Expected Uses of Probability

Evan Chen
evanchen@mit.edu

August 11, 2014

Sorry for the bad title. This is mostly about expected value, both in its own right and in the context of the probabilistic method.

1 Definitions and Notation

Nothing tricky here, just setting up notation. I'll try to not be overly formal.

A **random variable** is just a quantity that we take to vary randomly. For example, the outcome of a standard six-sided dice roll, say D_6, is a random variable. We can now discuss the **probability** of certain events, which we'll denote $\mathbb{P}(\bullet)$. For instance, we can write

$$\mathbb{P}(D_6 = 1) = \mathbb{P}(D_6 = 2) = \cdots = \mathbb{P}(D_6 = 6) = \frac{1}{6}$$

or $\mathbb{P}(D_6 = 0) = 0$ and $\mathbb{P}(D_6 \geq 4) = \frac{1}{2}$.

We can also discuss the **expected value** of a random variable X, which is the "average" value. The formal definition is

$$\mathbb{E}[X] \stackrel{\text{def}}{=} \sum_x \mathbb{P}(X = x) \cdot x.$$

But an example for our dice roll D_6 makes this clearer:

$$\mathbb{E}[D_6] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \cdots + \frac{1}{6} \cdot 6 = 3.5.$$

In natural language, we just add up all the outcomes weighted by probability they appear.

We'll assume the reader has some familiarity with basic graph theory terms; see `http://en.wikipedia.org/wiki/Graph_theory#Definitions` otherwise. One term we'll define here that may not be so known – given a graph G, an **independent set** is a set of vertices for which no two are connected by an edge.

2 Properties of Expected Value

2.1 A Motivating Example

It is an unspoken law that any introduction to expected value begins with the following classical example.

Example 2.1. At MOP, there are n people, each of who has a name tag. We shuffle the name tags and randomly give each person one of the name tags. Let S be the number of people who receive their own name tag. Prove that the expected value of S is 1.

This result might seem surprising, as one might intuitively expect $\mathbb{E}[S]$ to depend on the choice of n.

For simplicity, let us call a person a *fixed point* if they receive their own name tag.[1] Thus S is just the number of fixed points, and we wish to show that $\mathbb{E}[S] = 1$. If we're interested in the expected value, then according to our definition we should go through all $n!$ permutations, count up the total number of fixed points, and then divide by $n!$ to get the average. Since we want $\mathbb{E}[S] = 1$, we expect to see a total of $n!$ fixed points.

Let us begin by illustrating the case $n = 4$ first, calling the people W, X, Y, Z.

	W	X	Y	Z	Σ
1	**W**	**X**	**Y**	**Z**	4
2	**W**	**X**	Z	Y	2
3	**W**	Y	X	**Z**	2
4	**W**	Y	Z	X	1
5	**W**	Z	X	Y	1
6	**W**	Z	**Y**	X	2
7	X	W	**Y**	**Z**	2
8	X	W	Z	Y	0
9	X	Y	W	**Z**	1
10	X	Y	Z	W	0
11	X	Z	W	Y	0
12	X	Z	**Y**	W	1
13	Y	W	X	**Z**	1
14	Y	W	Z	X	0
15	Y	**X**	W	**Z**	2
16	Y	**X**	Z	W	1
17	Y	Z	W	X	0
18	Y	Z	X	W	0
19	Z	W	X	Y	0
20	Z	W	**Y**	X	1
21	Z	**X**	W	Y	1
22	Z	**X**	**Y**	W	2
23	Z	Y	W	X	0
24	Z	Y	X	W	0
Σ	6	6	6	6	24

We've listed all $4! = 24$ permutations, and indeed we see that there are a total of 24 fixed points, which I've bolded in red. Unfortunately, if we look at the rightmost column, there doesn't seem to be a pattern, and it seems hard to prove that this holds for larger n.

However, suppose that *rather than trying to add by rows, we add by columns*. There's a very clear pattern if we try to add by the columns: we see a total of 6 fixed points in each column. Indeed, the six fixed W points correspond to the $3! = 6$ permutations of the remaining letters X, Y, Z. Similarly, the six fixed X points correspond to the $3! = 6$ permutations of the remaining letters W, Y, Z.

This generalizes very nicely: if we have n letters, then each letter appears as a fixed point $(n-1)!$ times.

[1] This is actually a term used to describe points which are unchanged by a permutation. So the usual phrasing of this question is "what is the expected number of fixed points of a random permutation?"

Thus the expected value is

$$\mathbb{E}[S] = \frac{1}{n!}\left(\underbrace{(n-1)! + (n-1)! + \cdots + (n-1)!}_{n \text{ times}}\right) = \frac{1}{n!} \cdot n \cdot (n-1)! = 1.$$

Cute, right? Now let's bring out the artillery.

2.2 Linearity of Expectation

The crux result of this section is the following theorem.

Theorem 2.2 (Linearity of Expectation)**.** *Given any random variables X_1, X_2, ..., X_n, we always have*

$$\mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n].$$

This theorem is obvious if the X_1, X_2, ..., X_n are independent of each other – if I roll 100 dice, I expect an average of 350. Duh. The wonderful thing is that this holds even if the variables are not independent. And the basic idea is just the double-counting we did in the earlier example: even if the variables depend on each other, if you look only at the expected value, you can still add just by columns. The proof of the theorem is just a bunch of sigma signs which say exactly the same thing, so I won't bother including it.

