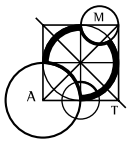
PROBLEM SOLVING TACTICS

A DI PASQUALE, N DO & D MATHEWS

Published by



AMT Publishing

Australian Mathematics Trust
University of Canberra Locked Bag 1
Canberra GPO ACT 2601
AUSTRALIA

Telephone: +61 2 6201 5137
www.amt.edu.au

AMTT Limited ACN 083 950 341

National Library of Australia Card Number and ISSN
Australian Mathematics Trust Enrichment Series ISSN 1326-0170

Problem Solving Tactics
ISBN 978-1-876420-75-8

THE AUSTRALIAN MATHEMATICS TRUST

ENRICHMENT SERIES

The books in this series are selected for their motivating, interesting and stimulating sets of quality problems, with a lucid expository style in their solutions. Typically, the problems have occurred in either national or international contests at the secondary school level.

They are intended to be sufficiently detailed at an elementary level for the mathematically inclined or interested to understand but, at the same time, be interesting and sometimes challenging to the undergraduate and the more advanced mathematician. It is believed that these mathematics competition problems are a positive influence on the learning and enrichment of mathematics.

THE AUSTRALIAN MATHEMATICS TRUST

ENRICHMENT SERIES

BOOKS IN THE SERIES

1 AUSTRALIAN MATHEMATICS COMPETITION 1978–1984 BOOK 1
WJ Atkins, JD Edwards, DJ King, PJ O'Halloran & PJ Taylor

2 MATHEMATICAL TOOLCHEST
AW Plank & NH Williams

3 INTERNATIONAL MATHEMATICS TOURNAMENT OF TOWNS 1984–1989 BOOK 2
PJ Taylor

4 AUSTRALIAN MATHEMATICS COMPETITION 1985–1991 BOOK 2
PJ O'Halloran, G Pollard & PJ Taylor

5 PROBLEM SOLVING VIA THE AMC
WJ Atkins

6 INTERNATIONAL MATHEMATICS TOURNAMENT OF TOWNS 1980–1984 BOOK 1
PJ Taylor

7 INTERNATIONAL MATHEMATICS TOURNAMENT OF TOWNS 1989–1993 BOOK 3
PJ Taylor

8 ASIAN PACIFIC MATHEMATICS OLYMPIAD 1989–2000
H Lausch & C Bosch Giral

9 METHODS OF PROBLEM SOLVING BOOK 1
JB Tabov & PJ Taylor

10 CHALLENGE! 1991–1998 BOOK 1
JB Henry, J Dowsey, AR Edwards, LJ Mottershead,
A Nakos, G Vardaro & PJ Taylor

11 USSR MATHEMATICAL OLYMPIADS 1989–1992
AM Slinko

12 AUSTRALIAN MATHEMATICAL OLYMPIADS 1979–1995 BOOK 1
H Lausch & PJ Taylor

13 CHINESE MATHEMATICS COMPETITIONS AND OLYMPIADS 1981–1993
A Liu

14 POLISH & AUSTRIAN MATHEMATICAL OLYMPIADS 1981–1995
ME Kuczma & E Windischbacher

15 INTERNATIONAL MATHEMATICS TOURNAMENT OF TOWNS 1993–1997 BOOK 4
PJ Taylor & AM Storozhev

16 Australian Mathematics Competition 1992–1998 Book 3
WJ Atkins, JE Munro & PJ Taylor

17 Seeking Solutions
JC Burns

18 101 Problems in Algebra
T Andreescu & Z Feng

19 Methods of Problem Solving Book 2
JB Tabov & PJ Taylor

20 Hungary-Israel Mathematics Competition: The First Twelve Years
S Gueron

21 Bulgarian Mathematics Competition 1992–2001
BJ Lazarov, JB Tabov, PJ Taylor & A Storozhev

22 Chinese Mathematics Competitions and Olympiads 1993–2001 Book 2
A Liu

23 International Mathematics Tournament of Towns 1997–2002 Book 5
AM Storozhev

24 Australian Mathematics Competition 1999–2005 Book 4
WJ Atkins & PJ Taylor

25 Challenge! 1999–2006 Book 2
JB Henry & PJ Taylor

26 International Mathematics Tournament of Towns 2002–2007 Book 6
A Liu & PJ Taylor

27 International Mathematical Talent Search Part 1
G Berzsenyi

28 International Mathematical Talent Search Part 2
G Berzsenyi

29 Australian Mathematical Olympiads 1996–2011 Book 2
H Lausch, A Di Pasquale, DC Hunt & PJ Taylor

30 Methods of Problem Solving Book 3
JB Tabov, EM Kolev & PJ Taylor

31 Australian Mathematics Competition 2006–2012 Book 5
WJ Atkins & PJ Taylor

32 Australian Intermediate Mathematics Olympiads 1999–2013
JB Henry & KL McAvaney

33 Problem Solving Tactics
A Di Pasquale, N Do & D Mathews

About this book

What is this book about?

Each year the Australian Mathematical Olympiad Committee (AMOC) runs two training schools. These are designed to extend and challenge the mathematical skills of the 25 secondary school students who are invited to attend. Particular emphasis is given to honing the skill of problem solving.

This book is based on past and present lectures given at the two annual AMOC training schools. As such it is suitable for

- anyone who wishes to qualify for an Olympiad training school in mathematics, either in Australia or overseas
- anyone who has attended an Olympiad training school in mathematics and who would like to be better prepared should they qualify again for an invitation
- interested students, teachers and parents, as it will give an idea of the sorts of mathematics considered there
- any mathematically able students, hobbyists or problem solvers, whether local or abroad, who would find this publication enriching.

What is in this book?

The authors have gone to considerable care to showcase many of the tricks and problem-solving tactics they consider to be important for Olympiad mathematics and problem solving in general. Apart from the first chapter the topics are grouped into the four broad traditional Olympiad divisions of number theory, geometry, algebra and combinatorics.

Most of the sections within each main chapter highlight a particular idea important for problem solving, thus providing over 150 such ideas in total. Each idea is illustrated with one or two problems along with solutions. For extra practice, most chapters begin with a list of problems. Although they are not necessarily in order of difficulty, we have tried to arrange them so that the first few problems tend to be easier than the later ones.

Mathematics can be quite hard to read and digest and so the style has purposely been kept rather informal and conversational. The book often gives the impression that it is conversing with the reader.

How do you use this book?

Most chapters do not depend much on other chapters. Therefore, apart from the first chapter, most can be studied almost independently of each other. However, where dependencies arise there are cross references.

It is the opinion of the authors and many others involved in AMOC training schools that the chief way to improve one's problem-solving ability is to go through the struggle of trying to solve problems oneself. So we recommend that:

> The focus of the user of this book should not be on reading solutions but on *trying to solve problems*.

That is why solutions are not provided to the problems at the beginning of each chapter. It is also why we recommend that a problem be tried thoroughly with the showcased idea of the section in mind, before the solution is studied.

We recognise that some problems are relatively easy exercises while others are of the difficulty of the International Mathematical Olympiad—the pinnacle of problem-solving mathematics for high school students the world over. So the reader definitely should not expect to be able to solve all of the problems straight away.

Acknowledgments

Some problems are the inventions of staff members at AMOC training schools. However, many of the problems have come from contests such as the Australian Mathematical Olympiad (AMO), national mathematical Olympiads of some other countries, the Asian Pacific Mathematics Olympiad (APMO), the International Mathematical Olympiad (IMO) and problems shortlisted for the IMO. Since many of these problems have appeared in multiple contests, in many cases it has been hard to identify their true origin and so we would simply like to acknowledge all of the above sources.

Although this book has three listed authors, the ideas it contains are the product of many AMOC staff interacting with each other and with students over many years.

We also express our appreciation to Ross Atkins, Andrew Elvey Price, Ivan Guo, Konrad Pilch, Chaitanya Rao, Sally Tsang, Graham White and Sampson Wong, who assisted in proofreading for mathematical content and accuracy, and provided other feedback.

The document was typeset using LaTeX. The drawings were done with the help of GeoGebra and Ti*k*Z.

About the authors

Angelo Di Pasquale was twice a contestant at the International Mathematical Olympiad. He completed a PhD in mathematics at The University of Melbourne in 1999. His research combined topology, combinatorics and algebra to better understand the relationship between algebraic curves and their complements. Since 2000 he has been Director of Training for the Australian Mathematical Olympiad Committee, and since 2002 he has been Australian Team Leader at the International Mathematical Olympiad. He enjoys composing Olympiad problems for Australian and international mathematics contests.

Norman Do represented Australia at the 1997 International Mathematical Olympiad and was Deputy Leader for the Australian team on four occasions. He obtained a PhD in mathematics from The University of Melbourne in 2010. He has worked at McGill University, Quebec, and is now a lecturer at Monash University, where he researches problems that combine geometry, topology, combinatorics, and mathematical physics. He is currently the Chair of the Australian Mathematical Olympiad Committee's Senior Problems Committee, which sets national mathematics Olympiad papers and proposes problems for international competitions.

Daniel Mathews was twice a contestant at the International Mathematical Olympiad, and three times Deputy Leader of the Australian team. He studied mathematics, law and languages at The University of Melbourne before obtaining a PhD in mathematics from Stanford University, California, in 2009. He has worked as a mathematician in France and the US and is currently a lecturer in the School of Mathematical Sciences at Monash University. His mathematical research studies the relationship between geometry, topology, and mathematical physics.

Contents

About this book i

1 Methods of proof 1

1.0 Problems . . . 1
1.1 Logic and deduction . . . 5
1.2 Converse . . . 5
1.3 If and only if . . . 6
1.4 Contrapositive . . . 7
1.5 Proof by contradiction . . . 8
1.6 Proof by induction . . . 9
1.7 Strong induction . . . 11
1.8 Proof by exhaustion (case bashing) . . . 12
1.9 Pigeonhole principle . . . 13
1.10 Advanced pigeonhole principle . . . 15
1.11 Extremal principle . . . 16
1.12 Telescoping . . . 18

2 Number theory 19

2.0 Problems . . . 19
2.1 Fundamental theorem of arithmetic . . . 24
2.2 Pigeonhole principle . . . 25
2.3 Dealing with digits . . . 25
2.4 Floor function . . . 26
2.5 Square roots and conjugates . . . 27
2.6 Powers of two . . . 28
2.7 Euclid's algorithm . . . 29
2.8 Integers base-n . . . 31
2.9 Construction problems . . . 32

2.10 Modular arithmetic . . . 34
2.11 Chinese remainder theorem . . . 34
2.12 From Fermat to Euler . . . 35
2.13 The gcd trick . . . 37
2.14 Existence of a generator . . . 37

3 Diophantine equations **39**
3.0 Problems . . . 39
3.1 Factorisation . . . 43
3.2 Monotonicity . . . 43
3.3 Bounding arguments . . . 44
3.4 Polynomial modulus . . . 45
3.5 Quadratic discriminants . . . 46
3.6 Modular arithmetic . . . 46
3.7 Divisibility and gcds . . . 47
3.8 Reduction of variables . . . 48
3.9 Infinite descent . . . 49
3.10 Vieta jumping . . . 50
3.11 Cyclotomic recognition . . . 51

4 Plane geometry **53**
4.0 Problems . . . 54
4.1 Angle chasing . . . 59
4.2 Cyclic quadrilaterals . . . 59
4.3 One step at a time . . . 60
4.4 Triangle centres . . . 62
4.5 Constructions . . . 64
4.6 Exploit symmetry . . . 66
4.7 Extend to the circumcircle . . . 66
4.8 Reverse reconstruction . . . 68
4.9 Trigonometry . . . 69
4.10 Areas . . . 70
4.11 Relate to known diagrams . . . 72
4.12 Create beautiful pictures . . . 73

5 Important configurations in geometry **75**
5.1 **A-List: extremely useful** . . . 76
A1 Angle bisector and perpendicular bisector . . . 76
A2 Pivot theorem . . . 77
A3 Radical axis theorem . . . 78
A4 Similar switch . . . 79
5.2 **B-List: very useful** . . . 80
B1 Radical axis bisects common tangent . . . 80
B2 Perpendicularity . . . 81

B3 Alternate segment switch . . . 82
B4 Ratios for collinearity . . . 83
B5 Points of contact of incircle and excircle . . . 84
B6 Circumcircle, incentre and excentre . . . 85
B7 Simson line . . . 86
B8 Pascal's theorem . . . 87
B9 Desargues' theorem . . . 88
B10 Quadrilateral and incircle . . . 89
5.3 **C-List: useful** . . . 90
C1 Nine-point circle . . . 90
C2 Euler line . . . 91
C3 Four lines and four circles . . . 92
C4 Newton–Gauss line . . . 93
C5 Alternative characterisation of symmedian . . . 94
C6 Convex cyclic hexagon and diagonals . . . 95
C7 Quadrilateral, triangles and incircles . . . 96
C8 Median, inradius and chord of incircle . . . 97
C9 Incentre and midpoints . . . 98
C10 Incentre, excentre, midpoint and contact points . . . 99
C11 Incentre and chord of incircle . . . 100
C12 Harmonic quadrilateral . . . 101
C13 Incentre and mixtilinear incircle . . . 102
C14 Butterfly theorem . . . 103

6 Incidence geometry **105**
6.0 Problems . . . 105
6.1 Collinear points . . . 108
6.2 Menelaus' theorem . . . 109
6.3 Concurrent lines . . . 110
6.4 Ceva's theorem . . . 111
6.5 Concyclic points . . . 113
6.6 Power of a point . . . 117
6.7 Radical axes . . . 118
6.8 Ellipses . . . 119
6.9 Pascal's theorem . . . 122

7 Transformation geometry **125**
7.0 Problems . . . 126
7.1 Translations . . . 130
7.2 Rotations . . . 130
7.3 Dilations . . . 131
7.4 Spiral symmetries . . . 133
7.5 Affine transformations . . . 134

8 Complex numbers **137**
8.0 Problems . . . 137
8.1 Addition ideas . . . 140
8.2 Angles . . . 141
8.3 Multiplication ideas . . . 142
8.4 Similarity ideas . . . 144
8.5 Roots of unity . . . 146

9 Polynomials **149**
9.0 Problems . . . 150
9.1 Identity theorem . . . 154
9.2 Division algorithm . . . 154
9.3 Fundamental theorem of algebra . . . 155
9.4 Vieta's formulas . . . 156
9.5 Integer polynomials . . . 158
9.6 Complex numbers . . . 160
9.7 Algebraic trickery . . . 160
9.8 Irreducibility . . . 162
9.9 Factorisation . . . 163
9.10 Polynomials modulo p (upstairs–downstairs) . . . 163
9.11 Polynomials modulo $P(x)$ . . . 166
9.12 Lagrange interpolation . . . 166
9.13 Root focus . . . 167

10 Functional equations **169**
10.0 Problems . . . 169
10.1 Cauchy's functional equation . . . 173
10.2 Guess and hope . . . 174
10.3 Substitutions . . . 175
10.4 Injective, surjective and bijective . . . 176
10.5 The associative trick . . . 178
10.6 Exploit symmetry . . . 179
10.7 Involutions . . . 179
10.8 Fixed points . . . 180
10.9 Somewhere versus everywhere . . . 182
10.10 Completely multiplicative functions . . . 183
10.11 Well-ordering of $\mathbb{N}^+$ . . . 184

11 Inequalities **187**
11.0 Problems . . . 187
11.1 Squares are non-negative . . . 190
11.2 AM–GM inequality . . . 191
11.3 Rearrangement inequality . . . 191
11.4 Cauchy–Schwarz inequality . . . 193

11.5 Power means inequality . . . 195
11.6 Jensen's inequality . . . 196
11.7 Substitutions . . . 197
11.8 Addition and multiplication of inequalities . . . 198
11.9 Expand and conquer . . . 200
11.10 Homogeneous inequalities . . . 201
11.11 Muirhead's inequality . . . 202
11.12 Weighted inequalities . . . 203

12 Geometric inequalities **207**
12.0 Problems . . . 207
12.1 Triangle inequality . . . 210
12.2 Reflection principle . . . 210
12.3 Transformations . . . 213
12.4 Trigonometry . . . 214
12.5 Parametrisation . . . 215
12.6 Ptolemy's inequality . . . 216
12.7 Locus and tangency . . . 219
12.8 Isoperimetric inequalities . . . 221
12.9 Incircle substitution . . . 222
12.10 Triangle formulas . . . 222

13 Combinatorics **225**
13.0 Problems . . . 225
13.1 Addition and multiplication . . . 228
13.2 Subtraction and division . . . 228
13.3 Binomial identities . . . 229
13.4 Bijections . . . 231
13.5 The supermarket principle . . . 232
13.6 Pigeonhole principle . . . 233
13.7 Principle of inclusion–exclusion . . . 234
13.8 Double counting . . . 235
13.9 Injections . . . 237
13.10 Recursion . . . 238
13.11 Double counting via tables . . . 240
13.12 Combinatorial reciprocal principle . . . 241

14 Graph theory **243**
14.0 Problems . . . 244
14.1 Degree . . . 247
14.2 Directed graphs . . . 248
14.3 Connected graphs, cycles and trees . . . 248
14.4 Complete graphs and bipartite graphs . . . 250
14.5 Pigeonhole principle . . . 250

14.6 Euler trails . . . 251
14.7 Paths . . . 253
14.8 Extremal principle . . . 254
14.9 Count and count again . . . 255
14.10 Planar graphs . . . 256
14.11 Polyhedra . . . 257
14.12 Graph theory and inequalities . . . 258

15 Games and invariants **261**
15.0 Problems . . . 261
15.1 Number invariants . . . 267
15.2 Parity . . . 268
15.3 Modular arithmetic invariants . . . 269
15.4 Colouring invariants . . . 270
15.5 Monovariants . . . 272
15.6 Invariants as cost . . . 273
15.7 Permutation parity . . . 274
15.8 Combinatorial games . . . 275
15.9 Position analysis . . . 276
15.10 The copycat strategy . . . 277
15.11 Pairing strategies . . . 277
15.12 Strategy stealing . . . 278

16 Combinatorial geometry **281**
16.0 Problems . . . 281
16.1 Proof by contradiction . . . 285
16.2 Extremal principle . . . 285
16.3 Perturbation . . . 287
16.4 Induction . . . 288
16.5 Discrete intermediate value theorem . . . 288
16.6 Convex hull . . . 289
16.7 Euler's formula . . . 290
16.8 Pigeonhole principle . . . 291
16.9 Colouring . . . 292

17 Appendices **293**
17.1 How do complex numbers work? . . . 293
17.2 Function notation . . . 296
17.3 Directed angles . . . 297
17.4 Some useful triangle formulas . . . 298

Index **299**

1

Methods of proof

Proofs are the essence of mathematics. They are nothing more than logical arguments which present the solution to a mathematical problem beyond all possible doubt. There is no set format for a proof, no particular way in which it must be presented on a page. Some people may try to tell you that proofs must appear in two columns, with statements on the left and explanations on the right—but this is complete nonsense! As long as you have sufficiently clear assumptions, a logical argument that is fully explained, and the correct conclusion, then you have a proof.[1]

So how do you know whether or not you have a rigorous proof to a mathematical problem? Well, this is a difficult question to answer. Being able to write correct proofs without leaving out important details is a skill which can only be learned through experience—that is, by reading and writing them yourself. However, the following little test might help you on your way. Imagine that you have an inquisitive younger sibling who is looking over your shoulder as you write your proof and constantly interrogating you. 'What does this mean?' she might ask, or 'Why are you doing that?' or 'How does that make sense?' If you claim that something is obvious, she will ask why and if you are tempted to be vague, she just won't understand. Your goal, of course, is to make sure that she understands and to answer all of her questions before she has even asked them. In order to accomplish this, you should provide a clear explanation for every single fact you write down that is not blatantly obvious.

Apart from the logical aspect of your proof, there are a few other points to keep in mind. Proofs should always be clear, concise, and most definitely legible. Keep in mind that a nice proof to a problem is usually shorter than a messy one. This is often achieved by using good notation and providing the right definitions. A common problem is for people to provide extraneous material, which does not contribute to the main argument, or to write long essays, which are not enlightening at all. One way to avoid these pitfalls is to break a problem into smaller chunks, which can be individually proved and then reassembled to provide the complete proof. In general, a good proof should take the reader on a pleasant mathematical journey ending at the desired result.

1.0 Problems

1. Prove that if $a + b$ is an irrational number, then at least one of a or b is irrational.

[1] There is a rigorous notion of what constitutes a mathematical proof, but for our purposes and for the purposes of most modern mathematics, our informal explanation is sufficient.

2. Show that at any party, there are always at least two people with exactly the same number of friends at the party.

3. The *equal temperament tuning*[2] of musical instruments is based on the fact that $2^{\frac{19}{12}}$ is very close to 3.

   Show that there can be no *perfect tuning*[3] by proving that if $2^x = 3$, then x must be irrational.

4. If m and n are positive integers, prove that $\sqrt[m]{n}$ is either a positive integer or irrational.

5. Prove that there are infinitely many prime numbers of the form $6n + 5$, where n is a positive integer.

6. For every positive integer n, prove that

$$\frac{1}{1 \times 2} + \frac{1}{2 \times 3} + \frac{1}{3 \times 4} + \cdots + \frac{1}{(n-1) \times n} = \frac{n-1}{n}.$$

7. Prove that $n^2 < 2^n$ for every integer $n \geq 5$.

8. For every positive integer n, prove that

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

9. For every positive integer n, prove that

$$1^3 + 2^3 + 3^3 + \cdots + n^3 = \frac{n^2(n+1)^2}{4},$$

   and go on to conclude that

$$1^3 + 2^3 + 3^3 + \cdots + n^3 = (1 + 2 + 3 + \cdots + n)^2.$$

10. Recall that the Fibonacci sequence is defined recursively by $F_1 = 1$, $F_2 = 1$ and $F_{n+1} = F_n + F_{n-1}$, for $n \geq 2$.

    Prove the following identity for Fibonacci numbers: for all $n \geq 1$,

$$F_1^2 + F_2^2 + \cdots + F_n^2 = F_n F_{n+1}.$$

11. Prove that every positive integer can be uniquely expressed as a sum of different numbers, where each number is of the form 2^n for some non-negative integer n.

12. Suppose x is a real number such that $x + \frac{1}{x}$ is an integer.

    Prove that $x^n + \frac{1}{x^n}$ is also an integer for any positive integer n.

13. Any finite collection of lines in the plane divides the plane up into regions.

    Prove that it is possible to colour each of these regions either black or white in such a way that no two regions which share a common edge have the same colour.

[2] This tuning has perfect octaves and almost perfect fifths. Octaves are tuned so that the ratio of the frequency of the pitch of the higher note to that of the note an octave lower is $2:1$. Perfect fifths are tuned so that the ratio of the frequency of the pitch of the higher note to the note a fifth below is $3:2$. Standard equal temperament tuning says that rising by an almost perfect fifth 12 times should be the same as rising by a perfect octave seven times. Thus $\left(\frac{3}{2}\right)^{12} \approx 2^7$.

[3] By perfect tuning, we mean that all octaves and fifths are perfect.

14. Show that if there are five points in a square with side length 1 metre, then there exist two of them which are less than 75 centimetres apart.

15. Four points are given inside a square with side length 8 metres.

    (a) Prove that two of them are less than $\sqrt{65}$ metres apart.

    (b) Can you prove, beyond a shadow of a doubt, that two of them are less than 8 metres apart?

16. (a) Prove that if x and y can each be written as the sum of the squares of two integers, then so can xy.

    (b) Prove that if x and y are both of the form $a^2 + 2b^2$ $(a, b \in \mathbb{Z})$, then so is xy.

    (c) Let k be a fixed integer. Prove that if x and y are both of the form $a^2 + kb^2$ $(a, b \in \mathbb{Z})$, then so is xy.

17. Consider the non-empty subsets of $\{1, 2, \ldots, n\}$. For each of these subsets, consider the reciprocal of the product of its elements.

    Determine the sum of all of these numbers.

18. A finite set of chords is drawn in a circle such that each of them passes through the midpoint of another chord.

    Prove that all of the chords must be diameters.

19. Show that if we take $n + 1$ numbers from the set $\{1, 2, 3, \ldots, 2n\}$, there must exist two which have no common factor greater than 1.

    Does this remain true if we take n numbers?

20. A circular island is divided into states by a number of chords of the circle. Consider a tour that starts and ends in the same state without passing through the intersection of any two borders.

    Prove that the tour must involve an even number of border crossings.

21. Suppose you are given a balance scale and a collection of weights whose masses are $1, 3, 3^2, 3^3, \ldots$.

    (a) Prove that using these masses you can determine the weight of any object whose mass is a positive integer.

    (b) Prove that apart from interchanging the contents of the left and right pans of the scale, the configuration of masses on the pans that correctly determines the weight is unique.

    (c) Prove that the weights $1, 3, 3^2, 3^3, \ldots$ are the only integral weights that uniquely determine the weight of every integral mass.

22. A group of people played in a tennis tournament where each person played exactly one match against every other person.

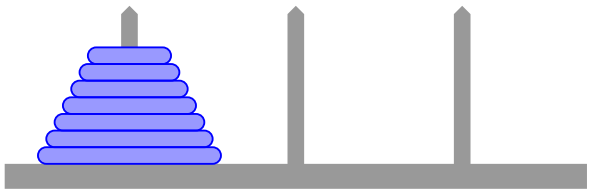
    Prove that it is always possible to put the players in a line so that the first player beat the second, the second player beat the third, all the way down to the last player.

23. A polygon is divided into triangles by diagonals whose endpoints are the vertices of the polygon in such a way that no two of the diagonals intersect inside the polygon.

    Prove that it is possible to colour the vertices of the polygon with three colours so that the three vertices of each triangle have different colours.

24. The *Tower of Hanoi* is a mathematical puzzle consisting of three rods and n discs of distinct sizes, which can slide onto any of the three rods. The puzzle starts with the discs neatly stacked in order of size on one rod, the smallest at the top, as shown in the diagram below. The aim is to transfer the entire stack to another rod by moving a disc from the top of one stack to the top of another stack in such a way that no disc is placed on top of a smaller disc.

    Prove that the task can be accomplished in $2^n - 1$ moves but not in fewer moves.



25. Thirty coins lie on a table, with 17 of them showing heads. Your task, should you choose to accept it, is to separate the coins into two piles, not necessarily of the same size, each of which has the same number of heads showing. Unfortunately, you happen to be blindfolded and cannot feel the difference between the two sides of a coin.

    How can you perform the task?

26. Each of the numbers $1, 2, 3, \ldots, n^2$ is written in one of the squares of an $n \times n$ chessboard.

    Show that there exist two squares which share a vertex or an edge whose entries differ by at least $n + 1$.

27. Each square of an 8×8 chessboard has a real number written in it in such a way that each number is equal to the geometric mean[4] of all the numbers a knight's move away from it.

    Is it true that all of the numbers must be equal?

28. There are 1000 positive numbers written at different points on the circumference of a circle. If the numbers x, y, z appear in a row in that order, then it is known that $xz = y^2$.

    Prove that all of the numbers are equal.

29. There are m horizontal lines and n vertical lines drawn in the plane. Each point of intersection between a pair of lines is coloured in one of 100 colours.

    Find values of m and n such that, no matter how the colouring is performed, there always exists a rectangle whose vertices are the same colour.

30. Prove that from any set of 10 distinct two-digit numbers, it is possible to select two disjoint subsets whose members have the same sum.

31. Does there exist a convex polyhedron such that no two of its faces have the same number of edges?

[4] See section 11.2 if you don't know what this is.

1.1 Logic and deduction

Mathematics is brimming with statements of the form

$$\text{If } X, \text{ then } Y.$$

For example, we could take X to be the statement 'Rex is a dog' and Y to be the statement 'Rex is an animal'. Being such lazy creatures, we mathematicians have invented the following shorthand for such statements, which is often read as 'X implies Y'.

$$X \Rightarrow Y$$

Now if $X \Rightarrow Y$ and $Y \Rightarrow Z$, then you can automatically deduce that $X \Rightarrow Z$. One of the easiest ways to write a proof is to string together a chain of deductions in this manner, starting with the assumptions of the problem and ending with the conclusion of the problem. This is often called a *direct proof*, an example of which follows.

Problem Prove the quadratic formula, which states that if $ax^2 + bx + c = 0$ and $a \neq 0$, then

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{or} \quad x = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Solution Since $a \neq 0$, we can divide both sides by a to obtain

$$ax^2 + bx + c = 0 \quad \Rightarrow \quad x^2 + \frac{b}{a}x + \frac{c}{a} = 0.$$

Now we use an algebraic trick, known as *completing the square*, to write the left-hand side as

$$x^2 + \frac{b}{a}x + \frac{c}{a} = \left(x + \frac{b}{2a}\right)^2 - \left(\frac{b^2 - 4ac}{4a^2}\right).$$

This is a good time to mention that you should never ever take equations like this for granted. So grab a pen and some paper and check that it's true for yourself! Once you've done that, you should be convinced that the quadratic equation now takes the following form.

$$\begin{aligned}\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2} \quad &\Rightarrow \quad x + \frac{b}{2a} = \frac{\pm\sqrt{b^2 - 4ac}}{2a} \\ &\Rightarrow \quad x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\end{aligned}$$

Of course, we were careful to consider both the positive and the negative square roots, as one should always do, and this completes the proof. □

1.2 Converse

For every statement of the form $X \Rightarrow Y$, there is something known as the *converse*, which is the statement $Y \Rightarrow X$. You may be tempted to think that these two statements mean the same thing, that is, if $X \Rightarrow Y$ is true, then $Y \Rightarrow X$ is true. But this is most definitely not the case. For example, using our two example statements from earlier, it is clear that

$$\text{Rex is a dog} \quad \Rightarrow \quad \text{Rex is an animal}$$

is a true statement, whereas the statement

$$\text{Rex is an animal} \quad \Rightarrow \quad \text{Rex is a dog}$$

is false, since there is the possibility that Rex could be a lizard or some other animal. So if we know that a statement is true, there is no guarantee whatsoever that its converse is also true. But sometimes it is, as in the following well-known example.

Problem Pythagoras' theorem states that, if a right-angled triangle has side lengths a, b, c, where c is the length of the hypotenuse, then $a^2 + b^2 = c^2$.

Assuming that Pythagoras' theorem is true, prove the converse of Pythagoras' theorem.

Solution Of course, the first thing we must do is write down what the converse actually is.

> If a triangle has side lengths a, b, c, where $a^2 + b^2 = c^2$, then the triangle is right-angled and c is the length of the hypotenuse.

Let's construct a right-angled triangle whose legs have lengths a and b, and let the hypotenuse have length d. The reason for this is because we can now invoke Pythagoras' theorem, which we already know to be true. It tells us that $a^2 + b^2 = d^2$. Using this in conjunction with our assumption that $a^2 + b^2 = c^2$, we deduce that $c^2 = d^2$, which implies that $c = d$.

Therefore, the triangle with side lengths a, b, c that we were given possesses exactly the same side lengths as the right-angled triangle that we have constructed. This means that the two triangles are, in fact, congruent. So the given triangle was indeed right-angled, as we intended to prove. Furthermore, the equation $a^2 + b^2 = c^2$ implies that c is the longest side length in the triangle, and hence is the length of the hypotenuse. □

In the previous solution, we relied on Pythagoras' theorem to prove its converse. This is a rather general strategy, so keep the following point in mind. If you are given a true statement and asked to prove its converse, then it is often advantageous, sometimes crucial, to use the original statement itself.

1.3 If and only if

Is there some way to combine Pythagoras' theorem and its converse into one super-duper Pythagorean statement? Yes, there most certainly is!

Pythagoras' theorem and its converse Suppose that a triangle has side lengths a, b, c, where c is the longest side. Then the triangle is right-angled if and only if $a^2 + b^2 = c^2$.

In general, the statements 'If X, then Y' and 'If Y, then X' can be combined to create the single statement 'X if and only if Y'. You can probably guess that the mathematical notation for this is simply

$$X \Leftrightarrow Y.$$

Now if someone actually asks you to prove a statement of the form $X \Leftrightarrow Y$, then what do you do? The simplest approach is to split the problem into two parts. First prove $X \Rightarrow Y$, then prove $Y \Rightarrow X$. The next problem not only demonstrates this point but also provides us with a useful way to test whether or not a number is divisible by 7. You should think about why this is so.

Problem If a and b are integers, prove that $10a + b$ is divisible by 7 if and only if $a - 2b$ is divisible by 7.

Solution As with most 'if and only if' statements, the proof naturally divides into two parts.

- *If $10a + b$ is divisible by 7, then $a - 2b$ is divisible by 7.*

  If $10a + b$ is divisible by 7, then certainly $5(10a + b) = 50a + 5b$ is divisible by 7. And if $50a + 5b$ is divisible by 7, then certainly $(50a + 5b) - 7(7a + b) = a - 2b$ is divisible by 7. This proves the statement in one direction.

- *If $a - 2b$ is divisible by 7, then $10a + b$ is divisible by 7.*

  If $a - 2b$ is divisible by 7, then certainly $(a - 2b) + 7(7a + b) = 50a + 5b$ is divisible by 7. And if $50a + 5b$ is divisible by 7, then certainly $(50a + 5b) \div 5 = 10a + b$ is divisible by 7. This proves the statement in the opposite direction and completes the proof. □

Hopefully, you will have noticed that the two parts are very similar in nature. Although this is reasonably common, there will be times when one direction is significantly easier to prove than the other. And, as we mentioned in the previous section, once you've proved the statement in one direction, you can often use it to your advantage to prove the statement in the other direction.

1.4 Contrapositive

The contrapositive is a way of turning a logical statement on its head to give an equivalent logical statement. For example, instead of saying

If Rex is a dog, then Rex is an animal,

we could say the equivalent statement

If Rex is *not* an animal, then Rex is *not* a dog.

In general, the contrapositive of the statement $X \Rightarrow Y$ is the equivalent statement

'not Y' $\Rightarrow$ 'not X',

where 'not X' is the opposite of X and 'not Y' is the opposite of Y. By calling a statement and its contrapositive equivalent, we mean that if the statement is true, then its contrapositive is true, while if the statement is false, then its contrapositive is false. In other words, proving either one will automatically prove the other. Note that if you take the contrapositive of the contrapositive, then you actually end up with the statement you started with. As an example, consider the following two statements, which are the contrapositives of each other.

- If a shape is a rectangle, then it has four sides.
- If a shape does not have four sides, then it is not a rectangle.

Problem If a and b are real numbers such that ab is irrational, then at least one of a and b must be irrational.

Solution The contrapositive of this statement is the following.

If a and b are both rational, then ab is rational.

This is a fact so simple that you probably take it for granted all the time! Let's prove it by writing $a = \frac{a_1}{a_2}$ and $b = \frac{b_1}{b_2}$, where a_1, a_2, b_1, b_2 are integers with a_2 and b_2 non-zero. The product is simply $ab = \frac{a_1 b_1}{a_2 b_2}$, which is clearly rational since $a_1 b_1$ and $a_2 b_2$ are integers with $a_2 b_2$ non-zero. □

1.5 Proof by contradiction

Proof by contradiction is a most useful method of proof, one that you have almost definitely used before without even realising it. For example, how can you prove that the Sun is far away from the Earth? This is easy, because if the Sun were close to the Earth, then it would be too hot for you to be alive reading this sentence![5]

The idea is that if you assume the opposite of what you are trying to prove, use only correct logical deductions, and arrive at a nonsensical conclusion, then something must be wrong somewhere. And since all of your logical deductions were correct, what must be wrong is your initial assumption. So the opposite of what you are trying to prove is false. In other words, what you are trying to prove is true.

The two examples we'll look at are mathematical gems which were known to the ancient Greeks over two thousand years ago. The first appears in a work by the mathematician Euclid while the second is usually attributed to Hippasus, a philosopher and disciple of Pythagoras.

Problem Prove that there are infinitely many primes.

Solution Let's suppose that there are only finitely many primes, so that we can list them all as $p_1, p_2, \ldots, p_n$. Euclid asks us to consider the number $N = p_1 \times p_2 \times \cdots \times p_n + 1$. The main fact that we'll use is that two consecutive integers cannot be divisible by the same prime. Since $N - 1$ is divisible by p_1, it's impossible for N to be divisible by p_1 as well. Similarly, it's impossible for N to be divisible by p_2 or p_3 or any of the primes in our supposedly complete list. So what does this all mean? It means that if we look at the prime factors of N, they must come from outside our list. This contradicts the fact that we started with a complete list of the primes. Since we've shown that no finite list of primes can be complete, it follows that there are infinitely many primes. □

If you're seeing proof by contradiction for the first time, then it can be a little baffling. Make sure you understand the previous argument completely before looking at the next example. Eventually, you should not only be able to understand proofs by contradiction, but also come up with your own.

Problem Prove that $\sqrt{2}$ is irrational.

Solution Let us assume on the contrary that $\sqrt{2}$ is actually a rational number and hope to find a contradiction. If $\sqrt{2}$ were rational, then we could express it as $\frac{p}{q}$, where p and q are integers which have no common factors greater than 1. Start with the equation $\frac{p}{q} = \sqrt{2}$ and remove square roots and fractions to obtain

$$\frac{p}{q} = \sqrt{2} \quad \Rightarrow \quad \frac{p^2}{q^2} = 2 \quad \Rightarrow \quad p^2 = 2q^2.$$

[5] Of course from a different point of view, the Sun *is* close to the Earth, relative to other astronomical objects!

Since the right-hand side is even, the left-hand side is even. So p^2, and hence p, is even. Therefore, we can write $p = 2m$ for some integer m and substitute this back into the previous equation.

$$(2m)^2 = 2q^2 \quad \Rightarrow \quad 2m^2 = q^2$$

Now it's time for the left-hand side to be even, which forces the right-hand side to be even. So q^2, and hence q, is even. But hold on a second! We've assumed that p and q have no common factors greater than 1 and also proved that both p and q are even! This contradiction means that our original assumption, that $\sqrt{2}$ is rational, was incorrect. Therefore, we must conclude that $\sqrt{2}$ is irrational, which was what we set out to do. □

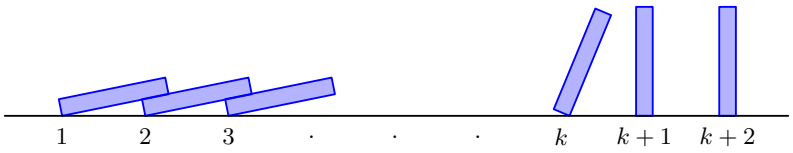
Already, we've learnt about the *con*verse, the *con*trapositive and *con*tradiction, which may create some *con*fusion. So take some time to learn the differences between these three *con*cepts!

1.6 Proof by induction

Suppose that you have an infinite supply of dominoes, numbered 1, 2, 3, and so on. They are all standing upright on a table and you want them all to fall over. One way of achieving this goal is to knock over the domino labelled 1, then knock over the domino labelled 2, then knock over the domino labelled 3, and so on. This would take infinitely long and is simply far too energy consuming for your average lazy mathematician. A much smarter way to do this would be to

- knock over the domino labelled 1
- make sure that if the domino labelled k falls, then it knocks over the domino labelled $k + 1$.

This is, in essence, what proof by induction is all about.



Analogously, suppose that you want to prove that some statement is true for all positive integers n. An incredibly silly way to do this would be to prove it for $n = 1$, then for $n = 2$, then for $n = 3$, and so on, but you would never get to the end of it. Instead, all you need to do is

- prove the *base case*, that is, prove the statement for $n = 1$
- prove the *inductive step*, i.e. prove that whenever the statement is true for $n = k$ (the *inductive hypothesis*), then it is also true for $n = k + 1$.

Using this idea to prove a statement for all positive integers n is called *proof by induction*. This may all seem pretty confusing to you at the moment, but the following example might help to clarify the situation.

Problem For every positive integer n, prove that

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}.$$

Solution For the base case, it is a simple matter to verify that the statement is true for $n = 1$. Indeed, the left-hand side is simply 1 while the right-hand side is $\frac{1(1+1)}{2}$. These two being equal takes care of the base case.

Next, for the inductive step, we can make use of the inductive hypothesis

$$1 + 2 + 3 + \cdots + k = \frac{k(k+1)}{2}$$

in order to prove that the statement is true for $n = k + 1$. But if we know the value of $1 + 2 + 3 + \cdots + k$, then surely it must be an easy matter for us to determine the value of $1 + 2 + 3 + \cdots + k + (k + 1)$. In fact, we have

$$\begin{aligned} 1 + 2 + 3 + \cdots + k + (k+1) &= \frac{k(k+1)}{2} + (k+1) \\ &= \frac{(k+1)(k+2)}{2}. \end{aligned}$$

And this is exactly the statement that we are trying to prove, with $n = k + 1$. This takes care of the inductive step.

But wait a moment! Do we actually know that the statement is true for $n = k$? Well, no we don't. But what we have shown is that *if* the statement is true for $n = k$, *then* it must also be true for $n = k + 1$. And since we already know it's true for $n = 1$, then it must also be true for $n = 2$. And since it's true for $n = 2$, then it must also be true for $n = 3$. And since it's true for $n = 3$, then it must also be true for $n = 4$, and so on. In this way we have managed to prove by induction that, for every positive integer n,

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}. \qquad \square$$

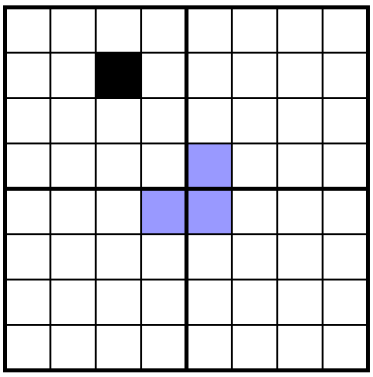
Problem An *L-tromino* is a shape formed by joining three unit squares along their edges to form an L shape.

Prove that, if any square of a $2^{100} \times 2^{100}$ chessboard is removed, then the part of the board which remains can be tiled by L-trominoes.

Solution Hopefully, one of the first things you realise when reading this problem is the fact that the number 100 is not important at all! In fact, we will prove that the statement is true for a $2^n \times 2^n$ chessboard, where n is any positive integer. Needless to say, we will proceed by induction.

The base case $n = 1$ is, as usual, extremely simple though entirely necessary to take care of. We simply note that removing a square from a 2×2 chessboard always leaves an L-tromino.

Next, for the inductive step, we assume the inductive hypothesis. In other words, that it is possible to tile a $2^k \times 2^k$ chessboard with any square removed. Our goal now is to prove that it is possible to tile a $2^{k+1} \times 2^{k+1}$ chessboard with any square removed. How can we relate these two facts? The idea is to divide the larger $2^{k+1} \times 2^{k+1}$ board into four blocks, each one a copy of the smaller $2^k \times 2^k$ board. After rotating the board, we may assume that the removed square lies in the top-left block. The inductive hypothesis guarantees that we can tile the remaining part of the top-left block, but what to do with the remaining three blocks?



If we want to rely on the inductive hypothesis again, it would be nice to remove one square from each of the remaining three blocks. We can do this by placing an L-tromino as shown in blue in the diagram above. Once again, the inductive hypothesis guarantees that we can tile the remaining parts of the remaining three blocks. This completes the inductive step.

Therefore we have proved by induction that, for every positive integer n, a $2^n \times 2^n$ chessboard with any square removed can be tiled by L-trominoes. □

Proof by induction is such an extremely important concept that we will finish this section with one further example.

Problem For every integer $n \geq 4$, prove that

$$2^n < n!.$$

(Recall $n! = n \times (n-1) \times \cdots \times 3 \times 2 \times 1$ for any positive integer n, and $0! = 1$ by convention.)

Solution This time, the statement we are trying to prove is only true for integers $n \geq 4$. This causes no problem since we can simply start from the base case $n = 4$ instead.

This is easy enough to verify, because $16 = 2^4 < 4! = 24$.

Next, for the inductive step, we assume the inductive hypothesis. In other words, that $2^k < k!$ for some value of $k \geq 4$. From this assumption, we would like to obtain the fact that the statement is true for $n = k + 1$, that is, $2^{k+1} < (k+1)!$. This is easily done, by starting with the inductive hypothesis and multiplying both sides by 2.

$$2^k < k! \quad \Rightarrow \quad 2^{k+1} < 2 \times k! < (k+1) \times k! = (k+1)!$$

And there we have it! Knowing that the statement is true for $n = 4$ tells us that it is true for $n = 5$. And knowing that the statement is true for $n = 5$ tells us that it is true for $n = 6$, and so on. So we have managed to prove by induction that $2^n < n!$, for every integer $n \geq 4$. □

1.7 Strong induction

One aspect of induction that makes it so powerful is that you are allowed to assume the statement is true for some positive integer k in order to prove it for the positive integer $k + 1$. However, sometimes it helps to know that the statement is true for $1, 2, \ldots, k$ in order to

prove it for $k+1$. This type of induction is certainly legitimate and is often called *strong induction*. This is a more advanced method of proof which can be extremely useful, such as in the solution to the following problem.

Problem You might already know that the Fibonacci sequence is defined by $F_1 = 1$, $F_2 = 1$ and $F_{m+1} = F_m + F_{m-1}$ for $m \geq 2$.

Prove that every positive integer can be expressed as a sum of terms from the Fibonacci sequence, no two of which appear consecutively.

Solution The Fibonacci sequence begins

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \ldots.$$

Let us call a positive integer *representable* if it can be expressed as a sum of non-consecutive Fibonacci numbers.

It is easy enough to show that the first few positive integers are representable: $1 = 1$, $2 = 2$, $3 = 3$, $4 = 3 + 1$, $5 = 5$, $6 = 5 + 1$, $7 = 5 + 2$, $8 = 8$, $9 = 8 + 1$, and $10 = 8 + 2$. In order to take care of the base case, we don't need to prove the statement for all of these values, but it certainly doesn't hurt.

Next, assume that every number from 1 up to k is representable. This is the *strong inductive hypothesis*. We are now going to show that $k+1$ is representable. If $k+1$ is actually a Fibonacci number, then it is, of course, representable. Otherwise, let F_m be the largest Fibonacci number less than it. Then $k + 1 - F_m$ is a positive integer which must be smaller[6] than F_{m-1} and, by the strong induction hypothesis, must be representable. Appending F_m onto the representation of $k + 1 - F_m$ gives a representation of $k + 1$. So we have managed to prove by strong induction that every positive integer is representable. □

This result is often called *Zeckendorf's theorem*. It gives us a pretty interesting way to represent positive integers, certainly more interesting than the customary base 10!

1.8 Proof by exhaustion (case bashing)

It is often a good strategy to divide and conquer. By this, we mean splitting a large problem into several smaller, simpler problems. Handling all of these smaller cases can be tough work, and that's why we refer to such a proof as a *proof by exhaustion* or as a *case bash*. Whenever you opt for this sort of proof, you should make absolutely sure that your cases are listed clearly and that they are indeed exhaustive.

Problem Prove that if n is an integer, then $n^7 - n$ is divisible by 7.

Solution First, we notice the factorisation

$$n^7 - n = (n-1)n(n+1)(n^2 - n + 1)(n^2 + n + 1).$$

We can divide the problem into seven cases according to the remainder that n leaves after division by 7.

- Case 1: $n = 7q + 1$
  In this case, the factor $n - 1 = 7q$ is divisible by 7.

[6] Your inquisitive younger sibling is asking why $k + 1 - F_m$ is smaller than F_{m-1}. Please provide a good reason.

- Case 2: $n = 7q + 2$
  In this case, the factor $n^2 + n + 1 = 49q^2 + 35q + 7$ is divisible by 7.
- Case 3: $n = 7q + 3$
  In this case, the factor $n^2 - n + 1 = 49q^2 + 35q + 7$ is divisible by 7.
- Case 4: $n = 7q + 4$
  In this case, the factor $n^2 + n + 1 = 49q^2 + 63q + 21$ is divisible by 7.
- Case 5: $n = 7q + 5$
  In this case, the factor $n^2 - n + 1 = 49q^2 + 63q + 21$ is divisible by 7.
- Case 6: $n = 7q + 6$
  In this case, the factor $n + 1 = (7q + 6) + 1 = 7q + 7$ is divisible by 7.
- Case 7: $n = 7q$
  In this case, the factor $n = 7q$ is divisible by 7.

Therefore, in all seven cases, at least one of the factors of $n^7 - n$ is divisible by 7. Since these cases account for every possible situation, we can conclude that if n is an integer, then $n^7 - n$ is divisible by 7, as desired.[7] □

1.9 Pigeonhole principle

The *pigeonhole principle* is a hugely powerful technique in the hands of an experienced problem solver. In its simplest form, it states that if you place $n + 1$ pigeons into n pigeonholes, then at least one pigeonhole contains at least 2 pigeons. You might think that this is completely self-evident, although the more astute and sadistic reader might have noticed that it isn't even correct. Take, for example, $1\frac{1}{5}$ pigeons placed in each of 5 pigeonholes! Anyway, it's a simple matter to amend the problem and the result is the following.

Pigeonhole principle If you place $n + 1$ pigeons into n pigeonholes, then at least one pigeonhole will contain at least 2 pigeons . . . as long as you don't cut them up!

This is now correct and, despite being blatantly obvious, is surprisingly useful. As with many of the techniques that we will learn, the power of the pigeonhole principle arises from the ingenious ways in which it can be applied, and what better way to demonstrate this than with an example.

The pigeonhole principle is particularly useful for problems where you have to show the existence of something without being able to construct it explicitly.

For example, suppose that you were asked to prove that there are two people in Australia with the same number of hairs on their head. Your first instinct might be to go and find two completely bald people! But if that seems too difficult for you, then you can also solve the problem with a little thought and an application of the pigeonhole principle. For there are over twenty million people in Australia, and each one of them has fewer than one million hairs on their head. In fact, the average head full of hair will have approximately one hundred thousand hairs. So the population of Australia will form our pigeons and we will put them into pigeonholes numbered from 0 up to 999999 according to the number of hairs on their head. The pigeonhole principle then guarantees that there will be at least two who end up in the same pigeonhole.

[7] A much quicker way to deal with this question is provided by Fermat's little theorem, found in section 2.12.

The next couple of problems illustrate how the pigeonhole principle can be used in number theory and combinatorics.

Problem Prove that any set of n integers has a non-empty subset whose sum is divisible by n.

Solution Let the integers be $a_1, a_2, \ldots, a_n$ and consider the following numbers.

$$\begin{aligned} s_1 &= a_1 \\ s_2 &= a_1 + a_2 \\ &\vdots \\ s_n &= a_1 + a_2 + \cdots + a_n \end{aligned}$$

We treat these n numbers as our pigeons and place them into pigeonholes according to their value modulo n. If the numbers $s_1, s_2, \ldots, s_n$ are all distinct modulo n, then there must be one of them, say s_k, which is divisible by n. In other words, $a_1 + a_2 + \cdots + a_k$ is divisible by n.

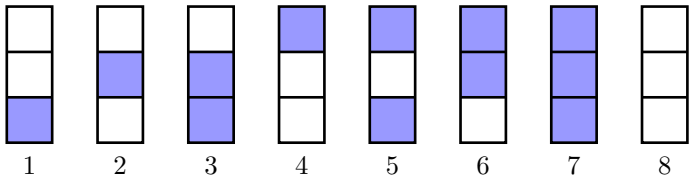
Otherwise, the pigeonhole principle guarantees that two of the numbers, say s_i and s_j, are the same modulo n. This means that $s_j - s_i$ is divisible by n, where we may assume without loss of generality that $i < j$. In other words, $a_{i+1} + a_{i+2} + \cdots + a_j$ is divisible by n. □

Problem Suppose that each square in a 3×7 grid is coloured blue or white.

Prove that there exist two rows and two columns such that the four squares contained in their intersection are the same colour.

Solution One rather foolish way to solve this problem would be to consider all of the possible colourings of the grid, of which there are $2^{3\times 7} = 2097152$. Of course, this is a rather inelegant solution, not to mention the fact that it could take months! We'll take a shortcut to the solution by making use of the pigeonhole principle. One feature of the problem which suggests that we should do so is the fact that we are asked to prove the existence of something, without demonstrating precisely what that something is. But even though we know that we would like to invoke the pigeonhole principle, some ingenuity is still required to know exactly how to proceed.

Let us call four squares formed by the intersection of two rows and two columns a *quartet* and refer to such a quartet as *monochromatic* if all four squares are the same colour. Our first observation is that there are eight possibilities for each column, as shown in the diagram below.



Our second observation is that if one of these occurs twice, then we must necessarily have a monochromatic quartet. Our third observation is that if you use column 7 in conjunction with columns 3, 5 or 6, then you must have a monochromatic quartet and similarly, if you use column 8 in conjunction with columns 1, 2 or 4, then you must have a monochromatic quartet. The problem now divides naturally into three cases.

- Case 1: Column 7 is used.
  If any of the remaining six columns are of type 3, 5, 6 or 7, we have a monochromatic quartet. If, on the other hand, each of the remaining six columns are of type 1, 2, 4 or 8, then by the pigeonhole principle, two are the same, thereby forcing a monochromatic quartet.
- Case 2: Column 8 is used.
  A similar argument to case 1 shows that a monochromatic quartet occurs.
- Case 3: Columns 7 and 8 are not used.
  Therefore, the seven columns must be of type 1, 2, 3, 4, 5 or 6. So, by the pigeonhole principle, two are the same, thereby forcing a monochromatic quartet. □

1.10 Advanced pigeonhole principle

The pigeonhole principle can be applied in quite an advanced manner. For example, the pigeonholes don't have to appear to be the same size, as the next problem demonstrates.

Problem Show that if we take $n+1$ numbers from the set $\{1, 2, 3, \ldots, 2n\}$, there must exist one which is divisible by another.

Solution Since this problem asks for the existence of two numbers satisfying certain conditions, it seems like a prime candidate for the pigeonhole principle. We are given $n+1$ numbers, which will most likely be our pigeons, so it seems sensible to look for n pigeonholes. Furthermore, these should be constructed so that given any two numbers from the same pigeonhole, one is divisible by the other. If we can construct n pigeonholes from the set $\{1, 2, 3, \ldots, 2n\}$ which satisfy these conditions, then we will be done.

Let's play around with some small numbers then. Clearly, 1 and 2 can be in the same pigeonhole, although 3 cannot lodge with them. The next number which can join them is 4, and after that 8. In fact, we can put all numbers of the form 2^a into one pigeonhole. If we start with 3 in a pigeonhole, then the next number which can join it is 6, then 12, then 24, and so on. So the numbers which are of the form 3×2^a can all be put into one pigeonhole. Similarly, the numbers which are of the form 5×2^a can all be put into one pigeonhole. Continuing in this way, each positive integer is in exactly one pigeonhole.[8]

$$
\begin{aligned}
X_1 &= \{1, 2, 4, 8, 16, \ldots\} \\
X_2 &= \{3, 6, 12, 24, 48, \ldots\} \\
X_3 &= \{5, 10, 20, 40, 80, \ldots\} \\
&\vdots \\
X_k &= \{2k-1, 2(2k-1), 4(2k-1), 8(2k-1), 16(2k-1), \ldots\} \\
&\vdots
\end{aligned}
$$

Now that we've constructed the pigeonholes, the rest is easy. Given $n+1$ numbers from the set $\{1, 2, 3, \ldots, 2n\}$, they all belong to one of the n pigeonholes $X_1, X_2, \ldots, X_n$. But by the pigeonhole principle, two of them must end up in the same pigeonhole. These two numbers are of the form $m \times 2^a$ and $m \times 2^b$, for some positive integer m. Thus the larger one is divisible by the smaller one. □

[8] Your younger sibling wants to know why this is the case.

