Math 130B - Projects

In this writing assignment, you are each assigned one of the problems below, and your task is to write a formal solution as if this was an article for an undergraduate mathematics journal. You will need to decide how to structure the article and how to guide your readers while remaining appropriately formal and concise. If you find it appropriate, you are welcome to discuss additional extensions to the question asked (such as generalizations or open questions). This would increase your grade. For some of these problems, there can be many different proofs, and it may be worth presenting alternative proofs as well. The length of the paper is not fixed; you should try to give justice to the problem being solved. Nevertheless, in no case can the paper be longer than 10 pages or shorter than 5 pages.

If you cannot figure out the mathematics for these problems yourself within a reasonable time, you can certainly ask us for help. We will also hold Zoom meetings, one for each project, so that all students working on the same project can ask questions and exchange ideas. Feel free to work in groups, but the write-up should be done individually.

1 No monochromatic sumsets

For every subset $A := \{a_1, \ldots, a_k\} \subseteq \mathbb{N}$ define its *sum-set* to be

$$S(A) := \left\{ \sum_i a_i x_i \mid x_i \in \{0,1\} \text{ for all } 1 \leq i \leq k \right\}.$$

For example, for $A = \{1,3,4\}$ we have

$$S(A) = \{1,3,4,5,7,8\},$$

where for $A = \{1,2\}$ we have

$$S(A) = \{1,2,3\}.$$

Consider the set of integers $[n] := \{1, \ldots, n\}$. We call $A \subseteq [n]$ *relevant* if and only if $S(A) \subseteq [n]$ (for example, if you take a set that contains both $n-10$ and 12, then since $(n-10)+12 \notin [n]$, this set is not relevant). Now, given a function $f : [n] \to \{Red, Blue\}$ (we refer to f as a 2-*coloring* of $[n]$), we say that f is k-*valid* if for every relevant subset $A \subseteq [n]$ of size k we have that $S(A)$ is **not** monochromatic; that is, for every relevant set $A \subseteq [n]$ of size k there exist $x \neq y \in S(A)$ with $f(x) \neq f(y)$.

Next we define, for every $k \in \mathbb{N}$, the function $F(k)$ as the maximum integer n for which there exists a k-valid 2-coloring of $[n]$. In this project we will prove that $F(k)$ has (at least) a double exponential type behavior. There are multiple ways to approach this problem, one good option is as follows:

(a) Try to understand the problem by considering the case $k = 2$.

(b) Prove that there exists a constant c (independent of k) such that for every subset $A \subseteq [n]$ of size k we have $|S(A)| \geq ck^2$.

(c) Find a bound on n using a random coloring strategy (that is, what is the largest n for which you can still prove the existence of a k-valid coloring using a random coloring?), and identify the weakest point in the proof.

(d) Show that if A is a set of size k with $|S(A)| \leq 2^k - 2$, then there exists $x \in \mathbb{N}$ with both x and $2x$ in $S(A)$.

(e) For each odd number $x \in [n]$, let $G_x = \{x, 2x, 4x, 8x, \ldots\} \cap [n]$ be the geometric sequence starting at x. Prove that $\bigcup_x G_x = [n]$ and that $G_x \cap G_y = \emptyset$ for all $x \neq y$.

(f) Suppose that you take any coloring of the odd numbers, and for each odd x you extend the coloring by alternating colors in G_x. Show that no relevant subset $A \subseteq [n]$ with $|S(A)| \leq 2^k - 2$ is monochromatic. Moreover, for each relevant set A with $|S(A)| = 2^k - 1$, show that $S(A)$ intersects at least 2^{k-1} distinct G_x's.

(g) Now, instead of taking an arbitrary coloring, take a random 2-coloring of the odd numbers and extend it to the even numbers as described in (f). What is the largest n for which you can prove that the expected number of relevant A's with a monochromatic $S(A)$ is smaller than 1. Conclude that, for such an n, a k-valid coloring exists.

The list above is a guide for how to gain understanding of the problem, but the paper must not be a list; it should be cohesive, with sections that are motivated and flow well.

You could provide alternate proofs for your results, if appropriate. You could also consider extensions such as adding more colors or considering other sets instead of sum sets.

2 Number of Prime Divisors of a Random Number

For a positive integer n, let $\nu(n)$ denote the number of primes dividing n. The goal of this project is to study the behavior of the parameter $\nu(n)$. First, we aim to prove the following theorem.

Theorem 1. *Let $\omega(n)$ be any function such that $\omega(n) \to \infty$ when n goes to infinity. Then the number of x in $[n]$ such that*

$$|\nu(x) - \log\log n| > \omega(n)\sqrt{\log\log n} \tag{1}$$

is $o(n)$.

In a sense, this theorem says that for large n, most integers in $[n]$ have around $\log\log n$ prime divisors.

(a) Choose x randomly from $\{1, 2, \ldots, n\}$ and set $M = n^{1/100}$. Let X be the number of primes less than M dividing x. Show that

$$E[X] = \sum_{p \leq M} \left(\frac{1}{p} + O\left(\frac{1}{n}\right)\right). \tag{2}$$

Using the fact (can you prove it?) that

$$\sum_{p \le n} \frac{1}{p} = \log\log n + O(1), \tag{3}$$

conclude that $E[X] = \log\log n + O(1)$.

