Equation Systems and Roots

Yanheng Wang

1 Linear System

This section exploits a simple fact from linear algebra:

> Every linear system $A\boldsymbol{x} = \mathbf{0}$ with more variables than equations has a non-trivial solution.

1.1 Equal unions and intersections

Suppose we are given sets $S_1,\ldots,S_m \subseteq [n]$. If m is large enough (say $m \geqslant 2^n$) then it is always possible to select two groups $I \uplus J$ such that $\bigcup_{i\in I} S_i = \bigcup_{j\in J} S_j$. But this bound is a gross overestimate. We will prove that $m \geqslant n+1$ suffices!

Suppose $m \geqslant n+1$. Let $\boldsymbol{v}_i \in \mathbb{R}^n$ be the characteristic vector of S_i. Then the linear system

$$\sum_{i=0}^{m} x_i \boldsymbol{v}_i = \mathbf{0}$$

has a non-trivial solution $\boldsymbol{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$. Therefore the sets $I := \{i \in [m] : x_i > 0\}$ and $J := \{j \in [m] : x_j < 0\}$ are both non-empty. Then we have

$$\sum_{i\in I} x_i \boldsymbol{v}_i = \sum_{j\in J} (-x_j)\, \boldsymbol{v}_j \quad =: \boldsymbol{v}$$

and the coefficients on both sides are, by definition, strictly positive. (They could be fractional numbers, of course.) Now observe that $\boldsymbol{v}(k) = 0$ iff $\boldsymbol{v}_i(k) = 0$ for all $i \in I$, namely $k \notin \bigcup_{i\in I} S_i$. By the same reasoning, $\boldsymbol{v}(k) = 0$ iff $k \notin \bigcup_{j\in J} S_j$. Therefore the two unions must coincide.

The proof can actually be strengthened to take care of intersections at the same time – given that $m \geqslant n+2$. To see this, we invent new vectors $\boldsymbol{u}_i := \begin{pmatrix} \boldsymbol{v}_i \\ 1 \end{pmatrix}$ and replay the game. This time we get

$$\begin{aligned} \sum_{i\in I} x_i \boldsymbol{v}_i &= \sum_{j\in J} (-x_j)\, \boldsymbol{v}_j \quad =: \boldsymbol{v} \\ \sum_{i\in I} x_i &= \sum_{j\in J} (-x_j) \quad =: \alpha \end{aligned}$$

We construct the indices I and J as before. The equal-union property holds as before, and the additional formula further implies $\bigcap_{i\in I} S_i = \bigcap_{j\in J} S_j$! The argument is: $\boldsymbol{v}(k) = \alpha$ iff $\boldsymbol{v}_i(k) = 1$ for all $i \in I$, namely $k \in \bigcap_{i\in I} S_i$. The same applies to the J part.

Thinking over again, we can interpret the proof geometrically. The characteristic vectors are lying on an n-dimensional hypercube. Taking (non-degenerate) conic combination of vectors in which some component is absent corresponds to moving on a face of the hypercube.

It is thus not surprising to see in the next section that the idea acts well on high-dimensional geometry problems.

1.2 Classical theorems in discrete geometry

Lemma. (Radon) Let $P = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\} \subseteq \mathbb{R}^d$ be a point set. If $m \geqslant d+2$ then we may partition $P = Q \uplus R$ so that $\operatorname{conv}(Q) \cap \operatorname{conv}(R) \neq \emptyset$.

The proof is a repetition of the equal-union-intersection proof: Simply normalise the shared vector $\boldsymbol{v}$ by the shared constant α to obtain *convex* coefficients!

The lemma below displays the same argument yet once more:

Lemma. Let $S_1, \dots, S_m \subseteq \mathbb{R}^d$ be point sets with $S_i := \{\boldsymbol{u}_i, \boldsymbol{v}_i\}$. If $m \geqslant d+1$ then there is a way to colour one point black and the other white for each S_i, such that $\operatorname{conv}\{\text{black points}\} \cap \operatorname{conv}\{\text{white points}\} \neq \emptyset$.

Proof. When $m \geqslant d+1$, the linear system $\sum_{i=1}^m x_i(\boldsymbol{u}_i - \boldsymbol{v}_i) = \boldsymbol{0}$ has a non-trivial solution $(x_1, ..., x_m)$. As usual, let $I := \{i \in [m] : x_i > 0\}$ and $J := \{j \in [m] : x_j < 0\}$ which are non-empty. Then

$$\sum_{i \in I} x_i \boldsymbol{u}_i + \sum_{j \in J} (-x_j)\boldsymbol{v}_j = \sum_{i \in I} x_i \boldsymbol{v}_i + \sum_{j \in J} (-x_j)\boldsymbol{u}_j.$$

So naturally we colour points on the left black, and the points on the right white. Normalising the above equation by the constant $\sum_{i \in I} x_i + \sum_{j \in J}(-x_j)$ gives the lemma. □

Theorem. (Helly) Let $C_1, \dots, C_m$ be convex sets in $\mathbb{R}^d$ where $m \geqslant d+2$. If any $d+2$ of them intersect, then all of them intersect.

