Pigeonhole Principle and the Probabilistic Method

Lecturer: Michel Goemans

In these notes, we discuss two techniques for proving the existence of certain objects (graphs, numbers, sets, etc.) with certain properties.

1 The Pigeonhole Principle

We first discuss the pigeonhole principle and its applications. A basic version states:

> *If m objects (or pigeons) are put in n boxes (or pigeonholes) and $n < m$, then at least one box contains more than one object.*

Further, one can see that at least one box contains at least $\lceil \frac{m}{n} \rceil$ objects. This "principle" is so basic, it is natural to suspect one cannot deduce anything interesting from this basic principle. However, this suspicion is incorrect.

Same birthday

The number of MIT undergraduate students is 4528. We know (without checking) that there are 13 of them that share the same birthday. Indeed, there are 366 possible birthdays, and $4528 > 12 \cdot 366$.

Two equal degrees

A graph G consists of a set V (the elements of V are called *vertices*) and a set E of pairs of vertices (the elements of E are called *edges*). The *degree* of a vertex v in a graph is the number of edges containing v.

Theorem 1. *In any finite graph, there are two vertices of equal degree.*

As a story, this means that at a party with n persons, there exist two persons who know the same number of people at the party.

Proof. For any graph on n vertices, the degrees are integers between 0 and $n-1$. Therefore, the only way all degrees could be different is that there is exactly one vertex of each possible degree. In particular, there is a vertex v of degree 0 (with no neighbors) and a vertex w of degree $n-1$ adjacent to all other vertices). However, if there is an edge (v, w), then v cannot have degree 0, and if there is no edge (v, w) then w cannot have degree $n-1$. This is a contradiction. □

Equal sum subsets

Here is a more profound application of the pigeonhole principle. Suppose we have 30 7-digit numbers. We claim that there are two disjoint subsets of these numbers which have the same sum. How do we prove this? There are $2^{30} - 1$ distinct nonempty subsets of these numbers; these will be

the pigeons. For each of these subsets, the sum of the numbers in the subset will be the pigeonhole. Since our numbers are all between 0 and 10^7, the sum of thirty of them is at most $3 \cdot 10^8$, which is less than $2^{30} - 1 \approx 10^9$. Thus, since there are more subsets (pigeons) than sums (holes), there must be two subsets that have the same sum. Thus, we have an equation

$$x_{i_1} + x_{i_2} + \ldots + x_{i_k} = x_{j_1} + x_{j_2} + \ldots + x_{j_l}$$

These two subsets may not disjoint, but we can eliminate any variable that appears on both sides of this equation. If we do this, then we get two disjoint subsets which both have the same sum. There is just one last thing to check, which is that we don't get the equation $0 = 0$ after eliminating the common variables. This cannot happen since the original two subsets were different.

One more comment on this question. While it is quite easy to prove that these numbers exist, they are quite hard to find. With 30 7-digit numbers, the problem is in the range of modern computers, but if you increase the number of numbers to 100, and make them correspondingly larger, with all known methods, the computation time to find two subsets of with the same sum is enormous.

Monotone subsequences

The next result shows that every very long sequence of distinct real numbers contains a long subsequence which is monotone, which means the subsequence is increasing or decreasing. Consider for example the following sequence of length 12 (i.e., has 12 terms):

$$4, 3, 2, 1, 8, 7, 6, 5, 12, 11, 10, 9.$$

The subsequence $4, 8, 11$ (consisting of the first, fifth, and tenth terms) is increasing of length 3, and the subsequence $4, 3, 2, 1$ (consisting of the first four terms) is decreasing of length 4. One can check that, in the sequence above, there is no longer increasing subsequence and no longer decreasing subsequence. The following theorem implies that any longer sequence of distinct real numbers has to have a longer increasing subsequence or a longer decreasing subsequence.

