.

§1.2 The sum-difference conjecture

We've set things up in $\mathbb{R}^2$ so far, but we'll actually instead be working in $W \times W$ for a finite-dimensional real vector space W.

> **Definition 1.4.** For any $r \in \mathbb{R}$, we define $\pi_r \colon W \times W \to W$ as the map $(g_1, g_2) \mapsto g_1 + rg_2$. We also define $\pi_\infty \colon W \times W \to W$ as the map $(g_1, g_2) \mapsto g_2$.

Intuitively, we think of r as a 'slope'; when $W = \mathbb{R}$, the map π_r keeps track of which line with slope r the point (g_1, g_2) lies on (e.g., by recording its y-intercept).



Definition 1.5. Given a finite set of 'slopes' $R = \{r_1, \dots, r_m\}$, none of which is -1, we say that $\mathsf{SD}(R, \alpha)$ holds if for all $G \subseteq W \times W$ such that π_{-1} is injective on G, if $|\pi_r(G)| \leq n$ for all $r \in R$, then $|G| \leq n^\alpha$.

The condition that π_{-1} is injective on G corresponds to the statement (in the case $W = \mathbb{R}$) that every line of slope -1 intersects G at most once. And we're given that for each of the directions $r \in R$, we only need n lines in the direction r in order to cover G; and we want to conclude that G has at most n^α points.

Example 1.6

Our argument from earlier (with two slopes) shows that $\mathsf{SD}(\{r_1, r_2\}, 2)$ holds (for any $r_1 \neq r_2$).

Definition 1.7. We say the statement $\mathsf{SD}(\alpha)$ holds if for every $\varepsilon > 0$, we can find some set of slopes R_ε for which the statement $\mathsf{SD}(R_\varepsilon, \alpha + \varepsilon)$ holds.

In other words, $\mathsf{SD}(\alpha)$ states that by having information about G in sufficiently many directions, we can push our bounds on $|G|$ arbitrarily close to n^α.

Conjecture 1.8 (Sum-difference conjecture) — The statement $\mathsf{SD}(1)$ holds.

Remark 1.9. Why is this called the sum-difference conjecture? First, as an alternative definition of $\mathsf{SD}(R, \alpha)$, we could remove the condition that π_{-1} is injective on G and instead ask for the conclusion that $|\pi_{-1}(S)| \leq n^\alpha$. Clearly this formulation implies the original; meanwhile, the original formulation also implies this one, as we can delete points with repeated values of π_{-1}.

Then we're given a set of points (g_1, g_2), and we know there aren't too many sums $g_1 + rg_2$ for several 'slopes' r; and we want to deduce that there aren't too many differences $g_1 - g_2$.

§1.3 History

We saw just that with two slopes, the best bound we can get is $\alpha = 2$. Bourgain showed that we can do a bit better using *three* slopes instead.

Theorem 1.10 (Bourgain)

The statement $\mathsf{SD}(\{0, 1, \infty\}, 2 - 1/13)$ holds.

(The specific slopes 0, 1, and ∞ are somewhat important — we can modify them by performing a projective transformation, but the argument relies on having a certain cross ratio.)

And we can do even better by allowing *four* slopes.

Theorem 1.11 (Katz–Tao 1999)

The statement $\mathsf{SD}(\{0, 1, 2, \infty\}, 7/4)$ holds.

And the best-known bound comes from two years later.

Theorem 1.12 (Katz–Tao 2001)

The statement $\mathsf{SD}(\alpha)$ holds, where $\alpha \approx 1.675$ is the root of $\alpha^3 - 4\alpha + 2 = 0$ in $[1, 2]$.

In the two decades since this bound, we haven't really been able to improve it.

Remark 1.13. These bounds are true for W of any dimension. In fact, the general setting should be equivalent to the one-dimensional case (as we can take a generic projection down to one dimension).

§2 Connection to the Kakeya conjecture

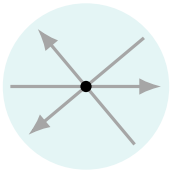
The way we've stated the sum-difference conjecture makes it look like a natural additive combinatorics problem, but the reason people first started studying it was actually because of a connection to the Kakeya conjecture.

§2.1 The Kakeya conjecture

The story of the Kakeya conjecture begins in 1917.

Question 2.1 (Kakeya)**.** Suppose we have a unit needle in $\mathbb{R}^2$, and we're allowed to translate and rotate it (inside some shape). What's the smallest possible shape (in terms of area) that we need to be able to make the needle point in every direction?

As a simple example, we can imagine rotating the needle about its center; this gives a circle.



Kakeya came up with a slightly better construction using a cycloid, and he thought that this would be tight. But surprisingly, it turns out the real answer — for the minimum area necessary — is (almost) zero!

Theorem 2.2 (Besikovich 1918)

For any $\varepsilon > 0$, there exists a set of area at most ε which works (meaning that we can move a unit needle around in the set and make it point in every direction).

