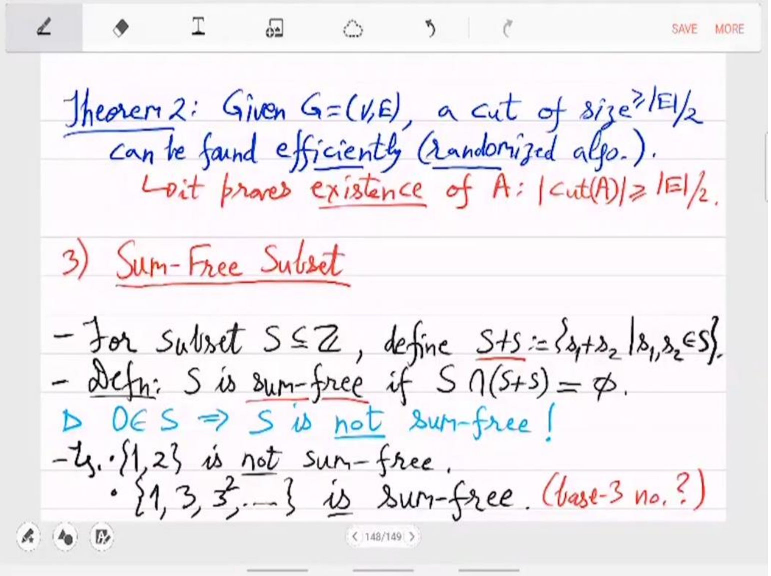
Probability for Computer Science
Prof. Nitin Saxena
Department of Computer Science and Engineering
Indian Institute of Technology - Kanpur

Module - 7
Lecture - 28
Sum-Free Subsets, Discrepancy

Let us now do a third example of probabilistic methods. So, we will actually see many examples; hopefully, it will give you a good understanding of the applications of probability in computer science and combinatorics.

(Refer Slide Time: 00:38)



Theorem 2: Given $G=(V,E)$, a cut of size $\geq |E|/2$ can be found efficiently (randomized algo.).

↳ it proves existence of $A: |cut(A)| \geq |E|/2$.

3) Sum-Free Subset

- For subset $S \subseteq \mathbb{Z}$, define $S+S := \{s_1+s_2 \mid s_1, s_2 \in S\}$.
- Defn: S is sum-free if $S \cap (S+S) = \phi$.

▷ $0 \in S \Rightarrow S$ is not sum-free!

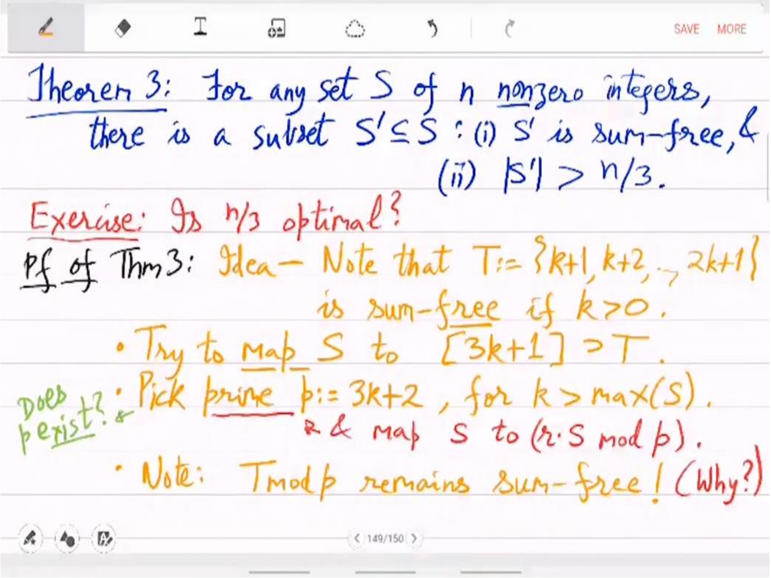
- Eg. • $\{1,2\}$ is not sum-free.
  • $\{1, 3, 3^2, \ldots\}$ is sum-free. (base-3 no.?)