Theorem 2. *Any sequence of $mn + 1$ distinct real numbers $a_1, a_2, \ldots, a_{mn+1}$ has an increasing subsequence of length $m + 1$ or a decreasing subsequence of length $n + 1$.*

Proof. Let $t(i)$ denote the length of the longest increasing subsequence ending with a_i. It $t(i) > m$ for some i, then we have an increasing subsequence of length at least $m + 1$ and we are done. So we may assume $t(i) \in \{1, 2, \ldots, m\}$ for all i. As we have $mn + 1$ numbers with m possible values, by the pigeonhole principle, there must be some $s \in \{1, \ldots, m\}$ such that $t(i_j) = s$ for at least $n + 1$ indices $i_1 < \ldots < i_{n+1}$. We claim that $a_{i_1} > a_{i_2} > \cdots > a_{i_{n+1}}$, so that we have a decreasing subsequence of length $n + 1$ and we are done. Indeed, if this was not the case, then there is a pair such that $a_{i_j} < a_{i_{j+1}}$. We could extend the increasing subsequence of length s ending at a_{i_j} by adding the term $a_{i_{j+1}}$ at the end to get an increasing subsequence of length $s + 1$ ending at $a_{i_{j+1}}$. However, this contradicts $t(i_{j+1}) = s$, which says that longest increasing subsequence ending at $a_{i_{j+1}}$ has length s. □

Subsets without divisors

Let $[2n] = \{1, \ldots, 2n\}$. Suppose you want to pick a subset $S \subset [2n]$ so that no number in S divides another. How many numbers can you pick? You can certainly take $S = \{n+1, n+2, \ldots, 2n\}$ as the ratio of any two numbers $a < b$ in this set satisfy $b/a < 2$. This set has size n. Can you pick more then n numbers? The answer is negative.

Theorem 3. *For any subset $S \subset [2n]$ of size $|S| > n$, there are two distinct numbers $b, c \in S$ such that b divides c.*

Proof. For each odd number $a \in [2n]$, we make a box $C_a = \{2^k a : k \text{ is a nonnegative integer}\}$. There are n such boxes. Further, every $b \in [2n]$ is in exactly one of these boxes as we can write each such integer uniquely in the form $b = 2^k a$ with k a nonnegative integer and $a \in [2n]$ odd by factoring out from b the largest power of 2 that divides b. Note that any two distinct elements in the same box C_a have the property that one of them divides the other. Consider any $S \subset [2n]$ of size $|S| > n$. By the pigeonhole principle, there is a class C_a that contains at least two elements of S. These two elements form the desired pair. □

2 The Probabilistic Method

Ramsey numbers

In the 1950s, the Hungarian sociologist S. Szalai studied friendship relationships between children. He noticed that among 20 children, he was always able to find four children each pair of which were friends, or four children such that no two of them were friends. Before deducing any sociological conclusions, Szalai asked three distinguished mathematicians from Hungary: Erdős, Turán, and Sós. A brief discussion revealed that indeed this is a mathematical phenomenon rather than a sociological one. One can make the friendship graph between children, any two children that are friends form an edge. In any graph on 20 vertices (in fact, 18 suffices) there are four vertices which form a clique (a subset of vertices, each pair of which form an edge) or an independent set (a subset of vertices, no pair of which form an edge). Let's look at a smaller case:

Proposition 4. Among any six people, there are three any two of whom are friends, or there are three such that no two of them are friends.

This is not a sociological claim, but a very simple graph-theoretic statement. In other words, in any graph on 6 vertices, there is a triangle or three vertices with no edges between them.

Proof. Let $G = (V, E)$ be a graph and $|V| = 6$, i.e. G has six vertices. Fix a vertex $v \in V$. We consider two cases.

Case 1: The degree of v is at least 3. In this case, consider three neighbors of v, call them x, y, z. If any two among $\{x, y, z\}$ are friends, we are done because they form a triangle together with v. If not, no two of $\{x, y, z\}$ are friends and we are done as well.