In the previous section, we proved that there are at least two people in Australia with the same number of hairs on their heads. But it seems likely that, with so many Australians around, we should be able to find a larger group of people, all with the same number of hairs on their heads. This idea is the basis for the following more advanced version of the pigeonhole principle.

Pigeonhole principle If you place $kn + 1$ pigeons into n pigeonholes, then at least one pigeonhole will contain at least $k + 1$ pigeons ... as long as you don't cut them up!

Problem Seventeen points are given inside a cube of side length 2.

Show that there exist three of them which form a triangle of area at most $\frac{\sqrt{3}}{2}$.

Solution We have 17 objects and need to prove the existence of three of them satisfying certain conditions. Given this information alone, it seems that the problem is begging for the pigeonhole principle to be used on it! Not only that, the numbers 17 and 3 imply that we should take the given points to be our pigeons and look for eight pigeonholes. Fortunately, there is a particularly obvious way to divide a cube of side length 2 into eight pieces, that is, by slicing it into eight unit cubes.

From here, the solution to the problem is evident, since the pigeonhole principle tells us that three of our points must lie in the same unit cube. These three points would form a triangle with the largest area if they were at non-adjacent vertices of the unit cube.[9] Note that three such points form an equilateral triangle with side length $\sqrt{2}$ which has area $\frac{\sqrt{3}}{2}$, as required. □

There is yet another version of the pigeonhole principle which you might as well know about, even if we won't require it to solve any of the problems in this particular chapter.

Infinite pigeonhole principle If you place infinitely many pigeons into finitely many pigeonholes, then at least one pigeonhole will contain infinitely many pigeons.

1.11 Extremal principle

When solving mathematics problems, it is often useful to look at the maximum or minimum of some value. For example, you might consider the largest, the smallest, the longest, the shortest, the coolest, the most colourful, and so on. This is precisely the *extremal principle*. It isn't so much a mathematical result, but a useful way of thinking. And thinking in this way can often make difficult problems seem easy and can make long, complicated proofs turn into simple, concise ones, as you will soon see.

Problem Consider an $m \times n$ rectangular grid, where each of the mn unit squares is labelled with a real number. Suppose that the number in each square is the average of the numbers in all neighbouring squares.

Prove that all of the labels must be the same.

Solution We start with the simple fact that every finite set of real numbers has a largest element, a fact which is not true for infinite sets. So, we can consider the largest number appearing in the grid and call it M. Consider the numbers in the adjacent squares. They are

[9] Your inquisitive younger sibling is asking why this produces a triangle of maximal area. Can you provide an explanation?

all less than or equal to M and yet their average is M. The only way that this can happen is if all of the numbers in the adjacent squares are also M. So what we've proved is that every square adjacent to one containing the number M also contains the number M. It now follows that every square in the grid is labelled with the number M. □

The next problem illustrates the fact that the extremal principle is particularly valuable when used in conjunction with proof by contradiction.

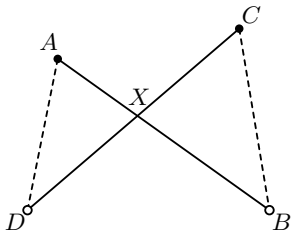
Problem Consider 10 black points and 10 white points in the plane, no three of which lie on a line.

Prove that it is possible to connect each black point with a white point by a line segment such that no two line segments intersect.

Solution Let us call a set of 10 line segments which connect each black point with a white point a *matching*. If we don't care whether or not the line segments intersect, there are only finitely many (actually 10!) matchings possible. Let us call the sum of the lengths of the 10 line segments the *length* of the matching. The idea now is that if you don't want your lines to cross, then you probably don't want to connect points faraway from each other. With this in mind, we consider a matching which has the shortest length and prove that it avoids intersecting line segments.

In order to obtain a contradiction, suppose that the line segments AB and CD intersect at a point X, where A and C are black points while B and D are white points. Then we can replace these with the line segments AD and BC. By the triangle inequality, which essentially states that the shortest distance between two points is a straight line segment, we have

$$AD < AX + XD \quad \text{and} \quad BC < BX + XC.$$



Adding these two inequalities, we obtain

$$AD + BC < AB + CD.$$

So we have constructed a new matching which is shorter than the one which was supposed to be the shortest. This contradiction implies that the shortest matching cannot contain two line segments which intersect, so we are done. □

There are two important points we should make about the previous proof. First, it was necessary to mention that there are only finitely many matchings. For if there were infinitely many, then it might not have been possible to choose one with the shortest length. Second, we were careful to choose *a* shortest matching rather than *the* shortest matching, since it is certainly possible that there could be many of them. Our proof works no matter which of the shortest matchings we decide to consider.

1.12 Telescoping

This is a useful method of evaluating sums. The idea is simply to rewrite things so that cancellation systematically occurs. Specifically, if we wish to find a closed form for the sum $a_1 + a_2 + \cdots + a_n$, then we seek a function F which satisfies

$$a_i = F(i+1) - F(i) \qquad \text{for } i = 1, 2, \ldots, n.$$

The sum *telescopes* because all the middle terms cancel out as follows.

$$\begin{aligned} a_n + a_{n-1} + \cdots + a_1 &= F(n+1) - F(n) + F(n) - F(n-1) + \cdots + F(2) - F(1) \\ &= F(n+1) - F(1) \end{aligned}$$

Problem Find a formula for

$$\sum_{i=1}^{n} (2i - 1) = 1 + 3 + \cdots + (2n - 1).$$

Solution This is quite easy once it is discovered that

$$2i - 1 = i^2 - (i-1)^2.$$

This corresponds to $F(i) = (i-1)^2$. It follows that the given sum telescopes to

$$F(n+1) - F(1) = n^2 - 0^2 = n^2. \qquad \square$$

As a general principle, if a_i is a polynomial function of i of degree m, try $F(i)$ to be a polynomial function of i of degree $m+1$. If a_i is of the form $p(i)c^i$, where p is a polynomial of degree m and c is a constant, try $F(i) = q(i)c^i$ where q is a polynomial of degree at least m. If a_i involves binomial coefficients or factorials, try $F(i)$ to be something similar.

2

Number theory

The great mathematician Carl Friedrich Gauss once stated that 'Mathematics is the queen of the sciences and number theory is the queen of mathematics'. You get the feeling that he considered number theory to be a pretty important area. Indeed, the study of such simple objects as the integers leads to incredibly deep problems for which a whole assortment of ideas and techniques have been developed. This chapter comprises a collection of some of the more simple, though still very useful, ideas and techniques that are used in number theory.

2.0 Problems

1. For any right-angled triangle whose side lengths are positive integers, we define the *tri-product* of the triangle to be the product of its three side lengths.

   Find with proof the greatest common divisor of all tri-products of all such right-angled triangles.

2. Find all solutions in positive integers to the equation

   $$m! + 5 = n^3.$$

3. The numbers from 1 to 81 are written in a 9×9 square array.

   Prove that there exists a positive integer k such that the product of the numbers in row k does not equal the product of the numbers in column k.

4. Find all positive integers p such that

   $$p, \quad 4p^2 + 1 \quad \text{and} \quad 6p^2 + 1$$

   are all primes.

5. Let A be the sum of the digits of 4444^{4444}, written in decimal notation.

   If B is the sum of the digits of A, find the sum of the digits of B.

6. Let n be a positive integer such that $2^n - 1$ is a prime number.

   Prove that n is a prime number.

7. Let n be a positive integer such that $2^n + 1$ is a prime number.

   Prove that $n = 2^k$ for some integer k.

8. Find all 5-digit natural numbers such that after deleting any one digit, the remaining number is a 4-digit number which is divisible by 7.

   (Note that leading zeros do not count as digits, e.g. 00123 counts as a 3-digit number.)

9. Show that there do not exist prime numbers p and q which differ by 2 and have the property that

$$pq + 10^k$$

   is also a prime number for some non-negative integer k.

10. Determine all positive integers $x_1 > x_2 > x_3 > x_4 > x_5$ such that

$$\left\lfloor \frac{x_1 + x_2}{3} \right\rfloor^2 + \left\lfloor \frac{x_2 + x_3}{3} \right\rfloor^2 + \left\lfloor \frac{x_3 + x_4}{3} \right\rfloor^2 + \left\lfloor \frac{x_4 + x_5}{3} \right\rfloor^2 = 38.$$

    (Note that $\lfloor x \rfloor$ is the largest integer not exceeding x.)

11. Let φ be the Euler phi function.[1]

    (a) Prove that $\varphi(n)$ is odd if and only if $n = 1$ or $n = 2$.

    (b) Find all solutions to the equation $\varphi(n) = 4$.

12. The numbers $1!, 2!, 3!, \ldots, 100!$ are written on a blackboard.

    Is it possible to erase one of the numbers so that the product of the remaining 99 numbers is a perfect square?

13. Can you find a set of 2000 distinct positive integers such that the sum of the members of every subset is not a perfect square?

14. For each positive integer k let

$$a_k = 2^{2^k} + 1.$$

    (a) Prove that if $i \neq j$, then $\gcd(a_i, a_j) = 1$.

    (b) Use the result of part (a) to prove that there are infinitely many primes.

15. Show that for every positive integer n,

$$121^n - 25^n + 1900^n - (-4)^n$$

    is divisible by 2000.

16. Show that there are infinitely many positive integers m for which

$$18^m + 45^m + 50^m + 125^m$$

    is divisible by 2006.

17. Let n be a given positive integer. Prove that the sequence

$$a, a^a, a^{a^a}, a^{a^{a^a}}, \ldots$$

    is eventually constant modulo n for any positive integer a.

[1] The Euler phi function is defined by the property that $\varphi(n)$ is equal to the number of integers from the set $\{1, 2, \ldots, n\}$ which have greatest common divisor 1 with n.

18. Let $V = \{1, 2, \ldots, 24, 25\}$. Alexander is looking for a subset W of V with the property that no two different members of W have a product which is a perfect square.

    (a) Find the maximal possible size for W.

    (b) Find the number of such sets W with this maximal size.

19. Prove that for each positive integer n, there exists a unique number divisible by 2^n whose decimal representation consists of n digits, each of them equal to 1 or 2.

20. Prove the following useful lemma. Suppose that positive integers a, b, x, y satisfy

$$a^x = b^y.$$

    Prove there exist positive integers r, m, n such that

$$a = r^m \quad \text{and} \quad b = r^n.$$

    (Note that in particular this implies that either $a \mid b$ or $b \mid a$.)

21. Call a set of positive integers *funky* if every pair of elements has greatest common divisor not equal to 1, while every triple of elements has greatest common divisor equal to 1.

    Show that there does not exist a funky set with infinitely many elements.

22. Prove that the sequence $a_1, a_2, a_3, \ldots$ defined by

$$a_n = \left\lfloor n + \sqrt{n} + \frac{1}{2} \right\rfloor$$

    contains every positive integer exactly once apart from the perfect squares.

23. Prove that every prime factor of $2^{2^k} + 1$ is of the form $2^{k+1}x + 1$, for some positive integer x, and hence prove that there are infinitely many primes of the form $2^w x + 1$ for each positive integer w.

24. Let n be a positive integer such that $4^n + 2^n + 1$ is a prime number.

    Prove that $n = 3^k$ for some integer k.

25. The squares of an infinite grid are numbered as illustrated. The number 0 is placed in the top-left corner. Each remaining square is numbered with the smallest non-negative integer that does not already appear to the left of it in the same row or above it in the same column.

    Which number will appear in the 1003rd row and 1980th column?

0	1	2	3	4	
1	0	3	2	5	
2	3	0	1	6	
3	2	1	0	7	
4	5	6	7	0	

26. Determine the value of the following sum, where p and q are relatively prime positive integers.

$$\left\lfloor \frac{p}{q} \right\rfloor + \left\lfloor \frac{2p}{q} \right\rfloor + \cdots + \left\lfloor \frac{(q-1)p}{q} \right\rfloor$$

27. For any positive integer n, let $d(n)$ denote the number of positive divisors of n.

    For which n does the sequence

    $$n, d(n), d(d(n)), \ldots$$

    not contain any perfect squares?

28. Let $S = \{2!, 3!, 4!, \ldots\}$. Some members of S may be written as a product of smaller members of S, such as $4! = 3! \times 2! \times 2!$. Let's call such a number a *factorial-composite*. If a member of S cannot be written as a product of smaller members of S, such as $3!$, let's call such a number a *factorial-prime*.

    (a) Prove that S contains infinitely many factorial-composites.

    (b) Prove that S contains infinitely many factorial-primes.

    (c) Prove that S contains infinitely many factorial-composites which have at least two different ways of factoring them into factorial-primes.[2]

29. Prove that the sequence

    $$\left\lfloor \sqrt{2}+1 \right\rfloor, \left\lfloor (\sqrt{2}+1)^2 \right\rfloor, \left\lfloor (\sqrt{2}+1)^3 \right\rfloor, \ldots$$

    alternates between even and odd integers.

30. If a is an integer greater than 1, prove that

    $$\gcd(a^m - 1, a^n - 1) = a^{\gcd(m,n)} - 1$$

    for all positive integers m and n.

31. Show that for every positive integer n, there are infinitely many terms of the Fibonacci sequence which are divisible by n.

32. The geometric mean of m non-negative real numbers is the mth root of their product.

    (a) Prove that, for every positive integer n, there exists a set of n distinct positive integers such that the geometric mean of every subset is an integer.

    (b) Does there exist an infinite set of distinct positive integers such that the geometric mean of every finite subset is an integer?

33. (a) For an irrational number x, consider the sequence $\{x\}, \{2x\}, \{3x\}, \ldots$, where as usual $\{x\} = x - \lfloor x \rfloor$ denotes the *fractional part* of x.

    For every positive integer n, show that there exists a term of the sequence which lies between 0 and $\frac{1}{n}$.

    (b) Hence, for any real numbers a and b satisfying $0 \leq a < b \leq 1$, prove that there exists a term of this sequence which lies between a and b.

    (c) Use this fact to show that there exists a power of two whose decimal representation starts with the digits 123456789.

34. Find all powers of two with the property that, after deleting the first digit of its decimal representation, one again obtains a power of two.

[2] Thus the analogue of the fundamental theorem of arithmetic (factorisation into primes is unique up to the order of factors) is not true in this setting.

35. Find all solutions in positive integers to the equation

$$a^{b^2} = b^a.$$

36. Determine all positive integers which are relatively prime to all terms of the infinite sequence

$$a_n = 2^n + 3^n + 6^n - 1, \quad n \geq 1.$$

37. (a) Prove the following useful lemma. Let p be an odd prime and a a positive integer such that $p \mid a - 1$.
If $p^\alpha \parallel a - 1$, prove that

$$p^{\alpha+\beta} \parallel a^n - 1 \quad \text{if and only if} \quad p^\beta \parallel n.$$

(The notation $p^j \parallel k$ means that $p^j \mid k$ but $p^{j+1} \nmid k$.)

(b) What is the corresponding result if $p = 2$?

38. A positive integer n is called *groovy* if, for every positive integer a, n^2 divides $a^n - 1$ whenever n divides $a^n - 1$.

Show that there are infinitely many composite numbers which are groovy.

39. Find all Carmichael numbers[3] of the form $3pq$, where 3, p and q are distinct primes.

40. Find all pairs of positive integers (m, n) such that

$$n \mid 1 + m^{3^n} + m^{3^n \cdot 2}.$$

41. Find all positive integers n such that

$$\left\lfloor \frac{2^n}{n} \right\rfloor$$

is a power of 2.

42. Prove there are infinitely many positive integers n which can be expressed as the sum of the squares of k positive integers for every integer $1 \leq k \leq n - 14$.

43. Show that for every positive integer n there exist n distinct positive integers such that their sum is a perfect 2009th power, and their product is a perfect 2010th power.

44. Let $b, n > 1$ be integers. Suppose that for each $k > 1$ there exists an integer a_k such that

$$k \mid b - a_k^n.$$

Prove that $b = A^n$ for some integer A.

45. A *wobbly number* is a positive integer whose digits are alternately non-zero and zero with the units digit non-zero.

Determine all positive integers which do not divide any wobbly number.

46. Prove that, for every positive integer n, there exists a positive integer X such that

$$X, 2X, 3X, \ldots, nX$$

are all nontrivial perfect powers.

[3] See the problem in section 2.14 for the definition of a Carmichael number.

2.1 Fundamental theorem of arithmetic

Let's start with one of the most basic, though one of the most important, results in number theory.

Fundamental theorem of arithmetic Every positive integer has a unique prime factorisation.

Perhaps the most important word appearing in the above statement is the word *unique*. It tells us that a prime factorisation acts like a fingerprint. In other words, it provides a way to identify positive integers. And it's quite a useful one, because if you're given numbers in prime factorised form, then it's a simple matter to test for divisibility, count the number of divisors, calculate greatest common divisors, determine lowest common multiples, and so on.

Problem Call a set of positive integers *funky* if every pair of elements has greatest common divisor not equal to 1, while every triple of elements has greatest common divisor equal to 1.

Show that there exists a funky set with n elements for each positive integer n.

Solution The easiest way to show the existence of a funky set with n elements is simply to construct a funky set with n elements. This is exactly what we are going to do, and our final set of numbers will be called $\{a_1, a_2, \ldots, a_n\}$. Since the problem involves greatest common divisors, it makes sense to consider prime factorisations. In fact, let's create a table with n rows, one for each of the numbers $a_1, a_2, \ldots, a_n$, and infinitely many columns, one for each of the primes $p_1, p_2, p_3, \ldots$. We could then place a tick in the row corresponding to a and the column corresponding to p if and only if p appears in the prime factorisation of a. Your job now, if you choose to accept it, is to place ticks in the table in order to guarantee that we end up with a funky set.

For each pair of elements to have greatest common divisor larger than 1 simply means that for $1 \leq i < j \leq n$, there exists a column which contains ticks in the rows corresponding to a_i and a_j. For each triple of elements to have greatest common divisor equal to 1 simply means that no column can contain three or more ticks. However, this is easy to do! For each pair $1 \leq i < j \leq n$, just find a column—any single one of the infinitely many columns—in which to put exactly two ticks in the rows corresponding to a_i and a_j. For example, the construction for $n = 5$ looks something like the following.

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	$\cdots$
a_1	✓	✓	✓	✓							
a_2	✓				✓	✓	✓				
a_3		✓			✓			✓	✓		
a_4			✓			✓		✓		✓	
a_5				✓			✓		✓	✓	

This would give us the funky set consisting of $a_1 = p_1p_2p_3p_4$, $a_2 = p_1p_5p_6p_7$, $a_3 = p_2p_5p_8p_9$, $a_4 = p_3p_6p_8p_{10}$ and $a_5 = p_4p_7p_9p_{10}$. Since this construction can be carried out for any number of rows, the problem is solved. □

It's worth noting that we could have recorded in the table the highest power of p which divides a, although this particular problem didn't require us to do so. Another lesson we can learn from this problem is that thinking in terms of tables can often help to simplify a problem immensely.

2.2 Pigeonhole principle

You just can't get enough of the pigeonhole principle! It was introduced in chapter 1 and will rear its head again in chapter 13. For now, we'll use it to attack an interesting number theory problem.

Problem Show that for every positive integer n not divisible by 2 or 5, there exists a multiple of n all of whose digits are ones.

Solution Our pigeons will be the first $n+1$ numbers from the sequence $1, 11, 111, 1111, \ldots$, and our pigeonholes will be the n possible remainders modulo n. So the pigeonhole principle asserts that two of these numbers must be congruent to each other modulo n. Suppose that they consist of i ones and j ones, where we can assume without loss of generality that $i > j$. Then

$$\underbrace{111\ldots1}_{i} - \underbrace{111\ldots1}_{j} = \underbrace{111\ldots1}_{i-j}\underbrace{000\ldots0}_{j} \equiv 0 \pmod{n}.$$

However, since n is not divisible by 2 or 5, we are allowed to divide both sides of this congruence equation by 10^j. This leaves us with

$$\underbrace{111\ldots1}_{i-j} \equiv 0 \pmod{n},$$

so there certainly does exist a multiple of n all of whose digits are ones. □

2.3 Dealing with digits

It's reasonably common to write $\overline{a_n a_{n-1}\ldots a_0}$ for the number whose decimal digits, from left to right, are $a_n, a_{n-1}, \ldots, a_0$. One of the first tricks in dealing with digits is to express this number as

$$\overline{a_n a_{n-1}\ldots a_0} = a_n \times 10^n + a_{n-1} \times 10^{n-1} + \cdots + a_0 \times 10^0.$$

Another trick is to use the following facts from modular arithmetic. They're both easy to prove and provide simple tests for divisibility by 9 and 11.

$$\begin{aligned} \overline{a_n a_{n-1}\ldots a_1 a_0} &\equiv a_0 + a_1 + a_2 + \cdots + a_n && \pmod{9} \\ \overline{a_n a_{n-1}\ldots a_1 a_0} &\equiv a_0 - a_1 + a_2 - \cdots + (-1)^n a_n && \pmod{11} \end{aligned}$$

Problem Find all integers $n > 1$ for which there exist distinct positive integers a and b such that $n^a + 1$ can be obtained from $n^b + 1$ by reversing the order of its decimal digits, and vice versa.

Solution Let's assume without loss of generality that $a < b$. The first observation that we can make about the two numbers $n^a + 1$ and $n^b + 1$ is that they must have the same number of digits. It seems that we should be able to use this fact to show that n can't be all that large. Indeed, if we consider that the larger of these two numbers is less than 10 times the smaller, we obtain the inequality

$$\begin{aligned} & n^b + 1 < 10(n^a + 1) \\ \Rightarrow \quad & n^a(n^{b-a} - 10) < 9. \end{aligned}$$

By inspection, the latter inequality simply cannot hold if $n \geq 11$. Furthermore, if $n = 10$, then $n^a + 1$ takes the form $100\ldots01$, so reversing the order of its digits returns exactly the same number.

Thus, we are left to consider the case $2 \leq n \leq 9$. Since $n^a + 1$ and $n^b + 1$ have exactly the same digits, they must leave the same remainders modulo 9. Therefore,

$$\begin{aligned} & n^a + 1 \equiv n^b + 1 \pmod 9 \\ \Rightarrow \quad & n^a(n^{b-a} - 1) \equiv 0 \pmod 9. \end{aligned}$$

At this stage, it makes sense to divide the problem into cases.

- Case 1: Suppose that n is not divisible by 3.

  Then it must be the case that $n^{b-a} - 1 \equiv 0 \pmod 9$ or equivalently, $n^{b-a} \equiv 1 \pmod 9$. Earlier we found that $n^a(n^{b-a} - 10) < 9$, from which it follows that $n^{b-a} - 10 < \frac{9}{n^a} < \frac{9}{2}$. Therefore, $n^{b-a} < 14\frac{1}{2}$. Since $n^{b-a} \equiv 1 \pmod 9$ we deduce that $n^{b-a} = 10$. This is clearly impossible, since 10 is certainly not the power of any integer from 2 to 9.

- Case 2: Suppose that n is equal to 3.

  This provides us with one solution to the problem since $3^3 + 1 = 28$ and $3^4 + 1 = 82$.

- Case 3: Suppose that n is equal to 6.

  Note that $6^a + 1$ and $6^b + 1$ both end in the digit 7, so they must both also start with the digit 7. However, it's easy to show that $6^b + 1$ is at least five times larger than $6^a + 1$, which then prevents them from having the same number of digits.

- Case 4: Suppose that n is equal to 9.

  Note that $9^a + 1$ and $9^b + 1$ must end in the digit 0 or 2, so they must both also start with the digit 0 or 2. Since the former doesn't yield any solutions, we may assume that both $9^a + 1$ and $9^b + 1$ start and end in the digit 2. However, it's easy to show that $9^b + 1$ is at least eight times larger than $9^a + 1$, which then prevents them from having the same number of digits.

In conclusion, the one and only integer which satisfies the conditions of the problem is 3. □

2.4 Floor function

Recall that $\lfloor x \rfloor = n$ simply means that $n \leq x < n+1$, where n is an integer. So whenever you are confronted with the floor function, you can immediately write down inequalities which capture the same information. The floor function pops up in all sorts of places, including the following amazing result known as *Beatty's theorem*.

Problem Let α and β be positive irrational numbers satisfying

$$\frac{1}{\alpha} + \frac{1}{\beta} = 1.$$

Prove that the sequences

$$\lfloor \alpha \rfloor, \lfloor 2\alpha \rfloor, \lfloor 3\alpha \rfloor, \ldots \quad \text{and} \quad \lfloor \beta \rfloor, \lfloor 2\beta \rfloor, \lfloor 3\beta \rfloor, \ldots$$

together include every positive integer exactly once.[4]

Solution First, we note that $\alpha > 1$ and $\beta > 1$, which implies that the two sequences must be strictly increasing. Next, we'll show that no integer can occur in both sequences. To do this, suppose that $\lfloor i\alpha \rfloor = \lfloor j\beta \rfloor = n$. Since α and β are irrational, it cannot be the case that $i\alpha = n$ or $j\beta = n$, so we have the strict inequalities

$$n < i\alpha < n+1 \qquad \text{and} \qquad n < j\beta < n+1.$$

Rearranging these inequalities yields

$$\frac{n}{\alpha} < i < \frac{n+1}{\alpha} \qquad \text{and} \qquad \frac{n}{\beta} < j < \frac{n+1}{\beta},$$

which add to give

$$n = \frac{n}{\alpha} + \frac{n}{\beta} < i + j < \frac{n+1}{\alpha} + \frac{n+1}{\beta} = n+1.$$

However, this is a contradiction because $i + j$ is an integer, so it can't lie between the consecutive integers n and $n+1$. Therefore, no integer can occur in both sequences.

Finally, we'll show that every integer occurs in one of the sequences. To do this, suppose that the integer n doesn't occur in either sequence. Then there must exist integers i and j such that

$$i\alpha < n < n+1 < (i+1)\alpha \qquad \text{and} \qquad j\beta < n < n+1 < (j+1)\beta.$$

Rearranging these inequalities yields

$$i < \frac{n}{\alpha}, \qquad i+1 > \frac{n+1}{\alpha}, \qquad j < \frac{n}{\beta} \qquad \text{and} \qquad j+1 > \frac{n+1}{\beta}.$$

These add to give

$$i + j < \frac{n}{\alpha} + \frac{n}{\beta} = n \qquad \text{and} \qquad i + j + 2 > \frac{n+1}{\alpha} + \frac{n+1}{\beta} = n+1.$$

However, this is a contradiction because it implies that the integer $i + j$ must lie between the consecutive integers $n-1$ and n. Therefore, no integer fails to occur in both sequences. Piecing all the information together, we find that the sequences together include every positive integer exactly once. □

2.5 Square roots and conjugates

In this section, we consider numbers of the form $a + b\sqrt{d}$, where a and b are rational numbers, while d is an integer which is not a perfect square. You should be able to prove that $\sqrt{d}$ is irrational and that $a_1 + b_1\sqrt{d} = a_2 + b_2\sqrt{d}$ implies that $a_1 = a_2$ and $b_1 = b_2$. If you find the number $r = a + b\sqrt{d}$ appearing in a problem, it's almost always useful to consider its conjugate $\overline{r} = a - b\sqrt{d}$. The beauty of conjugates is that their sum, difference and product are all nice expressions.

$$\begin{aligned}
(a + b\sqrt{d}) + (a - b\sqrt{d}) &= 2a \\
(a + b\sqrt{d}) - (a - b\sqrt{d}) &= 2b\sqrt{d} \\
(a + b\sqrt{d}) \times (a - b\sqrt{d}) &= a^2 - db^2.
\end{aligned}$$

[4] The converse is also true! If the two sequences $\lfloor \alpha \rfloor, \lfloor 2\alpha \rfloor, \lfloor 3\alpha \rfloor, \ldots$ and $\lfloor \beta \rfloor, \lfloor 2\beta \rfloor, \lfloor 3\beta \rfloor, \ldots$ together include every positive integer exactly once, then α and β are positive irrational numbers satisfying $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. See if you can prove it!

Conjugates also behave nicely under addition, subtraction, multiplication and division. More explicitly, if r_1 and r_2 are numbers of the form $a + b\sqrt{d}$, then we have the following facts.

$$\overline{r_1 + r_2} = \overline{r_1} + \overline{r_2} \qquad \overline{r_1 - r_2} = \overline{r_1} - \overline{r_2}$$
$$\overline{r_1 \times r_2} = \overline{r_1} \times \overline{r_2} \qquad \overline{r_1 \div r_2} = \overline{r_1} \div \overline{r_2}$$

One particularly useful corollary is the fact that

$$\overline{r^n} = \overline{r}^n,$$

where on the left, the exponent appears inside the conjugate, while on the right, the exponent appears outside the conjugate.

Problem Show that if we raise $\sqrt{2} - 1$ to the power of a positive integer, then the result is of the form $\sqrt{m} - \sqrt{m-1}$ for some positive integer m.

Solution If we expand $(\sqrt{2} - 1)^n$ by using the binomial theorem[5], we obtain

$$(\sqrt{2} - 1)^n = a + b\sqrt{2}$$

for some integers a and b. Now consider the conjugate expression to obtain

$$(-\sqrt{2} - 1)^n = a - b\sqrt{2}.$$

Multiplying these two equations together gives

$$\begin{aligned} & (a + b\sqrt{2})(a - b\sqrt{2}) = (\sqrt{2} - 1)^n(-\sqrt{2} - 1)^n \\ \Rightarrow \quad & a^2 - 2b^2 = (-1)^n. \end{aligned}$$

So for n even, we have $a^2 - 2b^2 = 1$ and we can take $m = a^2$ to obtain

$$\begin{aligned} (\sqrt{2} - 1)^n &= a + b\sqrt{2} \\ &= \pm\sqrt{m} \pm \sqrt{m-1}. \end{aligned}$$

However, since $\sqrt{2} - 1$ is between 0 and 1, it follows that $(\sqrt{2} - 1)^n$ must also be between 0 and 1. So it must be the case that $(\sqrt{2} - 1)^n = \sqrt{m} - \sqrt{m-1}$.

The case for n odd can be handled in a similar way. □

2.6 Powers of two

The numbers $1, 2, 4, 8, 16, \ldots$ seem to appear in a seeming infinitude of number theory problems. So in this section, let's pay our respects to the powers of two by considering the following problem.

Problem For each positive integer n, determine the remainder when $3^{2^n} - 1$ is divided by 2^{n+3}.

[5] See section 13.3 if you don't know what this is.

Solution Given an expression like $3^{2^n}-1$, you can't help but want to use the *difference of perfect squares* factorisation, otherwise known by the acronym DOPS. This gives

$$3^{2^n}-1=(3^{2^{n-1}}+1)(3^{2^{n-1}}-1)$$

and once again, DOPS can be happily applied to the second factor. This gives

$$3^{2^n}-1=(3^{2^{n-1}}+1)(3^{2^{n-2}}+1)(3^{2^{n-2}}-1).$$

So if we continue with this veritable feast of DOPS, we will finally be left with

$$3^{2^n}-1=(3^{2^{n-1}}+1)(3^{2^{n-2}}+1)\cdots(3^2+1)(3^1+1)(3^1-1).$$

We'd like to know the remainder after this number is divided by 2^{n+3}. But with so many brackets, all being even, it seems that our number might actually be divisible by 2^{n+3}. Let's see if this is the case. A keen observation will reveal that for integers m,

$$3^{2m}+1\equiv(-1)^{2m}+1\equiv 2 \pmod 4.$$

So all of the brackets, apart from the final two, are divisible by 2, but not 4. Since there are $n-1$ such brackets, and the final two brackets contribute a factor of 8, our number must be divisible by $2^{n-1}\times 8=2^{n+2}$, but not by 2^{n+3}. It follows that $3^{2^n}-1$ leaves a remainder of 2^{n+2} when divided by 2^{n+3}. □

2.7 Euclid's algorithm

One of the gems of number theory is *Euclid's algorithm*. Ostensibly, it provides a recipe for calculating the greatest common divisor of two positive integers, but it is so much more than that! The basis for it is the *division algorithm*, which we now state.

Division algorithm For any two integers a and $b\neq 0$, there is a unique way to write $a=qb+r$, where $0\leq r<|b|$.

A particular corollary of the division algorithm is that, once we write $a=qb+r$ as specified, then we have $\gcd(a,b)=\gcd(b,r)$. Iterating this process until we are left with a greatest common divisor which we can calculate by inspection is what Euclid's algorithm is all about. As with many algorithms, the best way to demonstrate Euclid's algorithm is through an example, which we do in the following problem.

Problem Find a number N which is divisible by 39 and such that $N+1$ is divisible by 106.

Solution We'd like to solve the equations $N=39a$ and $N+1=106b$ for integers a and b. However, these imply that $106b-39a=1$, which is known as a linear Diophantine equation and can be solved with the help of Euclid's algorithm. In the left column below, we apply Euclid's algorithm, by repeatedly applying the division algorithm as shown.

$$\begin{aligned}106&=2\times 39+28\\39&=1\times 28+11\\28&=2\times 11+6\\11&=1\times 6+5\\6&=1\times 5+1\end{aligned}\qquad\qquad\begin{aligned}1&=1\times 6-1\times 5\\&=2\times 6-1\times 11\\&=2\times 28-5\times 11\\&=7\times 28-5\times 39\\&=7\times 106-19\times 39\end{aligned}$$

On the left, we have applied Euclid's algorithm to the numbers 106 and 39, thereby showing that $\gcd(106, 39) = 1$. On the right, we have sneakily reversed Euclid's algorithm. The last line of Euclid's algorithm allows us to write the number 1 as a multiple of 5 (namely -5), plus a multiple of 6 (namely 6). We can express this idea by saying that 1 is a *combination* of 5 and 6. The second last line of Euclid's algorithm allows us to exchange the number 5 for a combination of 6 and 11. Therefore, we can write the number 1 as a combination of 6 and 11. The third last line of Euclid's algorithm allows us to exchange the number 6 for a combination of 11 and 28. Therefore, we can write the number 1 as a combination of 11 and 28. We can continue in this way until we have finally expressed the number 1 as a combination of 39 and 106, which is what we set out to do. Given that $7 \times 106 - 19 \times 39 = 1$, it is clear that we should take $N = 19 \times 39 = 741$. □

Next, we'll look at a problem which mixes Euclid's algorithm with some Fibonacci fun! The result is a rather amazing fact concerning the greatest common divisor of two Fibonacci numbers.

Problem Recall that the *Fibonacci sequence* is defined by

$$F_0 = 0, \quad F_1 = 1 \quad \text{and} \quad F_{m+1} = F_m + F_{m-1} \quad \text{for } m \geq 1.$$

Prove that $\gcd(F_m, F_n) = F_{\gcd(m,n)}$.

Solution Let's break the problem into bite-sized pieces which we'll finally put together to produce a complete proof.

- *For every positive integer n, $\gcd(F_n, F_{n-1}) = 1$.*
  We have $\gcd(F_n, F_{n-1}) = \gcd(F_{n-1} + F_{n-2}, F_{n-1}) = \gcd(F_{n-1}, F_{n-2})$, by the division algorithm. Repeated use of this fact yields

  $$\begin{aligned}\gcd(F_n, F_{n-1}) = \gcd(F_{n-1}, F_{n-2}) = \gcd(F_{n-2}, F_{n-3}) = \cdots \\ \cdots = \gcd(F_3, F_2) = \gcd(F_2, F_1) = \gcd(F_1, F_0) = 1.\end{aligned}$$

- *For any positive integers m and n, $F_{m+n} = F_{m+1}F_n + F_m F_{n-1}$.*
  This statement will follow if we can prove by induction on k that

  $$F_N = F_{k+1}F_{N-k} + F_k F_{N-k-1}$$

  for all positive integers N and all $k = 0, 1, 2, \ldots, N$. The statement is easily seen to be true for $k = 0$ and the induction will be complete once we have verified that $F_{k+1}F_{N-k} + F_k F_{N-k-1} = F_{k+2}F_{N-k-1} + F_{k+1}F_{N-k-2}$. However, this is equivalent to $F_{k+1}(F_{N-k} - F_{N-k-2}) = F_{N-k-1}(F_{k+2} - F_k)$, which in turn is equivalent to $F_{k+1}F_{N-k-1} = F_{N-k-1}F_{k+1}$.

- *If $m \mid n$, then $F_m \mid F_n$.*
  We will prove by induction on k that $F_m \mid F_{km}$ for every positive integer k. The statement is easily seen to be true when $k = 1$. Furthermore, $F_m \mid F_{km}$ implies that $F_m \mid F_{km+1}F_m + F_{km}F_{m-1} = F_{(k+1)m}$, where we have used the previous result. This completes the induction.

- *If $n = qm + r$, then $\gcd(F_m, F_n) = \gcd(F_m, F_r)$.*
  We simply write $F_n = F_{qm+r} = F_{qm+1}F_r + F_{qm}F_{r-1}$, which implies that

  $$\begin{aligned}\gcd(F_m, F_n) &= \gcd(F_m, F_{qm+1}F_r + F_{qm}F_{r-1}) \\ &= \gcd(F_m, F_{qm+1}F_r) \\ &= \gcd(F_m, F_r).\end{aligned}$$

To pass from the first line to the second line, we simply use the division algorithm, while to pass from the second line to the third line, we use the fact that $F_m \mid F_{qm}$ and $\gcd(F_{qm+1}, F_{qm}) = 1$.

If we write $n = qm + r$, then the statement $\gcd(F_m, F_n) = \gcd(F_m, F_r)$ is remarkably similar to the statement $\gcd(m, n) = \gcd(m, r)$, which forms the basis of Euclid's algorithm. In fact, the only difference is that the letter F has been added into the picture. Therefore, applying Euclid's algorithm to two Fibonacci numbers gives the result $\gcd(F_m, F_n) = F_{\gcd(m,n)}$. □

If you're not quite sure about how the proof works, then the following example may be enlightening.

$$\begin{aligned}
& & \gcd(106, 39) & & \gcd(F_{106}, F_{39}) \\
106 &= 2 \times 39 + 28 & &= \gcd(39, 28) & &= \gcd(F_{39}, F_{28}) \\
39 &= 1 \times 28 + 11 & &= \gcd(28, 11) & &= \gcd(F_{28}, F_{11}) \\
28 &= 2 \times 11 + 6 & &= \gcd(11, 6) & &= \gcd(F_{11}, F_6) \\
11 &= 1 \times 6 + 5 & &= \gcd(6, 5) & &= \gcd(F_6, F_5) \\
6 &= 1 \times 5 + 1 & &= \gcd(5, 1) & &= \gcd(F_5, F_1) \\
5 &= 5 \times 1 + 0 & &= \gcd(1, 0) & &= \gcd(F_1, F_0)
\end{aligned}$$

2.8 Integers base-n

Just because most of us have 10 digits on our hands, why should we represent positive integers using 10 digits? Indeed in computer science, it's useful to work in base-2, also known as binary, with only two symbols. Some computer science settings work in base-16, also known as hexadecimal, with 16 symbols. And, in mathematics, it sometimes pays to work in base-n for other positive integers n.

Problem Is it possible to choose 2000 distinct non-negative integers, all less than 100000, no three of which are consecutive terms of an arithmetic progression?

Solution Since we want to keep our numbers small, let's use the following greedy algorithm.[6] Start with the number $a_0 = 0$ and let a_{n+1} be the next largest integer which doesn't form a three term arithmetic progression with any of $a_0, a_1, a_2, \ldots, a_n$. So it's clear that we can take $a_1 = 1$, but then we have to skip 2 to avoid the arithmetic progression $(0, 1, 2)$. However, we can take $a_2 = 3$ and continue in this fashion to obtain an increasing sequence of positive integers. Before you read on, try to write down the first 20 terms of the sequence and see if you can spot a pattern.

If you can't spot the pattern, then at least you should have noticed that there are large jumps at certain parts of the sequence. The first jump is at $a_2 = 3$ and the next largest jump occurs at $a_4 = 9$. There are two more large jumps at $a_8 = 27$ and $a_{16} = 81$. The numbers 3, 9, 27 and 81, are all powers of 3, which suggests that something interesting might be happening in base-3. So let's write out a table which displays the numbers of our sequence in base-3.

[6] Informally, a *greedy algorithm* makes an optimal choice at each step. Such a tactic is not generally guaranteed to produce an optimal set overall. However, in many questions, including this one, it is sufficient to solve the problem.

n	0	1	2	3	4	5	6	7	8
a_n	0	1	3	4	9	10	12	13	27
base-3	0	1	10	11	100	101	110	111	1000

n	9	10	11	12	13	14	15	16	17
a_n	28	30	31	36	37	39	40	81	82
base-3	1001	1010	1011	1100	1101	1110	1111	10000	10001

Rather amazingly, the base-3 representation of a_n only contains the digits 0 and 1, just like numbers written in binary. And even more amazingly, it seems that the ternary representation of a_n is simply the binary representation of n. Of course, everything we've accomplished so far has merely been pattern spotting, so let's try to prove the following conjecture.

> Let a_n be the positive integer whose base-3 representation is the binary representation of n. Then no three terms of the sequence $a_0, a_1, a_2, \ldots$ form an arithmetic progression.

In order to prove this, we assume that the numbers x, y, z are distinct terms of the sequence which form an arithmetic progression, that is, $x + z = 2y$. So x and z both have ternary[7] representations containing only the digits 0 and 1, while $2y$ has a ternary representation containing only the digits 0 and 2. It's clear that if we add x and z in ternary, then there can be no carries. So in order for their sum to contain only the digits 0 and 2, each occurrence of the digit 1 in x must match up with a corresponding digit 1 in z and vice versa. In other words, x and z must actually be the same number, contradicting the fact that x, y, z are distinct. Therefore, we can be sure that no three terms of the sequence $a_0, a_1, a_2, \ldots$ form an arithmetic progression.

Now it just suffices to show that the 2000 numbers $a_0, a_1, a_2, \ldots, a_{1999}$ of our sequence are all less than 100000. Since the sequence is increasing, we need only consider the number a_{1999}. However, 1999 in binary is 11111001111 and 11111001111 in ternary is only 88249, much less than 100000. □

2.9 Construction problems

A number theory problem might ask you to come up with a construction which satisfies certain conditions. Such problems can sometimes require a great deal of ingenuity, but will certainly be easier to solve if you are guided by intuition and experience.

Problem Does there exist an infinite increasing sequence $t_1 < t_2 < t_3 < \cdots$ of positive integers such that, for any integer c, the sequence

$$t_1 + c, t_2 + c, t_3 + c, \ldots$$

contains only a finite number of primes?

Solution Consider a particular value of c, such as 73. How can we easily guarantee that the sequence $t_1 + 73, t_2 + 73, t_3 + 73, \ldots$ contains only a finite number of primes? One way is to make sure that, eventually, the numbers in the sequence $t_1, t_2, t_3, \ldots$ are divisible by 73. In fact, we would like this to be true for almost any value of c, which means that the numbers

[7] This is another term for base-3.

$t_1, t_2, t_3, \ldots$ should probably have lots of factors. A good candidate for such a sequence is $t_n = n!$. In fact, for any c with $|c| \geq 2$, the number $n! + c$ is divisible by $|c|$ and larger than it whenever $n > |c|$. So it's certainly true that there are finitely many primes in the sequence $1! + c, 2! + c, 3! + c, \ldots$ as long as $|c| \geq 2$. Furthermore, when $c = 0$, there are clearly no primes in the sequence apart from 2.

So we're left to deal with the cases $c = -1$ and $c = +1$. Unfortunately, there is no easy way to see that $n! \pm 1$ is composite, so our initial choice of sequence may have to be tweaked. We would like to change the sequence somehow, maintaining the abundance of factors in our numbers, but using numbers which are obviously composite when 1 is added or subtracted. So what sort of numbers are obviously composite when 1 is added or subtracted? One answer to this question is perfect cubes, since we have the factorisations

$$X^3 - 1 = (X - 1)(X^2 + X + 1) \quad \text{and} \quad X^3 + 1 = (X + 1)(X^2 - X + 1).$$

So the sequence $t_n = (n!)^3$ certainly satisfies the conditions of the question. □

Sometimes, you may be asked to perform a particular construction for all positive integers n. In that case, it should be easier to find a construction in the case $n + 1$ if you use the fact that you have already found a construction in the case n. This means that your construction will involve induction, quite a common phenomenon in number theory problems.

Problem Let n be a positive integer.

Prove that there are infinitely many perfect squares of the form $2^n a - 7$, where a is a positive integer.

Solution In the language of modular arithmetic, the problem is asking us to prove that there are infinitely many integers m such that $m^2 \equiv -7 \pmod{2^n}$. Our first observation is that once you've found one such m, then infinitely many such m are easy to find. This is because

$$m^2 \equiv -7 \pmod{2^n} \quad \Rightarrow \quad (m + 2^n)^2 \equiv -7 \pmod{2^n}.$$

So the problem is really asking us to prove that there exists at least one integer m such that $m^2 \equiv -7 \pmod{2^n}$. For the first several values of n, we can simply list such values of m.

$$\begin{aligned} 1^2 &\equiv -7 \pmod{2} & 1^2 &\equiv -7 \pmod{4} & 1^2 &\equiv -7 \pmod{8} \\ 3^2 &\equiv -7 \pmod{16} & 5^2 &\equiv -7 \pmod{32} & 11^2 &\equiv -7 \pmod{64} \end{aligned}$$

The idea behind our construction is as follows: if $m^2 \equiv -7 \pmod{2^n}$, then it must be true that either

$$m^2 \equiv -7 \pmod{2^{n+1}} \quad \text{or} \quad m^2 \equiv -7 + 2^n \pmod{2^{n+1}}.$$

In the former case, we can simply use m again. But in the latter case, we can use $m + 2^{n-1}$ and in this case we have

$$\begin{aligned} (m + 2^{n-1})^2 &\equiv m^2 + m2^n + 2^{2n-2} && \pmod{2^{n+1}} \\ &\equiv -7 + 2^n + m2^n + 2^{2n-2} && \pmod{2^{n+1}} \\ &\equiv -7 + 2^n(m + 1 + 2^{n-2}) && \pmod{2^{n+1}} \\ &\equiv -7 && \pmod{2^{n+1}}. \end{aligned}$$

This is certainly true for $n \geq 3$, since we know that m is necessarily odd, which forces $(m + 1 + 2^{n-2})$ to be even. So given the fact that $m^2 \equiv -7 \pmod{2^n}$, we've shown that either $m^2 \equiv -7 \pmod{2^{n+1}}$ or $(m + 2^{n-1})^2 \equiv -7 \pmod{2^{n+1}}$. Therefore, by induction, for every positive integer n, we have a value of m for which $m^2 \equiv -7 \pmod{2^n}$. □

2.10 Modular arithmetic

Many questions in number theory involve modular arithmetic in one way or another. One of the advantages of modular arithmetic is that we can add, subtract and multiply residues. Better still, if we are working modulo n, then we can also divide residues, as long as they are relatively prime to n. The explicit statement of this result follows.

Division modulo n For an integer a satisfying $\gcd(a, n) = 1$,

$$ax \equiv ay \pmod{n} \quad \text{if and only if} \quad x \equiv y \pmod{n}.$$

For an integer a satisfying $\gcd(a, n) = d > 1$,

$$ax \equiv ay \pmod{n} \quad \text{if and only if} \quad \frac{a}{d}x \equiv \frac{a}{d}y \pmod{\frac{n}{d}}.$$

Problem If p is a prime, show that $\binom{2p}{p} - 2$ is a multiple of p.

Solution We start by writing

$$\binom{2p}{p} = \frac{(2p)(2p-1)(2p-2)\cdots(p+1)}{p(p-1)(p-2)\cdots 1} = 2\frac{(2p-1)(2p-2)\cdots(p+1)}{(p-1)(p-2)\cdots 1}.$$

So what we are required to prove is

$$2\frac{(2p-1)(2p-2)\cdots(p+1)}{(p-1)(p-2)\cdots 1} \equiv 2 \pmod{p}.$$

If $p = 2$, then the statement is obvious, since both sides are congruent to 0. On the other hand, for $p > 2$ we can divide both sides by 2 and then multiply by the denominator to obtain

$$(2p-1)(2p-2)\cdots(p+1) \equiv (p-1)(p-2)\cdots 1 \pmod{p}.$$

This statement is equivalent to the previous one precisely because we are working modulo a prime, so we may divide by any number which is not 0 modulo p. The last equation is clearly true, because

$$2p-1 \equiv p-1, \quad 2p-2 \equiv p-2, \quad \ldots, \quad \text{and} \quad p+1 \equiv 1 \pmod{p}. \qquad \square$$

In fact, more is true. If $p > 3$ is prime, then $\binom{2p}{p} - 2$ is a multiple of p^3. See if you can prove it!

2.11 Chinese remainder theorem

Suppose we are looking for solutions to the simultaneous congruences

$$x \equiv 1 \pmod{4}, \quad x \equiv 2 \pmod{3}, \quad x \equiv 3 \pmod{5}.$$

You can verify that $x \equiv 53 \pmod{60}$ is a solution. In fact it represents all of the solutions. This is a special case of the following.

Chinese remainder theorem Consider the n congruences

$$x \equiv a_1 \pmod{m_1}, \quad x \equiv a_2 \pmod{m_2}, \quad \ldots, \quad x \equiv a_n \pmod{m_n}.$$

A solution for x definitely exists if $m_1, m_2, \ldots, m_n$ are pairwise relatively prime. Furthermore, the solution is unique modulo $m_1 m_2 \cdots m_n$.

This theorem is particularly useful if you wish to show the existence of numbers satisfying certain congruence conditions.

Problem Prove that for each positive integer n, there exist n consecutive positive integers, none of which is an integral power of a prime number.

Solution Consider the consecutive integers $x+1, x+2, \ldots, x+n$. To guarantee that $x+1$ is not an integral power of a prime, we simply need to make sure that it is divisible by two distinct primes, say p_1 and q_1. Similarly, we would like $x+2$ to be divisible by two distinct primes p_2 and q_2 and, in general, $x+k$ should be divisible by two distinct primes p_k and q_k. So what we have to do now is show the existence of a positive integer x which satisfies the following n simultaneous congruences.

$$\begin{aligned} x+1 &\equiv 0 \pmod{p_1 q_1} \\ x+2 &\equiv 0 \pmod{p_2 q_2} \\ &\vdots \\ x+n &\equiv 0 \pmod{p_n q_n} \end{aligned}$$

And this is where the Chinese remainder theorem comes to our rescue! This is because it guarantees that there exists a solution as long as the numbers $p_1 q_1, p_2 q_2, \ldots, p_n q_n$ are relatively prime. Therefore, if we simply pick the primes $p_1, p_2, \ldots, p_n, q_1, q_2, \ldots, q_n$ to be distinct, then we can invoke the Chinese remainder theorem to find a solution for x modulo m, where $m = p_1 q_1 p_2 q_2 \cdots p_n q_n$. □

Note that x is a solution to the problem if and only if $x+m$ is. This shows that for a given n there are infinitely many solutions.

2.12 From Fermat to Euler

In this section, we consider results which involve perfect powers in modular arithmetic. The first is due to Fermat and, to distinguish it from his many others, is referred to as his *little theorem*. However, this is by no means an accurate indication of its importance!

Fermat's little theorem If p is a prime and $\gcd(a, p) = 1$, then

$$a^{p-1} \equiv 1 \pmod{p}.$$

A convenient equivalent form of Fermat's little theorem is: If p is a prime and a is *any* positive integer, then

$$a^p \equiv a \pmod{p}.$$

Problem A positive integer n is called *groovy* if, for every positive integer a, n^2 divides $a^n - 1$ whenever n divides $a^n - 1$.

Show that all primes are groovy.

Solution Let p be a prime and suppose that p divides $a^p - 1$. So we have

$$a^p \equiv 1 \pmod{p},$$

whereas Fermat's little theorem asserts that

$$a^p \equiv a \pmod{p}.$$

Putting together these two pieces of information, we find that

$$a \equiv 1 \pmod{p}$$

and we can write $a = mp + 1$ for some integer m. Therefore, by the binomial theorem, we have

$$a^p - 1 = (mp+1)^p - 1 = \sum_{k=0}^{p} \binom{p}{k} m^k p^k - 1.$$

If we consider this expression modulo p^2, then each term with $k \geq 2$ disappears, due to the appearance of p^k. Therefore, we conclude that

$$a^p - 1 \equiv \binom{p}{0} m^0 p^0 + \binom{p}{1} m^1 p^1 - 1 \equiv 1 + mp^2 - 1 \equiv 0 \pmod{p^2},$$

and that every prime is indeed groovy. □

Sometimes you're not working modulo a prime, in which case Fermat's little theorem just isn't good enough. Luckily Euler's theorem, which generalises Fermat's little theorem, comes to the rescue!

Euler's theorem If n is a positive integer and $\gcd(a, n) = 1$, then

$$a^{\varphi(n)} \equiv 1 \pmod{n}.$$

Here, $\varphi(n)$ denotes the number of integers $1 \leq k \leq n$ such that $\gcd(k, n) = 1$ and is usually referred to as the *Euler phi function*.[8] For example, $\varphi(12) = 4$, since the four numbers 1, 5, 7 and 11 are relatively prime to 12. We'll use Euler's theorem to provide a quick proof to a problem we previously encountered in section 2.2.

Problem Show that for every positive integer n not divisible by 2 or 5, there exists a multiple of n all of whose digits are ones.

Solution Consider the number $X = \frac{1}{9}(10^{\varphi(9n)} - 1)$. It's an integer whose decimal representation consists entirely of ones, specifically $\varphi(9n)$ ones, to be precise. However, since $\gcd(10, 9n) = \gcd(10, n) = 1$ by assumption, we can invoke Euler's theorem to show that

$$10^{\varphi(9n)} \equiv 1 \pmod{9n} \quad \Rightarrow \quad \frac{1}{9}(10^{\varphi(9n)} - 1) \equiv 0 \pmod{n}.$$

Therefore, X is divisible by n, as desired. □

Fermat's little theorem and Euler's theorem are most successfully applied in conjunction with the following useful result, which is worth a section on its own.

[8] There is a formula for $\varphi(n)$ in terms of the prime factorisation of n. See if you can find and prove it!

2.13 The gcd trick

The gcd trick If $a^x \equiv 1 \pmod{n}$ and $a^y \equiv 1 \pmod{n}$, then

$$a^{\gcd(x,y)} \equiv 1 \pmod{n}.$$

This useful result can be proven using the representation of $\gcd(x, y)$ as a linear combination of x and y using Euclid's algorithm. See section 2.7. The proof is not hard and you should try to work it out for yourself.

Problem Prove that if p and q are primes satisfying $q \mid 2^p - 1$, then $p < q$.