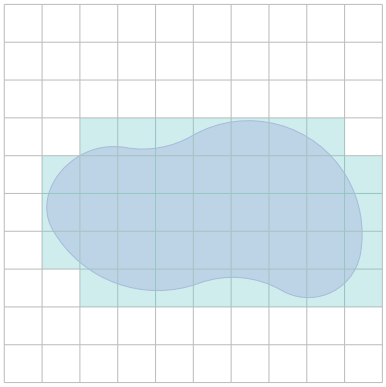
We can do this in higher dimensions as well; there, instead of thinking about moving a needle around, we'll just ask that the set contains a unit segment in each possible direction.

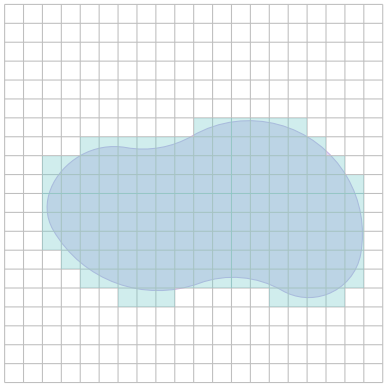
Definition 2.3. A set $K \subseteq \mathbb{R}^k$ is Kakeya if it is compact and contains a unit segment in every direction.

The same construction of Besikovich still works, giving that there is a Kakeya set of measure ε for any $\varepsilon > 0$; and in fact taking the limit as $\varepsilon \to 0$ gives a construction with measure 0. This seems very surprising — we might expect that containing a unit segment in every direction should force K to be 'large' in some sense, but this isn't true if we think of 'largeness' in terms of measure. So instead, we'll use a different way of quantifying the 'largeness' of a set, namely the Minkowski dimension. (This can also be done with the Hausdorff dimension, but today we'll only discuss the Minkowski dimension.)

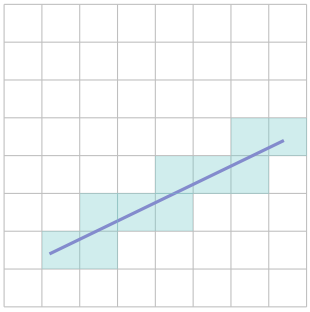
Definition 2.4 (Minkowski dimension)**.** Let $K \subseteq [0,1]^k$. For each $\varepsilon > 0$, consider the ε-grid over $[0,1]^k$, and let $b_\varepsilon(K)$ be the number of cubes of this grid needed to cover K. Then we define the Minkowski dimension of K as

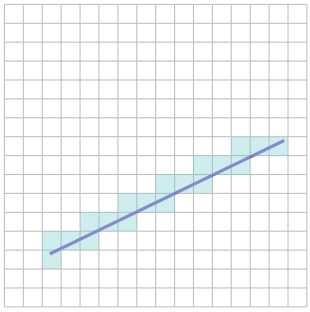
$$\dim(K) = \lim_{\varepsilon \to 0} \frac{\log(b_\varepsilon(K))}{\log(1/\varepsilon)}.$$





Intuitively, this means we'd expect $b_\varepsilon(K)$ to scale according to $(1/\varepsilon)^d$ for some d as ε shrinks, and we define the Minkowski dimension as this value of d. We can check that objects such as lines and planes do have the dimension we'd expect them to (1 and 2, respectively) — for example, for a line segment, halving the grid length should double the number of squares the segment passes through.





If a set in $\mathbb{R}^k$ has positive measure, then it definitely has Minkowski dimension k. But it's possible for a set to have measure zero but still have Minkowski dimension k.

And the Kakeya conjecture is that any Kakeya set should be large if we measure largeness by Minkowski dimension — specifically, it has to have the maximum possible dimension.

Conjecture 2.5 (Kakeya conjecture) **—** If $K \subseteq \mathbb{R}^k$ is a Kakeya set, then $\dim(K) = k$.

§2.2 History

It's known that when $d = 2$, the Kakeya conjecture is true — this is a result due to Davies (1971), which uses just the fact that two lines intersect at one point together with Cauchy–Schwarz.

Meanwhile, for large n, the best bound we have (on the dimension of a Kakeya set) is again due to Katz–Tao (from the same paper from 2001).

Theorem 2.6 (Katz–Tao 2001)
If $K \subseteq \mathbb{R}^k$ is a Kakeya set, then $\dim(K) \geq \beta k$, where $\beta \approx 0.59$ is some irrational constant.

In fact, the value of β here is precisely $1/\alpha$ for the value of α from Theorem 1.12, and this isn't a coincidence — there's a reduction from the sum-difference problem to the Kakeya one, and in fact solving the sum-difference conjecture would actually solve the Kakeya conjecture as well. (We don't have a reduction in the other direction, and don't expect one to exist.)