Case 2: The degree of v is at most 2, then there are at least three other vertices which are not neighbors of v, call them x, y, z. In this case, the argument is complementary to the previous one.

If $\{x, y, z\}$ are mutual friends, then we are done. Otherwise, there are two among $\{x, y, z\}$ who are not friends, for example x and y, and then no two of $\{v, x, y\}$ are friends.

□

The number 6 in the above proposition cannot be replaced by 5. This is because the 5-cycle C_5 has neither a triangle nor three vertices with no edges between them.

The above proposition is a special case of *Ramsey's theorem* proved in 1930, which is a foundational result in *Ramsey theory*. This consists of a large body of deep results in mathematics that which roughly say, according to Motzkin, that "complete disorder is impossible." In other words, any very large system contains a large well-organized subsystem.

Definition 1. The Ramsey number $R(s, t)$ is the minimum n such that ever graph with n vertices contains a clique of order s or an independent set of order t.

So Proposition 4 can be stated as $R(3, 3) = 6$.

By replacing a graph by its complement, we can deduce that $R(s, t) = R(t, s)$. Further, we have $R(2, t) = t$ as any graph on t vertices either contains an edge or is an independent set of order t, and any smaller empty graph has neither an edge nor an independent set of order t.

Theorem 5. *For any positive integers s and t, the Ramsey number $R(s, t)$ exists. Further, it satisfies*

$$R(s, t) \leq \binom{s + t - 2}{s - 1}.$$

The bound given here is due to Erdős and Szekeres, and is considerably better then the bound in Ramsey's proof.

Proof. We claim that

$$R(s, t) \leq R(s - 1, t) + R(s, t - 1), \tag{1}$$

and then deduce the desired theorem. To show this, let $n = R(s - 1, t) + R(s, t - 1)$, and consider any graph G on n vertices. Fix any vertex v. We consider two cases:

Case 1: The degree of v is at least $R(s - 1, t)$. Then, by the definition of $R(s, t - 1)$, the set of neighbors of v either contains a clique of order $s-1$, or an independent set of order t. In the second case, we are done as this is an independent set in G. In the first case, we can extend the clique by adding v, and hence G contains a clique of order s, completing this case.

Case 2: The degree of v is at most $R(s-1, t)-1$. In this case, v has at least $n-1-(R(s-1, t)-1) = R(s, t - 1)$ nonneighbors. Then, by the definition of $R(s, t - 1)$, the set of nonneighbors of v either contains a clique of order s, or an independent set of order $t - 1$. In the first case, we are done as this is a clique in G. In the second case, we can extend the independent set by adding v, and hence G contains an independent set of order t, completing the proof of the claim.

Given (1), it follows by induction that these Ramsey numbers are finite. Moreover, we get an explicit bound. First $R(s, t) \leq \binom{s+t-2}{s-1}$ holds in the base case $s = 1$ or $t = 1$ since every graph contains a clique of order 1 and an independent set of order 1. The inductive step is:

$$R(s, t) \leq R(s - 1, t) + R(s, t - 1) \leq \binom{s + t - 3}{s - 2} + \binom{s + t - 3}{s - 1} = \binom{s + t - 2}{s - 1},$$

where the equality is Pascal's identity for binomial coefficients. □

How good is the above bound for the diagonal case $s = t$? We get the upper bound

$$R(s,s) \leq \binom{2s-2}{s-1} \leq \frac{4^s}{\sqrt{s}}.$$

This upper bound has not been significantly improved in roughly 70 years! All we know currently is that the exponential growth is the right order of magnitude, but the base of the exponential is not known. The following is an old lower bound of Erdős. Note that to get a lower bound, we need to show that there is a large graph without cliques and independent sets of a certain order. This is quite difficult to achieve by an explicit construction. (The early lower bounds on $R(s,s)$ were only polynomial in s.)

