Sum-free sets.

DEFINITION 1: A subset A of an abelian group $(G, +)$ is said to be *sum-free* if $A \cap (A + A) = \phi$, in other words, if there are no solutions in A to the equation $x = y + z$.

The abelian groups which are of most interest to number theorists are $\mathbb{Z}$ and the groups $\mathbb{Z}_p$, where p is a prime.

EXAMPLE 1: Let $n \in \mathbb{N}$ and let A be a sum-free subset of $\{1, ..., n\}$. If a is the largest element of A, and

$$B := \{a - a_1 : a_1 \in A,\ a_1 \neq a\},$$

then A and B are disjoint subsets of $\{1, ..., n\}$. It follows that $|A| \leq \lceil n/2 \rceil$. There are essentially two different examples of a sum-free subset of this size, namely

$$A_1 = \{\text{odd numbers in } [1, n]\}, \quad A_2 = \left(\frac{n}{2}, n\right].$$

EXAMPLE 2: Let p be a prime, say $p = 3k + i$, where $k \in \mathbb{N}_0$ and $i \in \{0, 1, 2\}$. If $i \in \{0, 1\}$, then $A := \{k + 1, ..., 2k\}$ is a sum-free set modulo p, whereas if $i = 2$, then $A := \{k + 1, ..., 2k + 1\}$ is sum-free modulo p. Thus, if $p \equiv 2 \pmod 3$, there exists a sum-free set A in $\mathbb{Z}_p$ such that $|A| = \frac{p+1}{3}$. This is best-possible, but a proof is not as simple as in Example A. It is an easy consequence of the *Cauchy-Davenport theorem*, which is also in this week's lecture notes. We will now apply a probabilistic argument to prove the following result, which apparently was first proven by Erdős in 1965 and rediscovered by Alon and Kleitman in 1990:

Theorem 1.1. *Let S be any finite subset of $\mathbb{Z}$, not containing zero. Then there exists a sum-free subset A of S such that $|A| \geq \frac{|S|+1}{3}$.*

Proof. Let S be given and choose a prime p satisfying the following two conditions :

(i) $p > \max_{s \in S} |s|$,
(ii) $p \equiv 2 \pmod 3$.

Corollary 7.3(i) in the notes for Week 47 guarantees the existence of such a prime. Say $p = 3k + 2$ and let $C := \{k + 1, ..., 2k + 1\}$. As noted in Example 2 above, the set C is sum-free modulo p. We shall work in the probability space (Ω, μ), where $\Omega = \{1, 2, ..., p - 1\}$ and μ is uniform measure. For each $s \in S$ let $f_s : \Omega \to \Omega$ be the map given by

$$f_s : \omega \mapsto \omega s \pmod p.$$

The choice of p (property (i)) guarantees that each of the maps f_s is one-to-one. Let $X_s := \mathcal{X}_{f_s, C}$. Then for every s we have

$$\mathbb{E}[X_s] = \frac{|C|}{p - 1} > \frac{1}{3}.$$

Let $X = \sum_{s \in S} X_s$. By linearity of expectation,

$$\mathbb{E}[X] > \frac{|S|}{3}.$$

Hence there exists some $\omega \in \Omega$ such that $X(\omega) > |S|/3$. But, unwinding the definitions, we see that

$$X(\omega) = \#\{s \in S : \omega s \pmod{p} \in C\}. \tag{1.1}$$

Let A be the subset of S on the right of (1.1). This is a sum-free subset of S, since a dilation of it lies, modulo p, entirely within C, and hence is sum-free. Since $|A| > |S|/3$ and $|A|$ is an integer, we must have $|A| \geq (|S| + 1)/3$. □

Remark 1.2. One can reformulate the above argument in non-probabilistic language, in which case it basically employs the well-known method in combinatorics of *counting pairs*. In the proof, we are basically counting in two different ways the ordered pairs (ω, s) which satisfy (i) $\omega \in \Omega$ (ii) $s \in S$ (iii) $\omega s \in C \pmod{p}$. I leave it as a voluntary exercise to fill out the details.