Solution We can write $q \mid 2^p - 1$ equivalently as $2^p \equiv 1 \pmod{q}$. But Fermat's little theorem tells us that $2^{q-1} \equiv 1 \pmod{q}$. So we can apply the gcd trick to obtain

$$2^{\gcd(p,q-1)} \equiv 1 \pmod{q}.$$

Since p is prime, we know that $\gcd(p, q-1)$ must be equal to 1 or p. The former case is absurd, since it would imply that $1 \equiv 2^{\gcd(p,q-1)} \equiv 2 \pmod{q}$.

We are left to ponder the latter case, when $\gcd(p, q-1) = p$. But this simply means that $p \mid q - 1$, which is only possible if $p < q$. □

The previous problem provides us with a beautiful proof that there are infinitely many primes. This is because if there were only finitely many, we could take p to be the largest one. Then $2^p - 1$ would be divisible by some prime q, which would have to be larger than p. Since this contradicts the fact that p was the largest prime, we conclude that there must be infinitely many primes.

2.14 Existence of a generator

Suppose we consider a modulus, say modulo 7, and a number, say 2, and compute powers of 2 modulo 7,

$$2^0, 2^1, 2^2, 2^3, \ldots.$$

We end up in a cycle

$$1, 2, 4, 1, 2, 4, \ldots$$

of three elements.

Yet if we chose to compute powers of 3 modulo 7 instead, we would end up with

$$1, 3, 2, 6, 4, 5, 1, 3, 2, 6, 4, 5, \ldots,$$

which is a 6-cycle.

The number 3 is called a generator modulo 7 because its successive powers generate everything which is relatively prime to 7 modulo 7.

More generally, a number a is called a *generator*[9] modulo n if the powers of a, namely,

$$1, a, a^2, a^3, \ldots,$$

[9]The term *primitive root* modulo n is used in some literature. It means the same thing.

generate the complete set of elements modulo n which are relatively prime to n.

Existence of a generator There exists a generator modulo n if and only if $n = 1$, 2, 4, p^k, or $2p^k$, where p is an odd prime and k is a positive integer.

This theorem is most often used in the case where n is a prime. It can be used very powerfully in conjunction with the gcd trick. The important point is that if g is a generator modulo n, then $d = \varphi(n)$ is the smallest positive integer for which $g^d \equiv 1 \pmod{n}$.

Problem A number n is called a *Carmichael number*[10] if it is composite and

$$a^n \equiv a \pmod{n}$$

for all positive integers a.

Prove that there are no Carmichael numbers which are the product of two distinct primes.

Solution Suppose $n = pq$ is a Carmichael number, where p and q are distinct primes. If a is any number coprime to n, then we know $a^{pq-1} \equiv 1 \pmod{pq}$, from which it follows that

$$a^{pq-1} \equiv 1 \pmod{p}.$$

By Fermat we know that

$$a^{p-1} \equiv 1 \pmod{p}.$$

Using the gcd trick we find that

$$a^d \equiv 1 \pmod{p},$$

where $d = \gcd(pq - 1, p - 1) = \gcd(q - 1, p - 1)$.

This is not particularly helpful until we realise that we may choose $a = g$ to be a generator modulo p. Since $p - 1$ is the smallest positive exponent e for which $g^e \equiv 1 \pmod{p}$, this means that $\gcd(q - 1, p - 1) = p - 1$. Thus $p - 1 \mid q - 1$.

A similar argument modulo q shows that $q - 1 \mid p - 1$. Thus $p = q$, in contradiction to $p \neq q$. □

For a long time it was suspected that there are infinitely many Carmichael numbers. A proof of this fact was finally found in 1994.

[10] These numbers are of interest because in the light of Fermat's little theorem, they look prime, but are not!

3

Diophantine equations

A *Diophantine equation* is an equation that asks for integer solutions. The equation is usually posed as, or at least can be reduced to, a polynomial expression in several variables. For example, $a^2 + b^3c = d^4$ is a Diophantine equation.

3.0 Problems

1. Show that for every integer $n > 1$ it is possible to write $\frac{1}{n}$ as a sum of two reciprocals of distinct positive integers.

2. Find all pairs of integers (x, y) such that

$$3x + 4y = 2xy.$$

3. Find all integer solutions to the equation

$$(x - y)^2 = x + y.$$

4. Find all ordered triples of positive integers such that each of them is a factor of the sum of the other two.

5. Show that if $u^2 + uv + v^2$ is divisible by 9, where u and v are integers, then so is uv.

6. Find all integers x and y such that

$$x^3y + x + y = xy + 2xy^2.$$

7. Let a and b be integers.

   (a) Prove that $13 \mid 2a + 3b$ if and only if $13 \mid 3a - 2b$.

   (b) Prove that $13 \mid a^2 + b^2$ if and only if $13 \mid 2a + 3b$ or $13 \mid 3a + 2b$.

8. Suppose that u, v are integers such that uv is not divisible by 7.

   (a) Prove that $u^2 + uv + 2v^2$ is not divisible by 49.

   (b) Prove that if $u^2 + uv + 2v^2$ is divisible by 7, then so is $u^2 + 2uv - v^2$.

(c) Find all pairs of integers (u, v) such that $u^2 + 2uv - v^2$ is divisible by 7 but $u^2 + uv + 2v^2$ is not divisible by 7.

9. Suppose that a and b are positive integers such that

$$11^{2011} \mid a^2 + b^2.$$

Prove that

$$11^{2012} \mid ab.$$

10. Find all positive integers of the form

$$k = \frac{n^2 - 29}{3n + 11},$$

where n is also an integer.

11. Determine all positive integers k such that for all positive integers a and b the following statement is true.

$$7a + 8b \text{ is divisible by } k \text{ if and only if } 8a + 7b \text{ is divisible by } k.$$

12. Find all positive integers a, b and c such that

$$ab + bc + ca = 2 + abc.$$

What if we drop the restriction that a, b and c are all positive?

13. Can the product of four consecutive positive integers be a perfect square?

14. Show that

$$x^4 + 131y^4 = 3z^4$$

has no solutions in positive integers.

15. Find all triples of positive integers (a, b, c) such that

$$a \mid bc - 1, \quad b \mid ac - 1 \quad \text{and} \quad c \mid ab - 1.$$

16. Determine all pairs (x, y) of positive integers such that

$$y^2(x - 1) = x^5 - 1.$$

17. Find all integers a, b and c with $1 < a < b < c$ such that

$$(a - 1)(b - 1)(c - 1)$$

is a divisor of $abc - 1$.

18. Suppose that a, b, c and d are positive integers satisfying $ab = cd$.
Prove that neither $a + b + c + d$ nor $a^2 + b^2 + c^2 + d^2$ can be prime numbers.

19. Find all integer solutions to the equation

$$8x^3 - 4 = y(6x - y^2).$$

20. Can the product of five consecutive positive integers be a perfect square?

21. Prove that the equation
$$6(6a^2 + 3b^2 + c^2) = 5n^2$$
has no solutions in integers apart from $a = b = c = n = 0$.

22. Prove that if p is prime and x is an integer, then every factor of the expression
$$x^{p-1} + x^{p-2} + \cdots + x + 1$$
is congruent to 0 or 1 modulo p.

23. Find all ordered triples of positive integers (a, b, c) such that for all positive integers t
$$\frac{(at+1)(bt+1)(ct+1) - 1}{\operatorname{lcm}(at, bt, ct)}$$
is also a positive integer.

24. Given any set $A = \{a_1, a_2, a_3, a_4\}$ of four distinct positive integers, we denote the sum $a_1 + a_2 + a_3 + a_4$ by s_A. Let n_A denote the number of pairs (i, j) with $1 \le i < j \le 4$ for which $a_i + a_j$ divides s_A.

    Find all sets A of four distinct positive integers which achieve the largest possible value of n_A.

25. Find all pairs (a, b) of positive integers such that
$$\frac{a^2b + a + b}{ab^2 + b + 7}$$
is also a positive integer.

26. The positive integers a and b are such that
$$15a + 16b \quad \text{and} \quad 16a - 15b$$
are both squares of positive integers.

    What is the smallest possible value of the smaller of the two squares?

27. Find all pairs of integers (a, b) such that
$$a^5 + a^3b + b^4 = 0.$$

28. Determine all pairs of positive integers (a, b) such that
$$\frac{a^2}{2ab^2 - b^3 + 1}$$
is a positive integer.

29. Find the smallest integer $n > 1$ such that
$$\frac{1}{n}(1^2 + 2^2 + \cdots + n^2)$$
is a perfect square.

30. Find all pairs of integers (a, b) such that
$$a^2b^2 - 4a - 4b$$
is a perfect square.

31. Find all pairs of integers (a, b) such that

$$a^2 + 4b \quad \text{and} \quad b^2 + 4a$$

are both perfect squares.

32. Determine the maximum value of $m^2 + n^2$, where m and n are integers $1 \leq m, n \leq 2015$ and

$$(n^2 - mn - m^2)^2 = 1.$$

33. Let a and b be positive integers. Show that if

$$\frac{a^2 + b^2}{ab + 1}$$

is a positive integer, then it is a perfect square.

34. Find all positive integers k of the form

$$k = \frac{x+1}{y} + \frac{y+1}{x},$$

where x and y are also positive integers.

35. Let a and b be positive integers such that

$$\frac{(4a^2 - 1)^2}{4ab - 1}$$

is also a positive integer.

Prove that $a = b$.

36. Find all integers x, y such that

$$y^2 - x^3 = 7.$$

37. Find all integer solutions of the equation

$$\frac{x^7 - 1}{x - 1} = y^5 - 1.$$

3.1 Factorisation

If someone asked you to solve the equation $xy = 100$, where x and y are integers, then hopefully you'd find the problem easy. That's because x must be a divisor of 100. After listing out all of the divisors of 100—remembering, of course, that they may be positive or negative—we find the values for x and then solve to find the corresponding values for $y = \frac{100}{x}$.

It's often possible to rearrange a Diophantine equation so that we have an integer on one side and a factored expression on the other side. It is then a simple matter of examining every possible factorisation of the integer and lining it up with the factored expression.

Problem Find all pairs of integers (b, c) such that

$$2b + 3c = 5bc.$$

Solution Perhaps you hoped to write $5bc - 2b - 3c$ in the form

$$5bc - 2b - 3c = (5b + *)(c + *) + *,$$

or something similar, where each asterisk is an integer. Unfortunately this is not possible. The '5' gets in the way. We can remove this difficulty by multiplying the whole equation by 5. In this case we quickly find that

$$(5b - 3)(5c - 2) = 6.$$

It only remains to check the pairs of integers which multiply to give 6. We check each of $(5b - 3, 5c - 2) = (-6, -1)$, $(-3, -2)$, $(-2, -3)$, $(-1, -6)$, $(1, 6)$, $(2, 3)$, $(3, 2)$ and $(6, 1)$. Of these only $(-3, -2)$ and $(2, 3)$ result in b and c being integers. These pairs correspond to the solutions $(b, c) = (0, 0)$ or $(b, c) = (1, 1)$. □

3.2 Monotonicity

Suppose, for example, we know that x is a perfect square. Suppose also we can show that $a^2 < x < (a + 3)^2$ by other means. Then we may conclude that the only possibilities we need pursue further are $x = (a + 1)^2$ and $x = (a + 2)^2$. This is because the function $f(t) = t^2$ is an increasing function on the positive integers. This sort of technique is valid for any monotonic (i.e. increasing or decreasing) function, as in the following example.

Problem Find all pairs of integers (x, y) such that

$$(x + 2)^4 - x^4 = y^3.$$

Solution Rewrite the equation as

$$y^3 = 8x^3 + 24x^2 + 32x + 16.$$

Thus y is even and we may write $y = 2z$ for some integer z. The equation now becomes

$$z^3 = x^3 + 3x^2 + 4x + 2.$$

Note that $x^3 + 3x^2 + 4x + 2$ is very close to $(x + 1)^3$.

It can be shown (and you should do this!) that

$$x^3 < x^3 + 3x^2 + 4x + 2 < (x + 2)^3$$

for all real numbers x. Thus if there is a solution, it must be $z = x + 1$. Substituting this in for z yields $x = -1$, and then $y = 0$. □

3.3 Bounding arguments

It often occurs that by examining the orderings of the variables you can bound one of them. What is left is usually a case bash. If the equation is symmetric, you can use a WLOG ('without loss of generality') argument.

Problem Find all ways of writing the number 1 as a sum of the reciprocals of three positive integers.

Solution The equation $\frac{1}{a} + \frac{1}{b} + \frac{1}{c} = 1$ is equivalent to the equation

$$abc = ab + ac + bc.$$

Note that none of a, b, c equal 1. Due to symmetry, we may order the variables and say WLOG that $2 \leq a \leq b \leq c$. Using this we note that the RHS $\leq 3bc$ and so $abc \leq 3bc$. Thus $a \leq 3$. We may now take the two cases $a = 2$ and $a = 3$.

- Case 1: $a = 3$

  Then we have $3b + 3c = 2bc$. This may be rewritten as

  $$(2b - 3)(2c - 3) = 9.$$

  As $c \geq b \geq 2$, both factors are positive and thus $2b - 3 = 1$, 3 or 9. Remembering that we had $2 \leq a \leq b \leq c$, we quickly find that only $b = 3$ is possible, which leads to the solution $(a, b, c) = (3, 3, 3)$.

- Case 2: $a = 2$

  In a similar way to case 1 we arrive at

  $$(b - 2)(c - 2) = 4.$$

  Remembering that $2 \leq b \leq c$ we find that only $b = 3$ and $b = 4$ are possible, which lead quickly to the solutions $(a, b, c) = (2, 3, 6)$ and $(a, b, c) = (2, 4, 4)$.

In summary we have solutions $(3, 3, 3)$, $(2, 3, 6)$ and $(2, 4, 4)$. However, remember that we used a WLOG argument at the beginning due to symmetry. So we must remember to include all permutations of our solutions too, giving us 10 distinct solutions in all. □

Problem Find all triples (a, b, c) of positive integers such that

$$a^2 + b + c = abc.$$

Solution Rearrange to get

$$b + c = a(bc - a).$$

We suspect that the RHS should dominate as the variables get larger, especially as b and c get larger. Note that if $a \geq bc$, then there are no solutions because the RHS is not positive but the LHS is.

Consider the graph of the parabola $f(x) = x(bc - x)$. It is an upside-down parabola that crosses the x-axis at $x = 0$ and $x = bc$. Thus if f is restricted to the interval $1 \leq x \leq bc - 1$, we see that f achieves its minimum value simultaneously at $x = 1$ and at $x = bc - 1$, the common

minimum value being $f(1) = bc - 1$. Since $1 \leq a \leq bc - 1$, we have $b + c = f(a) \geq f(1)$. Thus we must have

$$bc - 1 \leq b + c.$$

This certainly is a candidate for bounding! WLOG $c \geq b$. If $b \geq 3$, then $bc \geq c+c+c > b+c+1$. So we must have either $b = 1$ or $b = 2$.

- Case 1: $b = 1$

  The original equation can be rearranged to

  $$c = \frac{a^2 + 1}{a - 1} = a + 1 + \frac{2}{a - 1}.$$

  Since $\frac{2}{a-1}$ must be an integer, this leads to $a = 2$ or $a = 3$. The solutions corresponding to this case are then $(a, b, c) = (2, 1, 5)$ and $(3, 1, 5)$.

- Case 2: $b = 2$

  The original equation can be rearranged to

  $$c = \frac{a^2 + 2}{2a - 1}.$$

  Thus

  $$4c = \frac{(2a)^2 + 8}{2a - 1} = 2a + 1 + \frac{9}{2a - 1}.$$

  Hence $2a - 1 \mid 9$, which leads to $a = 1, 2, 5$. The solutions corresponding to this case are $(a, b, c) = (1, 2, 3)$, $(2, 2, 2)$ and $(5, 2, 3)$. □

If you are wondering how we mysteriously came up with the idea of going from $c = \frac{a^2+2}{2a-1}$ to $4c = \frac{(2a)^2+8}{2a-1} = 2a + 1 + \frac{9}{2a-1}$, please read on because section 3.4 will provide illumination!

3.4 Polynomial modulus

If we discover that one polynomial expression must be divisible by another, the use of modular arithmetic for polynomials can be very helpful.

Problem Find all integers a such that $a^2 + 2$ is divisible by $2a - 1$.

Solution Using a polynomial modulus we wish to solve $a^2 + 2 \equiv 0 \pmod{2a - 1}$.

$$\begin{aligned} & a^2 + 2 \equiv 0 && \pmod{2a - 1} \\ \Leftrightarrow\ & (2a)^2 + 8 \equiv 0 && \pmod{2a - 1} \\ \Leftrightarrow\ & (1)^2 + 8 \equiv 0 && \pmod{2a - 1} \end{aligned}$$

Thus $2a - 1 \mid 9$. Checking the six factors of 9, including the negative ones, gives rise to six solutions, namely, $a = -4$, -1, 0, 1, 2 or 5. □

3.5 Quadratic discriminants

Many Diophantine equations can be seen through quadratic eyes if looked at in the right way. If such an equation can be written as a quadratic in one of its variables, then we know that for an integer solution to exist, the discriminant of the quadratic must be a perfect square.

Problem Find all integers a such that $a^2 + 2$ is divisible by $2a + 3$.

You would be right in thinking that this would be a prime candidate for trying the method of using a polynomial modulus. And yes, it works just fine. However, study the following solution which showcases the quadratic discriminant method.

Solution We want the equation

$$a^2 + 2 = 2ak + 3k$$

to have integer solutions. Viewing this as a quadratic in a, this is only possible if the discriminant $4k^2 + 12k - 8$ is a perfect square. Thus

$$4k^2 + 12k - 8 = u^2,$$

for some non-negative integer u. Note that u^2 is even and so u is also. Write $u = 2v$. Rearranging the above relation yields

$$k^2 + 3k - 2 - v^2 = 0.$$

This is a quadratic in k. Once again its discriminant $9 + 4(2 + v^2)$ is a perfect square. Thus

$$4v^2 + 17 = w^2,$$

for some non-negative integer w. This may be rewritten as

$$(w - 2v)(w + 2v) = 17.$$

Since $w + 2v \geq 0$, only the positive factors of 17 need to be checked. The only possibility is $w - 2v = 1$ and $w + 2v = 17$. This yields $w = 9$ and $v = 4$. Thus $u = 8$. Solving for k yields $k = 3$ or -6. Finally solving for a yields the four solutions $a = -10$, -2, -1 or 7. □

3.6 Modular arithmetic

Suppose you want to show that a certain number divides a variable. Or you'd like to show that there are no integer solutions to an equation. Modular arithmetic might be just the tool you need! What is a good modulus to try? Anything that significantly simplifies the situation. For instance, if a nasty coefficient is divisible by p, then it may be useful to try modulo p. If there are mth powers involved, then since $a^{\varphi(n)} \equiv 1 \pmod{n}$ whenever $\gcd(a, n) = 1$ from Euler's theorem, it is good to try working modulo n, where $\varphi(n) = m$. If this does not work, try to choose n such that $m \mid \varphi(n)$ or at least so that m and $\varphi(n)$ have factors in common. Squares work well modulo 3, 4, 5, 8 and 16. Cubes work well modulo 7, 9 and 13.

Problem Show that the equation

$$x^4 + 131y^4 = 3z^4 + 2000$$

has no integer solutions.

Solution Since $\varphi(5) = 4$, we try considering the equation modulo 5. The equation simplifies as

$$x^4 + y^4 \equiv 3z^4 \pmod 5.$$

Observe that $a^4 \equiv 0, 1 \pmod 5$ for any integer a. It follows that LHS $\equiv 0$, 1 or 2 (mod 5) and RHS $\equiv 0$ or 3 (mod 5). So we must have LHS $\equiv$ RHS $\equiv 0$ (mod 5). But this is only possible if $x \equiv y \equiv z \equiv 0 \pmod 5$. However, this in turn implies that $5^4 \mid 2000$, which is a contradiction. □

Problem Show that the equation

$$5m^2 - 6mn + 7n^2 = 1985$$

has no integer solutions.

Solution Viewing the equation as a quadratic in m, if it did have solutions, then the discriminant

$$36n^2 - 20(7n^2 - 1985) = 4(9925 - 26n^2)$$

should be a perfect square. Thus

$$9925 - 26n^2$$

is a perfect square.

Squares can only be congruent to 0, 1 or 4 modulo 8. Hence considering this expression modulo 8 yields

$$\begin{aligned} 5 - 2n^2 &\equiv 0, 1, 4 && \pmod 8 \\ &\equiv 1 && \pmod 8 \quad \text{(since LHS is odd)} \\ \Rightarrow \quad 2n^2 &\equiv 4 && \pmod 8 \\ \Rightarrow \quad n^2 &\equiv 2 && \pmod 4. \end{aligned}$$

This is impossible because squares are congruent to 0 or 1 modulo 4. □

3.7 Divisibility and gcds

Considering the greatest common divisor (gcd) of a pair or triple of numbers can help.

Problem Find all pairs of integers (x, y) such that

$$x^4 + 2x^2y + y^3 = 0.$$

Solution Let $\gcd(x, y) = d$. Thus we may write $x = sd$ and $y = td$, where $\gcd(s, t) = 1$. The equation becomes

$$s^4d + 2s^2t + t^3 = 0.$$

Now if p is any prime factor of s, we see that $p \mid t^3$ and thus $p \mid t$. But since $\gcd(s, t) = 1$, this forces $s = \pm 1$ and thus $d = -2t - t^3$. Thus the complete set of solutions is given by

$$\begin{aligned} x &= \pm t(2 + t^2) \\ y &= -t^2(2 + t^2), \end{aligned}$$

for any integer t. We can verify these solutions by direct substitution. □

Problem Find all positive integral solutions to the equation

$$a^2 + 2b^2 = c^2.$$

Solution By dividing out by $\lambda = \gcd(a, b, c)$, we may assume that a, b and c are pairwise coprime. Considering the equation modulo 4 shows us that a and c are both odd and that b is even. Thus after writing $b = 2d$ and rearranging we have

$$2d^2 = \frac{c-a}{2} \cdot \frac{c+a}{2}$$

and that $\gcd(\frac{c-a}{2}, \frac{c+a}{2}) = \gcd(a, c) = 1$. Thus whichever one of $\frac{c-a}{2}$ and $\frac{c+a}{2}$ is odd must be a perfect square, say u^2, and the other must be twice a square, say $2v^2$. Adding these together yields

$$c = u^2 + 2v^2, \quad a = \pm(u^2 - 2v^2) \quad \text{and} \quad d = uv.$$

Thus the most general solutions are given by

$$\begin{aligned} a &= |u^2 - 2v^2|\lambda \\ b &= 2uv\lambda \\ c &= (u^2 + 2v^2)\lambda, \end{aligned}$$

where λ is any positive integer and (u, v) are any pair of relatively prime positive integers.[1] □

Note that the solutions are valid even if $\gcd(u, v) > 1$.

3.8 Reduction of variables

Some problems do not ask for all solutions but only ask for the existence of an infinite number of solutions. Reducing the number of variables by some well thought out substitutions can often simplify the matter. The idea is to make as many things cancel out as possible, while still leaving scope for infinitely many solutions.

Problem Show that the equation

$$a^3 + b^3 = c^3 + d^3 + e^3$$

has infinitely many integer solutions where $\gcd(a, b, c, d, e) = 1$ and a, b, c, d and e are all distinct.

Solution We try the substitution $a = x - s$, $b = x + s$. The left-hand side becomes $2x^3 + 6s^2x$. A little more thought encourages us to try a similar thing on the right-hand side, namely $c = x - t$, and $d = x + t$. Using these, the equation simplifies to

$$6x(s - t)(s + t) = e^3.$$

We have now reduced the equation to four variables. Trying $s = 2$ and $t = 1$ simplifies matters even further to

$$18x = e^3.$$

[1] A very similar approach yields a formula for all Pythagorean triples, that is, triples of positive integers that are the side lengths of a right-angled triangle.

Choosing $e = 18y$ finishes off the problem. We have discovered an infinite one-parameter family of solutions given by

$$\begin{aligned} a &= 18^2y^3 - 2 \\ b &= 18^2y^3 + 2 \\ c &= 18^2y^3 - 1 \\ d &= 18^2y^3 + 1 \\ e &= 18y. \end{aligned}$$

It is clear that they are distinct with $\gcd(a, b, c, d, e) = 1$. □

3.9 Infinite descent

The following statement may seem pretty obvious: It is impossible to have an infinitely descending sequence of positive integers. This is known as the principle of *infinite descent*. This principle is equivalent to the fact that any set of positive integers has a smallest member.

The principle of infinite descent is particularly useful for proving that a Diophantine equation has no solutions. For instance, suppose we could prove that given any solution to a Diophantine equation there is another solution which is even smaller (but still positive). Then if we applied this to a supposed *minimal* solution, we would have a contradiction.

By minimal we mean that we can measure the size of the solution in some sense: for example, the sum of the absolute values of the variables, or perhaps, the absolute value of a particular variable.

Problem Find all integral solutions to

$$x^3 + 3y^3 + 9z^3 = 9xyz.$$

Solution Assume we have a solution (x, y, z) with $|x|+|y|+|z| > 0$. Then there is a solution with $|x| + |y| + |z|$ minimal. The equation implies $3 \mid x^3$ and so $3 \mid x$. Thus $x = 3a$ for some integer a. Substitute this into the equation and divide the equation by 3 to find

$$9a^3 + y^3 + 3z^3 = 9ayz.$$

In a similar way we obtain $3 \mid y$ so that $y = 3b$ for some integer b and so we deduce

$$3a^3 + 9b^3 + z^3 = 9abz.$$

From this we obtain $3 \mid z$ so that $z = 3c$ for some integer c and so we deduce

$$a^3 + 3b^3 + 9c^3 = 9abc.$$

It only remains to note that (a, b, c) is also a solution to the original equation and satisfies $0 < |a| + |b| + |c| = \frac{|x|+|y|+|z|}{3}$. This contradicts the minimality of $|x| + |y| + |z|$. Thus there is no solution with $|x| + |y| + |z| > 0$. Hence the only solution is $x = y = z = 0$. □

3.10 Vieta jumping

Suppose that we have a Diophantine equation in two variables x and y which is a monic[2] quadratic in x. If we have one solution $(x, y) = (a, b)$ in integers, it turns out that we can apply Vieta's formula[3] for the sum of the roots to generate another integer solution.

Problem Show that the Diophantine equation

$$x^2 + y^2 - 4xy - 4 = 0$$

has infinitely many solutions.

Solution We try $x = 0$. This quickly yields $y = 2$ is a solution. So $(0, 2)$ is a solution.

Now set $y = 2$ into the equation. We obtain the quadratic equation

$$x^2 - 8x = 0.$$

Using Vieta's formula, the sum of the roots is 8. Since $x = 0$ is one of the roots, the other must be $x = 8$. So $(8, 2)$ is a solution. By symmetry, $(2, 8)$ is also a solution.

Next substitute $y = 8$ into the original equation. We obtain the quadratic equation

$$x^2 - 32x + 60 = 0.$$

Again by Vieta's formula, the sum of the roots is 32. Since $x = 2$ is one of the roots, the other must be $x = 30$. So $(30, 8)$ is a solution. By symmetry, so is $(8, 30)$.

This procedure may be continued indefinitely. In general if $(x, y) = (a, b)$ is a solution, we may construct another solution by considering the quadratic equation

$$x^2 - 4bx + b^2 - 4 = 0.$$

From Vieta's formula, the sum of the roots is $4b$. Since $x = a$ is one of the roots, the other must be $4b - a$. Thus $(x, y) = (4b - a, b)$ is also a solution and by symmetry, so is $(x, y) = (b, 4b - a)$.

If we assume that $0 \leq a < b$, then $b < 4b - a$. Thus any solution (a, b) leads to a bigger solution $(b, 4b - a)$. Since $(0, 2)$ is a solution, we can thus construct infinitely many solutions using this procedure starting from $(0, 2)$. □

By using the technique of infinite descent, it is possible to prove that all integer solutions (x, y), with $0 \leq x \leq y$, to the given equation occur as a consecutive pair of the generated sequence $0, 2, 8, 30, \ldots$, where the rule for the sequence is

$$x_0 = 0, \quad x_1 = 2 \quad \text{and} \quad x_{n+2} = 4x_{n+1} - x_n \quad \text{for } n = 0, 1, 2 \ldots.$$

See if you can prove this.

[2] A polynomial of degree n in x is *monic* if the coefficient of x^n is equal to 1.
[3] See section 9.4 if you don't know about Vieta's formulas.

3.11 Cyclotomic recognition

The cyclotomic polynomials $\Phi_1(x), \Phi_2(x), \Phi_3(x), \ldots$ are defined recursively by the relations

$$x^n - 1 = \prod_{d|n} \Phi_d(x)$$

for $n \geq 1$.

From this definition it is easy to compute the first few cyclotomic polynomials to be

$$\begin{aligned}
\Phi_1(x) &= x - 1 \\
\Phi_2(x) &= x + 1 \\
\Phi_3(x) &= x^2 + x + 1 \\
\Phi_4(x) &= x^2 + 1 \\
\Phi_5(x) &= x^4 + x^3 + x^2 + x + 1 \\
\Phi_6(x) &= x^2 - x + 1 \\
\Phi_7(x) &= x^6 + x^5 + x^4 + x^3 + x^2 + x + 1 \\
\Phi_8(x) &= x^4 + 1.
\end{aligned}$$

Additionally, it is not too hard to show directly from the definition that

$$\Phi_p(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$$

for any prime p.

Sometimes part of an algebraic expression can be seen to be part of a cyclotomic polynomial. Since any cyclotomic polynomial is a factor of an expression of the form $x^n - 1$, there is a fair chance that the theorems of Euler or Fermat from section 2.12 might be helpful.

Problem Prove that the equation

$$x^3 = y^{16} + y^{15} + \cdots + y^2 + y + 9$$

has no solutions in positive integers.

Solution We astutely recognise that the RHS is equal to 8 plus the cyclotomic polynomial $\Phi_{17}(y) = y^{16} + y^{15} + \cdots + y^2 + y + 1$. Rewrite the equation as

$$x^3 - 8 = y^{16} + y^{15} + \cdots + y^2 + y + 1.$$

Note that the RHS is a factor of $y^{17} - 1$.

Let p be any prime factor of the RHS. Note that this implies that $p \nmid y$. Then it follows that

$$y^{17} \equiv 1 \pmod{p}.$$

However, since $p \nmid y$, we also know from Fermat's little theorem that

$$y^{p-1} \equiv 1 \pmod{p}.$$

Using the gcd trick from section 2.13 we deduce that

$$y^d \equiv 1 \pmod{p},$$

where $d = \gcd(17, p - 1)$.

Thus there are two possibilities for d, namely, $d = 1$ or $d = 17$.

- Case 1: $d = 1$

  Then we have $y \equiv 1 \pmod{p}$, and so,

  $$y^{16} + y^{15} + \cdots + y^2 + y + 1 \equiv 17 \pmod{p}.$$

  Thus $p \mid 17$ and therefore, $p = 17$.

- Case 2: $d = 17$

  Then $17 \mid p - 1$ from which it follows that $p \equiv 1 \pmod{17}$.

We have shown that every prime factor p of the RHS satisfies either $p = 17$ or $p \equiv 1 \pmod{17}$. Hence it follows that *every* factor of the RHS is congruent to 0 or 1 modulo 17.

However, since

$$x^3 - 8 = (x - 2)(x^2 + 2x + 4)$$

we see that $x - 2 \equiv 0, 1 \pmod{17}$.

But if $x \equiv 2 \pmod{17}$, then $x^2 + 2x + 4 \equiv 12 \pmod{17}$.

If $x \equiv 3 \pmod{17}$, then $x^2 + 2x + 4 \equiv 2 \pmod{17}$.

In both cases the other factor is not congruent to 0 or 1 modulo 17, a contradiction. Thus the equation has no integer solutions. □

4

Plane geometry

Welcome dear reader to a domain where many fear to tread, and many stumble. The domain is plane geometry, surely one of the most beautiful areas of mathematics.

In one sense, geometry is the easiest mathematical topic to learn for it is the epitome of logical rigour. But in another sense, geometry is the most difficult subject to teach. It relies heavily on ingenuity. Solving a geometry problem is more often dependent on noticing something that is difficult to see, perhaps more so than elsewhere. But perseverance is an excellent character trait for problem solving.

We shall assume you know all the basic theorems of plane geometry: congruent triangles, similar triangles, basic circle theorems, cyclic quadrilaterals and special points of a triangle.

There are a number of pitfalls unique to geometry, into which many an inexperienced problem-solver falls. They are incredibly easy to avoid but far too many fail to avoid them, because they think these matters are not important.

- **Your chances of solving the problem are as good as your diagram.** If you don't draw a good diagram in a geometry problem, unless it is very, very easy, you have significantly lowered your chances of solving it.
- **Size matters.** We admire the responsible sentiment of those who don't want to use up too much paper. But with a big accurate diagram, you're likely to see things more clearly and solve the problem. With a dozen small terrible diagrams you're not going to get anywhere. A good diagram should take up most of the page.
- **Many-coloured diagrams stimulate the mind.** It is prettier to draw colourful diagrams. Consider that almost all of your work in mathematics will be done in vast swathes of monochromatic grey or black or blue. Wouldn't you appreciate a change of colour occasionally? A multicoloured diagram is often a better diagram, helping you to solve the problem. Drawing different parts or aspects of the diagram in different colours helps you to see these aspects independently, and see relationships between them. Unfortunately for technical reasons this book is black and white with a little blue and grey. But you do not have such excuses!
- **Accuracy is a virtue.** If you draw freehand, you draw sloppily. If you draw accurately, using ruler and compass, your diagram will be spot on the mark. And with a perfect diagram, you can see very concretely the *real* situation the problem is asking about. You can spot things, make guesses, get a feel for how all the points and lines relate, and

so on. Does one of your angles look like a right angle? Does one of your quadrilaterals look cyclic? Do those three points look collinear? Your accurate diagram will give you all sorts of suggestions, if you look hard enough at it, and you can try and prove them.

4.0 Problems

1. Two parallel lines are tangent to a circle with centre O. A third line, also tangent to the circle, meets the two parallel lines at A and B.

   Prove that AO is perpendicular to OB.

2. Suppose that two circles are externally tangent at P. Let a common tangent touch the circles at A and B.

   Prove that triangle APB is right-angled.

3. Let ABC be a triangle with incentre I. Suppose that X is the midpoint of the arc BC not containing A on the circumcircle of triangle ABC.

   Prove that X is the circumcentre of triangle BIC.

4. Let D, E, F be points on sides AB, BC, CA of triangle ABC such that $DE = BE$ and $FE = CE$.

   Prove that the circumcentre of triangle ADF lies on the bisector of $\angle DEF$.

5. Given a triangle ABC, let the median from vertex A intersect the circumcircle of the triangle again at K. A circle Γ passes through A and B such that BC is tangent to Γ. Let L be the intersection of AK and Γ different from A.

   Prove that $BLCK$ is a parallelogram.

6. Let two circles intersect at A and B. Suppose that a common tangent to the two circles meets them at P and Q.

   If the line AB meets PQ at M, show that M is the midpoint of PQ.

7. ABC is a triangle, right-angled at C. The internal angle bisectors of angle BAC and angle ABC meet BC and CA at P and Q, respectively. Let M and N be the feet of the perpendiculars from P and Q to AB, respectively.

   Find the size of $\angle MCN$.

8. Two circles Γ_1 and Γ_2 intersect at A and B. It is known that Γ_2 passes through point O, where O is the centre of Γ_1. Point X lies on the arc AOB, and AX intersects Γ_1 again at Y.

   Prove that $XY = XB$.

9. Let AD be an altitude and H be the orthocentre of $\triangle ABC$. Let line AD meet the circumcircle of $\triangle ABC$ at X.

   (a) Prove that $HD = DX$.

   (b) Prove that the circumradius of the triangles formed by H and any two vertices is equal to the circumradius of $\triangle ABC$.

10. Two circles of equal radius intersect at A and B. The point C is on one of the circles such that B is the midpoint of the arc AC.

    Prove that AC is tangent to the other circle.

11. Two circles C_1 and C_2 intersect at A and B. Let P be a point on C_1 and Q be a point on C_2, so that P and Q lie on opposite sides of the line AB. Suppose further that $\angle APB + \angle AQB = 90°$.

    Prove that if O_1 is the centre of C_1 and O_2 is the centre of C_2, then the triangles O_1AO_2 and O_1BO_2 are right-angled.

12. Let $ABCD$ be a quadrilateral such that CD bisects $\angle ACB$. Suppose that

$$\angle DAC + \angle DBC + \angle DCB = 90°.$$

    Prove that D is the incentre of triangle ABC.

13. Circles C_1 and C_2 intersect at two distinct points P and Q. A line ℓ through P meets the circles C_1 and C_2 in A and B, respectively. Let Y be the midpoint of AB and suppose QY meets the circles C_1 and C_2 in X and Z, respectively.

    Show that $XY = YZ$.

14. Let AB and CD be distinct parallel chords of a circle and let P be a point on their common perpendicular bisector. The lines AP and CP meet the circle again at Q and R, respectively. Let S be the intersection of the lines BR and QD.

    Prove that PS is parallel to AB.

15. Triangle ABC is a 40°-60°-80° triangle. The angle bisector at A meets BC at D. Let E be the midpoint of CD and let F be the foot of the altitude from C to AB.

    Prove that perpendicular bisector of EF intersects AC at its midpoint.

16. From a point O, two tangents are drawn to a given circle touching it at points A and B. The chord AC is drawn parallel to the tangent OB, and OC intersects the circle at E.

    Prove that the line AE bisects OB.

17. Let ABC be a triangle with incentre I, inradius r, circumcentre O, circumradius R and semiperimeter s. Let I_a, I_b and I_c be the centres of the excircles opposite A, B and C, respectively, and let r_a, r_b and r_c be the radii of the excircles opposite A, B and C, respectively.

    Prove the following formulas.

    (a) $\frac{1}{r} = \frac{1}{r_a} + \frac{1}{r_b} + \frac{1}{r_c}$

    (b) $\text{Area}(\triangle I_aI_bI_c) = 2sR$

    (c) $OI^2 = R^2 - 2Rr$

18. Let k be a semicircle with diameter AB. Let D be a point such that $AB = AD$ and AD intersects k at the point E. Let F be the point on the chord AE such that $DE = EF$. Let BF extended meet k at the point C.

    Show that $\angle BAE = 2\angle EAC$.

19. The inscribed circle of triangle ABC is tangent to the sides AB and BC of $\triangle ABC$ in the points P and Q. The line PQ intersects the bisector of $\angle BAC$ at the point S.

    Prove that AS is perpendicular to SC.

20. Two chords AB and CD of a circle intersect at a point E inside the circle. Let M be an interior point of the segment EB. The tangent line at E to the circle through D, E and M intersects the lines BC and AC at F and G, respectively.

    If $AM/MB = t$, find EG/EF in terms of t.

21. The bisectors of the angles A and B of the triangle ABC meet the sides BC and CA at the points D and E, respectively.

    Assuming that $AE + BD = AB$, determine the size of $\angle C$.

22. Suppose that P is a point inside triangle ABC that satisfies $\angle ABP = \angle ACP$ and $\angle CBP = \angle CAP$.

    Prove that P is the orthocentre of the triangle.

23. Let P be an interior point of $\triangle ABC$, and let AP, BP and CP intersect BC, CA and AB at D, E and F, respectively.

    Prove that
    $$\frac{AP}{AD} + \frac{BP}{BE} + \frac{CP}{CF} = 2 \quad \text{and} \quad \frac{AE}{EC} + \frac{AF}{FB} = \frac{AP}{PD}.$$

24. Convex quadrilateral $PQRS$ satisfies $PQ = QR$ and PQ is not parallel to SR. The diagonals PR and QS intersect at T. The perpendicular bisectors of PR and QS intersect at V.

    Prove that $PQTV$ is cyclic.

25. Triangle ABC satisfies $\angle ABC = 2\angle ACB$. Point P, located inside the triangle, satisfies $PB = PC$ and $AP = AB$.

    Prove that $\angle BAC = 3\angle PAC$.

26. A triangle ABC satisfies $\angle ACB > \angle ABC$. The internal bisector of $\angle BAC$ meets BC at D. The point E on AB is such that $\angle EDB = 90°$. The point F on AC is such that $\angle BED = \angle DEF$.

    Show that $\angle BAD = \angle FDC$.

27. Let A, B and C be three collinear points with B between A and C. Equilateral triangles ABD, BCE and CAF are constructed with D and E on one side of the line AC and F on the opposite side.

    (a) Prove that the centroids of the triangles are the vertices of an equilateral triangle.

    (b) Prove that the centroid of this triangle lies on the line AC.

28. Let ABC be an acute triangle. Let AD be the altitude on BC, and let H be any interior point on AD. Lines BH and CH, when extended, intersect AC and AB at E and F, respectively.

    Prove that $\angle EDH = \angle FDH$.

29. Let C be a circle with centre O and let A and B be points on the circle such that $\angle AOB = 90°$. Let C_1 and C_2 be two circles internally tangent to C at points A and B, respectively. Furthermore C_1 and C_2 are tangent to each other and have centres O_1 and O_2, respectively. Circle C_3 is located inside angle AOB. It has centre O_3 and is externally tangent to C_1 and C_2 and is internally tangent to C.

    Prove that $OO_1O_3O_2$ is a rectangle.

30. Let ABC be an acute triangle. Let M be the midpoint of BC and P be the point on AM such that $MB = MP$. Let H be the foot of the perpendicular from P to BC. The lines through H perpendicular to PB and PC meet AB and AC at Q and R, respectively.

    Show that BC is tangent to the circle through Q, H and R at H.

31. Triangle ABC is acute-angled with circumcircle Γ and orthocentre H so that $AB \neq AC$. Let AH meet BC and Γ at D and E, respectively. Let F be the midpoint of BC. The line tangent to circle DEF at D meets the lines AB and AC at M and L, respectively.

    Prove that $MD = DL$.

32. Let ABC be a triangle with a right angle at A and area $\triangle$, and let S be its circumcircle. Let S_1 be the circle tangent to sides AB and AC and internally tangent to S. Let S_2 be the circle tangent to rays AB and AC and externally tangent to S. Let r_1 and r_2 denote the respective radii of S_1 and S_2.

    Prove that $r_1 r_2 = 4\triangle$.

33. Angle A is the smallest in triangle ABC. The points B and C divide the circumference of the triangle into two arcs. Let U be an interior point of the arc between B and C which does not contain A. The perpendicular bisectors of AB and AC meet the line AU at V and W, respectively. The lines BV and CW meet at T.

    Show that $AU = TB + TC$.

34. Points X, Y and Z are located inside triangle ABC and satisfy

$$\begin{aligned} \angle YAC = \angle ZAB &= \tfrac{1}{3}\angle A \\ \angle ZBA = \angle XBC &= \tfrac{1}{3}\angle B \\ \angle XCB = \angle YCA &= \tfrac{1}{3}\angle C. \end{aligned}$$

    Prove that triangle XYZ is equilateral.[1]

35. Let ω be the incircle of triangle ABC. Let L, N and E be the points of tangency of ω with the sides AB, BC and CA, respectively. Lines LE and BC intersect at the point H and lines LN and AC intersect at the point J. (All the points H, J, N and E lie on the same side of the line AB.) Let O and P be the midpoints of EJ and NH, respectively.

    Prove that

$$\text{Area}(HJNE) = 4\sqrt{\text{Area}(ABOP) \times \text{Area}(COP)}.$$

36. Let $ABCD$ be a convex quadrilateral in which the diagonals AC and BD are perpendicular and the opposite sides AB and DC are not parallel. The perpendicular bisectors of AB and CD meet at point P inside $ABCD$.

    Prove that $ABCD$ is cyclic if and only if triangles ABP and CDP have equal areas.

37. Consider five points A, B, C, D and E such that $ABCD$ is a parallelogram and $BCED$ is a cyclic quadrilateral. Let ℓ be a line passing through A. Suppose that ℓ intersects the interior of the segment DC at F and intersects line BC at G. Suppose also that $EF = EG = EC$.

    Prove that ℓ is the bisector of angle DAB.

38. Let $ABCD$ be a convex quadrilateral which does not have any two sides of equal length.

    Prove that $ABCD$ is cyclic if and only if there exist points Q and R on line BD, one strictly between B and D, the other outside of the segment BD, such that

$$\angle DAQ = \angle DCQ = \angle BAR = \angle BCR.$$

[1] This result is known as *Morley's theorem*.

39. A circle with centre O passes through the vertices A and C of triangle ABC and intersects the segments AB and BC again at distinct points K and N, respectively. The circumscribed circles of the triangles ABC and KBN intersect at exactly two distinct points B and M.

    Prove that angle OMB is a right angle.

4.1 Angle chasing

One of the simplest techniques in geometry is to 'chase' angles. Given a diagram, insert all the angles you can. Label one angle as α or some other symbol, and then work out what you can in terms of α. Continuing this process disposes of many a geometry problem.

Warning! *Doing* an angle chase is much easier than *writing down* an angle chase in a proof, step by step. The former just involves writing stuff on your diagram. The latter involves explaining each step, in the correct order. As a result it's sometimes hard to read an angle chase when written down as a proof. (Just ask any Olympiad marker!) So don't worry if what's written here is a little hard to follow. It's best to draw your own diagram and work through the proof on that diagram.

Let's start by stating an underused but important theorem.

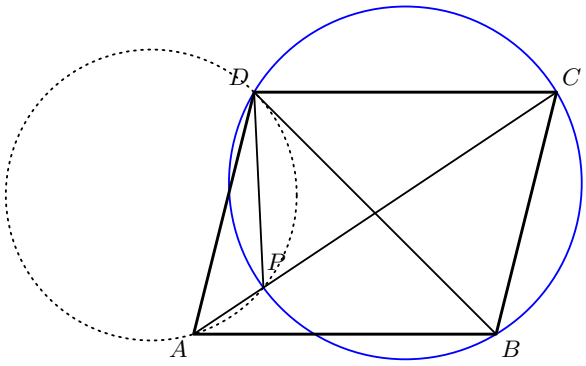
Alternate segment theorem Let A, B, C be points on a circle Γ, and let PA be a line segment such that P lies on the opposite side of line AB as C. Then the line PA is tangent to Γ at A if and only if $\angle ACB = \angle PAB$.

We will now see how angle chasing can make light work of a problem.

Problem In parallelogram $ABCD$, AC is longer than BD. Let P be a point on AC such that $BCDP$ is a cyclic quadrilateral.

Prove that BD is a common tangent to the circumcircles of triangle ADP and triangle ABP.

Solution



By the alternate segment theorem, it is sufficient to prove that $\angle PDB = \angle DAP$ and $\angle PBD = \angle BAP$.

Since the quadrilateral $BCDP$ is cyclic, we have $\angle PDB = \angle PCB$. Also $\angle PCB = \angle DAP$ because $ABCD$ is a parallelogram. Putting these two pieces of information together gives $\angle PDB = \angle DAP$, which is one of the statements that we wanted to prove. The other follows by an entirely analogous argument. □

4.2 Cyclic quadrilaterals

It's always great to find a cyclic quadrilateral. In fact it is positive progress, because they have useful properties.

Theorem

- A quadrilateral $ABCD$ is cyclic if and only if $\angle ABC + \angle ADC = 180°$.
- A quadrilateral $ABCD$ is cyclic if and only if $\angle ACB = \angle ADB$.

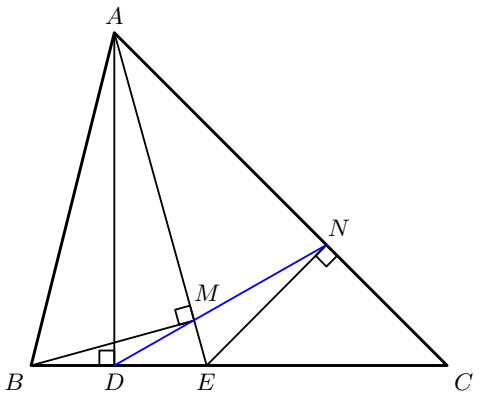
Even problems that don't seem to involve any circles often feature cyclic quadrilaterals.

It's worth remembering that discovering a cyclic quadrilateral is a valuable find because it gives you new information that you could not otherwise obtain by standard angle-chasing techniques.

Problem In triangle ABC, points D and E are located on the side BC such that AD is an altitude and AE is an angle bisector. The point M on AE is such that BM is perpendicular to AE and the point N on AC is such that EN is perpendicular to AC.

Prove that the points D, M, N are collinear.

Solution



With three right angles floating around the diagram, you can be confident that there are also cyclic quadrilaterals. In fact, we know that $ABDM$ is cyclic because $\angle ADB = \angle AMB = 90°$. We also know that $ADEN$ is cyclic because $\angle ADE + \angle ANE = 90° + 90° = 180°$.

Now there are many ways to show that the points D, M, N are collinear, but our particular approach will be to prove that $\angle BDM + \angle NDC = 180°$.

As with many plane geometry problems, we start by labelling a sensible angle. Here we will use $\angle BAC = 2\alpha$. Then we use this to label as many other angles in the diagram as possible. For a start, we have $\angle BAE = \angle CAE = \alpha$.

The cyclic quadrilateral $ABDM$ tells us that

$$\angle BDM = 180° - \angle BAM = 180° - \angle BAE = 180° - \alpha.$$

The cyclic quadrilateral $ADEN$ tells us that

$$\angle NDC = \angle NDE = \angle NAE = \alpha.$$

Therefore, $\angle BDM + \angle NDC = (180° - \alpha) + \alpha = 180°$, as required. □

4.3 One step at a time

With any type of problem, not just maths problems, often a good tactic is to examine different parts of the problem one by one and see what you can work out. Then putting it all together

you may obtain an answer.

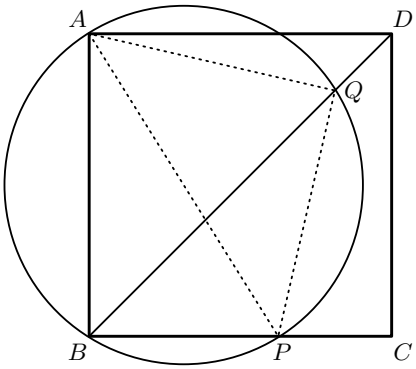
This is pretty obvious advice but it's often overlooked in geometry, since you usually draw the whole diagram at once, and the whole diagram is often too complicated to see anything, especially in harder problems.

Nobody can teach you how to separate out the parts of a difficult problem. Indeed getting to know how to look at things the right way is part of the skill of problem solving. This principle can be used to solve the following easier problem.

Problem Let $ABCD$ be a square and P be a point on its side BC. The circle passing through points A, B and P intersects BD once more at point Q. The circle passing through points C, P and Q intersects BD once more at point R.

Prove that points A, R and P are collinear.

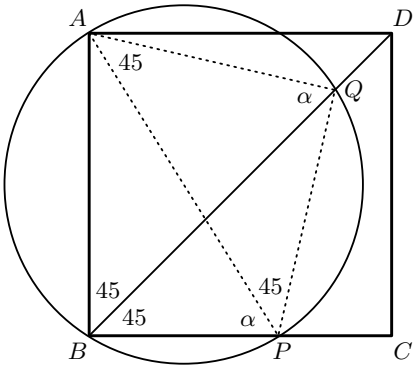
Solution First of all draw the entire diagram carefully. But after that it pays to draw just a part of the diagram and see what we can glean. We will examine the situation with each circle separately. So first let's just examine the set-up with circle $ABPQ$.



Since BD is the diagonal of a square, $\angle ABD = \angle CBD = 45°$.

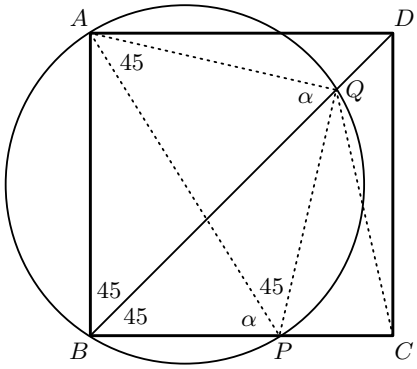
Clearly $ABPQ$ is cyclic, so $\angle APQ = \angle ABQ = 45°$ and $\angle PAQ = \angle PBQ = 45°$. Hence APQ is isosceles with $AQ = PQ$.

Similarly, letting $\angle APB = \alpha$ we can chase all the angles around.



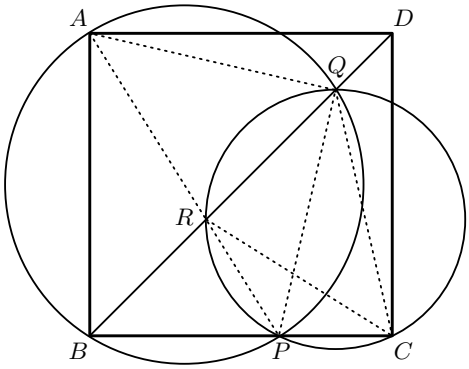
Some of the other angles in the diagram are $\angle BQP = \angle BAP = 90° - \alpha$ and $\angle QPC = 135° - \alpha$.

Well, that's all pretty interesting. Now let's think about drawing in the circle around C, P, Q. Actually, let's just think about drawing in CQ—one step at a time!



When you only look at this part of the diagram, it's pretty clear that by some sort of symmetry, $AQ = CQ$. You could prove this by saying that C is the reflection of A in the line BD, and Q is on BD. Or you could prove ADQ and CDQ are congruent. Anyway, we now have $AQ = CQ = PQ$. So CPQ is isosceles and $\angle QCP = \angle QPC = 135° - \alpha$.

Now we'll put the second circle in.



We now have $\angle PCR = \angle PQR = 90° - \alpha$. Since $\angle PCQ = 135° - \alpha$, we have $\angle RCQ = 45°$. Thus $\angle RPQ = \angle RCQ = 45°$. But now $\angle RPQ = \angle APQ = 45°$. Therefore, points A, R and P are collinear as required. □

4.4 Triangle centres

It is likely that you have come across the following special points of a triangle.[2] See if you can navigate through the following exercises.

1. **Centroid** Let ABC be a triangle with medians AX, BY and CZ meeting at centroid G.
   - Prove that the line segments XY, YZ and ZX divide triangle ABC into four smaller triangles which are congruent to each other.

[2] If you don't know about centroids, orthocentres, and so on, you can look them up on the internet.

- The triangle XYZ is called the *medial triangle*. Prove that it can be obtained by performing a dilation[3] on triangle ABC with centre G and factor $-\frac{1}{2}$.
- Prove that G is the centroid of the medial triangle.
- Prove that the three medians of a triangle divide it into six triangles of equal area.
- Prove that
$$\frac{AG}{GX} = \frac{BG}{GY} = \frac{CG}{GZ} = 2.$$

2. **Orthocentre** Let ABC be a triangle with altitudes AD, BE and CF meeting at orthocentre H.
   - Draw triangle ABC, the three altitudes AD, BE, CF and triangle DEF.
   - Write down all six cyclic quadrilaterals which appear in the diagram.
   - Label every angle in the diagram in terms of $\angle A$, $\angle B$ and $\angle C$.
   - What is the relationship between H and triangle DEF?
   - What are the orthocentres of triangles AHB, BHC and CHA?