The amazing thing about Erdős' proof is that he never presents a specific graph. He simply shows that one exists by considering a *random* graph almost always works. This was one of the first occurences of the *probabilistic method* in combinatorics. The basic idea is the following. Suppose one wants to show that a structure with certain desired properties exists. One then creates a probability space on the set of structures, and shows that the desired properties hold with positive probability, and hence there exists a structure with the desired properties. The probabilistic method has been used in discrete mathematics ever since with phenomenal success. The statements proved with the method have (typically) nothing to do with probability; probability is introduced for the purpose of the proof.

Theorem 6. *For $s \geq 3$,*

$$R(s,s) \geq 2^{s/2}.$$

Proof. Let n be the largest integer less than $2^{s/2}$. Consider a random graph G on n vertices, where each pair is an edge with probability $1/2$ chosen independently from the other edges. For any particular set S of s vertices, the probability that S forms a clique is $2^{-\binom{s}{2}}$, and the probability that S forms an independent set is $2^{-\binom{s}{2}}$. Since these are disjoint events, the probability that S forms a clique or independent set is $2^{1-\binom{s}{2}}$. The number of such sets S on s vertices is $\binom{n}{s}$. By linearity of expectation, the expected number of cliques or independent sets in the random graph G is

$$\binom{n}{s} 2^{1-\binom{s}{2}} = \frac{n!}{s!(n-s)!} \frac{2}{2^{s(s-1)/2}} < \frac{2n^s}{s! 2^{s(s-1)/2}} \leq \frac{2^{1+s/2}}{s!} < 1,$$

the inequality $\leq$ comes from the constraint on n and s. Since the number of cliques or independent sets of order s is a nonnegative integer, and the expected (i.e. average) number is less than one, there must be an instance for which there is no clique or independent set of order s. We conclude that $R(s,s) \geq 2^{s/2}$. □

Determining Ramsey numbers exactly, even for rather small values of s, is a notoriously difficult problem. It is known that $R(4,4) = 18$, but even $R(5,5)$ is not known (it is known to be between 43 and 49), and determining $R(6,6)$ seems hopeless (it is between 102 and 165). Paraphrasing a (rather silly) quote of Paul Erdős[1]:

Suppose aliens invade the earth and threaten to obliterate it in a year's time unless human beings can find the Ramsey number for five and five. We could marshal the world's best minds and

[1]From "Ramsey Theory" by R. L. Graham and J. H. Spencer, in *Scientific American* (July 1990), p. 112–117.

fastest computers, and within a year we could probably calculate the value. If the aliens demanded the Ramsey number for six and six, however, we would have no choice but to launch a preemptive attack.

We consider a few more examples of the probabilistic method.

Cuts in graphs

Theorem 7. *For every graph $G = (V, E)$, there is a vertex partition $V = U \cup W$ such that at least half of the edges of G have one vertex in U and the other in W.*

Proof. Consider a random partition $V = U \cup W$ with each vertex in U with probability $1/2$ and otherwise it is in W, picked independently of the other vertices. For a given edge $\{u, w\}$ the probability that one of its vertices is in U and the other is in W is $1/2$. Let X be the random variable counting the number of edges with one vertex in U and the other in W. By linearity of expectation, $\mathbb{E}[X] = |E|/2$. Since there is an instance for which $|X| \geq \mathbb{E}[X]$, there is a partition $V = U \cup W$ for which the number of edges with one vertex in each part is at least $|E|/2$. □

Tournaments with many Hamiltonian paths

A *tournament* is an orientation of a complete graph. That is, for any two vertices, the tournament contains exactly one of the two possible directed edges (u, v) and (v, u). You might want to think of a round robin tournament, where each pair of players play a game, always with a single winner, and (u, v) denotes that player u won against v in the game they played. A *Hamiltonian path* in a tournament is a directed path passing through all of the vertices. Szele in 1943 proved the following result showing the existence of tournaments with many Hamiltonian paths.

