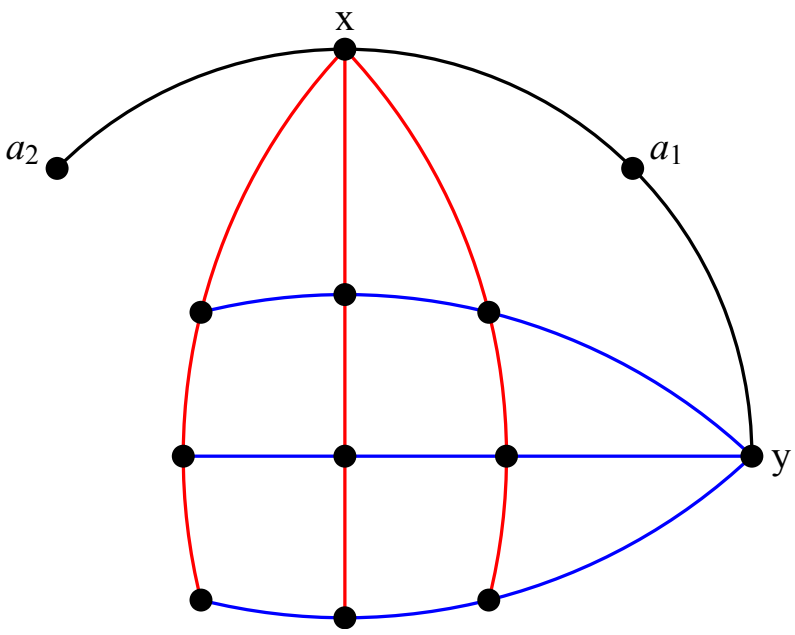
MA3J2 Combinatorics II

Keith Ball



Contents

Chapter 0. Introduction

This course continues the second year Combinatorics course. As in that course we shall cover several different topics which are related more by the style of argument than by the nature of the mathematical objects we study. The choice of topics will show that combinatorics is closely related to other areas of mathematics and has many uses. For example we shall introduce error-correcting codes which enable us to send messages accurately over noisy communication channels or to store information on media that may get corrupted. The CDs and DVDs that you play have information encoded on them in such a way that even if you scratch them or drive over a bump while playing the CD, the music comes out as it should.

Another topic will be graph colouring. This relates to the discussion of planar graphs last year. We shall construct a polynomial that describes how many ways we can colour a graph and which is related to polynomials that occur in statistical mechanics.

To provide a little more structure to the course I will divide it into two volumes whose chapters are loosely related: **Patterns** and **Graph Theory**.

The second volume will follow on quite naturally from several parts of last year's course: the chapters on planar graphs and colouring, trees, Hall's Theorem and Ramsey Theory. The first will introduce several new topics which can loosely be regarded as describing special mathematical structures which exhibit patterns that are important or useful. The latter probably needs some explanation by means of an example.

A Steiner system with parameters n, k and t is a set Ω with n elements and a family $\mathcal{F}$ of subsets of Ω each having k elements, with the property that each t-element subset of Ω is contained in exactly one member of $\mathcal{F}$. For example if $t = 1$ we just want a partition of Ω into sets of size k, which is possible if $k|n$. If $k = t$ then we can take all the sets of size t. If $k = n$ then we can take $\mathcal{F} = \{\Omega\}$.

Thus a Steiner system is a family of subsets which has a certain uniformity or symmetry to it. We shall be studying some special examples of Steiner systems in the case $t = 2$: the so-called finite projective planes. In this case we want a family of sets of size k with each pair of points contained in exactly one of them. Each one contains $\binom{k}{2}$ pairs and the total number of pairs is $\binom{n}{2}$. So the number of sets in the family is

$$\frac{\binom{n}{2}}{\binom{k}{2}}$$

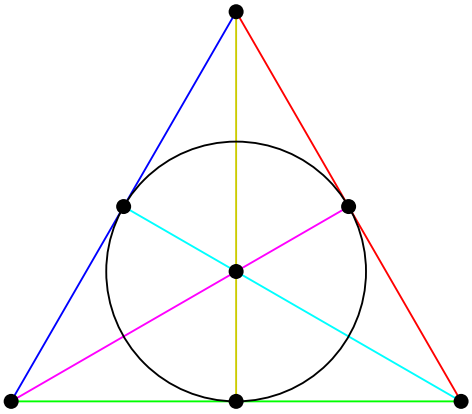
so we need $k(k-1)$ to be a factor of $n(n-1)$.

Now we can also consider how many sets contain a fixed element x. Each of these contain $k-1$ pairs $\{x, y\}$ and there are $n-1$ such pairs altogether. So the number of such sets is

$$\frac{n-1}{k-1}.$$

So we also need $k-1$ to be a factor of $n-1$.

The simplest non-trivial case is $t=2$, $k=3$ and $n>3$. We need $2|n-1$ so n is odd and $6|n(n-1)$ so $n=5$ won't work. The simplest case is thus $t=2$, $k=3$ and $n=7$. This example exists and is called the Fano plane.



Each pair of points is contained in exactly one "line". This Steiner system has the additional property that each pair of lines meet in exactly one point.

As we saw, for a Steiner system with parameters n, k and $t=2$ to exist we need $k-1|n-1$ and $k(k-1)|n(n-1)$ while if $t=1$ we just need $k|n$. For general t there are t conditions of a similar type. It is not obvious that a Steiner system exists just because these conditions hold. It was proved by Wilson in the 70's that Steiner systems exist for $t=2$ whenever the parameters satisfy the appropriate divisibility conditions and n is large enough and by Keevash very recently for general t.

The two volumes will be divided up as follows.

Patterns

1. Projective planes and Latin squares
2. Error-correcting codes
3. Discrete geometry
4. Partially-ordered sets

Graph Theory

5. Graph colouring: the chromatic polynomial
6. Matroids
7. Random graphs
8. The regularity method

Volume I. Patterns

Chapter 1. Projective planes and Latin squares

Projective planes

There is a deep connection between algebra and geometry but once one moves beyond linear algebra to questions about polynomial maps there is a certain inconvenience to the usual vector space structure in which we do geometry. This inconvenience stems from an asymmetry. Any two points in the plane belong to a line but it is not true that any two lines meet in a point: they could be parallel. Motivated in part by the theory of perspective in painting, mathematicians realised that they could study an extended geometry: projective geometry: in which two lines can meet "at infinity".

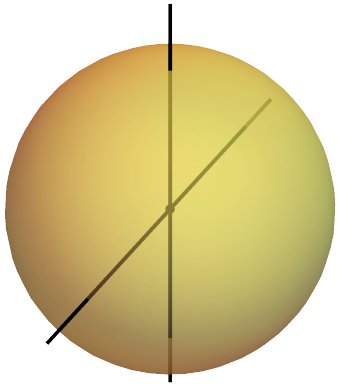
In this chapter we shall study the discrete analogue in which our projective planes have only finitely many points. The basic example is as follows. Let $\mathbf{F}$ be a field and consider the vector space $\mathbf{F}^3$ with the point $(0,0,0)$ removed: $\mathbf{F}^3 - \{0\}$. Regard two points (a,b,c) and (x,y,z) as equivalent if there is a (non-zero) element λ of the field so that

$$(a,b,c) = \lambda(x,y,z).$$

The projective plane over $\mathbf{F}$, denoted $\mathbf{FP}^2$ consists of the equivalence classes of non-zero points in $\mathbf{F}^3$.

Equivalently we can think of the elements of the projective plane as the lines in $\mathbf{F}^3$ that pass through the origin. Any two points which are multiples of one another lie on the same line through the origin. How can we picture this space of lines in the case $\mathbf{F} = \mathbf{R}$?

Consider a sphere centred at the origin. Each line exits the sphere at two antipodal points. So we can think of the collection of lines as being the sphere but with every pair of opposite points identified with one another.



Locally the set of lines looks like a patch of the sphere. But we can't see how to join the patches together in 3 dimensions.

A line in the projective plane corresponds to a great circle on the sphere (with opposite points identified). This in turn consists of all points which are perpendicular to a given direction: all points (x, y, z) satisfying an equation

$$\lambda x + \mu y + \nu z = 0$$

where λ, μ and ν are elements of $\mathbf{F}$ that are not all zero.

In this chapter we shall study the discrete analogue in which our projective planes have only finitely many points. If the field is finite with q elements then $\mathbf{F}^3 - \{0\}$ has $q^3 - 1$ elements. Each equivalence class has $q - 1$ elements because there are $q - 1$ non-zero elements of $\mathbf{F}$ to scale by. So there are

$$\frac{q^3 - 1}{q - 1} = q^2 + q + 1$$

elements of $\mathbf{FP}^2$.

A line in this space consists of the equivalence classes of all points satisfying an equation

$$\lambda x + \mu y + \nu z = 0$$

where λ, μ and ν are elements of $\mathbf{F}$ that are not all zero. Note that the definition makes sense because if two points are in the same equivalence class then either they

both satisfy

$$\lambda x + \mu y + \nu z = 0$$

or neither does.

Notice that if we multiply each of the coefficients λ, μ and ν by a non-zero element of the field then we get the same solutions: the same set of points satisfying the equation. We can thus think of the line as being given not by the coefficient sequence (λ, μ, ν) itself but by an equivalence class of sequences: a point of $\mathbf{FP}^2$ rather than of $\mathbf{F}^3 - \{0\}$. Each point of $\mathbf{FP}^2$ determines a line in $\mathbf{FP}^2$. It might be that two distinct points of the projective plane gave rise to the same line: the same set of solutions. We shall check that this doesn't happen.

Suppose (λ, μ, ν) and (α, β, γ) are non-zero points defining the same line. Assume that $\lambda \neq 0$. Note that $(\mu, -\lambda, 0)$ lies on the line so

$$\alpha\mu - \beta\lambda = 0.$$

Hence $\beta = \lambda^{-1}\alpha\mu$ and in the same way $\gamma = \lambda^{-1}\alpha\nu$. But this implies that

$$(\alpha, \beta, \gamma) = \lambda^{-1}\alpha(\lambda, \mu, \nu)$$

showing that our two points (λ, μ, ν) and (α, β, γ) belong to the same equivalence class.

Let's look at the simplest case in which $\mathbf{F} = \mathbf{Z}_2$: the field of two elements $\{0, 1\}$ in which $1 + 1 = 0$. In this case the equivalence classes only have one element each so we can think of the points in the projective plane as the seven points

$$(1,0,0),\ (0,1,0),\ (0,0,1),\ (0,1,1),\ (1,0,1),\ (1,1,0),\ (1,1,1).$$

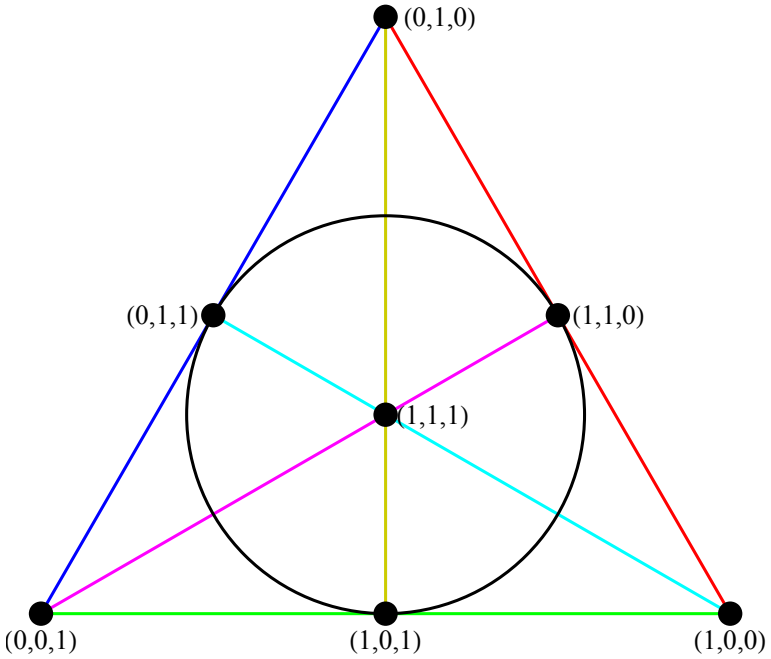
Each line contains 3 points: for example $x = 0$ is satisfied by the points

$$(0,1,0),\ (0,0,1),\ (0,1,1)$$

while $x + y + z = 0$ is satisfied by

$$(0,1,1),\ (1,0,1),\ (1,1,0).$$

This projective plane is called the Fano plane and is often drawn as shown. The lines have different colours.



As we would expect, any two distinct points determine one line.

Lemma (**Points in the projective plane**)**.** *Any two distinct points in* $\mathbf{FP}^2$ *belong to exactly one line.*

Proof Let the points be (equivalence classes of) (a, b, c) and (x, y, z). Form the vector product:

$$(\lambda, \mu, \nu) = (bz - cy, cx - az, ay - bx).$$

It is easy to check that $\lambda a + \mu b + \nu c = 0$ and $\lambda x + \mu y + \nu z = 0$ so both our points lie on the line defined by λ, μ and ν. We need to check that the coefficients are not all zero. Suppose that they were. We know that a, b and c are not all zero so we can assume that $a \neq 0$.

Now set $\theta = a^{-1}x$. Using the fact that $cx - az = ay - bx = 0$ we get

$$\begin{aligned} x &= \theta a \\ y &= \theta b \\ z &= \theta c \end{aligned}$$

showing that (a, b, c) and (x, y, z) belong to the same equivalence class and contradicting the fact that they are distinct.

Finally we want to show that there is only one such line. Suppose that

$$\begin{aligned} \lambda a + \mu b + \nu c &= 0 \quad (1) \\ \lambda x + \mu y + \nu z &= 0. \quad (2) \end{aligned}$$

We want to show that (λ, μ, ν) is equivalent to the vector product that we know gives one such line. We know that $bz - cy$, $cx - az$ and $ay - bx$ are not all zero so assume that $\phi = bz - cy \neq 0$. Multiplying (1) by z and (2) by c and subtracting we get

$$\lambda(cx - az) = \mu(bz - cy) = \mu\phi$$

and in a similar way we have

$$\lambda(ay - bx) = \nu(bz - cy) = \nu\phi.$$

We thus have

$$\begin{aligned} \lambda &= \phi^{-1}\lambda(bz - cy) \\ \mu &= \phi^{-1}\lambda(cx - az) \\ \nu &= \phi^{-1}\lambda(ay - bx) \end{aligned}$$

as required. □

What the preceding proof shows is that if (a, b, c) and (x, y, z) are non-equivalent non-zero elements of $\mathbf{F}^3$ then there is an unique equivalence class of elements (λ, μ, ν) of $\mathbf{F}^3 - \{0\}$ satisfying

$$\lambda a + \mu b + \nu c = \lambda x + \mu y + \nu z = 0.$$

This shows that any two distinct points lie on an unique line but it *also* shows that any two distinct lines meet in an unique point.

Lemma (**Lines in the projective plane**)**.** *Any two distinct lines in* $\mathbf{FP}^2$ *meet in exactly one point.*

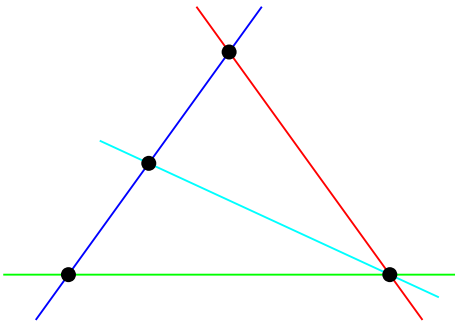
Based on the algebraic construction we shall define a combinatorial object that we shall call *a finite projective plane.*

Definition (**Finite Projective Plane**). *A finite projective plane is a finite set P of points and a set L of subsets of P called lines satisfying*

- *Each pair of points belong to exactly one line*
- *Each pair of lines contain exactly one common point*
- *There are 4 points, no three of which belong to a single line*

The third condition rules out certain degenerate cases which fail to possess the kind of symmetry properties that we would like. Finite projective planes have a very regular structure which we shall prove in the theorem below. To begin with we have a simple lemma.

Lemma (**Point-Line matching**). *Let (P, L) be a FPP, l one of its lines and p a point* not *on l. Then the number of points on l is equal to the number of lines through p.*



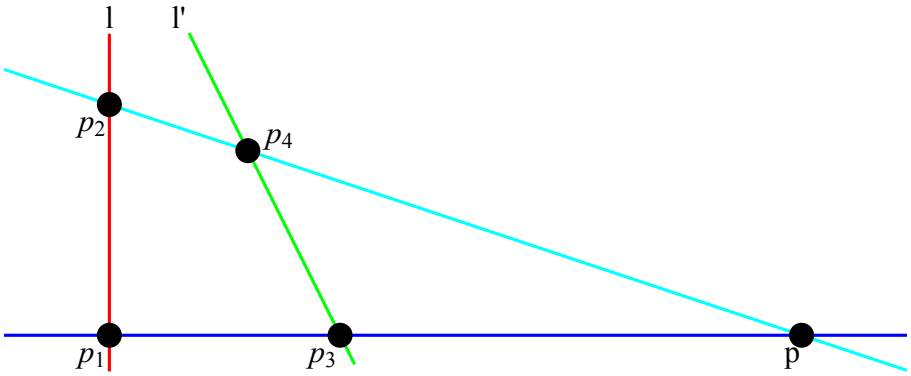
Proof Each point on l lies on exactly one line through p and each line through p meets l in exactly one point. □

Theorem (**FPP structure**). *Let (P, L) be a FPP. Then there is a number q with the property that*

1. *Each line contains $q + 1$ points*
2. *Each point lies on $q + 1$ lines*
3. *There are $q^2 + q + 1$ points*
4. *There are $q^2 + q + 1$ lines*

The number q is called the order *of the FPP.*

Proof 1. It suffices to show that any two lines contain the same number of points (and call this $q+1$). Suppose l and l' are two lines. By the lemma it is enough to find a point p not on either line since in that case each line has as many points as there are lines through p. Consider the 4 points guaranteed by the third property: p_1, p_2, p_3 and p_4. If one of them is in neither line we are done. So assume they are all in l or l' and observe that there must be two in each line: say $p_1, p_2 \in l$ and $p_3, p_4 \in l'$. Now consider the line l_{13} containing the points p_1 and p_3 and the line l_{24} containing the points p_2 and p_4. These lines meet at a point p.



I claim that p cannot lie on l or l'. If p were on l then l_{13} and l would be the same line since they also contain p_1. Then we would have p_3 on l as well as p_1 and p_2 contradicting the 4 point property. Similarly p cannot be on l'.

2. Let p be a point. By the lemma it suffices to find a line not containing p. Look at the 4 points p_1, p_2, p_3 and p_4 and assume that $p \neq p_1$. Then the line containing p_1 and p_2 and the line containing p_1 and p_3 cannot both contain p since they already meet at p_1.

3. Let p be a point and consider the $q+1$ lines passing through p. No two of them meet except at p, each contains q points other than p and they cover the plane. So the total number of points is $(q+1)q+1 = q^2+q+1$.

4. Each point belongs to $q+1$ lines and each line contains $q+1$ points so the number of lines must equal the number of points. □

The number q defined in the preceding theorem is called the order of the FPP. We have seen that if F is a field of q elements then there is an FPP of order q. There exist finite fields of every prime power order $q = p^k$ where p is prime and $k \geq 1$. There are FPPs that do not come from fields but it is conjectured that all FPPs have prime power order.

Latin squares

A Latin square is an $n \times n$ arrangement of symbols: each one of n symbols occurring once in every row and once in every column:

$$\begin{array}{cc} A & B \\ B & A \end{array} \qquad\qquad \begin{array}{ccc} A & B & C \\ B & C & A \\ C & A & B \end{array}$$

Any group table is an example. For C_4

	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

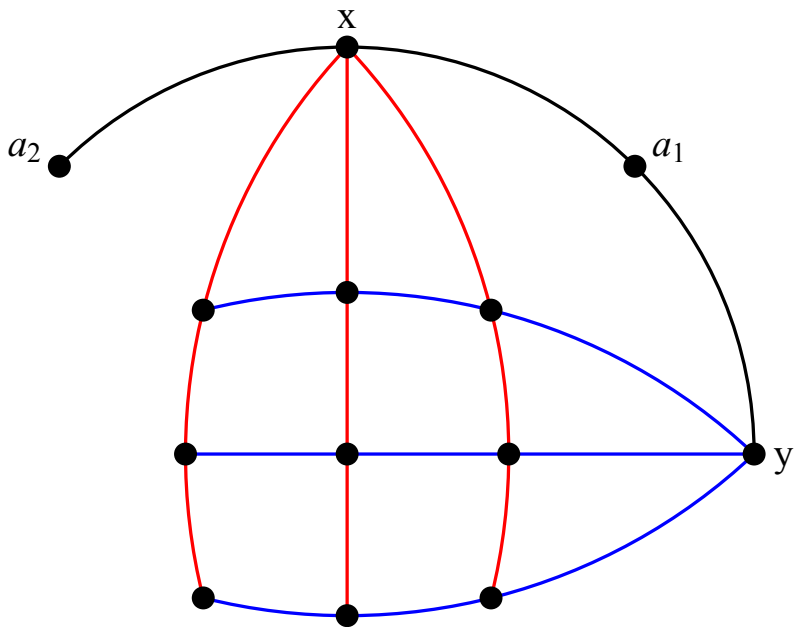
Two $n \times n$ Latin squares $A = (a_{ij})$ and $B = (b_{ij})$ are said to be orthogonal if the n^2 pairs (a_{ij}, b_{ij}) include all possible pairs of one symbol from A and one from B. For example if we look at

$$\begin{array}{ccc} A & B & C \\ B & C & A \\ C & A & B \end{array} \quad \text{and} \quad \begin{array}{ccc} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{array}$$

and merge the two squares then the pairs are

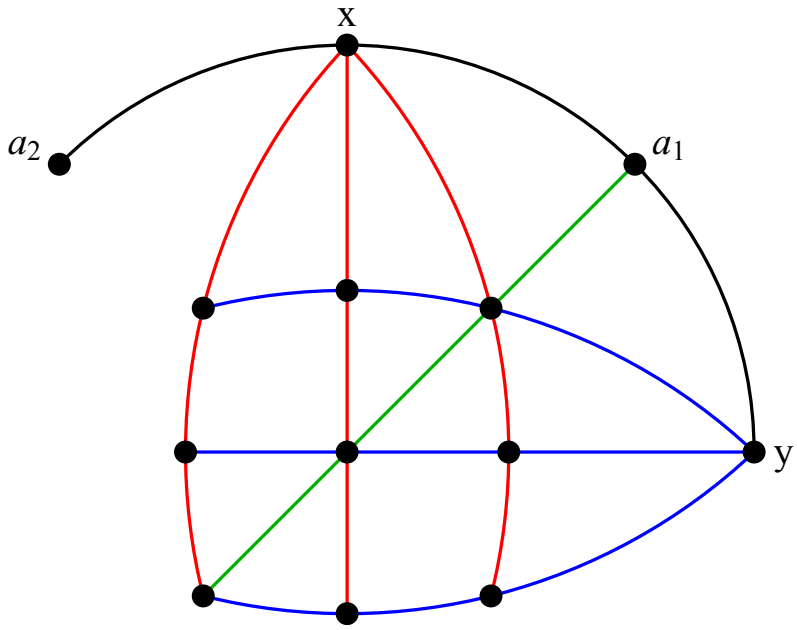
$$\begin{array}{ccc} (A,1) & (B,2) & (C,3) \\ (B,3) & (C,1) & (A,2) \\ (C,2) & (A,3) & (B,1) \end{array}$$

To finish this chapter we shall investigate a rather surprising link between FPPs and orthogonal Latin squares. In order to do so we build up the FPP from two points as follows.