§2.3 Reduction from sum-difference to Kakeya

Now we'll state and prove this reduction.

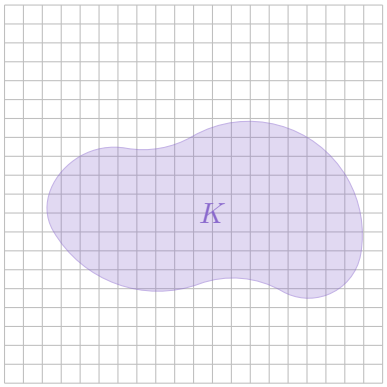
Lemma 2.7
If $\mathsf{SD}(\alpha)$ holds, then any Kakeya set $K \subseteq \mathbb{R}^k$ must satisfy $\dim(K) \geq k/\alpha$.

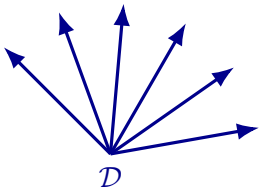
In particular, $\mathsf{SD}(1)$ would imply the Kakeya conjecture.

Proof. Suppose that $\mathsf{SD}(\{1, \dots, m\}, \alpha)$ holds. Here we're taking a very particular choice of slopes, but we can actually do this without loss of generality — it's possible to show that if $\mathsf{SD}(R, \alpha)$ holds for some R, then we can replace R with a (possibly bigger) set $\{1, \dots, m\}$. (Technically, we only get to assume that $\mathsf{SD}(\{1, \dots, m\}, \alpha + \varepsilon)$ holds for some m for each $\varepsilon > 0$ (where m depends on ε), but this is good enough — if we can show $\dim(K) \geq k/(\alpha + \varepsilon)$ for all $\varepsilon > 0$, then we get $\dim(K) \geq k/\alpha$. So we'll ignore this.)

First, since K is compact, it must be bounded, so we can scale so that $K \subseteq [0, 1]^n$.

Now take an ε-grid of $[0, 1]^k$, and let $\mathcal{D}$ be an ε-separated set of directions. (We can think of our directions as living in $(k-1)$-dimensional projective space, and we can assign a metric to this space — the choice of metric doesn't really matter, since we think of k as fixed; and then we're requiring all our directions to be at least ε apart under this metric.) The set of directions is $(k-1)$-dimensional, so we'll have $\mathcal{D} \asymp (1/\varepsilon)^{k-1}$.

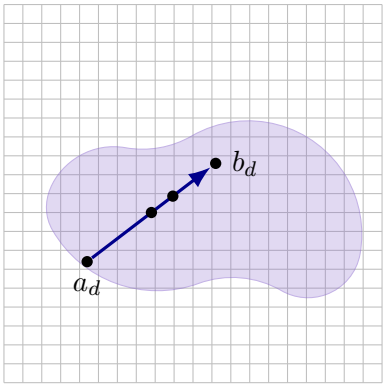




We want to get some set on which we can apply the sum-difference problem, and we'll do so by looking at the line segments inside K in just the directions given by $\mathcal{D}$, and taking appropriate linear combinations. Specifically, for each direction $d \in \mathcal{D}$, let a_d and b_d be the endpoints of a unit line segment in K with direction d. (Technically, since we scaled K down, we really only know that it has a line segment in each direction of some constant length — not necessarily unit length — but this doesn't matter.) We then consider the linear combinations

$$\frac{a_d + r b_d}{r+1} \text{ for } r \in [m],$$

which give us some m points on the line segment $[a_d, b_d]$.



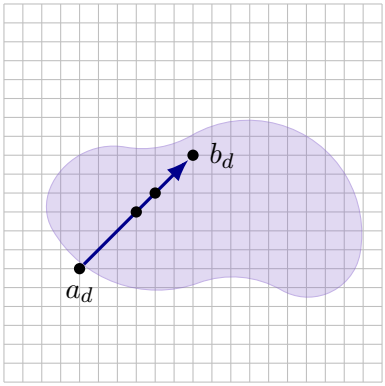
We'd like to lower-bound the number of these points, and relate this number to the covering number $b_\varepsilon(K)$ of our set. For this, things would work out most nicely if these points were all grid points, and we can actually ensure this is the case by wiggling around our line segment $[a_d, b_d]$ a bit (possibly moving them outside of K).

Claim 2.8 — For each d, we can move a_d and b_d by $O_m(\varepsilon)$ to produce points a'_d and b'_d such that a'_d, b'_d, and all the linear combinations

$$\frac{a'_d + rb'_d}{r+1}$$

(for $r \in [m]$) are grid points.

Proof sketch. We can first move a_d by at most roughly ε so that it becomes a grid point. Then we can move b_d by at most roughly $(m+1)!\,\varepsilon$ so that b'_d is also on the grid and the number of grid steps we take to go from a'_d to b'_d (in each direction) is a multiple of $2, \dots, m+1$ (i.e., $b'_d - a'_d$ is a multiple of $(m+1)!\,\varepsilon$). This ensures all the linear combinations we're working with lie on the grid as well.