Theorem 8. *There is a tournament on n vertices that has at least $2^{1-n}n!$ Hamiltonian paths.*

Proof. We calculate the expected number of Hamiltonian paths in a random tournament on n vertices, where every edge has a random direction, each possiblity with probability $1/2$, independently of the other edges. For a given ordering $v_1, \ldots, v_n$ of the n vertices, the probability that this ordering forms a Hamiltonian path with (v_i, v_{i+1}) an edge for $1 \leq i \leq n-1$ is 2^{1-n}. This is because each of the $n-1$ edges have a probability $1/2$ of having a certain direction, independently of the other edges. As there are $n!$ such orderings, by linearity of expectation, the expected number of Hamiltonian paths in this random tournament is $2^{1-n}n!$. So there is a tournament with at least the expected number, $2^{1-n}n!$, of Hamiltonian paths. □

Van der Waerden's theorem

As a final example, we consider an application about coloring integers. Recall the concept of an *arithmetic progression*; a subset of "equally spaced" numbers. For our purposes, a *k-term arithmetic progression* is a subset

$$\{a, a+b, a+2b, \ldots, a+(k-1)b\}$$

of $\mathbb{N}$, for some $a, b \in \mathbb{N}$. So for example, $1, 3, 5, 7$ and $7, 11, 15, 19$ are both 4-term arithmetic progressions.

Next, we define the idea of a *coloring* of a subset. Let S be a subset of $\mathbb{N}$ (e.g., $\{1, 2, \ldots, 100\}$). A 2-coloring of S simply assigns each element of S either "red" or "blue". (More formally, a 2-coloring is a map $\chi : S \to \{1, 2\}$, where 1 and 2 represent red and blue.)

The following theorem is known as "Van der Waerden's Theorem"; it's another example of a result in Ramsey Theory.

Theorem 9. *For every $k \in \mathbb{N}$, there exists an $n \in \mathbb{N}$ such that for every 2-coloring of $\{1, 2, \ldots, n\}$, there exists some k-term arithmetic progression which is monochromatic (i.e., whose elements all have the same color).*

We won't prove this theorem; it doesn't need a lot of background though, so you could try to read the proof if you're interested.

Now let $W(k)$ be the smallest possible choice of n in the theorem, for a particular value of k. In other words, $W(k)$ is such that any 2-coloring of $\{1, 2, \ldots, W(k)\}$ contains a monochromatic k-term arithmetic progression; but there exists some coloring of $\{1, 2, \ldots, W(k) - 1\}$ with *no* monochromatic k-term arithmetic progression. For example, $W(2) = 3$; this is easy to see just by trying all the possible colorings. It turns out that $W(3) = 9$; that's not as easy to prove, though it's not hard to find a coloring of $\{1, \ldots, 8\}$ with no k-term arithmetic progressions:

1 2 3 4 5 6 7 8.

If you try to color one more, you'll find you get stuck, no matter what you do! These van der Waerden numbers have received lots of interest; we don't know too much about them, and in particular we only have exact values for $W(k)$ for fairly small values of k.

We are going to prove a *lower bound* for $W(k)$:

Theorem 10. *For any $k \in \mathbb{N}$,*

$$W(k) \geq \sqrt{k-1} \cdot 2^{(k-1)/2}.$$

Proof. Let n be the largest integer strictly smaller than $\sqrt{k-1} \cdot 2^{(k-1)/2}$. Our goal is to show that there exists a "good" coloring of $\{1, 2, \ldots, n\}$: in other words, a coloring with *no* monochromatic k-term arithmetic progressions. However, we will not explicitly construct a good coloring. Instead, we are going to prove that a certain *randomly chosen* coloring has a positive probability of being good. This implies that there exists a good coloring; if there was no good coloring, the chance of getting a good coloring would have to be zero!