Choose two points in the FPP: say x and y. They lie on a line containing $q-1$ other points: $a_1, \ldots, a_{q-1}$. The FPP contains q^2 other points. The point x lies on q other lines, the x-lines, that cover the other q^2 points. The point y lies on q other lines, the y-lines, that also cover the other q^2 points. Each x-line meets each y-line. The q^2 intersections are the other points which thus have a Cartesian product structure: a grid.

a_1 also lies on q other lines that cover the grid. Each of these meets each x-line and each y-line. So each line through a_1 locates q places in the grid: one in each row and one in each column. Hence these q lines through a_1 generate a Latin square in the grid. Call the symbols in the grid a_{11} for the first line, a_{12} for the second line and so on. The $q-1$ points $a_1, \ldots, a_{q-1}$ on the xy-line thus generate $q-1$ Latin squares.



In the example above the point a_1 generates

$$\begin{array}{ccc} a_{13} & a_{11} & a_{12} \\ a_{11} & a_{12} & a_{13} \\ a_{12} & a_{13} & a_{11} \end{array}$$

while the point a_2 generates

$$\begin{array}{ccc} a_{22} & a_{21} & a_{23} \\ a_{23} & a_{22} & a_{21} \\ a_{21} & a_{23} & a_{22}. \end{array}$$

I claim that these $q-1$ Latin squares are pairwise orthogonal. Let a_1 and a_2 be two of the other points on the xy line. Each lies on q other lines corresponding to the symbols a_{1j} and a_{2j} used in their Latin squares. Each a_1 line meets each a_2 line at one of the q^2 points of the grid. So every possible pair of symbols appears when the grids are merged. This process can be reversed as you will show in the HW.

Theorem (**FPPs and Latin squares**)**.** *There is an FPP of order $q > 1$ if and only if there are $q-1$ pairwise orthogonal $q \times q$ Latin squares.*

Chapter 2. Error-correcting codes

Introduction

Suppose you want to send a message which you represent in binary: as a string of 0s and 1s. You can send each bit: a 0 or a 1. But there is a chance that the receiver will detect the opposite. To avoid this you could send each bit several times over: instead of sending 1 you send a string of ten: 1111111111. If the receiver detects 1111011101 s/he can be pretty sure that what was sent was meant to be 1. But if you do this, the rate at which you transmit your messages is decreased by a factor of ten.

The procedure just described of replacing

0 by 0000000000

1 by 1111111111

constitutes an *error-correcting code*. You insist that only certain strings of ten bits are allowed and you make sure that they look very different from one another so that there is no danger that one will be converted into another by noise.

It turns out, rather remarkably, that if you design your code carefully you can achieve almost perfect accuracy with only a limited decrease in the rate of transmission. To get an idea of why this is let us compare two different codes. The first will be the simple repetition code

0 is replaced by 00

1 is replaced by 11

This code detects one bit errors: if you receive 01 you know that there has been an error. In order for you to receive the wrong message the channel has to corrupt two bits instead of 1. The rate of this code is 1/2.

Now suppose we use this code to send two bits worth of information. We encode as follows

00 is replaced by 0000

10 is replaced by 1100

01 is replaced by 0011

11 is replaced by 1111

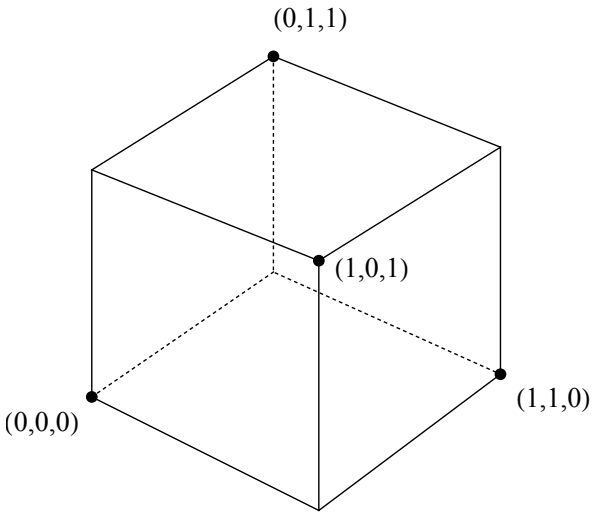
We use 4 bits to encode 2 bits worth of information and the channel can only fool us by switching 2 bits.

However using 3 bits we can get the same defence against noise, provided we encode carefully.

00 is replaced by 000
10 is replaced by 110
01 is replaced by 011
11 is replaced by 101

Any two of these strings differ in 2 places so to fool us, the channel has to corrupt 2 bits, but this code has rate 2/3 instead of 1/2. So by choosing the admissible strings carefully you can get the same accuracy at a better rate.

The collection of admissible strings used for this code 000 110 011 101 is a special structure within the collection of all strings of length 3.



The points we use are well spaced and exhibit a geometric pattern.

Block Codes

We shall examine only what are called *block codes*, in which the message you send is broken into blocks, each of which can be decoded without reference to the rest of the message. We called these blocks the *admissible strings* earlier. A more common term is *codeword*. In order to use the code you need to decide how to map the messages you want to send onto the codewords that you actually transmit. We saw an example in which we encoded the actual messages as

00 is replaced by 000
10 is replaced by 110
01 is replaced by 011
11 is replaced by 101

For our purposes it won't matter how you map your messages onto the codewords: all that matters is how well-separated the codewords are. The codewords are built out of "letters" from an "alphabet". For most communication the alphabet has just two letters: 0 and 1 and we call the code a *binary* code. We think of 0 and 1 as elements of the field $\mathbf{Z}_2$. Recall that in this field $1+1=0$. If the codewords have length n they can be thought of as elements of the vector space $\mathbf{Z}_2^n$.

The codewords we used above

$$000 \quad 110 \quad 011 \quad 101$$

are elements of $\mathbf{Z}_2^3$ and we can see that they actually form a linear subspace of this vector space. For example

$$110 + 011 = 101.$$

We shall be concerned only with examples like this. If the codewords form a subspace of $\mathbf{Z}_2^n$ we call the code a *linear* code.

In order to specify how good the code is we define the *Hamming distance* of two codewords as the number of positions in which they differ: so

$$110 \quad \text{and} \quad 011$$

are distance 2 apart. We want to find codes in which every pair of codewords are far from one another. Linear codes have a simple feature that makes this easier to achieve: there is a kind of "automatic" separation.

Lemma (**Separation of linear codes**)**.** *Suppose $C \subset \mathbf{Z}_2^n$ is a linear subspace and every element of C, other than $00\ldots0$, contains at least d coordinates equal to 1.*

Then the code is d-separated: every pair of codewords are at least Hamming distance d apart.

Proof Suppose u and v are two codewords that differ only in $d-1$ or fewer places.

Then $u+v=u-v$ is a codeword which has a 1 only in the places where u and v differ, contradicting the hypothesis. □

It is also very simple to determine the *rate* of a linear code. If the code uses n transmitted bits to send k-bits worth of information then the code has rate k/n. So if there

are 2^k codewords in $\mathbf{Z}_2^n$ the rate is k/n. More generally the rate of a binary block code C of length n is

$$\frac{\log_2 |C|}{n}.$$

Now if C is a linear subspace of $\mathbf{Z}_2^n$ of dimension k then it has 2^k elements: so the rate is just

$$\frac{\dim C}{n}.$$

For example if

$$C = \{000, \quad 110, \quad 011, \quad 101\}$$

then the minimum number of 1s in a non-zero codeword is 2. The code is 2-separated. C is a 2-dimensional subspace of $\mathbf{Z}_2^3$ and the rate is $2/3$.

So our aim is to look for k-dimensional subspaces of $\mathbf{Z}_2^n$ whose non-zero elements contain plenty of 1s to get a good separation and we want k to be pretty large to get a high rate. Since we want our subspace to have large dimension it is usually easier to define it by using a small number $n-k$ of linear equations rather than a basis of size k. For example, the code

$$\{000, \quad 110, \quad 101, \quad 011\}$$

consists of the elements (x_1, x_2, x_3) of $\mathbf{Z}_2^3$ which contain an even number of 1s. This means they are the elements satisfying

$$x_1 + x_2 + x_3 = 0.$$

The code is defined by a *parity check*.

The Hamming Codes

The aim of this section is to describe a family of quite efficient 3-separated binary codes. We will do it by using parity check rules. Let r be a positive integer and $n = 2^r - 1$ be the length of the code we are going to describe. The code will be an $n-r$ dimensional subspace of $\mathbf{Z}_2^n$ so we need r linear equations to specify it and its rate will be almost 1

$$\frac{n-r}{n} = 1 - \frac{r}{n} \approx 1 - \frac{\log_2 n}{n}.$$

The linear equations will be arranged into an $r \times n$ matrix B so the code will be

$$C = \{x \in \mathbf{Z}_2^n : Bx = 0\}.$$

The Hamming code is specified by the fact that the *columns* of B are all the binary numbers from 1 to n. If $r = 3$, $n = 7$:

$$B = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Observe that $(1, 1, 1, 0, \ldots, 0) \in C$ what ever the value of r since the first 3 columns of B have 0 below the top two rows. So the code cannot be more than 3-separated. In the HW you will see that it is 3-separated.

We saw that linear codes have the advantage that you can make them well-separated in a fairly simple way. Their structure also makes them relatively simple to decode and this is particularly true for the Hamming codes. What do we mean by "decode". We receive a message and break it into blocks of the right length n. For each block we want to decide which codeword was originally sent. If the block we receive *is* a codeword then we assume that it was the true message. If not we want to find the closest codeword to what we received and will then decode as that codeword. The Hamming code has the property that if x is *not* a codeword then it is distance 1 away from a codeword and from exactly one of them.

Lemma (**Decoding a Hamming code**)**.** *Let C be the Hamming code of length $n = 2^r - 1$. Then if $x \in \mathbf{Z}_2^n$ but $x \notin C$ there is one and only one element of C at distance 1 from x.*

Proof Suppose x is not a codeword. We want to modify one bit and arrive at a codeword. Consider the vector

$$u = Bx \in \mathbf{Z}_2^r.$$

If u were 0 it would tell us that $x \in C$. So we know that u is a non-zero element of $\mathbf{Z}_2^r$.

For each i, if u_i is zero it means that x satisfies the i^{th} parity rule while if $u_i = 1$ it means that x violates that rule. But u is one of the columns of B: indeed, if u is the binary representation of the number m then u is the m^{th} column. So if we change the m^{th} bit of x to obtain $\tilde{x}$ we will correct all the parity rules that x violated and leave unchanged all the ones that x satisfied. So $\tilde{x}$ is a codeword.

On the other hand the columns of B are all different so if we change a single bit other than the m^{th} we will not arrive at a codeword. Thus there is only one codeword at distance 1 from x. □

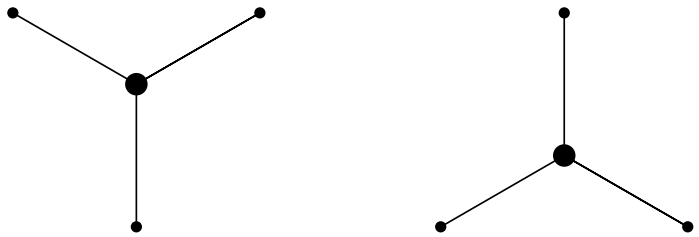
This tells us exactly how to decode the Hamming code. If we receive something other than a codeword we decode as the closest codeword and it is easy to find which that is by the procedure used in the proof.

The code has 2^{n-r} elements. Each one is the closest codeword to a variety of non-codewords: how many?

Lemma (**The domain of a codeword**)**.** *Let C be the Hamming code of length $n = 2^r - 1$. Then each codeword is the closest codeword to n other elements of $\mathbf{Z}_2^n$.*

Proof You have n possible bits to flip. □

Thus each codeword sits at the centre of a little cloud of 2^r strings: itself and $n = 2^r - 1$ others which are adjacent to it.



There are 2^{n-r} codewords, each one decodes the 2^r strings that belong to its cloud. This just covers all 2^n elements of $\mathbf{Z}_2^n$ (showing again that each string can only be adjacent to a single codeword). The Hamming code is a 3-separated code and can correct an error of one bit. If one bit of the original message is altered, we stay in the cloud of the original word and get the correct decoding. More generally, if you have a code that is $(2e + 1)$-separated, it can correct e errors. HW.

The little cloud picture that we have of the Hamming code can be used to get a bound on how good a code can be. Suppose we have a binary code of length n that is $(2e+1)$-separated. Then for each code word x, all the strings that are distance at most e are closer to x than to any other word. So the clouds (or balls) of radius e around the codewords are disjoint from one another. How many strings are there in each ball? How many strings can you get to by changing at most e bits?

If you change 0 bits you stay where you are: 1 string. If you change 1 bit you have a

choice of n to change: n strings. If you change 2 bits: $\binom{n}{2}$ strings. So the balls have

$$1 + n + \binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{e}$$

elements. Since all the balls are disjoint we have

$$|C| \sum_{i=0}^{e} \binom{n}{i} \leq 2^n.$$

We have proved the following.

Lemma (**Sphere-packing bound**). *Let C be a $(2e+1)$-separated binary code of length n. Then*

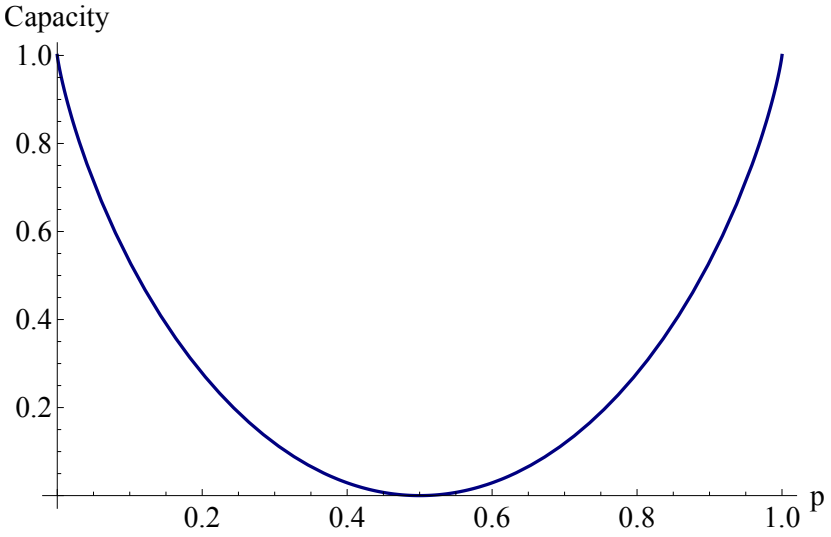
$$|C| \sum_{i=0}^{e} \binom{n}{i} \leq 2^n.$$

Shannon's Theorem

The subject of coding theory was really started by Claude Shannon who demonstrated the fact that was mentioned in the introduction: that by choosing a good code we can communicate with almost perfect accuracy up to a fixed rate that depends only upon how good our communication channel is.

Suppose we have a binary communication channel (sending 0 and 1) which flips each bit with probability p. Let $R = 1 + p \log_2 p + (1-p) \log_2 (1-p)$. This is called the capacity (or Shannon capacity) of the channel.

Theorem (**Shannon's Coding Theorem**). *Using a binary communication channel with probability p of flipping each bit we can find a code with rate almost $1 + p \log_2 p + (1-p) \log_2 (1-p)$ which achieves almost perfect accuracy.*



Proof (Sketch) Choose a large value of n for the length of a block code. The number of bits that get flipped has a binomial distribution with mean np and standard deviation $\sqrt{np(1-p)}$ which is much smaller than np. So there is almost no chance that we will flip many more than np bits. Let d be a bit bigger than np. It is tempting to look for a $(2d+1)$-separated code with the correct rate. We can't easily guarantee to achieve this but we don't have to.

Choose M codewords from $\mathbf{Z}_2^n$ independently at random. So for each choice we have probability $1/2^n$ of picking each of the possible elements of $\mathbf{Z}_2^n$ and the choices are independent. This will be our code.

A codeword is sent and we receive something. We look for the codeword closest to what we receive. What could go wrong? It is possible that our channel switches more than d bits in which case we can't hope to choose the correct codeword. If the channel switches at most d bits then the message we receive is within d bits of the original so the only problem would be that the message we receive is *also* within d bits of the wrong codeword. So the probability that we make a mistake is at most the sum of two probabilities:

1. The probability that we flipped more than d bits.
2. The probability that the message received whatever it is, is within d of one of the other $M-1$ codewords.

The first probability is very small because of the way in which we chose d. The second probability is at most $M-1$ times the probability that the string we receive is within

d of one particular codeword. The latter probability is

$$\frac{\text{the number of strings within } d \text{ of a particular string}}{2^n}$$

$$= \frac{\sum_{i=0}^{d} \binom{n}{i}}{2^n}.$$

If d is not too big then the sum of binomial coefficients is about the same size as the last one $\binom{n}{d}$. So we need $(M-1)\binom{n}{d}$ to be small compared to 2^n.

$$(M-1)\binom{n}{d} = \alpha 2^n.$$

The rate of the code is

$$\frac{\log_2 M}{n} \approx \frac{1}{n}\log_2\left(\frac{2^n}{\binom{n}{d}}\right) + \frac{\log_2 \alpha}{n} = 1 - \frac{1}{n}\log_2\binom{n}{d} + \frac{\log_2 \alpha}{n}.$$

The last term is small even if α is very small ($1/n^{10}$ say) so we want to show that

$$\frac{1}{n}\log_2\binom{n}{d} \approx -p\log_2 p - (1-p)\log_2(1-p).$$

By Stirling's formula

$$\begin{aligned}
\frac{1}{n}\log_2\binom{n}{d} &= \frac{1}{n}\log_2\left(\frac{n!}{d!(n-d)!}\right) \approx \frac{1}{n}\log_2\left(\frac{n^n}{d^d(n-d)^{n-d}}\right) \\
&= \frac{1}{n}\log_2\left(\frac{n^n}{(np)^{np}(n(1-p))^{n(1-p)}}\right) \\
&= \frac{1}{n}\log_2\left(\frac{1}{p^p(1-p)^{(1-p)}}\right)^n \\
&= -p\log_2 p - (1-p)\log_2(1-p).
\end{aligned}$$

□

Where does the free lunch come from? If we fix the rate then the number of bit flips we protect against, d, is proportional to n. Likewise if we fix the probability p the expected number of bits that we flip is np: also proportional to n. So it looks as though changing from n to $2n$ doesn't help: we might as well use an n-bit code and just follow one codeword by another. The key lies in the fact that the standard deviation

$\sqrt{np(1-p)}$ does *not* scale like n. As you increase n the chance that you flip $np + \varepsilon n$ bits decreases to 0. So by protecting against this number of flips (which you can do at a fixed rate) you cause the probability of error to approach 0.

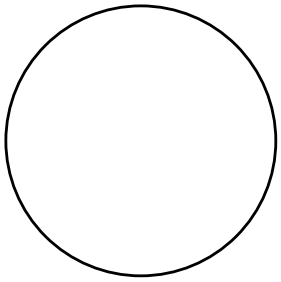
Shannon's Theorem tells us that there are good codes but it doesn't tell us how to find them. The theorem uses a random code which is problematic for two reasons. Generating random strings with many bits is costly (and strictly speaking impossible). If you choose the codewords at random the code will be very difficult to decode: there won't be a nice decoding algorithm as there was for the Hamming codes. From this point on, the theory of error-correcting codes turns almost exclusively to the construction of useful codes.

Chapter 3. Discrete geometry

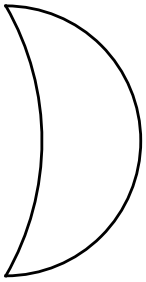
Introduction

In this chapter we study finite collections of points in $\mathbf{R}^d$ and convex combinations of them. What is there to say? Quite a lot.

A set C in $\mathbf{R}^d$ is called *convex* if whenever $x, y \in C$ and $\lambda \in [0,1]$ we have $(1-\lambda)x+\lambda y \in C$: the line segment joining x to y is also in C.

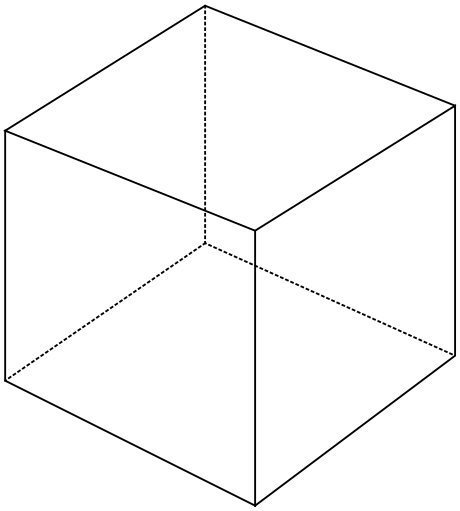


Convex



Not convex

The unit ball of any normed space $(\mathbf{R}^d, \|.\|)$ is convex. For example the Euclidean sphere or the cube



The expression $(1-\lambda)x + \lambda y \in C$ is called a convex combination of x and y. More generally if $x_1, x_2, \ldots, x_m$ are points in $\mathbf{R}^d$ then a convex combination of these points

is a point of the form

$$\sum_{1}^{m} \lambda_i x_i$$

where the λ_i are non-negative and add up to 1.