Remark 1.3. As shown in Example 2, the set C employed in the above proof is a sum-free subset of $\mathbb{Z}_p$ of maximum size. Hence, it is natural to conjecture that Theorem 1.1 cannot be improved upon. It turns out that this is not the case, but it seems to be non-trivial to show it. In a long and difficult paper, Bourgain [1] showed that, for any finite $S \subseteq \mathbb{Z}$, not containing zero, one can always find a sum-free subset A of S such that $|A| \geq \frac{|S|+2}{3}$. Nothing better than this is known, I think.

For upper bounds, it suffices to find examples of sets $S \leq \mathbb{N}$ without large sum-free subsets. I believe the current record is due to Lewko [2], who found, via computer search, a set of 28 positive integers with no sum-free subset of size 12. From such a single example, one can construct (I leave it as another exercise to determine how) arbitrarily large, finite sets $S \subseteq \mathbb{N}$ for which there are no sum-free subsets of size exceeding $\frac{11}{28}|S|$. The gap between $1/3$ and $11/28$ is a significant open problem.

References

[1] J. Bourgain, *Estimates related to sumfree subsets of sets of integers*, Israel J. Math. **97** (1997), no.1, 71–92.
[2] M. Lewko, *An improved upper bound for the sum-free subset constant*, J. Integer Seq. **13** (2010), no.8, Article 10.8.3, 15pp (electronic).

Second moment method and distinct subset sums.

Proposition 1.4. *Let X be a non-negative real-valued random variable, and $\alpha \geq 1$. Then*

$$\mathbb{P}(X \geq \lambda \mathbb{E}[X]) \leq \frac{1}{\lambda}. \tag{1.2}$$

Proof. Simple exercise. This result is called *Markov's inequality*. □

DEFINITION 2: Let X be a random variable. The *variance* of X, written as Var$[X]$, is defined as

$$\text{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2].$$

The square root of the variance is called the *standard deviation*.

Using linearity of expectation, it's easy to show that (exercise, if you have never done it before !)

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{1.3}$$

NOTATION : $\mathbb{E}[X] := \mu_X$, $\sqrt{\mathrm{Var}[X]} := \sigma_X$. We drop the subscripts when there can be no confusion about what random variable is being considered.

Remark 1.5. At this point it is worth clarifying the terminology *second moment method*. Let X be a random variable. The *exponential generating function* of X is the random variable e^X. Thus

$$e^X = \sum_{k=0}^{\infty} \frac{X^k}{k!}.$$

Under suitable convergence conditions, linearity of expectation yields that

$$\mathbb{E}[e^X] = \sum_{k=0}^{\infty} \frac{\mathbb{E}[X^k]}{k!}.$$

The quantity $\mathbb{E}[X^k]/k!$ in this expression is called the *k:th moment* of the random variable X. From (1.3) we see that the variance of X involves its second moment, hence the name.

A rough analogy to studying the 2nd moment of a random variable is to study the second derivative of a smooth function in calculus. And just as it is pretty hard to find a real-life situation where one is interested in the third derivative of a smooth function, so in probability theory it is pretty rare to study the third moment of a random variable. Basically, if you can't get a handle on the second moment, then you're probably in a whole lot of trouble !

Finally, it should now not come as a great shock that the term *first moment method* is applied when one just studies the expectation of a random variable itself. So this is the method we've been using in the applications up to now.