So, here, this I think I already mentioned when we started the course. So, we want to show that for any set containing n integers, you can find a subset which is large and which does not have equations like A + B = C. So, subsets that are sum-free. So, for subset of integers S, define S + S as the natural thing for any 2 elements in S, maybe with repetition, just add them and all set sums. That is S + S. Now, recall S sum-free if S and S + S, they are disjoint.

In other words, there is no S 3 in S such that S 1 + S 2 = S 3. So, one thing you can immediately deduce is that, if 0 is in S, then, S is not sum-free. Why is that? Well, simply because, I mean, 0 + 0 is 0. So, just having 0, immediately gives you something in common between S and S + S. But obviously, there might be other reasons. Even if 0 is not present, there might be other reasons why S is not sum-free. So, let us see more examples.

So, 1, 2 for example is not sum-free, because 1 + 1 is 2. What about sum-free? So, actually, if you look at a sequence of large enough numbers, growing numbers, like 1, 3, 3 square to infinity; this set is sum-free, because you can see that 1 + 1 is 2, which is smaller than 3; and similarly, 3 + 3 is 6, which is smaller than 9. So, when you add up 2 numbers here, you actually get a unique number. It is like base 3 representation. Think of this as base 3 representation. So, because of that, you can show that this set is sum-free.

(Refer Slide Time: 04:19)



Theorem 3: For any set S of n nonzero integers, there is a subset $S' \subseteq S$: (i) S' is sum-free, & (ii) $|S'| > n/3$.

Exercise: Is $n/3$ optimal?

Pf of Thm3: Idea— Note that $T := \{k+1, k+2, .., 2k+1\}$ is sum-free if $k > 0$.

- Try to map S to $[3k+1] \supset T$.

Does p exist? • Pick prime $p := 3k+2$, for $k > \max(S)$. r & map S to $(r \cdot S \bmod p)$.

- Note: $T \bmod p$ remains sum-free! (Why?)

So, what we will show here is the following theorem. For any set S of n non-zero integers, there exists a subset S prime of S that is sum-free. So, S prime is sum-free and S prime is large. How large? More than one-third. So, both these properties are true. So, every subset of integers, obviously, we have to avoid, we want to avoid 0, so, non-zero integers will have a; actually, in this case, you can also include 0, because we are only talking about the existence of a large subset; so, it will not matter; but anyway; so, every subset of integers, there is a large subset that is sum-free.

So, why n by 3? Why not less? Why not more? So, this thing, I leave as an exercise. Is n by 3 optimal? And let me continue with the proof of theorem 3. So, the idea is, it is a mapping idea. We want to map S into something that we understand, because S is given as an arbitrary subset of integers. So, this we do not understand, but we can map it to something which we understand, which we know is sum-free, and then, hopefully identify S prime from that map.

So, note that subset T, k + 1, k + 2 to 2k + 1, this window of numbers or this interval of numbers, this is always sum-free. This is a problem if you take k = 0, but then, k = 1 onwards, it is okay. So, if you look at 2, 3, ...; so, for example, k = 1 gives you 2 and 3. So, this is sum-free. And in general what is happening is, k + 1 when added to itself gives you 2k + 2, which actually goes outside the set T. So, that is the problem.

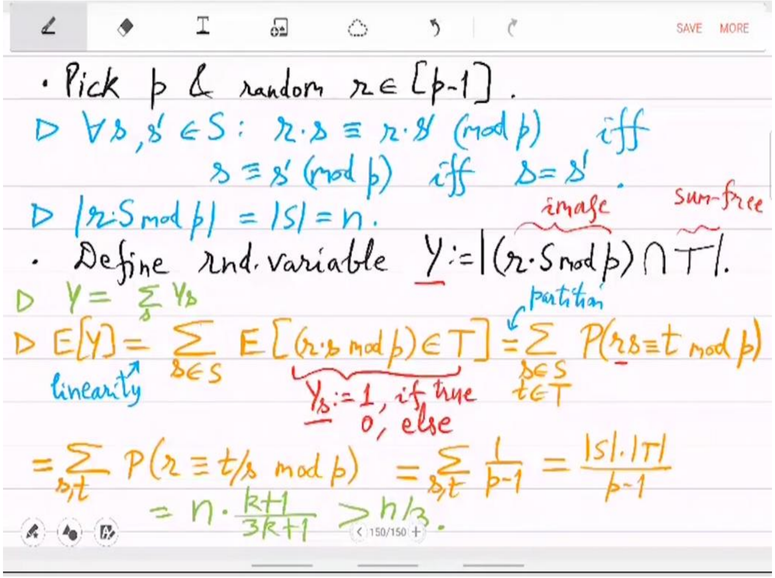
When you add 2 numbers, it becomes too large, including k + 1 with itself. So, this is clearly sum-free. So, why not identify S with this T as much as possible? Why not construct a map? So, that is what we will try to do. So, try to map S to 0 to 3k + 1, which contains T as a subset. So, let us try to map S to this range of size 3k + 2, and then identify which elements went to T. So, that would give us S prime.