(b) Show that

$$\mathrm{Var}[X] = \log\log n + O(1) \tag{4}$$

and use Chebyshev's inequality to deduce that

$$\Pr\left[|X - \log\log n| > t\sqrt{\log\log n}\right] < \frac{1}{t^2} + o(1)$$

for any $t > 0$. Conclude that the same holds for ν in place of X and that Theorem 1 follows.

Theorem 1 is already a pretty good concentration result, but we will now turn to prove something even more precise. We'll prove the following, stronger theorem.

Theorem 2. *Let λ be a fixed real number. Then*

$$\lim_{n\to\infty} \frac{1}{n}\left|\left\{x : 1 \le x \le n,\ \nu(x) > \log\log n + \lambda\sqrt{\log\log n}\right\}\right| = \int_\lambda^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

You might recognize the integral as the tail of a normal distribution. What this theorem really says is that ν behaves like a normal random variable with mean and variance $\log\log n$.

(c) Fix some function $s(n)$ with $s(n) \to \infty$ and $s(n) = o((\log\log n)^{1/2})$ and set $M = n^{1/s(n)}$. Like in part (a), pick x at random in $[n]$ and let X be the number of prime divisors of x less than M. Now for every prime p, define the independent random variables $Y_p \sim Bern(1/p)$. Set $Y = \sum_{p \le M} Y_p$. How are X and Y related?

Compute $\mu = E[Y]$ and $\sigma^2 = \mathrm{Var}[Y]$. By the central limit theorem $\tilde{Y} = (Y - \mu)/\sigma$ converges to a standard normal random variable $g \sim \mathcal{N}(0, 1)$ in distribution. One property of the normal distribution is that this implies convergence of the moments: $E[\tilde{Y}^k] \to E[g^k]$ for all positive integers k.

(d) Let $\tilde{X} = (X - \mu)/\sigma$. Show that

$$E[\tilde{X}^k] - E[\tilde{Y}^k] = o(1).$$

Using the fact that convergence to Gaussian moments implies convergence in distribution to a normal random variable, conclude that $\tilde{X}$ approaches g in distribution. Deduce Theorem 2.

3 Balls and bins

Suppose that there are n balls and n bins, and each ball is placed into exactly one bin independently at random.

(*a*) What is the probability that there are no unoccupied bins?

Since the probability to have unoccupied bins is quite high, there must be some bins loaded with more than one ball. Let $M(n)$ be the number of balls in the most loaded bin.

(*b*) Show that, assuming that n is sufficiently large, with probability at least (say) 9/10 we have that $M(n) = \Theta(\frac{\log n}{\log \log n})$ (note that you need to prove both a lower and an upper bound!).

Now, we show that the bound decreases by a lot if we slightly change the model. We still have n balls and n bins, but at each time step i, we pick *two* bins independently at random and then place ball i into the least loaded bin among these two.

Let $M_2(n)$ denote the size of the most loaded bin at the end of this process. Our main goal is to show that in this model, with high probability we have that $M_2(n) = O(\log \log n)$. There are many ways to prove it, and here we chose an approach which is based on random graphs.

Start with n isolated vertices (that is, there are no edges connecting them) corresponding to the bins. Now at each time step i, pick two vertices (bins) uniformly at random and add an edge e_i between them. Direct this edge towards the bin that we place the i-th ball into (think of e_i as an arrow pointing to that vertex). Note that two vertices might have many edges between them!

Our goal is to show that at the end of this process (meaning, after we add n random directed edges), no vertex has more than $C \log \log n$ many arrows pointing towards it for some $C > 0$. For technical reasons, it will be easier to assume that we only throw $n/1000$ balls (therefore, edges) into bins.

(*c*) Assuming we can prove this bound for $n/1000$ balls, show that we can prove a similar bound for n balls.

(*d*) Call the graph that this procedure produces G. Show that every connected component of G is of size $O(\log n)$.

(*e*) Show that there is a large constant K, such that every subset $X \subseteq V(G)$ of size at least K spans a subgraph with average degree at most 5 (recall that the sum of the degrees in a graph equals twice the number of edges).

Now, consider the following process, at each time step j, remove all the vertices of degree at most 10 from the graph and delete all eges connected to them. Repeat this process until no such vertices exist.

(*f*) Show that this process ends after $O(\log \log n)$ time steps, and that each remaining connected component is of size at most K.

(*g*) Show that every bin (or vertex) has load at most $10i + K$, where i is the last time step the vertex "survived" before being removed from the graph in the above deleting process.

(*h*) Deduce the theorem under the assumption that there are n balls and not the assumed $n/1000$.

4 Random graphs

Given a positive integer n and a probability $0 < p < 1$ (which could be a function of n), the random graph $G(n,p)$ is a graph $G = (V,E)$ on $|V| = n$ vertices in which every possible edge is present with probability p and these events are independent. The edge probability p could be a constant independent of n (say $p = \frac{1}{2}$), but p could also depend on n, like $p(n) = \frac{\log(n)}{n}$. (The sample space is the set of all graphs on n vertices and the probability of any graph depends on its number of edges.)