Proof. By induction on m. The base case $m = d+2$ is vacuously true. Now we go from m to $m+1$. For each $i \in [m+1]$, there exists a point $\boldsymbol{v}_i \in \bigcap_{k \neq i} C_k$ by induction hypothesis. So we may apply Radon's lemma to partition $[m] = A \uplus B$ such that there exists $\boldsymbol{y} \in \operatorname{conv}(\boldsymbol{v}_A) \cap \operatorname{conv}(\boldsymbol{v}_B)$.

Now we observe that $\boldsymbol{v}_A \subseteq \bigcap_{k \in B} C_k$ and $\boldsymbol{v}_B \subseteq \bigcap_{k \in A} C_k$ by definition of the points. Since the C_k's are convex, we must have $\operatorname{conv}(\boldsymbol{v}_A) \subseteq \bigcap_{k \in B} C_k$ and $\operatorname{conv}(\boldsymbol{v}_B) \subseteq \bigcap_{k \in A} C_k$. Therefore we may conclude $\boldsymbol{y} \in (\bigcap_{k \in B} C_k) \cap (\bigcap_{k \in A} C_k) = \bigcap_{k=1}^{m+1} C_k$, as desired. □

1.3 Tverberg's theorem*

Tverberg's theorem is a full generalisation of Radon's lemma:

Theorem. (Tverberg) Let $P = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_m\} \subseteq \mathbb{R}^d$ be a point set. If $m \geqslant (r-1)(d+1)+1$ then we may partition $P = \biguplus_{i=1}^r P_i$ so that $\bigcap_{i=1}^r \operatorname{conv}(P_i) \neq \emptyset$.

As in the proof of Radon's lemma, we augment each point $\boldsymbol{v}_i \in P$ into $\boldsymbol{u}_i := \begin{pmatrix} \boldsymbol{v}_i \\ 1 \end{pmatrix} \in \mathbb{R}^{d+1}$ to "synchronise" their coefficients in a linear equation. But we are facing a technical complication. How do we split $r > 2$ sets by "reading" the solution to the equation? In Radon's lemma we looked at the signs of coefficients, but now we need a new machinery to read out r parallel tracks.

To address the issue, we will *expand* the point set by associating each $\boldsymbol{u}_i$ with r "shadow points" $\{\boldsymbol{u}_i^j\}_{j=1}^r$ living in higher-dimensional space $\mathbb{R}^{(r-1)(d+1)}$. Then we work our way back, namely for each $\boldsymbol{u}_i$ we carefully *choose* exactly one of its r shadow points $\boldsymbol{u}_i^j$. The $j \in [r]$ that we chose encodes the partition that $\boldsymbol{v}_i$ belongs to.

First we elucidate the expansion. The r shadow points of $\boldsymbol{u}_i$ are given by

$$\boldsymbol{u}_i^1 := \begin{pmatrix} \boldsymbol{u}_i \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{u}_i^2 := \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{u}_i \\ \vdots \\ \boldsymbol{0} \end{pmatrix}, \quad \dots, \quad \boldsymbol{u}_i^{r-1} := \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{u}_i \end{pmatrix}, \quad \boldsymbol{u}_i^r := -\begin{pmatrix} \boldsymbol{u}_i \\ \boldsymbol{u}_i \\ \vdots \\ \boldsymbol{u}_i \end{pmatrix} \quad \in \mathbb{R}^{(r-1)(d+1)}$$

where we divided the $(r-1)(d+1)$ coordinates into $r-1$ "tracks" each containing $d+1$ coordinates. Note that $\sum_{j=1}^{r} \boldsymbol{u}_i^j = \mathbf{0}$, so in particular $\mathbf{0} \in \operatorname{conv}\{\boldsymbol{u}_i^1, \dots, \boldsymbol{u}_i^r\} =: \operatorname{conv}(S_i)$.

Next we choose one shadow point for each $i \in [m]$. The following lemma is crucial but kind of obvious. Its proof relies upon polytope theory; for a quick recap see https://yanhengwang.github.io/notes/geometry.

Lemma. (Barany) Suppose we have sets $S_1, \dots, S_{k+1} \subseteq \mathbb{R}^k$ with $\mathbf{0} \in \operatorname{conv}(S_i)$ for all i. Then it is always possible to pick one point $\boldsymbol{s}_i$ from each S_i such that $\mathbf{0} \in \operatorname{conv}\{\boldsymbol{s}_1, \dots, \boldsymbol{s}_{k+1}\}$.

Proof. We call $R \subseteq \bigcup_{i=1}^{k+1} S_i$ a *rainbow set* if $|R \cap S_i| = 1$ for all i. We want to prove that there exists a rainbow set R with $\mathbf{0} \in \operatorname{conv}(R)$.

