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 having 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* as

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

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:

1. Try to understand the problem by considering the case $k = 2$.
2. Prove that there exists a constant c (independent of k) such that for every subset $A \subseteq [n]$ of size k we have $|S(A)| \geq ck^2$.

3. 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 understand what is the weakest point in the proof.

4. 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 $S(A)$.

5. 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 $x \neq y$.

6. 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.

7. 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 Greatest Angle Among Points in $\mathbb{R}^d$ (Alon, Spencer)

The main goal of this project is to prove the following theorem.

Theorem 1. *For every $d \geq 1$, there is a set of at least $\lfloor \frac{1}{2}(\frac{2}{\sqrt{3}})^d \rfloor$ points in $\mathbb{R}^d$ such that all angles determined by three points from the set are acute.*

Here's one approach.

(a) Consider the unit cube, C_d, in $\mathbb{R}^d$ whose vertices are the the 0,1-vectors of length d. Treat each vertex as the characteristic vector of a subset of $[d]$. That is, associate the vertex a with the set $A = \{i : 1 \leq i \leq d,\ a_i = 1\}$. Prove that the vertices a, b, and c of C_d, corresponding to the sets A, B, and C, respectively, determine a right angle at c if and only if

$$A \cap B \subset C \subset A \cup B. \tag{1}$$

Deduce that it suffices to construct a set $S \subseteq C_d$ of cardinality at least the one stated int the theorem no three distinct members of which satisfy (1).

(b) Define $m = \lfloor \frac{1}{2}(\frac{2}{\sqrt{3}})^d \rfloor$ and choose, randomly and independently, $2m$ d-dimensional 0/1 vectors $a_1, \ldots, a_{2m}$, where each coordinate of each of the vectors is independently chosen to be either 0 or 1 with equal probability. Fix some triple a, b, c of the chosen points. Compute the probability that the corresponding sets satisfy (1), then compute the expected number of right angles determined by the a_i's. Deduce that there is a choice of as et X of $2m$ points in which the number of right angles is at most m. Complete the proof of the theorem from here.

3 Second Moment Method (Alon Spencer 4.2.1)

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

Theorem 2. *Let $\omega(n)$ be any function such that $\omega(n) \to \infty$. Then the number of x in $[n]$ such that*

$$|\nu(x) - \log\log n| > \omega(n)\sqrt{\log\log n} \tag{2}$$

is $o(n)$.

(a) Choose x randomly from $\{1, 2, \ldots, n\}$ and set $M = n^{1/00}$. Let X be the number of primes less than M dividing x. Show that

$$E[X] = \sum_{p \le M} \left(\frac{1}{p} + O\left(\frac{1}{n}\right) \right). \tag{3}$$

(b) Look up Abel's summation formula. Prove it too if you want – the proof just uses basic calculus. Use this along with Stirling's approximation ($\log(n!) = n\log n + O(n)$) to show that

$$\sum_{p \le n} \frac{1}{p} = \log\log n + O(1). \tag{4}$$

Hint: start by showing that

(c) Show that

$$\mathrm{Var}[X] = \log\log n + O(1) \tag{5}$$

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 2 follows.

Theorem 2 is already a pretty good concentration result, but we can be even more precise. We'll prove the following, stronger theorem.

Theorem 3. *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 integrand as the density of a normal distribution. What this theorem really says is that ν behaves like a normal random variable with mean and variance $\log\log n$.

(d) Fix some function $s(n)$ with $s(n) \to \infty$ and $s(n) = o((\log\log n)^{1/2})$ and set $M = n^{1/s(n)}$.

4 Recovering two mixed communities

Suppose there are two schools across the street from one another. Both schools participate in each other's bands, theater productions, and other after-school clubs, so the students interact with each other a fair amount. Say you knew each student's Facebook friend list but not what school they go to. Can you come up with a way separate the students by school from their friends lists alone? We can translate this problem into the language of graph theory.