Anyways, that means we can now nuke our original problem. The trick is to define **indicator variables** as follows: for each $i = 1, 2, \ldots, n$ let

$$S_i \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if person } i \text{ gets his own name tag} \\ 0 & \text{otherwise.} \end{cases}$$

Obviously,

$$S = S_1 + S_2 + \cdots + S_n.$$

Moreover, it is easy to see that $\mathbb{E}[S_i] = \mathbb{P}(S_i = 1) = \frac{1}{n}$ for each i: if we look at any particular person, the probability they get their own name tag is simply $\frac{1}{n}$. Therefore,

$$\mathbb{E}[S] = \mathbb{E}[S_1] + \mathbb{E}[S_2] + \cdots + \mathbb{E}[S_n] = \underbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}_{n \text{ times}} = 1.$$

Now that was a lot easier! By working in the context of expected value, we get a framework where the "double-counting" idea is basically automatic. In other words, linearity of expectation lets us only focus on small, local components when computing an expected value, without having to think about why it works.

2.3 More Examples

Example 2.3 (HMMT 2006)**.** At a nursery, 2006 babies sit in a circle. Suddenly, each baby randomly pokes either the baby to its left or to its right. What is the expected value of the number of unpoked babies?

Solution. Number the babies 1, 2, ..., 2006. Define

$$X_i \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if baby } i \text{ is unpoked} \\ 0 & \text{otherwise.} \end{cases}$$

We seek $\mathbb{E}[X_1+X_2+\cdots+X_{2006}]$. Note that any particular baby has probability $\left(\frac{1}{2}\right)^2 = \frac{1}{4}$ of being unpoked (if both its neighbors miss). Hence $\mathbb{E}[X_i] = \frac{1}{4}$ for each i, and

$$\mathbb{E}[X_1 + X_2 + \cdots + X_{2006}] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_{2006}] = 2006 \cdot \frac{1}{4} = \frac{1003}{2}. \quad \square$$

Seriously, this should feel like cheating.

2.4 Practice Problems

The first two problems are somewhat straightforward applications of the methods described above.

Problem 2.4 (AHSME 1989)**.** Suppose that 7 boys and 13 girls line up in a row. Let S be the number of places in the row where a boy and a girl are standing next to each other. For example, for the row $GBBGGGBGBGGGBGBGGBGG$ we have $S = 12$. Find the expected value of S.

Problem 2.5 (AIME 2006 #6)**.** Let $\mathcal{S}$ be the set of real numbers that can be represented as repeating decimals of the form $0.\overline{abc}$ where a, b, c are distinct digits. Find the sum of the elements of $\mathcal{S}$.

The next three problems are harder; in these problems linearity of expectation is not the main idea of the solution. All problems below were written by Lewis Chen.

Problem 2.6 (NIMO 4.3)**.** One day, a bishop and a knight were on squares in the same row of an infinite chessboard, when a huge meteor storm occurred, placing a meteor in each square on the chessboard independently and randomly with probability p. Neither the bishop nor the knight were hit, but their movement may have been obstructed by the meteors. For what value of p is the expected number of valid squares that the bishop can move to (in one move) equal to the expected number of squares that the knight can move to (in one move)?

Problem 2.7 (NIMO 7.3)**.** Richard has a four infinitely large piles of coins: a pile of pennies, a pile of nickels, a pile of dimes, and a pile of quarters. He chooses one pile at random and takes one coin from that pile. Richard then repeats this process until the sum of the values of the coins he has taken is an integer number of dollars. What is the expected value of this final sum of money, in cents?

Problem 2.8 (NIMO 5.6)**.** Tom has a scientific calculator. Unfortunately, all keys are broken except for one row: `1`, `2`, `3`, `+` and `-`. Tom presses a sequence of 5 random keystrokes; at each stroke, each key is equally likely to be pressed. The calculator then evaluates the entire expression, yielding a result of E. Find the expected value of E.

(Note: Negative numbers are permitted, so `13-22` gives $E = -9$. Any excess operators are parsed as signs, so `-2-+3` gives $E = -5$ and `-+-31` gives $E = 31$. Trailing operators are discarded, so `2++-+` gives $E = 2$. A string consisting only of operators, such as `-++-+`, gives $E = 0$.)

3 Direct Existence Proofs

In its simplest form, we can use expected value to show existence as follows: suppose we know that the average score of the USAMO 2014 was 12.51. Then there exists a contestant who got at least 13 points, and a contestant who got at most 12 points. This is similar in spirit to the pigeonhole principle, but the probabilistic phrasing is far more robust.

3.1 A First Example

Let's look at a very simple example, taken from the midterm of a class at the San Jose State University.[2]

Example 3.1 (SJSU M179 Midterm)**.** Prove that any subgraph of $K_{n,n}$ with at least $n^2 - n + 1$ edges has a perfect matching.

We illustrate the case $n = 4$ in the figure.

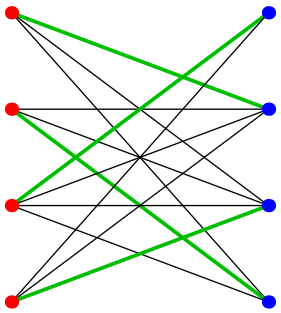

Figure 1: The case $n = 4$. There are $n^2-n+1 = 13$ edges, and the matching is highlighted in green.