So define a random coloring of $\{1, 2, \ldots, n\}$ as follows: for each i between 1 and n, color i red with probability $1/2$, and color it blue with probability $1/2$, and do this independently for each i. In other words, we toss an unbiased coin n times, one for each position, to determine the colors. Now define the random variable

$$f = \#\text{monochromatic } k\text{-term arithmetic progressions in } \{1, \ldots, n\}.$$

Our goal is to show that $\Pr(f = 0) > 0$. Our strategy to prove that is to show that $\mathbb{E}(f) < 1$; this is sufficient, by the following Lemma:

Lemma 1. *Let g be any random variable taking on only nonnegative integer values, and satisfying $\mathbb{E}(g) < 1$. Then $\Pr(g = 0) > 0$.*

Proof. We have

$$\begin{aligned}\mathbb{E}(g) &= \sum_{i=0}^{\infty} \Pr(g=i)i \\ &\geq \sum_{i=1}^{\infty} \Pr(g=i) \quad \text{(note the change in summation index)} \\ &= 1 - \Pr(g=0);\end{aligned}$$

the last line follows since $\sum_{i=0}^{\infty} \Pr(g=i) = 1$. Rearranging,

$$\Pr(g=0) \geq 1 - \mathbb{E}(g) > 0.$$

□

(If you think about it a bit, the lemma should seem quite intuitive; if the average is less than 1, and the random variable is nonnegative, it had better be less than 1 at least some of the time; since it's integral, it must be zero some of the time.)

In order to get a handle on $\mathbb{E}(f)$, we will use linearity of expectation. Define, for any $a, b \in \mathbb{N}$,

$$Q_{a,b} = \{a, a+b, \ldots, a+(k-1)b\}.$$

Then define for any a, b such that $Q_{ab} \subseteq \{1, \ldots, n\}$ the indicator random variable

$$I_{a,b} = \begin{cases} 1 & \text{if } Q_{a,b} \text{ is monochromatic} \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$f = \sum_{(a,b):Q_{a,b}\subseteq\{1,\ldots,n\}} I_{a,b};$$

every monochromatic k-term arithmetic progression will be counted once in the right hand side. Thus, using linearity of expectation,

$$\begin{aligned}\mathbb{E}(f) &= \sum_{(a,b):Q_{a,b}\subseteq\{1,\ldots,n\}} \mathbb{E}(I_{a,b}) \\ &= \sum_{(a,b):Q_{a,b}\subseteq\{1,\ldots,n\}} \Pr(Q_{a,b} \text{ is monochromatic}).\end{aligned}$$

Now it's not hard to see that for any a, b,

$$\Pr(Q_{a,b}) \text{ is monochromatic}) = \frac{1}{2^k} \cdot 2 = \frac{1}{2^{k-1}};$$

$Q_{a,b}$ has k elements, each red with probability $1/2$; so the probability that they are all red is $1/2^k$, and the same for all blue. So to get a bound on $\mathbb{E}(f)$, we just need to bound the number of terms in the sum.

So how many k-term arithmetic progressions are there, roughly? Well certainly a is between 1 and n (actually it can't be more than $n-k$, but we just want a rough bound). The final

term of the arithmetic progression must be at most n, and so $a + (k-1)b \leq n$. Thus certainly $b \leq n/(k-1)$. Since each k-term arithmetic progression is determined by the pair (a, b), there are at most $n^2/(k-1)$ of them.

So all in all, we have

$$\mathbb{E}(f) \leq \frac{n^2}{k-1} \cdot \frac{1}{2^{k-1}};$$

considering how we defined n, we see that we have reached our goal: $\mathbb{E}(f) < 1$. Thus, by the Lemma, $\Pr(f = 0) > 0$, and there must exist a good coloring.

□

The bound of the above theorem could be improved slightly, by doing a more precise count of the number of k-term arithmetic progressions.

Again, note that this proof does not tell us *how* to find a good coloring (except perhaps by trying random colorings until we find one that works), and doesn't tell us what these good colorings look like. There are many examples now of situations where we can prove that a certain structure exists, using the probabilistic method, but have no explicit deterministic construction!