   The four points A, B, C and H are called *orthocentric*.

3. **Circumcentre** Let ABC be a triangle with circumcentre O.
   - Draw triangle ABC, the perpendicular bisectors of the sides XO, YO, ZO and the medial triangle XYZ.
   - Write down all three cyclic quadrilaterals appearing in your diagram.
   - Label every angle in the diagram in terms of $\angle A$, $\angle B$ and $\angle C$.
   - What is the relationship between O and triangle XYZ?

4. **Incentre** Let the incircle of triangle ABC have centre I and touch AB, BC and CA at P, Q and R, respectively.
   - Draw triangle ABC, the angle bisectors AI, BI, CI, the points of tangency of the incircle P, Q, R, the segments PI, QI, RI and triangle PQR.
   - Write down all three cyclic quadrilaterals appearing in your diagram.
   - Label every angle in the diagram in terms of $\angle A$, $\angle B$ and $\angle C$.
   - Label the lengths AQ, AR, BR, BP, CP and CQ in terms of the side lengths a, b and c of the triangle.
   - Prove that the area of triangle ABC is given by rs, where r is the radius of the incircle and $s = \frac{a+b+c}{2}$ is the semiperimeter.

5. **Excentres** Let I_a, I_b and I_c be the excentres opposite A, B and C, respectively, of triangle ABC. Suppose that the excircle centred at I_a touches AB, BC and CA at P, Q and R, respectively.
   - Draw triangle ABC, the internal angle bisector AI_a, the external angle bisectors BI_a, CI_a, the points of tangency of the excircle P, Q, R, the segments PI_a, QI_a, RI_a and triangle PQR.
   - Write down all three cyclic quadrilaterals appearing in your diagram.
   - Label every angle in the diagram in terms of $\angle A$, $\angle B$ and $\angle C$.

[3] See section 7.3.

- Label the lengths AQ, AR, BR, BP, CP and CQ in terms of the side lengths a, b and c of the triangle.
- Prove that the area of triangle ABC is given by $r_a(s-a)$, where r_a is the radius of the excircle opposite A and s is the semiperimeter.
- What is the relationship between I and triangle $I_aI_bI_c$?
- What is the relationship between triangle $I_aI_bI_c$ and the circumcircle of triangle ABC?

4.5 Constructions

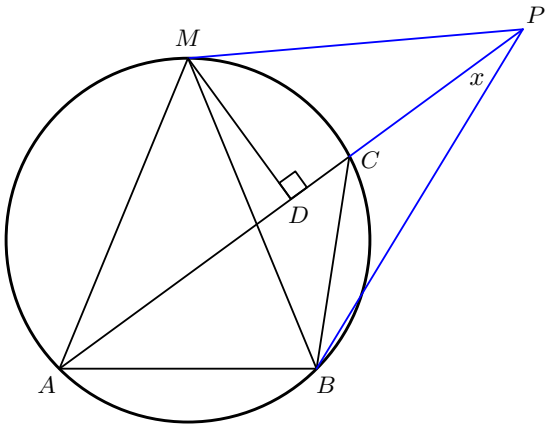
There's a simple rule of thumb for deciding whether a geometry question is hard, or *really* hard. If it can be solved without drawing any extra lines beyond the ones you're given (and maybe a couple more natural ones), it's not that hard. If it requires you to make some constructions of your own, it's really hard.

Again, nobody can teach you to look at a diagram and say where to make a construction—this is a truly creative business. But by doing enough problems it's possible to get a feel for how a certain situation works, enough that you can try inserting various extra bits, here and there. Don't be afraid to do so: if it doesn't work, or your diagram becomes too cluttered, just draw another one!

Problem Suppose that A, B, M are points on a circle such that M is the midpoint of the arc AB. Let C be an arbitrary point on the arc AMB such that AC is longer than BC. Let D be the foot of the perpendicular from M to AC.

Prove that $AD = DC + CB$.

Solution



A much easier task than proving one length is equal to the sum of two lengths is proving that one length is equal to another. With this in mind, we extend the line AC to the point P such that $CP = CB$. Of course, what we now need to prove is that $AD = DP$.

But if $AD = DP$, then we would know that M lies on the perpendicular bisector of AP. Since M also lies on the perpendicular bisector of AB, it must be the case that M is the circumcentre of triangle ABP. Let's aim to prove this using an angle chase.

First, we let $\angle APB = x$. Since we have constructed triangle BCP to be isosceles, we know that $\angle PBC = x$ and $\angle PCB = 180° - 2x$. From this, it follows that $\angle ACB = 2x$ and since $ABCM$ is a cyclic quadrilateral, we also have $\angle AMB = 2x$.

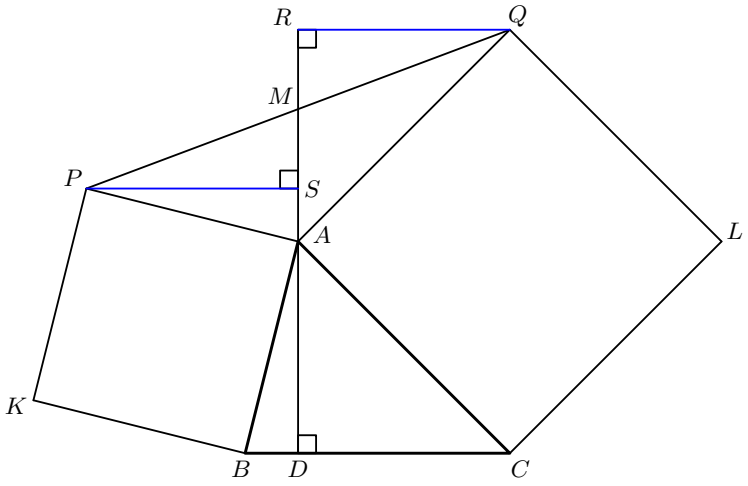
Now look at what we have here. The chord AB subtends an angle $2x$ at M with $AM = BM$ and an angle x at P. Since P and M lie on the same side of AB, the point M is indeed the circumcentre of triangle ABP. We now know that MD splits the isosceles triangle AMP into two congruent triangles, so $AD = DP$. □

By the way, as with most geometry problems, different solutions can be found. For the one just discussed, it is possible to define a point Q on segment AD such that $CD = DQ$. Thus it remains to show that $AQ = BC$. This can be done by proving that triangles AMQ and BMC are congruent. See if you can angle chase out the details!

Problem A triangle ABC has squares $PABK$ and $QACL$ constructed on its exterior. The altitude AD of triangle ABC is extended to meet PQ at point M.

Prove that M is the midpoint of PQ.

Solution



First, let's do a bit of angle chasing. If we let $\angle CAD = \alpha$ and $\angle BAD = \beta$, then we have $\angle ACD = 90° - \alpha$ and $\angle ABD = 90° - \beta$. Also, since angles on a straight line add to 180°, we have $\angle QAM = 90° - \alpha$ and $\angle PAM = 90° - \beta$.

There are several ways to proceed from here, but one way is as follows. Note that $QA = AC$ and $\angle QAM = \angle ACD$. These are quite similar situations, and by drawing a single line segment, we can create a pair of congruent triangles—certainly a useful thing to do! So let R be the foot of the perpendicular from Q to the line MD. This gives us the congruent triangles QAR and ACD, as desired. But what we have done on the right side of the diagram, we can similarly do on the left. So let S be the foot of the perpendicular from P to the line MD, so that we have congruent triangles PAS and ABD. In particular, we have managed to prove that $QR = AD = PS$.

Therefore, M is horizontally halfway between P and Q. This means that M must be the midpoint of PQ, and we are done. □

(You could make this last statement a little more rigorous by showing that triangle MQR is congruent to triangle MPS, so that $MP = MQ$.)

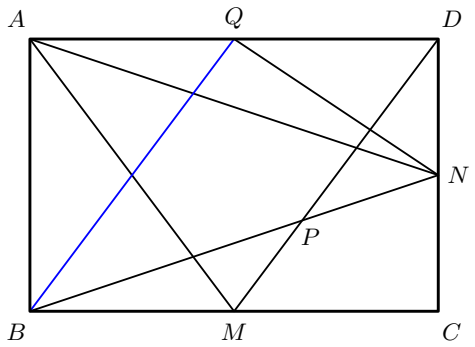
4.6 Exploit symmetry

When the diagram is symmetric, you can draw the same thing in different places.

Problem In rectangle $ABCD$, let M and N be the midpoints of BC and CD, respectively. Let DM and BN intersect at P.

Prove that $\angle MAN = \angle BPM$.

Solution We use the symmetry of the rectangle to redraw the situation near A, over near B. That is, let Q be the midpoint of AD and construct BQ. Then, by symmetry, $\angle QBN = \angle MAN$.



All that we need to prove now is that BQ and MD are parallel. But once we observe that BM and QD are equal and parallel, it follows that $BQDM$ is a parallelogram and so BQ is parallel to MD as desired. □

4.7 Extend to the circumcircle

When you've got a triangle involved and some cevians (any lines through the three vertices of the triangle such as medians, angle bisectors or altitudes), a useful construction is to extend them to the circumcircle of the triangle. This might sound like a pretty random thing to do, but it comes up sufficiently often that we're devoting a section to it. Nobody has ever explained to us a deep underlying reason as to why it should work so well, but it does.[4]

Here is a nice property associated with extending the angle bisectors to the circumcircle. See if you can prove it.

Extensions of angle bisectors to circumcircle Let I be the incentre of triangle ABC. If AI extended meets the circumcircle of triangle ABC at D, then $BD = CD = ID$. This means that D is the circumcentre of triangle IBC.

A diagram of this configuration is found on page 85.

[4]For example, this construction can be used with (i) the extended sine rule (section 4.9), (ii) Ceva's theorem (section 6.4) and (iii) a cyclic hexagon with concurrent diagonals as per the configuration found on page 95.

There is one diagram that is particularly exciting! Take a triangle and its circumcircle. Draw its altitudes and extend them to the circumcircle. Now, as an exercise, try to discover as many interesting properties of this diagram as you can.

Here are a couple of such properties that you should try to prove for yourself.

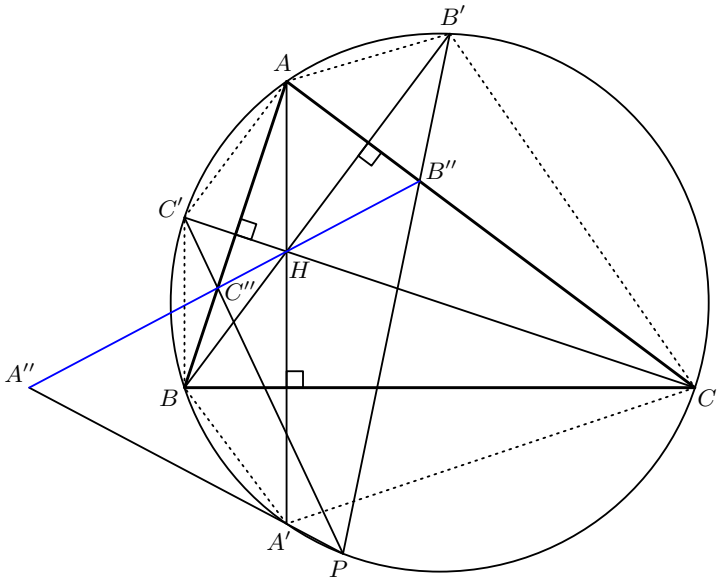
Extensions of altitudes to circumcircle Let AD, BE and CF be the altitudes and let H the orthocentre of triangle ABC. If AD, BE and CF are extended to meet the circumcircle of triangle ABC at A', B' and C', respectively, then $HD = DA'$, $HE = EB'$ and $HF = FC'$. This implies the following properties.

- The reflections of H in the three sides of the triangle lie on the circumcircle.
- The diagram has three kites in it, namely, $HBA'C$, $HCB'A$ and $HAC'B$.
- The circumradius of triangle HBC is equal to the circumradius of triangle ABC.

Problem With notation as above, let P be a point on the circumcircle of triangle ABC. Suppose that A'' is the intersection of lines PA' and BC, B'' is the intersection of lines PB' and CA, and C'' is the intersection of lines PC' and AB.

Prove that the points A'', B'' and C'' are collinear.

Solution Try drawing the diagram yourself. It looks quite complicated, unless it is large and multicoloured. A carefully drawn diagram suggests that H also lies on the line through A'', B'' and C''. If this were true, then it suffices to prove that any two of A'', B'', C'' are collinear with H. With this tactic in mind let's try and prove that B'', C'' and H are collinear. For this, it is enough to prove that $\angle BHC'' = \angle B'HB''$.



We know that AB is the perpendicular bisector of $C'H$. Furthermore, C'' lies on AB. Thus $\angle BHC'' = \angle BC'C'' = \angle BC'P$.

Similarly, AC is the perpendicular bisector of $B'H$. And B'' lies on AC. Therefore, we have $\angle B'HB'' = \angle HB'B'' = \angle BB'P$.

Finally, we note that $\angle BC'P = \angle BB'P$. Thus $\angle BHC'' = \angle B'HB''$ which implies B'', C'' and H are collinear.

A similar argument shows that A'', B'' and H are collinear. Thus A'', B'', C'' and H are collinear and we are done. □

Actually, we are not done because we have run into diagram dependence issues. Although the argument that A'', B'' and H are collinear is similar to the argument that B'', C'' and H are collinear, it is only similarish. There are some differences in the angle chase. Finally, the whole argument depended on the particular diagram drawn and the position of P relative to the other points on the circumcircle. All these issues need to be ironed out. See if you can do it![5]

As a final kicker, there is a *really* short solution to this problem involving Pascal's theorem.[6] See if you can find it.

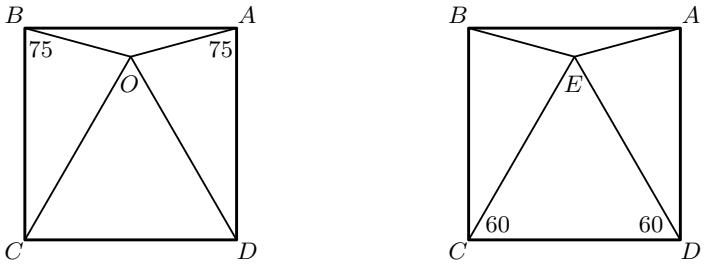
4.8 Reverse reconstruction

In a geometry problem, the situation you must deal with is fairly concrete. However, just by varying your logic a little, you can often find different approaches which work well. Sometimes you assume something slightly different to the problem, and show that actually you have the situation of the problem. Sometimes you can work backwards from the answer. Here is an example.

Problem Point O lies inside square $ABCD$ such that $\angle OAB = \angle OBA = 15°$.

Prove that triangle ODC is equilateral.

Solution Although there is a trigonometric approach to this problem, without trigonometry the problem is difficult to approach directly. So instead, we'll vary our assumptions by approaching the problem from the reverse end.



Let E be the point inside $ABCD$ such that EDC is equilateral. Then we have a diagram which looks the same, but by our assumption we know a different set of angles, and lots of lengths are equal. We now aim to show that $\angle EAB = \angle EBA = 15°$, so that E and O are the same point. (This follows since there is only one possible point O inside $ABCD$ satisfying the conditions $\angle OAB = \angle OBA = 15°$.)

As triangle CDE is equilateral we have $CE = CD = CB$. So triangle CBE is isosceles. But since $\angle BCE = 30°$ we have $\angle CEB = \angle CBE = 75°$ and so $\angle EBA = 15°$. Similarly, $\angle EAB = 15°$, as desired. Therefore $O = E$ and triangle $ODC = EDC$ is equilateral. □

[5] For help on how to do this you might like to consult section 17.3.
[6] See section 6.9.

4.9 Trigonometry

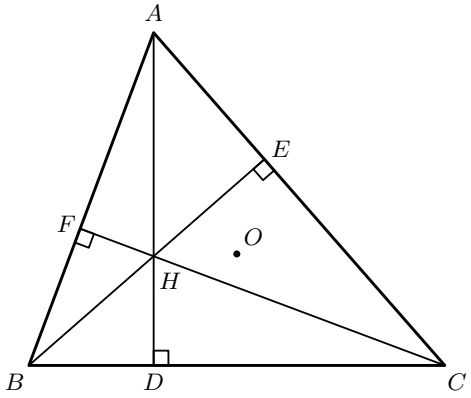
Sine rule Let ABC be a triangle with side lengths $a = BC$, $b = CA$, $c = AB$ and circumradius R. Then we have

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C} = 2R.$$

Problem Let ABC be an acute-angled triangle with circumcentre O and orthocentre H. If $AO = AH$, find all possible values of $\angle A$.

Solution The sine rule automatically gives us the value of AO in terms of the side lengths and angles of triangle ABC. The natural approach is to use trigonometry to express the length AH in terms of the side lengths and angles of triangle ABC.

As we discussed earlier in this chapter, any angles involving the orthocentre, the feet of the altitudes, and the vertices of the triangle can easily be found in terms of $\angle A$, $\angle B$ and $\angle C$. So our task shouldn't be too difficult!



Let D, E and F be the feet of the altitudes as per the diagram. Considering triangle AEH, we obtain

$$\frac{AE}{AH} = \sin \angle C \quad \Rightarrow \quad AH = \frac{AE}{\sin \angle C}.$$

Considering triangle AEB, we obtain

$$\frac{AE}{AB} = \cos \angle A \quad \Rightarrow \quad AE = AB \cos \angle A.$$

Piecing these two pieces of information together and invoking the sine rule yet again, we end up with

$$AH = \frac{AB \cos \angle A}{\sin \angle C} = 2R \cos \angle A = 2AO \cos \angle A.$$

Given that $AH = AO$, we can cancel this equation to give $\cos \angle A = \frac{1}{2}$, which in turn implies that $\angle A = 60°$. □

It turns out that if the triangle is not restricted to being acute, then there is another value of $\angle A$ for which $AH = AO$. See if you can find (and prove) what it is.

4.10 Areas

One simple approach to geometry problems, often neglected, is to consider areas. This means more than remembering the formulas for the area of a triangle

$$\Delta = \frac{1}{2}\text{base} \times \text{height} \quad \text{and} \quad \Delta = \frac{1}{2}ab\sin C.$$

The following ideas are useful.

Area properties

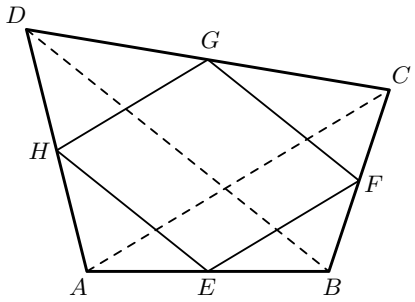
- Two triangles with the same base length and the same height have equal area.
- If two similar figures have corresponding lengths in the ratio $a : b$, then their areas are in the ratio $a^2 : b^2$.

The following problem is a variation on these ideas.

Problem Let $ABCD$ be a quadrilateral. Let the midpoints of AB, BC, CD and DA be E, F, G and H, respectively.

Prove that $EFGH$ has half the area of $ABCD$.

Solution The first thing to notice is that $EFGH$ is a parallelogram. Why? By the midpoint theorem, both EF and GH are parallel to AC and half its length. Therefore $EF \parallel GH$ and $EF = GH$. Similarly $FG = HE$ and $FG \parallel HE$. This is a commonly used fact. Now the diagram looks nicer!



Since clearly $\triangle AHE \sim \triangle ADB$ with ratio $1 : 2$ we have

$$|AHE| = \frac{1}{4}|ADB|.$$

(Here we have used the notation $|AHE|$ to mean the area of polygon AHE.)

Similarly,

$$|CFG| = \frac{1}{4}|CBD|, \quad |BEF| = \frac{1}{4}|BAC| \quad \text{and} \quad |DCH| = \frac{1}{4}|DAC|.$$

Note that

$$|ADB| + |CBD| = |BAC| + |DAC| = |ABCD|.$$

Adding together the equations above gives

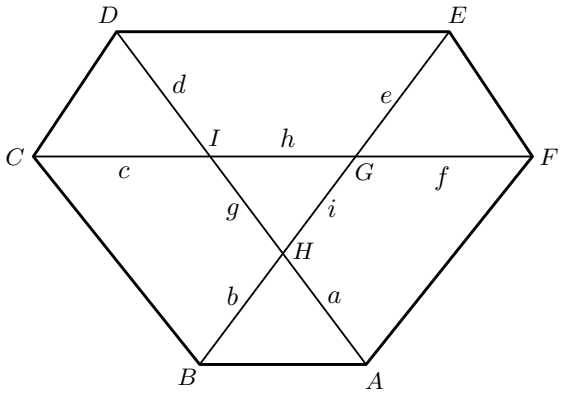
$$|AHE| + |CFG| + |BEF| + |DCH| = \frac{2|ABCD|}{4} = \frac{|ABCD|}{2}.$$

So these four triangles give half the area of $ABCD$. Now $EFGH$ is what you get if you remove these four triangles from $ABCD$. So $|EFGH| = \frac{1}{2}|ABCD|$ as required. □

Problem Suppose that each of the three main diagonals AD, BE and CF divide the convex hexagon $ABCDEF$ into two regions of equal area.

Prove that the three diagonals meet at a common point.

Solution Label the intersection points and lengths along the diagonals as shown. (The diagram has been drawn assuming that the diagonals don't meet at a common point.)



We want to show that the three points G, H and I coincide. Our strategy is to prove that $g = h = i = 0$.

Each of the main diagonals divides the hexagon into equal pieces of equal area. It follows that $|ABCD| = |DEFA| = |BCDE| = |EFAB| = |CDEF| = |FABC|$.

Now $ABCD$ and $BCDE$ overlap along $BCDH$, so we have

$$|ABH| = |DEH|.$$

Therefore,

$$\frac{1}{2}ab\sin\angle AHB = \frac{1}{2}(d+g)(e+i)\sin\angle DHE,$$

and since the angles are equal, this reduces to

$$ab = (d+g)(e+i).$$

Considering other pairs of quadrilaterals similarly gives

$$ef = (b+i)(c+h) \quad \text{and} \quad cd = (a+g)(f+h).$$

Multiplying all three equations together yields

$$abcdef = (a+g)(b+i)(c+h)(d+g)(e+i)(f+h).$$

All the quantities involved are non-negative, so this is ridiculous unless we have $g = h = i = 0$. Hence AD, BE and CF must meet at a point. □

Note that this solution is diagram dependent. There is one other way to draw the diagram. Specifically H could lie inside quadrilateral $CDEF$ (instead of $FABC$ as in the diagram). But this case can be handled in much the same way.

4.11 Relate to known diagrams

Most often, a construction is made because it relates one part of the diagram to another. If nothing extra is drawn, then often a section that is hard to deal with cannot be related to a well-understood portion. A construction is often the key. Knowing what construction to make is a matter of insight and experience. The underlying philosophy is always: make it nicer! If you know you can fit the diagram inside another diagram you have studied before, draw it in by all means!

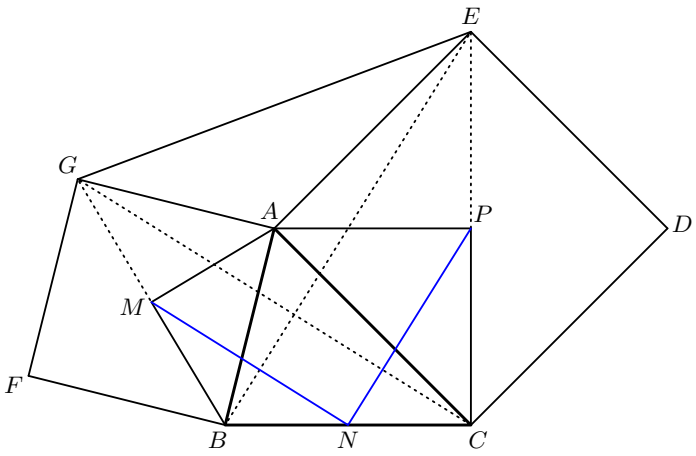
Problem Suppose that N is the midpoint of the side BC of triangle ABC. Construct right isosceles triangles AMB and APC on sides AB and AC outside the triangle where, $\angle AMB = \angle APC = 90°$.

Prove that MNP is also a right isosceles triangle.

Solution There is a construction which simplifies this problem enormously. Motivation is provided by the following considerations.

- It's hard to relate points from one of the isosceles triangles to the other.
- We are used to drawing squares, not isosceles triangles, on the sides of triangles, in some proofs of Pythagoras' theorem. This is a situation with lots of nice things going on.
- Right isosceles triangles lie nicely in squares.

So, construct squares $ABFG$ and $ACDE$ as shown.



We see that M and P are the centres of these two squares. In fact, since $BM = \frac{1}{2}BG$ and $BN = \frac{1}{2}BC$, triangles BMN and BGC are similar. So, $MN = \frac{1}{2}GC$, and $MN \parallel GC$. Similarly, $NP = \frac{1}{2}BE$ and $NP \parallel BE$. Thus it suffices to show that $GC = BE$ and $GC \perp BE$.

These facts are some of the nice properties of this diagram that you may already know from elsewhere. Either way, note that $\angle GAC = \angle GAB + \angle BAC = \angle EAC + \angle BAC = \angle BAE$ and $GA = BA$ and $AC = AE$. Therefore, triangles GAC and BAE are congruent. (In fact one is a rotation of the other about A by 90°.) It follows that $GC = BE$ and $GC \perp BE$ as required. □

4.12 Create beautiful pictures

In the end, geometry is one of the greatest art forms known to humankind. And in solving a geometry problem you are the artist, and you create beautiful pictures. Understand the picture; move the pieces around and see what happens; help the picture explain itself; it is, after all, confused, just like us, and seeks illumination.

Problem Let ABD be a triangle and let C be a point on the side BD, lying strictly between B and D. Suppose that $BC = 2CD$, $\angle ACB = 60^\circ$ and $\angle ADC = 45^\circ$.

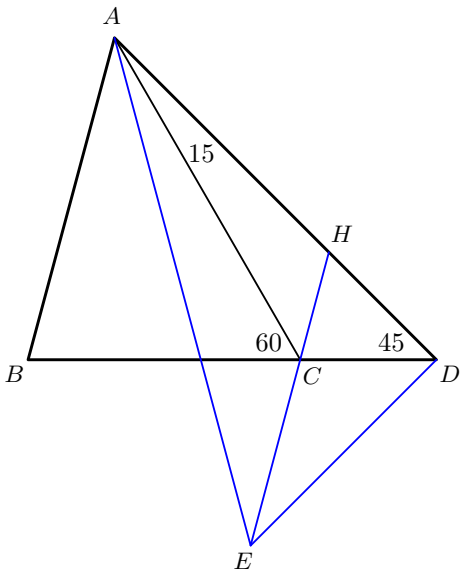
Determine $\angle BAD$.

Solution Simple techniques will not suffice here. Trigonometry can provide a solution but we'll avoid it, and take the artistic approach.

We get the feeling that there might be a construction which illuminates the diagram. We hope to discover something we didn't know before about the diagram. For instance, we might discover that some point is a centroid, or incentre, or another well-known point. We might discover a cyclic quadrilateral, a tangent line, or something else useful.

Since $DC : CB = 1 : 2$, it might be worthwhile to make a construction so that C will be a centroid of some triangle. But that does not seem to work, and so we will instead try to make C the *incentre* of a triangle. Incentres, being the intersection of angle bisectors, always give us lots of information about angles, and information about ratios, which is what we want.

So, paint a picture where C is the incentre of some triangle. A little angle chasing gives $\angle DAC = 15^\circ$. So construct a point E such that $\angle DAE = 30^\circ$ and $\angle ADE = 90^\circ$. Then C is the incentre of triangle AED. Furthermore, AED is a 30°-60°-90° triangle. Label points as shown with H being the intersection of EC and AD. Note that EHD is also a 30°-60°-90° triangle.



By the angle bisector theorem we have

$$\frac{AH}{HD} = \frac{AE}{DE} = 2.$$

We were also given

$$\frac{BC}{CD} = 2.$$

Hence triangles HCD and ABD are similar and therefore, $\angle BAD = 60°$. □

5

Important configurations in geometry

Solving problems in geometry very often involves recognising that a diagram contains a known configuration within it. What follows is a partial list of such configurations, a number of which you have already seen in this book and others that you will yet see. You should, in time, be able to prove them all. However, the art is in learning to recognise the configuration as part of a more complex diagram.

Some of these are well-known theorems and are quotable. Others are not so well known. In most cases it is permissible to quote the result. However, just making use of the result without explicitly saying that it is a *known* result has caused many students to lose up to three marks at the IMO.

The configurations are organised into three categories.

A-List	Extremely useful
B-List	Very useful
C-List	Useful

Each configuration is accompanied by one, two or three stars. This is an estimated rating of how difficult the result is to prove using geometric methods.[1] More stars mean that it is more difficult to prove.

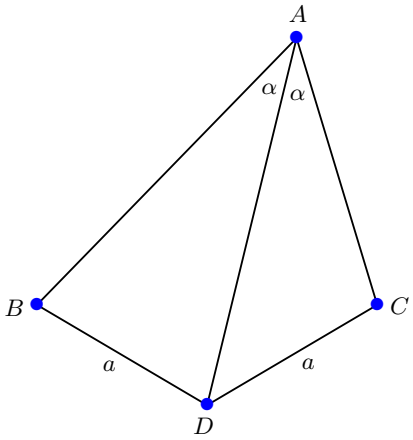
The configurations are purposely kept minimal since the idea is to be able to recognise the minimal elements needed to infer something. The resulting incidence is generally indicated by fattening and colouring the points or curves involved. Thus four fat points usually indicate that the four points are concyclic and so on.

This chapter is clearly very different to all the other chapters. Try to become very familiar with it!

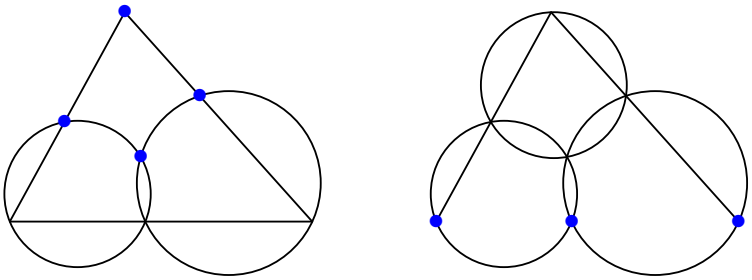
[1] That is, without resorting to grubby computational methods such as trigonometry, complex numbers or coordinate geometry!

A1 Angle bisector and perpendicular bisector ★

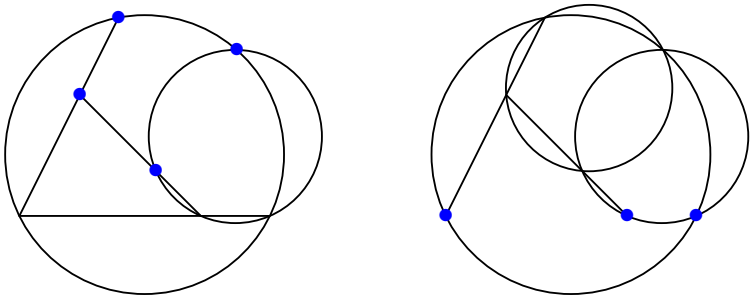
$AB \neq AC \quad \Rightarrow \quad ABCD$ is cyclic.
$AB = AC \quad \Rightarrow \quad ABCD$ is a kite.



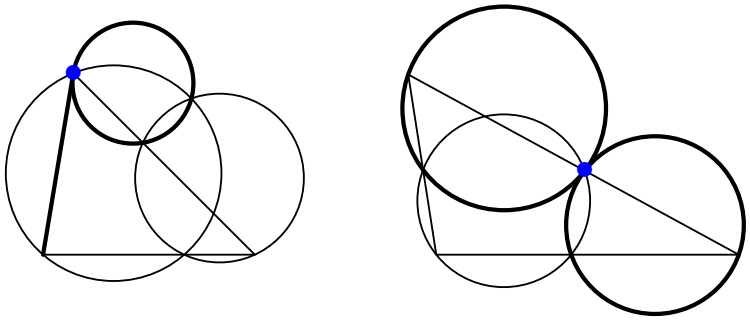
A2 Pivot theorem ★



Variations: lines extended

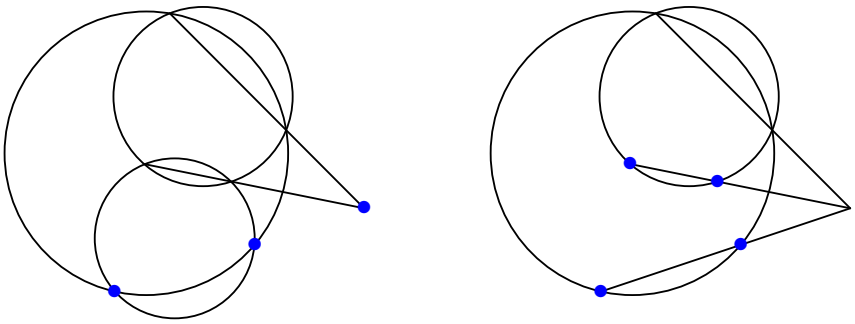


Variations: tangents[2]



[2]In both these variations, two points are allowed to coalesce into a single point resulting in a tangent incidence rather than two distinct intersection points.

A3 Radical axis theorem ★

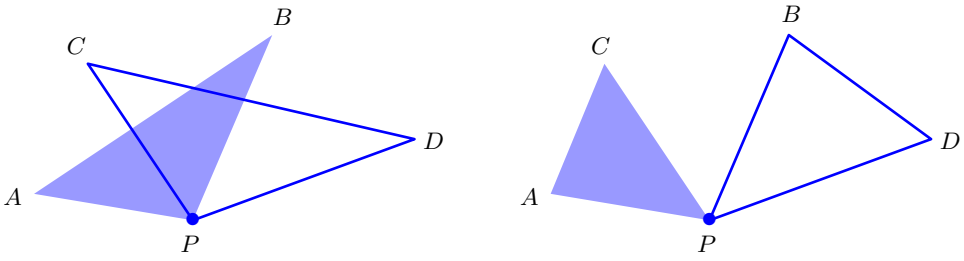


Variations

- Zero radius circle(s)
- Some circles tangent

A4 Similar switch ★

$\triangle APB \sim \triangle CPD \quad \Leftrightarrow \quad \triangle APC \sim \triangle BPD.$

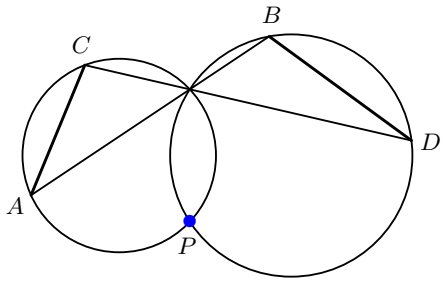


Equivalent formulation via spiral symmetries

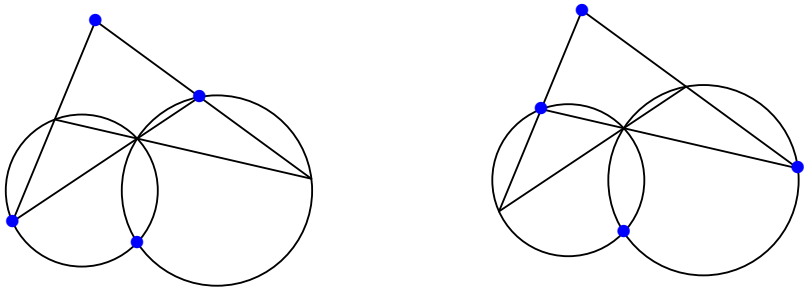
There is a spiral symmetry about P which sends segment AB to segment CD if and only if there is a spiral symmetry about P which sends segment AC to segment BD.

Additional information: presence of two circles

The above configuration is often seen as a byproduct of two intersecting circles. Point P is the centre of each spiral symmetry.

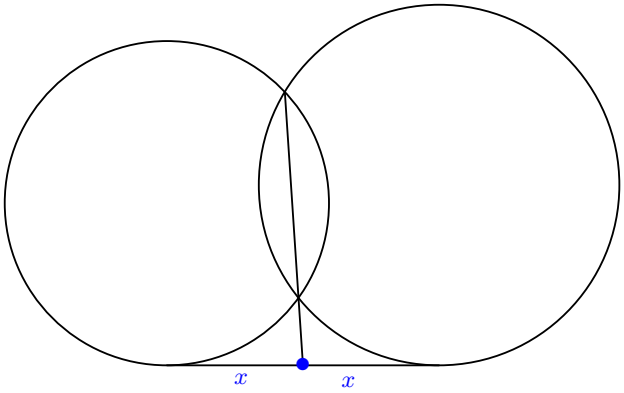


Additional information: presence of two more circles

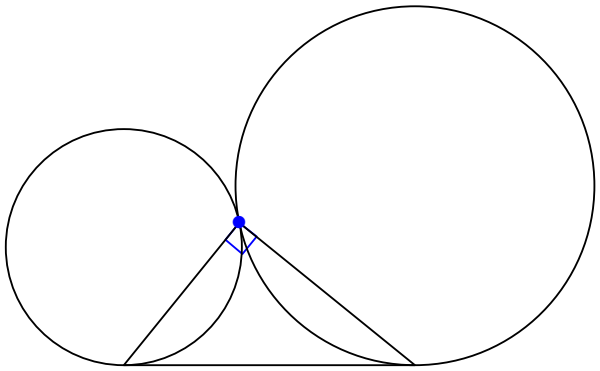


(Compare this with the four lines and four circles diagram in the C-List.)

B1 Radical axis bisects common tangent ★



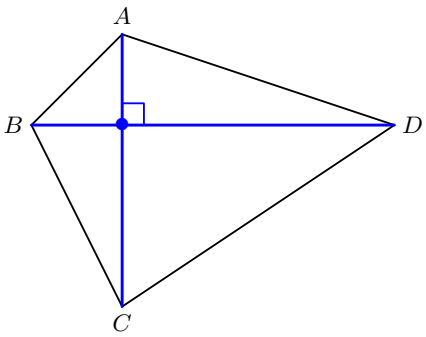
Variation: tangent circles



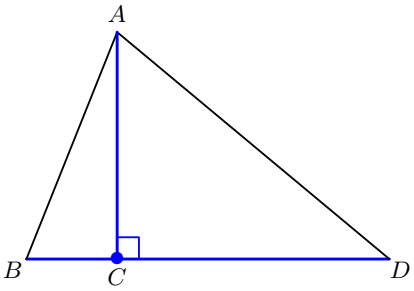
B2 Perpendicularity ★

$AC \perp BD \quad \Leftrightarrow \quad AB^2 + CD^2 = AD^2 + BC^2.$

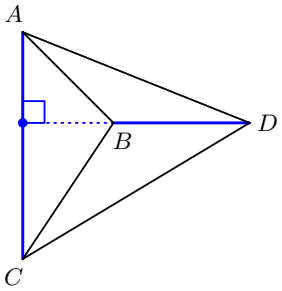
Convex quadrilateral

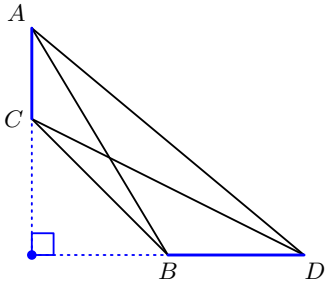


Triangle

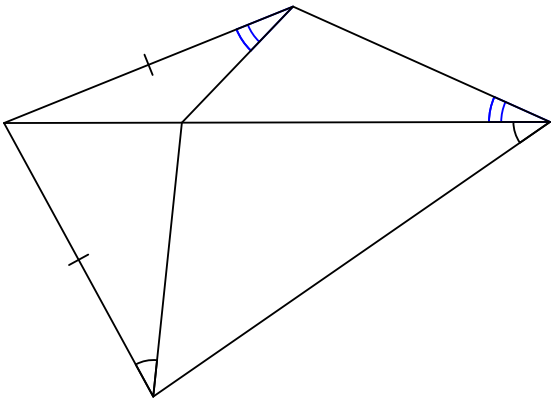


Non-convex quadrilateral





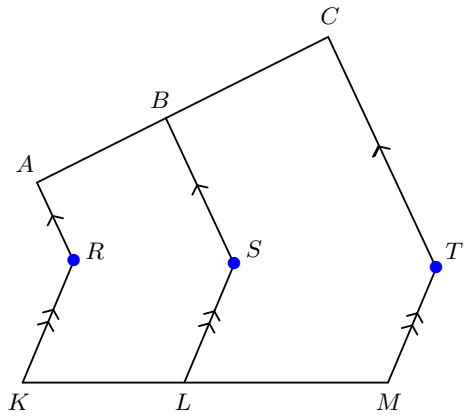
B3 Alternate segment switch ★



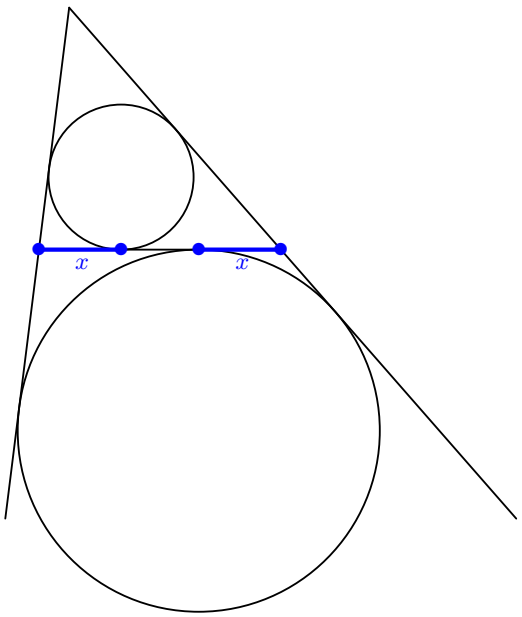
B4 Ratios for collinearity ★

(Points A, R, K are not permitted to be collinear for this diagram.)

R, S, T collinear $\Leftrightarrow$ $\dfrac{AB}{BC} = \dfrac{KL}{LM}$.



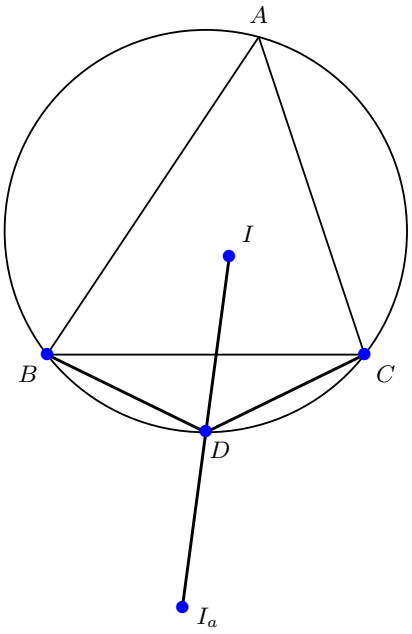
B5 Points of contact of incircle and excircle ★



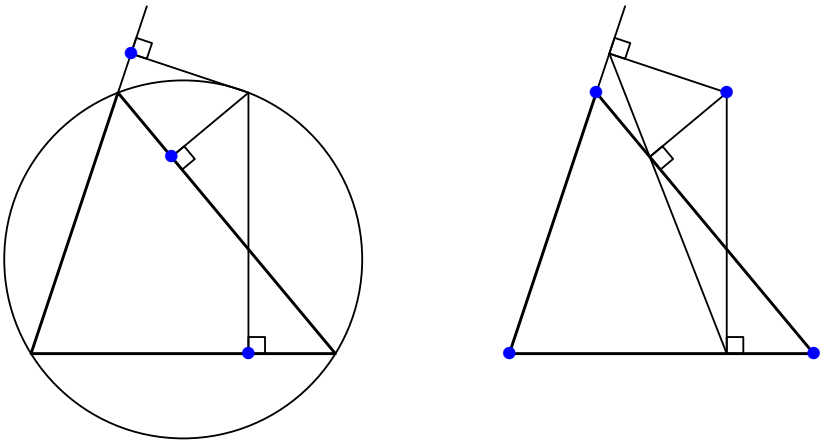
B6 Circumcircle, incentre and excentre ★

I = incentre
I_a = excentre

$DB = DC = DI = DI_a$.



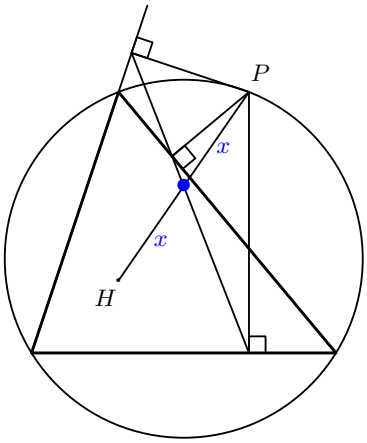
B7 Simson line ★



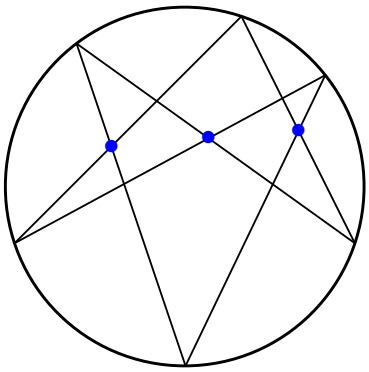
Additional property ★★★

H = orthocentre

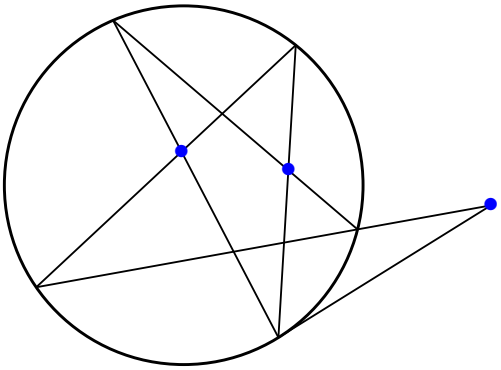
Simson line bisects PH.



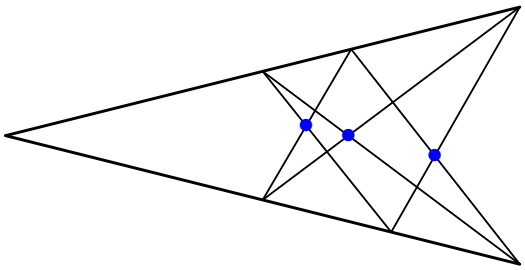
B8 Pascal's theorem ★★★



Two points coincident

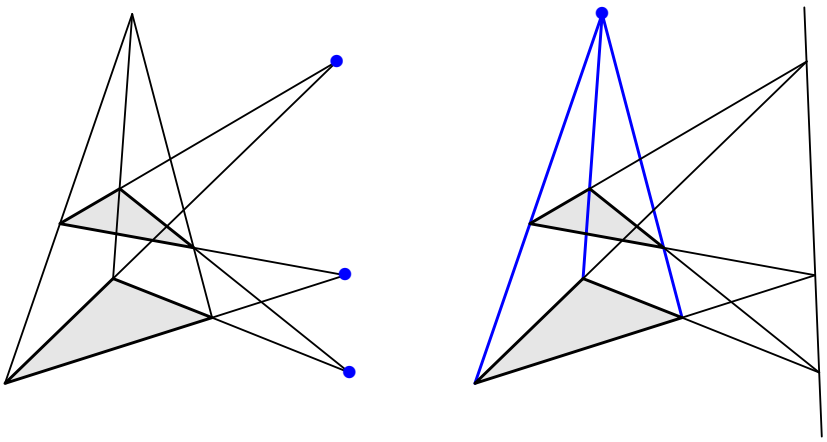


Pappus' theorem



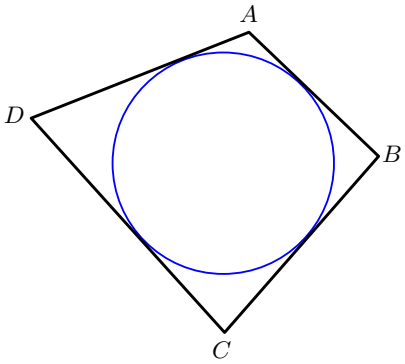
B9 Desargues' theorem ★★★

Triangles are in perspective from a point if and only if they are in perspective from a line.

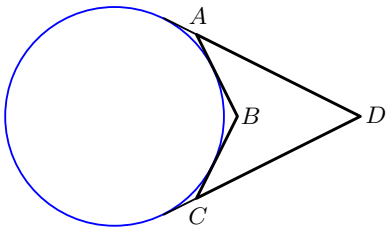


B10 Quadrilateral and incircle ★

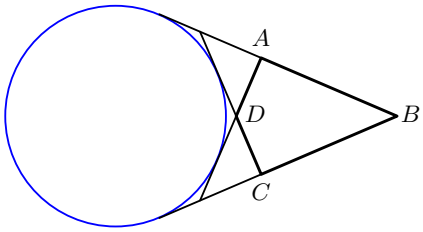
Convex quadrilateral $ABCD$ has an incircle $\Leftrightarrow$ $AB + CD = BC + AD$.



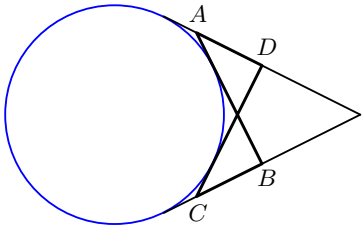
Variations



$AB + CD = BC + AD$



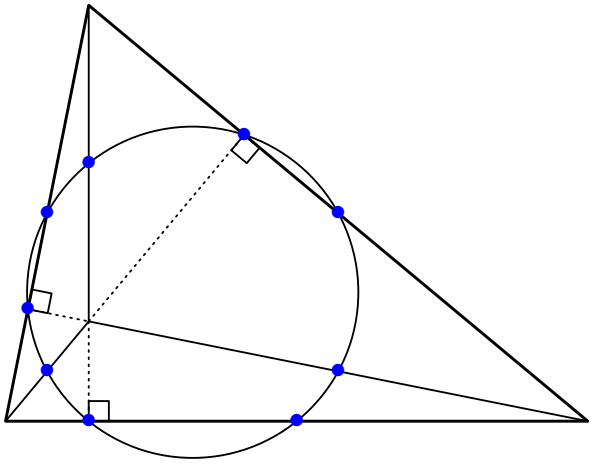
$AB - CD = BC - AD$



$AB - CD = BC - AD$

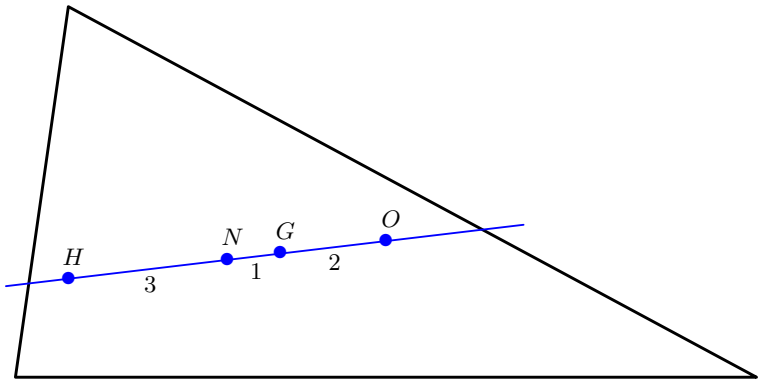
C1 Nine-point circle ★

Midpoints of sides, feet of altitudes and midpoints of segments connecting vertices to orthocentre are all concyclic.

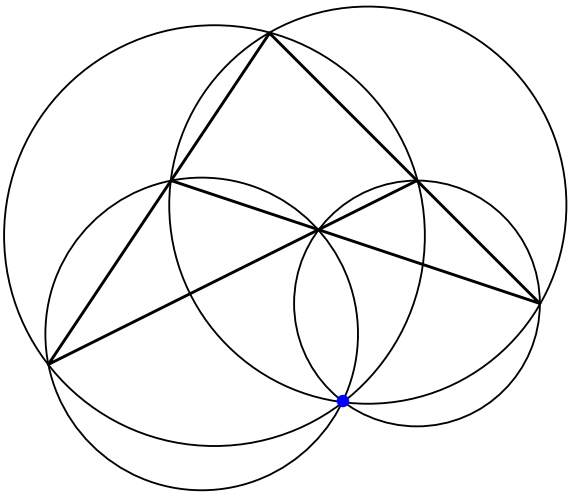


C2 Euler line ★

H = orthocentre
N = nine-point centre
G = centroid
O = circumcentre

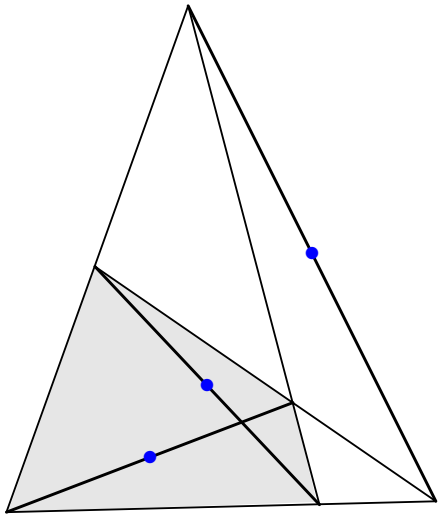


C3 Four lines and four circles ★



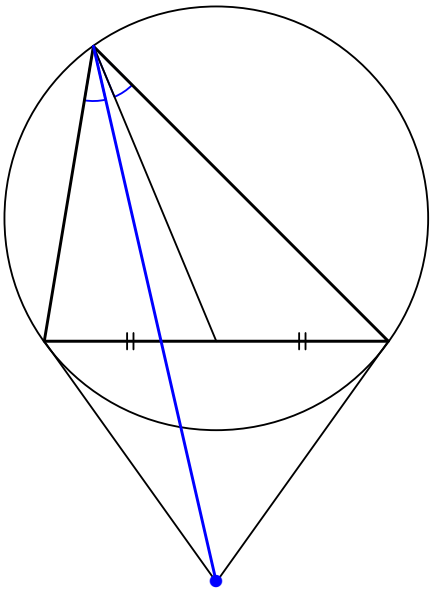
C4 Newton–Gauss line ★★

Midpoints of the diagonals of a complete quadrilateral[3] are collinear.

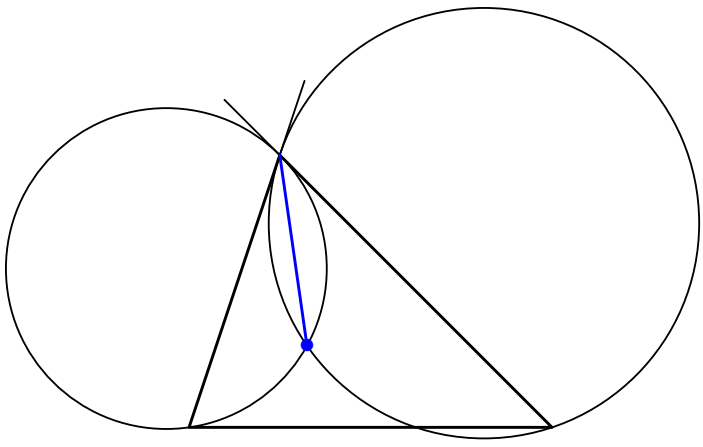


[3] A *complete quadrilateral* is the figure determined by four lines, no three of which are concurrent, and their six points of intersection. The three pairs of points which are not already connected by the original four lines determine the diagonals of the complete quadrilateral.