In this project you'll investigate various properties of random graphs. Your results should include:

(a) Show that for every $c \leq 1/1000$ and $p = c/n$, with probability that tends to 1 as n tends to infinity, all the connected components of $G(n,p(n))$ are of size $O(\log n)$. (this is actually true for all $c < 1$, can you prove it?)

(b) Show that for every $c \geq 1000$ and $p = c/n$, with probability that tends to 1 as n tends to infinity, there exists a connected component of a linear size. (this is true for all $c > 1$, can you prove something like that?)

(c) For which p (as a function $p(n)$ of n), does the random graph $G(n,p(n))$ have no isolated vertex (i.e., a vertex with no edge incident to it)? The type of result you should prove is that if $p(n) > (1+\epsilon)\frac{\log n}{n}$ then the probability that $G(n,p(n))$ has an isolated vertex tends to 0 as n tends to infinity (say for fixed $\epsilon > 0$), while if $p(n) < (1-\epsilon)\frac{\log n}{n}$ then the probability that $G(n,p(n))$ has an isolated vertex tends to 1. It is useful to look at the random variable representing the number of isolated vertices, and to compute its expectation, variance, etc.

(d) Show that for $p \geq (1+\epsilon)\frac{\log n}{n}$, then the probability that $G(n,p)$ is *connected* tends to 1 as n tends to infinity.

(e) Write (and prove) a result stating the size of the largest clique in the random graph $G(n, \frac{1}{2})$. Again this needs to be formalized.

(f) For which values of p do we start seeing triangles with probability that tends to 1?

Extensions you could investigate would be when (i.e. for which values of p) a random graph has at least one triangle (a clique of size 3), or when a random graph is connected, or the maximum (or minimum) degree of a random graph. There are countless other questions that could be considered. People make careers out of this.

5 Anti concentration

Let $a = (a_i)_{i=1}^n$ be a sequence of n non-zero integers, and let

$$S_n(a) = \sum_{i=1}^{n} a_i X_i,$$

where the X_i's are i.i.d (independent, and identically distributed) random variables with

$$\Pr[X_i = 1] = \Pr[X_i = -1] = \frac{1}{2}.$$

In this project we'll try to understand the following general problem, whose simplest variant was shown in class.

Given a sequence $a = (a_i)_{i=1}^n \in (\mathbb{R} \setminus \{0\})^n$, give a non-trivial upper bound for its *atom probability*. That is, try to get the best possible upper bound for the following parameter:

$$\rho(a) := \max_{m \in \mathbb{R}} \Pr\left[\sum_i a_i X_i = m\right].$$

The proofs here will involve some combinatorial tricks, tools from probability, and some very basic Fourier analysis (in order to be able to solve it, all you need to know is that $e^{it} = \cos t + i \sin t$, where i is a (complex) square root of -1). There are multiple ways to approach this problem and many variants of possible questions, and here we will study few of them in an increasing order of difficulty.

1. Suppose that $a_i = 1$ for all $1 \leq i \leq n$ and that n is even. What is the probability that $S_n(a) = 0$?

2. Prove that every integer n has at most one representation as $\sum_{i=1}^{\infty} x_i 2^i$, where the x_i's are in $\{-1, 1\}$. Conclude that for the sequence $a_i = 2^i$, $1 \leq i \leq n$, we have $\Pr[S_n = m] \in \{0, 2^{-n}\}$ for all $m \in \mathbb{Z}$.

3. Convince yourself that $\Pr[S_n(a) = 0] = \mathbb{E}[\delta_0(S_n(a))]$, where

$$\delta_0(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

   Moreover, observe that

$$\delta_0(x) = \int_{-1/2}^{1/2} e^{2\pi i t x} dt,$$

   write a formula for $\mathbb{E}[\delta_0(S_n(a))]$, and simplify it as much as you can.

4. If you are not familiar with it, Google the "AM-GM inequality" (Arithmetic Mean - Geometric Mean inequality) and state it (feel free to prove it, although it is not mandatory). Using this inequality, prove that $\Pr[S_n(a) = 0] \leq \Pr[\sum_{i=1}^n X_i = 0]$ for every sequence of non-zero a_i's and an even n.

5. Prove a better bound for the case where all the a_i's are distinct. For example, what if $a_i = i$ for all $1 \leq i \leq n$? Can you prove a decent bound using Sperner's theorem?

6. Can you obtain sharper results by assuming less structure on the sequence (a_i)? Can you come up with a sequence that has an atom probability around (say) $n^{-2.5}$?

Ideas for extensions/generalizations: consider other families of vectors a (for example, what if you know that a is contained in an arithmetic progression of length (say) n^2?). Try to obtain some non-trivial bounds for higher dimensions; that is, assume that the a_i's are vectors in $\mathbb{Z}^d$ and consider $\sum_{i=1}^n a_i X_i$, where the X_i's are i.i.d. ± 1 balanced random variables. What is the probability that $\sum_i a_i X_i = \overline{0}$?