Suppose to the contrary that no such rainbow set exists. Then, among all rainbow sets we choose the R that minimises $\operatorname{dist}(\mathbf{0}, \operatorname{conv}(R))$. Let $\boldsymbol{x} \in \operatorname{conv}(R)$ be the closest point to $\mathbf{0}$, which must lie on a face F of the polytope $\operatorname{conv}(R)$. Moreover, F is supported by a hyperplane H perpendicular to the line $\overline{\mathbf{0}\boldsymbol{x}}$. Assume $\operatorname{conv}(R) \subseteq H^-$ and $\mathbf{0} \in H^+$. Without loss of generality we may also assume that the vertices of F are $\boldsymbol{p}_1 \in S_1, \dots, \boldsymbol{p}_\ell \in S_\ell$ where $\ell \leqslant k$.

For all $i > \ell$, recall that $\mathbf{0} \in \operatorname{conv}(S_i)$, so there must exist $\boldsymbol{q}_i \in S_i \cap H_i^+$. Then we consider the rainbow set $R' := \{\boldsymbol{p}_1, \dots, \boldsymbol{p}_\ell, \boldsymbol{q}_{\ell+1}, \dots, \boldsymbol{q}_{k+1}\}$. Clearly $\operatorname{dist}(\mathbf{0}, \operatorname{conv}(R')) < \operatorname{dist}(\mathbf{0}, \boldsymbol{x}) = \operatorname{dist}(\mathbf{0}, \operatorname{conv}(R))$, contradicting the choice of R. □

With Barany's lemma, one could choose for each $i \in [m]$ a $j(i) \in [r]$ such that $\mathbf{0} \in \operatorname{conv}\{\boldsymbol{u}_i^{j(i)}\}_{i=1}^m$. That is, $\mathbf{0} = \sum_{i=1}^m x_i \boldsymbol{u}_i^{j(i)}$ for some convex coefficients x_i's. Now let's decipher the information hidden in this $(r-1)(d+1)$-dimensional vector equation.

By construction of the shadow points, the equation naturally splits into $r-1$ parallel tracks; we denote $I_j := \{i \in [m] : j(i) = j\}$ and read from the equation

$$\sum_{i \in I_j} x_i \begin{pmatrix} \boldsymbol{v}_i \\ 1 \end{pmatrix} = \sum_{i \in I_r} x_i \begin{pmatrix} \boldsymbol{v}_i \\ 1 \end{pmatrix} =: \begin{pmatrix} \boldsymbol{v} \\ \alpha \end{pmatrix}$$

for all $j \in [r-1]$. Normalising by α, we conclude

$$\frac{1}{\alpha}\boldsymbol{v} = \sum_{i \in I_j} \frac{x_i}{\alpha} \boldsymbol{v}_i \quad \in \operatorname{conv}\{\boldsymbol{v}_i\}_{i \in I_j}$$

for all $j \in [r]$, thus $\bigcap_{j=1}^r (\operatorname{conv}\{\boldsymbol{v}_i\}_{i \in I_j}) \neq \emptyset$.

2 Polynomial System

An indispensable theme in combinatorics is "making choices". In such scenarios, a *polynomial system* comes more handy than linear systems. Luckily, just as in linear systems, we can reveal the number of solutions of a polynomial system given the comparison betwen the number of variables and the (weighted) number of equations.

Theorem. (Chevalley-Warning) Let $\mathbb{F}$ be a finite field of of characteristic p. Consider the polynomial system where $f_1, \dots, f_n : \mathbb{F}^m \to \mathbb{F}$:

$$\begin{cases} f_1(x_1, \dots, x_m) = 0, \\ \cdots\cdots \\ f_n(x_1, \dots, x_m) = 0. \end{cases}$$

If $m > \sum_{i=1}^{n} \deg(f_i)$ then the number of solution of the system is a multiple of p.

In typical applications, we design a polynomial system whose non-trivial solutions correspond to our desired objects. We also ensure that the system has a trivial solution, say $\mathbf{0}$. Then Chevalley-Warning will imply another, non-trivial solution, thus proving the existence of our desired object.

Towards the proof, we need a basic lemma about finite fields:

Lemma. Let $\mathbb{F}$ be a finite field of size q. Then $\sum_{x \in \mathbb{F}} x^k = 0$ for all $0 \leqslant k < q-1$.