C5 Alternative characterisation of symmedian[4] ★★



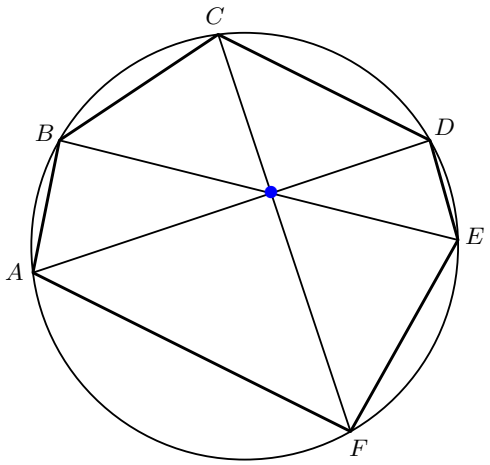
Another characterisation of symmedian via tangent circles ★★



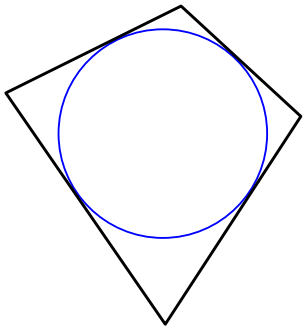
[4]The *symmedian* is the reflection of the median about the angle bisector.

C6 Convex cyclic hexagon and diagonals ★

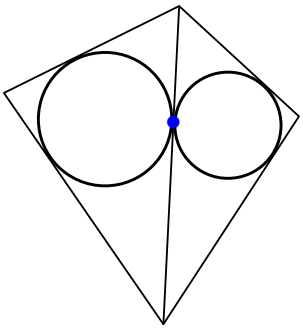
Diagonals concurrent $\Leftrightarrow$ $AB \cdot CD \cdot EF = BC \cdot DE \cdot FA$.



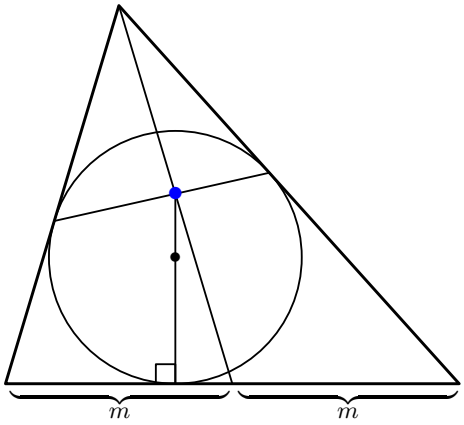
C7 Quadrilateral, triangles and incircles ★



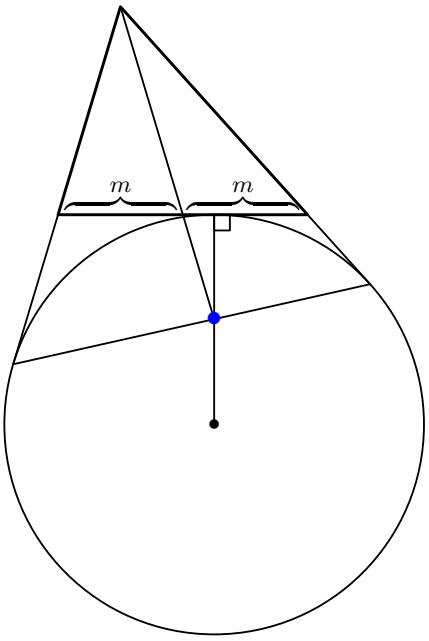
if and only if



C8 Median, inradius and chord of incircle ★★

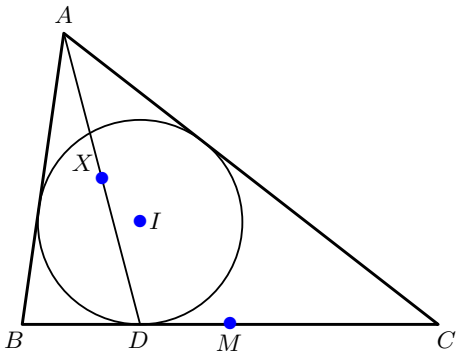


Variation: excentre



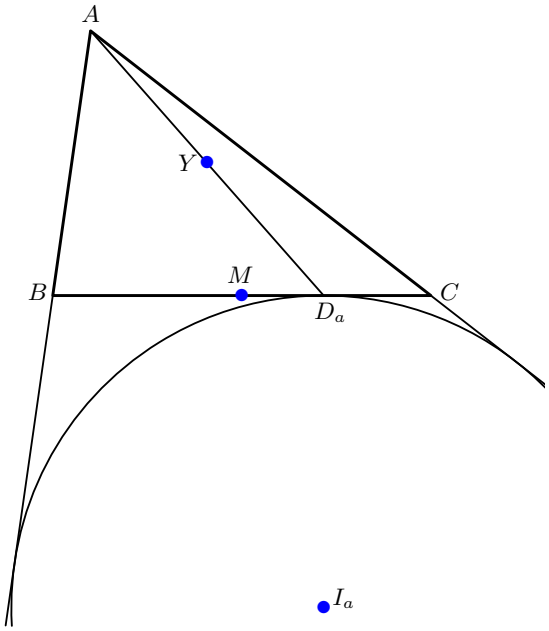
C9 Incentre and midpoints ★★

M = midpoint BC
X = midpoint AD



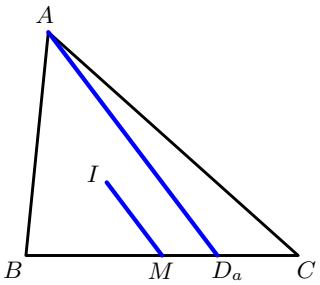
Variation: excentre

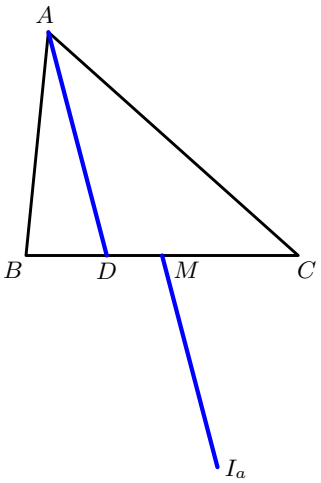
M = midpoint BC
Y = midpoint AD_a



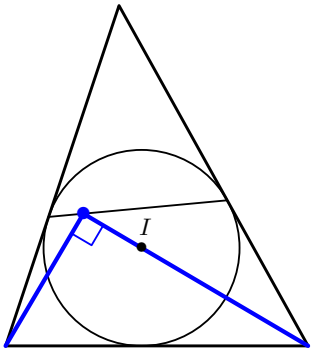
C10 Incentre, excentre, midpoint and contact points ★★

M = midpoint BC
D = incircle contact point
D_a = excircle contact point

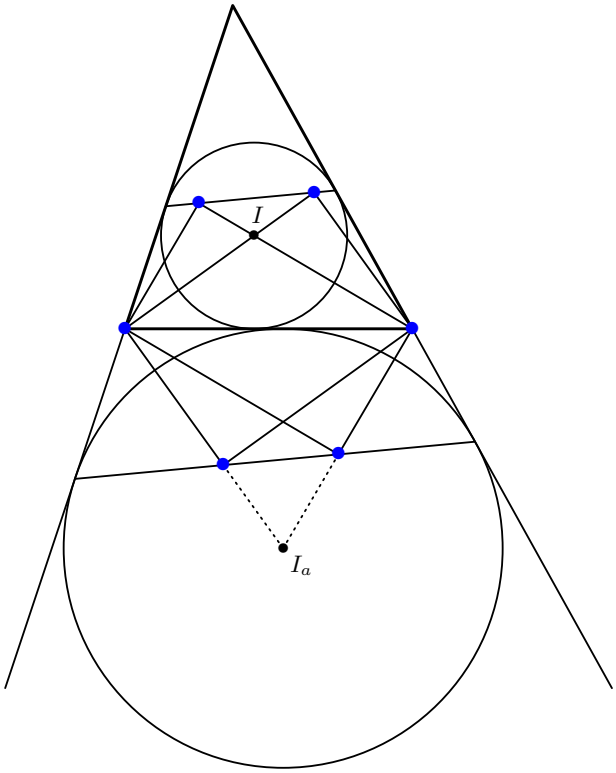




C11 Incentre and chord of incircle ★★

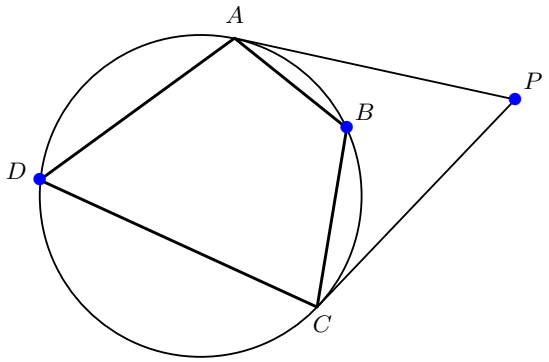


Variation: incentre and excentre



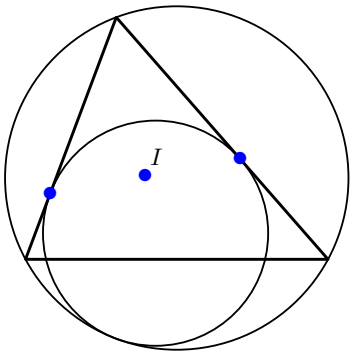
C12 Harmonic quadrilateral[5] ★★★

P, B, D collinear $\Leftrightarrow$ $AB \cdot CD = AD \cdot BC$.

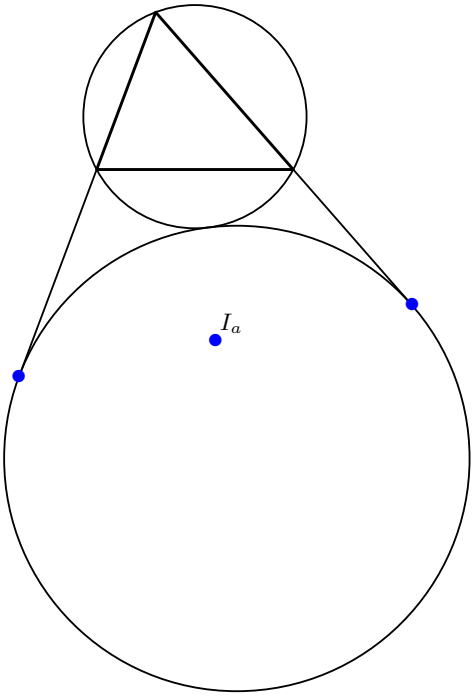


[5] A cyclic quadrilateral is said to be *harmonic* if the products of its opposite sides are equal.

C13 Incentre and mixtilinear incircle[6] ★★★

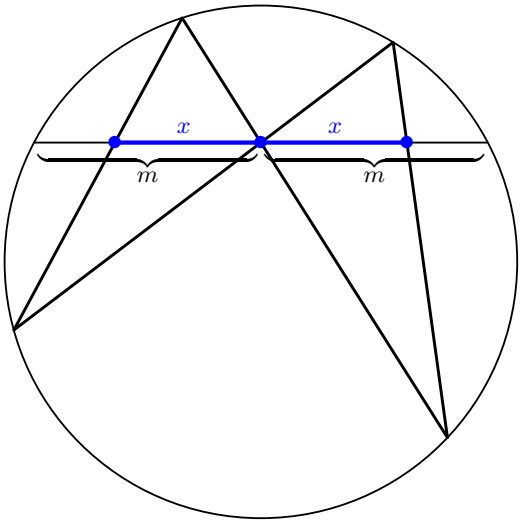


Variation: excentre and mixtilinear excircle



[6] A *mixtilinear incircle* is a circle which is internally tangent to a triangle's circumcircle and two of its sides. There are three mixtilinear incircles associated with any triangle. A mixtilinear excircle is similar but is instead externally tangent to the triangle's circumcircle.

C14 Butterfly theorem ★★★



6

Incidence geometry

An *incidence* basically describes a situation when something more specific occurs than might otherwise be thought. For example, three lines are usually not concurrent, but sometimes you may be asked to prove that three lines are concurrent. So 'three lines concurrent' is an example of an incidence. Other examples include 'three points collinear' and 'four points concyclic'.

Closely related to incidences is the concept of locus. For our purposes, a *locus* is a set of points satisfying some condition. We often think of a locus as the curve traced out by a point following some rule. For example, a circle may be thought of as the locus of points equidistant from a given point. A line may be thought of as the locus of points equidistant from a given line and lying on a given side of the given line.

When asked to find the locus of a set of points it is important to know beforehand what the answer is likely to be. This can be done by drawing a careful diagram and building up a picture of the locus. Most loci tend to be lines or circles, but sometimes ellipses or even whole regions can occur.

An example might be to find the locus of points P such that $\angle APB = 45°$, where A and B are given fixed points. We know that the angle subtended in a circle on one side of a chord is constant, so we can see that the locus is the union of two circular arcs each of which is three-quarters of a full circle.

6.0 Problems

1. Show that the orthocentre and the circumcentre of a triangle are isogonal conjugates of each other.[1]

2. (a) Prove that the medians of a triangle are concurrent.

   (b) Prove that the altitudes of a triangle are concurrent.

   (c) Prove that the angle bisectors of a triangle are concurrent.

   (d) Prove that the lines joining the vertices of a triangle with the points of tangency of the incircle with the opposite sides are concurrent.

[1] See the isogonal conjugates theorem in section 6.3 if you don't know what isogonal conjugates are.

3. Find the locus of points P such that for fixed points A and B,
$$\frac{AP}{BP} = k,$$
where k is a positive constant.

4. Let A, B, C and D be four distinct points on a line, in that order. The circles with diameters AC and BD intersect at the points X and Y. The line XY meets BC at the point Z. Let P be a point on the line XY different from Z. The line CP intersects the circle with diameter AC at the points C and M, and the line BP intersects the circle with diameter BD at the points B and N.
Prove that the lines AM, DN and XY are concurrent.

5. The segment AB is fixed and point M is a variable point on that segment. Squares S_A and S_B are constructed on AM and MB and on the same side of AB.
Find the locus of the midpoint of the segment joining the centres of the two squares.

6. Let circles C_1 and C_2 intersect at A and B. The tangent to C_1 at A meets C_2 again at P. The tangent to C_2 at A meets C_1 again at Q. Let M be the point on line AB such that $AB = BM$.
Prove that A, P, M and Q are concyclic.

7. Let A be a fixed point on a fixed circle Γ. Let B and C be variable points on Γ and let D be a point on BC such that $\frac{AD^2}{BD \cdot DC}$ is a fixed constant.
Find the locus of D as B and C vary over Γ.

8. Of all triangles with given base length and height, which one has the largest inradius?

9. Points A, B and C range over the circumference of a fixed circle.
Find the locus of the incentre of triangle ABC.

10. Let K, L, M and N be four collinear points on the possibly extended sides AB, BC, CD and DA, respectively, of quadrilateral $ABCD$.
Prove that
$$\frac{AK}{KB} \cdot \frac{BL}{LC} \cdot \frac{CM}{MD} \cdot \frac{DN}{NA} = +1,$$
where this is an equation in directed lengths.[2]

11. Let $ABCDEF$ be a cyclic convex hexagon with vertices labelled clockwise.
Prove that
$$AB \cdot CD \cdot EF = BC \cdot DE \cdot FA,$$
if and only if the diagonals AD, BE and CF are concurrent.

12. Let $ABCD$ be a cyclic quadrilateral. Let A_1 and C_1 be the respective feet of the perpendiculars from A and C to BD, and let B_1 and D_1 be the respective feet of the perpendiculars from B and D to AC.
Prove that $A_1B_1C_1D_1$ is cyclic.

13. Let $ABCD$ be a quadrilateral whose diagonals AC and BD are perpendicular. Let P, Q, R and S be the midpoints of the sides, and let T, U, V and W be the feet of the altitudes from these midpoints to the opposite sides.
Prove that P, Q, R, S, T, U, V and W all lie on a circle.

[2] By this we mean that $\frac{AK}{KB}$ is positive if K is between A and B, and negative otherwise.

14. Triangle ABC has sides which satisfy

$$2AB = AC + BC.$$

Prove that the midpoints of AC and BC and the incentre and circumcentre of triangle ABC are concyclic.

15. Let M be the midpoint of the side AC of a triangle ABC and let H be the foot of the altitude from B. Let P and Q be the orthogonal projections of A and C onto the bisector of the angle at B.

Prove that the four points H, P, M and Q lie on the same circle.

16. Show that the tangents to the circumcircle of a triangle at two of its vertices meet on the symmedian from the third vertex.

17. In an acute-angled triangle ABC let AD and BE be altitudes and let AP and BQ be internal angle bisectors. Denote by I and O the incentre and the circumcentre of ABC, respectively.

Prove that D, E and I are collinear if and only if P, Q and O are collinear.

18. Let ABC be an acute-angled triangle and let D, E and F be the feet of the perpendiculars from A, B and C onto the sides BC, CA and AB, respectively. Let P, Q and R be the feet of the perpendiculars from A, B and C onto the lines EF, FD and DE, respectively.

Prove that the lines AP, BQ and CR are concurrent.

19. Let $ABCD$ be a quadrilateral such that its opposite sides AB and CD meet at P and its other opposite sides AD and BC meet at Q. Let X be the intersection of PQ and CA. Let Y be the intersection of PQ and DB.

Show that

$$PX \cdot QY = PY \cdot XQ.$$

20. Let $ABCD$ be a given convex quadrilateral with sides BC and AD equal in length and not parallel. Let E and F be variable points on the sides BC and AD, respectively, and which satisfy $BE = DF$. The lines AC and BD meet at P, the lines BD and EF meet at Q and the lines EF and AC meet at R. Consider all triangles PQR as E and F vary as mentioned earlier.

Show that the circumcircles of all these triangles have a common point other than P.

21. Let I be the incentre of triangle ABC. A circle which is tangent to the circumcircle of triangle ABC on the inside also touches CA and BC at D and E, respectively.

Show that I is the midpoint of DE.

22. Let $ABCD$ be a convex quadrilateral that just happens to have both an incentre I and a circumcentre O.[3]

Prove that the intersection of the diagonals of the quadrilateral lies on the line OI.

[3]Such a quadrilateral is called *bicentric*.

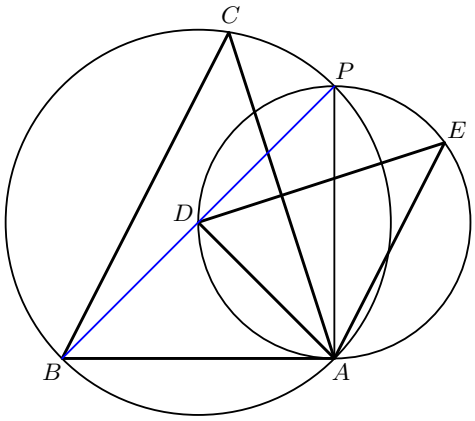
6.1 Collinear points

One way to prove that three points A, B, C are collinear is to prove that $\angle ABC = 0^\circ$ or 180°.

Problem Let ABC and ADE be similar triangles whose vertices are labelled clockwise. Let P be the second common point of the circumcircles of the triangles besides A.

Show that P must lie on the line connecting B and D.

Solution



Look at the diagram. Equal angles abound and we have the chain of equalities

$$\angle BPA = \angle BCA = \angle DEA = \angle DPA.$$

The first follows from the cyclic quadrilateral $ABCP$, the second follows from the similar triangles ABC and ADE, while the third follows from the cyclic quadrilateral $ADEP$. But seeing that B and D lie on the same side of the line AP, the equality $\angle BPA = \angle DPA$ tells us that P must lie on the line passing through B and D. □

Are we done? No, we certainly are not! This is an opportune time to mention a common pitfall in geometry known as *diagram dependence*. We only solved the problem for the diagram shown. It is possible to have other diagrams where the relative positions of the points are different, and our angle chase is a bit different. For instance, if triangle ADE were rotated clockwise until D lay on ray AP beyond P, then it is no longer true that $\angle DEA = \angle DPA$, but instead we would have $\angle DEA = 180^\circ - \angle DPA$. See if you can identify all the different configurations possible and solve in each case.[4]

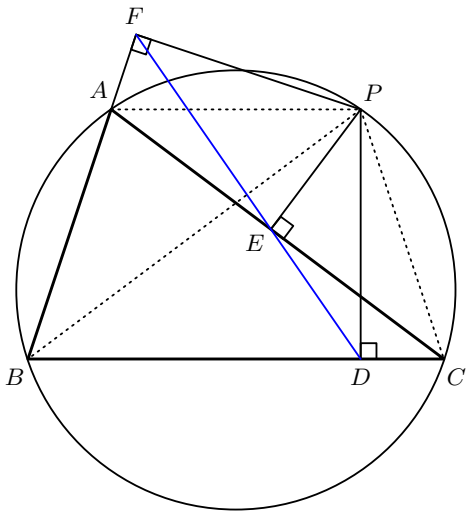
Three other points in the diagram turn out to be collinear. What are they? Prove it!

Problem Let P be a point on the circumcircle of triangle ABC. Let D, E and F be the feet of the perpendiculars from P to the lines BC, AC and AB, respectively.

Show that D, E and F are collinear.

[4] You might like to consult section 17.3 for a possible way to deal with this without resorting to a large number of case distinctions.

Solution



Since BAF is a straight line, it suffices to show that $\angle BFD = \angle AFE$. From cyclic quadrilaterals $PFBD$ and $PFAE$ we know that $\angle BFD = \angle BPD$ and $\angle AFE = \angle APE$. If we subtract $\angle BPE$ from both these angles, then it suffices to prove that $\angle EPD = \angle APB$. However, both these angles are equal to $\angle BCA$—the first is because $CDEP$ is cyclic, and the second is because $BAPC$ is cyclic. □

This solution is not complete since we have not addressed issues of diagram dependence. We leave this for the reader to complete.

The result from the preceding problem can be extended. It is known as *Simson's theorem*.

Simson's theorem Let P be any point in the plane of triangle ABC. Then the feet of the perpendiculars from P to the (possibly extended) sides of the triangle are collinear if and only if P lies on the circumcircle of the triangle. In such a case the line containing the three feet of the perpendiculars is called the *Simson line*.

6.2 Menelaus' theorem

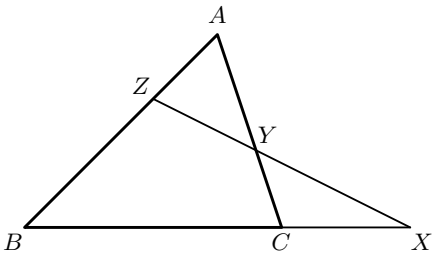
Menelaus' theorem can be used to prove collinearity of points.

Menelaus' theorem If X, Y and Z lie on the three (possibly extended) sides BC, AC and AB of a triangle ABC, then the three points X, Y and Z are collinear if and only if

$$\frac{AZ}{ZB} \cdot \frac{BX}{XC} \cdot \frac{CY}{YA} = -1,$$

where the segments are considered to have directed length.

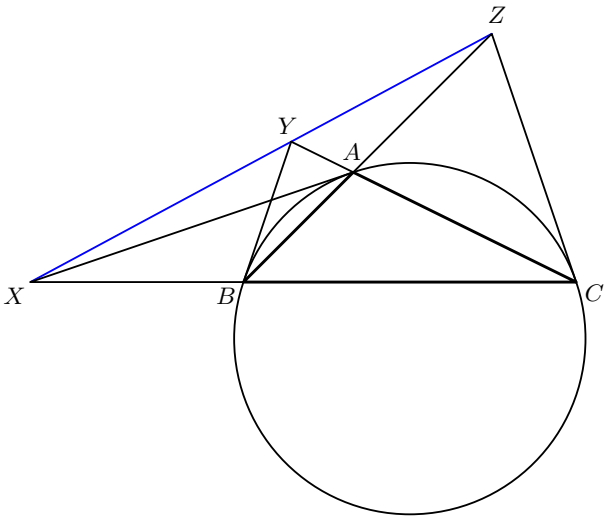
The part about directed lengths in the statement of Menelaus' theorem simply means that the ratios take into account the directions of the vectors $\overrightarrow{AZ}$, $\overrightarrow{ZB}$, and so forth. Thus $\frac{AZ}{ZB}$ is a positive ratio if Z lies on segment AB, and is a negative ratio otherwise.



Problem Suppose that ABC is a triangle with circumcircle Γ in which the three tangents to Γ at A, B and C meet the three opposite sides at X, Y and Z, respectively.

Prove that X, Y and Z are collinear.

Solution



First, triangles XAB and XCA are similar. This follows from the alternate segment theorem, which asserts that $\angle XAB = \angle BCA$. Thus we may write

$$\frac{XA}{XC} = \frac{XB}{XA} = \frac{AB}{AC}.$$

Combining some of these equalities yields

$$\frac{BX}{XC} = -\frac{AB^2}{AC^2}.$$

We compute similar expressions for the other two ratios. These all multiply together and cancel out to give -1. Thus X, Y and Z are collinear by Menelaus' theorem. □

6.3 Concurrent lines

One way of showing that three lines k, ℓ, m are concurrent is to show that the point of intersection of two of the lines lies on the third.

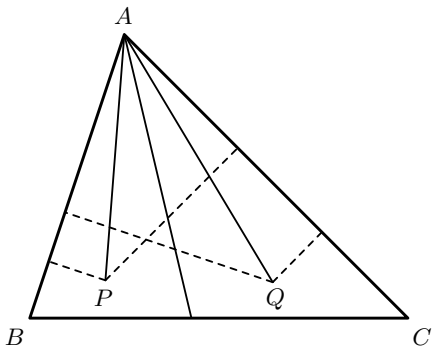
Problem Let P be a point in the plane of a triangle ABC. Reflect the lines PA, PB and PC through the angle bisectors at A, B and C, respectively.

Prove that these three reflected lines are concurrent.

Solution Let p_a, p_b and p_c denote the distances from P to the lines BC, CA and AB, respectively. Similarly, for any point Q in the plane of triangle ABC, we let q_a, q_b and q_c denote the distances from Q to the lines BC, CA and AB, respectively.

Observe that Q lies on the line obtained by reflecting the line PA through the angle bisector at A if and only if

$$\frac{q_b}{q_c} = \frac{p_c}{p_b}.$$



(Can you locate the appropriate similar triangles to establish why this is true?)

Next, define Q to be the point which lies on the reflected line through A and the reflected line through B. Therefore, the previous observation gives us the two equations

$$\frac{q_b}{q_c} = \frac{p_c}{p_b} \quad \text{and} \quad \frac{q_c}{q_a} = \frac{p_a}{p_c}.$$

Multiplying these two equations together, we obtain

$$\frac{q_b}{q_a} = \frac{p_a}{p_b},$$

which implies that Q lies on the reflected line through C as well. Therefore, the three reflected lines are concurrent at Q. □

The result of this problem is a known theorem.

Isogonal conjugates theorem If ℓ_A is a cevian through A of triangle ABC, the *isogonal conjugate* of ℓ_A is found by reflecting ℓ_A in the angle bisector at A to get the line ℓ'_A. The theorem states that three cevians ℓ_A, ℓ_B and ℓ_C are concurrent if and only if their isogonal conjugates ℓ'_A, ℓ'_B and ℓ'_C are concurrent. The respective points of concurrency are called isogonal conjugate points.

6.4 Ceva's theorem

Ceva's theorem is a highly useful criterion often used to prove that three lines concur.

Ceva's theorem If X, Y and Z lie on the three (possibly extended) sides BC, AC and AB of a triangle ABC, then the three lines (called *cevians*) AX, BY and CZ are concurrent if and only if

$$\frac{AZ}{ZB} \cdot \frac{BX}{XC} \cdot \frac{CY}{YA} = +1,$$

where the segments are considered to have directed length.

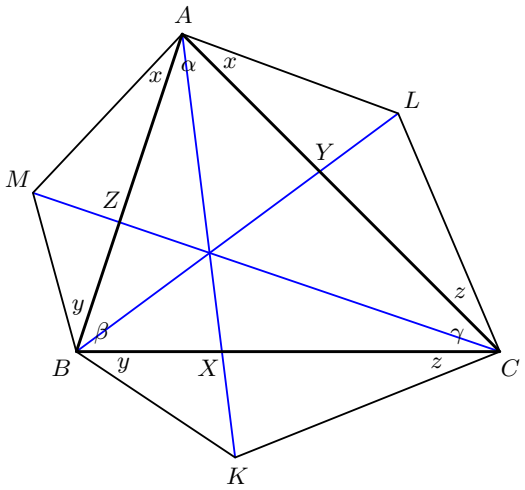
Note that this is exactly the same expression as for Menelaus' theorem except that we have $+1$ on the right-hand side instead of -1.

Problem Outside triangle ABC, points K, L and M are constructed in such a way that

$$\angle MAB = \angle LAC, \quad \angle KBC = \angle MBA \quad \text{and} \quad \angle LCA = \angle KCB.$$

Prove that the three lines AK, BL and CM are concurrent.

Solution Let $x = \angle MAB$, $y = \angle KBC$ and $z = \angle LCA$ and let AK, BL and CM intersect BC, CA and AB at points X, Y and Z, respectively, as in the diagram.



By Ceva's theorem it would suffice to prove that

$$\frac{AZ}{ZB} \cdot \frac{BX}{XC} \cdot \frac{CY}{YA} = +1.$$

Let P denote this product. Note that

$$\frac{BX}{XC} = \frac{|\triangle ABX|}{|\triangle ACX|} = \frac{|\triangle KBX|}{|\triangle KCX|}.$$

Using addendo[5] we have

$$\frac{BX}{XC} = \frac{|\triangle ABK|}{|\triangle ACK|} = \frac{\frac{1}{2}AB \cdot BK \sin(\beta + y)}{\frac{1}{2}AC \cdot CK \sin(\gamma + z)}.$$

[5] *Addendo* is a simple yet highly useful little result. If $r = \frac{a}{b} = \frac{c}{d}$, then also $r = \frac{a+c}{b+d}$.

We obtain similar expressions for the other two ratios and thus compute after much cancelling out that

$$P = \frac{KB}{KC} \cdot \frac{LC}{LA} \cdot \frac{MA}{MB}.$$

Finally, we use the sine rule in triangle KBC to find

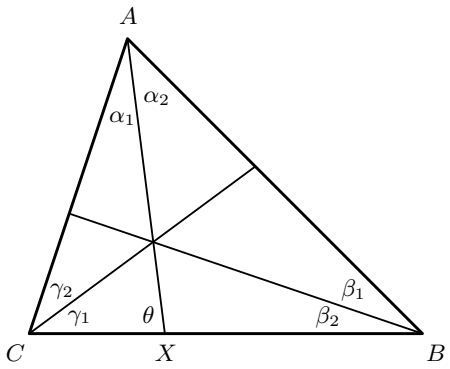
$$\frac{KB}{KC} = \frac{\sin z}{\sin y}.$$

We obtain similar expressions for the other two ratios so that we finally compute that $P = 1$. Therefore, AX, BY and CZ are concurrent by Ceva's theorem. □

There is also a trigonometric version of Ceva's theorem.

Trigonometric form of Ceva's theorem If angles are marked as in the figure, then the cevians are concurrent if and only if

$$\frac{\sin \alpha_1}{\sin \alpha_2} \cdot \frac{\sin \beta_1}{\sin \beta_2} \cdot \frac{\sin \gamma_1}{\sin \gamma_2} = +1.$$



The proof of this is quite straightforward and may be carried out by using the sine rule six times. For example,

$$\frac{CX}{\sin \alpha_1} = \frac{AC}{\sin \theta} \quad \text{and} \quad \frac{XB}{\sin \alpha_2} = \frac{AB}{\sin(180° - \theta)}$$

and so we obtain equations such as

$$\frac{\sin \alpha_1}{\sin \alpha_2} = \frac{AB}{AC} \cdot \frac{CX}{XB}.$$

6.5 Concyclic points

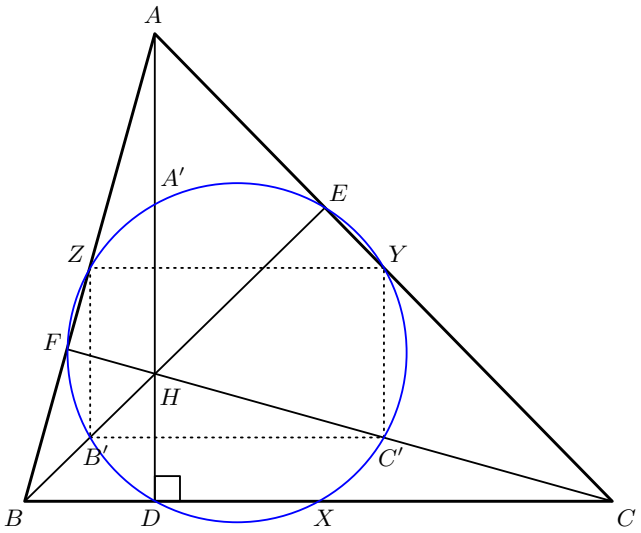
How can you show that four points A, B, C and D are concyclic?

- Find a point O such that $AO = BO = CO = DO$.
- Show that $\angle ABC + \angle ADC = 180°$ and that C and D lie on opposite sides of the line AB.
- Show that $\angle ACB = \angle ADB$ and that C and D lie on the same side of the line AB.

Problem Let ABC be a triangle with altitudes AD, BE, CF, medians AX, BY, CZ, and orthocentre H. Let A', B' and C' be the midpoints of AH, BH and CH, respectively.

Prove that the nine points A', B', C', D, E, F, X, Y and Z all lie on a circle.[6]

Solution



There are midpoints galore in this problem. In fact, six of the nine points which we will prove are concyclic are defined as midpoints. Therefore, it seems like a prime opportunity to use the *midpoint theorem*, which states that if Z is the midpoint of AB and Y is the midpoint of AC, then YZ is parallel to BC and half its length. Applied in triangle ABH, we obtain that $B'Z$ is parallel to AH, while applied in triangle ACH, we obtain that $C'Y$ is also parallel to AH. Applied in triangle ABC, we obtain that YZ is parallel to BC, while applied in triangle HBC, we obtain that $B'C'$ is parallel to BC.

In summary, $B'Z$ and $C'Y$ are parallel to each other and to AH. Furthermore, YZ and $B'C'$ are parallel to each other and to BC. However, since AH is perpendicular to BC, $B'C'YZ$ must be a rectangle. Similar arguments lead to the fact that $C'A'ZX$ and $A'B'XY$ are also rectangles.

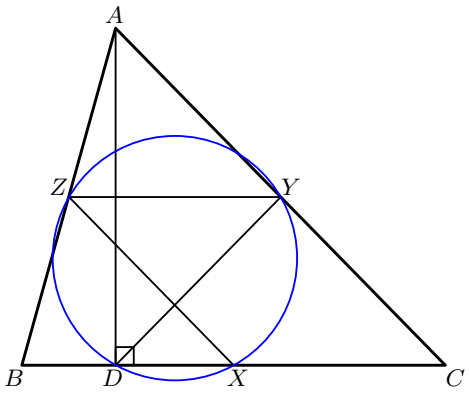
Therefore, if we let N be the midpoint of $B'Y$, then N is the centre of the circle circumscribing the rectangle $B'C'YZ$, as well as the centre of the circle circumscribing $A'B'XY$, both of which have $B'Y$ as diameters. It follows that A', B', C', X, Y and Z all lie on a circle.

Now note that the line YZ bisects AD and is perpendicular to it. In other words, the reflection of A in the line YZ is the point D. To paraphrase again, triangle AYZ is congruent to triangle DYZ. However, triangle AYZ is also congruent to triangle XZY. In particular, we have the equal angles $\angle YDZ = \angle YXZ$, so that the quadrilateral $XYZD$ is cyclic. Therefore, the point D—and by a similar argument, the points E and F—lie on the circumcircle of triangle XYZ. It follows that A', B', C', D, E, F, X, Y and Z all lie on a circle. □

Here is an alternative solution which illustrates that there are many different ways to complete an angle chase. It is also highlights the 'one step at a time' method described in section 4.3.

[6] For obvious reasons, this circle is called the *nine-point circle* of triangle ABC.

Solution Draw the triangle with only the extra points X, Y, Z and D marked.



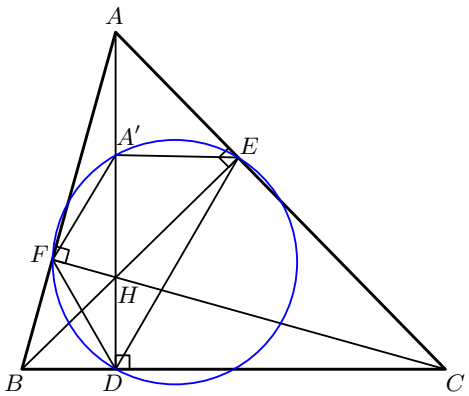
Thus

$$\angle YDC = \angle YCD = \angle XZY.$$

The first equality comes from the fact that triangle ADC is right-angled at D. Hence its circumcentre is the midpoint of AC, namely Y, and so $YA = YD = YC$. The second equality comes from the fact that $XZYC$ is a parallelogram, and thus has opposite angles equal.

Hence $\angle YDX = \angle YDC = \angle XZY$, which establishes that D lies on circle XYZ. Similarly, points E and F lie on circle XYZ. Thus the six points X, Y, Z, D, E and F all lie on the same circle.

Now draw the triangle with only the extra points D, E, F, A' and H marked.



The circle with diameter AH passes through points E and F due to the right angles at E and F. Thus A' is the centre of circle $EAFH$, and so $\angle EA'F = 2\angle A$.

Furthermore, $\angle HDF = \angle HBF = 90° - \angle A$ from cyclic quadrilateral $HDBF$. Similarly, $\angle HDE = \angle HCE = 90° - \angle A$. Therefore,

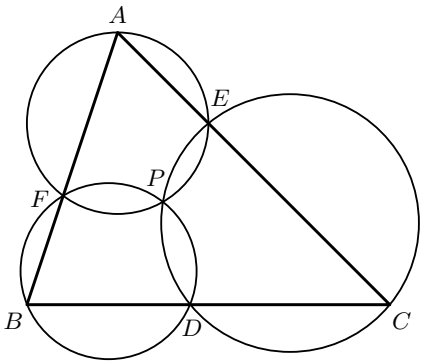
$$\angle EDF = \angle HDF + \angle HDE = 180° - 2\angle A = 180° - \angle EAF.$$

This means that A' lies on circle DEF. Similarly, points B' and C' also lie on circle DEF.

Since the circle through D, E and F is unique and contains the points X, Y, Z, A', B' and C', we conclude that all nine points lie on this same circle. □

Problem Let D, E and F be points on the sides BC, CA and AB of triangle ABC. Prove that the circumcircles of triangles AEF, DBF and DEC are concurrent.

Solution



Let the circumcircles of triangles AEF and DEC meet at P. Then since the quadrilaterals $AFPE$ and $DPEC$ are cyclic, we have

$$\begin{aligned}\angle BFP &= 180^\circ - \angle AFP \\ &= \angle AEP \\ &= 180^\circ - \angle CEP \\ &= \angle CDP \\ &= 180^\circ - \angle BDP.\end{aligned}$$

Since $\angle BFP + \angle BDP = 180^\circ$, the quadrilateral $BFPD$ is also cyclic. Therefore, the circumcircle of triangle DBF also passes through P.[7] □

The result of this problem can be generalised as follows.

Generalised pivot theorem Let A, B, C, D, E, F and P be any seven distinct points in the plane. Consider the following six statements.

- D lies on line BC.
- E lies on line AC.
- F lies on line AB.
- P lies on circle AEF.
- P lies on circle BDF.
- P lies on circle CDE.

If any five of these statements are true, then so is the sixth.

This is an extremely useful result and is well worth remembering. Many geometry problems happen to include this set-up lurking as a subdiagram.

[7] This argument is diagram dependent because it did not address the possibility of P lying outside of the triangle. See section 17.3 to see how to deal with this and more.

6.6 Power of a point

Another useful condition for concyclicity is via the *power of a point theorem*. This is also known as the intersecting chords theorem.

Power of a point theorem If lines AB and CD intersect at P, then A, B, C and D are concyclic if and only if

$$PA \cdot PB = PC \cdot PD,$$

where the lengths are considered to be directed.

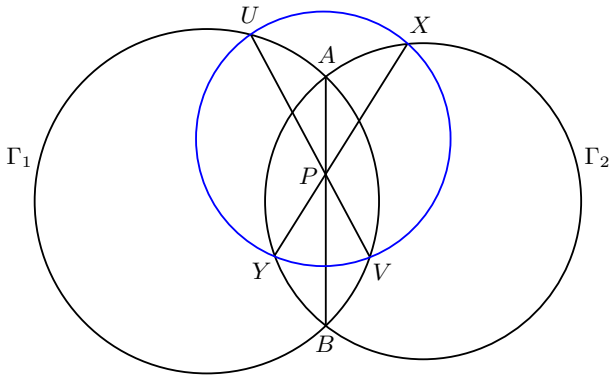
In addition, we define the *power* of a point P with respect to a circle Γ to be the real number $PA \cdot PB$, where P, A and B are collinear and A and B are points on Γ. We treat the lengths as being directed so that the power is positive if P is outside Γ, and negative if P is inside Γ.

Note that the power of a point theorem ensures that this is a well-defined real number. That is, we get the same result no matter how A and B are chosen on Γ, provided that P, A and B are collinear.

Problem Two circles Γ_1 and Γ_2 intersect in two points A and B. Point P lies on the line AB. A line passing through P intersects Γ_1 at U and V. Another line passing through P intersects Γ_2 at X and Y.

Prove that the four points U, V, X and Y are concyclic.

Solution



Using power of a point on Γ_1 we have

$$PU \cdot PV = PA \cdot PB.$$

Using power of a point on Γ_2 we have

$$PX \cdot PY = PA \cdot PB.$$

Thus

$$PU \cdot PV = PX \cdot PY,$$

and so using the power of a point theorem again we deduce that U, V, X and Y are concyclic. □

6.7 Radical axes

Sometimes locus considerations help us to establish incidences as in the following.

We define the *radical axis* of two circles to be the set (locus) of points which have equal power (see section 6.6) with respect to both circles.

Theorem The radical axis of a pair of circles is a straight line.

In the case that the two circles intersect, then you can probably see that their radical axis is simply the line joining their two points of intersection. However, the theorem is still true even if the circles don't intersect.

The radical axis concept is very important, especially in light of the following.

Radical axis theorem Given three circles, the three radical axes associated with the three pairs of circles are either concurrent or parallel.

Problem Prove the radical axis theorem.

Solution Let the three circles be Γ_1, Γ_2 and Γ_3. Let λ be the radical axis of Γ_1 and Γ_2 and let μ be the radical axis of Γ_2 and Γ_3. Suppose that λ and μ intersect at P.[8]

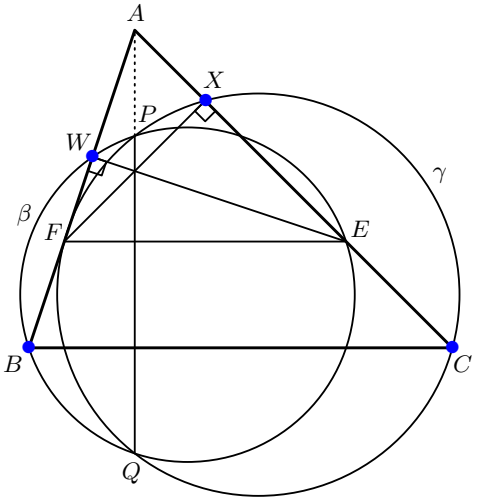
Since P lies on λ, it has equal power with respect to Γ_1 and Γ_2. Similarly, since P lies on μ, it has equal power with respect to Γ_2 and Γ_3. It follows that P has equal power with respect to Γ_1 and Γ_3 and consequently lies on the radical axis of Γ_1 and Γ_3.

Thus the three radical axes all pass through P. □

The concurrence of three radical axes is highly useful. Look out for them!

Problem A line parallel to the side BC of triangle ABC meets AB at F and AC at E. Prove that A lies on the common chord of the circles whose diameters are BE and CF.

Solution



[8] If λ and μ are parallel, then P is undefined and so this argument is faulty. Can you deal with this case?

Let β be the circle with diameter BE and let γ be the circle with diameter CF. Let P and Q be the two points where β and γ intersect. We want to show that line PQ passes through A.

We have two circles, and their common chord which is supposed to pass through A. Note that there are two other obvious lines, namely AB and AC, also passing through A. This looks like a candidate for the radical axis theorem!

We seek a third circle whose radical axes with β and γ are the lines AB and AC, respectively. There is only one such candidate. It should pass through the intersection, W say, of line AB and β and it should also pass through the intersection, X say, of line AC and γ. If we can show that $BWXC$ is cyclic, then applying the radical axis theorem to circles $BWXC$, β and γ will tell us that BW, CX and PQ are concurrent. Since BW and CX intersect at A, this means that PQ would also pass through A.

Thus it remains to show that $BWXC$ is cyclic. Since BE is a diameter of β, we know that $\angle FWE = \angle BWE = 90°$. Thus W lies on the circle with diameter FE. Similarly X lies on the circle with diameter FE. So $FWXE$ is a cyclic quadrilateral.

Therefore,

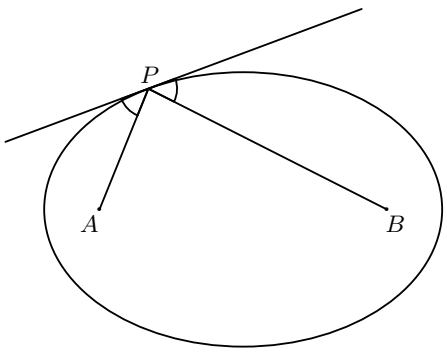
$$\begin{aligned} \angle AWX &= \angle XEF \quad (FWEX \text{ cyclic}) \\ &= \angle ECB \quad (FE \parallel BC). \end{aligned}$$

Thus $\angle AWX = \angle XCB$ implying that $BWXC$ is cyclic as desired. □

If you look at an accurately drawn diagram, it seems that WE, FX and PQ are also concurrent. This is the same as saying the orthocentre of triangle AEF also lies on line PQ. See if you can prove it! You only need the radical axis theorem applied once more. But which circles should you apply it to?

6.8 Ellipses

Ellipses have certain highly useful properties. For example, an ellipse can be thought of as the locus of points P such that the sum $PA + PB$ is a given constant for fixed points A and B. The points A, B are called the *foci*. These two points have a nice optical property: if the inside boundary of an ellipse is made of reflective material, then any light emitted from one focus, A say, will always pass through the other focus B. Thus if P is any point on the ellipse and ℓ is a tangent at P, then the angle that AP makes with ℓ equals the angle that BP makes with ℓ, that is, 'angle of incidence equals angle of reflection'.



Problem Let ABC be an acute triangle.

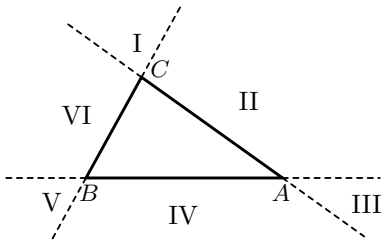
(a) For which point(s) P in the plane is the sum $PA + PB + PC$ minimal?[9]

(b) Show how to construct such a point using a straightedge and compass.

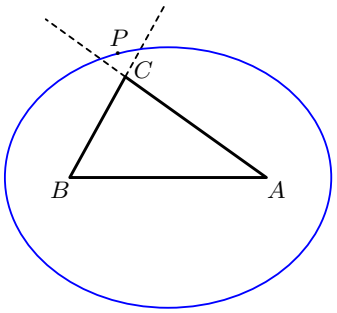
Solution

(a) For any point P in the plane write $f(P) = PA + PB + PC$. We wish to find the minimum value for $f(P)$.

First we investigate the possibility that P might lie outside the triangle. The outside of the triangle can be divided into six regions defined by the lines AB, AC and BC and P could lie in any of these.



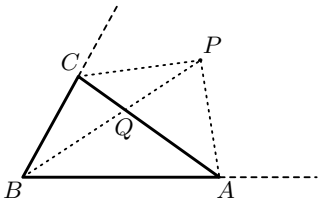
- Case 1: The point P lies in one of the regions I, III or V.
  WLOG P lies in region I.



Consider the ellipse with foci at A and B such that P is on the ellipse. Since triangle PAB lies inside the ellipse and C lies inside triangle PAB, we conclude that C lies strictly inside the ellipse. Consequently, we have $AP + BP > AC + BC$. Therefore,

$$f(P) = AP + BP + CP > AP + BP > AC + BC = f(C).$$

- Case 2: The point P lies in one of the regions II, IV or VI.
  WLOG P lies in region II.



[9] This problem is considered again in section 12.6.

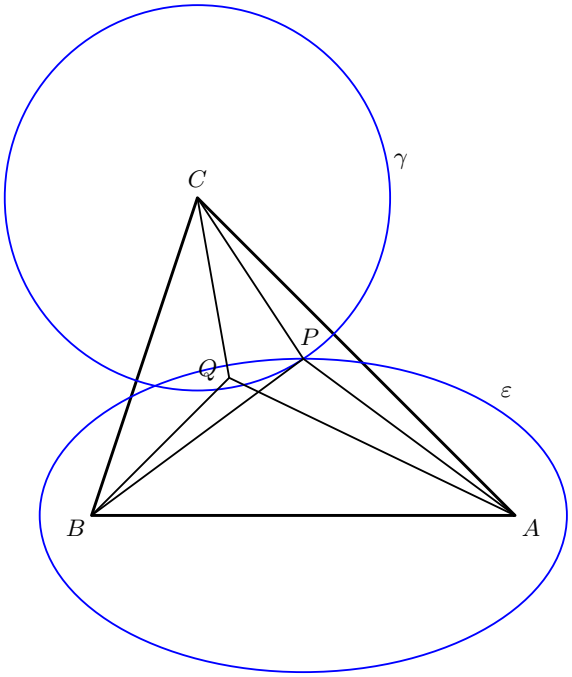
Consider the point Q which is the intersection of segments AC and BP. Then with the help of the triangle inequality $AP + CP > AC$, we have

$$f(P) = AP + CP + BP > AC + BP > AC + BQ = AQ + CQ + BQ = f(Q).$$

So in both cases we have shown that for any point P lying outside the triangle, there exists another point Q on the boundary of the triangle such that $f(P) > f(Q)$. So from here on we may restrict our attention to points not lying outside the triangle.

Consider any position of a point P inside or on the boundary of triangle ABC such that $f(P) = PA + PB + PC$ is minimal.[10] If P were on the boundary of the triangle, it is easy to see that P would have to be a foot of an altitude of the triangle. Consequently, P cannot be at a vertex.

Consider the ellipse ε with foci at A and B and such that P is on ε. Consider also the circle γ centred at C passing through P.



If γ is not tangent to ε, then it intersects ε in at least two points. In this case, any point Q lying strictly inside the intersection of γ and ε would necessarily have $CQ < CP$ and $AQ + BQ < AP + BP$ and so $f(Q) < f(P)$, a contradiction. (Even if Q lay outside the triangle, this would still be a contradiction, because $f(P)$ was taken to be the global minimum.)

Therefore, γ is tangent to ε at P. Let ℓ be the common tangent of γ and ε at P. Since we know the angle of incidence is equal to the angle of reflection at P in ε and we also know that $\ell \perp PC$, this allows us to deduce that $\angle APC = \angle BPC$. Similarly, we deduce that $\angle BPA = \angle CPA$.

Hence all three angles around P are equal. Therefore, P is a point where all three angles are equal to 120°. □

[10] A minimal value exists because f is a continuous function whose domain is now restricted to a closed and bounded subset of the plane.

(b) We construct the point as follows. Erect equilateral triangles and their circumcircles outwards on AB and BC. These circles both subtend $120°$ angles on AB and BC inside triangle ABC and hence their intersection point is the desired point P. □

6.9 Pascal's theorem

Here is another useful tool for proving collinearity or concurrence.

Pascal's theorem Let A, B, C, D, E and F be any six points on any conic section. Let $X = AB \cap DE$, $Y = BC \cap EF$ and $Z = CD \cap FA$. Then X, Y and Z are collinear.

Note that the theorem is true no matter what order the six points appear on the conic. The theorem is also true in its limiting cases. For example, if B is permitted to approach and then coincide with A, then the line AB becomes the tangent at A. In the context of using Pascal's theorem, the tangent at A is often written as AA. Some illustrative diagrams can be found on page 87.

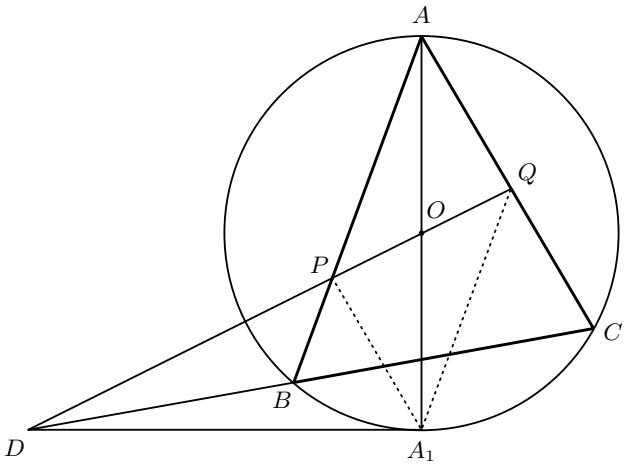
In most applications the conic section[11] will be a circle. Sometimes it might be a pair of lines.[12]

A useful way to remember what intersects what is that AB, DE are pairs of opposite sides of the hexagon $ABCDEF$, as are BC, EF and CD, FA.

Problem Let ABC be an acute triangle with circumcentre O. Let AA_1 be a diameter of the circumcircle of triangle ABC. The tangent line to the circumcircle at A_1 intersects the line BC at point D. The line OD intersects the sides AB and AC of the triangle at points P and Q, respectively.

Prove that $OP = OQ$.

Solution

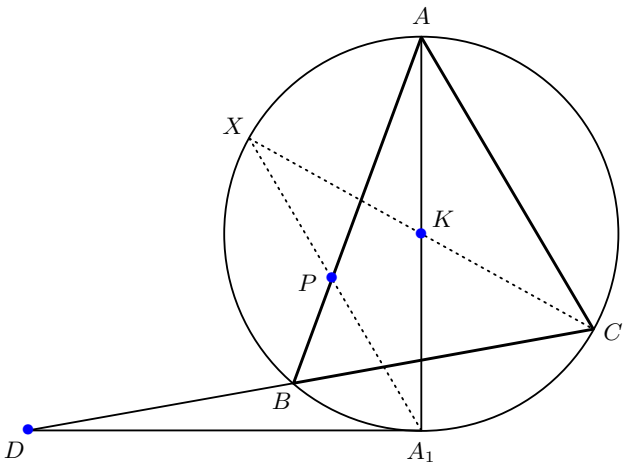


Since $OA = OA_1$, the problem is equivalent to proving that AQA_1P is a parallelogram. To this end it suffices to prove that $PA_1 \parallel AC$. A similar argument will show that $QA_1 \parallel AP$.