It is easy to check that if C is convex and $x_1, x_2, \ldots, x_m \in C$ then any convex combination of these points also belongs to C. For example if $x, y, z \in C$ then

$$\frac{x+y}{2} \in C$$

and hence

$$\frac{x+y+z}{3} = \frac{2}{3}\left(\frac{x+y}{2}\right) + \frac{1}{3}(z) \in C.$$

It is also easy to see that a "convex combination of convex combinations is a convex combination". We shall be particularly interested in convex sets with only finitely many "corners". We shall start gently.

It is clear that the intersection of any number of convex sets is again convex. This means that if I give you any set E in $\mathbf{R}^d$ you can consider all convex sets that include E, and their intersection will be the smallest convex set including E. This is called the *convex hull* of E or $\mathrm{conv}(E)$. The convex hull $\mathrm{conv}(E)$ has the following properties:

- $E \subset \mathrm{conv}(E)$
- If C is convex and $E \subset C$ then $\mathrm{conv}(E) \subset C$.

We have an easy theorem.

Theorem (The convex hull)**.** *If $E \subset \mathbf{R}^d$ then* $\mathrm{conv}(E)$ *consists of all convex combinations of points in E.*

Proof Let CC be the set of convex combinations of points in E.

Any convex combination of points in E is automatically a convex combination of points in $\mathrm{conv}(E)$. This set is convex so every point of CC belongs to $\mathrm{conv}(E)$. Hence $CC \subset \mathrm{conv}(E)$.

In the other direction it is clear that $CC \supset E$ since each point of E is a combination of itself. So it suffices to check that CC is convex and therefore includes $\mathrm{conv}(E)$. But we already remarked that a convex combination of convex combinations is again a convex combination: so CC is convex. □

Separation

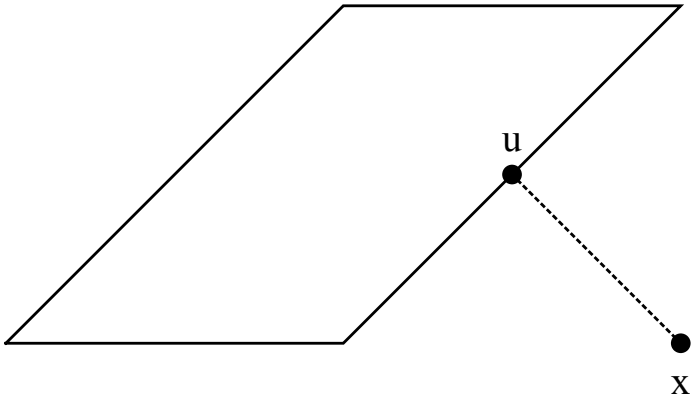
We shall be concerned almost entirely with convex sets that are closed and in almost all cases they will also be bounded and hence compact. The most fundamental fact related to convexity is the separation principle. This has many forms: the simplest is the following.

Theorem (**The separation principle**)**.** *If $C \subset \mathbf{R}^d$ is a compact, convex set and $x \notin C$ then there is a hyperplane separating x from C. In other words there is a linear functional ϕ on $\mathbf{R}^d$ and a number α so that*

- $\phi(x) > \alpha$ *but*
- $\phi(c) < \alpha$ *for each* $c \in C$.

This theorem is closely related to the Hahn-Banach Theorem that you may see in Functional Analysis II but the finite-dimensional version is easier to prove.

Proof Consider the function $c \mapsto \|x - c\|$ on C. This function is continuous and so has a minimum on C: there is a closest point of C to x. Call it u. Since $x \notin C$ we know that $u \neq x$.



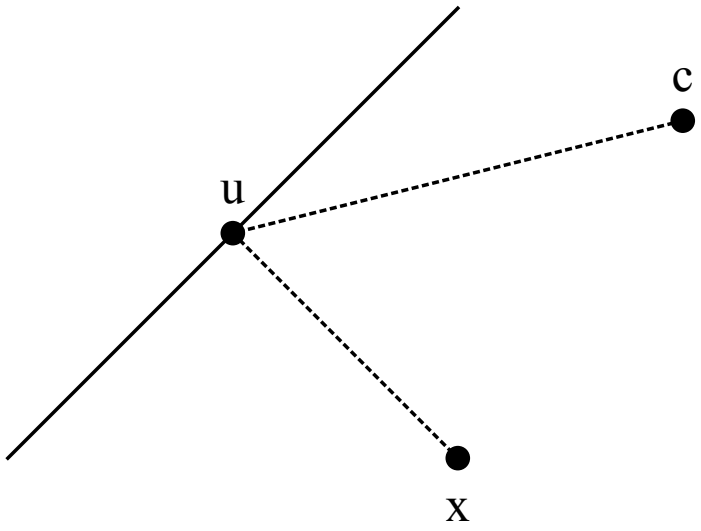
Now define $\phi : \mathbf{R}^d \to \mathbf{R}^d$ by

$$\phi(y) = \langle x - u, y \rangle.$$

Since $\phi(x) - \phi(u) = \langle x - u, x - u \rangle = \|x - u\|^2 > 0$ we can choose a number α with $\phi(u) < \alpha < \phi(x)$.

To complete the proof we just need to check that if $c \in C$ then $\phi(c) \leq \phi(u)$. Notice that up till now we have not used the convexity of C. Suppose that $c \in C$ and $\phi(c) > \phi(u)$. Consider the combinations $(1 - \delta)u + \delta c$ which belong to C for each

$\delta \in [0,1]$. I claim that if δ is small enough the combination will be closer to x than u is, giving us a contradiction.



The square of the distance to x is

$$\begin{aligned} \|x - ((1-\delta)u + \delta c)\|^2 &= \|x - u - \delta(c-u)\|^2 \\ &= \|x-u\|^2 - 2\delta\langle x-u, c-u\rangle + \delta^2\|c-u\|^2 \\ &= \|x-u\|^2 - 2\delta\phi(c-u) + \delta^2\|c-u\|^2. \end{aligned}$$

Since $\phi(c-u) > 0$ this expression will be less than $\|x-u\|^2$ if $\delta > 0$ is small enough. □

We can immediately get a corollary which we use repeatedly:

Theorem (**Supporting hyperplanes**)**.** *If $C \subset \mathbf{R}^d$ is a compact, convex set and x is on the boundary of C then there is a hyperplane supporting C at x.*

In other words there is a non-zero linear functional ϕ on $\mathbf{R}^d$ so that $\phi(c) \leq \phi(x)$ for each $c \in C$.

Proof Homework. □

Corollary (**Half-spaces**)**.** *If $C \subset \mathbf{R}^d$ is a compact, convex set then C can be expressed as an intersection of half-spaces.*

Proof For each point x outside C there is a half-space containing C and not x. So the intersection of all half-spaces containing C, is C. □

We can relax the hypothesis of the Separation Theorem: we don't need a compact convex set: a closed one will do.

Theorem (**The separation principle**)**.** *If $C \subset \mathbf{R}^d$ is a closed, convex set and $x \notin C$ then there is a hyperplane separating x from C. In other words there is a linear functional ϕ on $\mathbf{R}^d$ and a number α so that*

- $\phi(x) > \alpha$ *but*
- $\phi(c) < \alpha$ *for each* $c \in C$.

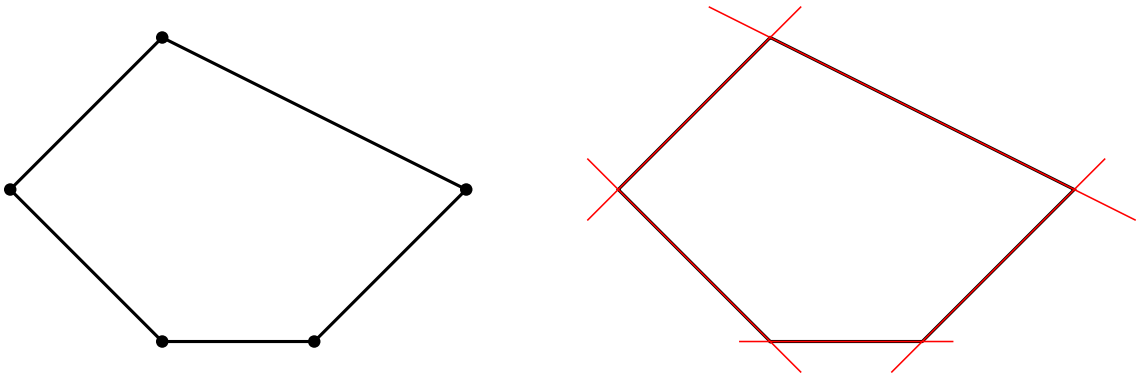
Proof Homework. □

It is not difficult to see that every linear functional on $\mathbf{R}^d$ is of the form

$$x \mapsto \sum_1^d x_i y_i = \langle x, y \rangle$$

(just like the one in the preceding proof). This will be HW. From now on when I talk about linear functionals on $\mathbf{R}^d$ I will automatically think of them this way.

Some of the most important convex sets in $\mathbf{R}^d$ are the polyhedra or *polytopes* as they are often called. A polytope is the convex hull of finitely many points: its corners.



Equivalently it is the intersection of finitely many half-spaces: $\{x : \langle x, u \rangle \leq \alpha\}$ (assuming this intersection is bounded). The fact that these two definitions are equivalent is not quite trivial: it will emerge from a more detailed study of convex sets.

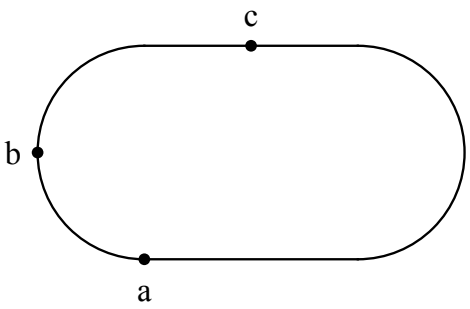
Extreme points

Definition (**Extreme points**). *If C is a convex set in $\mathbf{R}^d$, an* extreme point *of C is a point $c \in C$ which is not in the interior of any line segment contained in C. In other words, if x and y belong to C, $0 < \lambda < 1$ and*

$$c = (1 - \lambda)x + \lambda y$$

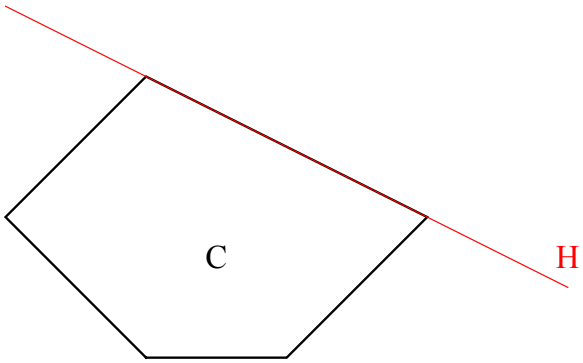
then in fact $x = y = c$.

In the figure, a and b are extreme but c is not.



We shall see that every compact convex set has plenty of extreme points. To begin with we need to relate the two concepts we have.

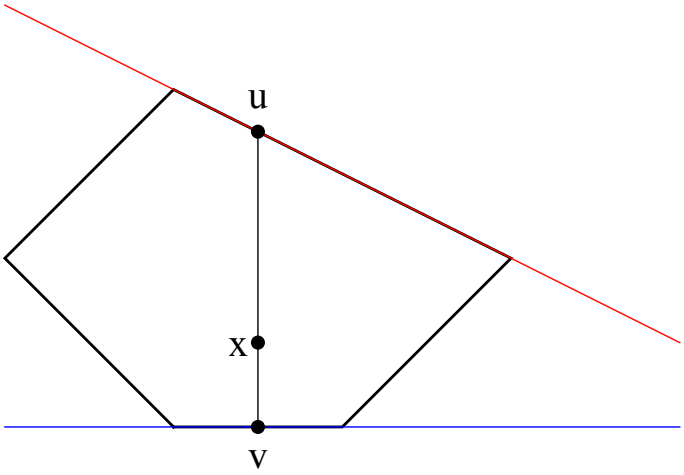
Lemma (**Extreme points of faces**). *Let H be a supporting hyperplane to the compact convex set C. Then $H \cap C$ is compact and convex and every extreme point of $H \cap C$ is an extreme point of C.*



Proof The fact that $H \cap C$ is compact and convex is clear. Suppose x is an extreme point of $H \cap C$. Then it does not lie in the interior of a line segment in $H \cap C$. Suppose it did lie in one inside C. Then the segment would have points on either side of H. But this is impossible since C is on one side of H. □

Theorem (**Extreme point theorem**). *A compact convex subset of* $\mathbf{R}^d$ *is the convex hull of its extreme points.*

Proof We shall use induction on the dimension. Let x be a point in a compact convex C. If x is an extreme point of C there is nothing to prove. If not, it lies on a line segment inside C which we may extend in each direction until it hits the boundary of C: say at u and v.

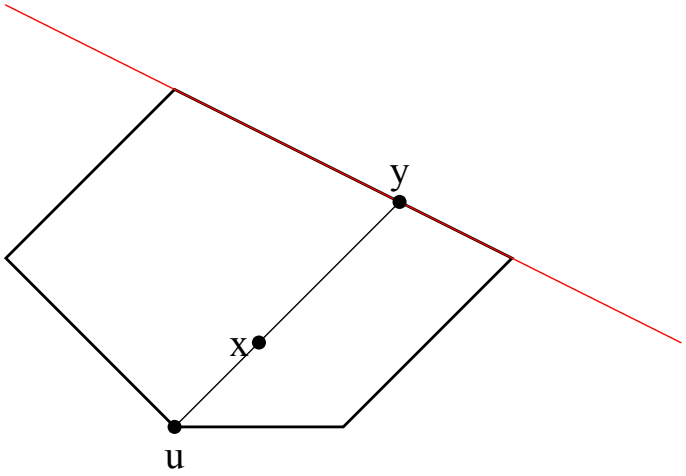


Let H be a supporting hyperplane to C at u. Then $H \cap C$ is a compact convex set in $d-1$ dimensions so u belongs to the convex hull of its extreme points and hence the extreme points of C. The same is true of v and hence x. □

This theorem has an immediate strengthening.

Theorem (**Caratheodory**). *Each point of a compact convex subset of* $\mathbf{R}^d$ *is a convex combination of* $d+1$ *of its extreme points.*

Proof We use induction on the dimension. Assume the corresponding result for sets in $d-1$ dimensions. Let x be a point in a compact convex $C \subset \mathbf{R}^d$. Choose an extreme point u of C and look at the line passing through u and x. It exits C at u on one side of x and some other point y on the other.



Let H be a supporting hyperplane to C at y. By the inductive hypothesis y is a convex combination of d extreme points in $H \cap C$. Then x is a convex combination of these and u. □

Polytopes and duality

For the moment, let's call C a *polyhedron* if it is a bounded intersection of finitely many half-spaces and a *polytope* if it is the convex hull of finitely many points. We want to show that every polyhedron is a polytope and vice versa. Let's start with the first.

Theorem (**Polyhedra are polytopes**)**.** *Every polyhedron in* $\mathbf{R}^d$ *is a polytope.*

Proof Let C be the intersection of half-spaces $S_1, S_2, \ldots, S_n$ bounded by hyperplanes $H_1, H_2 \ldots, H_n$. Since C is compact, it suffices to show that it has only finitely many extreme points. Hence it suffices to show that each extreme point of C is the intersection of some d of the hyperplanes which meet only at that point: because there are at most $\binom{n}{d}$ of these intersections.

We use induction on dimension yet again to prove this statement. In dimension 1 it is obvious. Assume it is true (mutatis mutandis) for polyhedra in lower dimensions than d. Let c be an extreme point of C. If it lay in the interior of all the half-spaces there would be a ball around it inside C and so c would not be extreme. So c lies on at least one of the H_i: say H_1. Now $H_1 \cap C$ is a polyhedron inside the $(d-1)$-dimensional hyperplane H_1 and c is one of its extreme points. So by the inductive hypothesis, c is the intersection of $d-1$ of the H_i which (inside H_1) meet only at c. Therefore c is the intersection of d of the H_i which meet only at c. □

We shall deduce the reverse implication from what we just proved using what is called duality. If C is a compact convex set then we define its *polar* to be

$$C^\circ = \left\{ y \in \mathbf{R}^d : \langle x, y \rangle \leq 1, \text{ for all } x \in C \right\}.$$

Under appropriate conditions, polarity creates a pairing between C and C° which is why we call it duality.

Lemma (Polarity)**.** *If C is a compact convex set containing the point 0 then*

$$C^{\circ\circ} = (C^\circ)^\circ = C.$$

Proof For each $y \in C^\circ$ we know that each point x of C satisfies the inequality $\langle x, y \rangle \leq 1$. So $C \subset C^{\circ\circ}$. The only danger is that there might be a point x outside C satisfying

$$\langle x, y \rangle \leq 1, \text{ for all } y \in C^\circ.$$

If x is such a point use the separation theorem to find a linear functional $u \mapsto \langle u, y \rangle$ on $\mathbf{R}^d$ and a number α with $\langle x, y \rangle > \alpha$ but $\langle c, y \rangle < \alpha$ for each $c \in C$. Since $0 \in C$ we have $\alpha > 0$ and by rescaling y we may assume $\alpha = 1$. The vector y therefore belongs to C°. But we know that $\langle x, y \rangle > 1$ so indeed $x \notin C^{\circ\circ}$. □

Let $C = \{x \in \mathbf{R}^d : \|x\|_2 \leq 1\}$ be the Euclidean ball in $\mathbf{R}^d$. What is its polar?

$$C^\circ = \{y : \langle x, y \rangle \leq 1 \text{ for all } x \text{ with } \|x\|_2 \leq 1\}.$$

If $\|y\| \leq 1$ then $y \in C^\circ$ because

$$\langle x, y \rangle \leq \|x\|_2 \|y\|_2$$

by the Cauchy-Schwarz inequality. On the other hand, if $\|y\|_2 = r > 1$ then $x = y/r \in C$ and

$$\langle x, y \rangle = \langle y/r, y \rangle = \frac{\|y\|_2^2}{r} = r > 1$$

so $y \notin C^\circ$. The Euclidean unit ball is its own polar.

More generally, if C is the unit ball of a normed space $X = (\mathbf{R}^d, \|.\|)$ then C° is the unit ball of X^*.

We need one further very small lemma.

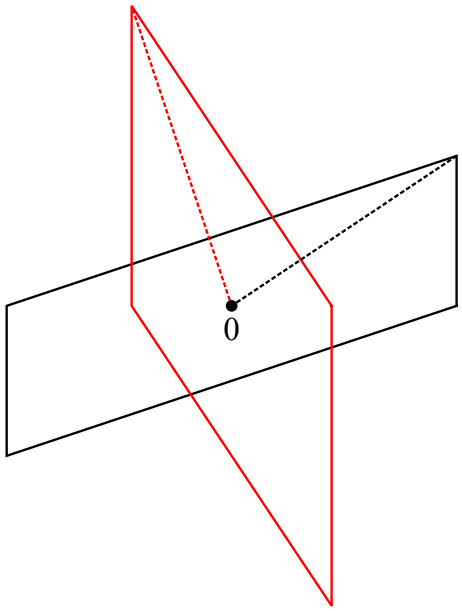
Lemma (**The polar of a polytope**). *If C is the convex hull of a finite set $\{x_1, x_2, \ldots, x_m\}$ then*

$$C^\circ = \left\{ y \in \mathbf{R}^d : \langle x_i, y \rangle \leq 1, \text{ for all } i \right\}.$$

In other words we don't have to check that $\langle x, y \rangle \leq 1$ for *every* $x \in C$, just the vertices.

Proof Homework. □

Thus the vertices of C correspond to the facets of C°.



x is a vertex of C if the corresponding $(d-1)$-dimensional face of C° lies in the hyperplane $\{y : \langle x, y \rangle = 1\}$.

Examples of polars

C is the convex hull of $(1,0)$, $(0,1)$ and $(1,1)$. By the Polar of a Polytope Lemma C° is the intersection of half-spaces

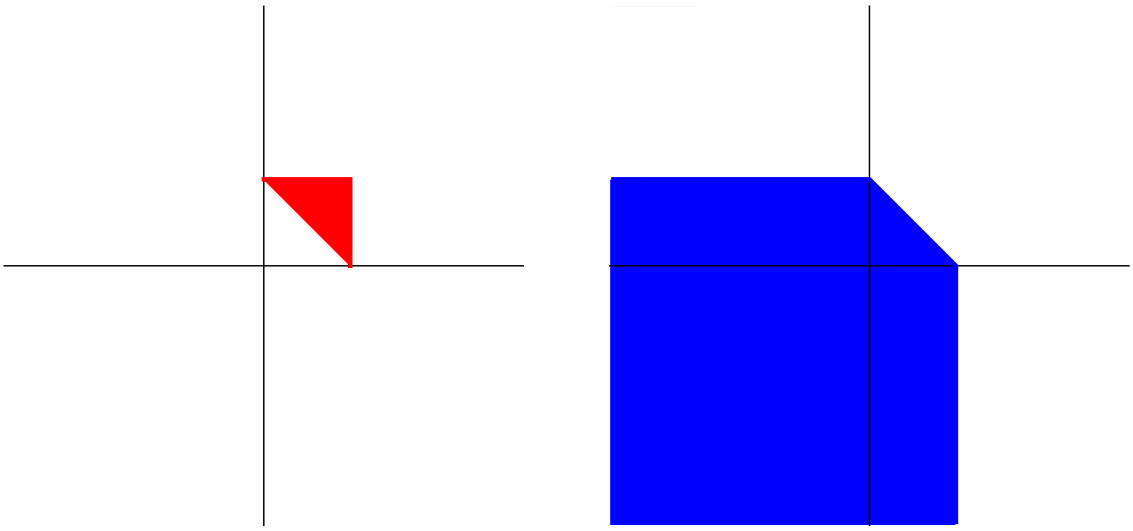
$$\{(x,y) : (1,0).(x,y) \leq 1\} \text{ or equivalently } \{(x,y) : x \leq 1\}$$

$$\{(x,y) : y \leq 1\}$$

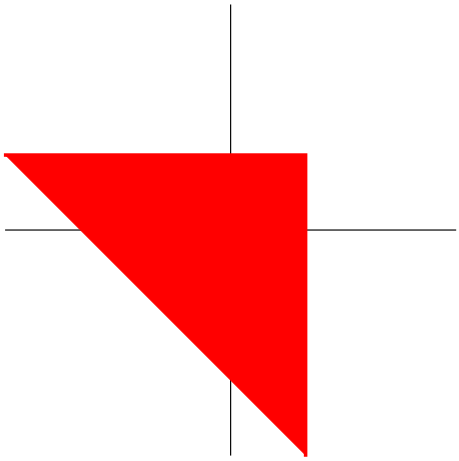
$$\{(x,y) : x + y \leq 1\}.$$

So

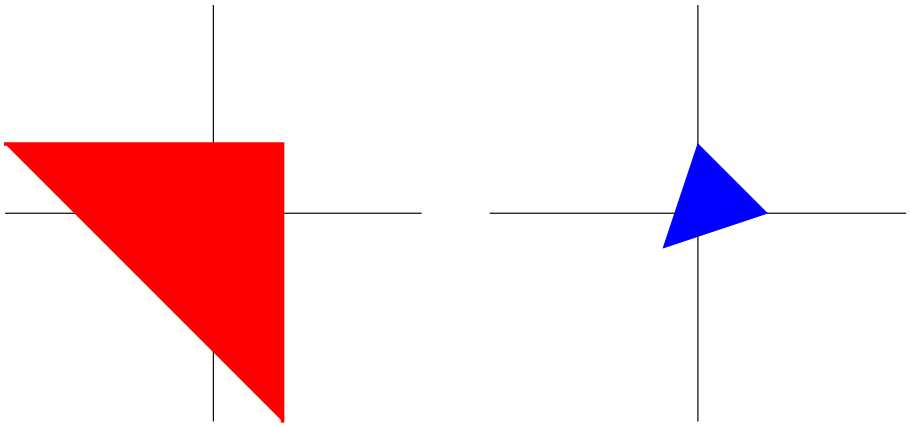
$$C^\circ = \{(x,y) : x \leq 1\} \cap \{(x,y) : y \leq 1\} \cap \{(x,y) : x + y \leq 1\}$$



What are the vertices of the polar of C: the convex hull of $(1,-3)$, $(-3,1)$ and $(1,1)$?