Proof. Denote $s := \sum_{x \in \mathbb{F}} x^k$. Take any $y \in \mathbb{F} \setminus \{0\}$ such that $y^k \neq 1$. Such y always exists because the equation $t^k = 1$ has at most $k \leqslant q-2$ solutions. Observe $s \cdot y^k = \sum_{x \in \mathbb{F}} (x\,y)^k = \sum_{z \in \mathbb{F}} z^k = s$, namely $s \cdot (y^k - 1) = 0$. Since $y^k \neq 1$ we conclude $s = 0$. □

Proof of Chevalley-Warning Theorem. Let q be the size of the finite field $\mathbb{F}$. It is well known that $x^{q-1} = 1$ for all $x \in \mathbb{F} \setminus \{0\}$, and 0 otherwise. So the number of solutions (modulo p) of the system is counted by

$$N = \sum_{x_1, \ldots, x_m \in \mathbb{F}} \prod_{i=1}^{n} (1 - f_i^{q-1}(x_1, \ldots, x_m)).$$

Shattering the product into monomials, we get

$$\begin{aligned} N &= \sum_{x_1, \ldots, x_m \in \mathbb{F}} \sum_{k_1, \ldots, k_m} C_{k_1, \ldots, k_m} \cdot \prod_{i=1}^{n} x_i^{k_i} \\ &= \sum_{k_1, \ldots, k_m} C_{k_1, \ldots, k_m} \cdot \prod_{i=1}^{n} \left(\sum_{x_i \in \mathbb{F}} x_i^{k_i} \right) \end{aligned}$$

for some coefficients $C_{k_1, \ldots, k_m}$. By assumption of the theorem the degree of each monomial is at most $(q-1) \cdot \sum_{i=1}^{n} \deg(f_i) < (q-1)m$. In other words, $k_1 + \cdots + k_m$ is always smaller than $(q-1)m$ in the summation. Hence for each combination of $k_1, \ldots, k_m$ there always exists some $i : k_i < q-1$. By the lemma we see the inner summation $\sum_{x_i \in \mathbb{F}} x_i^{k_i}$ is zero, thus the entire product vanishes. Therefore, $N = 0$. □

2.1 Erdős-Ginzburg-Ziv theorem

Given $a_1, \ldots, a_m \in \mathbb{Z}_n$ $(m \geqslant n)$, can we always choose n of them which sum up to 0? The EGZ theorem gives an affirmative answer whenever $m \geqslant 2n-1$.

Let us first assume $n = p$ is a prime. We shall utilise Fermat's little theorem $x^{p-1} = \begin{cases} 1 & x \neq 0 \\ 0 & x = 0 \end{cases} \pmod{p}$ to model binary choices. This is an important trick that should be kept in mind.

We naturally introduce a variable x_i for each $i \in [m]$, modelling whether or not we choose a_i. Then our requirements translate to the following:

$$\begin{aligned} \sum_{i=1}^{m} x_i^{p-1} &= 0, \\ \sum_{i=1}^{m} a_i\, x_i^{p-1} &= 0. \end{aligned}$$

The sum of degrees is $2(p-1) < 2p-1 \leqslant m$. Also $\mathbf{0}$ is a trivial solution to the system. Hence by Chevalley-Warning theorem there exists a non-trivial solution $\boldsymbol{x}$, which encodes the choices we made.

Next we proceed to general n by induction. Suppose $n = s\,t$ and we have $m \geqslant 2st-1$ numbers at hand. We apply induction hypothesis to find $\{b_{11}, \ldots, b_{1s}\} \subseteq \{a_1, \ldots, a_m\}$ such that $\sum_{i=1}^{s} b_{1s} = 0 \pmod{\text{s}}$; here the numbers are interpreted as in $\mathbb{Z}_s$. Then we delete these numbers from the list and repeat the procedure. In round i we find and delete numbers $\{b_{i1}, \ldots, b_{\text{is}}\}$. The procedure lasts for $2t-1$ rounds; by then there are only $m - s\,(2t-1)$ numbers left, so we may no longer apply induction hypothesis.

For each round $i \in [2t-1]$, we define a "summary value" $b_i := \frac{1}{s}\sum_{j=1}^{s} b_{is}$. Since by construction the sum $\sum_{j=1}^{s} b_{is} = 0 \pmod{s}$, we know $b_i \in \mathbb{Z}$. We interpret it as a number in $\mathbb{Z}_t$ and apply induction hypothesis to find $I \subseteq [2t-1] : |I| = t$ such that $\sum_{i \in I} b_i = 0 \pmod{t}$. Expanding the definition:

$$0 = \sum_{i \in I} b_i = \frac{1}{s} \sum_{i \in I} \sum_{j=1}^{s} b_{is} \pmod{t}$$

Therefore $\sum_{i \in I} \sum_{j=1}^{s} b_{is} = 0 \pmod{st}$. This gives us exactly n numbers that sum up to 0 in $\mathbb{Z}_n$.

2.2 Davenport constant

Continuing the EGZ theorem but with setting slightly modified: Given $a_1, \ldots, a_m \in G$ for a group G with $|G| = n$. Can we always choose *some* (at least one) of them which sum up to 0?

The minimum m that guarantees an affirmative answer is called the *Davenport constant* of G, denoted $S(G)$.

It is relatively easy to see $S(\mathbb{Z}_n) = n$. Indeed, if someone maliciously gives us $n-1$ copies of $1 \in \mathbb{Z}_n$, then it is impossible to sum them up to 0. On the other hand, if we are given $m \geqslant n$ numbers $a_1, \ldots, a_m \in \mathbb{Z}_n$, then we append some dummy $a_{m+1}, \ldots, a_{2n-1} := 0$ and conclude from EGZ that we can choose n of them summing up to 0. Since there are less than n dummy numbers, we must have chosen at least one number from $a_1, \ldots, a_m$, as required.