The basic concentration estimate involving variance is *Chebyshev's inequality*:

Proposition 1.6. *Let X be a random variable with mean μ and standard deviation σ. Let $\lambda \geq 1$. Then*

$$\mathbb{P}(|X-\mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}. \tag{1.4}$$

Proof. Define a new random variable Y by $Y := |X-\mu|^2$. Then the left-hand side of (1.4) is just, by definition of variance, $\mathbb{P}(Y \geq \lambda^2 \mathbb{E}[Y])$. Markov's inequality (1.2) now gives the result immediately. □

We now specialise to the case where

$$X = X_1 + \cdots + X_n$$

is a sum of indicator variables. We do not assume the X_i to be identically distributed though. Indeed let us denote by A_i the event indicated by X_i and $p_i := \mathbb{P}(A_i)$. Thus

$$X_i = \begin{cases} 1, & \text{with probability } p_i, \\ 0, & \text{with probability } 1-p_i. \end{cases}$$

Also denote $\mu_i := \mathbb{E}[X_i]$, $\sigma_i^2 :=$ Var$[X_i]$. Clearly, $\mu_i = p_i$. Also, by (1.3) and the fact that $X_i^2 = X_i$ since X_i only takes on the values 0 and 1, we have

$$\sigma_i^2 = p_i - p_i^2 = p_i(1 - p_i). \tag{1.5}$$

We thus have the inequality

$$\sigma_i^2 \leq \mu_i. \tag{1.6}$$

Since in applications the individual probabilities p_i are usually very small (even if the number of events A_i is usually very large), one does not lose much information in using (1.6).

We want an expression for the variance of X. Using (1.3) and several applications of linearity of expectation, we obtain that

$$\sigma^2 = \sum_{i=1}^{n} \sigma_i^2 + \sum_{i \neq j} \text{Cov}(X_i, X_j), \tag{1.7}$$

where the *covariance* of X_i and X_j is defined by

$$\text{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j].$$

Since the X_i are indicator variables, we have

$$\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] = \mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i)\mathbb{P}(A_j).$$

Hence Cov$(X_i, X_j) = 0$ if the events A_i and A_j are independent. In this case, (1.7) simplifies to

$$\sigma^2 = \sum_{i=1}^{n} \sigma_i^2, \quad \text{when the } X_i \text{ are independent.} \tag{1.8}$$

We now describe an application of the second moment method to a problem in number theory. It is a relatively simple application from a theoretical viewpoint, in that it only uses Chebyshev's inequality and (1.8).

DEFINITION 3: Let $A = \{a_1, ..., a_k\}$ be a finite set of integers. A is said to have *distinct subset sums* if, for every two distinct subsets I, J of $\{1, ..., k\}$, the sums $\sum_{i \in I} a_i$ and $\sum_{j \in J} a_j$ have different values[1].

Let $f(n)$ be the maximum possible size of a subset of $\{1, ..., n\}$ which has distinct subset sums.

LOWER BOUNDS:

Take $n = 2^k$ and $A = \{2^i : 0 \leq i \leq k\}$. This example shows that $f(n) \geq 1 + \lfloor \log_2 n \rfloor$. Erdős offered 500 dollars for a proof that there exists a universal constant C such that $f(n) \leq \log_2 n + C$. Note that he's not asking here for a computation of the optimal C or even a decent estimate of it, just a proof that some such constant exists, in other words that $f(n) = \log_2 n + O(1)$. The base-2 example shows that $C \geq 1$. If we confine ourselves to integer C then a number of authors, starting with John Conway

[1]If I is the empty set, the sum is assigned the value zero. The definition extends to infinite sets, but the notation will just become a bit more complicated.

and Richard Guy in 1969, have produced examples showing that $C \geq 2$. The point here is that the powers-of-2 example is not optimal. Note that, in order to get a better lower bound on C, it suffices to do so for a single n, because of the following trick: if $A = \{a_1, ..., a_k\}$ is a subset of $\{1, ..., n\}$ with distinct subset sums, and u is any odd number s.t. $1 \leq u \leq 2n$, then $A' = \{2a_1, ..., 2a_k, u\}$ is a subset of $\{1, ..., 2n\}$ with distinct subset sums and one additional element. This means that if $f(n) > \log_2 n + C$ then $f(N) > \log_2 N + C$ for every N of the form $N = 2^t n$.