This time 0 is in the interior of C. So $C = C^{\circ\circ}$ and C° is a polytope whose vertices correspond to the facets of C.



The facets lie in the lines $(x = 1)$, $(y = 1)$ and $(x + y = -2)$. The latter is

$$(-1/2x - 1/2y = 1).$$

So the vertices of the polar are $(1, 0)$, $(0, 1)$ and $(-1/2, -1/2)$.

Polytopes are polyhedra

Lemma (**Inversion of order**). *If C and D are convex sets with $D \subset C$ the $C^\circ \subset D^\circ$.*

Proof If $\langle x, y\rangle \leq 1$ for every $x \in C$ then automatically this is true for every $x \in D$. So if $y \in C^\circ$ then $y \in D^\circ$. □

Theorem (**Polytopes are polyhedra**). *Every polytope in $\mathbf{R}^d$ is a polyhedron.*

Proof We use induction on dimension. Let C be the convex hull of finitely many points. If it lies in a hyperplane then we can use the inductive hypothesis.

So assume that it does not lie in a hyperplane. Pick a point u in C and then d points $v_1, v_2, \ldots, v_d$ in C for which the set of points $\{v_i - u : 1 \leq i \leq d\}$ span $\mathbf{R}^d$. The convex hull of the points $u, v_1, \ldots, v_d$ has non-empty interior: it contains a ball of radius r say around the average

$$\frac{u + v_1 + v_2 + \cdots v_d}{d + 1}.$$

Assume by translating that this point is 0 and that C is the convex hull of $x_1, x_2, \ldots, x_m$. We know that if $\|x\| \leq r$ then $x \in C$. I claim that the polar C° is bounded.

If $y \in C^\circ$ and $\|y\| = K$ then ry/K has norm at most r so it belongs to C. Therefore

$$\langle ry/K, y \rangle = r/K \|y\|^2 = rK \leq 1.$$

So $K \leq 1/r$. Thus C° lies in the ball of radius $1/r$ and is bounded.

We know that

$$C^\circ = \left\{ y \in \mathbf{R}^d : \langle x_i, y \rangle \leq 1, \text{ for all } i \right\}$$

and have thus proved that C° is a polyhedron. By the previous theorem it is a polytope: the convex hull of $y_1, y_2, \ldots, y_m$ say. But this implies that

$$C = C^{\circ\circ} = \left\{ x \in \mathbf{R}^d : \langle x, y_i \rangle \leq 1, \text{ for all } i \right\}$$

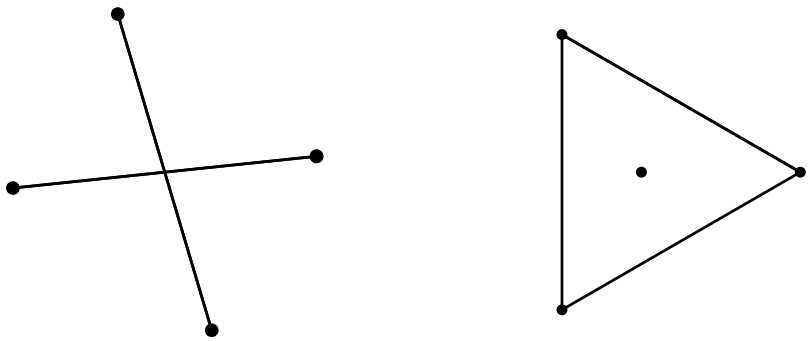
and is therefore a polyhedron. □

Radon's Lemma and Helly's Theorem

The last couple of results are more combinatorial in nature. The first is very simple but surprisingly useful.

Lemma (Radon)**.** *Suppose we have a set of $d+2$ points in $\mathbf{R}^d$. Then we can partition it into two subsets whose convex hulls overlap.*

Let's look at the case $d = 2$: 4 points in the plane. They might form the corners of a convex quadrilateral in which case the diagonals meet.



Alternatively they might be the corners of a triangle and a point inside it. In either case we can partition as dictated by Radon's Lemma and it is clear that 4 points is

exactly the right number in the sense that usually there will be exactly one point in the intersection.

The problem with proving the general result in $\mathbf{R}^d$ is that we have to decide how to partition the points. There are many different choices and different structures. With 5 points in $\mathbf{R}^3$ we could have a tetrahedron and a point inside it or a triangle and a line segment piercing it: how do we decide which?

Proof (**of Radon**) Let the points be $x_1, x_2, \ldots, x_{d+2}$. Consider the $d+2$ points in $\mathbf{R}^{d+1}$ given by

$$y_i = (x_i, 1)$$

obtained by adjoining an extra coordinate to each x_i and setting it equal to 1. There is a non-zero linear combination of the y_i which is equal to 0:

$$\sum a_i y_i = 0.$$

Each coefficient a_i has a sign: let's assume that the first k of them are positive (or 0) and the last $m = d+2-k$ are negative (or 0). Relabel the first k as $\lambda_1, \lambda_2, \ldots, \lambda_k$ and the last m as $-\mu_{k+1}, -\mu_{k+2}, \ldots, -\mu_{d+2}$ (so the λ_i and μ_j are non-negative).

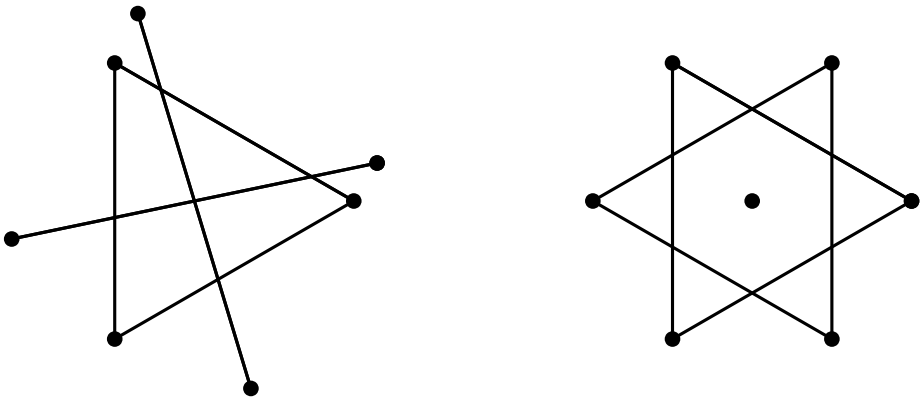
We have that $\sum \lambda_i y_i = \sum \mu_j y_j$. This means that $\sum \lambda_i x_i = \sum \mu_j x_j$ and by considering the final coordinate, that $\sum \lambda_i = \sum \mu_j = \beta > 0$. Then the vectors

$$\sum \frac{\lambda_i}{\beta} x_i \quad \text{and} \quad \sum \frac{\mu_j}{\beta} x_j$$

are convex combinations of two disjoint subsets of the x_i and these two vectors are equal. □

Suppose that instead of asking whether we can partition into two sets whose convex hulls intersect we ask about partitioning into 3 subsets whose convex hulls share a common point. If we are to have a hope of doing this we will need to start with more than $d+2$ points. The right number turns out to be $2d+3$ and in general if we are partitioning into k subsets we need $(k-1)(d+1)+1$.

For example if $d = 2$ and $k = 3$ we get 7 points and these can be partitioned into a triangle and two line segments which will usually meet in at most a single point or into two triangles and a point inside both.



The more general theorem does hold but is much harder to prove than Radon's Lemma.

Theorem (**Tverberg**)**.** *Each set of* $(k-1)(d+1)+1$ *points in* $\mathbf{R}^d$ *can be partitioned into* k *subsets whose convex hulls have a non-empty intersection.*

The simplest proof is by means of something called the Colourful Caratheodory Theorem. HW

The next theorem is a rather striking dual version of Caratheodory. It looks like a kind of quantitative compactness for convex sets. Recall the following characterisation of compactness. A space is compact if and only if whenever $\{K_\alpha\}$ is a family of closed sets in the space with empty intersection, there is some finite number that have empty intersection.

Theorem (**Helly**)**.** *Let* $C_1, C_2, \ldots, C_m$ *be a family of compact convex sets in* $\mathbf{R}^d$ *and suppose that any* $d+1$ *have a non-empty intersection. Then the whole family has a non-empty intersection.*

Proof We use induction on the number of sets m. If there are $d+1$ sets then there is nothing to prove. Now suppose that we have m sets and we know the result for fewer sets. If $m > d+2$ then the inductive step is easy. We know that every collection of $d+2$ of our sets have a common point. Let us look at the family of $m-1$ sets

$$C_1 \cap C_2, \quad C_3,\ C_4, \ldots, C_m.$$

The intersection of any $d+1$ of these sets is the intersection of at most $d+2$ of the original ones and so is non-empty. By the inductive hypothesis the intersection of these $m-1$ sets is non-empty. But this is just $\bigcap_{i=1}^m C_i$.

This doesn't work if $m = d + 2$ because we use the $d + 2$ case in the inductive step: this is really the crucial case. Suppose we have $d + 2$ sets $C_1, C_2, \ldots, C_{d+2}$. For each one of these, C_i we know that the remainder have a common point: call it x_i. So $x_1 \in C_2 \cap C_3 \cap \ldots \cap C_{d+2}$ and so on. By Radon's Lemma we can partition this collection of points into two subsets whose convex hulls intersect: let's say in the point u.

I claim that u belongs to all the C_i. It suffices to check that it belongs to C_{d+2}. The point x_{d+2} falls into one of the subsets in the Radon partition and u is a convex combination of the points in the other subset. So u can be written as a convex combination

$$u = \sum_{i=1}^{d+1} \lambda_i x_i.$$

Now for each $i < d + 2$ we know that $x_i \in C_{d+2}$. So $u \in C_{d+2}$. □

Chapter 4. Partially ordered sets and set systems

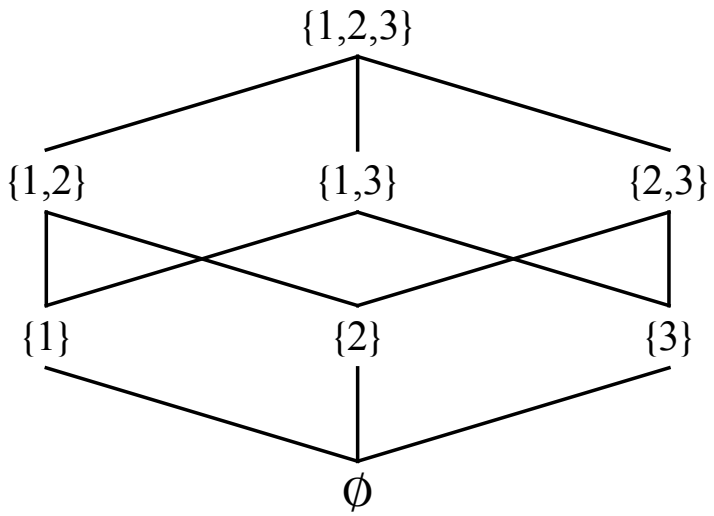
Introduction

A partially-ordered set or poset is a set Ω together with a relation $\leq$ between elements of the set satisfying

- $x \leq x$ for all $x \in \Omega$
- If $x, y \in \Omega$ and $x \leq y$ and $y \leq x$ then $x = y$
- If $x, y, z \in \Omega$ and $x \leq y$ and $y \leq z$ then $x \leq z$.

These are the relations you expect from an ordering but whereas in the real numbers we insist that any two elements are related, here we do not. The reals are *totally ordered* but we are looking at *partial* orders.

The basic example is the set of subsets of $[n] = \{1, 2, \ldots, n\}$, ordered by inclusion. We call this $\mathcal{P}([n])$ the power set of $[n]$.



A *chain* in a poset is a totally ordered subset: eg.

$$\emptyset \subset \{1\} \subset \{1, 2\} \subset \{1, 2, 3\}.$$

An *antichain* in a poset is a collection of elements no two of which are related: eg.

$$\{1\}, \{2\}, \{3\}$$

or

$$\{1\}, \{2, 3\}.$$

Clearly within $\mathcal{P}([n])$ the sets of any fixed size form an antichain. If two different sets have the same number of elements then neither can be included in the other. So for each k there is an antichain in the poset with $\binom{n}{k}$ members. The largest of these will be the one or ones in the middle: if n is even then it has size

$$\binom{n}{n/2}$$

while if n is odd there are two largest ones at level $(n-1)/2$ and $(n+1)/2$. In either case the largest has size

$$\binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

Sperner's Theorem

We shall start by proving the following theorem of Sperner.

Theorem (**Sperner's Theorem**). *Let $(\Omega, \subset)$ be the poset $\mathcal{P}([n])$ partially ordered by inclusion.*

Then the largest antichain in Ω has

$$\binom{n}{\lfloor \frac{n}{2} \rfloor}$$

elements.

It is clear that an antichain of this size exists: just pick all the subsets with $\lfloor \frac{n}{2} \rfloor$ elements. We will deduce Sperner's Theorem from a stronger statement due to Lubell, Yamamoto and Meshalkin.

Theorem (**The LYM inequality**). *Let $(\Omega, \subset)$ be the poset $\mathcal{P}([n])$ partially ordered by inclusion and $\mathcal{F}$ an antichain in Ω.*

Then

$$\sum_{A \in \mathcal{F}} \frac{1}{\binom{n}{|A|}} \leq 1.$$

Thus each set of size k in the antichain is given a weight $\frac{1}{\binom{n}{k}}$ and the total weight of all sets in the antichain is at most 1. Since each weight is at least

$$\frac{1}{\binom{n}{\lfloor \frac{n}{2} \rfloor}}$$

there can be no more than

$$\binom{n}{\lfloor \frac{n}{2} \rfloor}$$

elements of the antichain proving Sperner's Theorem.

Proof (**of LYM**) Each of the $n!$ orders of the symbols $1, 2, \ldots, n$ corresponds to a maximal chain in Ω: for example the order $2, 4, 3, 1$ gives the chain

$$\emptyset \subset \{2\} \subset \{2, 4\} \subset \{2, 4, 3\} \subset \{2, 4, 3, 1\}.$$

Now regard the collection of maximal chains as a probability space by giving them equal probability $1/n!$. For each $A \subset \{1, 2, \ldots, n\}$ let E_A be the event consisting of those maximal chains in which A appears. If A and B belong to $\mathcal{F}$ then E_A and E_B are disjoint because A and B cannot belong to the same chain. So

$$\sum_{A \in \mathcal{F}} \mathrm{P}(E_A) \leq 1.$$

To finish we just need to check that if A has k elements then its probability is $\frac{1}{\binom{n}{k}}$. How many maximal chains does A belong to? $k!(n-k)!$ because we put in the elements of A in any order and then the remaining $n-k$ elements in any order. So the probability of E_A is just

$$\frac{k!(n-k)!}{n!} = \frac{1}{\binom{n}{k}}.$$

□

It is easy to see that if you can cover the poset with m chains then you cannot find an antichain with more than m elements. Hence another way to prove Sperner's Theorem would be to find a cover of the poset using

$$\binom{n}{\lfloor \frac{n}{2} \rfloor}$$

chains. This is a bit trickier. Our proof of the LYM inequality used a similar idea but was simpler because we looked at a multiple covering by all the maximal chains and then counted how often each set was covered.

Interestingly enough it is possible to go in the other direction: to deduce that there is a covering by chains from the fact that antichains are not too big: as we shall see in the next section.

Dilworth's Theorem

Theorem (**Dilworth**). *Let $(\Omega, \leq)$ be a poset in which every antichain has at most m elements. Then Ω can be covered by m chains.*

Recall from last year Hall's Marriage Theorem.

Theorem (**Hall's Marriage Theorem**). *Let G be a bipartite graph with vertex classes A and B. For each subset $U \subset A$ let $\Gamma(U)$ be the set of neighbours of vertices in U:*

$$\Gamma(U) = \{b : ab \text{ is an edge for some } a \in U\}.$$

If for every $U \subset A$ the set $\Gamma(U)$ is at least as large as U then G contains a complete matching from A into B.

Dilworth's theorem has the same flavour as Hall's Theorem in that the condition is trivially necessary and turns out rather surprisingly to be sufficient. More is true: Hall's Theorem can be deduced very simply from Dilworth's Theorem. HW

The proof below is taken from Bollobás' book.

Proof (**of Dilworth**) We shall use induction on the number of elements of the poset. If it has 1 element there is nothing to prove.

Suppose $(\Omega, \leq)$ has more than one element and the result is true for smaller posets. Let m be the size of the largest antichain in Ω. Choose a maximal chain C in Ω. It might be that $\Omega - C$ contains no antichain larger than $m-1$ in which case we can cover it with $m-1$ chains by the inductive hypothesis and we have finished.

If not then let $\{a_1, a_2, \ldots, a_m\}$ be an antichain in $\Omega - C$. Let $S^- = \{x \in \Omega : x \leq a_i, \text{ for some } i\}$ and $S^+ = \{x \in \Omega : x \geq a_i, \text{ for some } i\}$. Neither of these can be the whole of Ω since for example the maximal element of C is not in S^-. So, by the inductive hypothesis again each can be covered by m chains:

$$S^- = \bigcup_1^m C_i^- \quad \text{and} \quad S^+ = \bigcup_1^m C_i^+.$$

Each of these decompositions covers all a_i and the a_i lie in different chains so assume that $a_i \in C_i^-$ and $a_i \in C_i^+$ for each i.

If we can show that a_i is the maximal element in C_i^- and the minimal element in C_i^+ then we can join these chains together and cover Ω with m chains. If not then for example we can find an i and an $x \in C_i^-$ with $a_i < x$. Since $x \in S^-$ there is a $j \neq i$ with $x \leq a_j$. But then $a_i < a_j$ contradicting the fact that we had an antichain. □

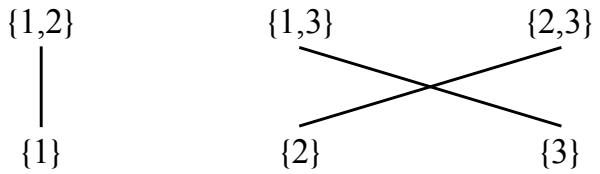
Covering by chains

Dilworth's Theorem and the trivial observation that the largest antichain cannot be more than the number of chains in a covering, show that the size of the largest antichain is equal to the minimum number of chains with which we can cover.

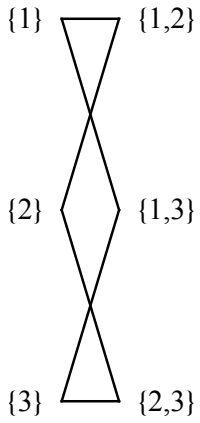
Dilworth's Theorem and Sperner's Theorem together show that the poset $\mathcal{P}([n])$ can be covered by $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ chains. It is natural to ask whether we can show this directly. The answer is yes and the most obvious argument uses Hall's Theorem. Choose $r < n/2$ and look at the sets of size r and those of size $r+1$. I claim that there is a 1-1 map f from the first collection into the second so that for each set A of size r

$$A \subset f(A).$$

For example



Proof Consider the bipartite graph in whose vertex classes are the sets of size r and the sets of size $r+1$ respectively and in which A is adjacent to B if $A \subset B$.



Each set of size r (on the left) has $n-r$ neighbours since we can extend it to a set of size $r+1$ using any of the remaining symbols. Each set of size $r+1$ (on the right) has $r+1$ neighbours since we can throw out any one of its elements. Note that $n-r \geq r+1$

since $r \leq (n-1)/2$. Now suppose we have a collection $\mathcal{F}$ of sets on the left. The total number of edges coming out is $(n-r)|\mathcal{F}|$ and they must be incident to at least

$$\frac{(n-r)|\mathcal{F}|}{r+1} \geq |\mathcal{F}|$$

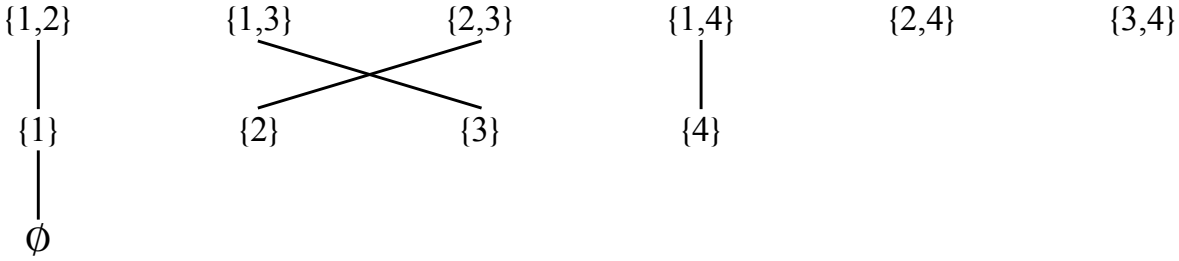
sets on the right.