[11] Conic sections include circles, ellipses, hyperbolas, parabolas and pairs of straight lines.
[12] This special case of Pascal's theorem is known as *Pappus' theorem*.

Thus we may now forget about point Q.

Let X be the point where A_1P meets the circumcircle for a second time. Note that proving $PA_1 \parallel AC$ is now equivalent to proving that ACA_1X is a rectangle or equivalently that CX passes through point O. Let K be the intersection of AA_1 and CX. We shall prove that $K = O$.



It would suffice to prove that K also lies on DP for then K would lie on both DP and AA_1, ensuring that $K = O$. This is where Pascal's theorem makes an entry. Indeed $K = CX \cap AA_1$, $D = A_1A_1 \cap BC$ and $P = AB \cap A_1X$. It only remains to see if there is a Pascal configuration corresponding to these lines and points.

The Pascal configuration virtually reveals itself! A_1A_1 is a side (corresponding to the tangent of the circle at A_1), A_1X and AA_1 are sides, as are XC and CB. So if the hexagon is AA_1A_1XCB with points in that order, then we apply Pascal's theorem to find that the points $K = AA_1 \cap XC$, $D = A_1A_1 \cap CB$ and $P = A_1X \cap BA$ are collinear as desired. □

7

Transformation geometry

Geometric transformations are often used in an effort to understand a particular geometry problem better. The transformations considered in this chapter are spiral symmetries and affine transformations. Reflections are discussed in section 12.2.

Every spiral symmetry may be described as a linear complex function.[1] Understanding what functions correspond to what spiral symmetries helps to understand their compositions.

An anticlockwise rotation of angle θ composed with a dilation of factor $r \neq 0$ about the same point is called a *spiral symmetry*. If it is performed about the origin, it may be described by

$$f(z) = az,$$

where $a = re^{i\theta} = r\cos\theta + ir\sin\theta$. If it is performed about a general point P represented by the complex number p, then we have $f(z) - p = a(z - p)$. Thus

$$f(z) = az + (1 - a)p.$$

Hence a general spiral symmetry is described by

$$f(z) = az + b$$

for any complex number b provided that $a \neq 0$ or 1. The centre of the spiral symmetry is given by $p = \frac{b}{1-a}$. If $a = 1$, we also have the *identity* spiral symmetry $f(z) = z$. Furthermore, if we throw in the translation functions $f(z) = z + b$ for each complex number b, we have the complete set of linear functions of the complex plane. In loose terms a translation can be thought of as a spiral symmetry whose centre is infinitely far away and whose angle of rotation is 0°. From here on, whenever we speak of a spiral symmetry, we implicitly include the translations amongst them.

To summarise, the full family of spiral symmetries, now also including the translations, is given by the set of functions

$$f(z) = az + b,$$

where $a \neq 0$. The rotation angle is $\theta = \arg(a)$ and the dilation factor is $r = |a|$.

Consider what happens if we compose two such spiral symmetries. Suppose we do

$$f_1(z) = a_1 z + b_1$$

[1] You might like to check out chapter 8 and section 17.1 if you would like a refresher on complex numbers.

followed by

$$f_2(z) = a_2 z + b_2.$$

We obtain

$$f_3(z) = f_2(f_1(z)) = a_3 z + b_3,$$

where $a_3 = a_1 a_2$ and $b_3 = a_2 b_1 + b_2$.

If $a_1 = r_1 e^{i\theta_1}$, $a_2 = r_2 e^{i\theta_2}$ and $a_3 = r_3 e^{i\theta_3}$, then

$$r_3 = r_1 r_2 \quad \text{and} \quad \theta_3 = \theta_1 + \theta_2.$$

Thus we have the following.

Theorem When composing spiral symmetries, we have the following.

- Dilation factors multiply.
- Rotation factors add.

Note also that in general the order of composition *does matter*. So, it would not necessarily be that case that $f_1(f_2(z)) = f_2(f_1(z))$. For instance, the respective centres of the spiral symmetries are usually different.

There are some important subgroups of the family of spiral symmetries.

- *Translations*

  A general translation is represented by $f(z) = z + \alpha$ for any complex number α. It is easy to verify that the composition of two translations is a translation.

- *Rotations*

  A rotation is represented by $f(z) = az + b$ for $|a| = 1$. Note that for our purposes, the family of rotations includes all the translations. It is easy to verify that the composition of two rotations is a rotation.

- *Dilations*

  A dilation is represented by $f(z) = az + b$ for $a \in \mathbb{R}$ and $a \neq 0$. Note that for our purposes, the family of dilations includes all the translations. It is easy to verify that the composition of two dilations is a dilation.

There are also geometric transformations, known as affine transformations, which we will discuss later in the chapter.

7.0 Problems

1. Given three parallel lines, show how to construct with straightedge and compass a point on each line so that the points form an equilateral triangle.
2. If the opposite sides of a hexagon are equal and parallel, prove that the diagonals joining opposite vertices are concurrent.
3. Show that the three medians of a triangle are concurrent.
4. Given three circles in the plane, show how to construct a point on each so that the points form an equilateral triangle. Discuss under what circumstances this is even possible.

5. Given two triangles in the plane which have corresponding sides parallel, show that the lines joining the corresponding vertices are concurrent.

6. Let ABC be a triangle and let D, E and F be the midpoints of BC, AC and AB, respectively.

   Prove that the vectors $\overrightarrow{AD}, \overrightarrow{BE}$ and $\overrightarrow{CF}$ form a triangle.

7. Show how to construct, using straightedge and compass, a square whose vertices all lie on the sides of a given triangle.

8. Two circles are internally tangent at T. A chord AB of the outer circle is tangent to the inner circle at P.

   Prove that TP bisects $\angle ATB$.

9. Two common tangents of two intersecting circles meet at a point A. Let B be a point of intersection of the two circles, and C and D be the points in which one of the tangents touches the circles.

   Prove that the line AB is tangent to the circumcircle of triangle BCD.

10. Let ABC be a triangle. Triangles $A'BC$, $B'CA$ and $C'AB$ are erected externally on the sides of triangle ABC such that

$$\angle AC'B + \angle BA'C + \angle CB'A = 180^\circ.$$

    Let O_A, O_B and O_C be the circumcentres of triangles $A'BC$, $B'CA$ and $C'AB$, respectively.

    Prove that $\angle O_A O_C O_B = \angle AC'B$.

11. Let X and Y be the centres of the squares erected externally on sides AB and AC of triangle ABC. Let M be the midpoint of BC.

    Prove that MX and MY are equal and perpendicular.

12. Triangle ABC has squares $PABK$ and $QACL$ constructed on the exterior of the sides AB and AC, respectively. Let AH be an altitude of triangle ABC with H on BC.

    Prove that A, H and the midpoint of PQ are collinear.

13. Given triangle ABC, we erect similar isosceles triangles BCP, ACR and ABQ externally on sides BC and AC and internally on side AB.

    Prove that $PQRC$ is a parallelogram.

14. Given a quadrilateral $ABCD$, we erect the four equilateral triangles ABP and CDR externally and BCR and ADS internally on its sides.

    Show that $PQRS$ is a parallelogram.[2]

15. Chords AB and CD of circle Γ intersect at a point E inside Γ. The circle ω is internally tangent to the figure bounded by segments AE and EC and arc AC (not containing B or D) of Γ, touching arc AC at point F. A line ℓ containing the centre O of Γ, intersects segments AE and DE at points P and Q, respectively, and satisfies $EP = EQ$. Line EF intersects ℓ at point M.

    Prove that the line through M parallel to the line AB is tangent to Γ.

[2] This is a generalisation of the previous problem.

16. Let M be the midpoint of the altitude of triangle ABC from vertex A. Let I be the incentre of the triangle. Let Y be the point of tangency of the excircle opposite vertex A with side BC.

    Prove that M, Y and I are collinear.

17. The incircle of triangle ABC has centre I and touches BC at D. Let E be the midpoint of BC.

    Prove that the line through E and I passes through the midpoint of the segment AD.

18. Two circles Γ_1 and Γ_2 lie inside and are internally tangent to a third circle Γ at points A_1 and A_2, respectively. A common external tangent touches Γ_1 at T_1 and Γ_2 at T_2. Let P be the intersection of lines A_1T_1 and A_2T_2.

    (a) Prove that P lies on Γ.

    (b) Hence prove that $A_1T_1T_2A_2$ is cyclic.

    (c) Hence also prove that if Γ_1 and Γ_2 intersect at two points, then the extension of their common chord also passes through P.

19. Let $\mathcal{Q} = A_0A_1A_2A_3$ be a quadrilateral in the plane. Given a point M_0 in the plane we define the sequence $M_0, M_1, M_2, \ldots$ of points in the plane as follows. If $n \equiv i \pmod 4$, then M_{n+1} is the point obtained from rotating the point M_n anticlockwise about A_i by $90°$.

    (a) Prove that if $M_{2008} = M_0$ for one point M_0 in the plane, then we also have $M_{2008} = M_0$ for all points M_0 in the plane.

    (b) Suppose that $\mathcal{Q}$ is a parallelogram. Prove that if $M_{2008} = M_0$, then $\mathcal{Q}$ is a square. Under what circumstances is the converse true?

    (c) Suppose that $\mathcal{Q}$ is a general quadrilateral. Find simple necessary and sufficient conditions on $\mathcal{Q}$ such that $M_{2008} = M_0$.

20. Let O, A, B, C, D and E be six points in the plane such that the 10 triangles which have O as a vertex all have area at least 1.

    (a) Prove that one of those 10 triangles has area at least $\sqrt{2}$.

    (b) Is the result still true if we only consider O plus four more points?

21. A point P is permitted to vary over the interior of a fixed triangle ABC. The line AP meets BC in A_1. Points B_1 and C_1 are defined similarly.

    As P varies over the interior of the triangle, what is the maximum possible area of triangle $A_1B_1C_1$?

22. Circles Γ_1 and Γ_2 intersect at points A and B. A common tangent to the two circles touches Γ_1 at P_1 and Γ_2 at P_2. Points M_1 and M_2 are such that line M_1M_2 is the perpendicular bisector of AB, and

    $$\angle P_1M_1M_2 = \angle P_2M_2M_1 = 90°.$$

    Line AM_1 intersects Γ_1 again at N_1. Line AM_2 intersects Γ_2 again at N_2. Lines N_1P_1 and N_2P_2 intersect at point P.

    Prove that BP_1PP_2 is a parallelogram.

23. Let ε be an ellipse inscribed in $\triangle ABC$.

    Prove that the foci of ε are isogonal conjugates.

24. Let $\mathcal{H}$ be a strictly convex hexagon inscribed in an equilateral triangle. Suppose that $\mathcal{H}$ has all of its sides of equal length.

    Prove that the three main diagonals of $\mathcal{H}$ are concurrent.

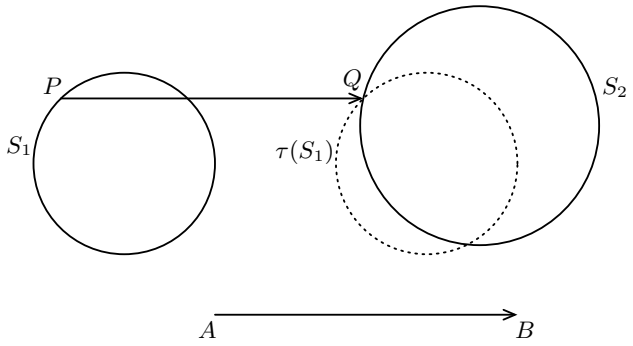
25. Let $ABCD$ be a convex quadrilateral with $BA = BC$. Denote the incircles of triangles ABC and ADC by ω_1 and ω_2, respectively. Suppose that there exists a circle ω tangent to the ray BA beyond A and to the ray BC beyond C, which is also tangent to the lines AD and CD.

    Prove that the common external tangents of ω_1 and ω_2 intersect on ω.

7.1 Translations

Problem Given a segment AB and circles S_1 and S_2, using a straightedge and compass, show how to construct points P and Q, one on each circle, satisfying $PQ \parallel AB$ and $PQ = AB$.

Solution



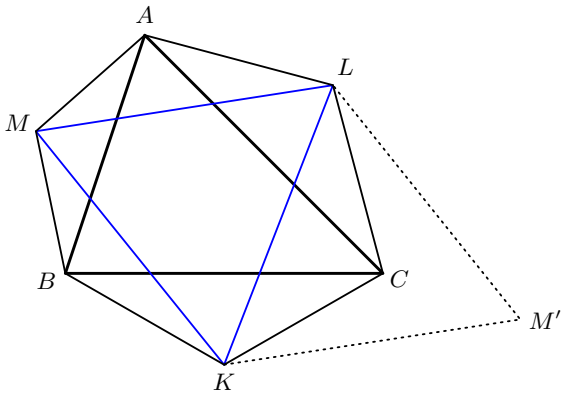
Suppose that P and Q are located as required on the two circles with P on S_1, say. Then the translation τ, which takes A to B, will also take $P \in S_1$ to $Q \in S_2$. Thus if we apply the translation τ to S_1,[3] any place where $\tau(S_1)$ intersects S_2 will be a candidate for Q. Finally, we can invert the translation to recover P from Q. □

In the above problem there may be 0, 1, 2, 3, 4, or infinitely many possible positions for the segment PQ. Can you analyse under what conditions these occur?

7.2 Rotations

Problem Suppose ABC is a triangle. We erect three equilateral triangles PBC, QCA and RAB externally on the sides of triangle ABC. Let K, L and M be their respective centroids. Show that triangle KLM is equilateral.

Solution Note that triangles AMB, BKC and CLA are all isosceles and have 120° angles at M, K and L, respectively. We capitalise on the fact that these three angles sum to 360°.



[3] It is routine to translate a circle using straightedge and compass.

Consider the three 120° clockwise rotations T_M, T_K and T_L about M, K and L, respectively. Note that

$$T_M(A) = B, \quad T_K(B) = C \quad \text{and} \quad T_L(C) = A.$$

We know that the composition $T_L \circ T_K \circ T_M$ of these rotations has angle sum 360° and thus must be a translation. However,

$$T_L(T_K(T_M(A))) = A.$$

Thus the composition is a translation which fixes the point A. Hence the composition must be the identity transformation.

Consider now, what happens to the point M.

- Under T_M, point M remains fixed.
- Under T_K, point M is rotated to some point M', say.
- Finally under T_L, point M' is rotated back to M.

From these observations we see that

$$MK = KM', \quad ML = LM' \quad \text{and} \quad \angle MKM' = \angle MLM' = 120°.$$

It follows that triangles MKL and $M'KL$ are congruent (SSS). Consequently,

$$\angle MKL = \angle M'KL = 60° \quad \text{and} \quad \angle MLK = \angle M'LK = 60°.$$

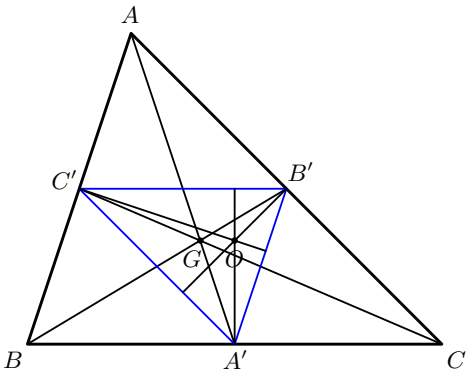
This is sufficient to show that triangle KLM is equilateral. □

7.3 Dilations

One of the important things to remember about dilations is that if X is the centre of a dilation f, then X, Y and $f(Y)$ are always collinear for any point Y.

Problem Prove that the centroid, orthocentre and circumcentre of a triangle are collinear.[4]

Solution



[4]The line common to all three points is called the *Euler line*.

Consider the dilation f with scale factor $-\frac{1}{2}$ about the centroid G of triangle ABC. Note that f maps triangle ABC to its medial triangle $A'B'C'$ which has the same centroid $G' = G$ as triangle ABC.

Let H be the orthocentre of triangle ABC and let $H' = f(H)$ be the orthocentre of triangle $A'B'C'$. The perpendicular bisectors of the sides of ABC are the altitudes of the medial triangle and so $O = H'$. But H, G and $H' = f(H)$ are collinear thanks to the dilation. Thus H, G and O are collinear. □

In fact we have proven more! Namely that H, G and O occur in that order on the line and that $HG : GO = 2 : 1$.

Continuing these ideas, it is possible to deduce that if N is the circumcentre of the medial triangle, then N also lies on the same line in between H and G. Furthermore,

$$HN : NG : GO = 3 : 1 : 2.$$

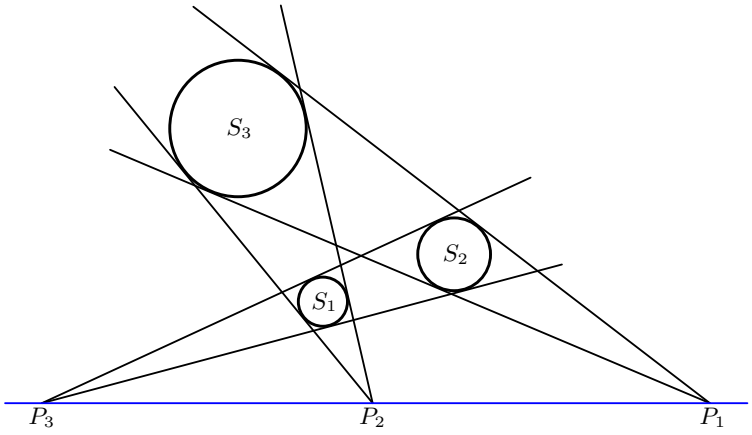
Point N is in fact the centre of the nine-point circle of the triangle.

With a little more thought can you see that there are *two* dilations which transform the nine-point circle into the circumcircle?! One has centre G and dilation factor -2. The other has centre H and dilation factor $+2$.

Problem For a pair of non-congruent circles neither of which lies inside the other, we define their *focal point* to be the point of intersection of their pair of external common tangents. For three circles in the plane, none of which lies inside the other, we thus have three focal points, one for each pair of circles.

Prove that these three focal points are collinear.[5]

Solution



Let the circles be S_1, S_2 and S_3 and let r_1, r_2 and r_3 be their respective radii. Let P_1, P_2 and P_3 be the focal points of the pairs (S_2, S_3), (S_3, S_1) and (S_1, S_2) of circles, respectively.

The intersection of the two common external tangents to a pair of circles is a centre of dilation of the circles, and the dilation factor is equal to the ratio of the two radii.

[5] This is known as *Monge's theorem*. If you understand the proof of it, you should be able to state and prove a similar result involving two internal focal points (the intersection point of the pair of internal tangents) and an external focal point. In fact you should be able to state and prove a result even when one circle is inside another.

With this in mind, let D_1 be the dilation with factor $+\frac{r_3}{r_2}$, centred at P_1. Note that D_1 sends S_2 to S_3. Similarly, let D_2 be the dilation with factor $+\frac{r_1}{r_3}$, centred at P_2 and let D_3 be the dilation with factor $+\frac{r_2}{r_1}$, centred at P_3. Thus

$$D_1(S_2) = S_3, \quad D_2(S_3) = S_1 \quad \text{and} \quad D_3(S_1) = S_2.$$

Note that the composition $D_3 \circ D_2 \circ D_1$ has dilation factor equal to

$$\frac{r_3}{r_2} \cdot \frac{r_1}{r_3} \cdot \frac{r_2}{r_1} = +1$$

and so $D_3 \circ D_2 \circ D_1$ must be a translation. However,

$$D_3(D_2(D_1(S_2))) = S_2.$$

Thus the composition is a translation which fixes the circle S_2. Hence the composition must be the identity transformation.

Consider now the line $\ell = P_1P_2$. Since ℓ passes through P_1, we see that the dilation D_1 leaves ℓ fixed (although points within ℓ do move along ℓ). After next applying D_2 to ℓ we see that ℓ still remains fixed because ℓ passes through P_2. Finally applying D_3 to ℓ must leave ℓ fixed because $D_3 \circ D_2 \circ D_1$ is the identity transformation.

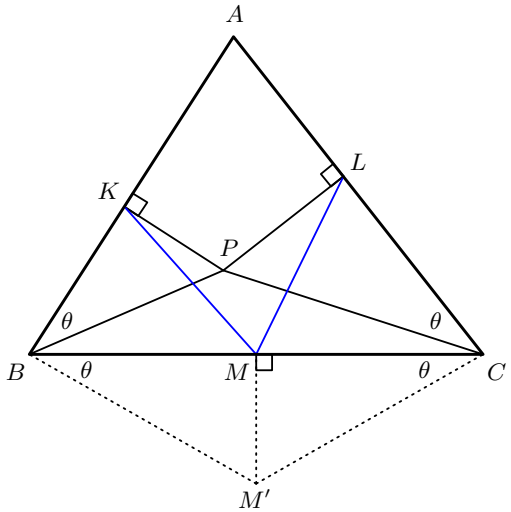
However, if P_3 were not on ℓ, then D_3 would move ℓ. Thus P_3 is also on ℓ and hence collinear with P_1 and P_2. □

7.4 Spiral symmetries

Problem Let P be a point inside triangle ABC such that $\angle PBA = \angle PCA$. Let K and L be the feet of the perpendiculars from P to AB and AC, respectively. Let M be the midpoint of BC.

Prove that $KM = LM$.[6]

Solution Let $\theta = \angle PBK = \angle LCP$, and let $r = \frac{PB}{KB} = \frac{PC}{LC}$. We exploit the fact that triangles PKB and PCL are similar, but oppositely oriented.



[6]This problem is solved using complex numbers in section 8.4.

Consider the spiral symmetry S_C centred at C which takes L to P. Note that S_C has dilation factor r and rotation factor θ. Consider also the spiral symmetry S_B centred at B which takes P to K. Note that S_B has dilation factor $\frac{1}{r}$ and rotation factor θ.

The composition $S_B \circ S_C$ has dilation factor 1 and rotation factor 2θ. It is thus a rotation about some point. If we can show that this point is M, then we are done because the composition takes L to K.

Since the only point fixed by a rotation is the centre of rotation, it suffices to show that M is fixed under the composition. This is rather easy. Indeed suppose that $S_C(M) = M'$, then triangles CMM' and CLP are similar. Thus $MM' \perp BC$. Since M is the midpoint of BC we also have that triangles BMM' and CMM' congruent. Thus $S_B(M') = M$. Hence M is indeed fixed under the composition. □

In fact we have proven not only that $KM = LM$ but also that $\angle KML = 2\theta$, which is even more than we were asked to prove!

7.5 Affine transformations

A *simple affine transformation* can be thought of as a stretch of the plane in one direction. For example, the transformation $(x, y) \mapsto (x, ry)$, where $r > 0$ is a stretch in the direction of the y-axis.

Composing such transformations with spiral symmetries yields the full set of affine transformations. These may be represented as the set of functions $F : \mathbb{R}^2 \to \mathbb{R}^2$, where $(x, y) \mapsto (ax + by + c, dx + ey + f)$ and $ad - bc \neq 0$.

Unlike the family of spiral symmetries, affine transformations do not generally preserve angles. Nevertheless, affine transformations have a number of useful properties.

Properties of affine transformations

1. Collinear points remain collinear.
2. Concurrent lines remain concurrent.
3. Parallel lines remain parallel.
4. Ratios of lengths of **parallel** segments are unchanged.
5. Ratios of **areas** are unchanged.
6. It is possible to transform any ellipse into any circle.
7. It is possible to transform any triangle into any triangle.
8. It is possible to transform any convex quadrilateral into a cyclic quadrilateral whose diagonals intersect at any given angle.

Note that properties 6, 7 and 8 are not independent of each other. For example, if we choose to transform an ellipse into a circle, then we lose the freedom to transform any triangle into any triangle.

Problem Let G be the centroid of triangle ABC and let M be the midpoint of BC. Let X and Y be on AB and AC, respectively, such that the points X, G and Y are collinear and so

that XY is parallel to BC. Suppose that XC and BG intersect at Q and that YB and GC intersect at P.

Show that triangle MPQ is similar to triangle ABC.

Solution It is sufficient to prove that $QM \parallel AC$, $PM \parallel AB$ and $QP \parallel BC$. Conveniently these statements are all invariant under an affine transformation.

Furthermore, the whole problem set-up may be defined purely in terms of information that is invariant under affine transformations. For example, the centroid is just the intersection of the three medians. But a median requires that a midpoint of a side be chosen and a midpoint of a segment remains a midpoint after an affine transformation. The segment XY remains parallel to BC, and so on.

Consider the affine transformation which turns ABC into an equilateral triangle. By the foregoing discussion it suffices to prove the result only for the special case in which ABC is equilateral! This is not particularly difficult and we leave it as an exercise. □

Note that we could just have easily assumed that ABC was a right isosceles triangle and used coordinate geometry.

8

Complex numbers

Many geometry problems can be solved by simply tossing them onto the complex number plane and then doing a few routine calculations. Arithmetical operations involving complex numbers[1] can be seen to have geometric interpretations. Indeed if $\alpha \in \mathbb{C}$ is a constant, we have the following table of interpretations for any complex number z.

Algebra	**Geometry**
$z \pm \alpha$	Translation
$z \cdot \alpha$, $\dfrac{z}{\alpha}$ (for $\alpha \neq 0$)	Spiral symmetry Dilation if $\alpha \in \mathbb{R}$ Rotation if $\lvert\alpha\rvert = 1$
$\overline{z}$	Reflection
$\lvert z\rvert$	Length
$\arg(z)$	Angle
Roots of $z^n = 1$	Vertices of regular n-gon

8.0 Problems

1. Suppose that points A and B in the plane are represented by the complex numbers α and β.

   Find geometric interpretations for the arithmetic, geometric and harmonic means of α and β.

2. Show how to code the following properties using complex numbers.

   (a) Two lengths are equal.

   (b) Two lengths are parallel.

   (c) Two lengths are perpendicular.

   (d) Two angles are equal, or one is twice another.

[1] Please refer to section 17.1 if you would like to brush up on complex number basics.

(e) A length is constant.

(f) An angle is constant.

(g) Two triangles are similar. (Be careful! There are two cases.)

(h) A quadrilateral is cyclic.

3. Given a triangle ABC, find complex number expressions for the triangle's centroid, circumcentre and orthocentre in terms of the vertices. The expressions are particularly simple if the triangle is inscribed in the unit circle of the complex plane.

4. Given triangle ABC, we erect similar isosceles triangles BCP, ACR and ABQ externally on sides BC and AC and internally on side AB.

   Prove that $PQRC$ is a parallelogram.[2]

5. Given a quadrilateral $ABCD$, we erect the four equilateral triangles ABP and CDR externally and BCR and ADS internally on its sides.

   Show that $PQRS$ is a parallelogram.

6. Devise a geometry problem based on the relation

$$(2+i)(3+i) = 5+5i.$$

7. Draw any quadrilateral. On each side draw a square lying outside the given quadrilateral. Draw line segments joining the centres of opposite squares.

   Show that the two line segments are equal in length and perpendicular.

8. Let $\mathcal{Q} = A_0A_1A_2A_3$ be a quadrilateral in the plane. Given a point M_0 in the plane we define the sequence $M_0, M_1, M_2, \ldots$ of points in the plane as follows. If $n \equiv i \pmod 4$, then M_{n+1} is the point obtained from rotating the point M_n anticlockwise about A_i by $90°$.

   (a) Prove that if $M_{2008} = M_0$ for one point M_0 in the plane, then we also have $M_{2008} = M_0$ for all points M_0 in the plane.

   (b) Suppose that $\mathcal{Q}$ is a parallelogram. Prove that if $M_{2008} = M_0$, then $\mathcal{Q}$ is a square. Under what circumstances is the converse true?

   (c) Suppose that $\mathcal{Q}$ is a general quadrilateral. Find simple necessary and sufficient conditions on $\mathcal{Q}$ such that $M_{2008} = M_0$.

9. Let $ABCDE$ be a convex pentagon such that

$$\angle BAC = \angle CAD = \angle DAE \quad \text{and} \quad \angle ABC = \angle ACD = \angle ADE.$$

   The diagonals BD and CE meet at P.

   Prove that the line AP bisects the side CD.

10. Consider a regular n-gon $A_0A_1 \ldots A_{n-1}$ and a point P on its circumcircle.

    (a) Compute the value of

$$\sum_{i=0}^{n-1} PA_i^4.$$

[2] Some of these problems were also seen in *Transformation geometry*. It should not be a surprise that some geometry problems are susceptible to both tactics. After all, the basic properties of spiral symmetries can be explained using complex numbers.

(b) For which points P is the product

$$\prod_{i=0}^{n-1} PA_i$$

maximised or minimised?

(c) What results can you derive if P is allowed to vary over a circle concentric with the circumcircle?

11. (a) Show there is an equiangular (i.e. having all internal angles equal) hexagon with side lengths 1, 2, 3, 4, 5 and 6 in some order.

(b) Do the same for a 15-gon.

(c) Generalise to any n-gon, where n is not a power of a prime.

(d) If n has at least three different prime factors, show that we can construct an equiangular n-gon with side lengths $1^2, 2^2, \ldots, n^2$ in some order.

12. For which integers $n \geq 3$ is it possible to find a convex n-gon with all its interior angles equal but all its side lengths different positive integers?

13. If $P(x)$, $Q(x)$, $R(x)$ and $S(x)$ are all polynomials such that

$$P(x^5) + xQ(x^5) + x^2R(x^5) = (x^4 + x^3 + x^2 + x + 1)S(x),$$

prove that $x - 1$ is a factor of $P(x)$.

14. Suppose

$$(1 + x + x^2 + \cdots + x^{10})^{200} = a_0 + a_1x + \cdots + a_{2000}x^{2000}.$$

Determine the value of

$$a_1 + a_{12} + a_{23} + \cdots + a_{1992}.$$

8.1 Addition ideas

In some contexts, complex numbers may be thought of as vectors. Expressions for the midpoint and centroid are exactly the same as for vectors. Specifically, the midpoint M of AB can be described as

$$\mu = \frac{\alpha + \beta}{2},$$

where α, β and μ are the complex numbers representing the points A, B and M, respectively.

From here on, in a minor abuse of notation, we will simply identify the points with their corresponding complex number. So, for example, not only will M represent the point M of the geometrical configuration, but it will also denote the complex number represented by M after tossing the figure onto the complex plane. Thus, in the above example,

$$M = \frac{1}{2}(A + B).$$

The centroid of a triangle ABC is given by

$$G = \frac{1}{3}(A + B + C).$$

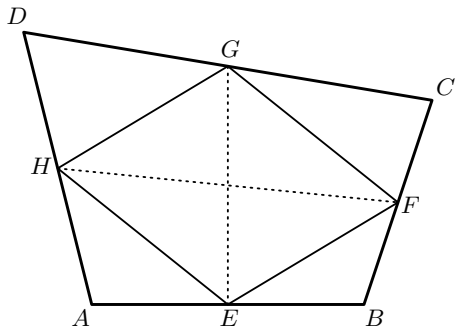
More generally, the centroid of an n-gon $A_1A_2 \ldots A_n$ is given by

$$\frac{1}{n}(A_1 + A_2 + \cdots + A_n).$$

Problem Show that the midpoints of a quadrilateral form a parallelogram whose diagonals intersect at the centroid of the original quadrilateral.

Solution Let the quadrilateral be $ABCD$. Let the midpoints of AB, BC, CD and DA be E, F, G and H, respectively. If we toss the quadrilateral onto the complex plane, then we have

$$E = \frac{1}{2}(A + B), \quad F = \frac{1}{2}(B + C), \quad G = \frac{1}{2}(C + D) \quad \text{and} \quad H = \frac{1}{2}(D + A).$$



We know that $EFGH$ is a parallelogram if and only if EF is equal and parallel to HG. That is, if

$$F - E = G - H.$$

A routine calculation shows in fact that they both equal $\frac{1}{2}(C - A)$. Thus $EFGH$ is indeed a parallelogram.

Finally, we know that the centroid of a general quadrilateral $ABCD$ is just

$$\frac{1}{4}(A+B+C+D).$$

The centroid of a parallelogram is also the intersection of its diagonals.[3] For $EFGH$ this point is

$$\frac{1}{4}(E+F+G+H),$$

which easily simplifies to the same expression as we obtained for $ABCD$. □

8.2 Angles

Given an angle $\angle AOB$, where O is at the origin of the complex plane, we can express

$$\angle AOB = \arg B - \arg A = \arg\left(\frac{B}{A}\right).$$

(This is only correct if $\angle AOB$ is measured anticlockwise. If $\angle AOB$ is measured clockwise, then the correct expression is $\angle AOB = \arg A - \arg B = \arg \frac{A}{B}$.)

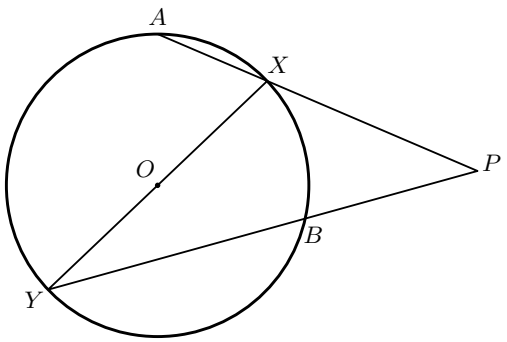
Given four points A, B, X and Y in the plane, how can we express the angle between segments AB and XY? If we translate both these segments to the origin of the complex plane, we find that the required angle is just

$$\arg\left(\frac{B-A}{Y-X}\right).$$

Given a complex number z, how can we find $\arg z$? If $z = re^{i\theta}$, then $\overline{z} = re^{-i\theta}$. Thus $\frac{z}{\overline{z}} = e^{2i\theta}$. This form can be useful when you want to prove that two angles are equal.

Problem Let A and B be two fixed points on a circle Γ and let XY be a variable diameter of Γ. Find the locus of points P defined by the intersection of AX and BY as XY varies over Γ.

Solution A careful diagram suggests that P lies on a circle passing through A and B. We can verify this by showing that $\angle APB$ is constant. This is equivalent to showing that the complex number $z = \frac{A-X}{B-Y}$ has constant argument. This in turn is equivalent to showing that $\frac{z}{\overline{z}}$ is constant.



[3] Try and prove this fact for yourself. It's not hard.

We may assume that the centre O of Γ is the origin of the complex plane. Thus XY being a diameter means that $Y = -X$. Let r be the radius of the circle. Thus for any complex number α lying on the circle, we have $\alpha\overline{\alpha} = r^2$.

We compute

$$\begin{aligned}\frac{z}{\overline{z}} &= \frac{A-X}{B-Y}\cdot\frac{\overline{B}-\overline{Y}}{\overline{A}-\overline{X}}\\ &= \frac{A-X}{B-Y}\cdot\frac{\frac{r^2}{B}-\frac{r^2}{Y}}{\frac{r^2}{A}-\frac{r^2}{X}}\\ &= \frac{AX}{BY}\\ &= -\frac{A}{B}.\end{aligned}$$

But $\arg z = \frac{1}{2}\arg\left(\frac{z}{\overline{z}}\right)$. Thus $\angle APB = \frac{1}{2}\angle AOB$, which is a constant. □

Do not be content to finish here with the problem! The only instance of where we actually used $Y = -X$ was to simplify $\frac{A\cdot X}{B\cdot Y}$ to $-\frac{A}{B}$. In fact, if $\frac{X}{Y}$ is constant, we still derive that $\frac{z}{\overline{z}}$ is constant. But $\frac{X}{Y}$ is constant is the same as saying that XY is a chord of constant length. Thus our solution by complex numbers shows that the problem can in fact be generalised.

Can you discuss the significance of the negative sign?

Can you discuss how we account for both parts, that is, the major and minor arcs of the circle which form the locus of P?

8.3 Multiplication ideas

A rotation about the origin of the complex plane corresponds to multiplication by a complex number of unit magnitude. If we wish to rotate about a point other than the origin, we must first translate to the origin, do the rotation then translate back. Specifically, if A, X and Y are three points such that Y is obtained by rotating X about A through an angle of θ, then we may write

$$Y = A + (X - A)e^{i\theta}.$$

Problem Devise a geometry problem[4] based on the equality

$$(2+i)^3 = 2 + 11i.$$

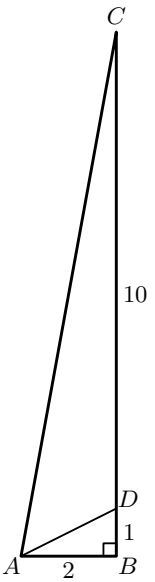
Solution We observe that the relation immediately tells us that

$$\arg(2+11i) = 3\arg(2+i).$$

Thus we may pose the following problem.

> **Problem** Let ABC be a triangle such that $AB \perp BC$, $AB = 2$ and $BC = 11$. Let D be a point on side BC such that $BD = 1$.
>
> Prove that $\angle DAB = \frac{1}{3}\angle CAB$.

[4] There are lots of complex number relations that give rise to potential geometry problems. You might like to try and find some more.



The solution would then run as follows.

Solution Toss triangle ABC onto the complex plane so that

$$A = 0, \quad B = 2 \quad \text{and} \quad C = 2 + 11i.$$

Then $D = 2 + i$. We can compute that

$$D^3 = (2+i)^3 = 2 + 11i = C.$$

Therefore,

$$\arg(C) = \arg(D^3) = 3\arg(D).$$

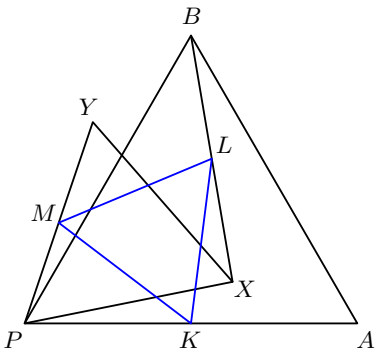
This means that $\angle CAB = 3\angle DAB$. □

Problem Let PAB and PXY be equilateral triangles oriented the same way. Let K, L and M be the midpoints of PA, BX and PY, respectively.

Prove that triangle KLM is equilateral.

Solution Toss the figure onto the complex plane so that $P = 0$. Let $\omega = e^{\frac{\pi i}{3}}$. Multiplication by ω corresponds to a rotation of 60° anticlockwise. Thus we may write

$$B = \omega A \quad \text{and} \quad Y = \omega X.$$



We may also write

$$K = \frac{1}{2}A, \quad L = \frac{1}{2}(\omega A + X) \quad \text{and} \quad M = \frac{1}{2}\omega X.$$

Triangle KLM is equilateral if and only if

$$\begin{aligned} & L - M = \omega(K - M) \\ \Leftrightarrow \quad & \frac{1}{2}\omega A + \frac{1}{2}X - \frac{1}{2}\omega X = \frac{1}{2}\omega A - \frac{1}{2}\omega^2 X \\ \Leftrightarrow \quad & X(\omega^2 - \omega + 1) = 0. \end{aligned}$$

Since

$$\omega = e^{\frac{\pi i}{3}} = \cos 60^\circ + i \sin 60^\circ = \frac{1}{2}(1 + i\sqrt{3}),$$

it is easy to check that $\omega^2 - \omega + 1 = 0$, as desired. □

8.4 Similarity ideas

Suppose we wish to specify a triangle ABC up to similarity. Then it is enough to specify one ratio of side lengths and the measure of the angle between these sides.

Let's toss ABC onto the complex plane so that A is situated at the origin of the complex plane. The angle at A is the same as $\arg B - \arg C = \arg(\frac{B}{C})$. The ratio $\frac{AB}{AC}$ is the same as $\frac{|B|}{|C|} = |\frac{B}{C}|$. These two pieces of information are simply the polar form of $\frac{B}{C}$. Thus it suffices to know only the value of $\frac{B}{C}$. If A is not at the origin, then translating the triangle so that A is at the origin, shows that the similarity type of ABC depends only on $\frac{B-A}{C-A}$.

In particular, suppose that we wish to locate B such that $\angle BAC = \theta$ and $AB = rAC$. If A were at the origin, we would have $B = zC$, where $|z| = r$ and $\arg z = \theta$. If A is not at the origin, we find after translating to the origin that $B - A = z(C - A)$. Thus

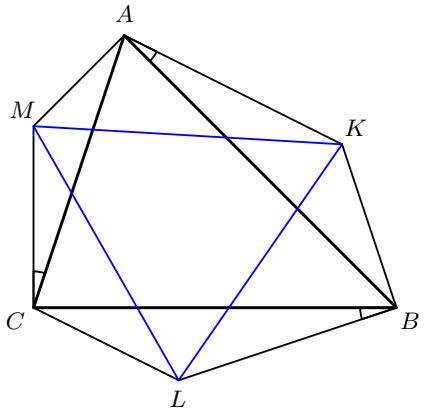
$$B = zC + (1 - z)A$$

and the complex number z determines the similarity type of triangle ABC.

Problem Let similar triangles AKB, BLC and CMA be constructed on the exterior of triangle ABC.

Prove that the centroids of triangles ABC and KLM coincide.

Solution



Let z be the complex number such that $\arg z = \angle BAK$ and $|z| = \frac{AK}{AB}$. Due to the similarity of the three triangles we may write

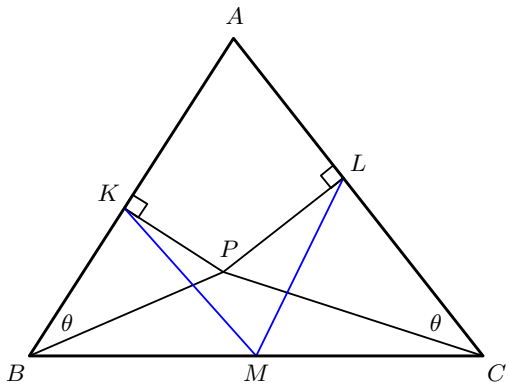
$$K = zB + (1-z)A, \quad L = zC + (1-z)B \quad \text{and} \quad M = zA + (1-z)C.$$

The centroid of triangle KLM is just $\frac{1}{3}(K+L+M)$, which upon substitution in terms of A, B, C and z, simplifies to $\frac{1}{3}(A+B+C)$ and this is the centroid of triangle ABC. □

Problem Let P be a point inside triangle ABC such that $\angle PBA = \angle PCA$. Let K and L be the feet of the perpendiculars from P to AB and AC, respectively. Let M be the midpoint of BC.

Prove that $KM = LM$.[5]

Solution Toss the figure onto the complex plane.



Triangles PKB and PLC are similar but are oppositely oriented. Assume that $\frac{BK}{PK} = \frac{CL}{PL} = r$. Point B is obtained by rotating P about K by 90° clockwise and dilating by a factor r. This corresponds to multiplying by the complex number $-ir$. Thus

$$B = (-ir)P + (1-(-ir))K = -irP + (1+ir)K.$$

Similarly, point C is obtained by rotating P about L by 90° anticlockwise and dilating by a factor r. This corresponds to multiplying by the complex number ir. Thus

$$C = irP + (1-ir)L.$$

Finally, $M = \frac{1}{2}(B+C)$, because it is the midpoint of BC. Therefore,

$$M = \frac{1}{2}((1+ir)K + (1-ir)L).$$

Now $KM = LM$ if and only if

$$\begin{aligned} & |K-M| = |L-M| \\ \Leftrightarrow \quad & \frac{1}{2}\,|(1-ir)K - (1-ir)L| = \frac{1}{2}\,|(1+ir)L - (1+ir)K| \\ \Leftrightarrow \quad & |K-L||1-ir| = |K-L||1+ir|. \end{aligned}$$

This is true because r is a real number and so $1-ir$ and $1+ir$ are complex conjugates with common magnitude $\sqrt{1+r^2}$. □

[5]This problem was solved using spiral symmetries in section 7.4.

Note that in the above solution point A was not used at all in the calculation. This is because the important part consisted of the two similar triangles PKB and PLC. In such a calculational solution it is probably better only to draw in what you need to complete the calculation. In this case point A and the segments AK and AL could be omitted.

8.5 Roots of unity

The *nth roots of unity*, that is, all n complex roots of the equation $x^n = 1$, are the vertices of a regular n-gon centred at 0. Thus questions about regular polygons can often be posed in terms of roots of unity.

If $\omega \neq 1$ is a root of unity, then ω is a root of

$$x^{n-1} + x^{n-2} + \cdots + 1 = 0.$$

Since $|\omega| = 1$, it is also true that

$$\overline{\omega} = \frac{1}{\omega} = \omega^{n-1}.$$

The set of nth roots of unity may be given by $\{e^{\frac{2\pi ik}{n}} \mid k = 0, 1, \ldots, n-1\}$. Thus if $\omega = e^{\frac{2\pi i}{n}}$, then all the other nth roots of unity are just the powers of ω, that is, $1, \omega, \omega^2, \ldots, \omega^{n-1}$.

Problem Let $A_0A_1 \ldots A_{n-1}$ be a regular n-gon of circumradius r.

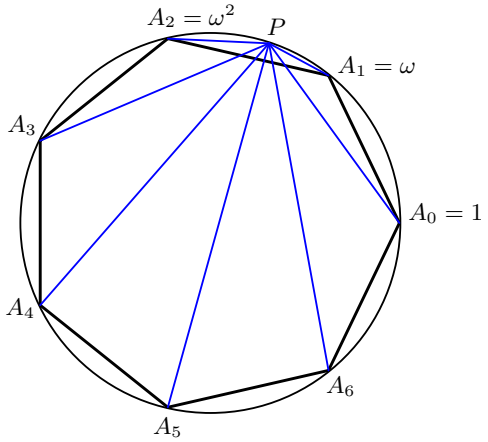
Prove that for any point P on the circumcircle of the n-gon, the value of

$$\sum_{k=0}^{n} PA_k^2$$

is independent of P and find this value.

Solution Dilate the n-gon by a factor of $\frac{1}{r}$ and toss it onto the complex plane so that each A_k is an nth root of unity. Specifically, $A_k = \omega^k$, where $\omega = e^{\frac{2\pi i}{n}}$. Furthermore, since P is a point on the circumcircle we may identify P with a complex number ρ satisfying

$$\rho\overline{\rho} = |\rho|^2 = 1.$$



Recall that

$$1 + \omega + \omega^2 + \cdots + \omega^{n-1} = 0.$$

We may now compute as follows.

$$\begin{aligned} PA_k^2 &= |\rho - \omega^k|^2 \\ &= (\rho - \omega^k)\overline{(\rho - \omega^k)} \\ &= (\rho - \omega^k)(\overline{\rho} - \overline{\omega}^k) \\ &= \rho\overline{\rho} + (\omega\overline{\omega})^k - \rho\overline{\omega}^k - \overline{\rho}\omega^k \\ &= |\rho|^2 + |\omega|^{2k} - \rho\overline{\omega}^k - \overline{\rho}\omega^k \\ &= 2 - \rho\overline{\omega^k} - \overline{\rho}\omega^k. \end{aligned}$$

Thus

$$\begin{aligned} \sum_{k=0}^{n-1} PA_k^2 &= \sum_{k=0}^{n-1} \left(2 - \rho\overline{\omega^k} - \overline{\rho}\omega^k\right) \\ &= 2n - \rho\sum_{k=0}^{n-1} \omega^{n-k} - \overline{\rho}\sum_{k=0}^{n-1} \omega^k \\ &= 2n - \rho\sum_{k=0}^{n-1} \omega^k - \overline{\rho}\sum_{k=0}^{n-1} \omega^k \\ &= 2n - 0 - 0 \\ &= 2n. \end{aligned}$$

Thus for a regular n-gon of unit circumradius the sum required is $2n$. If the circumradius is r, then the answer is $2nr^2$, which is independent of P. □

Problem Let $p \geq 3$ be a prime number and let $\mathcal{P}$ be a convex p-gon with all its interior angles equal and all its side lengths positive integers.

Prove that $\mathcal{P}$ is regular.

Solution Suppose that the side lengths of P in order were $a_0, a_1, \ldots, a_{p-1}$. It would then follow that

$$\sum_{k=0}^{p-1} a_k\omega^k = 0,$$

where $\omega = e^{\frac{2\pi i}{p}}$. (Can you see why?)

So ω is a root of the polynomial

$$f(x) = \sum_{k=0}^{p-1} a_k x^k = 0.$$

Yet ω, being a root of unity, is also a root of

$$g(x) = \sum_{k=0}^{p-1} x^k = 0.$$

Thus ω is a root of

$$d(x) = \gcd(f(x), g(x)).$$

However, $g(x)$ is irreducible.[6] Thus $d(x) = g(x)$ and so $g(x)$ is a factor of $f(x)$. Since $f(x)$ and $g(x)$ have the same degree, we must have $f(x) = a_0 g(x)$. Consequently, $a_0 = a_1 = \cdots = a_{p-1}$ and therefore, P is regular. □

Problem Suppose that

$$(1 + x + x^2 + \cdots + x^{10})^{200} = a_0 + a_1 x + \cdots + a_{2000} x^{2000}.$$

Determine the sum

$$a_0 + a_{11} + a_{22} + \cdots + a_{1991}.$$

Solution Let

$$p(x) = \sum_{j=0}^{2000} a_j x^j = (1 + x + \cdots + x^{10})^{200}.$$

Let $\omega = e^{\frac{2\pi i}{11}}$. Note that $p(w^k) = 0$ provided that $11 \nmid k$.

We now substitute $x = \omega^k$ for $k = 0, 1, 2, \ldots, 10$ into $p(x)$. This yields the following 11 equations.

$$\begin{aligned}
\sum_{j=0}^{2000} a_j &= 11^{200} \\
\sum_{j=0}^{2000} a_j \omega^j &= 0 \\
\sum_{j=0}^{2000} a_j \omega^{2j} &= 0 \\
&\vdots \\
\sum_{j=0}^{2000} a_j \omega^{10j} &= 0
\end{aligned}$$

On the LHS we see that the coefficient of a_j is

$$\sum_{k=0}^{10} \omega^{jk} = \sum_{k=0}^{10} (\omega^j)^k.$$

If $11 \nmid j$, then this sum is just zero. If $11 \mid j$, then this sum is 11.

Thus

$$\text{LHS} = 11(a_0 + a_{11} + a_{22} + \cdots + a_{1991}).$$

On the RHS we simply get 11^{200}. Thus

$$a_0 + a_{11} + a_{22} + \cdots + a_{1991} = 11^{199}.$$ □

[6] To prove that $g(x) = 1 + x + x^2 + \cdots + x^{p-1}$ is irreducible over $\mathbb{Z}$ whenever p is a prime, take the change of variables $x = y + 1$, expand the brackets, then use the *upstairs–downstairs* technique modulo p from section 9.10.

9

Polynomials

Polynomials are just algebraic expressions of a particular type. However, they actually constitute much more than a boring branch of algebra. In fact, polynomials have many interesting properties and many amazing connections to other flavoursome areas of mathematics. This need for lateral thinking provides a fertile breeding ground for mathematical problems, and in this chapter we'll sample a variety of those goodies. But in order to proceed, you'll need some basic knowledge about polynomials, but not too much. Let's start by introducing a bit of jargon, while various results concerning polynomials will appear, mostly without proof, scattered throughout the chapter.

- A *polynomial* is an expression of the form

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0,$$

  where $a_0, a_1, \ldots, a_n$ are numbers called *coefficients* and x is a variable. If the coefficients are integers, then we call $P(x)$ an *integer polynomial* or a polynomial over $\mathbb{Z}$, and similarly for $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$.

- We refer to a_0 as the *constant term* and we refer to a_n as the *leading coefficient*, provided that it is non-zero. If a_n is equal to 1, then we say that the polynomial is *monic*.

- If a_n is non-zero, then we refer to n as the *degree* of the polynomial, and denote it by $\deg P(x)$. By convention, the zero polynomial has degree equal to $-\infty$. Note that the degrees of two polynomials behave nicely under multiplication and addition.

$$\begin{aligned} \deg P(x)Q(x) &= \deg P(x) + \deg Q(x) \\ \deg(P(x) + Q(x)) &\leq \max(\deg P(x), \deg Q(x)) \end{aligned}$$

  Equality almost always holds in the last line. The only time it does not is when the leading terms of $P(x)$ and $Q(x)$ cancel each other out.

- If $P(r) = 0$, then we say that r is a *root* or a *zero* of $P(x)$.

- We say that $P(x)$ is a *divisor* or a *factor* of $Q(x)$ if there exists a polynomial $M(x)$ such that $P(x)M(x) = Q(x)$.

9.0 Problems

1. Let $P(x)$ be a real polynomial satisfying

$$P(1) = 3, \quad P(2) = 5 \quad \text{and} \quad P(3) = 2.$$

Determine the remainder when $P(x)$ is divided by $(x-1)(x-2)(x-3)$.

2. Let $P(x)$ be an integer polynomial such that $P(2)$ is divisible by 5 and $P(5)$ is divisible by 2.

Prove that $P(7)$ is divisible by 10.

3. If p, q, r are the roots of the polynomial $x^3 - x - 1$, compute

$$p^2 + q^2 + r^2 \qquad \text{and} \qquad \frac{1+p}{1-p} + \frac{1+q}{1-q} + \frac{1+r}{1-r}.$$

4. Find the cubic monic polynomial whose roots are the cubes of the roots of

$$x^3 - x^2 + x - 2 = 0.$$

5. For which positive integers n is the polynomial

$$(x+1)^n + x^n + 1$$

divisible by the polynomial $x^2 + x + 1$?

6. Let $P(x)$ be an integer polynomial whose leading coefficient is odd. Suppose that $P(0)$ and $P(1)$ are also odd.

Prove that $P(x)$ has no rational roots.

7. Let $P(x)$ be a monic polynomial of degree four with distinct integer roots a, b, c and d.

If $P(r) = 4$ for some integer r, prove that

$$r = \frac{1}{4}(a+b+c+d).$$

8. Does there exist an integer polynomial $P(x)$ such that

$$P(10) = 400, \quad P(14) = 440 \quad \text{and} \quad P(18) = 520?$$

9. Find a polynomial $P(x)$ such that $P(x)$ is divisible by $x^2 + 1$ and $P(x) + 1$ is divisible by $x^3 + x^2 + 1$.

10. Find all real polynomials $P(x)$ such that

$$xP(x-1) = (x-2)P(x).$$

11. Let $P(x)$ be an integer polynomial such that the equation $P(x) = 5$ has five distinct integer solutions.

Prove that the equation $P(x) = 8$ has no integer solutions.

12. Find all real polynomials $P(x)$ such that

$$P(x)P(x+1) = P(x^2).$$

13. (a) Does there exist a non-constant integer polynomial $P(x)$ such that the sequence $P(1), P(2), P(3), \ldots$ consists only of primes?

    (b) Does there exist a non-constant real polynomial $P(x)$ such that the sequence $P(1), P(2), P(3), \ldots$ consists only of primes?

14. Find all real polynomials $P(x)$ such that, if a is a real number and $P(a)$ is an integer, then a must also be an integer.

15. (a) Prove that the polynomial

$$x^5 - 5x^3 + 10x^2 - 15x + 20$$

    is irreducible over $\mathbb{Q}$.