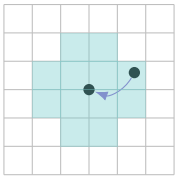
(The dependence on m is huge, but this is fine — all that matters is the dependence on ε.) □

Now we define K' to be the set of all these grid points — i.e., we let

$$K' = \left\{ a'_d, b'_d, \frac{a'_d + rb'_d}{r+1} \mid d \in \mathcal{D},\, r \in [m] \right\}.$$

Then since we obtained K' by taking a bunch of points in K and just moving them over by $O_m(\varepsilon)$, we have $|K'| \asymp_{m,k} b_\varepsilon(K)$. This is because if we've moved all our points by at most $c\varepsilon$, then each point in K' comes from a point in K which is in a grid cube at most c steps away, and there's a constant number of such boxes (roughly $(2c)^k$) — in particular, only a constant number of boxes in the covering of K can be collapsed to

the same point in K'. (This is why $|K'| \gtrsim_{m,k} b_\varepsilon(K)$, but that's the direction we need; the other direction can be shown similarly.)



Now K' is a finite set of points, and we'll apply the sum-difference problem to it — let

$$S = \{(a'_d, b'_d) \mid d \in \mathcal{D}\} \subseteq \mathbb{R}^k \times \mathbb{R}^k.$$

Then for each $r \in [m]$ we have

$$\pi_r(S) = \#\{a'_d + rb'_d \mid d \in \mathcal{D}\} \leq |K'|$$

(since K' contains a fixed multiple of each of the points $a'_d + rb'_d$). Furthermore, since the directions $d \in \mathcal{D}$ are sufficiently separated, the values of $b'_d - a'_d$ are all distinct — this is because $b'_d - a'_d$ is very close to $b_d - a_d$, which has direction d. (To be more precise, we've moved each $b_d - a_d$ by $O_m(\varepsilon)$ to get $b'_d - a'_d$, so we actually need to take the directions $d \in \mathcal{D}$ to be $\Theta_m(\varepsilon)$-separated — not just ε-separated — but this doesn't affect the rest of the argument.)

So then the statement $\mathsf{SD}(\{1, \dots, m\}, \alpha)$ (which we assumed is true) implies that $|\mathcal{D}| \leq |K'|^\alpha$ (since we have one element of S for each direction $d \in \mathcal{D}$). And $|\mathcal{D}| \asymp_{m,k} (1/\varepsilon)^{k-1}$ (since $\mathcal{D}$ is an $\Theta_m(\varepsilon)$-separated set in $k-1$ dimensions), so we get that $(1/\varepsilon)^{k-1} \lesssim_{m,k} |K'|^\alpha$, and therefore $|K'| \lesssim_{m,k} (1/\varepsilon)^{(k-1)/\alpha}$. And finally, since $|K'| \asymp_{m,k} b_\varepsilon(K)$, we get that $b_\varepsilon(K) \gtrsim_{m,k} (1/\varepsilon)^{(k-1)/\alpha}$ (as $\varepsilon \to 0$), so $\dim(K) \geq (k-1)/\alpha$.

Now we're almost done; but we wanted $\dim(K) \geq k/\alpha$, so it just remains to get rid of the extra $1/\alpha$. And this is not hard — consider the t-fold product $K^t = K \times \cdots \times K \subseteq \mathbb{R}^{kt}$. This is a Kakeya set in kt dimensions, and we'll have $\dim(K^t) = t \dim(K)$. Then the above proof applied to K^t gives that $\dim(K^t) \geq kt/\alpha - 1/\alpha$, and taking $t \to \infty$ gives that $\dim(K) \geq k/\alpha$. □

So bounds for the sum-difference problem also give bounds for the Kakeya problem, which is why Bourgain and Katz–Tao were working on this problem.

§3 Proving sum-difference bounds

We'll now discuss how to prove bounds for the sum-difference problem. We'll prove the bound with $\alpha = 7/4$ (as in Theorem 1.11, but possibly with different slopes), and then sketch how to improve it to reach $\alpha = 1 + 1/\sqrt{2} \approx 1.7$. (For comparison, the best bound — in Theorem 1.12 — is $\alpha \approx 1.67$.) This will illustrate the main ideas of the Katz–Tao arguments; we'll then see that even if we could take these arguments to their limit, there's no hope of getting all the way to $\alpha = 1$.