One can then use a computer to help find individual examples ... For up-to-date information on lower bounds see, for example,

http://garden.imacs.sfu.ca/?q=op/sets_with_distinct_subset_sums

UPPER BOUNDS:

If A has size k and is contained in $\{1, ..., n\}$ then there are 2^k distinct subset sums and each is among $\left\{0, ..., nk - \frac{k(k-1)}{2}\right\}$. Thus

$$2^{f(n)} \leq 1 + nf(n) - \frac{f(n)(f(n)-1)}{2}.$$

Taking base-2 logs, we have

$$f(n) \leq \log_2 n + \log_2 f(n) + O(1),$$

which leads to a bound of the form

$$f(n) \leq \log_2 n + \log_2 \log_2 n + O(1). \tag{1.9}$$

Erdős improved this to the following

Theorem 1.7.

$$f(n) \leq \log_2 n + \frac{1}{2} \log_2 \log_2 n + O(1). \tag{1.10}$$

Proof. The idea is to refine the basic counting argument which leads to (1.9) by using the fact that the 2^k subset sums for a set $A = \{a_1, ..., a_k\}$ are not "uniformly distributed" in the interval $\left[0, nk - \frac{k(k-1)}{2}\right]$, but that there is a higher concentration of sums close to the mean. To make this precise requires a second moment analysis, which we now perform in detail.

Let $A = \{a_1, ..., a_k\}$ be a subset of $\{1, ..., n\}$ with distinct subset sums. For each $i = 1, ..., k$, let X_i be the r.v. given by

$$X_i = \begin{cases} a_i, & \text{with probability } 1/2, \\ 0, & \text{with probability } 1/2. \end{cases} \tag{1.11}$$

The X_i:s are assumed to be independent, and we let $X := \sum_{i=1}^k X_i$. In words, X is the value of a subset sum of A, where the subset is chosen uniformly at random from all 2^k subsets of A. Though it is of no interest for the proof, note that, by linearity of expectation,

$$\mu = \mathbb{E}[X] = \frac{1}{2}\left(\sum_{i=1}^k a_i\right). \tag{1.12}$$

What we are interested in is the variance. By (1.5) and (1.8), we have

$$\sigma^2 = \mathrm{Var}(X) = \frac{1}{4}\left(\sum_{i=1}^{k} a_i^2\right) \leq \frac{kn^2}{4},$$

hence $\sigma \leq n\sqrt{k}/2$. Now let $\lambda \geq 1$. By Chebyshev's inequality,

$$\mathbb{P}\left(|X-\mu| \geq \frac{\lambda n\sqrt{k}}{2}\right) \leq \frac{1}{\lambda^2}.$$

This is equivalent to saying that

$$\mathbb{P}\left(|X-\mu| < \frac{\lambda n\sqrt{k}}{2}\right) \geq 1 - \frac{1}{\lambda^2}. \tag{1.13}$$

Now, on the one hand, X is integer-valued, and the number of integers satisfying $|X - \mu| < \frac{\lambda n\sqrt{k}}{2}$ is less than $1+\lambda n\sqrt{k}$. On the other hand, (1.13) says that the probability that a uniformly randomly chosen subset sum satisfies this inequality is at least $1 - 1/\lambda^2$. Since there are 2^k subset sums, and they are assumed to be all distinct, it follows that there must be at least $\left(1 - \frac{1}{\lambda^2}\right) 2^k$ integers satisfying the inequality. We conclude that

$$\left(1 - \frac{1}{\lambda^2}\right) 2^{f(n)} < 1 + \lambda n\sqrt{f(n)}.$$

Taking base-2 logs, we have

$$f(n) \leq \log_2 n + \frac{1}{2}\log_2 f(n) + O(1),$$

where the $O(1)$-term depends on λ. From this one easily deduces (1.10). $\square$