This problem doesn't "feel" like it should be very hard. After all, there's only a total of n^2 possible edges, so having $n^2 - n + 1$ edges means we have practically all edges present.[3]

So let's be really careless and just *randomly* pair off one set of points with the other, regardless of whether there is actually an edge present. We call the *score* of such a pairing the number of pairs which are actually connected by an edge. We wish to show that some pairing has score n, as this will be the desired perfect matching.

So what's the expected value of a random pairing? Number the pairs $1, 2, \dots, n$ and define

$$X_i \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if the } i\text{th pair is connected by an edge} \\ 0 & \text{otherwise.} \end{cases}$$

Then the score of the configuration is $X = X_1 + X_2 + \cdots + X_n$. Given any red point and any blue point, the probability they are connected by an edge is at least $\frac{n^2-n+1}{n^2}$. This means that $\mathbb{E}[X_i] = \frac{n^2-n+1}{n^2}$, so

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] \\ &= n \cdot \mathbb{E}[X_1] \\ &= \frac{n^2 - n + 1}{n} \\ &= n - 1 + \frac{1}{n}. \end{aligned}$$

Since X takes only integer values, there must be some configuration which achieves $X = n$. Thus, we're done.

[2]For a phrasing of the problem without graph theory: given n red points and n blue points, suppose we connect at least $n^2 - n + 1$ pairs of opposite colors. Prove that we can select n segments, no two of which share an endpoint.

[3]On the other hand, $n^2 - n + 1$ is actually the best bound possible. Can you construct a counterexample with $n^2 - n$?

3.2 Ramsey Numbers

Let's do another simple example. Before we begin, I will quickly introduce a silly algebraic lemma, taken from [5, page 30].

Lemma 3.2. *For any positive integers n and k,*

$$\binom{n}{k} < \frac{1}{e}\left(\frac{en}{k}\right)^k.$$

Here $e \approx 2.718\ldots$ is Euler's constant.

Proof. Do $\binom{n}{k} < \frac{n^k}{k!}$ and then use calculus to prove that $k! \geq e(k/e)^k$. Specifically,

$$\ln 1 + \ln 2 + \cdots + \ln k \geq \int_{x=1}^{k} \ln x \, dx = k \ln k - k + 1$$

whence exponentiating works. □

Algebra isn't much fun, but at least it's easy. Let's get back to the combinatorics.

Example 3.3 (Ramsey Numbers)**.** Let n and k be integers with $n \leq 2^{k/2}$ and $k \geq 3$. Then it is possible to color the edges of the complete graph on n vertices with the following property: one cannot find k vertices for which the $\binom{k}{2}$ edges among them are monochromatic.

Remark. In the language of Ramsey numbers, prove that $R(k,k) > 2^{k/2}$.

Solution. Again we just randomly color the edges and hope for the best. We use a coin flip to determine the color of each of the $\binom{n}{2}$ edges. Let's call a collection of k vertices *bad* if all $\binom{k}{2}$ edges are the same color. The probability that any collection is bad is

$$\left(\frac{1}{2}\right)^{\binom{k}{2}-1}.$$

The number of collections in $\binom{n}{k}$, so the expected number of bad collections is

$$\mathbb{E}[\text{number of bad collections}] = \frac{\binom{n}{k}}{2^{\binom{k}{2}-1}}.$$

We just want to show this is less than 1. You can check this fairly easily using Lemma 3.2; in fact, we have a lot of room to spare. □

3.3 A Tricky Application

To cap off this section, we give a tricky proof (communicated to me via [3]) of the following result.

Theorem 3.4 (Ajtai-Komlós-Szemerédi)**.** *Given a triangle-free graph G with average degree d and N vertices, we can find an independent set with size at least $0.01\frac{N}{d}\log d$.*

Here, triangle-free just means there are no three vertices which are all adjacent to each other.[4] Another phrase for this is *locally sparse*.

[4] If you're familiar with the notation $R(m,n)$, here's some food for thought: what's the connection between this and $R(3,t)$?

Our first move is to try and replace the "average degree" d with "maximum degree" Δ. Here's the trick: notice that at most half of the vertices have degree greater than $2d$. So if we throw away these vertices, we still have half the vertices and left, and now the maximum degree is $\Delta \le 2d$. If we let $n = N/2$, then we just need an independent set of size $0.04\frac{n}{\Delta}\log\Delta$ in our new graph.

So now we have n vertices with maximum degree Δ. Here's the trick: consider all possible independent sets, and pick one set S uniformly at random (!). For this set S, we define a *score* X as follows:

- For each vertex u in S, we write a $+\Delta$ at that u.
- For each vertex v adjacent to something in S, we write $+1$ at that vertex. A vertex can receive $+1$ multiple times. However, note that since S is independent, this means $v \notin S$.
- Define the score X to be the sum of all numbers written.

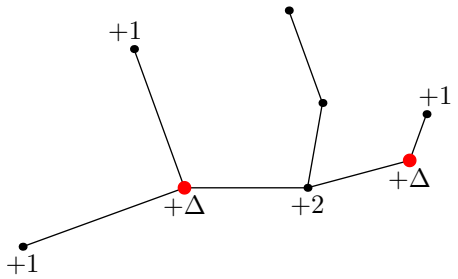

Figure 2: Assigning scores. The elements of S are the large red vertices.