Also without much effort we find $S(\mathbb{Z}_p^k) = p^k = n$. The proof of EGZ carries over here (but with the second equation only).

However, for many other groups, say $\mathbb{Z}_2 \times \mathbb{Z}_3$, the proof technique deteriorates because different coordinates do not "synchronise". Still, we may characterise the Davenport constant of a rich class of groups:

Theorem. (Olson) The Davenport constant of group $G = \prod_{i=1}^{r} \mathbb{Z}_p^{k_i}$ is $1 + \sum_{i=1}^{r} (p^{k_i} - 1)$.

The proof is beyond the scope of the note. We remark that any Abelian group could be decomposed into $\prod_{i=1}^{r} \mathbb{Z}_{n_i}$, so Olson's theorem essentially applies to *all* Abelian groups of size p^k.

2.3 4-regular graph contains 3-regular subgraph

Although this sounds obvious at first glance, it is not so trivial after careful thoughts. There is a combinatorial proof by Tashkinov in 1982. But here we present a extremely simple algebraic proof that gives a slight weakening:

> Every 4-regular (multi)graph + an edge contains a 3-regular subgraph.

The idea is straightforward: we want to choose a subset of edges such that every vertex is covered either 0 or 3 times. Hence we introduce a variable $x_e \in \mathbb{Z}_3$ for each $e \in E$, indicating if we selected edge e. Then the constraint writes

$$\sum_{e \ni v} x_e^2 = 0 \qquad \forall v \in V.$$

We remark that the equation says "every vertex is covered a multiple of 3 times". But since a 4-regular graph+an edge has maximum degree 5, the only possible coverage is 0 or 3.

The system has $\frac{4n}{2}+1=2n+1$ variables, and the total degree is $2n$. Also $\mathbf{0}$ is a trivial solution. So applying Chevalley-Warning finishes the proof.

3 Roots of a Single Polynomial

As we have seen, polynomials can encode combinatorial constructions in their roots. Chevalley-Warning theorem helps extract these constructions. But it has weaknesses. First, it applies only to finite fields. Second, it entails an undesirable "modulo characteristic p" even if the field has size p^{100}, say.

This and the next sections will develop more general tools in polynomial theory. In particular, we elucidate the relations between the degree and the number of roots of a single polynomial – either univariate or multivariate, in finite fields or in infinite fields. These tools are not only usable in *constructing objects* but also in *deriving bounds*.

The fundamental theorem of algebra states that a non-trivial *univariate* polynomial of degree d has at most d roots. (If the polynomial is over $\mathbb{C}$ then it has exactly d roots; but in other fields, e.g. $\mathbb{R}, \mathbb{Z}_p$, the number could be less.) In simple words: "low-degree polynomial has few roots".

This principle generalises to *multivariate* polynomials as well:

Lemma. A polynomial $f\colon \mathbb{F}_q^n \to \mathbb{F}_q$ on finite field $\mathbb{F}_q$ with $\deg(f) \leqslant d$ has at most $d\,q^{n-1}$ roots.

Proof. By induction on n. For $n=1$ the claim is clear. Now we proceed from n to $n+1$.

We decompose $f(x_1,\dots,x_n,y)=\sum_{i=0}^{k} g_i(x_1,\dots,x_n)\,y^i$ where $k\leqslant d$ is the largest degree that the variable y has. The polynomial $g_i\colon \mathbb{F}_q^n \to \mathbb{F}_q$ has degree at most $d-i$. For a particular choice of $x_1,\dots,x_n$, there are two cases:

- If $g_i(x_1,\dots,x_n)=0$ for all i, then $f(x_1,\dots,x_n,y)=0$ whatever we choose for $y\in\mathbb{F}_q$. This case happens $\leqslant (d-k)\,q^{n-1}$ times by induction hypothesis (considering $g_k(x_1,\dots,x_n)=0$ alone). So it contributes at most $(d-k)\,q^n$ roots of f.
- Otherwise, $f(x,\dots,x_n,y)$ is a non-trivial univariate polynomial, hence having $\leqslant k$ roots. This case happens (trivially) $\leqslant q^n$ times. So it contributes $k\,q^n$ roots of f.

Summing these up we have at most $d\,q^n$ roots, finishing the induction. □

3.1 Finite field Kakeya set

$S\subseteq\mathbb{F}_q^n$ is called a *Kakeya set* if S contains a "line" in every "direction". Formally, for any $\boldsymbol{u}\in\mathbb{F}_q^n\setminus\{0\}$, there exists $\boldsymbol{b}\in\mathbb{F}_q^n$ such that $\boldsymbol{b}+t\,\boldsymbol{u}\in S$ for all $t\in\mathbb{F}_q$.

How small can a Kakeya set be? Dvir proved the following lower bound: $|S|\geqslant\binom{n+q-1}{q-1}$.