So, for this map, what we will do is, use some algebra. In fact, you can also call, we will use some number theory. So, we will actually reduce the numbers S modulo a prime, a big prime. So, pick a prime number p, 3k + 2, for k bigger than the size of, or the value of numbers in S. So, you want p to be significantly bigger than the numbers you have in S, so that when you reduce modulo p, they will not change. So, this is the mod p arithmetic.

Advantage of mod p arithmetic is that you can for example now multiply, you can try to scale up numbers in S. You multiply each number with some number r. And then, modulo p, bring it back to this range 0 to 3k + 1. And map S to r dot S mod p. This is what we will do. We will use the prime number p to actually first multiply each element in S with a number r and then reduce modulo prime. So, it will bring you back to 0 to 3k + 1, the remainder.

And here, note that, if you reduce the numbers in T mod p, even then it is sum-free. It cannot be the case that 2 numbers in T, that is k + 1 to 2k + 1; then you add 2 numbers. It cannot be the case that the sum is equal to a third number modulo p. So, what we are saying is, even modulo p, it remains sum-free. So, think about this. This is an important point; why is this true? So, with these sequence of ideas, now I hope you are ready to see the formal proof.

(Refer Slide Time: 11:57)



- Pick p & random $r \in [p-1]$.

▷ $\forall s, s' \in S$: $r \cdot s \equiv r \cdot s' \pmod{p}$ iff $s \equiv s' \pmod{p}$ iff $s = s'$.

▷ $|r \cdot S \bmod p| = |S| = n$.

- Define rnd. variable $\underline{Y} := |\overbrace{(r \cdot S \bmod p)}^{\text{image}} \cap \overset{\text{sum-free}}{T}|$.

▷ $Y = \sum_{s} Y_s$

▷ $E[Y] \underset{\text{linearity}}{=} \sum_{s \in S} E[\underbrace{(r \cdot s \bmod p) \in T}_{\underline{Y_s} := 1, \text{ if true};\ 0, \text{ else}}] \overset{\text{partition}}{=} \sum_{\substack{s \in S \\ t \in T}} P(\underline{r}s \equiv t \bmod p)$

$= \sum_{s,t} P(r \equiv t/s \bmod p) = \sum_{s,t} \frac{1}{p-1} = \frac{|S| \cdot |T|}{p-1}$

$= n \cdot \frac{k+1}{3k+1} > n/3$.

Actually, let me remark. Here also, you have to show, why does it exist? This numbers, prime numbers of the type 3k + 2, they exist. No matter what S is, and no matter how large n is, you will always be able to find this prime number p. So, show this using number theory arguments or treat this up in a textbook. I will not go into that. So, prime numbers p exist. And using such a p; so, pick p and a random number r in 1 to p - 1.

And note that for all elements s 1, s 2 in S, r dot s is same as; let me call it s and s prime. So, r dot s is the same as r dot s prime mod p, if and only if; so, since r is not divisible by p, you can cancel it out. So, this is really the same as s congruent to s prime mod p, which is the same as; well, note that p has been chosen 3 times bigger than the values in s. So, when you reduce s by p and s prime by p, you get the same number.

So, this is really same as s and s prime equal. In other words, when you multiply the set s with r, when you scale up all the numbers modulo; well, scale up by r and then reduce modulo p, you still get distinct numbers. Let us just remember that. So, this means that r S mod p is the same as before; they are n many; the number is the same as before; cardinality is still n. This will be very important actually. So, now define random variable Y.