Obviously, $X \le 2\Delta\,|S|$, since each vertex in S bestows Δ to itself and at most Δ among its neighbors. Now, we will place a bound on $\mathbb{E}[X]$, which will give us the result.

Consider any vertex v, and consider its set of neighbors. Note that by the triangle-free condition, no neighbors are adjacent to each other. Let X_v denote the sum of the scores given to vertex v (so $X = \sum_v X_v$). We are going to show that $\mathbb{E}[X_v] \ge 0.08\log\Delta$. This is enough, because then $\mathbb{E}[X] \ge 0.08n\log\Delta$, and for a good choice of X, we then have $|S| \ge 0.04n\frac{\log\Delta}{\Delta}$.

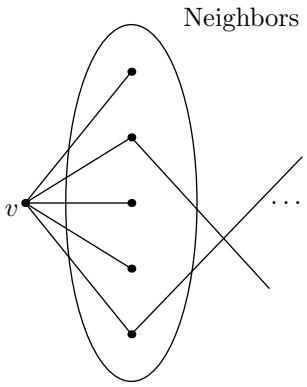

Figure 3: Ignoring things.

Suppose we're selecting an independent set, and we're done selecting everything aside from v and its neighbors. We'll prove that regardless of how the stuff outside is chosen,

$\mathbb{E}[X_v] \geq 0.08 \log \Delta$ still holds. Assume that, not including v, there are m other vertices in the neighborhood which we can still pick (i.e. they are not adjacent to anything outside that has been selected).

There are a few ways we can pick the remaining set:

- We can pick v, but then we can no longer pick any of its neighbors.
- We can pick any nonempty subset of the m remaining vertices, but then we can no longer pick v.
- We can pick no vertices.

There are a total of $1 + (2^m - 1) + 1$ possibilities. In the first scenario, the $X_v = +\Delta$. In the second and third scenario, $X_v = \mathbb{E}[\#\text{ neighbors chosen}] = \frac{1}{2}m$. So,

$$\mathbb{E}[X_v] = \frac{1 \cdot \Delta + 2^m \cdot \frac{1}{2}m}{2^m + 1} = \frac{\Delta}{2^m + 1} + \frac{1}{1 + 2^{-m}} \cdot \frac{m}{2} > \frac{1}{4} \max\left\{ \frac{\Delta}{2^m}, m \right\}.$$

It remains to prove this is at least $0.08 \log \Delta$. You can check this, because if $m \geq \frac{1}{2} \log_2 \Delta$, then $\frac{1}{4}m$ is enough; otherwise, $\frac{\Delta}{2^m} \geq \sqrt{\Delta}$ which is certainly sufficient.

3.4 Practice Problems

The first two problems are from [2]; the last one is from [4].

Problem 3.5. Show that one can construct a (round-robin) tournament with more than 1000 people such that in any set of 1000 people, some contestant beats all of them.

Problem 3.6 (BAMO 2004)**.** Consider n real numbers, not all zero, with sum zero. Prove that one can label the numbers as a_1, a_2, $\ldots$, a_n such that

$$a_1a_2 + a_2a_3 + \cdots + a_na_1 < 0.$$

Problem 3.7 (Russia 1996)**.** In the Duma there are 1600 delegates, who have formed 16000 committees of 80 people each. Prove that one can find two committees having no fewer than four common members.

4 Heavy Machinery

Here are some really nice ideas used in modern theory. Unfortunately I couldn't find many olympiad problems that used them. If you know of any, please let me know!

4.1 Alteration

In previous arguments we often proved a result by showing $\mathbb{E}[\text{bad}] < 1$. A second method is to select some things, find the expected value of the number of "bad" situations, and subtract that off. An example will make this clear.

Example 4.1 (Weak Turán)**.** A graph G has n vertices and average degree d. Prove that it is possible to select an independent set of size at least $\frac{n}{2d}$.

Proof. Rather than selecting $\frac{n}{2d}$ vertices randomly and hoping the number of edges is 1, we'll instead select each vertex with probability p. (We will pick a good choice of p later.)

That means the expected number of vertices we will take is np. Now there are $\frac{1}{2}nd$ edges, so the expected number of "bad" situations (i.e. an edge in which both vertices are taken) is $\frac{1}{2}nd \cdot p^2$.

Now we can just get rid of all the bad situations. For each bad edge, delete one of its endpoints arbitrarily (possibly with overlap). This costs us at most $\frac{1}{2}nd \cdot p^2$ vertices, so the expected value of the number of vertices left is

$$np - \frac{1}{2}ndp^2 = np\left[1 - \frac{1}{2}dp\right].$$

It seems like a good choice of p is $\frac{1}{d}$, which now gives us an expected value of $\frac{n}{2d}$, as desired. □

A stronger result is Problem 6.5.

4.2 Union Bounds and Markov's Inequality

A second way to establish existence is to establish a nonzero probability. One way to do this is using a union bound.

Proposition 4.2 (Union Bound)**.** *Consider several events A_1, A_2, ..., A_k. If*

$$\mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_k) < 1$$

then there is a nonzero probability that none of the events occur.

The following assertion is sometimes useful for this purpose.