Suppose to the contrary that $|S|<\binom{n+q-1}{q-1}$. Intuitively, it is possible to cook up a low-degree polynomial that vanishes on S because S is small. Indeed, there are $\binom{n+q-1}{q-1}$ monomials on n variables of degree $\leqslant q-1$. So we can solve non-trivial coefficients from the linear equation

$$\sum_{k_1+\cdots+k_n\leqslant q} C_{k_1,\dots,k_n}\cdot x_1^{k_1}x_2^{k_2}\cdots x_n^{k_n}=0 \qquad \forall (x_1,\dots,x_n)\in S.$$

Hence, there exists a polynomial $f: \mathbb{F}_q^n \to \mathbb{F}_q$ of degree $d < q$ that vanishes on the entire S.

Next we use the Kakeya property to deduce that f must vanish everywhere, which leads to a contradiction.

Write $f =: g_1 + \cdots + g_d$ where g_i groups together all terms of degree i. Fix an arbitrary $\boldsymbol{u} \in \mathbb{F}_q^n \setminus \{\mathbf{0}\}$. By the Kakeya property, we may find $\boldsymbol{b} \in \mathbb{F}_q^n$ such that $h(t) := f(\boldsymbol{b} + t\,\boldsymbol{u})$ vanishes for all $t \in \mathbb{F}_q$. In other words, h is identically zero. Note that the $h(t) = g_d(\boldsymbol{u})\, t^d + \cdots$. Since $\deg(h) = d < q$, this initial term can only be 0; otherwise h would be non-trivial and has up to $d < q$ roots, not enough to cover the entire space. So we conclude that $g_d(\boldsymbol{u}) = 0$. The argument works for every $\boldsymbol{u} \neq \mathbf{0}$. Therefore, g_d has $q^n - 1 > d\,q^{n-1}$ roots, and by the lemma it has to be trivial. But this contradicts the degree assumption $\deg(f) = d$.

3.2 Number of joints

Given a set of n lines in $\mathbb{R}^3$, a joint is where three or more non-coplanar lines intersect. How many joints can there be?

The following lemma is a stepping stone for an upper bound.

Lemma. For a set of n lines with m joints, there exists a line that contains $\leqslant 2\sqrt[3]{m}$ joints.

Assuming the lemma, we upper bound m as follows. Whenever there remains a joint, we choose the line from the lemma and remove it. The procedure lasts for at most n steps since no joint can survive after a complete removal. So $m \leqslant 2n\sqrt[3]{m}$. From this we solve $m \leqslant (2n)^{3/2}$.

Proof of the lemma. The main builk of the proof mimics the Kakeya one, though we no longer have a bound for the number of roots of the (now real-valued) multivariate polynomial.

Assume to the contrary that all lines contain $> 2\sqrt[3]{m}$ joints. The plan is to get a *minimal degree* polynomial f that vanishes on all joints. Then, since each line contains excessively many joints, we can show that f in fact vanishes on all lines. Finally, we take directional derivatives to contradict the minimality of f.

Because there are $\binom{2+2\sqrt[3]{m}}{3} > \frac{8(m-1)}{6} > m$ monomials of degree $< 2\sqrt[3]{m}$ on 3 variables, there exists a polynomial f of degree $< 2\sqrt[3]{m}$ that vanishes on all joints. We take one with minimal degree, and denote the degree $d < 2\sqrt[3]{m}$.

By our assumption, f has more than d roots along each line. Fix any line $l : \boldsymbol{a}\,t + \boldsymbol{b}\,(t \in \mathbb{R})$. Then the univariate polynomial $h(t) := f(\boldsymbol{a}\,t + \boldsymbol{b})$ has more than d roots – more than its degree! Hence h has to be trivial, and thus $f = 0$ on the any line l.

Finally we take derivatives; namely let $g := \frac{\mathrm{d}f}{\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z}$. Note that this is a polynomial with strictly smaller degree. Note that the directional derivatives of f w.r.t. the three lines defining a joint are all root, since f is constantly root on those directions. Moreover the lines are non-coplanar and thus form a basis. This means $g = 0$ on all joints, contradicting the minimality of f. □

4 Roots of Polynomial in a Box

In general we cannot bound the number of roots of a multivariate polynomial over infinite fields. Consider one of the simplest example $f(x, y) := x + y$ over $\mathbb{R}^2$; there are infinitely many roots!

However, if we focus only on a finite "box" in the field, then it is possible to reproduce our slogan "low-degree polynomial has few roots". (Check it with the polynomial above!)

Theorem. (combinatorial Nullstellensatz) Let $f : \mathbb{F}^n \to \mathbb{F}$ be a polynomial with $\deg(f) = \sum_{i=1}^{n} d_i$. If the term $\prod_{i=1}^{n} x_i^{d_i}$ has non-zero coefficient, then for any box $S_1 \times \cdots \times S_n$ with $|S_i| \geqslant d_i + 1$, there exists an $\boldsymbol{x} \in S_1 \times \cdots \times S_n$ such that $f(\boldsymbol{x}) \neq 0$.