§3.1 Some setup

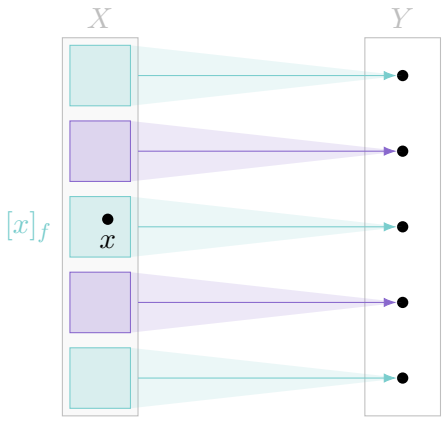
We'll first set up a few pieces of notation and facts that will be useful for the proof. We'll state things pretty generally here — imagine we've got two finite sets X and Y, and a function $f: X \to Y$. We'll think of f as defining an equivalence relation on X.

Definition 3.1. We write $x_1 \sim_f x_2$ to denote that $f(x_1) = f(x_2)$.

We can imagine f chops up X into a bunch of parts; we'll use the following notation to refer to these parts.

Definition 3.2. Let $x \in X$. We use $[x]_f$ to denote the equivalence class of x — i.e.,

$$[x]_f = \{v \in X \mid f(v) = f(x)\}.$$



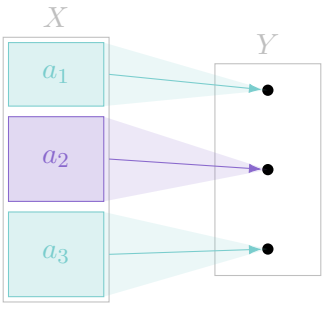
We'll use two simple facts. The first gives a lower bound for the size of the 'collision set' of f (the number of pairs (x_1, x_2) with $x_1 \sim_f x_2$).

Lemma 3.3

For any $f\colon X \to Y$, we have

$$\#\{(x_1, x_2) \in X \times X \mid x_1 \sim_f x_2\} \geq \frac{|X|^2}{|Y|}.$$

Proof. This follows from Cauchy–Schwarz — let $a_1, \ldots, a_m$ be the sizes of the equivalence classes of X (i.e., the number of elements mapped to each $y \in Y$).



Then by Cauchy–Schwarz we have

$$\#\{(x_1, x_2) \in X \times X \mid x_1 \sim_f x_2\} = \sum a_i^2 \geq \frac{(\sum a_i)^2}{|Y|} = \frac{|X|^2}{|Y|}.$$ □

The proof will involve taking refinements where we throw away some points in X and only keep the ones whose parts (under a certain function f) are not too small; the other fact bounds how much we lose by doing so.

Definition 3.4. For $f: X \to Y$, we define the refinement of X corresponding to f as

$$X^{(f)} = \left\{ x \in X \mid |[x]_f^X| \geq \frac{|X|}{2\,|Y|} \right\}.$$

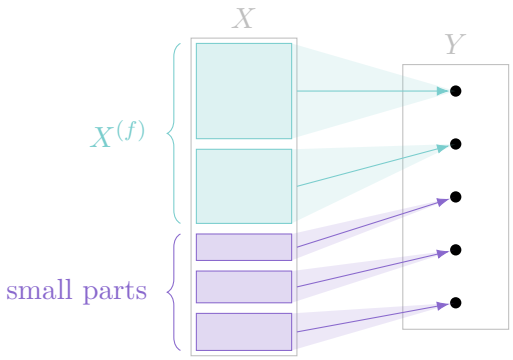
Lemma 3.5

For any $f: X \to Y$, we have $|X^{(f)}| \geq \frac{1}{2}\,|X|$.

Proof. This follows from Markov (or an averaging argument) — if we consider all the equivalence classes $|x|_f^X$ which are too small (and therefore not included in $X^{(f)}$), their total size is at most

$$|Y| \cdot \frac{|X|}{2\,|Y|} = \frac{1}{2}\,|X|$$

(since there's at most $|Y|$ equivalence classes in total).



So the refinement throws away at most $\frac{1}{2}\,|X|$ elements, which means $|X^{(f)}| \geq \frac{1}{2}\,|X|$. □

§3.2 Defining the slopes

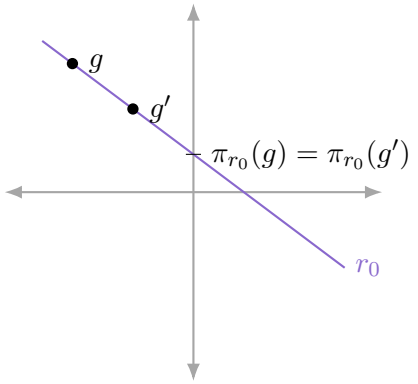
In order to prove that $\mathsf{SD}(R, \alpha)$ holds (where $\alpha = 7/4$, and R is some set of slopes), the intuition is that if our set G is too large, then the condition that the set $\pi_r(G) = \{g_1 + rg_2 \mid (g_1, g_2) \in G\}$ is not too large means that there's lots of collisions. And we're going to take these collisions along various slopes $r \in R$ and use linear algebra to generate a collision along the slope -1 (i.e., two points with the same value of $g_1 - g_2$), which will contradict the assumption that π_{-1} is injective on G.