Theorem 4.3 (Markov's Inequality)**.** *Let X be a random variable taking only nonnegative values. Suppose $\mathbb{E}[X] = c$. Then*

$$\mathbb{P}(X \geq rc) \leq \frac{1}{r}.$$

This is intuitively obvious: if the average score on the USAMO was 7, then at most $\frac{1}{6}$ of the contestants got a perfect score. The inequality is also sometimes called *Chebyshev's inequality* or *the first Chebyshev inequality.*

4.3 Lovász Local Lemma

The Lovász Local Lemma (abbreviated LLL) is in some sense a refinement of the union bound idea – if the events in question are "mostly" independent, then the probability no events occur is still nonzero.

We present below the "symmetric" version of the Local Lemma. An asymmetric version also exists (see Wikipedia).

Theorem 4.4 (Lovász Local Lemma)**.** *Consider several events, each occurring with probability at most p, and such that each event is independent of all the others except at most d of them. Then if*

$$epd \leq 1$$

the probability that no events occur is positive.

Note that we don't use the number of events, only the number of dependencies.

As the name implies, the local lemma is useful in situations where in a random algorithm, it appears that things do not depend much on each other. The following Russian problem is such an example.

Example 4.5 (Russia 2006)**.** At a tourist camp, each person has at least 50 and at most 100 friends among the other persons at the camp. Show that one can hand out a T-shirt to every person such that the T-shirts have (at most) 1331 different colors, and any person has 20 friends whose T-shirts all have pairwise different colors.

The constant $C = 1331$ is extremely weak. We'll reduce it to $C = 48$ below.

Solution. Give each person a random T-shirt. For each person P, we consider the event $E(P)$ meaning "P's neighbors have at most 19 colors of shirts". We wish to use the Local Lemma to prove that there is a nonzero probability that no events occur.

If we have two people A and B, and they are neither friends nor have a mutual friend (in graph theoretic language, the distance between them is at least two), then the events $E(A)$ and $E(B)$ do not depend on each other at all. So any given $E(P)$ depends only on friends, and friends of friends. Because any P has at most 100 friends, and each of these friends has at most 99 friends other than P, $E(P)$ depends on at most $100+100\cdot 99 = 100^2$ other events. Hence in the lemma we can set $d = 100^2$.

For a given person, look at their $50 \le k \le 100$ neighbors. The probability that there are at most 19 colors among the neighbors is clearly at most

$$\binom{C}{19} \cdot \left(\frac{19}{C}\right)^k.$$

To estimate the binomial coefficient, we can again use our silly Lemma 3.2 to get that this is at most

$$\frac{1}{e}\left(\frac{eC}{19}\right)^{19} \cdot \left(\frac{19}{C}\right)^k = e^{18} \cdot \left(\frac{19}{C}\right)^{k-19} \le e^{18}\left(\frac{19}{C}\right)^{31}.$$

Thus, we can put $p = e^{18}\left(\frac{19}{C}\right)^{31}$. Thus the Lemma implies we are done as long as

$$e^{19}\left(\frac{19}{C}\right)^{31} \cdot 100^2 \le 1.$$

It turns out that $C = 48$ is the best possible outcome here. Needless to say, establishing the equality when $C = 1331$ is trivial. □

5 Grand Finalé – IMO 2014, Problem 6

This article was motivated by the following problem, given at the 55th International Mathematical Olympiad, and the talk by Po-Shen Loh [3] given on it.

Example 5.1 (IMO 2014/6)**.** A set of lines in the plane is in *general position* if no two are parallel and no three pass through the same point. A set of lines in general position cuts the plane into regions, some of which have finite area; we call these its *finite regions*. Prove that for all sufficiently large n, in any set of n lines in general position it is possible to colour at least $\sqrt{n}$ lines blue in such a way that none of its finite regions has a completely blue boundary.

Note: Results with $\sqrt{n}$ replaced by $c\sqrt{n}$ will be awarded points depending on the value of the constant c.

We'll present two partial solutions ($c < 1$), one using Local Lovász, and one using alteration. For completeness we also present the official solution obtaining $c = 1$, even though it is not probabilistic. Then, we will establish the bound $O(\sqrt{n \log n})$ using some modern tools (this was [3]).

5.1 Partial Solution Using LLL

We'll show the bound $c\sqrt{n}$ where $c = (6e)^{-\frac{1}{2}}$.

Split the n lines into $c\sqrt{n}$ groups of size $\frac{\sqrt{n}}{c}$ each, arbitrarily. We are going to select one line from each of the groups at random to be blue. For each of the regions R_1, R_2, $\dots$, R_m we will consider an event A_k meaning "three of the lines bounding R_k are blue"; we designate these lines beforehand. We will show there is a nonzero probability that no events occur.

The probability of A_k is clearly at most $\left(\frac{c}{\sqrt{n}}\right)^3$.

For each R_k, we have three groups to consider. Each group consists of $\frac{\sqrt{n}}{c}$ lines. Each line is part of at most $2n-2$ regions. Hence A_k depends on at most $3 \cdot \frac{\sqrt{n}}{c} \cdot (2n-2)$ events.

Thus,

$$e\left(\frac{c}{\sqrt{n}}\right)^3 \left(3 \cdot \frac{\sqrt{n}}{c} \cdot (2n-2)\right) < 6ec^2 = 1$$

and we are done by LLL.

5.2 Partial Solution Using Alteration

We'll show the bound $c\sqrt{n}$ for any $c < \frac{2}{3}$.

First, we need to bound the number of triangles.

Claim. There are at most $\frac{1}{3}n^2$ triangles.