    (b) Let $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ be an integer polynomial and suppose that there exists a prime p such that

$$p^2 \nmid a_0, \quad p \mid a_0, \quad p \mid a_1, \quad \ldots, \quad p \mid a_{n-1} \quad \text{and} \quad p \nmid a_n.$$

    Prove that $P(x)$ is irreducible over $\mathbb{Q}$.[1]

    (c) Prove that the polynomial

$$2x^n + 4x^{n-1} + 12x + 3$$

    is irreducible over $\mathbb{Q}$.

16. (a) Show that there is a unique polynomial $p(x)$ with integer coefficients in the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ such that $p(-2) = p(-5) = 10$.

    (b) What happens if $p(-2) = p(-5) = 2009$ instead?

17. Let $p(x)$ be a polynomial with integer coefficients. Suppose that there are integers $x_1, x_2, \ldots, x_n$ such that

$$p(x_1) = x_2, \quad p(x_2) = x_3, \quad \ldots, \quad p(x_{n-1}) = x_n \quad \text{and} \quad p(x_n) = x_1.$$

    Prove that $x_1 = x_3$.

18. (a) Which polynomials can be written as a finite sum of cubes of polynomials with real coefficients?

    (b) Which polynomials can be written as a finite sum of cubes of polynomials with integer coefficients?

19. Let $a_1, a_2, \ldots, a_{100}, b_1, b_2, \ldots, b_{100}$ be distinct real numbers. For each i and j the number $a_i - b_j$ is written in the ith row and jth column of a 100×100 table. Suppose that the product of the numbers in each column is equal to 1.

    Prove that the product of the numbers in each row is equal to -1.

20. Show that the set of real numbers x which satisfy the inequality

$$\sum_{k=1}^{70} \frac{k}{x-k} \geq \frac{5}{4}$$

    is a union of disjoint intervals, the sum of whose lengths is 1988.

[1] This result is commonly known as *Eisenstein's criterion*.

21. Let $p(x)$ be a polynomial of degree n with the property that

$$p(k) = \frac{k}{k+1}, \quad \text{for } k = 0, 1, 2, \ldots, n.$$

Determine $p(n+1)$.

22. Let $p(x)$ be a polynomial of degree n, with the property that

$$p(k) = 2^k, \quad \text{for } k = 0, 1, 2, \ldots, n.$$

Determine $p(n+1)$.

23. Let $p(x)$ be a polynomial of degree 10. Suppose that

$$p(0) = p(2) = p(4) = p(6) = p(8) = p(10) = 0$$

and

$$p(1) = p(3) = p(5) = p(7) = p(9) = 1.$$

What is the value of $p(11)$?

24. Let $p(x)$ be a polynomial of degree $3n$ such that

$$\begin{aligned} p(0) = p(3) = \cdots = \quad p(3n) \quad &= 2 \\ p(1) = p(4) = \cdots = p(3n-2) &= 1 \\ p(2) = p(5) = \cdots = p(3n-1) &= 0, \end{aligned}$$

and $p(3n+1) = 730$.
Find n.

25. For any polynomial $p(x)$ of degree at most n, prove that

$$p(n+1) = \sum_{k=0}^{n} (-1)^{n-k} \binom{n+1}{k} p(k).$$

26. Let k be a given positive integer and let $F(x)$ be an integer polynomial satisfying

$$0 \le F(c) \le k, \quad \text{for } c = 0, 1, \ldots, k+1.$$

(a) If $k \ge 4$, prove that $F(0) = F(1) = \cdots = F(k+1)$.
(b) Find counterexamples for when $k < 4$.

27. Let k be an integer and let

$$p(x) = x^6 + x^5 + x^4 + x^3 + x^2 + k.$$

(a) Prove that $p(x)$ does not contain a cubic factor.
(b) If $k = 9$ or 43, prove that $p(x)$ is irreducible.

28. Let $P(x)$ and $Q(x)$ be polynomials whose coefficients are all equal to 1 or 7.
If $P(x)$ divides $Q(x)$, prove that $1 + \deg P(x)$ divides $1 + \deg Q(x)$.

29. Find all polynomials $p(x, y)$ with real coefficients such that

$$p(u, p(v, w)) = p(u + v, w)$$

for all real u, v, w.

30. The graph of a monic cubic polynomial contains the vertices of exactly one square.

What is the area of the square?

31. Prove that every real polynomial can be multiplied by a non-zero real polynomial to obtain a polynomial whose exponents are all divisible by 1000.

32. Find all integer polynomials $f(x)$ such that $f(a) = f(b)$ for infinitely many pairs of integers a, b with $a \neq b$.

33. Let $F(x)$ be a real polynomial of degree n.

Show that $F(x)$ has n real roots if and only if it is not possible to write

$$F(x)^2 = G(x)^2 + H(x)^2$$

for non-zero real polynomials $G(x)$ and $H(x)$ of different degrees.

34. Let $P(x)$ be a polynomial of degree $n > 1$ with integer coefficients and let k be a positive integer. Consider the polynomial

$$Q(x) = P(P(\cdots P(P(x))\cdots)),$$

where P occurs k times.

Prove that there are at most n integers t such that $Q(t) = t$.

9.1 Identity theorem

We start with a very basic proposition: a polynomial which is zero infinitely often must actually be the zero polynomial. This is stated as the identity theorem below, along with a couple of simple variants.

Identity theorem

- If a polynomial has infinitely many roots, then it is the zero polynomial.
- If two polynomials satisfy $P(x) = Q(x)$ for infinitely many values of x, then the polynomials $P(x)$ and $Q(x)$ are equal.
- If two polynomials of degree at most n satisfy $P(x) = Q(x)$ for $n+1$ values of x, then the polynomials $P(x)$ and $Q(x)$ are equal.

You should definitely try your hand at proving the identity theorem. It should be well within your grasp, particularly if you are acquainted with the following important results.

Factor theorem and remainder theorem The factor theorem states that the number r is a root of $P(x)$ if and only if $P(x)$ is divisible by $x - r$. More generally, the remainder theorem states that $P(r) = c$ if and only if the remainder after dividing $P(x)$ by $x - r$ is c.

Problem Find all real polynomials $P(x)$ which satisfy

$$P(0) = 0 \quad \text{and} \quad P(x^2 + 1) = P(x)^2 + 1.$$

Solution If we substitute $x = 0$ into the given equation, we obtain $P(1) = 1$. If we substitute $x = 1$, we obtain $P(2) = 2$. If we substitute $x = 2$, we obtain $P(5) = 5$.

Continuing in this fashion, we will find infinitely many values of x for which $P(x) = x$. Thus, $P(x) - x$ is a polynomial with infinitely many roots so, by the identity theorem, it's the zero polynomial. Therefore, $P(x) = x$. □

This proof is a little sloppy. The following is a brief sketch of how you should write it down, given that you want to produce a completely rigorous proof.

You might start by explicitly defining the sequence $0, 1, 2, 5, \ldots$ that we are interested in by defining a sequence $a_0, a_1, \ldots$ which satisfies

$$a_0 = 0 \quad \text{and} \quad a_{n+1} = a_n^2 + 1 \quad \text{for } n = 0, 1, 2 \ldots.$$

Next, you would show that this is an increasing sequence. Then you would prove that $P(a_n) = a_n$ for each n, by induction.

If you have done all of this correctly, then you now have infinitely many values of x for which $P(x) = x$ and can invoke the identity theorem just as we did above.

9.2 Division algorithm

If you are given two polynomials, then you can add them, subtract them or multiply them to obtain another polynomial. However, you cannot in general divide them to obtain another polynomial. This situation should remind you of the one we have when we deal with integers. In that case, we know that for any two integers a and $b \neq 0$, there is a unique way to write

$a = qb + r$, where $0 \leq r < |b|$. One of the lessons you will hopefully learn from this chapter is the fact that many concepts from number theory have natural analogues in the world of polynomials. The *division algorithm* provides just one example.

Division algorithm for polynomials For any two polynomials $A(x)$ and $B(x) \neq 0$, there is a unique way to write

$$A(x) = Q(x)B(x) + R(x),$$

where $\deg R(x) < \deg B(x)$.

You can convince yourself that this is true by performing polynomial long division on a few well chosen pairs of polynomials.

Problem Let $P(x)$ be a real polynomial satisfying $P(a) = A$ and $P(b) = B$, where $a \neq b$. Determine the remainder when $P(x)$ is divided by $(x - a)(x - b)$.

Solution The division algorithm asserts that there is a unique way to write

$$P(x) = Q(x)(x - a)(x - b) + R(x),$$

where $R(x)$ is linear or a constant. Our goal, of course, is to determine $R(x)$. Plugging the values a and b into this equation, we find that

$$R(a) = P(a) = A \quad \text{and} \quad R(b) = P(b) = B.$$

Therefore, the two points (a, A) and (b, B) lie on the graph of $y = R(x)$, which we know to be a line. So we can easily deduce that the desired remainder is

$$R(x) = \frac{A - B}{a - b}x + \frac{aB - Ab}{a - b}.$$

Of course, the answer is nicely symmetric, as we should have expected. □

9.3 Fundamental theorem of algebra

One of the most important results concerning polynomials, as evidenced by its grand name, is the *fundamental theorem of algebra*. This hugely powerful theorem provides us with a first route of attack against many polynomial problems.

Fundamental theorem of algebra A complex polynomial $P(x)$ of degree n can be completely factorised as

$$P(x) = A(x - r_1)(x - r_2) \cdots (x - r_n),$$

where $r_1, r_2, \ldots, r_n$ are complex numbers. The roots $r_1, r_2, \ldots, r_n$ may include repetition, and the number of times a root occurs is called its *multiplicity*.

For the remainder of this chapter, we will never again mention polynomials with complex coefficients. So, you might be wondering, why is the fundamental theorem of algebra so useful to us? It's simply because we'll be dealing with polynomials whose coefficients are integers, rationals and real numbers, and these are all just particular examples of complex numbers.

Problem Find all real polynomials $P(x)$ which satisfy the equation

$$(x - 16)P(2x) = 16(x - 1)P(x).$$

Solution By the fundamental theorem of algebra, we may factorise the polynomial as

$$P(x) = A(x - r_1)(x - r_2) \cdots (x - r_n).$$

Now substitute this into both sides of the given equation—a good way to start many a polynomial problem. The left-hand side is

$$A(x - 16)(2x - r_1) \cdots (2x - r_n) = 2^n A(x - 16) \left(x - \frac{r_1}{2}\right) \cdots \left(x - \frac{r_n}{2}\right),$$

while the right-hand side is

$$16A(x - 1)(x - r_1) \cdots (x - r_n).$$

Since these should be equal, we can compare leading coefficients to obtain

$$2^n A = 16A,$$

from which it follows that $n = 4$. We can now compare roots to obtain the fact that the following two sets contain exactly the same elements.

$$\left\{16, \frac{r_1}{2}, \frac{r_2}{2}, \frac{r_3}{2}, \frac{r_4}{2}\right\} = \{1, r_1, r_2, r_3, r_4\}$$

Without loss of generality, let $r_1 = 16$ which implies that $\frac{r_1}{2} = 8$. Then, without loss of generality, let $r_2 = 8$ which implies that $\frac{r_2}{2} = 4$.

Proceeding in this fashion yields

$$P(x) = A(x - 2)(x - 4)(x - 8)(x - 16),$$

where A can be any real number. As usual, you should substitute this back into the original equation to check that it actually works, and you will find that it certainly does. □

9.4 Vieta's formulas

If you want to describe a polynomial to somebody, then you can either give them all of its coefficients or give them just the leading coefficient and all of its complex roots.

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 \qquad \text{or} \qquad a_n (x - r_1)(x - r_2) \cdots (x - r_n)$$

There are nice relations between these two descriptions, which you obtain by expanding out the second expression and comparing coefficients with the first.

Vieta's formulas If the polynomial $a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$ has roots $r_1, r_2, \ldots, r_n$, then the following formulas hold.

$$\begin{aligned} r_1 + r_2 + \cdots + r_n &= -\frac{a_{n-1}}{a_n} \\ r_1 r_2 + r_1 r_3 + \cdots + r_{n-1} r_n &= \frac{a_{n-2}}{a_n} \\ &\vdots \\ r_1 r_2 \cdots r_n &= (-1)^n \frac{a_0}{a_n} \end{aligned}$$

Observe that the signs on the right-hand sides of these equations alternate between negative and positive. The first equation involves the sum of the roots, the second equation involves

the sum of the products of the roots taken two at a time, the third equation involves the sum of the products of the roots taken three at a time, and so on. Such expressions, known as *elementary symmetric functions*, appear in the following important result.

Fundamental theorem of symmetric polynomials A polynomial in the variables $r_1, r_2, \ldots, r_n$ is called *symmetric* if it remains the same when you swap r_i with r_j for any i and j.
The theorem tells us that every symmetric polynomial in the variables $r_1, r_2, \ldots, r_n$ can be expressed in terms of the elementary symmetric functions.

Problem Suppose that $P(x)$ is a degree three polynomial with roots p, q, r such that

$$P\left(\frac{1}{5}\right) + P\left(-\frac{1}{5}\right) = 82P(0).$$

Determine the value of

$$\frac{1}{pq} + \frac{1}{qr} + \frac{1}{rp}.$$

Solution If we let $P(x) = ax^3 + bx^2 + cx + d$, then the given equation becomes

$$\left(\frac{a}{125} + \frac{b}{25} + \frac{c}{5} + d\right) + \left(-\frac{a}{125} + \frac{b}{25} - \frac{c}{5} + d\right) = 82d.$$

This simplifies to

$$\frac{2}{25}b + 2d = 82d,$$

from which we quickly obtain

$$\frac{b}{d} = 1000.$$

But from Vieta's formulas, we know that

$$\frac{1}{pq} + \frac{1}{qr} + \frac{1}{rp} = \frac{-a(p+q+r)}{-apqr} = \frac{b}{d} = 1000. \qquad \square$$

Now let's turn our attention to something just a little more difficult.

Problem Given a positive integer n, let $1, r_1, r_2, \ldots, r_n$ be the roots of the polynomial $P(x) = x^{n+1} - 1$.

Show that

$$\frac{1}{1-r_1} + \frac{1}{1-r_2} + \cdots + \frac{1}{1-r_n} = \frac{n}{2}.$$

Although the left-hand side of the equation appears to involve a symmetric expression of the roots of $P(x)$, the root 1 is missing. So our approach might be to find a polynomial whose roots are precisely $r_1, r_2, \ldots, r_n$. That polynomial would be

$$\frac{P(x)}{x-1} = x^n + x^{n-1} + x^{n-2} + \cdots + 1.$$

Subsequently, your strategy might be to expand out the given expression, rely on Vieta's formulas, and hope for the best. But such an approach would create a huge mess, requiring you to put the given expression over a common denominator, followed by excessive amounts of algebra. However, a much slicker approach is available.

Solution Rather than considering the polynomial with $r_1, r_2, \ldots, r_n$ as roots, we consider the polynomial with $1 - r_1, 1 - r_2, \ldots, 1 - r_n$ as roots.

For ease of notation, let's use the substitution $s_k = 1 - r_k$ for $k = 1, 2, \ldots, n$. Since each r_k satisfies the equation $r_k^{n+1} = 1$, each s_k satisfies

$$(1 - s_k)^{n+1} = 1.$$

Therefore, $s_1, s_2, \ldots, s_n$ are roots of the polynomial

$$(1 - x)^{n+1} - 1.$$

Since this polynomial has degree $n+1$, it must have $n+1$ complex roots counting multiplicity. So just as we previously excluded the root 1, we now must exclude the root 0. In summary, $s_1, s_2, \ldots, s_n$ are precisely the roots of the polynomial

$$Q(x) = \frac{(1 - x)^{n+1} - 1}{x}.$$

Now if we write $Q(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$, then Vieta's formulas tell us that

$$\frac{1}{s_1} + \frac{1}{s_2} + \cdots + \frac{1}{s_n} = \frac{s_2 s_3 \cdots s_n + \cdots + s_1 s_2 \cdots s_{n-1}}{s_1 s_2 \cdots s_n} = -\frac{a_1}{a_0}.$$

In order to obtain the coefficients of $Q(x)$, we simply need to expand $\frac{(1-x)^{n+1}-1}{x}$ using the good old binomial formula.

$$Q(x) = -\binom{n+1}{1} + x\binom{n+1}{2} - x^2\binom{n+1}{3} + \cdots + (-1)^{n+1} x^n$$

Therefore, we conclude that

$$\frac{1}{s_1} + \frac{1}{s_2} + \cdots + \frac{1}{s_n} = \frac{\binom{n+1}{2}}{\binom{n+1}{1}} = \frac{n}{2}.$$

□

9.5 Integer polynomials

Within the class of real polynomials lies the class of integer polynomials, which is a fascinating world of its own. There are many new techniques and results which only apply to integer polynomials. One example is the following, which is known as the *rational root theorem*.

Rational root theorem Let $P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$ be an integer polynomial. If $p(x)$ has a rational root which can be expressed as $\frac{r}{s}$ in lowest terms, then $r \mid a_0$ and $s \mid a_n$.

Problem Prove the rational root theorem.

Solution If $\frac{r}{s}$ is a root of the polynomial, then we have the equation

$$a_n \left(\frac{r}{s}\right)^n + a_{n-1} \left(\frac{r}{s}\right)^{n-1} + \cdots + a_1 \left(\frac{r}{s}\right) + a_0 = 0.$$

It's only natural now to clear the equation of fractions. Multiplying both sides by s^n gives

$$a_n r^n + a_{n-1} r^{n-1} s + \cdots + a_1 r s^{n-1} + a_0 s^n = 0.$$

Since r divides every term of the equation except $a_0 s^n$, it follows that r also divides $a_0 s^n$ itself. However, we've assumed that r and s are relatively prime, so it must be the case that $r \mid a_0$. An analogous argument shows that $s \mid a_n$. □

The following is another useful little result, very easy to prove, which applies only to integer polynomials.

Lemma If $P(x)$ is a polynomial with integer coefficients, then

$$a - b \mid P(a) - P(b)$$

for all integers a and b.

Problem Suppose that $P(x)$ is an integer polynomial and a is an integer satisfying

$$P(P(P(P(a)))) = a.$$

Prove that $P(P(a)) = a$.

Solution Consider the sequence defined by $a_0 = a$ and $a_{n+1} = P(a_n)$ for all non-negative integers n. The sequence consists entirely of integers and we have $a_4 = P(P(P(P(a)))) = a_0$. It follows that $a_5 = a_1$, $a_6 = a_2$, $a_7 = a_3$, and so on. In other words, the sequence is periodic with period four.

By the lemma above, we have

$$a_k - a_{k-1} \mid P(a_k) - P(a_{k-1}) = a_{k+1} - a_k$$

for every positive integer k. In particular, we have the following chain of divisibilities.

$$a_1 - a_0 \mid a_2 - a_1 \mid a_3 - a_2 \mid a_0 - a_3 \mid a_1 - a_0$$

If a term appearing in this chain is equal to 0, then they must all be, since 0 doesn't divide any non-zero integer. In this case, we would have $a_1 = a_0$ or equivalently, $P(a) = a$, which immediately implies that $P(P(a)) = a$.

Otherwise, all terms appearing in the chain are non-zero. Since $a \mid b$ implies that $b = 0$ or $|a| \leq |b|$, we have the following chain of inequalities.

$$|a_1 - a_0| \leq |a_2 - a_1| \leq |a_3 - a_2| \leq |a_0 - a_3| \leq |a_1 - a_0|$$

Naturally, this implies the following chain of equalities.

$$|a_1 - a_0| = |a_2 - a_1| = |a_3 - a_2| = |a_0 - a_3| = |a_1 - a_0|$$

If $a_1 - a_0 = a_2 - a_1 = a_3 - a_2 = a_0 - a_3 = m$, then it's simple to deduce that $m = 0$, contradicting the fact that each term in the chain is non-zero.

So there must be a value of k for which $a_k - a_{k-1} = -(a_{k+1} - a_k)$, which implies that $a_{k+1} = a_{k-1}$. However, this implies that $a_{k+2} = a_k$, $a_{k+3} = a_{k+1}$, $a_{k+4} = a_{k+2}$, and so on. Sooner or later, since the sequence is periodic, we obtain an equation which is equivalent to $a_2 = a_0$, which implies that $P(P(a)) = a$. □

If you fully understand the previous solution, then you should have no problem in proving the following, more general, statement.

> If $P(x)$ is an integer polynomial and a is an integer satisfying
>
> $$P(P(\cdots P(P(a))\cdots)) = a,$$
>
> then $P(P(a)) = a$.

9.6 Complex numbers

It's an amazing fact that, even if you only care about real polynomials, it pays to know something about complex numbers. For example, the fundamental theorem of algebra asserts that every complex polynomial of degree n has precisely n complex roots. In the case when the polynomial has real coefficients, these roots must obey the following theorem.

Conjugate root theorem If $P(x)$ is a real polynomial, then z is a root of multiplicity m if and only if its conjugate $\overline{z}$ is a root of multiplicity m. Therefore, the roots of a real polynomial are real or occur in complex conjugate pairs.

If you are familiar with complex conjugation, you should be able to prove the conjugate root theorem without too much trouble.

Problem Prove that any real polynomial can be written as a product of real linear and real quadratic polynomials.

Solution The conjugate root theorem allows us to list the real roots of $P(x)$ as $r_1, r_2, \ldots, r_k$ and the non-real roots as $z_1, \overline{z_1}, z_2, \overline{z_2}, \ldots, z_m, \overline{z_m}$. In other words, for some real number A, we can write

$$P(x) = A(x - r_1) \cdots (x - r_k)(x - z_1)(x - \overline{z_1}) \cdots (x - z_m)(x - \overline{z_m}).$$

However, if $z = a + bi$ for real numbers a and b, then

$$(x - z)(x - \overline{z}) = (x - a - bi)(x - a + bi) = (x - a)^2 + b^2.$$

This means that each complex conjugate pair of roots contributes a real quadratic factor to $P(x)$. Clearly, each real root contributes a real linear factor to $P(x)$, so we're done. □

9.7 Algebraic trickery

Sometimes, solving a polynomial problem just comes down to thinking of the right algebraic identity or manipulation. Therefore, you should try to become an algebraic mathemagician and know all the tricks! The first one we'll look at involves a simple substitution which you can think of as shifting the graph of the polynomial to the left or right.

Problem Show that the polynomial

$$P(x) = x^7 - 2x^5 + 10x^2 - 1$$

has no real root greater than 1.

Solution Normally it's quite difficult to tell, simply by looking at a polynomial, whether or not it has a real root greater than 1. However, it's often easier to tell whether or not a polynomial has a real root greater than 0, as we'll soon see.

Motivated by these reasons, let's set $x = y + 1$ and substitute into the expression for $P(x)$ to obtain

$$\begin{aligned} P(y+1) &= (y+1)^7 - 2(y+1)^5 + 10(y+1)^2 - 1 \\ &= y^7 + 7y^6 + 19y^5 + 25y^4 + 15y^3 + 11y^2 + 17y + 8. \end{aligned}$$

Note that $P(x)$ has a real root greater than 1 if and only if $P(y+1)$ has a real root greater than 0. However, it's clear by inspection that $P(y+1)$ is positive whenever y is positive. So we can conclude that $P(y+1)$ has no real root greater than 0. Therefore, $P(x)$ has no real root greater than 1. □

For the next problem, we use the fact that the product of two sums of two perfect squares is another sum of two perfect squares. This follows from the algebraic identity

$$(A^2+B^2)(C^2+D^2) = (AC-BD)^2 + (AD+BC)^2.$$

You could prove this simply by expanding both sides and verifying that they are equal. For the complex numbers expert, a more insightful approach is to recognise that the result is equivalent to

$$|A+Bi| \times |C+Di| = |(A+Bi)\times(C+Di)|.$$

Either way, this is one of those remarkable formulas that every good mathemagician should know!

Furthermore, this identity is not just true when A, B, C and D are numbers, but any quantities that you can add, subtract and multiply. In particular it is true if A, B, C and D are functions, and of course these include polynomials.

Problem Let $P(x)$ be a real polynomial such that $P(x) \geq 0$ for all real x. Prove that it's possible to write

$$P(x) = F(x)^2 + G(x)^2$$

for real polynomials $F(x)$ and $G(x)$.

Solution If $P(x)$ has a real root r with odd multiplicity, then the sign of the graph $y = P(x)$ will change from positive to negative or vice versa at $x = r$. Therefore, every real root of $P(x)$ must have even multiplicity.

Now we use the result from section 9.6 to write

$$P(x) = A(x-r_1)^2 \cdots (x-r_k)^2 (x-z_1)(x-\overline{z_1}) \cdots (x-z_m)(x-\overline{z_m}).$$

Here, the real number A must be non-negative to ensure that $P(x) \geq 0$ for all real x. Therefore, we can write

$$P(x) = R(x)^2 Z_1(x) Z_2(x) \cdots Z_m(x),$$

where

$$R(x) = \sqrt{A}(x-r_1)(x-r_2)\cdots(x-r_k)$$

and

$$Z_i(x) = (x-z_i)(x-\overline{z_i}).$$

However, if $z = a+bi$ for real numbers a and b, then

$$(x-z)(x-\overline{z}) = (x-a-bi)(x-a+bi) = (x-a)^2 + b^2.$$

Therefore, we can write each of $Z_1(x), Z_2(x), \ldots, Z_m(x)$ as a sum of two perfect squares. Since $R(x)^2 = R(x)^2 + 0^2$ is also a sum of two perfect squares, we can repeatedly use our mathemagical identity

$$(A^2+B^2)(C^2+D^2) = (AC-BD)^2 + (AD+BC)^2,$$

to write the product $P(x) = R(x)^2 Z_1(x) Z_2(x) \cdots Z_m(x)$ as a sum of two perfect squares. □

9.8 Irreducibility

By the fundamental theorem of algebra, every polynomial can be factorised into linear factors, if we are allowed to use complex numbers. But given a polynomial with integer coefficients, one might ask whether or not it can be factorised into two polynomials, each with integer coefficients themselves. Of course, you can take out a factor of 1 or perhaps some larger integer, but this is totally boring. So we say that a polynomial $F(x)$ is *reducible* over $\mathbb{Z}$ if it can be written as a product $F(x) = G(x)H(x)$, where $G(x)$ and $H(x)$ are integer polynomials of positive degree. If this is not possible, then we say that $F(x)$ is *irreducible* over $\mathbb{Z}$. Similarly if $F(x) = G(x)H(x)$, where $G(x)$ and $H(x)$ are rational polynomials of positive degree, we say that $F(x)$ is reducible over $\mathbb{Q}$ and otherwise, that it is irreducible over $\mathbb{Q}$.

Clearly, if an integer polynomial is reducible over $\mathbb{Z}$, then it's reducible over $\mathbb{Q}$, because every integer is certainly rational. More surprisingly, the converse is true as well.

Gauss' lemma If an integer polynomial is reducible over $\mathbb{Q}$, then it's reducible over $\mathbb{Z}$. So, for an integer polynomial, reducibility over $\mathbb{Z}$ and reducibility over $\mathbb{Q}$ are the same thing.

Showing that a given polynomial is irreducible can be very difficult. We'll see a few different techniques as we go on, but for the following problem, we'll use a rather direct approach.

Problem If $a_1, a_2, \ldots, a_n$ are distinct integers, prove that the polynomial

$$F(x) = (x - a_1)(x - a_2) \cdots (x - a_n) - 1$$

is irreducible over $\mathbb{Z}$.

Solution Suppose that $F(x) = G(x)H(x)$, where $G(x)$ and $H(x)$ are integer polynomials of positive degree. Clearly, $F(x)$ has degree n, so we know that

$$\deg G + \deg H = n.$$

Now substituting $x = a_k$ for some value of k yields

$$G(a_k)H(a_k) = -1.$$

Therefore, $G(a_k)$ and $H(a_k)$ must be integers which multiply to give -1, which leads to $G(a_k) = 1$ and $H(a_k) = -1$ or vice versa. In either case, we have

$$G(a_k) + H(a_k) = 0.$$

In summary, we have deduced that the polynomial

$$P(x) = G(x) + H(x)$$

has the distinct roots $a_1, a_2, \ldots, a_n$, and possibly more. So, if $P(x)$ is a non-zero polynomial, then it would have to have degree at least n. But the degree of $P(x)$ is at most the maximum of $\deg G$ and $\deg H$, which is strictly less than n by assumption. So this case cannot occur.

The other possibility is that $P(x)$ is the zero polynomial. But then $G(x) = -H(x)$ and so,

$$F(x) = -G(x)^2.$$

There is a problem here: the leading coefficient of $-G(x)^2$ is clearly negative, while the leading coefficient of $F(x)$ is clearly positive.

From these contradictions, we can conclude that $F(x)$ is irreducible over $\mathbb{Z}$. □

We finish this section with a useful fact concerning irreducible integer polynomials which you might like to prove.

Proposition Suppose that r is a root of an irreducible rational polynomial $P(x)$. If r is also a root of a rational polynomial $Q(x)$, then it must be the case that $P(x) \mid Q(x)$.

9.9 Factorisation

Yet another piece of algebraic trickery that arises in polynomial problems is the *difference of perfect powers* factorisation.

$$x^n - y^n = (x-y)(x^{n-1} + x^{n-2}y + \cdots + y^{n-1})$$

This often appears when one is dealing with geometric series.

Problem Prove that there are no primes in the integer sequence

$$10001, 100010001, 1000100010001, \ldots.$$

Solution The trick is to write $x = 10$, so that we can express the terms of the sequence as

$$1 + x^4 + x^8 + \cdots + x^{4k}.$$

Now we apply two tactics. The first is to use the factorisation above to write this geometric series in closed form, and the second is to further factorise the result, which features differences of perfect squares.

$$\frac{x^{4k+4}-1}{x^4-1} = \frac{(x^{2k+2}+1)}{(x^2+1)} \times \frac{(x^{2k+2}-1)}{(x^2-1)}$$

Now the roots of $x^2 - 1$ are ± 1 and it's easy to check that these are both roots of $x^{2k+2} - 1$. Similarly, the roots of $x^2 + 1$ are $\pm i$, and it's easy to check that these are both roots of $x^{2k+2} + 1$ if k is even, or $x^{2k+2} - 1$ if k is odd.

In either case, after dividing the numerator by the denominator, we are left with the product of two integer polynomials. For $k \geq 2$, both of these polynomials are guaranteed to have positive degree. Furthermore, it's easy to check (and you should do so right now) that they cannot be equal to 1 when $x = 10$. Thus, every term of the sequence after the first is a product of two numbers greater than 1 and hence, not prime.

For the first term of the sequence, our clever polynomial argument doesn't work. However, we can compute directly that $10001 = 73 \times 137$. □

9.10 Polynomials modulo p (upstairs–downstairs)

Many of the tricks and techniques from number theory carry over to the study of polynomials. We've already witnessed the division algorithm for polynomials in action and now it's time to consider modular arithmetic for polynomials. We will be reducing the coefficients—but never the exponents—of an integer polynomial modulo p. For example, it makes sense to write a statement like

$$5x^6 - x^4 + 2x^3 - 10x^2 + x + 3 \equiv 4x^4 + 2x^3 + x + 3 \pmod 5.$$

In this way, polynomials can be added, subtracted, and multiplied modulo p.

It's almost always useful to take p to be prime, in which case you can rely on the division algorithm along with many of the results you already know and love which hold for polynomials in general. For example, we require p to be a prime for the following simple, though useful, statement to be true.

Theorem Let p be a prime. Suppose that

$$F(x)G(x) \equiv 0 \pmod{p},$$

then

$$F(x) \equiv 0 \pmod{p} \quad \text{or} \quad G(x) \equiv 0 \pmod{p}.$$

Another particularly useful true statement that requires p to be prime is the following analogue of the fundamental theorem of arithmetic.

Unique factorisation for polynomials modulo p If p is a prime, then any factorisation of a polynomial into irreducible factors modulo p is unique up to the order of its factors.

(We did not mention it earlier, but unique factorisation also holds for ordinary integer polynomials.)

One of the most useful reasons to consider polynomials modulo p is to prove irreducibility. If $F(x)$ can be factorised as $F(x) = G(x)H(x)$, then we know that $F(x) \equiv G(x)H(x) \pmod{p}$. In other words, if a polynomial is reducible over the integers, then it's reducible over the integers modulo p. The contrapositive of this statement is particularly useful: if a polynomial is irreducible over the integers modulo p, then it's irreducible over the integers.

The beauty of this technique is that modulo p there are only finitely many polynomials of degree lower than a given polynomial. This means that you can check them all! Furthermore, you are free to choose whichever value of p happens to work for you. However, one must beware! If a polynomial is reducible over the integers modulo p, then it definitely does not follow that it's reducible over the integers.

Problem Prove that the polynomial

$$F(x) = x^5 - x^2 + 1$$

is irreducible over $\mathbb{Z}$.

Solution We simply consider $F(x)$ modulo 2. For the remainder of this proof, all congruences are assumed to be considered modulo 2.

Suppose that $F(x)$ is reducible over the integers modulo 2. Then we can write

$$F(x) \equiv G(x)H(x),$$

where $G(x)$ and $H(x)$ have positive degree.

Since $\deg G + \deg H = 5$, one of $G(x)$ and $H(x)$ must have degree at most 2. Without loss of generality, assume that it's $G(x)$. So $G(x)$ has degree 1 or 2 and, furthermore, we may assume that $G(x)$ is irreducible.

Working modulo 2, there are only two polynomials of degree 1, namely, x and $x+1$. It's easy to check that neither is a factor of $F(x)$ modulo 2. For example, you can use polynomial division to obtain the following.

$$x^5 - x^2 + 1 \equiv (x^4 + x) \cdot x + 1$$
$$x^5 - x^2 + 1 \equiv (x^4 + x^3 + x^2) \cdot (x + 1) + 1$$

Similarly, there are only four polynomials modulo 2 and of degree 2, namely, x^2, x^2+1, x^2+x and x^2+x+1. To make things even simpler, we have

$$x^2 \equiv x \cdot x, \quad x^2+1 \equiv (x+1)\cdot(x+1) \quad \text{and} \quad x^2+x \equiv x\cdot(x+1),$$

so that x^2+x+1 is the only irreducible quadratic modulo 2. Since

$$x^5 - x^2 + 1 \equiv (x^3+x^2)\cdot(x^2+x+1)+1,$$

this is also not a factor of $F(x)$. So $F(x)$ is irreducible over the integers modulo 2 and hence, irreducible over the integers. □

Often, we think of working with integers as *upstairs* and working with integers modulo p as *downstairs*. For more difficult problems, we may need to move between upstairs and downstairs quite often, transferring information between the two levels.

Problem Prove that the polynomial

$$F(x) = x^n + 5x^{n-1} + 3$$

is irreducible over $\mathbb{Z}$ for every integer $n > 1$.

Solution In this solution, downstairs will refer to working modulo 3. Suppose that $F(x)$ is reducible, so that we can write $F(x) = G(x)H(x)$ for polynomials $G(x)$ and $H(x)$ of positive degree.

Downstairs, we have

$$F(x) \equiv x^{n-1}(x+5),$$

which, by unique factorisation modulo p, can only be non-trivially factorised into two polynomials as

$$x^k(x+5) \times x^{n-1-k},$$

for some $k = 0, 1, 2, \ldots, n-2$. So, without loss of generality, we have

$$G(x) \equiv x^k(x+5) \quad \text{and} \quad H(x) \equiv x^{n-1-k}.$$

If k is positive, then moving upstairs tells us that the constant terms of both $G(x)$ and $H(x)$ are divisible by 3. This implies that the constant term of $F(x)$ is divisible by 9, a rather blatant contradiction.

So, $k = 0$ and our equations downstairs become

$$G(x) \equiv x+5 \quad \text{and} \quad H(x) \equiv x^{n-1}.$$

Going back upstairs we must have $\deg G(x) \geq 1$ and $\deg H(x) \geq n-1$. But $G(x)H(x) = F(x)$, which has degree n. Thus, $\deg G(x) = 1$ and $\deg H(x) = n-1$.

Since $G(x)$ is linear it has a rational root, which we can express as $\frac{r}{s}$ in lowest terms. But then $\frac{r}{s}$ is a root of $F(x)$ and the rational root theorem implies that $r \mid 3$ and $s \mid 1$. In short, $F(x)$ must have ± 1 or ± 3 as a root.

We can substitute these four values into the expression $x^n + 5x^{n-1} + 3$ to verify that they are certainly not roots, which grants us our desired contradiction. □

9.11 Polynomials modulo $P(x)$

Perhaps surprisingly, we can also reduce a polynomial modulo another polynomial. All the standard rules of modular arithmetic still apply; but, as usual, it's best to see an example.

Problem Let a, b, c and d be positive integers.

Show that the polynomial

$$x^{4a+3} + x^{4b+2} + x^{4c+1} + x^{4d}$$

is divisible by $x^3 + x^2 + x + 1$.

Solution There are a number of ways to solve this problem. One is to find the roots of $x^3 + x^2 + x + 1$ and show that they're also roots of the given polynomial. But we're here to learn about polynomial modular arithmetic, so our approach will be to show that the given polynomial is congruent to zero modulo $x^3 + x^2 + x + 1$.

Since $x^4 - 1 = (x - 1)(x^3 + x^2 + x + 1)$, we have

$$x^4 \equiv 1 \pmod{x^3 + x^2 + x + 1}.$$

So by working modulo $x^3 + x^2 + x + 1$, we obtain

$$\begin{aligned} x^{4a+3} + x^{4b+2} + x^{4c+1} + x^{4d} &= x^3(x^4)^a + x^2(x^4)^b + x(x^4)^c + (x^4)^d \\ &\equiv x^3 + x^2 + x + 1 \\ &\equiv 0. \end{aligned}$$

□

9.12 Lagrange interpolation

The identity theorem tells us that, in general, two points determine a line, three points determine a quadratic, four points determine a cubic, and so on. So somebody could say to you that $P(x_1) = y_1, P(x_2) = y_2, \ldots, P(x_{n+1}) = y_{n+1}$ and challenge you to work out the degree n of the polynomial $P(x)$. Of course, you could just get scared and run away, but it would be much better if you knew that there is a big fat formula called the *Lagrange interpolation formula* to answer this question precisely. Even better than knowing the formula itself, which is quite unwieldy, is to understand where it comes from.

Problem If $x_1, x_2, \ldots, x_{n+1}$ are distinct numbers, find a polynomial of degree at most n which satisfies

$$P(x_1) = P(x_2) = \cdots = P(x_{n+1}) = 0,$$

except that $P(x_k) = y_k$, for one particular value of k.

Solution If $y_k = 0$, then the problem is easy, since the identity theorem forces us to take $P(x) = 0$. Otherwise, the problem is almost as easy, since we know all of the roots of the polynomial.

$$P(x) = A(x - x_1) \cdots \widehat{(x - x_k)} \cdots (x - x_{n+1}).$$

That strange looking hat notation is telling the term $(x - x_k)$ to politely leave the building—it is excluded from the product. We can then substitute $x = x_k$ to derive an expression for A. The final answer is

$$P(x) = \frac{(x - x_1) \cdots \widehat{(x - x_k)} \cdots (x - x_{n+1})}{(x_k - x_1) \cdots \widehat{(x_k - x_k)} \cdots (x_k - x_{n+1})}\, y_k.$$

□

Problem If $x_1, x_2, \ldots, x_{n+1}$ are distinct numbers, find a polynomial of degree at most n which satisfies

$$P(x_1) = y_1, \quad P(x_2) = y_2, \quad \ldots, \quad P(x_{n+1}) = y_{n+1}.$$

Solution Simple! We just add up the polynomials that we obtained in the previous problem. So the final answer is

$$P(x) = \sum_{k=1}^{n+1} \frac{(x-x_1)\cdots\widehat{(x-x_k)}\cdots(x-x_{n+1})}{(x_k-x_1)\cdots\widehat{(x_k-x_k)}\cdots(x_k-x_{n+1})}\, y_k. \qquad \square$$

It's this result that is known as the Lagrange interpolation formula. Once again, remember that it's not the actual formula itself that you should commit to memory, but the method used to arrive at the formula.

9.13 Root focus

It just might be possible to get some information about a root of a polynomial. Then with some insight you just might be able to use this to your advantage as in the following problem.

Problem If p is an odd prime and n is a positive integer, prove that the polynomial

$$F(x) = x^n + x + p$$

is irreducible over $\mathbb{Z}$.

Solution Suppose that $F(x)$ is reducible, so that we can write

$$F(x) = G(x)H(x)$$

for integer polynomials $G(x)$ and $H(x)$ of positive degree.

Since $F(0) = G(0)H(0) = p$, we may assume without loss of generality that

$$G(0) = \pm 1 \quad \text{and} \quad H(0) = \pm p.$$

So, if the complex roots of $G(x)$ are $r_1, r_2, \ldots, r_k$, then with the help of Vieta's formulas, we have the equation

$$|r_1 r_2 \cdots r_k| = 1.$$

If all these complex roots had magnitude greater than 1, then their product would be greater than 1, a contradiction. Hence there must be some complex root r of $G(x)$ which satisfies $|r| \leq 1$.

But any root of $G(x)$ is also a root of $F(x)$. So $F(x)$ has a complex root r which satisfies $|r| \leq 1$. This is a very interesting piece of information indeed! This was certainly not something that was obvious from the problem's statement.

Continuing, this implies that

$$r^n + r + p = 0$$

for some $r \in \mathbb{C}$ with $|r| \leq 1$.

Now think about what this equation means geometrically. It means that the triangle in the complex plane whose vertices are r^n, $r^n + r$ and $r^n + r + p = 0$ has side lengths $|r^n| \leq 1$, $|r| \leq 1$ and $p \geq 3$. But this directly contradicts the triangle inequality, so we conclude that $x^n + x + p$ is irreducible over $\mathbb{Z}$. $\square$

10

Functional equations

Functional equations are simply equations involving functions! To read this chapter, you'll need to have some familiarity with the concept of a function and other basic notions, some of which may be found in section 17.2. To solve functional equations, you'll need an assortment of standard and not-so-standard techniques, many of which may be found in this chapter. As is common for mathematical Olympiad problems, the concepts themselves are generally quite simple, but they will be applied in particularly tricky ways.

Hopefully, you are already familiar with the following standard notation, which we'll use consistently throughout this chapter.

$\mathbb{Z} = \{\text{integers}\}$	$\mathbb{Q} = \{\text{rationals}\}$	$\mathbb{R} = \{\text{reals}\}$
$\mathbb{N}^+ = \{\text{positive integers}\}$	$\mathbb{Q}^+ = \{\text{positive rationals}\}$	$\mathbb{R}^+ = \{\text{positive reals}\}$

10.0 Problems

1. Find all functions $f\colon \mathbb{R} \to \mathbb{R}$ such that
$$f(x - f(y)) = 1 - x - y$$
for all real numbers x and y.

2. Find all functions $f : \mathbb{R} \to \mathbb{R}$ such that
$$xf(x) + f(1 - x) = x^3 - x$$
for all real numbers x.

3. Find all functions $f : \mathbb{R} \to \mathbb{R}$ such that
$$f(x + y)f(x - y) = 2x + f(x^2 - y^2)$$
for all real numbers x and y.

4. Find all functions $f : \mathbb{R} \to \mathbb{R}$ which satisfy the equations
$$f(xy) = xf(y) + yf(x) \quad \text{and} \quad f(x + y) = f(x^{1001}) + f(y^{1001})$$
for all real numbers x and y.

5. Can you find a continuous function $g : \mathbb{R} \to \mathbb{R}$ such that

$$g(g(x)) = x$$

for all $x \in \mathbb{R}$, and such that g is not a linear function?

6. Find all functions $f : \mathbb{Z} \to \mathbb{Z}$ such that

$$f(m + f(n)) = f(m) + n$$

for all integers m and n.

7. Find all functions $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ such that

$$f(x) = xf\left(\frac{1}{x}\right)$$

for all non-zero real numbers x and

$$f(x) + f(y) = 1 + f(x + y)$$

for all non-zero real numbers x and y with non-zero sum.

8. Prove that there exists no function $f\colon \mathbb{Z} \to \mathbb{Z}$ which satisfies the equation

$$f(m + f(n)) = f(m) - n$$

for all integers m and n.

9. (a) Find all strictly increasing or strictly decreasing functions $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x + f(y)) = f(x) + y$$

for all real numbers x and y.

(b) Prove that there is no strictly increasing or strictly decreasing function $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x + f(y)) = f(x) + y^2$$

for all real numbers x and y.

10. Find all continuous functions $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$f(x + y) = f(x)f(y)$$

for all $x, y \in \mathbb{R}$.

11. Prove that there exists no function $f : \mathbb{Z} \to \mathbb{Z}$ which satisfies the equation

$$f(f(n)) = n + 1$$

for all integers n.

12. Find all functions $f\colon \mathbb{R} \to \mathbb{R}$ such that

$$f(x + y) + f(x + z) + f(y + z) \geq 3f(x + 2y + 3z)$$

for all $x, y, z \in \mathbb{R}$.

13. Find all functions $f : \mathbb{Z} \to \mathbb{Z}$ such that

$$f(m+n) + f(mn) = f(m)f(n) + 1$$

for all integers m and n.

14. Find all functions $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x^2 - y^2) = (x - y)(f(x) + f(y))$$

for all real numbers x and y.

15. Find all functions $f : \mathbb{Q} \to \mathbb{R}$ such that

$$f(xy) = f(x)f(y) - f(x+y) + 1$$

for all rational numbers x and y.

16. Find all functions $f : \mathbb{N}^+ \to \mathbb{N}^+$ such that

$$f(f(m) + f(n)) = m + n$$

for all positive integers m and n.

17. Find all functions $f : \mathbb{N}^+ \to \mathbb{N}^+$ such that

$$f(f(m)f(n)) = mn$$

for all positive integers m and n.

18. Prove that there exists no function $f : \mathbb{N}^+ \to \mathbb{N}^+$ which satisfies the equation

$$f(n) = f(f(n-1)) + f(f(n+1))$$

for all integers $n > 1$.

19. Find all functions $f : \mathbb{N}^+ \to \mathbb{N}^+$ such that

$$f(n) + f(f(n)) = 2n$$

for all positive integers n.

20. Find all functions $f : \mathbb{Q} \to \mathbb{Q}$ such that

$$f(x + f(y)) = f(x)f(y)$$

for all rational numbers x and y.

21. Find all monotonic functions $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$f(4x) - f(3x) = 2x$$

for all $x \in \mathbb{R}$.

22. Let T denote the set of all ordered pairs (a, b) of non-negative integers. Find all functions $f : T \to \mathbb{R}$ such that

$$f(a,b) = \begin{cases} 0 & \text{if } ab = 0, \\ 1 + \dfrac{f(a+1,b-1) + f(a-1,b+1)}{2} & \text{otherwise.} \end{cases}$$

23. Let $\mathbb{R}_0^+$ be the set of non-negative real numbers.
Find all functions $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ such that
$$f(x+y-z) + f(2\sqrt{xz}) + f(2\sqrt{yz}) = f(x+y+z)$$
for every $x, y, z \in \mathbb{R}_0^+$ such that $x + y \geq z$.

24. Find all functions $f\colon \mathbb{R}^+ \to \mathbb{R}^+$ such that
$$\frac{f(w)^2 + f(x)^2}{f(y^2) + f(z^2)} = \frac{w^2 + x^2}{y^2 + z^2}$$
for all positive real numbers w, x, y, z satisfying $wx = yz$.

25. Let n be a given positive integer.
Find all functions $f\colon \mathbb{R} \to \mathbb{R}$ such that
$$x^n f(y) - y^n f(x) = f\left(\frac{y}{x}\right)$$
for all real x, y with $x \neq 0$.

26. Find all functions $f : \mathbb{R} \to \mathbb{R}$ such that
$$f(f(x) + y) = 2x + f(f(y) - x)$$
for all real numbers x and y.

27. Determine all bijections $f : \mathbb{R} \to \mathbb{R}$ such that
    (i) $f(x)$ is strictly increasing, and
    (ii) $f(x) + f^{-1}(x) = 2x$ for all real x.

28. Find all functions $f : \mathbb{R} \to \mathbb{R}$ such that
$$f(x^4 + y) = x^3 f(x) + f(y)$$
for all real numbers x and y.

29. Let S be the set of real numbers greater than -1. Find all functions $f : S \to S$ such that
    (i) $f(x + f(y) + xf(y)) = y + f(x) + yf(x)$ for all x and y in S, and
    (ii) $\frac{f(x)}{x}$ is strictly increasing on the intervals $-1 < x < 0$ and $0 < x$.

30. Find all non-decreasing functions $f : \mathbb{R} \to \mathbb{R}$ such that
    (i) $f(0) = 0$, $f(1) = 1$, and
    (ii) $f(a) + f(b) = f(a)f(b) + f(a + b - ab)$ for all real numbers a, b such that $a < 1 < b$.

31. Let $\mathbb{R}_0^+$ be the set of non-negative real numbers. Suppose f is a function $f\colon \mathbb{R}_0^+ \to \mathbb{R}_0^+$ satisfying the conditions
    (i) $f(a) = 0$ if and only if $a = 0$
    (ii) $f(ab) = f(a)f(b)$
    (iii) $f(a + b) \leq f(a) + f(b) + |f(a) - f(b)|$
    for all $a, b \in \mathbb{R}_0^+$.
    Prove that
$$f(a + b) \leq f(a) + f(b)$$
    for all $a, b \in \mathbb{R}_0^+$.

10.1 Cauchy's functional equation

Cauchy's functional equation asks us to find functions which satisfy

$$f(x+y) = f(x) + f(y).$$

However, the choice of domain and codomain that we are interested in will have a massive effect on what the solutions look like. This is one of the simplest and most fundamental functional equations, and its solution should be familiar to any competent problem solver!

Problem Let $\mathbb{N}_0$ denote the set of non-negative integers.

Find all functions $f : \mathbb{N}_0 \to \mathbb{R}$ such that

$$f(x+y) = f(x) + f(y)$$

for all non-negative integers x and y.

The idea is to work out the value of $f(x)$ for more and more values of x until we have the entire function.

Before we proceed, here is a useful strategy for solving functional equations. It's often a good idea to guess what the solutions are and, in this case, you should be able to find at least one without too much trouble! You should keep the suspected answer in the back of your mind while solving the functional equation, yet remain open to other possibilities. For this particular problem, one of the easiest solutions to find is $f(x) = x$. But then you might also notice that $f(x) = 2x$ is a solution and, in fact, so is $f(x) = cx$ for any real number c.

Solution So how do you start on a problem like this? For almost any functional equation, a good starting point is the substitution of values for x and y. It's a free world, so you can use whatever numbers you like, provided they are in the function's domain. In practice, however, simple substitutions such as setting the variables to be 0, 1, -1 or equal to each other are the best to start with. For this problem, the substitution $x = y = 0$ yields $f(0) + f(0) = f(0)$ which then implies that $f(0) = 0$.

Next, you might try to figure out the value of $f(1)$. But, try as you might, you won't be able to do it. If you are too narrow-minded and believe that $f(x) = x$ is the only solution, then you could get stuck at this point. That's why you need to remain open-minded to other possibilities, because if $f(x) = cx$ is a solution for any real number c, then $f(1)$ can take on any value whatsoever!

With this in mind, let $f(1) = c$ so that we can try to prove that $f(x) = cx$ for all non-negative integers x. In order to obtain the value of $f(2)$, it's natural to substitute $x = y = 1$, which leads to

$$f(2) = f(1) + f(1) = 2c.$$

In fact, putting $y = 1$ in the functional equation gives

$$f(x+1) = f(x) + f(1) = f(x) + c.$$

Therefore, if $f(x) = cx$, then $f(x+1) = cx + c = c(x+1)$.

From here you should be able to prove by induction that $f(x) = cx$ for all non-negative integers x and some real number c.

It's easy, and absolutely necessary, to check that every function of this type satisfies the given functional equation. As a courtesy to the reader, we will not usually present such routine checks, but leave them as exercises! □

Problem Find all functions $f : \mathbb{Q} \to \mathbb{R}$ such that

$$f(x + y) = f(x) + f(y)$$

for all rational numbers x and y.

Solution Note that all the hard work required to solve the previous problem carries over to this problem as well. So we already know that $f(x) = cx$ for all non-negative integers x and some real number c.

Since we have already deduced that $f(0) = 0$, it makes sense to try the substitution $y = -x$ in the functional equation. This leads to

$$f(0) = f(x) + f(-x) \quad \Rightarrow \quad f(-x) = -f(x).$$

This piece of information tells us that $f(x) = cx$ holds for every integer, whether positive, negative or zero.

Now that we have the function on all of $\mathbb{Z}$, it's time to extend our net to all of $\mathbb{Q}$. The trick is to notice that Cauchy's functional equation implies that

$$f(x + y + z) = f(x + y) + f(z) = f(x) + f(y) + f(z).$$

Hence, we can expand out sums of not only two numbers, but three numbers, or four numbers, or even n numbers.[1] In particular, for any integer m and positive integer n, we have

$$f\Bigg(\underbrace{\frac{m}{n} + \frac{m}{n} + \cdots + \frac{m}{n}}_{n \text{ times}}\Bigg) = \underbrace{f\left(\frac{m}{n}\right) + f\left(\frac{m}{n}\right) + \cdots + f\left(\frac{m}{n}\right)}_{n \text{ times}}.$$

Therefore, $f(m) = nf(\frac{m}{n})$ which implies that

$$f\left(\frac{m}{n}\right) = \frac{f(m)}{n} = \frac{cm}{n}.$$

We've now deduced that $f(x) = cx$ for all rational numbers x and some real number c. Once again, you should check that functions of this type do indeed satisfy the given functional equation. □

We have solved Cauchy's functional equation where the domain is $\mathbb{N}_0$, $\mathbb{Z}$ and $\mathbb{Q}$. So what are the solutions to Cauchy's functional equation when the domain is $\mathbb{R}$? You could be forgiven for believing that all the solutions are linear, as they were in the previous cases, but that just isn't true! Although $f(x) = cx$ is a solution, there are in fact many, many more, all of which are crazy and none of which can be described by any compact formula. To eliminate these crazy solutions, it's necessary to impose an extra condition. For example, it's known that the solutions to Cauchy's functional equation on the real numbers are given by $f(x) = cx$, if we are also given that f is continuous, that f is monotonic, or that f is bounded on some interval.

10.2 Guess and hope

We earlier advocated the strategy of trying to guess the solutions to a functional equation. Although guessing itself is rarely a substantial step toward a proof, a correct guess can sometimes make the problem a great deal simpler.

[1] To be rigorous you could use induction to get the corresponding result for n numbers.

Here's the idea. Say you have a functional equation and you think that $f(x) = x^3$ is the only solution. Then the function $g(x) = f(x) - x^3$ would be pretty simple because it would be the zero function. Another simple function would be $h(x) = \frac{f(x)}{x^3}$ or perhaps even $k(x) = \sqrt[3]{f(x)}$. This motivates us to use one of these three substitutions to simplify the original functional equation. Sometimes, but certainly not always, you may end up with a much easier problem.

Problem Find all functions $f : \mathbb{Q} \to \mathbb{R}$ such that

$$f(x + y) = f(x) + f(y) + 2xy$$

for all rational numbers x and y.