So the hypothesis of Hall's Theorem holds and we have a matching from the sets of size r into the sets of size $r+1$. □

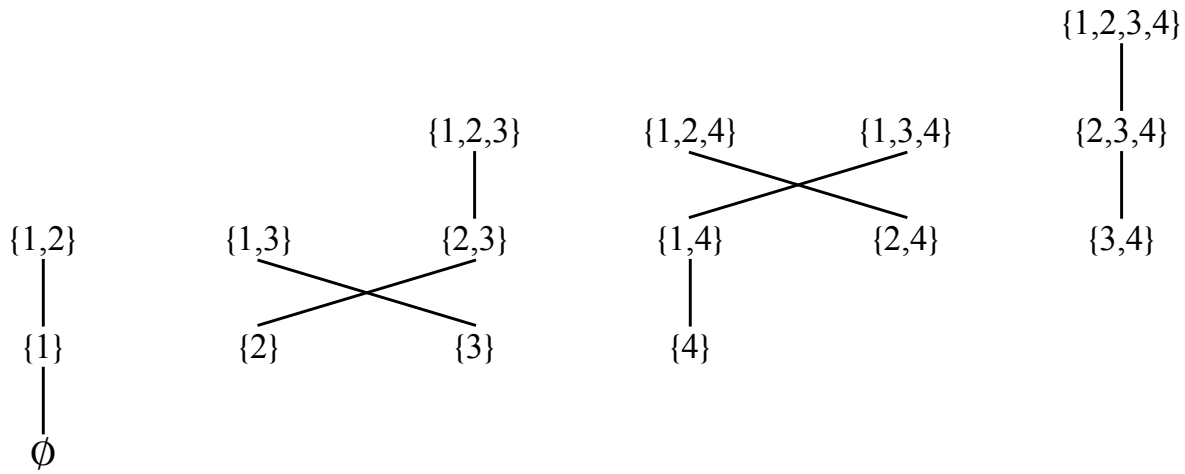
Now we can build chains covering the collection of subsets by putting together all these matchings. Suppose k is $\lfloor \frac{n}{2} \rfloor$. We can build chains covering all the sets of size at most k in the following way. $\emptyset$ is matched to a singleton: say $\{a\}$. This starts a chain. Now consider the matching of all singletons into sets of size 2. The set $\{a\}$ is matched to say $\{a, b\}$ so we extend the earlier chain to

$$\emptyset \subset \{a\} \subset \{a, b\}$$

and start new chains for each of the other singletons. Now we match all pairs into triples: some of these extend the chains containing the pairs that we already used: others start new chains. Continuing in this way we get up all the way to level k. If n is odd we can go a step further up to $k+1$: if n is even we stop at k.

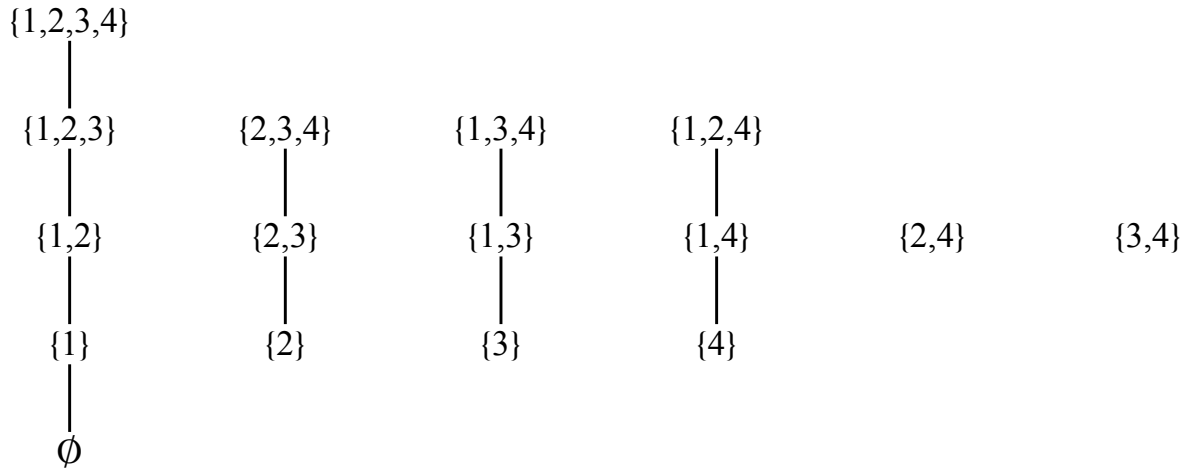


Now we can carry out the mirror image process from the top downwards. If we matched $\emptyset$ to $\{a\}$ we can match $\{1, 2, \ldots, n\}$ to $\{1, 2, \ldots, n\} - \{a\}$ and so on.



We can join the up chains and the down chains together at level k or $k+1$. Each chain contains a set at level k so we have $\binom{n}{k}$ chains.

This construction gives another proof of Sperner's Theorem since as we already remarked it is obvious that if you can cover with m chains then every antichain has at most m elements. The construction has a nice "improvement" in which we insist that each chain is "symmetric": that it consists of sets of size $0, 1, 2, \ldots, n$ or of size $1, 2, \ldots, n-1$ and so on. This improvement cannot be deduced just from Hall's Theorem as we did the earlier construction because (as we saw) the chains don't match up.



Theorem (**de Bruijn, Tengbergen and Kruyswijk**). *The poset $\mathcal{P}([n])$ can be decomposed into $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ symmetric chains.*

Although the Hall Theorem argument used above is probably the most obvious one,

the proof of the symmetric result is actually shorter although it is rather clever.

Proof We use induction on n. For $n = 1$ we have the single chain $(\emptyset, \{1\})$ which is symmetric.

Now suppose we have a decomposition of the subsets of $\{1, 2, \ldots, n-1\}$ into symmetric chains. For each such chain $(A_1, A_2, \ldots, A_k)$ we form two new chains:

$$(A_1, A_2, \ldots, A_k, A_k \cup \{n\})$$

and

$$(A_1 \cup \{n\}, A_2 \cup \{n\}, \ldots, A_{k-1} \cup \{n\}).$$

These chains cover the subsets of $\{1, 2, \ldots, n\}$ because for each old subset A we include A and $A \cup \{n\}$.

The first type of chain is one set longer so it becomes symmetric in the new poset. The second type is one set shorter but starts with a set which is one element larger than the original chain so it too is symmetric in the new poset. The construction ensures that in each chain each set has one more element than the set before it so each one contains an element of size $\lfloor \frac{n}{2} \rfloor$ so there are exactly $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ chains. □

On the face of it something is wrong with the proof. We seem to double the number of chains at each step. In going from $n-1$ odd to n even the middle binomial coefficient does double in size:

$$1, 3, 3, 1$$

becomes

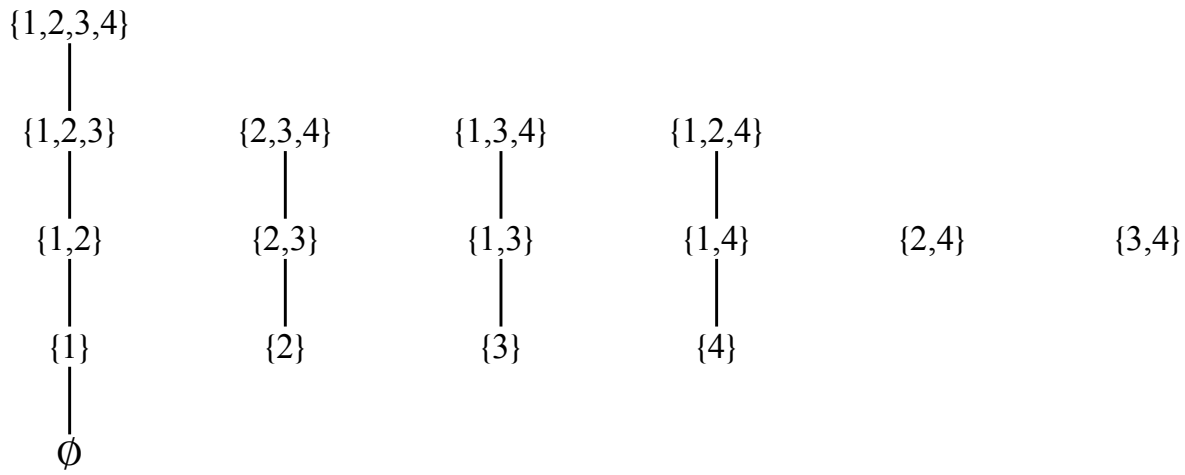
$$1, 4, 6, 4, 1$$

for example. But in going from $n-1$ even to n odd the coefficient does not quite double.

What happens is that if $n-1$ is even there are some chains containing only 1 element at level $(n-1)/2$. In that case the second type of chain

$$(A_1 \cup \{n\}, A_2 \cup \{n\}, \ldots, A_{k-1} \cup \{n\})$$

isn't really there because $k = 1$.



When we move to $n = 5$ we get 6 chains of the first type but only 4 of the second type giving 10 as required.

VC dimension and the Sauer-Shelah Lemma

Let Ω be a set such as the plane $\mathbf{R}^2$ or $[n]$, and $\mathcal{F}$ a family of its subsets. Can we measure how complicated $\mathcal{F}$ is?

The question arose originally (in a specific form) in connection with statistical sampling. If (X_i) are independent samples of a random variable X then the sequence of values should be distributed like X itself. What does this mean? If I is an interval of the line then with probability 1, the proportion of our sample that falls into I is close to what it should be

$$\frac{|\{i \leq n : X_i \in I\}|}{n} \to \mathrm{P}(X \in I).$$

If $\mathrm{P}(X \in I) = p$ then the quantity $|\{i \leq n : X_i \in I\}|$ is the sum of n independent Bernoulli random variables Y_i

$$\mathrm{P}(Y_i = 1) = p, \qquad \mathrm{P}(Y_i = 0) = 1 - p.$$

By the strong law of large numbers

$$\frac{1}{n}\sum_{1}^{n} Y_i \to p$$

with probability 1.

$$\frac{|\{i \leq n : X_i \in I\}|}{n} \to \mathrm{P}(X \in I).$$

In fact much more is true. The Glivenko-Cantelli theorem says that this convergence occurs uniformly over all intervals:

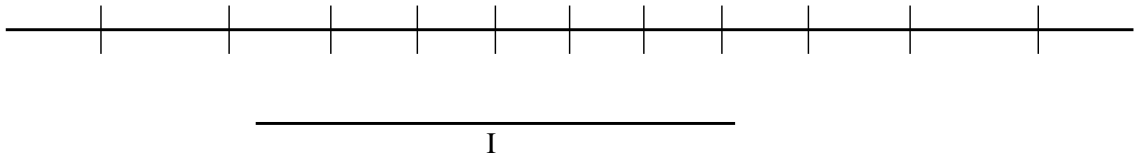
$$\sup_I \left| \frac{|\{i \leq n : X_i \in I\}|}{n} - \mathrm{P}(X \in I) \right| \to 0.$$

Break the line up into m intervals $I_1, I_2, \ldots, I_m$, with probabilities $1/m$. For large n we will have

$$\left| \frac{|\{i \leq n : X_i \in I_j\}|}{n} - \mathrm{P}(X \in I_j) \right| \leq \frac{1}{m^2}$$

for all j.

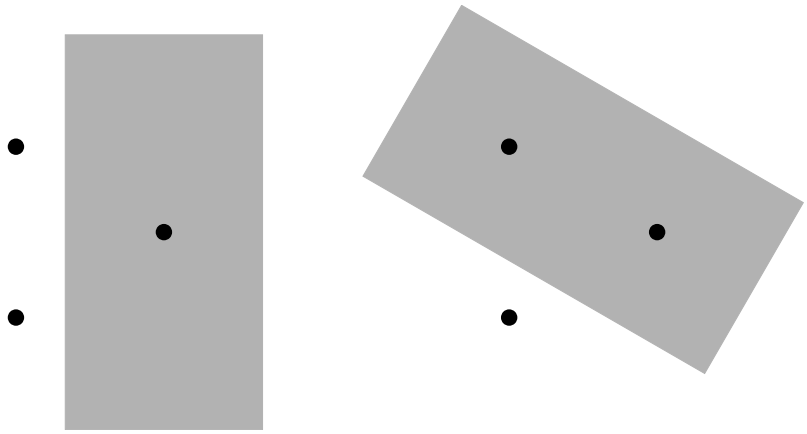
Now for a general interval we can approximate it by unions of our special intervals:



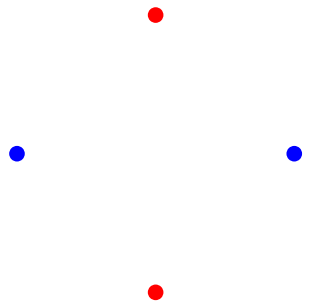
For which families of sets can we guarantee this uniformity? Does it work for the family of all sets? No because if you give me the sequence of values $X_1, X_2, \ldots$ I can look at the set $A = \{X_1, X_2, \ldots\}$ which gets probability 1 from each sample but may have probability zero for the original random variable.

It turns out that the GC Theorem holds for any family which is not too complicated in a sense that we will describe. For a finite set $U = \{u_1, u_2, \ldots u_m\}$ we say that the family $\mathcal{F}$ *shatters* U if for every subset $V \subset U$ we can find an element $A \in \mathcal{F}$ which cuts V out of U in the sense that $A \cap U = V$.

For example, suppose U consists of 3 corners of a triangle in the plane and $\mathcal{F}$ is the family of half-spaces. Each of the 8 subsets of U can be obtained by intersecting U with an appropriate half-space.



However if we pick 4 points in the plane we cannot shatter the set with half-spaces.



Any half-space containing the red points must contain at least one of the blue points. This is really a consequence of Radon's Lemma. Exercise: check that you cannot shatter 4 points one of which is in the convex hull of the other three. So in this sense the family of half-spaces is not too complicated. It can shatter (certain) sets of size 3 but no set of size 4.

We can ask for a general family $\mathcal{F}$ "What is the biggest set we can shatter?" The answer is called the Vapnik-Cervonenkis dimension of $\mathcal{F}$. The *VC-dimension* of a family $\mathcal{F}$ of subsets of Ω is the largest k for which there is a subset of Ω of size k, which is shattered by $\mathcal{F}$. The VC-dimension of the family of half-spaces in the plane is 3.

Suppose that Ω is $[n]$. How large can a family of subsets be before it shatters a set of size k? If $\mathcal{F}$ consists of all sets of size at most $k-1$ then it will not shatter a set of size k. So the family can certainly have as many as

$$\sum_{i=0}^{k-1} \binom{n}{i}$$

elements. It turns out that this is the largest number of members the family can have.

Theorem (**Sauer-Shelah Lemma**). *Let $\mathcal{F}$ be a family of subsets of $[n]$ with more than*

$$\sum_{i=0}^{k-1} \binom{n}{i}$$

members (for some $k \leq n$). Then $\mathcal{F}$ shatters a subset of $[n]$ of size k.

Proof We use induction on n. If $n = 1$ the theorem is obvious. So suppose that $n > 1$ and the result holds for smaller ground sets.

Given the family $\mathcal{F}$ we create two families of sets in $[n-1]$. $\mathcal{F}_1$ will consist of all subsets A of $[n-1]$ for which at least one of A and $A \cup \{n\}$ belongs to $\mathcal{F}$. $\mathcal{F}_2$ will consist of all subsets A of $[n-1]$ for which both A and $A \cup \{n\}$ belong to $\mathcal{F}$.

I claim that $|\mathcal{F}| = |\mathcal{F}_1| + |\mathcal{F}_2|$. Clearly

$$|\mathcal{F}_1| = \sum_{A \subset [n-1]} \mathbf{1}_{A \in \mathcal{F}_1} \quad \text{and} \quad |\mathcal{F}_2| = \sum_{A \subset [n-1]} \mathbf{1}_{A \in \mathcal{F}_2}$$

and

$$|\mathcal{F}| = \sum_{A \subset [n-1]} \mathbf{1}_{A \in \mathcal{F}} + \sum_{A \subset [n-1]} \mathbf{1}_{A \cup \{n\} \in \mathcal{F}}.$$

So it suffices to check that for every $A \subset [n-1]$

$$\sum_{A \subset [n-1]} \mathbf{1}_{A \in \mathcal{F}} + \sum_{A \subset [n-1]} \mathbf{1}_{A \cup \{n\} \in \mathcal{F}} = \sum_{A \subset [n-1]} \mathbf{1}_{A \in \mathcal{F}_1} + \sum_{A \subset [n-1]} \mathbf{1}_{A \in \mathcal{F}_2}.$$

We can do it by checking cases

- If $A \in \mathcal{F}$ but $A \cup \{n\} \notin \mathcal{F}$ then A is in $\mathcal{F}_1$ but not in $\mathcal{F}_2$.
- If $A \notin \mathcal{F}$ but $A \cup \{n\} \in \mathcal{F}$ then A is in $\mathcal{F}_1$ but not in $\mathcal{F}_2$.
- If $A \in \mathcal{F}$ and $A \cup \{n\} \in \mathcal{F}$ then A is in $\mathcal{F}_1$ and in $\mathcal{F}_2$.

In each case the set A is counted the same number of times on each side.

Now we know that

$$\sum_{i=0}^{k-1} \binom{n}{i} = \sum_{i=0}^{k-1} \binom{n-1}{i} + \sum_{i=1}^{k-1} \binom{n-1}{i-1} = \sum_{i=0}^{k-1} \binom{n-1}{i} + \sum_{i=0}^{k-2} \binom{n-1}{i}$$

by the Pascal triangle property of binomial coefficients.

So if $|\mathcal{F}| > \sum_{i=0}^{k-1} \binom{n}{i}$ then either $|\mathcal{F}_1| > \sum_{i=0}^{k-1} \binom{n-1}{i}$ or $|\mathcal{F}_2| > \sum_{i=0}^{k-2} \binom{n-1}{i}$. In the first case the inductive hypothesis tells us that the family $\mathcal{F}_1$ shatters a subset of $[n-1]$ of size k. In that case $\mathcal{F}$ will shatter the same subset. In the second case the family $\mathcal{F}_2$ shatters a subset σ of $[n-1]$ of size $k-1$. For each set B in $\mathcal{F}_2$ both B and $B \cup \{n\}$ belong to $\mathcal{F}$ which ensures that $\mathcal{F}$ shatters the set $\sigma \cup \{n\}$ which has size k. □

What is the number of different sets that we can cut out of m points in the plane with half-spaces? Since we can't shatter a set of size 4 the maximum is

$$\binom{m}{0} + \binom{m}{1} + \binom{m}{2} + \binom{m}{3}$$

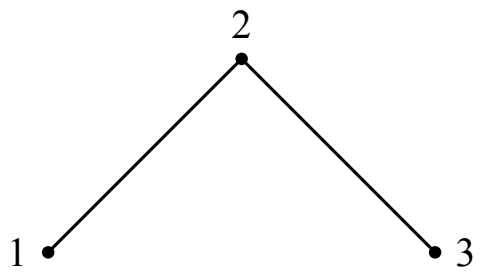
or roughly $m^3/6$. Actually the true number is less than this if m is large. When $m = 5$ the sum is $1 + 5 + 10 + 10 = 26$ but in fact you can only cut out 22 sets. For general m the number is $m(m-1) + 2$. HW

Volume II. Graph Theory

Chapter 5. Graph colouring

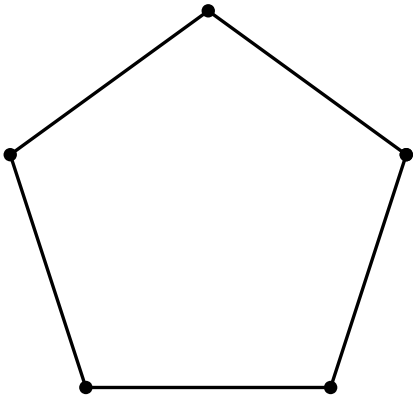
Recap

A graph G is a collection of **vertices** $V = \{v_1, v_2, \ldots, v_n\}$ together with a set of **edges** E each of which is a pair of vertices. For example if $V = \{1, 2, 3\}$ and $E = \{\{1, 2\}, \{2, 3\}\}$ we get

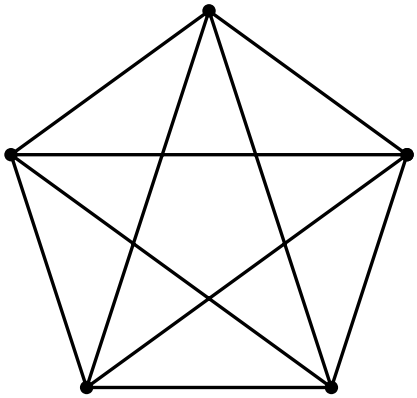


We don't allow loops, multiple edges or directed edges unless we explicitly say so. A number of special graphs turn up a lot: The **path** P_n of length n which has $n + 1$ vertices:

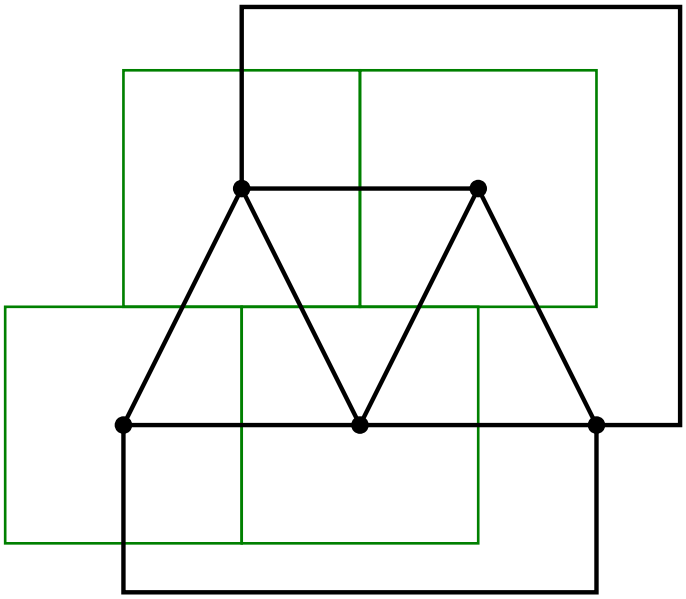
The **cycle** C_n of length n:



The **complete graph** K_n:



One of the most famous problems in combinatorics was the 4-colour problem. Is it possible to colour every plane map with 4 colours so that neighbouring countries have different colours? We can rephrase this in terms of plane graphs. Is it possible to colour the **vertices** of a plane graph with 4 colours so that adjacent vertices have different colours?



This was answered in the affirmative by Appel and Haken in 1976. They used a computer aided reduction of hundreds of critical cases. In last year's course we proved the 5-colour theorem.