Proof. Consider each of the $\binom{n}{2}$ intersection of two lines. One can check it is the vertex of at most two triangles. Since each triangle has three vertices, this implies there are at most $\frac{2}{3}\binom{n}{2} < \frac{1}{3}n^2$ triangles. □

It is also not hard to show there are at most $\frac{1}{2}n^2$ finite regions[5].

Now color each line blue with probability p. The expected value of the number of lines chosen is

$$\mathbb{E}[\text{lines}] = np.$$

The expected number of completely blue triangles is less than

$$\mathbb{E}[\text{bad triangles}] < \frac{1}{3}n^2 \cdot p^3.$$

For the other finite regions, of which there are at most $\frac{1}{2}n^2$, the probability they are completely blue is at most p^4. So the expected number of completely blue regions here is at most

$$\mathbb{E}[\text{bad polygons with 4+ sides}] < \frac{1}{2}n^2 \cdot p^4.$$

Note that the expected number of quadrilaterals (and higher) is really small compared to any of the preceding quantities; we didn't even bother subtracting off the triangles

[5]Say, use $V - E + F = 2$ on the graph whose vertices are the $\binom{n}{2}$ intersection points and whose edges are the $n(n-2)$ line segments.

that we already counted earlier. It's just here for completeness, but we expect that it's going to die out pretty soon.

Now we do our alteration – for each bad, completely blue region, we un-blue one line. Hence the expected number of lines which are blue afterwards is

$$np - \left(\frac{n^2}{3} \cdot p^3\right) - \left(\frac{n^2}{2} \cdot p^4\right) = np\left(1 - \frac{np^2}{3} - \frac{np^3}{2}\right).$$

Ignore the rightmost $\frac{np^3}{2}$ for now, since it's really small. We want $p = k/\sqrt{n}$ for some k; the value is roughly $k \cdot (1 - k^2/3)$ at this point, so an optimal value of p is $p = n^{-1/2}$ (that is, $k = 1$); this gives

$$\sqrt{n} \cdot \left(\frac{2}{3} - \frac{27}{16}\frac{1}{\sqrt{n}}\right) = \frac{2}{3}\sqrt{n} - \frac{81}{32}.$$

For n sufficiently large, this exceeds $c\sqrt{n}$, as desired.

5.3 Interlude – Sketch of Official Solution Obtaining $c = 1$

This is not probabilistic, but we include it for completeness anyways. It is in fact just a greedy algorithm.

Suppose we have colored k of the lines blue, and that it is not possible to color any additional lines. That means any of the $n - k$ non-blue lines is the side of some finite region with an otherwise entirely blue perimeter. For each such line ℓ, select one such region, and take the next counterclockwise vertex; this is the intersection of two blue lines v. We'll say ℓ is the *eyelid* of v.

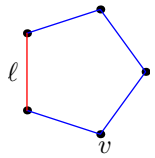

Figure 4: Here ℓ is the eyelid of v.

You can prove without too much difficulty that every intersection of two blue lines has at most two eyelids. Since there are $\binom{n}{2}$ such intersections, we see that

$$n - k \le 2\binom{k}{2} = k^2 - k$$

so $n \le k^2$, as required.

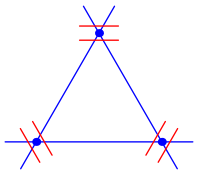

Figure 5: The greedy algorithm cannot do better than $\sqrt{n}$.

It's interesting to note that the greedy algorithm cannot be extended to achieve a result better than $\sqrt{n}$. To show this, note that if $n = m^2$, we can consider m arbitrary blue lines in general position, and then add $2\binom{m}{2}$ lines, two on either side of a given

intersection point. (Po-Shen Loh called these "tubes" in his talk.) Thus each of the new lines is the edge of a triangle with two blue sides, and so the greedy algorithm must stop here.

5.4 Overkill Solution

This solution is due to Po-Shen Loh [3]. We are now going to establish the bound $\sqrt{cn \log n}$. The heart is the following theorem.

Theorem 5.2 (Duke-Lefmann-Rödl)**.** *Given a hypergraph G with N vertices and with edges all of size 3, suppose that for any two vertices at most one 3-edge joins them. Then we can find an independent set with size at least $c \cdot \frac{N}{\sqrt{d}} \sqrt{\log d}$.*

Here a *hypegraph* is a graph in which an "edge" is any subset of vertices, as opposed to just two vertices. In the above theorem, all edges have three endpoints, and we require that any two vertices are joined by at most one edge.

In the context of the IMO problem, suppose we consider each of the n lines as a vertex and each finite region as a hyper-edge. Like in the previous solution, we treat pentagons, hexagons, ... as just quadrilaterals; hence we can assume all edges have size either 3 or 4. Once again we use a coin flip weighted with probability p to pick whether a vertex is chosen. Define the following random variables:

- Let W be the number of vertices remaining. Then $\mathbb{E}[W] = pn$.
- Let Y be the number of 4-edges. There are at most n^2 such edges, so $\mathbb{E}[Y] \le p^4 n^2$.
- Let Z be the number of pairs (u, v) with two 3-edges containing both (in the context of geometry, there are at most two such edges). Then $\mathbb{E}[Z] \le \binom{n}{2} p^4 < p^4 n^2$.

If we eliminate the situations in Y and Z then we reach a situation in which the theorem can be applied.

Finally, let X be the number of edges altogether remaining. Since each edge has ≥ 3 vertices and there are $\le n^2$ edges, $\mathbb{E}[X] \le n^2 p^3$.