Proof. By induction on n. The base case $n = 1$ is obvious. Now we step from n to $n+1$.

If the maximum degree of variable x_{n+1} happens to be $d_{n+1} =: \delta$ then we are basically done. The reasoning is as follows. We single out the last variable and write $f(x_1, \ldots, x_{n+1}) = \sum_{i=1}^{\delta} g_i(x_1, \ldots, x_n)\, x_{n+1}^i$. Note that $\deg(g_\delta) = \deg(f) - \delta = \sum_{i=1}^{n} d_i$. So we may apply induction hypothesis and conclude $g_\delta(x_1, \ldots, x_n) \neq 0$ for some particular choice of $x_1, \ldots, x_n$. We fix this choice and then regard f as a univariate polynomial about x_{n+1}. It has degree $\delta < |S_{n+1}|$, hence there indeed exists a value on which the polynomial does not vanish.

Now we deal with the case when the maximum degree of x_{n+1} exceeds δ. The idea is to drop the excessive degree without changing the evaluation on the box. Whenever we see an occurrence $x_{n+1}^{\delta+1}$, we replace it by $\big(x_{n+1}^{\delta+1} - \prod_{s \in S_{n+1}} (x_{n+1} - s)\big)$. Observe that the product always evaluates to zero if we plug in any $x_{n+1} \in S_{n+1}$. Moreover this operation cancels the $x_{n+1}^{\delta+1}$. If we have an occurrence like $x_{n+1}^{\delta+2}$, then we write it as $x_{n+1} \cdot x_{n+1}^{\delta+1}$ and apply the procedure. When the procedure terminates the degree of x_{n+1} is dropped to δ as desired.

One last thing to check: the term $\prod_{i=1}^{n+1} x_i^{d_i}$ itself is not killed. (If not, what will break?) Luckily this is indeed the case since we need a non-existent term $\prod_{i=1}^{n+1} x_i^{d_i} x_{n+1}$ to kill it! □

We can use this powerful theorem in two ways:

Deriving bounds. We design a polynomial that vanishes on a big box. Via the theorem, this transfers to a lower bound on the degree – typically a function of problem parameters.

Proving existence. We design a low-degree polynomial f whose roots encode structure of interest. In addition our "search space" is a box. We then use the theorem to show that f does not vanish in the box.

4.1 Cauchy-Davenport theorem

We define the sum of two sets as $A + B := \{a + b : a \in A, b \in B\}$. As a rule of thumb, $|A + B|$ is small if A and B are largely overlapping. It is easy to show $|A + B| \geqslant |A| + |B| - 1$, with equality when $A = [s]$ and $B = [t]$. Does the same bound holds in modular arithmetic?

Theorem. (Cauchy-Davenport) For sets $A, B \subseteq \mathbb{Z}_p$ we have $|A + B| \geqslant \min\{p, |A| + |B| - 1\}$.

Proof. First we consider the case $|A| + |B| \geqslant p + 1$. For any $x \in \mathbb{Z}_p$ we know $|\{x\} - A| = |A|$, hence $|\{x\} - A| + |B| \geqslant p + 1$, implying $(\{x\} - A) \cap B \neq \emptyset$. Therefore $x = a + b$ for some $a \in A, b \in B$. This proves $|A + B| = p$.

Next we may assume $|A| + |B| \leqslant p$. Suppose to the contrary that $|A + B| \leqslant |A| + |B| - 2$. Then we take a superset $C \supseteq A + B : |C| = |A| + |B| - 2$. We define a polynomial $f : \mathbb{Z}_p^2 \to \mathbb{Z}_p$ that vanishes on the whole box $A \times B$:

$$f(x, y) := \prod_{c \in C} (x + y - c).$$

Next we will show that the highest-order term $x^{|A|-1} y^{|B|-1}$ has non-zero coefficient. (There are many highest-order terms, we choose this particular one because we want to apply combinatorial Nullstellensatz.) Computing the coefficient gives $\binom{|A|+|B|-2}{|A|-1} \neq 0 \pmod{p}$. Therefore, we conclude from combinatorial Nullstellensatz that f cannot vanish on the box $A \times B$, a contradiction. □

There is many variants of the problem:

- Define $A \oplus B := \{a+b : a \in A, b \in B, a \neq b\}$. Can we find a lower bound for $|A \oplus B|$?
- What if $A \oplus B := \{a+b : a \in A, b \in B, ab \neq 1\}$?

All these variants can be answered in a similar approach, with a careful construction of the polynomial. In the first variant, for instance, we may construct $f(x,y) := (x-y)\prod_{c \in C}(x+y-c)$ to purposely zero the polynomial on the entire box $A \times B$.

4.2 Covering hypercube by hyperplanes

How many hyperplanes do we need to cover all vertices in the hypercube $\{0,1\}^n$? The answer is trivially two: the hyperplanes $x_1 = 0$ and $x_1 = 1$ will do the job, and there is no hope for less.