Our purpose in this course will be to study graph colouring as a question in its own right independent of planarity. If I give you a graph, can you colour its vertices with k colours so that any pair of adjacent vertices get different colours? How many colours do you need for the answer to be yes? The *Chromatic Number* $\chi(G)$ of a graph G is the smallest number of colours with which the vertices of G can be coloured so that adjacent vertices have different colours.

There are some simple examples. The complete graph K_n on n vertices needs n colours. A cycle of even length can be coloured with 2 colours. More generally, a graph is bipartite if and only if its chromatic number is at most 2. A cycle of odd length needs 3 colours. The triangle is the obvious example.

A graph whose maximum degree is Δ needs at most $\Delta+1$ colours. Start with a vertex and choose a colour. Now look at another vertex and choose a colour for it. We can keep doing this because we always have a spare colour to use on each new vertex. A planar graph has chromatic number at most 4: the 4-Colour Theorem.

Brooks' Theorem

The bound $\Delta + 1$ for the number colours needed to colour a graph with maximum degree Δ is rarely sharp.

Theorem (**Brooks' Theorem**)**.** *Let G be a connected graph with maximum degree Δ. Unless G is a complete graph or an odd cycle, the chromatic number of G is at most Δ.*

We shall give a proof of this theorem due to Lovász and we shall need a bit of preparation. Some observations are obvious: if $\Delta = 1$ the only connected graph is K_2. If $\Delta = 2$ the graph is a path or a cycle.

Lemma (**Lemma to Brooks' Theorem**)**.** *Let G be a connected graph with maximum degree Δ which has a vertex of degree less than Δ. Then the chromatic number of G is at most Δ.*

Proof Let x be a vertex of degree less than Δ. For each vertex determine its distance to x: the length of the shortest path to x. Now consider the vertices furthest from x. Each one is adjacent to a vertex which is closer to x so the graph induced by the furthest vertices has degree at most $\Delta - 1$. So we can colour it with Δ colours. Now consider the vertices one step closer to x. Again each is adjacent to a vertex closer to

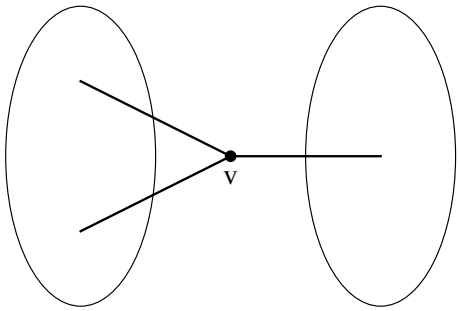
x so we can colour each of these with Δ colours taking into account the colours used for the previous layer. Continue until only x remains. It has degree at most $\Delta - 1$ so we have a colour left for x. □

The proof we shall give of Brooks' Theorem is a slight variant of the one found by Lovász and breaks into cases according to how well-connected the graph is. A graph is called k-connected if we need to remove at least k vertices in order to disconnect it. So a connected graph is 1-connected. A 2-connected graph remains connected even if you take away a vertex: whichever one you take away.

Recall that an *induced* subgraph of a graph G is a subgraph that consists of a subset of the vertices of G and *all* the edges of G that connect vertices in this subset.

Theorem (**Brooks' Theorem**)**.** *Let G be a connected graph with maximum degree Δ. Unless G is a complete graph or an odd cycle, the chromatic number $\chi(G)$ is at most Δ.*

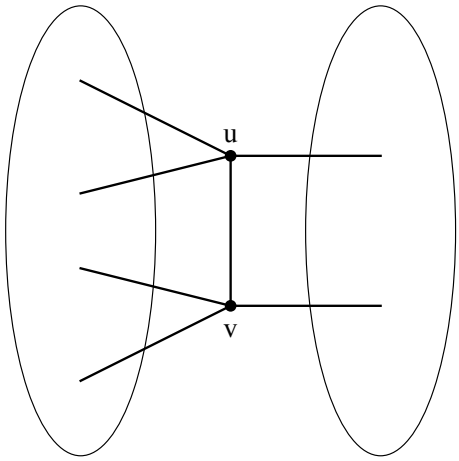
Proof We may assume that $\Delta \geq 3$ and that G is not a complete graph. We consider 3 cases.
Case 1 Suppose G is (connected but) not 2-connected and let v be a vertex whose removal disconnects G.



Consider the subgraph of G induced by a connected component of $G - \{v\}$ together with v. It has a vertex, namely v, with degree at most $\Delta - 1$ and so can be coloured with our Δ colours using the lemma. By permuting the colourings for the different components we can ensure that v gets the same colour in each case and then put them together to give a colouring for G.

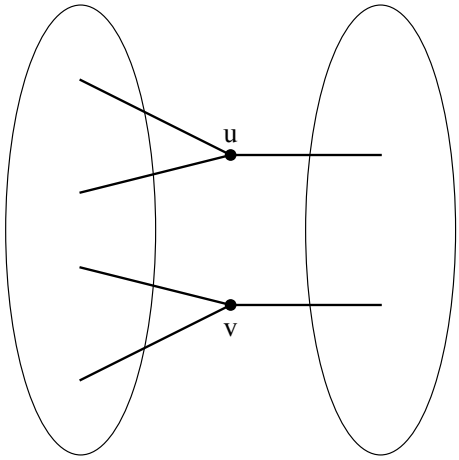
Case 2 Suppose G is 2-connected but not 3-connected and let (u, v) be a pair of vertices

whose removal disconnects G. I want to colour the subgraph of G induced by each component of $G - \{u, v\}$ together with u and v and then put the colourings together.

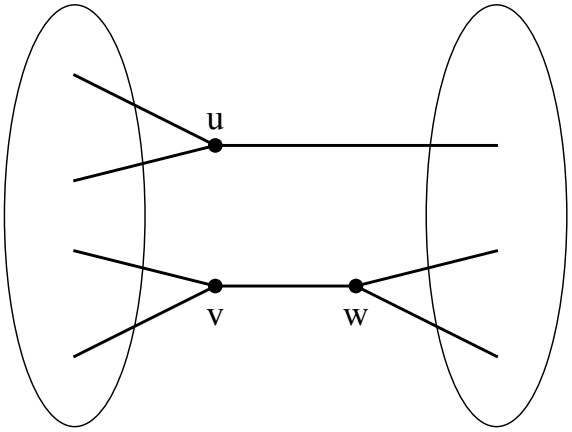


If u and v are adjacent then in each of the colourings they get different colours and so by permuting colours in the different colourings we can put them together.

If u and v are *not* adjacent then we can add in the edge uv each time we colour a component, without increasing the maximum degree beyond Δ. However we might increase the degrees of both u and v back to Δ and then have no vertex of degree less than Δ to apply the lemma.

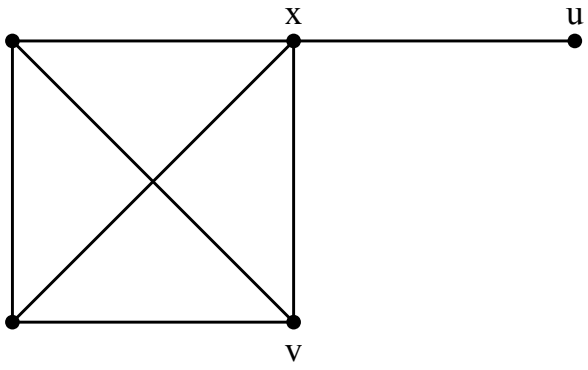


This happens if and only if $G - \{u, v\}$ has just two components L and R and u and v each have just one edge going into one of them: say R. In this case, replace v by its neighbour w in R. The vertices u and w disconnect G and now in each component of $G - \{u, w\}$ at least one of u and w has at most $\Delta - 2$ neighbours.



Therefore when we add in the edge uw each piece still has maximum degree Δ and at least one vertex of smaller degree.

Case 3 This is really the crucial case. Suppose G is 3-connected. I claim that we can find an induced path in G of length 2, uxv say. Let S be a maximal complete subgraph of G. Since G is connected but not complete there is a vertex u of G that is not in S but is adjacent to a vertex x in S. By the maximality of S there is another vertex v of S to which u is not adjacent.



Colour the vertices u and v with the same colour. The graph $G-\{u,v\}$ is connected so each vertex in it has a distance from x in this graph. Colour the vertices furthest from x respecting the colour of u and v: each of these has a neighbour closer to x so we can always colour with our Δ colours. Continue colouring vertices closer and closer to x. Finally, when we reach x it has two neighbours u and v with the same colour so we have a spare colour for x. □

The Chromatic Polynomial

For a graph G we define the function $P_G : \mathbf{N} \to \mathbf{N} \cup \{0\}$ as follows. $P_G(k)$ is the number of different ways to colour the graph G with k colours in such a way that adjacent vertices receive different colours. (Two colourings are different if there is a vertex which gets different colours in the two colourings).

Example (The chromatic polynomial of a complete graph)**.**

$$P_{K_n}(k) = k(k-1)(k-2)\dots(k-n+1).$$

If $k < n$ then we cannot colour K_n with k colours so $P_{K_n}(k) = 0$. If $k \geq n$ then we can colour: how many ways are there to do it? We can choose any one of k colours for the first vertex, any one of the remaining $k-1$ for the second and so on. So $P_{K_n}(k) = k(k-1)(k-2)\dots(k-n+1)$. Notice that this formula gives 0 if $1 < k < n$. So it is the correct value for all positive integers.

Example (The chromatic polynomial of a path)**.**

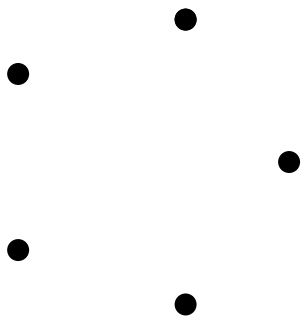
$$P_{P_n}(k) = k(k-1)^n.$$

If $k = 1$ and $n \geq 1$ then we cannot colour. If $k > 1$ we can choose any colour for the first vertex. Now colour each of the vertices in turn, moving along the path. At each point we have a choice of $k-1$ colours since we just have to avoid the one next door that we already coloured. So $P_{P_n}(k) = k(k-1)^n$. Again this is the correct value for $k = 1$.

The empty graph E_n of order n is the graph with n vertices and no edges.

Example (The chromatic polynomial of an empty graph)**.**

$$P_{E_n}(k) = k^n.$$



We have k choices for each vertex so $P_{E_n}(k) = k^n$.

So we have seen that

$$\begin{aligned} P_{K_n}(k) &= k(k-1)(k-2)\dots(k-n+1) \\ P_{P_n}(k) &= k(k-1)^n \\ P_{E_n}(k) &= k^n \end{aligned}$$

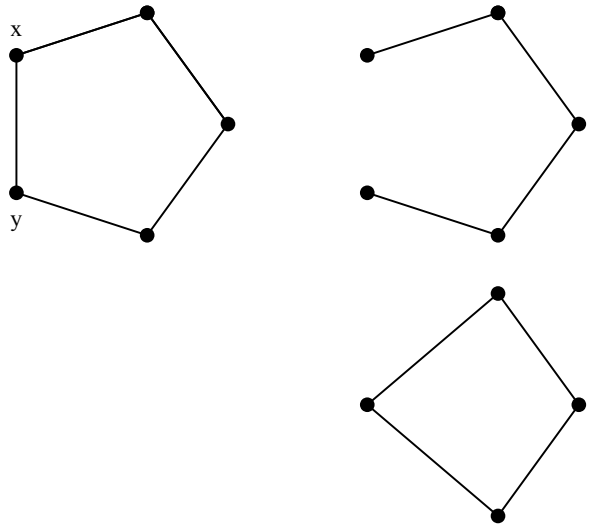
In each case the function P_G is a polynomial. This turns out to be true for all graphs. This fact will be the first theorem of the section. It is not something obvious.
It looks as though we should use induction on the number of vertices. Pick a vertex: colour the rest. Now ask how many choices we have for the last vertex. The problem is that it depends not only on the number of neighbours but also on whether those neighbours got the same colour. Instead we shall use induction on the number of edges.

Suppose G is a graph on n vertices and xy is an edge. Let $G - \{xy\}$ be the new graph we get by deleting that edge. Every colouring of G with k colours is also an admissible colouring of $G - \{xy\}$. But $G - \{xy\}$ has a few more colourings: namely those in which x and y are given the same colour. It turns out that we can count the number of colourings of $G - \{xy\}$ in which x and y get the same colour by using a different graph which we will call $G/\{xy\}$. We then have

$$P_G(k) = P_{G-\{xy\}}(k) - P_{G/\{xy\}}(k).$$

As long as this other graph has fewer edges than G we have a viable inductive step.

What is this other graph? How can we construct a graph which is like $G - \{xy\}$ except that when we colour it we force the vertices x and y to have the *same* colour? Suppose G is the 5-cycle with x and y as shown. We can replace the two vertices x and y by a single vertex z and join it to each vertex that used to be adjacent to either of x or y. We "pinch" the vertices x and y together.



Now, if you have a colouring of $G - \{xy\}$ with x and y getting the same colour (say red) you can turn it into a colouring of the new graph: colour z red and all the other vertices as they were. Any vertex that was adjacent to either x or y can't have been red. On the other hand, if you have a colouring of the new graph you can transfer it back to $G - \{xy\}$ by choosing the colour of z for both of x and y.

What happens in a more complicated situation in which some of the neighbours of x are also neighbours of y? Nothing changes. We can still pinch together x and y and join the new vertex to all the vertices that were adjacent to either x or y. Again, each colouring of $G - \{xy\}$ in which x and y have the same colour corresponds to exactly one colouring of $G/\{xy\}$.

If G is a graph and $\{x, y\}$ an edge of G we define $G/\{xy\}$ to be the graph obtained from G by replacing x and y by a single vertex adjacent to all neighbours of either x or y and leaving all other vertices and the edges between them, the same.

Theorem (**The chromatic polynomial**)**.** *For any graph G in which $\{x, y\}$ is an edge and any $k \geq 1$*

$$P_G(k) = P_{G-\{xy\}}(k) - P_{G/\{xy\}}(k).$$

Consequently P_G is a polynomial.

Proof Every colouring of $G - \{xy\}$ with k colours either assigns different colours to x and y in which case it corresponds to a colouring of G or assigns them the same colour.

So it suffices to check that the number of colourings of $G - \{xy\}$ in which x and y receive the same colour is equal to the number of colourings of $G/\{xy\}$.

If you have a colouring of $G - \{xy\}$ with x and y getting the same colour (say red) you can turn it into a colouring of the new graph: colour the new vertex red and all the other vertices as they were. Any vertex that was adjacent to either x or y can't have been red. On the other hand, if you have a colouring of the new graph you can transfer it back to $G - \{xy\}$ by choosing the colour of the new vertex for both x and y.

We now deduce that P_G is a polynomial by induction on the number of edges. If G has no edges it is the empty graph and we already saw that its $P_{E_n}(k) = k^n$. If G has an edge $\{x, y\}$ we can write P_G as a sum of chromatic polynomials of graphs with fewer edges so it too is a polynomial. □

Let's have an example in which we use the recurrence to calculate a chromatic polynomial.

Example (**The chromatic polynomial of a cycle**)**.** *The chromatic polynomial of the n-cycle is*

$$P_{C_n}(k) = (k-1)^n + (-1)^n(k-1).$$

Proof If $n = 3$ then the n-cycle is K_3 so we know its chromatic polynomial. The formula gives

$$\begin{aligned}(k-1)^3 + (-1)^3(k-1) &= (k-1)^3 - (k-1) \\ &= (k-1)(k^2 - 2k + 1 - 1) = k(k-1)(k-2)\end{aligned}$$

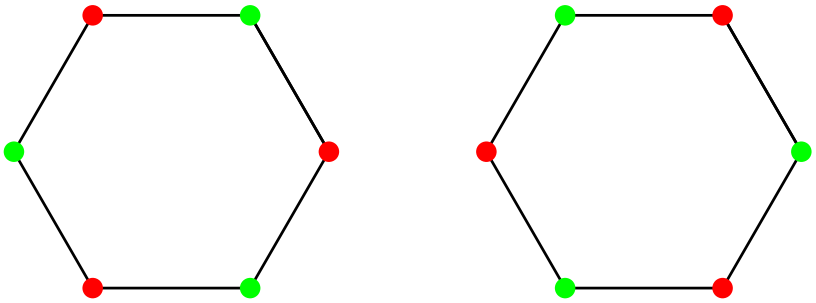
as required.

Now suppose that $n > 3$ and the result holds for smaller cycles. If $G = C_n$ then $G - \{xy\} = P_{n-1}$ and $G/\{xy\} = C_{n-1}$. Consequently

$$\begin{aligned}P_G(k) &= k(k-1)^{n-1} - (k-1)^{n-1} - (-1)^{n-1}(k-1) \\ &= (k-1)^n + (-1)^n(k-1)\end{aligned}$$

completing the inductive step. □

Note that if $k = 2$ this expression is 0 if n is odd but 2 if n is even.



There are indeed two ways to colour an even cycle with 2 colours.

When the chromatic polynomial was introduced there was a hope that it might solve the 4-colour problem. This turned out to be hopelessly optimistic. However the polynomial is of considerable interest. Among other things it is intimately connected to statistical mechanics. There are a number of properties of the polynomial that we can immediately deduce from the recurrence relation.

Theorem (**Simple properties of the chromatic polynomial**)**.** *If G is a graph of order n then P_G has degree n and its leading coefficient is 1. The coefficients of P_G alternate in sign. For any graph G, $P_G(0) = 0$.*

Proof For the first part we use induction on the number of edges. We already saw that for the empty graph $P_{E_n} = k^n$ which is a monic polynomial of degree n. Now suppose G has some edges and the result holds for graphs with fewer edges. We have

$$P_G(k) = P_{G-\{xy\}}(k) - P_{G/\{xy\}}(k).$$

The graph $G - \{xy\}$ has order n and fewer edges so $P_{G-\{xy\}}$ is a monic polynomial of degree n. The graph $G/\{xy\}$ has order $n-1$ and fewer edges so its chromatic polynomial does not contribute to the k^n term in P_G.

For the second we can also use induction. For the empty graph the coefficients are all zero after the first so they have the correct signs. Since $P_{G/\{xy\}}$ has degree $n-1$ its coefficients have the opposite sign to those of $P_{G-\{xy\}}$ so when we subtract we maintain the signs.

For the third we again use induction. $P_{E_n}(k) = k^n$ is 0 if $k = 0$ so the induction starts. The inductive step is obvious. □

Thus if G is a graph of order n then

$$P_G(k) = k^n - c_{n-1}k^{n-1} + c_{n-2}k^{n-2} - \cdots$$

In the HW I will ask you to determine the coefficients c_{n-1} and c_{n-2} in terms of properties of G. As was mentioned in the introduction and a couple of slides earlier, the chromatic polynomial is related to statistical mechanics. The partition function for the Potts model of quantum spin systems is equivalent via a change of variable to the Tutte polynomial and a special case to the chromatic polynomial.

Chapter 6. Matroids

Introduction

In the combinatorics course we proved Kirchoff's Matrix Tree Theorem by using a link between acyclic graphs and independent sets in a vector space. This link can be put into a more abstract setting which is sometimes useful: the theory of matroids. Recall the Exchange Lemma stating that if A and B are two linearly independent sets in a vector space and A is larger than B then we can transfer an element a from A to B so as to produce a linearly independent set $\{a\} \cup B$.

Definition (Matroid)**.** *A matroid is a set E together with a collection $\mathcal{I}$ of its subsets satisfying*

- $\emptyset \in \mathcal{I}$
- *If $A \subset B$ and $B \in \mathcal{I}$ then $A \in \mathcal{I}$*
- *If $A, B \in \mathcal{I}$ and $|A| > |B|$ then there is an element $a \in A$ for which $\{a\} \cup B \in \mathcal{I}$.*

The sets in $\mathcal{I}$ are called the *independent* sets of the matroid. While this section is not strictly speaking graph theory, one can think of matroids as generalisations of graphs.

The basic example of a matroid is any collection E of vectors in a vector space with $\mathcal{I}$ being the collection of linearly independent subsets of E. Very simple examples of matroids are the uniform matroids. If E is a set, r is a positive integer and $\mathcal{I}$ is the collection of subsets of E that have at most r elements then $(E, \mathcal{I})$ is a matroid. Exercise.

The next example is the one that we already met. Let G be a graph and E the set of its edges. The family $\mathcal{I}$ will consist of all the sets of edges that contain no cycles. Suppose G has n vertices and for each pair $\{i, j\}$ with $i < j$ let e_{ij} be the vector

$$(0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)$$

which has a 1 in the i^{th} place and -1 in the j^{th}.

We saw in the combinatorics course that the vectors corresponding to a set of edges are linearly independent if and only if the edges form no cycle. So the pair $(E, \mathcal{I})$ is a matroid. Let's see a direct proof of this. We shall need the following from last year:

Lemma (**Tree properties**). *For a graph G with n vertices, any two of the following implies the third (and hence that G is a tree).*

- *G has $n-1$ edges*
- *G is connected*
- *G is acyclic*

Lemma (**Graphic matroids**). *If G is a graph, E is its edge set and $\mathcal{I}$ is the family of acyclic subsets of E then $(E, \mathcal{I})$ is a matroid.*

Proof The first two matroid properties are obvious. Suppose that A and B are two acyclic sets of edges with $|A| > |B|$. Look at the graph G_B whose vertex set is that of G and whose edges are those in B. It is acyclic so its components are trees. We want to find an edge e in A that crosses between different components of G_B since then $\{e\} \cup B$ is acyclic. If there is no such edge then all edges in A lie within the components of G_B. But since A is acyclic it cannot have more edges in each component than a tree does: so it cannot have more elements than B. Here we used the Tree Properties Theorem. □

As explained earlier we know from last year's course that Graphic Matroids can be represented by systems of vectors. This situation has a formal name. A matroid $(E, \mathcal{I})$ is said to be *representable* over a field $\mathbf{F}$ if there is a map $\phi : E \to V$ into a vector space V over $\mathbf{F}$ such that for each $A \subset E$, $A \in \mathcal{I}$ if and only if $\phi(A)$ is a linearly independent set in V.

Graphic matroids can be represented over any field. The uniform matroid $U_{4,2}$ cannot be represented over $\mathbf{Z}_2$. HW However $U_{n,r}$ is representable over the reals for every n and r. HW

Rado's Theorem

Recall from last year Hall's Marriage Theorem.