So, what are interested in? We are interested in, as given in the idea, we are interested in how many elements from T did s map to? So, did you get lot of elements of T or none, or somewhere, exactly where in the middle? So, that is given by r dot S mod p set intersection with T. So, how many elements of T did you hit via this transformation? That is the image of the map; and this one is sum-free.

So, in other words, Y is counting or considering a sum-free subset. Random variable, right? So, we actually have to take the size of this. So, what is the sum-free subset that this process is giving you? That is what we are interested in? And let us do that next? The expectation, right? What is the expectation? So, expectation is the elements of; so, r s mod p, they are n elements; so, go over all of them and see whether this particular element is in T.

So, we have to also look at the elements in T now. It becomes a double summation. So, what is the probability that r s is t mod p? That is the expectation. In fact, I should have already called this probability. It is an event. So, this, we are looking at these disjoint events, whether r s is in T, whether r s prime is in T? That is a problem. I need expectation here, because of linearity of expectation. So, let me go back. So, in fact, I have to say it is a random variable.

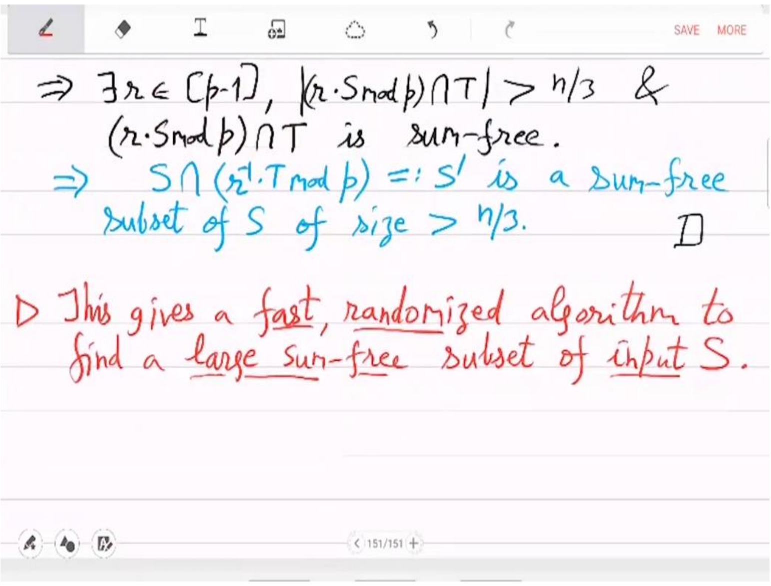
So, this Y s is a random variable. This is 1 if the above is true, and 0 otherwise. So, this is actually the thing which we do always. So, Y is equal to sigma Y s. And by linearity of expectation, expectation of Y is sigma expectation of Y s. And then, that becomes probability. This expectation of the indicator is nothing but probability of this event happening, which is whether r s is t mod p.

Now, r s can be only 1 T, a unique T. So, that is by partition. So, let me also write that. So, this is by linearity, and this is by partition. So, now, what is this probability? So, note that in this, everything is fixed; actually s is fixed, t is fixed, p is fixed in this event. It is only r that the probability is over; r is this random choice which was made. So, let us work with that. So, this is equal to sum of course. Then sum over probability that r is t by s mod p.

You can divide by s, because we have taken s to be non-zero. And no element in s should become 0 mod p; that also we want. So, let me add it somewhere. It should be added here. S should be mapped to non-zero numbers. That is how you should pick the prime p, because, well, p is quite large, so, if originally, s was zero-free, it will remain zero-free. So, hence, here you are allowed to divide by s.

And that actually forces r to be just one value; that is, there is only one favourable value. And we were picking r to be random from 1 to p - 1. So, you get probability 1 over p - 1. So, what is this ultimately? So, that is size of S, size of T, p - 1, which is equal to n times size of T is k + 1 and size of p - 1 is 3k + 1. So, that is more than n by 3. That is what we have learnt till now. Under the transformation, you will be hitting more than n by 3 elements of T. And T sum-free. So, this means that, if you look at the pre-image of the map, you have found a sum-free subset.