Using Markov's Inequality,

$$\mathbb{P}(Y > 4p^4 n^2) < \frac{1}{4}.$$

Similarly,

$$\mathbb{P}(Z > 4p^4 n^2) < \frac{1}{4} \text{ and } \mathbb{P}(X > 4n^2 p^3) < \frac{1}{4}.$$

Meanwhile, W is a binomial distribution, so one can actually show that,

$$\mathbb{P}(W < 0.99pn) \to 0 \text{ as } n \to \infty.$$

Consequently, the union bound implies there is a nonzero chance that all these inequalities fail, meaning $Y \le 4p^4 n^2$, $Z \le 4p^4 n^2$, and $X \le 4n^2 p^3$, and $W \ge 0.99pn$.

Now using alteration again, we delete the "bad" situations in Y and Z. Then the number of vertices, N, is at least

$$N \ge W - Y - Z \ge 0.99pn - 8p^4 n^2 \sim pn(1 - 8p^3 n)$$

Let's pick $p = 0.01n^{-1/3}$. Now $N \sim pn$.

The average degree is at most

$$d = \frac{3X}{N} \le \frac{\sim n^2 p^3}{\sim np} \sim np^2.$$

The theorem then gives us a bound of

$$\frac{N}{\sqrt{d}} \log d \sim \frac{pn}{p\sqrt{n}} \sqrt{\log \sqrt{pn^2}} \sim \sqrt{n \log n}$$

as desired.

6 Practice Problems

These problems are mostly taken from [2, 4].

Problem 6.1 (IMC 2002)**.** An olympiad has six problems and 200 contestants. The contestants are very skilled, so each problem is solved by at least 120 of the contestants. Prove that there exist two contestants such that each problem is solved by at least one of them.

Problem 6.2 (Romania 2004)**.** Prove that for any complex numbers z_1, z_2, ..., z_n, satisfying $|z_1|^2 + |z_2|^2 + \cdots + |z_n|^2 = 1$, one can select $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \in \{-1, 1\}$ such that

$$\left| \sum_{k=1}^{n} \varepsilon_k z_k \right| \le 1.$$

Problem 6.3 (Shortlist 1999 C4)**.** Let A be a set of N residues $\pmod{N^2}$. Prove that there exists a set B of of N residues $\pmod{N^2}$ such that $A + B = \{a + b | a \in A, b \in B\}$ contains at least half of all the residues $\pmod{N^2}$.

Problem 6.4 (Iran TST 2008/6)**.** Suppose 799 teams participate in a round-robin tournament. Prove that one can find two disjoint groups A and B of seven teams each such that all teams in A defeated all teams in B.

Problem 6.5 (Caro-Wei Theorem)**.** Consider a graph G with vertex set V. Prove that one can find an independent set with size at least

$$\sum_{v \in V} \frac{1}{\deg v + 1}.$$

Remark. Note that, by applying Jensen's inequality, our independent set has size at least $\frac{n}{d+1}$, where d is the average degree. This result is called **Turán's Theorem** (or the complement thereof).

Problem 6.6 (USAMO 2012/6)**.** For integer $n \ge 2$, let $x_1, x_2, \ldots, x_n$ be real numbers satisfying $x_1 + x_2 + \ldots + x_n = 0$ and $x_1^2 + x_2^2 + \ldots + x_n^2 = 1$. For each subset $A \subseteq \{1, 2, \ldots, n\}$, define

$$S_A = \sum_{i \in A} x_i.$$

(If A is the empty set, then $S_A = 0$.) Prove that for any positive number λ, the number of sets A satisfying $S_A \ge \lambda$ is at most $2^{n-3}/\lambda^2$. For which choices of $x_1, x_2, \ldots, x_n, \lambda$ does equality hold?

Problem 6.7 (Online Math Open, Ray Li)**.** Kevin has $2^n - 1$ cookies, each labeled with a unique nonempty subset of $\{1, 2, \ldots, n\}$. Each day, he chooses one cookie uniformly at random out of the cookies not yet eaten. Then, he eats that cookie, and all remaining cookies that are labeled with a subset of that cookie. Compute the expected value of the number of days that Kevin eats a cookie before all cookies are gone.

Problem 6.8. Let n be a positive integer. Let a_k denote the number of permutations of n elements with k fixed points. Compute

$$a_1 + 4a_2 + 9a_3 + \cdots + n^2 a_n.$$

Problem 6.9 (Russia 1999)**.** In a certain school, every boy likes at least one girl. Prove that we can find a set S of at least half the students in the school such that each boy in S likes an odd number of girls in S.

Problem 6.10 (Sperner)**.** Consider N distinct subsets S_1, S_2, ..., S_N of $\{1, 2, \ldots, n\}$ such that no S_i is a subset of any S_j. Prove that

$$N \le \binom{n}{\lfloor \frac{1}{2}n \rfloor}.$$

Problem 6.11. Let n be a positive integer. Suppose $11n$ points are arranged in a circle, colored with one of n colors, so that each color appears exactly 11 times. Prove that one can select a point of every color such that no two are adjacent.

Problem 6.12 (Sweden 2010, adapted)**.** In a town with n people, any two people either know each other, or they both know someone in common. Prove that one can find a group of at most $\sqrt{n \log n} + 1$ people, such that anyone else knows at least one person in the group.