To get these cancellations to happen, we'll want the slopes we use to be 'nice' in some way. So we'll start by making some definitions regarding our slopes.

First, we're going to take a special slope r_0 (which we can think of as 0; its value won't matter for this argument, but we leave it as a variable so that we'll be able to change its value when discussing improvements). We'll then define

$$V = \{(g, g') \in G \times G \mid \pi_{r_0}(g) = \pi_{r_0}(g')\}. \tag{1}$$

In words, V is the set of pairs of points in G which are on the same line of slope r_0 (for example, if $r_0 = 0$, then it's the set of pairs of points which are vertically on top of each other).



Note that Lemma 3.3 means V is reasonably large (here we're taking X to be G, f to be π_{r_0}, and Y to be $\pi_{r_0}(G)$ — V is by definition the collision set of π_{r_0} on G) — specifically, we have

$$|V| \geq \frac{|G|^2}{|\pi_{r_0}(G)|} \geq \frac{|G|^2}{n}. \tag{2}$$

We're also going to have a special slope r_∞ (which we can think of as ∞). We'll also have some constant s (which we'll later fix to be a generic constant, so that a certain determinant doesn't vanish).

To capture what's happening with cancellation along the slope -1, we'll define a function ν that combines information about one point along the slope r_∞ and another along the slope -1.

Definition 3.6. We define the function $\nu\colon G \times G \to W$ (on an input $(g, g') \in G \times G$) as

$$\nu(g, g') = s \cdot \pi_{r_\infty}(g) + \pi_{-1}(g').$$

Note that $\pi_r\colon (g_1, g_2) \mapsto g_1 + rg_2$ is a map $W \times W \to W$, so we can view it as a map $G \to W$; and here we're taking a *pair* of points in G and applying different functions π_r to each (and combining their results). We'll also use the following notation for convenience (to apply different functions π_r to two points).

Definition 3.7. For slopes r_1 and r_2, we define $\pi_{r_1 \otimes r_2}\colon G \times G \to W \times W$ as

$$\pi_{r_1 \otimes r_2}(g, g') = (\pi_{r_1}(g), \pi_{r_2}(g')).$$

Then the sense in which we'll want cancellation is the following.

Lemma 3.8

For any slope $r \notin \{-1, r_0, r_\infty\}$, there exists a slope r' such that $\pi_{r \otimes r'}$ determines ν over V.

What does this mean? We originally defined ν as a function on $G \times G$, so we can also view it as a function on $V \subseteq G \times G$. And this statement means that the value of $\pi_{r \otimes r'}$ on a point $(g, g') \in V$ is enough to figure out the value of ν on this point — in other words, if we know that (g, g') is in V and we also know $\pi_r(g)$ and $\pi_{r'}(g')$, then we can figure out $\nu(g, g')$.

Proof. We want to choose r' such that if we're given $\pi_r(g)$ and $\pi_{r'}(g')$, and that $\pi_{r_0}(g) = \pi_{r_0}(g')$, then we can compute $\nu(g, g') = s\pi_{r_\infty}(g) + \pi_{-1}(g')$. And these are all linear functions of g and g', so in order to do this, we want to be able to find constants x, y, and z (depending on the slopes, but not g and g') such that

$$s\pi_{r_\infty}(g) + \pi_{-1}(g') = x \cdot \pi_r(g) + y \cdot \pi_{r'}(g') + z \cdot (\pi_{r_0}(g) - \pi_{r_0}(g')). \tag{3}$$

(This is what we mean by 'cancellation.') We can imagine writing $g = (g_1, g_2)$ and $g' = (g_1', g_2')$ and expanding out the definition of π_r (as $\pi_r(g_1, g_2) = g_1 + rg_2$) for each term in (3); then the equation we get is

$$(s - x - z)g_1 + (sr_\infty - rx - r_0 z)g_2 + (1 - y + z)g_1' + (-1 - r'y + r_0 z)g_2' = 0.$$

We want this to hold for all g_1, g_2, g_1', and g_2', so we want each of these coefficients to be 0 — this gives the system of four equations

$$\begin{aligned} 0 &= s - x - z \\ 0 &= sr_\infty - rx - r_0 z \\ 0 &= 1 - y + z \\ 0 &= -1 - r'y + r_0 z \end{aligned}$$

(where we think of the variables as x, y, and z). This is a system of four equations in three variables, so it has a solution if and only if the corresponding determinant vanishes — i.e.,

$$\begin{vmatrix} s & -1 & 0 & -1 \\ sr_\infty & -r & 0 & -r_0 \\ 1 & 0 & -1 & 1 \\ -1 & 0 & -r' & r_0 \end{vmatrix} = 0.$$

And finally, given r, this gives some linear equation for r', which we can solve (this linear equation might not have a solution if s is some weird value such that the coefficient of r' is 0, but if we choose s generically then this won't happen, and we'll be able to solve for r'). □

Finally, we'll use the following set of slopes.