Theorem (**Hall's Marriage Theorem**). *Let G be a bipartite graph with vertex classes A and B. For each subset $U \subset A$ let $\Gamma(U)$ be the set of neighbours of vertices in U:*

$$\Gamma(U) = \{b : ab \text{ is an edge for some } a \in U\}.$$

If for every $U \subset A$ the set $\Gamma(U)$ is at least as large as U then G contains a complete matching from A into B.

Another way to state this is as follows. Suppose that $S_1, S_2, \ldots, S_n$ are subsets of a set E and for each set $\sigma \subset \{1, 2, \ldots, n\}$ of indices

$$\left| \bigcup_{i \in \sigma} S_i \right| \geq |\sigma|.$$

Then we can find a *transversal*: a set $\{e_1, e_2, \ldots, e_n\}$ of n distinct elements of E with $e_i \in S_i$ for each i. To prove it we apply Hall's Theorem to the bipartite graph in which the S_i are vertices on the left and E the set of vertices on the right with an edge $\{S_i, e\}$ if $e \in S_i$. Exercise.

Now suppose that we have sets S_i in a matroid $(E, \mathcal{I})$. Under what conditions can we find a transversal that is *independent*? For a set $A \subset E$ let the *rank* $r(A)$ of A be the size of the largest independent set inside A. From the matroid properties we can see that if I is an independent subset of A then we can extend I to an independent set in A with $r(A)$ elements. Among other things, all maximal independent sets have the same size.

Theorem (**Rado's Theorem**)**.** *Let $(E, \mathcal{I})$ be a matroid and $S_1, S_2, \ldots, S_n$ be subsets of E. Suppose that for each set $\sigma \subset \{1, 2, \ldots, n\}$ of indices*

$$r\left(\bigcup_{i \in \sigma} S_i \right) \geq |\sigma|.$$

Then there is an independent transversal: a set $\{e_1, e_2, \ldots, e_n\} \in \mathcal{I}$ of n distinct elements of E with $e_i \in S_i$ for each i.

Just as in Hall's Theorem the sufficient condition is obviously necessary. If an independent transversal exists then the union

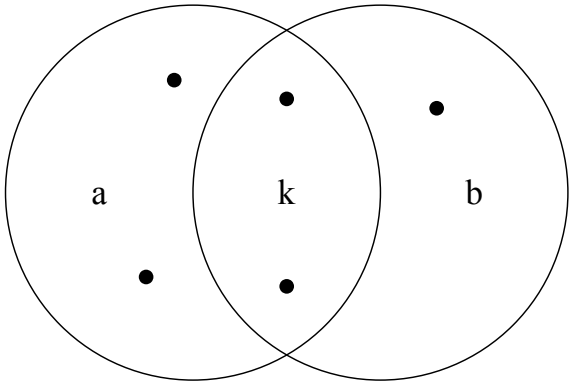
$$\bigcup_{i \in \sigma} S_i$$

includes $\{e_i : i \in \sigma\}$ which is independent and so the rank of the union is at least $|\sigma|$. The problem is the sufficiency.

The key to the proof is to observe that the rank function on a matroid is *submodular* which for our purposes means that for any $A, B \subset E$

$$r(A \cup B) + r(A \cap B) \leq r(A) + r(B).$$

To see this let $r(A \cup B) = m$ and $r(A \cap B) = k$, choose an independent set of size k in $A \cap B$ and extend it to an independent set of size m in $A \cup B$. The extended set contains no further elements in $A \cap B$: suppose it contains a elements in $A - B$ and b in $B - A$.



Then $m = a + k + b$ and hence

$$r(A \cup B) + r(A \cap B) = m + k = (a + k) + (b + k) \leq r(A) + r(B)$$

as required.

Proof (**of Rado**) Suppose each S_i had just one element. Then

$$\bigcup_i S_i$$

is a set of size n whose rank is n. So it contains an independent set of size n, itself, which consists of the single element from each S_i. We shall show that if some S_i has more than one element we can remove one of them without violating the hypothesis. By doing this repeatedly we eventually arrive at the case we already considered.

So suppose S_1 contains two elements a and b. If we can't throw out either one then there must be two sets of indices σ and τ not containing 1 for which

$$r\left((S_1 - \{a\}) \cup \bigcup_{i \in \sigma} S_i\right) = |\sigma|$$

and

$$r\left((S_1 - \{b\}) \cup \bigcup_{i \in \tau} S_i\right) = |\tau|.$$

Let $A = (S_1 - \{a\}) \cup \bigcup_{i \in \sigma} S_i$ and $B = (S_1 - \{b\}) \cup \bigcup_{i \in \tau} S_i$.

Now

$$A \cup B = \bigcup_{i \in \sigma \cup \tau \cup \{1\}} S_i$$

$$A \cup B = \bigcup_{i \in \sigma \cup \tau \cup \{1\}} S_i$$

so

$$r(A \cup B) \geq |\sigma \cup \tau| + 1.$$

$$A \cap B \supset \bigcup_{i \in \sigma \cap \tau} S_i$$

so

$$r(A \cap B) \geq |\sigma \cap \tau|.$$

Hence

$$\begin{aligned} |\sigma| + |\tau| + 1 &= |\sigma \cup \tau| + |\sigma \cap \tau| + 1 \\ &\leq r(A \cup B) + r(A \cap B) \\ &\leq r(A) + r(B) \\ &= |\sigma| + |\tau| \end{aligned}$$

giving a contradiction. □

This proof has an amusing history. It was found by Welsh but actually combines a proof that Rado gave, not for his theorem but for Hall's Theorem, with the submodularity idea.

We will use Rado's Theorem to prove a lovely theorem of Horn.

Theorem (**Horn's Theorem**)**.** *Suppose $x_1, x_2, \ldots, x_n$ are vectors in a vector space and for each set σ of indices*

$$\dim(\text{span}(\{x_i : i \in \sigma\})) \geq \frac{|\sigma|}{2}.$$

Then the set of vectors can be partitioned into two linearly independent sets.

Proof Define a matroid $(E, \mathcal{I})$ as follows: the set E consists of 2 copies of the vectors $x_1, x_2, \ldots, x_n$. Call them $x_1, x_2, \ldots, x_n$ and $x'_1, x'_2, \ldots, x'_n$. A subset will be independent in the matroid if its x elements are linearly independent and its x' elements are linearly independent. We need to check that this *is* a matroid. If A and B are in $\mathcal{I}$ and A is larger than B then either A has more x elements or more x' elements than B (or both). In the first case we can choose an x element from A to add to B so that the enlarged collection of x elements is still linearly independent (and the x' elements are unchanged) so the enlarged B is still in $\mathcal{I}$. Similarly in the second case.

Now choose subsets $S_1 = \{x_1, x'_1\}$ and so on. An independent transversal consists of one vector with each index, some of them x and some x'. The x vectors are linearly independent and so are the x' vectors so this is just a partition of the vectors into two linearly independent sets. So it suffices to check that the S_i satisfy the hypothesis of Rado's Theorem. Let σ be a set of indices. By the hypothesis of the theorem there is a subset $\tau \subset \sigma$ of at least half the size for which the vectors $\{x_i : i \in \tau\}$ are linearly independent. But then

$$\bigcup_{i \in \sigma} S_i$$

contains the set

$$\bigcup_{i \in \tau} S_i$$

which has at least $|\sigma|$ elements and is independent in the matroid. □

Matroids and greed

Matroids have an intriguing connection with greedy algorithms. Suppose we have a finite set E whose elements have non-negative weights $\{w_i : i \in E\}$. Suppose also we have a family $\mathcal{I}$ of subsets of E which is monotone: $A \in \mathcal{I}$ and $B \subset A$ implies that $B \in \mathcal{I}$. We want to choose a set in $\mathcal{I}$ whose total weight is as large as possible.

One way to try to do it is greedily. Order the elements of E by weight:

$$w_1 \geq w_2 \geq \cdots \geq w_n.$$

Go through the elements in turn and include an element if it does not throw you out of $\mathcal{I}$.

Theorem (**Matroid Greedy Theorem**)**.** *Suppose $(E, \mathcal{I})$ is a matroid with a weight function $w : E \to [0, \infty)$. Then the greedy algorithm yields the highest weight element of $\mathcal{I}$.*

The proof will show that if there is an independent set with larger weight than the set created by our algorithm then the algorithm would have run differently.

Proof Suppose that the greedy algorithm chooses a set $b_1, b_2, \ldots, b_k$ in that order and note that this is a maximal element of $\mathcal{I}$ otherwise there would be an element we could include that would have been included when it was considered.

Suppose that there is another maximal element of $\mathcal{I}$ with a larger sum: say $\{c_1, c_2, \ldots, c_k\}$ necessarily with the same number of elements. Let j be the first index for which $w(c_j) > w(b_j)$. The set $\{c_1, c_2, \ldots, c_j\}$ is larger than $\{b_1, b_2, \ldots, b_{j-1}\}$ so there is an element c_i of the former that can be included into the latter without leaving $\mathcal{I}$.

We have $w(c_i) \geq w(c_j) > w(b_j)$. So there is some $r \leq j$ with $w(b_{r-1}) \geq w(c_i) > w(b_r)$. But that means that when c_i was considered it would have been included by the algorithm instead of b_r. □

There is a converse to this theorem. If a family of sets is not a matroid then we can cook up a weight function for which the greedy algorithm does not work. Suppose A and B are sets in a monotone family $\mathcal{F}$ with $|A| > |B|$ but that for all $x \in A - B$ the augmented set $B \cup \{x\}$ is not in $\mathcal{F}$.

Define the weight function as follows

$$w(e) = \begin{cases} |A| + 1 & \text{if } e \in B \\ |A| & \text{if } e \in A - B \\ 0 & \text{otherwise.} \end{cases}$$

The elements of our set will be ordered as follows:

$$B, (A - B), E - A.$$

$$w(e) = \begin{cases} |A| + 1 & \text{if } e \in B \\ |A| & \text{if } e \in A - B \\ 0 & \text{otherwise.} \end{cases}$$

The algorithm will choose all elements of B and then no element of $A - B$. The total weight will be $(|A| + 1)|B| \leq (|A| + 1)(|A| - 1) < |A|^2$. But the set A is in the family and all its elements have weight at least $|A|$.

An obvious application of the Matroid Greedy Theorem is to the identification of minimal spanning trees. Suppose we are given a graph with weights on the edges. An important problem is to find the spanning tree with the smallest total weight. For example if the vertices are towns and the weights are the distances between towns you might want to find the cheapest way to connect all the towns with an electrical grid. We know that the spanning trees are the maximal independent sets in the graphic matroid of the graph.

The fact that we are trying to find the minimal spanning tree rather than maximal makes no difference because we can replace each weight w by $M - w$ where M is some number bigger than all the weights. We know that the greedy algorithm will work. We order the edges in increasing order of weight and then pick them one by one, never including an edge that makes a cycle.

Chapter 7. Random graphs

Introduction

In the chapter on Ramsey Theory in last year's course we saw an example of the use of random constructions in graph theory: namely the following theorem.

Theorem (**Erdös lower bound for $R(s,s)$**). *Let $s \geq 3$. Then*

$$R(s,s) \geq 2^{(s-1)/2}.$$

The idea is to colour the edges of K_n randomly. We colour each one red with probability $1/2$ and blue with probability $1/2$ and we colour the edges independently: the choice for one edge is not affected by the choices for the others.

Proof Colour the edges of K_n red or blue independently at random with probability $1/2$ each. What is the chance that a fixed set of s vertices form a red K_s? There are $s(s-1)/2$ edges to go red: so the chance is

$$2^{-s(s-1)/2}.$$

The chance that a fixed set of s vertices is monochromatic is just twice as big:

$$2 \times 2^{-s(s-1)/2}.$$

What is the expected number of monochromatic K_s graphs? There are $\binom{n}{s}$ places where it could occur. So the expected number is

$$\binom{n}{s} 2 \times 2^{-s(s-1)/2} < \frac{2n^s}{s! 2^{s(s-1)/2}}.$$

If we arrange for this number to be less than 1 then there will be colourings without any monochromatic K_s. If $n \leq 2^{(s-1)/2}$ then we succeed. □

Historically this argument was the start of a mathematical field called "Random Graph Theory" which has now become closely associated with the mathematical study of models in statistical physics. The theory has two slightly different themes.

- The construction of graphs with special properties (as in the argument above)
- The study of random graphs in their own right, sometimes as models for real world networks

This chapter will contain an example of each theme. In the first we shall see that a graph can have large chromatic number even if it is quite sparse. In the second we shall discuss the connectedness of random graphs. Both parts will use the following model: fix the order n of the graph and vertices labelled $1, 2, \ldots, n$ say. Fix a probability $p \in (0, 1)$ and include each edge with probability p independently of all the other edges. (So the earlier argument corresponded to $p = 1/2$ but now we just look at the graph consisting of say the red edges.)

The Poisson Distribution

Before we discuss this model further we shall recall the Poisson distribution. A Poisson random variable with mean μ is a random variable X which takes the values $0, 1, 2, \ldots$ with probabilities

$$\begin{aligned} \mathrm{P}(X = 0) &= e^{-\mu} \\ \mathrm{P}(X = 1) &= e^{-\mu}\,\mu \\ \mathrm{P}(X = 2) &= e^{-\mu}\,\tfrac{\mu^2}{2} \\ \mathrm{P}(X = 3) &= e^{-\mu}\,\tfrac{\mu^3}{6} \\ \vdots & \qquad \vdots \end{aligned}$$

Observe that

$$\sum_{k=0}^{\infty} \mathrm{P}(X = k) = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = e^{-\mu} e^{\mu} = 1$$

which is required for a probability distribution. The mean is indeed μ:

$$\mathrm{E}X = e^{-\mu} \sum_{k=0}^{\infty} k \frac{\mu^k}{k!} = e^{-\mu} \sum_{k=1}^{\infty} k \frac{\mu^k}{k!} = e^{-\mu} \mu \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} = \mu.$$

One can calculate the variance in a similar way but an alternative is to look at the moment generating function:

$$m(t) = \mathrm{E}e^{Xt} = \sum_{k=0}^{\infty} \frac{\mathrm{E}X^k}{k!} t^k.$$

We have

$$m(t) = \sum_{k=0}^{\infty} \frac{\mathrm{E}X^k}{k!} t^k = \mathrm{E}e^{Xt} = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} e^{kt} = e^{-\mu} e^{\mu e^t}.$$

Now

$$m(0) = 1.$$

$$m'(0) = \mathrm{E}X.$$

$$m''(0) = \mathrm{E}X^2.$$

We have

$$m''(t) = e^{-\mu} e^{\mu e^t} \left(\mu e^t + \mu^2 E^{2t} \right).$$

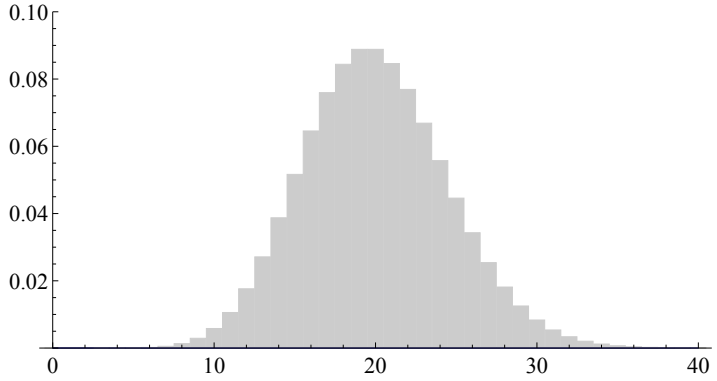
and so

$$\mathrm{E}X^2 = \mu + \mu^2.$$

Hence the variance is

$$\mathrm{E}X^2 - (\mathrm{E}X)^2 = \mu + \mu^2 - \mu^2 = \mu.$$

The variance of the Poisson is the same as its mean. Since the standard deviation is the square root of the variance, this tells us that when μ is large the Poisson concentrates around its mean. The figure shows the Poisson distribution with mean 20.



The Poisson distribution was originally created as an approximation to the binomial distribution. If n is large and X is a $B(n, p)$ random variable then the probability that $X = k$ is well approximated by $e^{-\mu}\mu^k/k!$ as long as k is small compared to $\sqrt{n}$ and kp is small. However the distribution is now of fundamental importance "in its own right". It measures the number of random events that occur in a given time interval for a variety of processes, in particular radioactive decay. For a sample of a radio-isotope, the number of decay events occurring in a time interval T has a Poisson distribution with mean λT where λ is a parameter depending upon the substance.

If μ is large then the probability that $X = 0$ is very small: $e^{-\mu}$. We shall see that just using the mean and variance we can get a much weaker estimate that is occasionally

useful. If the mean is μ and there were a fair chance of being at zero then the variance would be around μ^2 not just μ.

$$\text{var}(X) = \text{E}(X - \text{E}X)^2 \geq \text{Prob}(X = 0)(-\text{E}X)^2 = \text{Prob}(X = 0)(\text{E}X)^2.$$

Hence

$$\text{Prob}(X = 0) \leq \frac{\text{var}(X)}{(\text{E}X)^2} = \frac{1}{\mu}.$$

The Poisson Picture

Let us return to the random graph model. Fix the order n of the graph and vertices labelled $1, 2, \ldots, n$ say. Fix a probability $p \in (0, 1)$ and include each edge with probability p independently of all the other edges. We use $G_{n,p}$ to denote a random graph in this model and write for example

$$\text{Prob}(G_{n,p} \text{ is connected})$$

for the probability that a randomly chosen graph will be connected.

Each vertex has $n - 1$ potential edges coming out of it so if $p = K/n$ we expect the vertex to have degree about K. As p increases the graph is likely to become denser so it will be more likely to be connected or have a higher chromatic number. The graph certainly won't be connected if it has isolated vertices: vertices of degree 0. It is easy to calculate the expected number of these.

For each vertex the probability that it has no edges is $(1 - p)^{n-1}$. So the expected number of isolated vertices is $n(1 - p)^{n-1}$. This will be equal to Q if

$$(1 - p)^{n-1} = \frac{Q}{n}$$

or

$$p = 1 - \left(\frac{Q}{n}\right)^{\frac{1}{n-1}} = 1 - \exp\left(\frac{\log Q - \log n}{n - 1}\right) \approx \frac{\log n - \log Q}{n - 1}.$$

Thus if $p = \frac{\log n - \log Q}{n}$ the expected number of isolated vertices is about Q. The different vertices are not quite independent because there is an edge between each pair: if one is isolated the others are a bit more likely to be isolated. But they are almost independent. So we expect that the number of isolated vertices has roughly a Poisson distribution with mean Q. In that case the probability that no vertex is isolated would be e^{-Q}.

It turns out that for this type of random graph, the most likely way for it to be disconnected is to have isolated vertices. Once we choose p large enough to rule them out, the graph is almost bound to be connected. The most famous theorem in this direction is the following:

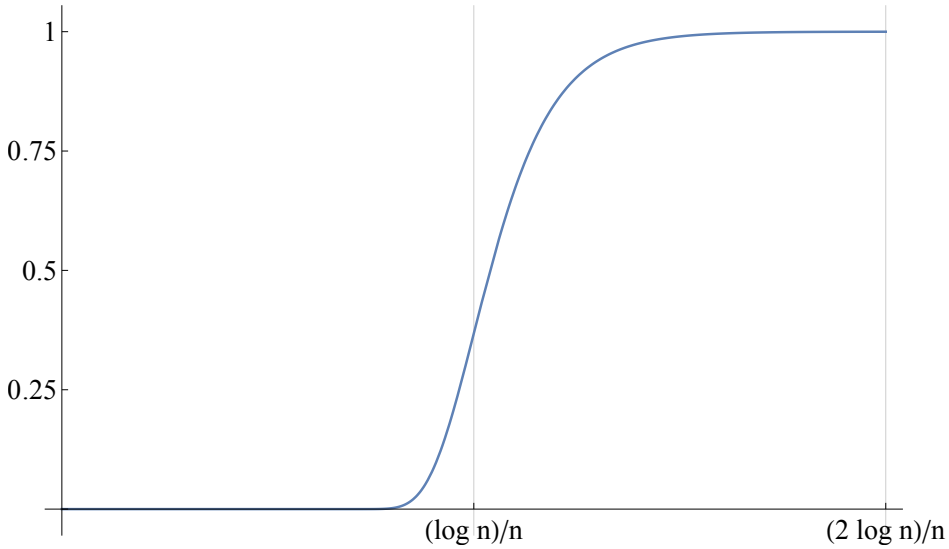
Theorem (**Erdös-Rényi**). *Let*

$$p = p(n) = \frac{\log n}{n} + \frac{c_n}{n}.$$

Then

$$\text{Prob}\,(G_{n,p} \text{ is connected}) \to \begin{cases} 0 & \text{if } c_n \to -\infty \\ e^{-Q} & \text{if } c_n \to -\log Q \\ 1 & \text{if } c_n \to \infty \end{cases}$$

The probability $p = \frac{\log n}{n}$ is called a *threshold* for connectedness. If you choose edges with slightly higher probability the graph is almost bound to be connected: if you choose edges with slightly lower probability the graph is almost bound *not* to be connected. The graph shows the probability of being connected as a function of p for $n = 10,000$. The probability changes suddenly from 0 to 1 as you cross the threshold.



The same phenomenon happens in models of quantum systems. As you cool a liquid you pass the freezing point and suddenly the whole structure changes into a crystal.

We shall prove the limiting cases of the theorem above.

Theorem (**Connectedness of random graphs**). *Let*

$$p = p(n) = \frac{\log n}{n} + \frac{c_n}{n}.$$

Then

$$\text{Prob}\,(G_{n,p} \text{ is connected}) \to \begin{cases} 0 & \text{if } c_n \to -\infty \\ 1 & \text{if } c_n \to \infty \end{cases}$$

The main difference between this and the theorem above is that we only get an estimate

$$\text{Prob}\,(G_{n,p} \text{ is connected}) \leq \frac{1}{Q}$$

rather than e^{-Q}.

The usual definition of threshold is rather weaker than we are proving here.

Definition (**Thresholds for random graphs**). *A sequence p_n is called a threshold for a property of $G_{n,p}$ if*

$$\text{Prob}\,(G_{n,\lambda_n p_n} \text{ has the property}) \to \begin{cases} 0 & \text{if } \lambda_n \to 0 \\ 1 & \text{if } \lambda_n \to \infty \end{cases}$$

As mentioned earlier, the other topic in this chapter will be the chromatic number of random graphs. Finding the chromatic number of a random graph was a long process involving subtle papers by Grimett and McDiarmid, Bollobás, Łuczak, Shamir and Spencer, Alon and Krivelevich, Friedgut, Achlioptas and Naor. The most recent result here is essentially the following:

Theorem (**The two values of the chromatic number**). *Let k be an integer and*

$$\frac{2k \log k - 2 \log k}{n} < p = p(n) < \frac{2k \log k}{n}.$$

Then

$$\text{Prob}\,(\chi(G_{n,p}) = k \text{ or } k+1) \to 1$$

as $n \to \infty$.