(Refer Slide Time: 22:43)



$\Rightarrow \exists r \in [p-1], |(r\cdot S \bmod p) \cap T| > n/3$ &
$(r\cdot S \bmod p) \cap T$ is sum-free.
$\Rightarrow S \cap (r^{-1}\cdot T \bmod p) =: S'$ is a sum-free subset of S of size $> n/3$. $\square$

▷ This gives a fast, randomized algorithm to find a large sum-free subset of input S.

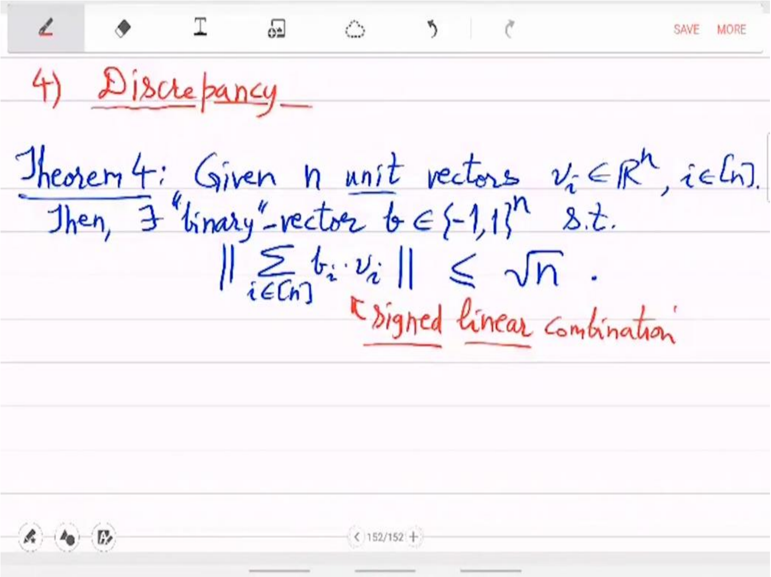
So, this implies that there exists an r, 1 to p - 1; there exists a transformation such that r S mod p transformation intersection T is large, and r S mod p intersection T is sum-free, which means that S intersection r inverse T mod p. So, the transformation was scaling up s. Now, the reverse, the inverse of that transformation will be scaling down by r mod p. So, if you do that, it will give you s prime; is a sum-free subset of S, and the size will not change.

So, if originally you had more than n by 3 elements, you still have more than n by 3 elements. That is it. So, that finishes the proof. It is a tricky proof. And still some number theory details are missing, which you can fill in if you are interested, but it is a powerful display of probabilistic method. So, you can see that in the original problem, there was no mention of number theory.

Well, you were looking at numbers, so, maybe there was some number theory, but there was certainly no mention of probability. And now, using probability, we are able to prove this beautifully. And not just prove, this is actually also a fast algorithm. So, this algorithm gives you s prime. So, this gives a fast randomised algorithm to find a large sum-free subset of input S.

You basically will pick this random r and p; and then you will transform; and after transformation, you will see that this one-third of the image; in the image, one-third will give you the sum-free subset. So, that finishes our third example. And we do not have much time. Let me anyway start the fourth application which is discrepancy.

(Refer Slide Time: 26:34)



4) Discrepancy

Theorem 4: Given n unit vectors $v_i \in \mathbb{R}^n$, $i \in [n]$.
Then, $\exists$ "binary"-vector $b \in \{-1,1\}^n$ s.t.

$$\Big\| \sum_{i \in [n]} b_i \cdot v_i \Big\| \leq \sqrt{n} .$$

↖signed linear combination

So, what we will show here is the following theorem, next time. So, suppose you are given n unit vectors, v i, and these are in the ambient space n as well, then there exist a bit string. They call it a bit string, but it actually + 1, - 1 coordinates. So, there exists a vector b; let me also call this then a binary vector, such that if you look at the combination, linear combination of v i's with these signs; i 1 to n; this length is at most square root n.

So, what they are saying is that, as long as you have unit vectors, there will always be a combination by just using signs. There is a signed linear combination; that is what you should remember; whose length is significantly smaller than n, square root n. So, we will prove this next time.