Theorem 3.9

For any slopes r_1 and r_2 (which are distinct from each other, as well as r_0 and r_∞), if we let r_1' and r_2' be as in Lemma 3.8, then $\mathsf{SD}(\{r_0, r_1, r_1', r_2, r_2', r_\infty\}, 7/4)$ holds.

§3.3 Proof of Theorem 3.9

We'll now prove the bound of $\alpha = 7/4$ (with slopes defined as in Theorem 3.9). This means we're given that

$$|\pi_r(G)| \le n \text{ for all } r \in \{r_0, r_1, r_1', r_2, r_2', r_\infty\}$$

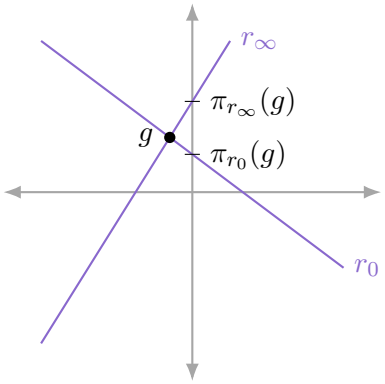
and that π_{-1} is injective on G, and we want to prove that $|G| \le n^{7/4}$.

Claim 3.10 — If we're given the values of $\nu(g, g')$ and $\pi_{r_\infty}(g)$ for some $(g, g') \in V$, then we can recover the points g and g'.

(Recall that we defined V in (1) as the 'collision set' of G along π_{r_0} — i.e., the set of pairs $(g, g') \in G \times G$ with $\pi_{r_0}(g) = \pi_{r_0}(g')$.)

Proof. First, if we know both $\nu(g, g')$ — which is defined as $s\pi_{r_\infty}(g) + \pi_{-1}(g')$ — and $\pi_{r_\infty}(g)$, then we can recover the value of $\pi_{-1}(g')$. Then, since π_{-1} is injective on G, we can recover the value of g'.

Now since $(g, g') \in V$, we know that $\pi_{r_0}(g) = \pi_{r_0}(g')$. And we know g', so we can compute $\pi_{r_0}(g')$; this lets us find $\pi_{r_0}(g)$. Finally, this means we have both $\pi_{r_0}(g)$ and $\pi_{r_\infty}(g)$, which means we have information about g in two directions; this is enough to recover g.



So we've recovered both g' and g, as desired. □

This is useful because it means any two pairs $(g, g') \in V$ with the same values of $\nu(g, g')$ must have different values of $\pi_{r_\infty}(g)$. And there's at most n possible values, so we get that

$$|[x]_\nu| \leq n \text{ for all } x \in V. \tag{4}$$

(We use $[x]_\nu$ to denote the equivalence class of x in V under ν.) We'll make use of this later.

Claim 3.11 — The functions $\pi_{r_1 \otimes r_1'}$ and $\pi_{r_2 \otimes r_2'}$ together are enough to parametrize all of $G \times G$ — i.e., if we're given $\pi_{r_1 \otimes r_1'}(g, g')$ and $\pi_{r_2 \otimes r_2'}(g, g')$ for some $(g, g') \in G \times G$, then we can recover (g, g').

Proof. By definition $\pi_{r \otimes r'}(g, g') = (\pi_r(g), \pi_{r'}(g'))$, so we're given $\pi_{r_1}(g)$ and $\pi_{r_2}(g)$, and information about g in two directions is enough to recover g; and similarly we're given $\pi_{r_1'}(g')$ and $\pi_{r_2'}(g')$, which is enough to recover g'. □

Now we'll take V and refine it based on first $\pi_{r_1 \otimes r_1'}$ (to get a set $V' \subseteq V$) and then $\pi_{r_2 \otimes r_2'}$ (to get a set $V'' \subseteq V'$) — so the equivalence class of every $x \in V''$ under $\pi_{r_2 \otimes r_2'}$ in V' is large, and the equivalence class of every $y \in V'$ under $\pi_{r_1 \otimes r_1'}$ in V is large. Then Lemma 3.5 guarantees that $|V''| \geq \frac{1}{2}|V'| \geq \frac{1}{4}|V|$.

Now we consider some $x \in V''$ and look at the set

$$T_x = \{(y, z) \in V \times V \mid \pi_{r_2 \otimes r_2'}(x) = \pi_{r_2 \otimes r_2'}(y),\ \pi_{r_1 \otimes r_1'}(y) = \pi_{r_1 \otimes r_1'}(z)\}.$$

So in other words, we're considering pairs where y is in the same equivalence class as x under $\pi_{r_2 \otimes r_2'}$, and z is in the same equivalence class as y under $\pi_{r_1 \otimes r_1'}$.