We shall prove something much weaker but at least we detect the correct order of the probability:

Theorem (**The chromatic number of a random graph**). *Suppose the probability p satisfies*

$$p \geq \frac{2k \log k + 4k}{n}$$

Then

$$\text{Prob}\left(\chi(G_{n,p}) \leq k\right) \leq e^{-n/(2k)}.$$

For this value of p with k quite a bit smaller than n the random graph $G_{n,p}$ is pretty sparse, not only in that its average degree is much less than n but also, as we shall see, in that it doesn't even have many short cycles.

Lemma (**Short cycles in a random graph**). *Let g and n be integers with,*

$$\frac{5}{n} \leq p \leq \frac{n^{1/g}}{n}$$

and let X be the number of cycles of length at most $g-1$ in $G_{n,p}$. Then

$$\text{E}X \leq \frac{n}{4}.$$

Using these two facts we shall be able to demonstrate the existence of graphs with large chromatic number and no short cycles.

We will repeatedly use the following facts which are left as exercises.

- For $0 < p < 1$ and any m we have $e^{-mp/(1-p)} \leq (1-p)^m \leq e^{-mp}$
- For $1 \leq k \leq n$
$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k.$$
- (**Markov or Chebyshev Inequality**) If X is a non-negative random variable and $t > 0$ then
$$\text{Prob}(X > t) \leq \frac{1}{t}\text{E}X.$$

Before we start let's recall how the linearity of the expectation enables us to compute expectations quite easily. For example let's calculate the expected number of triangles in $G_{n,p}$. There are $\binom{n}{3}$ possible triangles. Each triangle has probability p^3 of being included since we need all three of its edges to be included. For each triangle T let I_T be the random variable that is equal to 1 if T appears and 0 if not. Then the number of triangles in $G_{n,p}$ is

$$\sum I_T.$$

The random variables I_T are not independent: if one triangle is in, then any triangle that shares an edge with it is more likely to be in. But that is irrelevant as far as the expectation is concerned.

$$\mathrm{E}\sum I_T = \sum \mathrm{E}I_T.$$

The expectation of I_T is p^3 because the random variable is 1 with probability p^3 and 0 otherwise. So the expected number of triangles is $\binom{n}{3}p^3$.

In the 2018 exam, Question 4 asked for a proof that if $p = \frac{c_n}{n}$ with $c_n \to \infty$ then the probability that $G_{n,p}$ contains no triangles tends to 0. In fact $p = 1/n$ is a threshold for containment of a triangle.

We now return to the results that we shall prove. We begin with the chromatic number.

The chromatic number

Theorem (**The chromatic number of a random graph**). *Suppose the probability p satisfies*

$$p \geq \frac{2k\log k + 4k}{n}$$

Then

$$\mathrm{Prob}\left(\chi(G_{n,p}) \leq k\right) \leq e^{-n/(2k)}.$$

If a graph is coloured with k colours then one of the colour classes has at least n/k vertices in it. So it suffices to prove that if r is the integer satisfying

$$\frac{n}{k} \leq r < \frac{n}{k} + 1$$

then the probability that $G_{n,p}$ contains an independent set of r vertices is less than $e^{-r/2}$.

Theorem (**The independence number of a random graph**). *Suppose the probability p satisfies*

$$p \geq \frac{2k\log k + 4k}{n}$$

and $r \geq n/k$. Then

$$\mathrm{Prob}\left(G_{n,p} \textit{ has an independent set of size at least } r\right) \leq e^{-r/2}.$$

Proof There are $\binom{n}{r}$ options for this independent set and for it to be independent we need all the $\binom{r}{2}$ possible edges to be missing. So the probability is at most

$$\begin{aligned}\binom{n}{r}(1-p)^{r(r-1)/2} &\leq \left(\frac{en}{r}\right)^r (1-p)^{r(r-1)/2} \\ &\leq (ek)^r (1-p)^{r(r-1)/2} \\ &= \left(e^2k^2(1-p)^{r-1}\right)^{r/2}.\end{aligned}$$

Now

$$\begin{aligned}e^2k^2(1-p)^{r-1} &\leq e^2k^2e^{-(r-1)p} \leq e^3k^2e^{-rp} \\ &\leq e^3k^2e^{-np/k} \leq e^3k^2e^{-2\log k-4} = e^{-1}\end{aligned}$$

which implies that the probability is at most $e^{-r/2}$. □

We now move on to the lemma about cycles.

Lemma (**Short cycles in a random graph**). *Let g and n be integers with,*

$$\frac{5}{n} \leq p \leq \frac{n^{1/g}}{n}$$

and let X be the number of cycles of length at most $g-1$ in $G_{n,p}$. Then $\mathrm{E}X \leq \frac{n}{4}$.

Proof The number of different cycles of length j is at most n^j and for a sequence of j vertices to be a cycle we need all j edges to be included so the expected number of cycles of length j is at most n^jp^j. Hence

$$\mathrm{E}X \leq \sum_3^{g-1}(np)^j \leq \frac{(np)^g-1}{np-1} \leq \frac{(np)^g}{4} \leq \frac{n}{4}.$$

□

We can combine these two estimates quite easily to deduce that there exist graphs with large chromatic number and no short cycles. We take a random graph $G_{n,p}$ with $p = (\log n)/n$. This will satisfy the hypothesis of the Independent Set Theorem if $2k\log k + 4k \leq \log n$ which is true if n is large and

$$k \leq \frac{\log n}{2\log\log n}.$$

It will satisfy the hypothesis of the cycle lemma if $n^{1/g} \geq \log n$ which says that

$$g \leq \frac{\log n}{\log \log n}.$$

If X is the number of short cycles then $\mathrm{E}X \leq n/4$ and so by Markov's Inequality

$$\mathrm{Prob}(X > n/2) \leq \frac{1}{2}.$$

So with high probability our random graph has no independent set larger than n/k and with probability $1/2$ it has only $n/2$ short cycles. So there is a graph having both these properties. If we remove one vertex from each short cycle we remove all the short cycles and don't make the independent sets any bigger. We thus obtain a graph with at least $n/2$ vertices, no cycles of length less than g and chromatic number at least

$$\frac{n/2}{n/k} = \frac{k}{2}.$$

Corollary (**Large chromatic number and large girth**). *For large n there is a graph of order at most n with chromatic number at least*

$$\frac{\log n}{4 \log \log n} - 1$$

and no cycles shorter than

$$\frac{\log n}{\log \log n} - 1.$$

Now let's move on to the connectedness.

Connectedness

Theorem (**Connectedness of random graphs**). *Let*

$$p = p(n) = \frac{\log n}{n} + \frac{c_n}{n}.$$

Then

$$\mathrm{Prob}\left(G_{n,p} \text{ is connected}\right) \to \begin{cases} 0 & \text{if } c_n \to -\infty \\ 1 & \text{if } c_n \to \infty \end{cases}$$

Proof The case $c_n \to \infty$ is a straightforward calculation even though it is slightly messy. In order for $G_{n,p}$ to be disconnected there must be a number $k \leq n/2$ and a partition of the vertices into two sets of size k and $n-k$ with no edges between them. For each k the number of such partitions is $\binom{n}{k}$ and the number of edges between the two halves is $k(n-k)$. So the probability of such a partition is at most

$$\sum_{k=1}^{n/2} \binom{n}{k} (1-p)^{k(n-k)} \leq \sum_{k=1}^{n/2} \left(\frac{en}{k}\right)^k (1-p)^{k(n-k)}.$$

Now

$$\frac{en}{k}(1-p)^{n-k} \leq \frac{en}{k} e^{-p(n-k)}.$$

By differentiating with respect to k it is easy to check that this expression decreases until $k = 1/p$ and then increases until $k = n/2$. So it is never more than the larger of

$$ene^{-p(n-1)} \quad \text{and} \quad 2ee^{-pn/2}.$$

The first is $e^{1+p}ne^{-\log n - c_n} \leq e^2 e^{-c_n}$ and the second is $2ee^{-(\log n + c_n)/2}$. Both expressions are at most e^{-d_n} for some sequence $d_n \to \infty$ and so the probability of disconnectedness is at most

$$\sum_{k=1}^{n/2} (e^{-d_n})^k$$

and for large n this is at most

$$2e^{-d_n} \to 0.$$

For $c_n \to -\infty$ we will show that the graph is likely to contain an isolated vertex. Let X be the number of isolated vertices. We know that

$$\begin{aligned} \mathrm{E}X &= n(1-p)^{n-1} \geq ne^{-p(n-1)/(1-p)} \\ &= e^{p/(1-p)} ne^{-(\log n + c_n)/(1-p)} = e^{p/(1-p)} n^{-p/(1-p)} e^{-c_n/(1-p)} \end{aligned}$$

Since $p = p_n \leq \frac{\log n}{n} \to 0$ and $c_n \to -\infty$ the first factor converges to 1 while the last tends to ∞ with n. The middle factor is the reciprocal of

$$\exp\left(\frac{p \log n}{1-p}\right) \leq \exp\left(\frac{(\log n)^2}{n - \log n}\right) \to 1.$$

Hence

$$\mathrm{E}X \to \infty$$

as $n \to \infty$.

If we knew that X had a Poisson distribution we could conclude that $\text{Prob}(X = 0)$ is small. We don't have a simple way to describe the distribution of X but we can investigate its moments: in particular $\text{E}X^2$. The variance of a Poisson random variable is equal to its expectation: we shall show something similar for X.

For each k let I_k be 1 or 0 according to whether the vertex k is isolated. Then $X = \sum_1^n I_k$.

$$\text{E}X^2 = \text{E}\sum_{jk} I_j I_k = \sum_j \text{E}I_j^2 + \sum_{j\neq k} \text{E}I_j I_k = \sum_j \text{E}I_j + \sum_{j\neq k} \text{E}I_j I_k.$$

We already know that for each j, $\text{E}I_j = (1-p)^{n-1}$. For $j \neq k$ the product $I_j I_k$ is 1 only if the two vertices have no edges incident to them. There are $2n-3$ possible edges since each vertex has $n-1$ possible edges but one edge is shared. So

$$\text{E}I_j I_k = (1-p)^{2n-3}.$$

Hence

$$\text{E}X^2 = n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3} = \text{E}X + \frac{n-1}{n(1-p)}(\text{E}X)^2.$$

Therefore the variance of X is

$$\begin{aligned} \text{var}(X) &= \text{E}X^2 - (\text{E}X)^2 = \text{E}X + \left(\frac{n-1}{n(1-p)} - 1\right)(\text{E}X)^2 \\ &= \text{E}X + \frac{np-1}{n(1-p)}(\text{E}X)^2 \leq \text{E}X + \frac{p}{1-p}(EX)^2. \end{aligned}$$

Since $p \leq (\log n)/n$ is small this is roughly the expectation of X as we had hoped. Now

$$\text{var}(X) = \text{E}(X - \text{E}X)^2 \geq \text{Prob}(X = 0)(-\text{E}X)^2 = \text{Prob}(X = 0)(\text{E}X)^2.$$

So

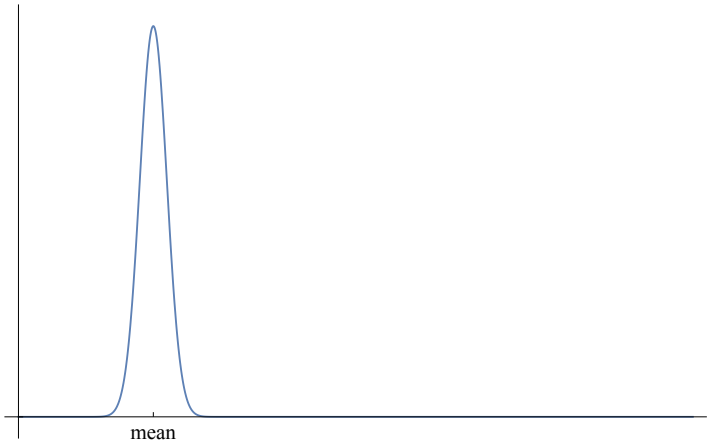
$$\text{Prob}(X = 0) \leq \frac{\text{var}(X)}{(\text{E}X)^2} \leq \frac{1}{\text{E}X} + \frac{p}{1-p} \to 0.$$

□

The estimate we got for $\text{Prob}(X = 0)$ is roughly $1/\text{E}X$ and is much weaker than $e^{-\text{E}X}$ which is what we expect from the Poisson distribution. In order to get a better estimate we would have to consider higher moments.

To finish this chapter we shall make some remarks concerning concentration of the number of edges which will be of interest when we look at regularity methods. Let us look at a random graph from our model and two disjoint sets of vertices with k and m members. The number of possible edges between them is km. The number selected is Binomial $B(km, p)$. So its mean is kmp and variance is $kmp(1-p)$ which is certainly not more than kmp.

Now suppose that our vertex sets are of reasonable size: say at least $n/10$ and p is not too small: say at least $1/(10n)$. Then the mean is at least $n/10^3$ and the variance is no larger. So the distribution looks like this.



In a random graph we can predict with considerable accuracy how many edges will cross from one set of vertices to another as long as those sets are fairly large.

Chapter 8. The regularity method

In 1975 Szemerédi proved that if a set of positive integers S has the property that

$$\limsup \frac{|S \cap \{1, 2, \ldots, n\}|}{n} > 0$$

then S contains arbitrarily long arithmetic progressions:

$$(a, a + b, a + 2b, \ldots, a + kb).$$

Thus if from time to time S contains a fixed proportion of the numbers between 1 and n then it contains long arithmetic progressions. In order to do this he developed a remarkable structure theorem for graphs (and hypergraphs): the Regularity Theorem. The theorem states (roughly) that every graph can be decomposed into a small number of pieces where the graph looks random and a small extra bit.

For sets $A, B \subset G$ we define the edge *density* between A and B to be

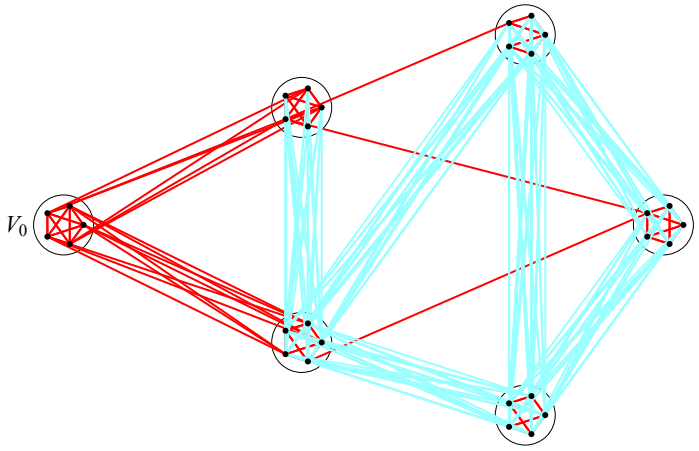
$$d(A, B) = \frac{\text{The number of edges from } A \text{ to } B}{|A|.|B|}.$$

We say that the pair (X, Y) of subsets of G is ε-regular if for all $A \subset X$ and $B \subset Y$ with $|A| \geq \varepsilon|X|$ and $|B| \geq \varepsilon|Y|$ we have

$$|d(A, B) - d(X, Y)| \leq \varepsilon.$$

This says that the edges between X and Y appear to have been chosen randomly with probability $p = d(X, Y)$. (Compare with the statement made at the end of the random graph chapter.)

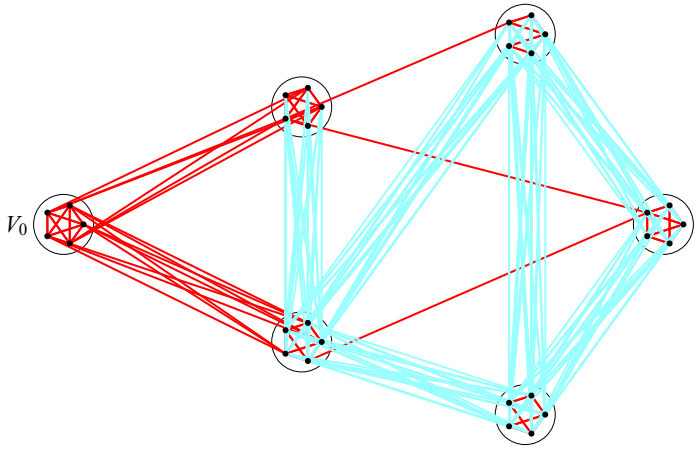
Theorem (**Szemerédi's Regularity Lemma**)**.** *For every $\varepsilon > 0$ and $m \in \mathbf{N}$ there are $M, N \in \mathbf{N}$ so that if G is a graph of order $n > N$ then we can partition the vertices of G as $V_0 \cup V_1 \cup \ldots \cup V_k$ where $|V_1| = |V_2| = \cdots = |V_k|$, $m < k < M$, $|V_0| < \varepsilon n$ and all but εk^2 of the pairs (V_i, V_j) are ε-regular.*



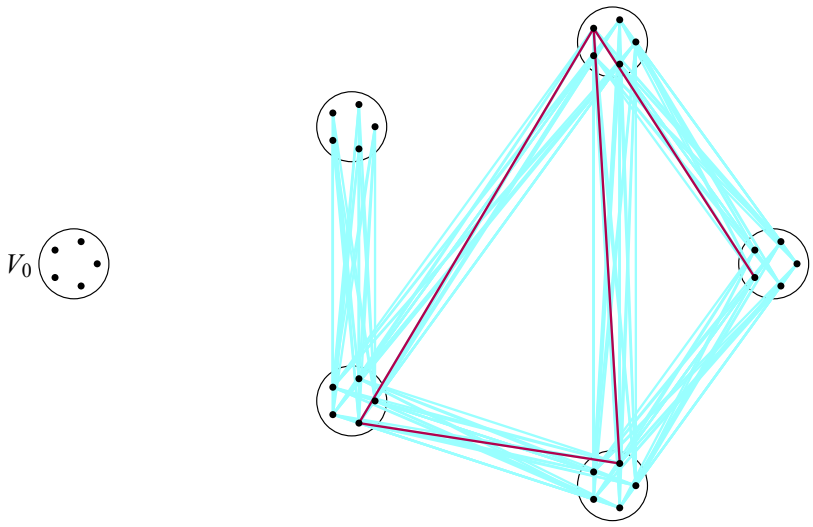
The Regularity Lemma is most often used via one of its consequences called the Graph Removal Lemma.

Theorem (**Graph Removal Lemma**). *For every $\varepsilon > 0$, every k and every graph H of order k, there is a $\delta > 0$ so that if G is a graph of order n, either we can remove εn^2 edges from G so as to leave no copy of H or G contains δn^k copies of H.*

Either we can remove a small proportion of the edges of G so as to eliminate all H subgraphs, or G contains about as many copies of H as it possibly could: $\binom{n}{k}$. How do we deduce it?



Partition G into about $1/\varepsilon$ pieces. Throw out all the edges inside the V_i, all the edges from V_0, all the edges in the irregular pairs and all the edges in sparse pairs. If we have a copy of H remaining it is formed by turquoise edges; edges between pairs (V_i, V_j) that are regular and pretty dense.



You have a copy of H using edges between pairs (V_i, V_j) that are regular and pretty dense. If you choose vertices from the pieces more or less randomly there is a good chance they will form a copy of H. So there will be many ways to do it.

To finish this short chapter we will deduce Roth's Theorem on arithmetic progressions.

Theorem (**Roth**). *If a set of positive integers S has the property that*

$$\limsup \frac{|S \cap \{1, 2, \ldots, n\}|}{n} > 0$$

then S contains arithmetic progressions of length 3:

$$(a, a + b, a + 2b).$$

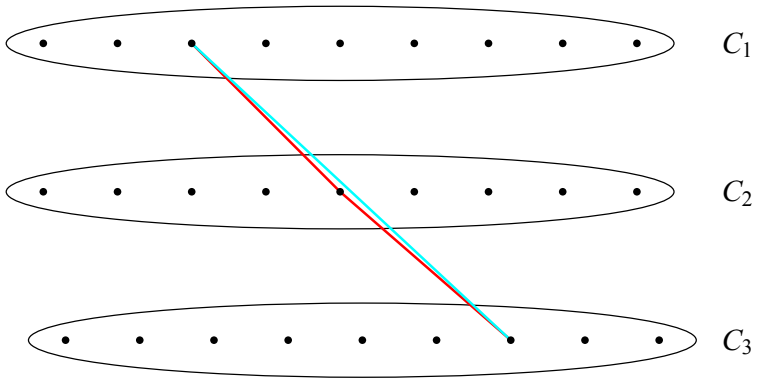
The theorem automatically implies that S must contain infinitely many such progressions.

Proof Choose n large for which

$$\frac{|S \cap \{1, 2, \ldots, n\}|}{n} > \varepsilon$$

and form a graph as follows. The vertices belong to 3 classes C_1, C_2 and C_3 each of which has $3n$ vertices labelled with the integers from 1 to $3n$. Edges will go only between vertex classes. Let $S_n = S \cap \{1, 2, \ldots, n\}$. An edge (u, v) between C_1 and C_2 will be included if $v - u \in S_n$ and similarly between C_2 and C_3. An edge (u, w) between C_1 and C_3 will be included if $(w - u)/2 \in S_n$.

We now look at triangles in G. If the triangle consists of $u \in C_1$, $v \in C_2$ and $w \in C_3$ with $v - u = w - v$ then we call the triangle trivial. The triangle looks like a line in a picture of G.



For each $a \in S_n$ we can find at least n trivial triangles

$$u, \quad u + a, \quad u + 2a$$

as u runs from 1 to n. These triangles are edge disjoint so to remove them all we would need to remove more than $\varepsilon n.n = \varepsilon n^2$ edges. By the removal theorem the graph must contain at least δn^3 triangles. But there can't be more than $3n^2$ trivial triangles so there must be a non-trivial one:

$$u \quad v \quad w$$

with $v - u$, $w - v$ and $(w - u)/2 \in S_n$ and not all the same. Let $a = v - u$, $b = w - v$ and $c = (w - u)/2 = (a + b)/2$. □

HW08
Show that for every ε if d is large enough, every subset A of $\mathbf{Z}_3^d$ with at least $\varepsilon 3^d$ elements contains distinct vectors x, y and z such that $x + y + z = 0$.