Problem 1. *Partition $[n]$ into two sets of the same size, $X \sqcup Y$. Then, choose probabilities $p > q$, and place edges between vertices according to the following rule: if $uv \subseteq X$ or $uv \subseteq Y$, then add uv with probability p. Otherwise add it with probability q. All choices are made independently at random. Given the resulting graph, G, can we find a way to recover the partition $X \sqcup Y$?*

a Suppose that G is a graph obtained according to the above distribution and that G' is a random graph where each edge is being added with probability $\frac{p+q}{w}$. Now, suppose that you see both graphs but don't know which one is G and which one is G'. Can you come up with a simple way to determine which one is which (assume for example that $p = 1/2$ and $q = 1/3$)?

b Going back to the original problem, show that if you classify vertices at random, then with probability at least (Say) $2/3$ we miss-classify around $n/2$ vertices.

c Now, we show that if p and q are far apart then we can actually classify all of them correctly with high probability. Suppose $p = 1/2$ and $q = 1/3$. Give an easy way to recover the partition (Hint: think about the number of common neighbors of two vertices from the same community VS number of common neighbors of vertices from different communities). Use a Chernoff bound to make your idea precise.

d Finally, we will show how to use the aid of some linear algebra in order to classify a better proportion than in the random strategy, even when $|p-q|$ is very small (depends on n). For example, let $p = 1/2$ and $q = p - \frac{1000}{\sqrt{n}}$. From now on, suppose $X = \{1, 2, \ldots, n/2\}$. Consider the block matrix

$$M = \begin{pmatrix} pJ_{n/2} & qJ_{n/2} \\ qJ_{n/2} & pJ_{n/2} \end{pmatrix},$$

where J_k is the $k \times k$ all-1 matrix. Compute the eigenvalues and eigenvectors of M. How is M related to our problem? *Hint: Think adjacency matrix of a graph.*

e Define the matrix B by

$$B = A(G) + pI,$$

where $A(G)$ is the adjacency matrix of G define adjacency matrix for them?. A problem showing that B and M are close would be cool, but Talagrand's inequality seems pretty complicated for them. Should we just give it to them?

f Let $\alpha_1 \geq \cdots \geq \alpha_n$ be the eigenvalues of B and let $\mu_1 > \mu_2 > 0 = \cdots = \mu_n$ be the eigenvalues of M (why can we assume that they're all real?). Problem on min-max (or give it to them) and show that $|\alpha_i - \mu_i| \leq \|R\|$, where $R = B - M$.

g If two matrices are close, then what can we say about their eigenvectors? Problem on Davis-Kahan maybe we will just state it for them (with a proof).

h Now to bring it home. Let v_2 be the eigenvector of B corresponding to α_2 and let w_2 be the eigenvector of M corresponding to μ_2, and let $\delta = v_2 - w_2$. Show that

$$\|\delta\| \leq \sqrt{2} \sin\theta,$$

where θ is the angle between v_2 and w_2. Apply Davis-Kahan. If k is the number of vertices that v_2 misclassifies, conclude that

$$k \leq \frac{Cp}{(p-q)^2},$$

for some constant C. Finally, deduce that for our choices of p and q, we misclassify at most a constant fraction of vertices.

5 Balls and bins

Suppose that there are n balls and n bins, and each ball picks exactly one bin independently at random. What is the probability that there are no unoccupied bins?

Since the probability to have unoccupied bins is quite high, it means in particular that there are bins loaded with more than one ball. Let $M(n)$ be the number of balls in the most loaded bin. 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 completely drops if we assume the following model: There are n balls and n bins, and now, in each time step i, Ball i picks *two* bins independently at random and then goes into the list loaded bin among the selected bins. 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 $M_2(n) = O(\frac{\log\log n}{})$

6 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 need to investigate various properties of random graphs. Your results should include:

1. 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 one 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.

2. A result stating what the size of the largest clique is in the random graph $G(n, \frac{1}{2})$. Again this needs to be formalized.

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.

7 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 are trying 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 as-good-as-you-can 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) 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.

1. Suppose that $a_i = 1$ for all $1 \le i \le n$ 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 $a_i = 2^i$, $1 \le i \le n$, 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 $\sum_i a_i X_i = \overline{0}$?