Claim 3.12 — We have $|T_x| \gtrsim |V|^2 / n^4$.

Proof. Imagine that we restrict y to be in V' (rather than just V — of course this can only shrink T_x). Then the number of choices we have for y is the size of the equivalence class of x in V' under $\pi_{r_2 \otimes r_2'}$, which we guaranteed to be large by taking x from V'' (which was defined as the refinement of V' under this function). Specifically, the number of possible values of $\pi_{r_2 \otimes r_2'}$ is at most n^2 (since $\pi_{r_2 \otimes r_2'}(g, g') = (\pi_{r_2}(g), \pi_{r_2'}(g'))$, and there's at most n possibilities for each coordinate), so we get

$$\#\{y \in V' \mid \pi_{r_2 \otimes r_2'}(x) = \pi_{r_2 \otimes r_2'}(y)\} \geq \frac{|V'|}{2n^2} \geq \frac{|V|}{4n^2}.$$

Then for each such y, the number of choices for z is the size of the equivalence class of y in V under $\pi_{r_1 \otimes r_1'}$, which we guaranteed to be large by taking y from V' (which is the refinement of V under this function) — again the range of $\pi_{r_1 \otimes r_1'}$ has size at most n^2, so we have

$$\#\{z \in V \mid \pi_{r_1 \otimes r_1'}(y) = \pi_{r_1 \otimes r_1'}(z)\} \geq \frac{|V|}{2n^2}.$$

Multiplying these two bounds gives the desired result. □

On the other hand, we can also get an *upper* bound on $|T_x|$.

Claim 3.13 — We have $|[x]_\nu| \geq |T_x|$.

Proof. First we'll show that if $(y, z) \in T_x$, then y and z are both in $[x]_\nu$ (i.e., $\nu(x) = \nu(y) = \nu(z)$). For y, we know that x and y have the same values of $\pi_{r_1 \otimes r_1'}$, and we defined r_1' such that $\pi_{r_1 \otimes r_1'}$ determines ν on V (as in Lemma 3.8), so then x and y must also have the same values of ν. Similarly, y and z have the same values of $\pi_{r_2 \otimes r_2'}$, and since $\pi_{r_2 \otimes r_2'}$ determines ν on V, they must also have the same values of ν.

Now we'll show that each choice of z corresponds to at most one choice of y, which will give the desired bound (since we have to choose $z \in [x]_\nu$, and then there's only one way to choose y). The point is that once we've fixed z, we know the values of both $\pi_{r_1 \otimes r_1'}$ and $\pi_{r_2 \otimes r_2'}$ on y (the first matches that of z, and the second matches that of x). But by Claim 3.11, this is enough information to determine y. □

And now we're essentially done — combining (4) with Claims 3.12 and 3.13 gives

$$n \geq |[x]_\nu| \geq |T_x| \geq \frac{|V|^2}{n^4},$$

which means $|V| \leq n^{5/2}$. And on the other hand, we saw in (2) that

$$|V| \geq \frac{|G|^2}{n}$$

(this came from Cauchy–Schwarz), so we get $|G| \leq n^{7/4}$.

§3.4 Ideas behind better bounds

We'll now briefly discuss how we improve this argument to get the better bound of $\alpha = 1 + 1/\sqrt{2}$. In the argument here, we had six slopes — the special slopes r_0 and r_∞, and two arbitrary slopes r_1 and r_2 and their 'duals' r_1' and r_2'. And what was important about having *two* slopes r_1 and r_2 was in some sense that π_{r_1} and π_{r_2} together parametrize $W \times W$ (which gave Claim 3.11); we can roughly think of this as corresponding to the statement $\mathsf{SD}(\{r_1, r_2\}, 2)$.

And so we can imagine iterating this argument — suppose we know that $\mathsf{SD}(R, \beta)$ holds (for some set of slopes R and some β). Then we can choose new values of r_0, r_∞, and s and define R' as the dual of R with respect to these new values, and show that then $\mathsf{SD}(\{r_0, r_\infty, R, R'\}, (4\beta - 1)/2\beta)$ holds (by a similar argument to the one we had here). We can keep doing this repeatedly; this gives us some recursion whose fixed point is $1 + 1/\sqrt{2}$.

To improve the bound to $\alpha \approx 1.67$, we do something similar, but with a bigger 'graph.' What does this mean? In our argument, we looked at finitely many slopes and tried to get some statement where if we have two things sharing values along all these slopes, then we can get a collision with -1. This can be interpreted as a constant-sized graph; and the proof of $\alpha \approx 1.67$ uses a bigger graph.

But arguments of this form can't get all the way to $\alpha = 1$ — Katz showed that they can't beat $\alpha = 1.5$. More specifically, there exist configurations that avoid all these constant-sized graphs, but have $n^{1.5}$ points. So in order to get past $\alpha = 1.5$, at some point we'll need to look at more than finitely many points at a time when looking for collisions.