Remark. In graph theoretic language – given a graph with diameter 2, prove that a dominating set of size at most $\sqrt{n \log n} + 1$ exists.

Problem 6.13 (Erdös)**.** Prove that in any set S of n distinct positive integers we can always find a subset T with $\frac{1}{3}n$ or more elements with the property that $a + b \neq c$ for any $a, b, c \in T$ (not necessarily distinct).

Remark. Such sets are called *sum-free*.

7 Solution Sketches

2.4 Answer: 9.1. Make an indicator variable for each adjacent pair.

2.5 Answer: 360. Pick a, b, c randomly and compute $\mathbb{E}[0.\overline{abc}]$. Then multiply by $|\mathcal{S}|$.

2.6 $8p = 4 \cdot \left(p + p^2 + p^3 + \ldots\right)$.

2.7 Let x_n be the EV at a state with $n \pmod{100}$. Then $x_0 = 0$ and

$$x_n = \frac{1}{4}\left((x_{n+1} + 1) + (x_{n+5} + 5) + (x_{n+10} + 10) + (x_{n+25} + 25)\right).$$

Do algebra.

2.8 Answer: 1866. Show that one can replace `+` or `-` buttons with `STOP`. Show that one can replace `1` and `3` buttons with `2`. Let $p = \frac{3}{5}$. Compute $2(p + 10p^2 + \cdots + 10^4 p^5)$.

3.5 Suppose there are n people, and decide each edge with a coin flip. Compute the expected number of 1000-subsets for which there is no one better than all. Check that this is less than 1 for very large n.

3.6 Show that a random permutations has expected value at most 0. Why are the inequalities strict?

3.7 Let n_i be the number of committees which the ith delegate is in. Pick two committees randomly and find the expected value of the number of common members. Use Jensen's inequality to get a lower bound on $\sum \binom{n_i}{2}$.

6.1 Pick the contestants randomly. Find the expected number of problems both miss.

6.2 Select each of the ε_i randomly with a coin flip. Square the left-hand side and use the fact that $|z|^2 = z\overline{z}$ for any z.

6.3 Randomly selecting B works; you can even permit repeated elements in B. You may need the inequality $\left(1 - \frac{1}{n}\right)^n \le \frac{1}{e}$.

6.4 Let d_k be the number of teams which defeat the kth team (here $1 \le k \le 799$). Select A randomly and compute the expected number of teams dominated by everyone in A. You need Jensen on the function $\binom{x}{7}$.

6.5 Use the following greedy algorithm – pick a random vertex, then delete it and all its neighbors. Repeat until everything is gone.

6.6 Compute $\mathbb{E}[S_A^2]$ for a random choice of A. Markov Inequality.

6.7 The number of days equals the number of times a cookie is chosen. Compute the probability any particular cookie is chosen; i.e. the expected value of the number of times the cookie is chosen. Sum up.

6.8 For a random permutation let X be the number of fixed points. We already know $\mathbb{E}[X] = 1$. Compute $\mathbb{E}[\binom{X}{2}]$. Use this to obtain $\mathbb{E}[X^2]$.

6.9 Use a coin flip to decide whether to select each girl, then take as many boys as possible. Show that any person, girl or boy, has exactly a 50% chance of being chosen.

6.10 First prove that

$$\sum_{k=1}^{N} \frac{1}{\binom{n}{|S_k|}} \le 1.$$

To do this, consider a random maximal chain of subsets

$$\emptyset = T_0 \subset T_1 \subset T_2 \subset \cdots \subset T_n = \{1, 2, \ldots, n\}.$$

Compute the expected number of intersections of this chain with $\{S_1, S_2, \ldots, S_N\}$.

6.11 LLL. Here $p = 11^{-2}$ and $d = 42$.

6.12 If any vertex has small degree, then its neighbors are already the desired set. So assume all degrees are greater than $\sqrt{n \log n}$. Pick each person with probability p for some well-chosen p; then we expect to pick np people. Show that the probability someone fails is less than $\frac{1}{n}$ and use a union bound. The inequality $1 - p \le e^{-p}$ is helpful.

6.13 Work modulo a huge prime $p = 3k + 2$. Find a nice sum-free (mod p) set U of size $k + 1$ first, and then consider $U_n = \{nx \mid x \in U\}$ for a random choice of n. Compute $\mathbb{E}[|S \cap U_n|]$.

References

[1] pythag011 at http://www.aops.com/Forum/viewtopic.php?f=133&t=481300

[2] Ravi B's collection of problems, available at:
http://www.aops.com/Forum/viewtopic.php?p=1943887#p1943887.

[3] Problem 6 talk ($c > 1$) by Po-Shen Loh, USA leader, at the IMO 2014.

[4] Also MOP lecture notes: http://math.cmu.edu/~ploh/olympiad.shtml.

[5] Lecture notes by Holden Lee from an MIT course:
http://web.mit.edu/~holden1/www/coursework/math/18997/notes.pdf

Thanks to all the sources above. Other nice reads that I went through while preparing this, but eventually did not use:

1. Alon and Spencer's *The Probabilistic Method*. The first four chapters are here:
http://cs.nyu.edu/cs/faculty/spencer/nogabook/.
2. A MathCamp lecture that gets the girth-chromatic number result:
http://math.ucsb.edu/~padraic/mathcamp_2010/class_graph_theory_probabilistic/lecture2_girth_chromatic.pdf