But let us change the statement. What if we disallow you from covering the origin? We can achieve it by n hyperplanes: $x_i = 1$ for $i \in [n]$. Is this the best we could do?

Yes, it is. Assume our hyperplanes are $H_j : \boldsymbol{a}_j^{\mathrm{T}} \boldsymbol{x} = 1$ for $j \in [m]$. We design a natural polynomial

$$f(\boldsymbol{x}) := \prod_{j=1}^{m} (\boldsymbol{a}_j^{\mathrm{T}} \boldsymbol{x} - 1) - (-1)^m \prod_{i=1}^{n} (1 - x_i)$$

which vanishes on $\{0,1\}^n$. (The first half of the definition just translates the problem constraint; the second half squeezes the origin in so that we have a complete vanishing box.) If we have too few hyperplanes (i.e. $m < n$), then the highest-order term $\prod_{i=1}^{n} x_i$ has coefficient $(-1)^{m+n+1} \neq 0$, contradicting combinatorial Nullstellensatz.

4.3 The permanent lemma

Let $A \in \mathbb{F}^{n \times n}$ with $\operatorname{per}(A) \neq 0$ and $\boldsymbol{b} \in \mathbb{F}^n$. You predefine some binary choices for each of $x_1, \ldots, x_n$, for instance $x_1 \in \{0,1\}$, $x_2 \in \{1,3\}$ and so on. No matter what options you provide, I can always choose $x_1, \ldots, x_n$ "maliciously", though conforming to your offer, such that $A\boldsymbol{x}$ and $\boldsymbol{b}$ differ in *every* coordinate!

The proof is short, given our knowledge of combinatorial Nullstellensatz. Construct polynomial $f(x_1, \ldots, x_n) := \prod_{i=1}^{n} (\boldsymbol{a}_i \boldsymbol{x} - b)$ where $\boldsymbol{a}_i$ is row i of A. The coefficient of the highest-order term, $\prod_{i=1}^{n} x_i$, is just $\operatorname{per}(A) \neq 0$, and we are done!

4.4 Existence of good permutation

Theorem. Let $A, B \subseteq \mathbb{F}_p$, $|A| = |B| = n$ and fix a permutation $(a_1, \ldots, a_n)$ of A. Then there exists a permutation $(b_1, \ldots, b_n)$ of B such that $a_i + b_i$ are distinct for all i.

Proof. Naturally, we define a polynomial that takes in a permutation of B and determines if it is good:

$$\begin{aligned} f(b_1, \ldots, b_n) &:= \prod_{1 \leqslant i < j \leqslant n} ((a_j + b_j) - (a_i + b_i)) \\ &= \operatorname{Vand}(a_1 + b_1, \ldots, a_n + a_n). \end{aligned}$$

where

$$\operatorname{Vand}(t_1, \ldots, t_n) := \begin{vmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_n \\ \vdots & \vdots & \ddots & \vdots \\ t_1^{n-1} & t_2^{n-1} & \cdots & t_n^{n-1} \end{vmatrix} = \prod_{1 \leqslant i < j \leqslant n} (t_j - t_i)$$

denotes the Vandermonde determinant on variables $t_1, \ldots, t_n$.

We want to apply combinatorial Nullstellensatz on the box B^n to detect a non-vanishing point. But be careful! The detected point might not encode a valid permutation; it could be something like $(1, 1, 4, 2, 3, 4)$. To explicitly rule out such case, we modify our polynomial into

$$f(b_1, \ldots, b_n) := \mathrm{Vand}(b_1, \ldots, b_n) \cdot \mathrm{Vand}(a_1 + b_1, \ldots, a_n + a_n),$$

so any non-vanishing point on B^n would be a valid good permutation.

Note that $\deg(f) = 2\binom{n}{2} = n(n-1)$. The remaining job is to verify the highest-degree monomial $\prod_{i=1}^{n} b_i^{n-1}$ has non-zero coefficient in f. This is as well the coefficient of the same monomial in

$$\begin{aligned}(\mathrm{Vand}(b_1, \ldots, b_n))^2 &= \left(\sum_{\pi \in S_n} (-1)^{\mathrm{sgn}(\pi)} \cdot \prod_{i=1}^{n} b_i^{\pi(i)-1}\right)^2 \\ &= \sum_{\pi \in S_n} \sum_{\sigma \in S_n} (-1)^{\mathrm{sgn}(\pi)+\mathrm{sgn}(\sigma)} \cdot \prod_{i=1}^{n} b_i^{\pi(i)+\sigma(i)-2}.\end{aligned}$$

In order to get the monomial $\prod_{i=1}^{n} b_i^{n-1}$, we must choose σ such that $\sigma(i) := n + 1 - \pi(i)$. Hence $\mathrm{sgn}(\pi) + \mathrm{sgn}(\sigma) = 0$ and the coefficient of interest is simply $\sum_{\pi \in S_n} 1 = n! \neq 0$. □