Solution The first thing you might notice is the fact that this looks like Cauchy's functional equation, except for that pesky $2xy$ term which ruins everything. You could try similar methods to those used to solve Cauchy's functional equation, but the $2xy$ term makes things a little difficult.

The second thing you might notice is the fact that $2xy$ arises when you expand $(x + y)^2$. In fact, the given functional equation bears an uncanny resemblance to the formula

$$(x + y)^2 = x^2 + y^2 + 2xy.$$

In fact, this verifies that $f(x) = x^2$ is a solution; maybe not the only solution, but a solution nonetheless.

Now it's time to guess and hope! Let's substitute $g(x) = f(x) - x^2$, that is, $f(x) = g(x) + x^2$, in the hope that $g(x)$ satisfies a simpler functional equation. This leads to

$$g(x + y) + (x + y)^2 = g(x) + x^2 + g(y) + y^2 + 2xy,$$

and we can cancel terms from both sides to leave

$$g(x + y) = g(x) + g(y).$$

You guessed it—Cauchy's functional equation! We already know that the solutions to this are given by $g(x) = cx$ for some real number c. Retracing our steps, we find that $f(x) = x^2 + cx$ for some real number c. You can easily substitute this back into the original functional equation to verify that it is indeed a solution. □

10.3 Substitutions

Suppose that you stumble upon a magic black box in which you can place an object and another object will emerge from the other side. If you want to know how the box works, you might try passing different objects through it and observing the output, until you can deduce something about the behaviour of the magic black box. Similarly, if you stumble upon a functional equation and want to solve it, a natural approach is to substitute particular numbers or, better still, algebraic expressions for the variables.

Problem Find all functions $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) \neq 0$ for $x \neq 0$ and

$$f(f(x) + y) = f(x^2 - y) + 4f(x)y$$

for all real numbers x and y.

Solution An obvious first substitution is $y = 0$, because it knocks out one of the terms on the right-hand side. In fact, it leads to the equation

$$f(f(x)) = f(x^2).$$

Trying the substitution $x = y = 0$, we discover that $f(f(0)) = f(0)$. You could continue trying various combinations of x and y equal to simple numbers like 0, 1 and -1, but you probably wouldn't get too far with this approach.

The idea here is to look for algebraic substitutions, rather than numerical ones. In particular, we would like to find algebraic substitutions which provide nice cancellation. For example, the appearance of $f(f(x) + y)$ on the left-hand side motivates us to try $y = -f(x)$, which yields a rather interesting looking equation.

$$f(0) = f(x^2 + f(x)) - 4f(x)^2$$

Furthermore, the term $f(x^2 - y)$ on the right-hand side motivates us to try $y = x^2$, which turns up yet another interesting looking equation.

$$f(f(x) + x^2) = f(0) + 4f(x)x^2$$

You probably can't help but notice that these two equations bear quite a resemblance to each other. In fact, we can use them to eliminate the unwieldy term $f(f(x) + x^2)$ and obtain

$$f(0) + 4f(x)^2 = f(0) + 4f(x)x^2 \quad \Rightarrow \quad f(x)^2 = f(x)x^2.$$

For $x = 0$, this equation simply tells us that $f(0) = 0$. But if $x \neq 0$, then $f(x) \neq 0$, and so we can divide through by $f(x)$ ending up with

$$f(x) = x^2.$$

Therefore, the only possible solution is $f(x) = x^2$ and you can check that it does indeed satisfy the original functional equation. □

This problem involved some algebraic trickery! Such deviousness is often the result of a long period of time spent playing around and substituting different combinations of the variables involved. In the end, this is a matter of intuition and experience. However, note what motivated us to substitute $y = x^2$. We turned $f(x^2 - y)$ into $f(0)$. You should always be on the lookout for substitutions which give similarly nice results.

10.4 Injective, surjective and bijective

A function is said to be *injective* or *one-to-one* if it doesn't take the same value twice. That is, $f(x) = f(y)$ implies that $x = y$. But how do you prove that a function is injective? The most direct way is just to assume that $f(x) = f(y)$ and use the functional equation somehow to deduce that $x = y$. But there are other ways, such as invoking the fact that a function which is strictly increasing or strictly decreasing must be injective. The big advantage of having an injective function f is that you can cancel f from both sides of an equation.

A function is said to be *surjective* or *onto* if it takes on all possible values in the codomain. That is, for every b in the codomain, there exists a in the domain such that $f(a) = b$. But how do you prove that a function is surjective? The most direct way is to take an arbitrary b in the codomain and use the functional equation somehow to find a value of a for which

$f(a) = b$. The big advantage of having a surjective function f is that you can then substitute $f(a) = b$ for any value of b in the codomain.

A function is said to be *bijective* if it is both injective and surjective. The big advantage of having a bijective function f is that there exists an inverse function f^{-1} which satisfies

$$f^{-1}(f(x)) = x \quad \text{and} \quad f(f^{-1}(x)) = x.$$

These three properties are supremely important when solving functional equations. It's a standard tactic, when confronted with a difficult problem, to try to prove that the function involved has some of these properties. All of the previous discussion might sound completely cryptic to you and, if that's the case, you'll probably feel more enlightened after considering the following example.

Problem The function $f : \mathbb{R} \to \mathbb{R}$ satisfies

$$f(xf(x) + f(y)) = f(x)^2 + y$$

for all real numbers x and y.

Prove that f is bijective.

Solution Think about holding x constant. For example, let $x = 0$ and vary y. Then the right-hand side can attain any real value. Since the left-hand side is f applied to some expression, this proves that f is surjective.

You might be a little suspicious of this argument, but we can also express it in more algebraic terms. To prove that f is surjective, for any real number b we must be able to find a real number a such that $f(a) = b$. This value of a is simply $f(b - f(0)^2)$, since the functional equation with $x = 0$ and $y = b - f(0)^2$ implies that

$$f(f(b - f(0)^2)) = b.$$

If the extra algebra didn't help your understanding, that's fine—it was probably just complicating what is actually quite a simple argument.

Now we turn our attention to showing that f is injective. To do this, we assume that $f(y_1) = f(y_2)$ and hope to deduce that $y_1 = y_2$. But if $f(y_1) = f(y_2)$, then we have

$$\begin{aligned} & f(xf(x) + f(y_1)) = f(xf(x) + f(y_2)) \\ \Rightarrow \quad & f(x)^2 + y_1 = f(x)^2 + y_2. \end{aligned}$$

Therefore, we can conclude that $y_1 = y_2$ and so f must be injective.

Since we have shown that f is both injective and surjective, we now know that f is bijective. □

If you examine the previous proof closely, you'll see that we relied on two particular features of the functional equation. The first is that the left-hand side involves only $f(y)$, but not y. The second is that the right-hand side involves only y, but not $f(y)$. Variables which appear in this way often provide the key to proving the fact that a function is injective, surjective or bijective. Unfortunately, this particular example doesn't really demonstrate just how helpful it is to know that f is bijective. But we'll see various illustrative examples of this as we progress.

10.5 The associative[2] trick

One of the most common strategies when solving a functional equation is to look for one expression which can be evaluated in two different ways. Doing so will often provide some extra information that is crucial to solving the problem.

One such possibility involves nested functions. For example, $f(g(h(x)))$ can be evaluated in two different ways

$$\underbrace{f(g}(h(x))) \quad \text{or} \quad f(\underbrace{g(h}(x)))$$

depending on which composition of functions is evaluated first.

Problem Do there exist functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ such that

$$f(g(x)) = x^2 \quad \text{and} \quad g(f(x)) = x^3$$

for all real numbers x?

Solution The following chain of implications tells us that the function f must be injective.

$$f(a) = f(b) \quad \Rightarrow \quad g(f(a)) = g(f(b)) \quad \Rightarrow \quad a^3 = b^3 \quad \Rightarrow \quad a = b$$

Now we notice that the given information about f and g allows us to calculate the expression $f(g(f(x)))$ in two different ways.

$$\underbrace{f(g}(f(x))) = f(x)^2 \quad \text{and} \quad f(\underbrace{g(f}(x)) = f(x^3)$$

Therefore,

$$f(x)^2 = f(x^3)$$

for all values of x. In particular, we know that

$$f(-1) = f(-1)^2, \quad f(0) = f(0)^2 \quad \text{and} \quad f(1) = f(1)^2.$$

It follows that $f(-1)$, $f(0)$ and $f(1)$ are all equal to 0 or 1 and so two of them are equal to each other. This directly contradicts the fact that f is injective, so we conclude that there cannot exist functions satisfying the conditions of the problem. □

Note that the associative trick is particularly useful where the expression $f(f(x))$ is known. For example, if

$$f(f(x)) = x + 1,$$

then applying the associative trick to $f(f(f(x)))$ tells us that

$$f(x + 1) = f(x) + 1.$$

[2] An operation $*$ is said to be *associative* if $f * (g * h) = (f * g) * h$ always holds true. In our setting here, the operation is function composition. Other examples of associative operations include the ordinary arithmetic operations of addition and multiplication.

10.6 Exploit symmetry

Another context for evaluating one expression in two different ways is when symmetry arises as illustrated in the following example.

Problem Find all functions $f : \mathbb{Z} \to \mathbb{Z}$ such that

$$f(m + f(n) + mf(n)) = m + mn + f(n)$$

for all integers m and n.

Solution The crucial observation is the following. If we substitute $m = f(p)$ for any integer p, the left-hand side is a symmetric expression in p and n. In other words, it remains the same when we swap p and n. However, the right-hand side is not symmetric in p and n.

$$f(f(p) + f(n) + f(p)f(n)) = f(p) + f(p)n + f(n)$$

Since swapping p and n in the above equation leaves the left-hand side unchanged, it must also leave the right-hand side unchanged. Thus

$$f(p) + f(p)n + f(n) = f(n) + f(n)p + f(p).$$

Hence

$$f(p)n = f(n)p$$

for all integers p and n.

Substituting $p = 1$, we deduce that $f(n) = f(1)n$. Thus any solution to this functional equation must be of the form $f(n) = cn$.

However, as with all functional equations, we must check our solutions. If we plug $f(n) = cn$ into the original functional equation, we deduce that $c = 1$. So, the only possible solution is $f(n) = n$. □

10.7 Involutions

An *involution* is a function which is its own inverse. That is, a function f which satisfies

$$f(f(x)) = x$$

for all x. Involutions can be surprisingly useful and crop up in the most mysterious ways.

Problem Find all functions $g : \mathbb{R} \setminus \{\frac{2}{3}\} \to \mathbb{R}$ such that

$$x - g(x) = \frac{1}{2}\, g\left(\frac{2x}{3x - 2}\right)$$

for all real numbers $x \neq \frac{2}{3}$.

Solution Though it might look scary and complicated, this functional equation is actually simpler than most. For example, it only contains one variable x. All the equation does is relate two different values of g, that is, the value at x and the value at $\frac{2x}{3x-2}$. This begs the

following question: if x is related to $\frac{2x}{3x-2}$, then what is $\frac{2x}{3x-2}$ related to? With this in mind, let's replace x with $\frac{2x}{3x-2}$ in the functional equation and hope for the best.

$$\left(\frac{2x}{3x-2}\right) - g\left(\frac{2x}{3x-2}\right) = \frac{1}{2}g\left(\frac{2(\frac{2x}{3x-2})}{3(\frac{2x}{3x-2})-2}\right)$$

Yes, it looks messy, but a little algebra should convince you that

$$\frac{2(\frac{2x}{3x-2})}{3(\frac{2x}{3x-2})-2} = \frac{4x}{6x-2(3x-2)} = x.$$

So we now have the equations

$$x - g(x) = \frac{1}{2}g\left(\frac{2x}{3x-2}\right) \quad \text{and} \quad \left(\frac{2x}{3x-2}\right) - g\left(\frac{2x}{3x-2}\right) = \frac{1}{2}g(x).$$

You can think of these as two simultaneous equations, which you can solve for both $g(x)$ and $g(\frac{2x}{3x-2})$. Do this and you'll find that

$$g(x) = \frac{4x(x-1)}{3x-2},$$

which is indeed the solution to the functional equation. □

What's going on here? Obviously there's something special about the expression $\frac{2x}{3x-2}$. In fact if you let $f(x) = \frac{2x}{3x-2}$, then $f(f(x)) = x$, which means that it's an involution!

10.8 Fixed points

A fixed point of a function f is a value of x for which $f(x) = x$. Despite being such a simple concept, it's often an extremely useful strategy to consider the fixed points of a function. In fact, interesting deductions often arise from the existence of fixed points.

Problem Find all functions $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that

(i) $f(xf(y)) = yf(x)$ for all positive real numbers x and y, and

(ii) $f(x) < 1000$ for all $x > 1000$.

Solution As a first step, let's prove that the function must be bijective. Think about holding x constant. For example, set $x = 1$ and vary y. Then the right-hand side can attain any positive real value. Since the left-hand side is f applied to some expression, this proves that f is surjective.

If we assume that $f(y_1) = f(y_2)$, then it is certainly true that

$$f(xf(y_1)) = f(xf(y_2)) \quad \Rightarrow \quad y_1 f(x) = y_2 f(x).$$

Therefore, we can deduce that $y_1 = y_2$ and so f must be injective.

Since we've shown that f is both injective and surjective, we now know that f is bijective.

Let's use these facts to determine the value of $f(1)$. The big advantage of having a surjective function lies in the fact that there exists a positive real number k for which $f(k) = 1$. Substituting $y = k$ in the functional equation gives

$$f(xf(k)) = kf(x) \quad \Rightarrow \quad f(x) = kf(x) \quad \Rightarrow \quad k = 1.$$

Therefore, we have $f(1) = f(k) = 1$. This is a fairly indirect way to find $f(1)$, but the technique is surprisingly common and useful. Finding just one value of the function, like $f(1)$, can often lead to a windfall of other benefits. For example, if we now set $x = 1$ in the functional equation, we obtain the fact that

$$f(f(y)) = y.$$

In other words, f is an involution.

A useful substitution in this case is $x = y$, which implies that

$$f(xf(x)) = xf(x).$$

This is interesting because it means that the number $xf(x)$ is a fixed point of f for any value of x. Such a fact seems to imply that f must have a whole lot of fixed points. On the other hand, you can't have too many fixed points, for the second condition guarantees that every fixed point is less than or equal to 1000. All this discussion really motivates us to concentrate on the fixed points of f. We have the following observations.

- Since $f(1) = 1$, the number 1 is a fixed point of f.
- Suppose that a is a fixed point of f, so that $f(a) = a$. Now substitute $x = \frac{1}{a}$ and $y = a$ into the functional equation to obtain

  $$f\left(\frac{f(a)}{a}\right) = af\left(\frac{1}{a}\right) \quad \Rightarrow \quad f(1) = af\left(\frac{1}{a}\right) \quad \Rightarrow \quad f\left(\frac{1}{a}\right) = \frac{1}{a}.$$

  Hence, $\frac{1}{a}$ is also a fixed point of f.
- Suppose that a and b are fixed points of f, so that $f(a) = a$ and $f(b) = b$. Now substitute $x = a$ and $y = b$ into the functional equation to obtain

  $$f(af(b)) = bf(a) \quad \Rightarrow \quad f(ab) = ab.$$

  Hence, ab is also a fixed point of f. It follows that, if a is a fixed point of f, then so are $a^2, a^3, a^4, \ldots$.

Now let's piece these three statements together in a clever way to show that 1 is the only fixed point of f.

Suppose that $a > 1$ is a fixed point, then the fixed points

$$a^2, a^3, a^4, \ldots$$

would eventually become greater than 1000, a contradiction.

Similarly, if $a < 1$ is a fixed point, then so is $\frac{1}{a} > 1$ and the fixed points

$$\frac{1}{a^2}, \frac{1}{a^3}, \frac{1}{a^4}, \ldots$$

would eventually become greater than 1000, another contradiction.

Thus, the only possible fixed point of f is 1. And yet we also know that $xf(x)$ is a fixed point for any value of x. The only way that this can happen is if $xf(x) = 1$ or, in other words, if

$$f(x) = \frac{1}{x}$$

for any value of x.

As always, you should substitute this solution back into the functional equation to make sure that it actually works. □

Of course, there was a great deal of algebraic trickery involved in this solution. Rather than remarking bewilderedly 'I could never think of that!', you should examine the solution again and instead ask 'What factors are present in this problem that would lead me to try that technique?' For all, well almost all, mathematicians are ordinary mortals. And most likely any solution you read to a difficult problem, however short, has involved a great deal of effort.

10.9 Somewhere versus everywhere

We come now to a common fallacy that is the bane of many a beginner functional equation solver! It concerns the relationship between individual points and the entire domain, between local information and global information, between somewhere and everywhere.

A prime example to illustrate this point is the functional equation we introduced in section 10.4. In that section, we demonstrated that the function is bijective. Now we capitalise on our previous work and solve the functional equation once and for all.

Problem Find all functions $f : \mathbb{R} \to \mathbb{R}$ which satisfy

$$f(xf(x) + f(y)) = f(x)^2 + y$$

for all real numbers x and y.

Solution Since we've already shown that f is bijective, we know that there exists a real number k for which $f(k) = 0$. Substituting $x = k$ into the functional equation immediately implies that

$$f(f(y)) = y,$$

which means that f is an involution.

The trick now is to employ the 'one expression, two ways' philosophy. Note that the expression $xf(x)$ is invariant—that is, does not change—when you replace x with $f(x)$. This follows from our newly obtained result which states that $f(f(x)) = x$. So when we replace x with $f(x)$ in the given functional equation, we obtain

$$f(f(x)f(f(x)) + f(y)) = f(f(x))^2 + y \quad \Rightarrow \quad f(xf(x) + f(y)) = x^2 + y.$$

Comparing this with the original functional equation leads directly to

$$f(x)^2 = x^2 \quad \Rightarrow \quad f(x) = \pm x.$$

And this is where we should tread carefully, for herein lies a trap for the unsuspecting functional equation enthusiast. A beginner might now conclude that there are two solutions, namely, $f(x) = x$ and $f(x) = -x$. But this reasoning is entirely incorrect, because all we have deduced is that for each real number x, the value of $f(x)$ is either x or $-x$. There are

still infinitely many possibilities, since the function could conceivably take the value x at some places and $-x$ at others!

Such crazy solutions to the functional equation seem improbable though, so let's try to eliminate them. To do this, suppose that there exist two distinct real numbers x and y such that $f(x) = x$ and $f(y) = -y$. In fact, we might as well assume that x and y are non-zero, since $f(0) = 0$ in either case. Now substitute these values into the functional equation to obtain

$$f(x^2 - y) = x^2 + y.$$

However, we know that $f(x^2 - y) = \pm(x^2 - y)$. If we take the positive sign, we obtain $y = 0$ and if we take the negative sign, we obtain $x = 0$. Both cases contradict the fact that x and y are non-zero.

Therefore, we can finally conclude that the only solutions to the functional equation are $f(x) = x$ and $f(x) = -x$. □

If you really want to make sure that you never fall into the 'somewhere versus everywhere' trap, then you should return to the problem from section 10.3 and solve it without the condition that $f(x) \neq 0$ for $x \neq 0$.

10.10 Completely multiplicative functions

Thus far, our focus has mainly been on functional equations where $\mathbb{R}$ is the domain. Many of the same techniques work when the domain is $\mathbb{N}^+$ or $\mathbb{Q}$, but sometimes you need to pull out a brand new trick. The one we'll look at here is based on the prime factorisation of positive integers. Recall the fundamental theorem of arithmetic, which asserts that every positive integer n can be uniquely expressed as a product of primes

$$n = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k},$$

where $p_1, \ldots, p_k$ are primes and $a_1, a_2, \ldots, a_k$ are positive integers. In fact, if we allow negative powers of primes, then we may uniquely express any positive rational number as well.

These concepts will be particularly useful for dealing with *completely multiplicative functions*. These are functions f which satisfy

$$f(mn) = f(m)f(n)$$

for every m and n in the domain.

Problem Find all completely multiplicative functions $f : \mathbb{N}^+ \to \mathbb{N}^+$.

The unsuspecting reader might assume that, much like other functional equations we've seen, the only solution is given by $f(n) = n$. But he or she could not be further from the truth.

Solution A first step would be to substitute simple numbers such as $m = n = 1$ into the functional equation $f(mn) = f(m)f(n)$. This would give us the fact that

$$f(1) = 1.$$

Now observe that the functional equation allows us to write

$$f(kmn) = f(km)f(n) = f(k)f(m)f(n).$$

More generally, you can prove by induction that

$$f(m_1 m_2 \cdots m_k) = f(m_1) f(m_2) \cdots f(m_k).$$

Thus, if we take the prime factorisation $n = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$, then we have

$$f(n) = f(p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}) = f(p_1)^{a_1} f(p_2)^{a_2} \cdots f(p_k)^{a_k}.$$

So the function can be completely described by its values at the primes, namely,

$$f(2), f(3), f(5), f(7), \ldots.$$

If you still believe that $f(n) = n$ is the only solution, then you could try to go further and show that $f(p) = p$ for every prime p. Unfortunately, you won't get anywhere! This is because we've deduced as much as we can possibly deduce.

In fact, as long as we take $f(1) = 1$, we can set $f(2), f(3), f(5), f(7), \ldots$ to be any arbitrary positive integer sequence and any such choice will give us a completely multiplicative function $f : \mathbb{N}^+ \to \mathbb{N}^+$.

You shouldn't just believe this, so check it for yourself. In conclusion, this functional equation has infinitely many solutions given by the previous description. □

10.11 Well-ordering of $\mathbb{N}^+$

One property of $\mathbb{N}^+$ not shared by $\mathbb{Q}$ or $\mathbb{R}$ is the fact that every subset of $\mathbb{N}^+$ has a smallest element. It seems so simple, yet this property can be incredibly useful, so much so that it has the grand sounding name of *well-ordering*.

Problem The function $f : \mathbb{N}^+ \to \mathbb{N}^+$ satisfies

$$f(n+1) > f(f(n))$$

for each positive integer n.

Prove that $f(n) = n$ for each positive integer n.

Solution The first thing you should do is try to construct such a function. You'll probably find that $f(n) = n$ is an easy one to come up with, but other possibilities tend to fail because you obtain various inequalities leading to other inequalities leading to contradictions. It's hard to see why these fail in general. Using the well-ordering of $\mathbb{N}^+$ allows us to consider minimality, which will aid us considerably here.

The idea is to suppose that $f(k)$ is the smallest member of the set $S = \{f(1), f(2), f(3), \ldots\}$. Of course, there might be more than one such k, but we only need one of them. The functional equation tells us that

$$f(k) > f(f(k-1)).$$

This seems to contradict the fact that $f(k)$ is the smallest member of S. The only way to avoid this contradiction is to have $k = 1$. In this way the inequality doesn't arise because the domain of f is the set of positive integers. This sleight of hand proves that $f(1)$ is the unique smallest member of S.[3]

[3] The term 'unique smallest' is a little imprecise. After all, S only has one smallest member anyway. What we mean is that if m is the smallest member of S, then the only number that satisfies $f(n) = m$, is $n = 1$. The term has an analogous meaning for 'unique second smallest', and so on, later in the proof.

Next, what's the second smallest member of S? It must be $f(k)$ for some $k \geq 2$. Again, there might be more than one such k, but we only need one of them. The functional equation tells us that

$$f(k) > f(f(k-1)).$$

If $f(k-1) > 1$, we would have $f(f(k-1)) > f(1)$. This would contradict the fact that $f(k)$ is the second smallest member of S. So we must have $f(k-1) = 1$. This implies that the smallest member of S is 1 and that $k = 2$. Thus another sleight of hand proves that $f(2)$ is the unique second smallest member of S.

What's the third smallest member of S? It is $f(k)$ for some $k \geq 3$. The functional equation tells us that

$$f(k) > f(f(k-1)).$$

If $f(k-1) > 2$, we would have $f(f(k-1)) > f(2) > f(1)$. This would contradict the fact that $f(k)$ is the third smallest member of S. We certainly can't have $f(k-1) = 1$ because this occurs only for $k = 2$. So we must have $f(k-1) = 2$. This implies that the second smallest member of S is 2 and that $k = 3$.

This argument can be continued inductively[4] to show that the members of S written in increasing order are

$$f(1) < f(2) < f(3) < \cdots.$$

It also shows that the members of S written in increasing order are

$$1 < 2 < 3 < \cdots.$$

From this we deduce that $f(n) = n$ is the only solution. □

[4]This proof cannot be considered complete as it currently stands. Can you complete the proof by writing out the details of the induction yourself?

11

Inequalities

An inequality is a mathematical problem which asks you to prove that some expression is always greater than (or equal to) some other expression. For example, you might have to show that for any real numbers a, b, c, the following holds.

$$a^2 + b^2 + c^2 \geq ab + bc + ca$$

Although daunting at first, you should be able to solve a whole variety of inequalities after mastering a handful of techniques and tricks. That is exactly what this chapter will help you to achieve.

11.0 Problems

1. Farmer Brown wants a rectangular paddock of area A next to a long straight river, with a fence on the other three sides.

   What is the minimum length of fencing that is required?

2. (a) Use the AM–GM inequality to find the maximum value of xyz, where x, y, z are positive real numbers satisfying $x + 2y + 3z = 3$.

   (b) Use the AM–GM inequality to find the minimum value of $x + y + z$, where x, y, z are positive real numbers satisfying $xy^2z^3 = 108$.

3. Prove that if a, b, c are positive real numbers satisfying $abc = 1$, then

$$(a + b)(b + c)(c + a) \geq 8.$$

4. Let x and y be positive numbers such that $x + y = 1$. Show that

$$\left(1 + \frac{1}{x}\right)\left(1 + \frac{1}{y}\right) \geq \frac{9}{4}.$$

5. If a, b, c are positive, prove in at least three different ways that

$$\frac{a}{b + c} + \frac{b}{c + a} + \frac{c}{a + b} \geq \frac{3}{2}.$$

6. Use the rearrangement inequality to prove the following, where a, b, c are positive real numbers.

   (a) $a^2b + b^2c + c^2a \leq a^3 + b^3 + c^3$

   (b) $\dfrac{1}{a} + \dfrac{1}{b} + \dfrac{1}{c} \leq \dfrac{a}{b^2} + \dfrac{b}{c^2} + \dfrac{c}{a^2}$

   (c) $a^{bc}b^{ca}c^{ab} \leq a^{ab}b^{bc}c^{ca}$

7. Use the Cauchy–Schwarz inequality to find the minimum value of

$$x^2 + y^2 + z^2,$$

   where x, y, z are real numbers satisfying $3x + 4y + 5z = 10$.

8. For positive real numbers a, b, c, prove that

$$\frac{a+b+c}{3} \geq \sqrt{\frac{ab+bc+ca}{3}}.$$

9. Let a, b, c be positive numbers satisfying

$$\frac{1}{a} + \frac{1}{b} + \frac{1}{c} = 1.$$

   Prove that

$$(a-1)(b-1)(c-1) \geq 8.$$

10. For positive real numbers a, b, c, d, prove that

$$\frac{1}{a} + \frac{1}{b} + \frac{4}{c} + \frac{16}{d} \geq \frac{64}{a+b+c+d}.$$

11. Let A, B, C be the angles of a triangle. If the triangle is acute, use Jensen's inequality to prove the following inequalities. Also determine which of these inequalities remain true if the triangle is obtuse.

   (a) $\sin A + \sin B + \sin C \leq \dfrac{3\sqrt{3}}{2}$

   (b) $\cos A + \cos B + \cos C \leq \dfrac{3}{2}$

   (c) $\cos A \cos B \cos C \leq \dfrac{1}{8}$

12. If $a_1, a_2, \ldots, a_n$ are distinct positive integers, prove that

$$\frac{a_1}{1^2} + \frac{a_2}{2^2} + \cdots + \frac{a_n}{n^2} \geq \frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n}.$$

13. If $a_1, a_2, \ldots, a_n$ are positive numbers whose product is 1, prove that

$$(2 + a_1)(2 + a_2) \cdots (2 + a_n) \geq 3^n.$$

14. Let a, b, c be positive real numbers such that $abc = 1$. Prove that

$$\frac{bc}{b^5 + c^5 + bc} + \frac{ca}{c^5 + a^5 + ca} + \frac{ab}{a^5 + b^5 + ab} \leq 1$$

    and determine when equality holds.

15. If a, b, c are positive real numbers, prove that

$$\frac{9}{a+b+c} \leq \frac{2}{a+b} + \frac{2}{b+c} + \frac{2}{c+a} \leq \frac{1}{a} + \frac{1}{b} + \frac{1}{c}.$$

16. Prove that if a, b, c are positive real numbers with sum less than 1, then

$$\frac{abc(1-a-b-c)}{(a+b+c)(1-a)(1-b)(1-c)} \leq \frac{1}{81}.$$

17. Suppose that the numbers x, y, z are greater than or equal to 1 and satisfy

$$\frac{1}{x} + \frac{1}{y} + \frac{1}{z} = 2.$$

Prove that

$$\sqrt{x+y+z} \geq \sqrt{x-1} + \sqrt{y-1} + \sqrt{z-1}.$$

18. For positive real numbers x, y, z, show that

$$\frac{x}{(x+y)(x+z)} + \frac{y}{(y+z)(y+x)} + \frac{z}{(z+x)(z+y)} \leq \frac{9}{4(x+y+z)}.$$

19. Let x, y, z be positive real numbers which satisfy $xyz = 1$. Show that

$$\frac{x^3}{(1+y)(1+z)} + \frac{y^3}{(1+z)(1+x)} + \frac{z^3}{(1+x)(1+y)} \geq \frac{3}{4}.$$

20. Let $a_1, a_2, \ldots, a_n$ be positive real numbers satisfying

$$\frac{1}{a_1+100} + \frac{1}{a_2+100} + \cdots + \frac{1}{a_n+100} = \frac{1}{100}.$$

Prove that

$$\sqrt[n]{a_1 a_2 \cdots a_n} \geq 100(n-1).$$

21. Let $n \geq 3$ be an integer, and let $a_2, a_3, \ldots, a_n$ be positive real numbers such that $a_2 a_3 \cdots a_n = 1$. Prove that

$$(1+a_2)^2 (1+a_3)^3 \cdots (1+a_n)^n \ > \ n^n.$$

11.1 Squares are non-negative

A simple though useful inequality is the fact that, when you multiply a real number by itself, the answer is always zero or positive. In short, *squares are non-negative*.

Problem Prove that

$$x^4 > 8x - 9$$

for every real number x.

Solution The given inequality is equivalent to

$$(x^2 - 1)^2 + 2(x - 2)^2 > 0.$$

You should check this right now with pen and paper to make sure that we're not lying to you! Once you've done that, the inequality should be almost self-evident since the left-hand side is a sum of squares, which we know to be non-negative. Furthermore, the first term is zero only when $x = \pm 1$, while the second term is zero only when $x = 2$. So it's impossible for them to be 0 simultaneously, and this guarantees that the left-hand side is not only non-negative, but always positive. □

Of course, the strategy we used here involved taking all terms to the left-hand side and then magically discovering that it could be expressed as a sum of squares. Although this may not have been apparent to you from the outset, spotting squares becomes much easier after one has had some square-spotting practice.

Problem Farmer Black wants a rectangular paddock of area A with a fence on each side. What is the minimum length of fencing that is required?

Solution Let the side lengths of Farmer Black's rectangular paddock be x and y. We are given that $xy = A$ and would like to find the minimum possible value of $2x + 2y$. Your intuition probably tells you that this will occur when x and y are equal to each other, in which case the rectangular paddock would actually be a square with perimeter $4\sqrt{A}$. So let's try to prove that

$$2x + 2y \geq 4\sqrt{A},$$

or equivalently,

$$\frac{x + y}{2} \geq \sqrt{xy}.$$

After squaring both sides[1], collecting all terms on the left-hand side, and factorising, the inequality becomes

$$x^2 + 2xy + y^2 \geq 4xy \quad \Leftrightarrow \quad x^2 - 2xy + y^2 \geq 0 \quad \Leftrightarrow \quad (x - y)^2 \geq 0.$$

Of course, this is true because squares are non-negative. Therefore, we have succeeded in proving that the minimum length of fencing that Farmer Black requires is $4\sqrt{A}$. □

[1] Generally LHS $\geq$ RHS is not equivalent to LHS$^2 \geq$ RHS^2. But if it is known that LHS ≥ 0 and RHS ≥ 0, then it is equivalent.

11.2 AM–GM inequality

In the previous section, we used the inequality $\frac{x+y}{2} \geq \sqrt{xy}$, which can be expressed as follows: the average of two non-negative numbers is at least as large as the square root of their product. More generally, it's true that the average of n non-negative numbers is at least as large as the nth root of their product.

AM–GM inequality If $a_1, a_2, \ldots, a_n$ are non-negative real numbers, then

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 a_2 \cdots a_n}.$$

Furthermore, equality occurs if and only if $a_1 = a_2 = \cdots = a_n$.

The inequality is so named because the left-hand side is known as the *arithmetic mean*—a fancy name for what we usually call the average—while the right-hand side is known as the *geometric*[2] *mean*.

Problem If a, b, c are positive and satisfy $a + b + c = 1$, prove that

$$\frac{1}{ab} + \frac{1}{bc} + \frac{1}{ca} \geq 27.$$

Solution Since we have the condition $a + b + c = 1$, the expression we wish to minimise is

$$\frac{1}{ab} + \frac{1}{bc} + \frac{1}{ca} = \frac{a+b+c}{abc} = \frac{1}{abc}.$$

In other words, we would like to maximise the product abc.

The AM–GM inequality is often the key when dealing with inequalities which involve sums and products. Applying it to a, b, c, we obtain

$$\frac{a+b+c}{3} \geq \sqrt[3]{abc} \quad \Rightarrow \quad \frac{1}{3} \geq \sqrt[3]{abc} \quad \Rightarrow \quad abc \leq \frac{1}{27}.$$

It follows that

$$\frac{1}{ab} + \frac{1}{bc} + \frac{1}{ca} = \frac{1}{abc} \geq 27,$$

as desired. □

11.3 Rearrangement inequality

Suppose that you are on a television game show and that there are three piles of money in front of you: one consisting of \$5 notes, one consisting of \$20 notes and one consisting of \$100 notes. For your prize, you are allowed to take 100 notes from one pile, 200 notes from another pile and 300 notes from the remaining pile. What should your strategy be? Well, the greedy capitalist inside you would probably take 300 \$100 notes, 200 \$20 notes and 100 \$5 notes, in order to maximise your winnings. And of course, in order to minimise your winnings, you would do the exact opposite, that is, take 300 \$5 notes, 200 \$20 notes and 100 \$100 notes. This greedy principle is the basis for the following.

[2]The arithmetic mean can be thought of as the average with respect to the addition operation, while the geometric mean can be thought of as the average with respect to the multiplication operation.

Rearrangement inequality Let $(a_1, a_2, \ldots, a_n)$ and $(x_1, x_2, \ldots, x_n)$ be two sequences of real numbers. Suppose we seek a rearrangement (i.e. permutation) $(b_1, b_2, \ldots, b_n)$ of $(x_1, x_2, \ldots, x_n)$ which maximises the expression

$$E = a_1b_1 + a_2b_2 + \cdots + a_nb_n.$$

Then E is maximised when the sequences $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ are sorted the same way.
If, on the other hand, we are trying to minimise E, then this occurs when the sequences $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ are sorted the opposite way.

Two sequences of numbers $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ are said to be *sorted the same way* if the largest a_i is in the same position as the largest b_i, the second largest a_i is in the same position as the second largest b_i, and so on. On the other hand, they are said to be *sorted the opposite way* if the largest a_i is in the same position as the smallest b_i, the second largest a_i is in the same position as the second smallest b_i, and so on.

For future convenience, we'll refer to the expression

$$a_1b_1 + a_2b_2 + \cdots + a_nb_n$$

as the *product* of the sequences $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$.

Problem If $x_1, x_2, \ldots, x_n$ are positive real numbers, show that

$$\frac{x_1^2}{x_2} + \frac{x_2^2}{x_3} + \cdots + \frac{x_{n-1}^2}{x_n} + \frac{x_n^2}{x_1} \geq x_1 + x_2 + \cdots + x_n.$$

Solution The left-hand side of the inequality looks suspiciously like the product of two sequences,

$$(x_1^2, x_2^2, \ldots, x_n^2) \quad \text{and} \quad \left(\frac{1}{x_2}, \frac{1}{x_3}, \ldots, \frac{1}{x_1}\right).$$

You should be able to see that the right-hand side can also be written as a product of the same two sequences rearranged, that is,

$$(x_1^2, x_2^2, \ldots, x_n^2) \quad \text{and} \quad \left(\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}\right).$$

So, by the rearrangement inequality, the desired result will follow once we prove that these latter two sequences are sorted the opposite way. But this is certainly true since

$$x_i^2 \leq x_j^2 \quad \text{if and only if} \quad \frac{1}{x_i} \geq \frac{1}{x_j}.$$

In other words, if x_i^2 is the largest term of the first sequence, then $\frac{1}{x_i}$ is the smallest term of the second sequence; if x_i^2 is the second largest term of the first sequence, then $\frac{1}{x_i}$ is the second smallest term of the second sequence; and so on. □

The rearrangement inequality is a surprisingly useful result, so you should take the time to really understand the previous argument before moving on.

Problem For positive real numbers a, b, c, prove that

$$a^a b^b c^c \geq a^b b^c c^a.$$

Solution This inequality involves products of powers, whereas the rearrangement inequality involves sums of products. This suggests that we should take the logarithm of both sides. And we can do this without any change of the inequality sign, since $\log x$ is an increasing function on the set of positive real numbers. Thus, we obtain the equivalent inequality

$$a \log a + b \log b + c \log c \geq b \log a + c \log b + a \log c.$$

Note that the left-hand side is a product of the two sequences

$$(a, b, c) \quad \text{and} \quad (\log a, \log b, \log c).$$

These are sorted the same way, since the log function is increasing, as we already mentioned. So by the rearrangement inequality, the left-hand side is certainly greater than or equal to the product of the two sequences

$$(a, b, c) \quad \text{and} \quad (\log c, \log a, \log b),$$

which just so happens to be the right-hand side of the inequality. □

As a final comment on this section we note that the rearrangement inequality generalises to more than two sequences of numbers. Here we state the version for three sequences of numbers.[3]

Rearrangement inequality for three sequences of numbers Let $(a_1, a_2, \ldots, a_n)$, $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ be three sequences of positive real numbers. Suppose we seek rearrangements $(b_1, b_2, \ldots, b_n)$ of $(x_1, x_2, \ldots, x_n)$ and $(c_1, c_2, \ldots, c_n)$ of $(y_1, y_2, \ldots, y_n)$ which maximise the expression

$$E = a_1 b_1 c_1 + a_2 b_2 c_2 + \cdots + a_n b_n c_n.$$

Then E is maximised if the three sequences are sorted the same way.

Note that in contrast to the two sequence version of the rearrangement inequality, there is no simple criterion for minimising E.

11.4 Cauchy–Schwarz inequality

Cauchy–Schwarz inequality If $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are two sequences of real numbers, then

$$(x_1^2 + x_2^2 + \cdots + x_n^2)(y_1^2 + y_2^2 + \cdots + y_n^2) \geq (x_1 y_1 + x_2 y_2 + \cdots + x_n y_n)^2.$$

Furthermore, equality occurs if and only if we have the equal ratios

$$x_1 : y_1 = x_2 : y_2 = \cdots = x_n : y_n.$$

The Cauchy–Schwarz inequality is often neglected by the budding inequality problem solver, mainly because it is difficult to know when and how to use it. However, it can be a mighty weapon in the hands of an expert. Here are some general tips.

- If there are squares or square roots, try Cauchy–Schwarz.

[3] Note that the numbers in the three sequence version must all be positive. This is in contrast to the two sequence version where they only need to be real.

- If there are products or fractions, try Cauchy–Schwarz.
- If all else fails, try Cauchy–Schwarz.

Problem Prove the AM–HM inequality[4], which states that if $a_1, a_2, \ldots, a_n$ are positive real numbers, then

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}}.$$

Furthermore, show that equality occurs if and only if $a_1 = a_2 = \cdots = a_n$.

Solution Since $a_1, a_2, \ldots, a_n$ are positive real numbers, we can simply take the two sequences

$$(\sqrt{a_1}, \sqrt{a_2}, \ldots, \sqrt{a_n}) \quad \text{and} \quad \left(\frac{1}{\sqrt{a_1}}, \frac{1}{\sqrt{a_2}}, \ldots, \frac{1}{\sqrt{a_n}}\right)$$

and substitute them into the Cauchy–Schwarz inequality.

We obtain

$$(a_1 + a_2 + \cdots + a_n)\left(\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}\right) \geq n^2,$$

which gives the desired result after some rearranging.

Equality occurs if and only if

$$\sqrt{a_1} : \frac{1}{\sqrt{a_1}} = \sqrt{a_2} : \frac{1}{\sqrt{a_2}} = \cdots = \sqrt{a_n} : \frac{1}{\sqrt{a_n}},$$

which is equivalent to the fact that $a_1 = a_2 = \cdots = a_n$. □

Problem Let a, b, c be positive real numbers such that $abc = 1$. Prove that

$$\frac{1}{a^3(b+c)} + \frac{1}{b^3(c+a)} + \frac{1}{c^3(a+b)} \geq \frac{3}{2}.$$

Solution A common starting strategy is to massage the given inequality into a nicer looking form by using substitutions or other algebraic trickery. In this case, it would be nice to get rid of the unsightly cubes appearing in the denominator of each term. We can improve the situation greatly if we opt for the clever substitution $(a, b, c) = \left(\frac{1}{x}, \frac{1}{y}, \frac{1}{z}\right)$. This will help to bring some of the terms into the numerator—almost always a good thing—and will also give us the equally nice constraint $xyz = 1$. As an example, the first term on the left-hand side would become

$$\frac{1}{a^3(b+c)} = \frac{1}{\frac{1}{x^3}(\frac{1}{y} + \frac{1}{z})} = \frac{x^3}{\frac{1}{y} + \frac{1}{z}} = \frac{x^3 yz}{y+z} = \frac{x^2}{y+z}.$$

Similar expressions hold for the other two terms.

So we now have the task of proving the inequality

$$\frac{x^2}{y+z} + \frac{y^2}{z+x} + \frac{z^2}{x+y} \geq \frac{3}{2},$$

for positive real numbers satisfying $xyz = 1$.

[4]The inequality is so named because the right-hand side is known as the *harmonic mean*.

This inequality not only looks much friendlier, but also seems to be a prime candidate for the Cauchy–Schwarz inequality. To obtain the left-hand side, it makes sense to let one of our sequences be

$$\left(\frac{x}{\sqrt{y+z}}, \frac{y}{\sqrt{z+x}}, \frac{z}{\sqrt{x+y}}\right).$$

When using Cauchy–Schwarz, you should always look for nice cancellation, and this is most easily achieved if we let the other sequence be

$$\left(\sqrt{y+z}, \sqrt{z+x}, \sqrt{x+y}\right).$$

Using these two sequences, the Cauchy–Schwarz inequality tells us that

$$\left(\frac{x^2}{y+z} + \frac{y^2}{z+x} + \frac{z^2}{x+y}\right)(2x+2y+2z) \geq (x+y+z)^2,$$

which in turn implies that

$$\frac{x^2}{y+z} + \frac{y^2}{z+x} + \frac{z^2}{x+y} \geq \frac{x+y+z}{2}.$$

All that remains is to prove that $x+y+z \geq 3$. But this is an immediate consequence of using the AM–GM inequality with the numbers x, y, z, along with the fact that $xyz = 1$. □

11.5 Power means inequality

Power means inequality For positive real numbers $a_1, a_2, \ldots, a_n$, define

$$M_0 = \sqrt[n]{a_1 a_2 \cdots a_n} \quad \text{and} \quad M_r = \left(\frac{a_1^r + a_2^r + \cdots + a_n^r}{n}\right)^{\frac{1}{r}}$$

for every non-zero real number r.

If $r > s$, then $M_r \geq M_s$, and equality occurs if and only if $a_1 = a_2 = \cdots = a_n$.

We often call M_2 the *quadratic mean* (QM), M_1 the *arithmetic mean* (AM), M_0 the *geometric mean*[5] (GM), and M_{-1} the *harmonic mean* (HM). You can see just how powerful this result is from the fact that the AM–GM inequality and the AM–HM inequality are simply special cases.

Problem Suppose that x, y, z are positive real numbers which satisfy $xyz = 1$.

If $r > s > 0$, prove that

$$x^r + y^r + z^r \geq x^s + y^s + z^s.$$

Solution The inequality seems reminiscent of the power means inequality with $n = 3$, which can be rearranged to give the following.

$$x^r + y^r + z^r \geq (x^s + y^s + z^s)\left(\frac{x^s + y^s + z^s}{3}\right)^{\frac{r}{s}-1}$$

[5] Actually, M_0 looks like the oddball here. Even though we can't directly put $r = 0$ into the definition of M_r, it turns out that M_r converges to the geometric mean as $r \to 0$.

So all that remains to be proved is that

$$\left(\frac{x^s+y^s+z^s}{3}\right)^{\frac{r}{s}-1} \geq 1.$$

However, this follows from the fact that $\frac{r}{s}-1>0$, which is true by assumption, and the fact that

$$\frac{x^s+y^s+z^s}{3} \geq \sqrt[3]{x^s y^s z^s} = 1,$$

which is true by the AM–GM inequality. □

In fact, this is a particular case of the following more general result, which you should now be able to prove on your own. It's definitely a handy little inequality to have under your belt and will reappear before the end of the chapter.

A useful inequality Suppose that $x_1, x_2, \ldots, x_n$ are positive real numbers which satisfy $x_1 x_2 \cdots x_n = 1$. If $r > s > 0$, then

$$x_1^r + x_2^r + \cdots + x_n^r \geq x_1^s + x_2^s + \cdots + x_n^s.$$

11.6 Jensen's inequality

Any time you have an expression in your inequality which can be written as

$$f(x_1) + f(x_2) + \cdots + f(x_n),$$

it's worth checking whether f is *convex* (like a smile) or *concave* (like a frown).[6] More often than not, Jensen's inequality can be used in such cases.

Jensen's inequality If the real numbers $x_1, x_2, \ldots, x_n$ lie on an interval where the function f is convex, then

$$\frac{f(x_1)+f(x_2)+\cdots+f(x_n)}{n} \geq f\left(\frac{x_1+x_2+\cdots+x_n}{n}\right).$$

If they lie on an interval where the function f is concave, then the inequality is reversed.

Problem Let $x_1, x_2, \ldots, x_n$ be positive real numbers whose sum is 1. What is the minimum value of the following expression?

$$\frac{x_1}{1+x_2+\cdots+x_n} + \frac{x_2}{1+x_1+x_3+\cdots+x_n} + \cdots + \frac{x_n}{1+x_1+\cdots+x_{n-1}}$$

Solution Note that this expression can be written more simply as

$$\frac{x_1}{2-x_1} + \frac{x_2}{2-x_2} + \cdots + \frac{x_n}{2-x_n}.$$

The form $f(x_1) + f(x_2) + \cdots + f(x_n)$ now appears, which suggests using Jensen's inequality with

$$f(x) = \frac{x}{2-x}.$$

[6] It is important to know that functions like $f(x) = x^2$ and $f(x) = e^x$ are convex while $f(x) = \log x$ is concave on the set of positive real numbers. For those who have learnt calculus, you can use the fact that a function f is convex when $f''(x) \geq 0$ and concave when $f''(x) \leq 0$.

Indeed, the function f is convex on the interval $[0,1]$ and since the numbers $x_1, x_2, \ldots, x_n$ all lie on this interval, we have

$$\begin{aligned}\frac{x_1}{2-x_1}+\frac{x_2}{2-x_2}+\cdots+\frac{x_n}{2-x_n} &= f(x_1)+f(x_2)+\cdots+f(x_n)\\ &\geq nf\left(\frac{x_1+x_2+\cdots+x_n}{n}\right)\\ &= nf\left(\frac{1}{n}\right)\\ &= \frac{n}{2n-1}.\end{aligned}$$

This shows that the expression is always greater than or equal to $\frac{n}{2n-1}$. To finish we need to demonstrate that this value can actually be attained. Simply taking $x_1 = x_2 = \cdots = x_n = \frac{1}{n}$ does the job. □

Problem If the numbers a, b, c are greater than 1, prove that

$$\frac{3}{1+\sqrt[3]{abc}} \leq \frac{1}{1+a}+\frac{1}{1+b}+\frac{1}{1+c}.$$

Solution You might think of trying Jensen's inequality with a function like $f(x) = \frac{1}{1+x}$. This helps with the RHS. By changing variables and modifying our function accordingly, we can deal with the LHS too.

Consider the function

$$f(x) = \frac{1}{1+e^x},$$

which is convex for x positive. Then Jensen's inequality allows us to deduce that

$$\frac{1}{1+e^{\frac{x+y+z}{3}}} \leq \frac{1}{3}\left(\frac{1}{1+e^x}+\frac{1}{1+e^y}+\frac{1}{1+e^z}\right).$$

If we now substitute $a = e^x$, $b = e^y$, $c = e^z$, which is allowed since a, b, c are greater than 1, we obtain the desired inequality. □

11.7 Substitutions

Substitutions can often help to simplify an inequality. In particular, you should always be on the lookout for useful substitutions, such as the following.

- Given $abc = 1$, consider the substitution $x = a^n$, $y = b^n$, $z = c^n$ for some positive or negative real number n. Other substitutions worth considering are $(a,b,c) = \left(\frac{x}{y}, \frac{y}{z}, \frac{z}{x}\right)$, which may break symmetry, or $(a,b,c) = \left(\frac{yz}{x^2}, \frac{xz}{y^2}, \frac{xy}{z^2}\right)$ which retains symmetry.
- If $a+b+c = 1$, you could try $(a,b,c) = \left(\frac{x}{x+y+z}, \frac{y}{x+y+z}, \frac{z}{x+y+z}\right)$.
- Consider a substitution which makes an ugly denominator look nicer, even if it makes the numerator look uglier.
- Consider a substitution which makes an ugly constraint look nicer, even if it makes the inequality look uglier.

- If a, b, c are the side lengths of a triangle, it's almost always useful to try the substitution $(a, b, c) = (y + z, z + x, x + y)$. In fact, a, b, c are the side lengths of a triangle if and only if $x = \frac{b+c-a}{2}$, $y = \frac{c+a-b}{2}$, $z = \frac{a+b-c}{2}$ are all positive.

  We call this the *incircle substitution* because the incircle of a triangle with side lengths a, b, c divides the sides into segments of lengths x, y, z. (See section 12.9.)

Problem If a, b, c are positive, show that

$$\frac{a}{b+2c} + \frac{b}{c+2a} + \frac{c}{a+2b} \geq 1.$$

Solution Whenever fractions are concerned, it's almost always better for ugliness to occur in the numerator rather than the denominator. In this particular inequality, we can introduce the following substitution to move the ugliness from the bottom of each fraction to the top.

$$\begin{aligned} x &= b + 2c \\ y &= c + 2a \\ z &= a + 2b \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} 9a &= 4y + z - 2x \\ 9b &= 4z + x - 2y \\ 9c &= 4x + y - 2z \end{aligned}$$

After performing these substitutions and tidying up (something which you should try on your own with pen and paper), the inequality looks a lot more palatable.

$$4\left(\frac{y}{x} + \frac{z}{y} + \frac{x}{z}\right) + \left(\frac{x}{y} + \frac{y}{z} + \frac{z}{x}\right) \geq 15$$

And it's not too difficult to see that this inequality is true, since each bracket is greater than or equal to 3 by the AM–GM inequality. □

11.8 Addition and multiplication of inequalities

Here is a neat way to create an interesting and difficult inequality from a boring and simple one. As an example, we'll start with the well-known inequality

$$x + y \geq 2\sqrt{xy}.$$

Now, pick a few variables such as a, b, c and write down all possible combinations of the inequality using these.

$$a + b \geq 2\sqrt{ab} \qquad b + c \geq 2\sqrt{bc} \qquad c + a \geq 2\sqrt{ca}$$

Adding or multiplying these together produces the following two bigger and better inequalities.

$$\begin{aligned} a + b + c &\geq \sqrt{ab} + \sqrt{bc} + \sqrt{ca} \\ (a + b)(b + c)(c + a) &\geq 8abc \end{aligned}$$

Many interesting inequalities are created using this technique. One way to solve them is to reverse engineer them. In other words, your task is to discover the original chunks which were added or multiplied together to create the inequality. As you will see, this does not involve guesswork alone, but also intelligence, experience, and sometimes luck.

Problem Let a, b, c be the lengths of the sides of a triangle.

Prove that

$$\sqrt{a + b - c} + \sqrt{b + c - a} + \sqrt{c + a - b} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}.$$

Solution We begin by using the incircle substitution

$$a = y + z, \quad b = z + x, \quad c = x + y,$$

which was mentioned in section 11.7. This is equivalent to

$$x = \frac{b + c - a}{2}, \quad y = \frac{c + a - b}{2}, \quad z = \frac{a + b - c}{2}$$

and transforms the inequality into the form

$$\sqrt{2x} + \sqrt{2y} + \sqrt{2z} \leq \sqrt{y + z} + \sqrt{z + x} + \sqrt{x + y}.$$

A very naive and incorrect approach is to guess that the first term on the left is less than or equal to the first term on the right, so that $\sqrt{2x} \leq \sqrt{y + z}$. However, this simply cannot be true for all positive x, y, z. So we have to dig a little deeper to find the chunks which are hiding. To uncover them, let's write the left-hand side as

$$\frac{\sqrt{2y} + \sqrt{2z}}{2} + \frac{\sqrt{2z} + \sqrt{2x}}{2} + \frac{\sqrt{2x} + \sqrt{2y}}{2}.$$

And now it seems far more plausible that each term on the left is smaller than the corresponding term on the right. All that remains is to show that

$$\sqrt{2x} + \sqrt{2y} \leq 2\sqrt{x + y},$$

which is easily proved by squaring, simplifying, and resorting to the AM–GM inequality. Then adding this inequality with the two others obtained by permuting the variables yields the desired result. □

Problem Prove that, for all positive real numbers a, b, c,

$$\frac{abc}{a^3 + b^3 + abc} + \frac{abc}{b^3 + c^3 + abc} + \frac{abc}{c^3 + a^3 + abc} \leq 1.$$

Solution This time, the chunks are much harder to uncover, but they look like the following.

$$\frac{abc}{a^3 + b^3 + abc} \leq \frac{c}{a + b + c} \quad \Leftrightarrow \quad a^2bc + b^2ca + c^2ab \leq ca^3 + cb^3 + c^2ab.$$

After collecting all terms on the left-hand side and factorising, this latter inequality simply becomes

$$-c(a + b)(a - b)^2 \leq 0,$$

which is obviously true. Now the solution to the original problem follows immediately, since we can simply sum the three corresponding inequalities.

$$\frac{abc}{a^3 + b^3 + abc} + \frac{abc}{b^3 + c^3 + abc} + \frac{abc}{c^3 + a^3 + abc} \leq \frac{a + b + c}{a + b + c} = 1$$

The trick in this solution was to break the number 1 into three chunks, each one depending on a, b and c. □

11.9 Expand and conquer

Most of the inequalities that we have seen are *symmetric*, which means that they look exactly the same after permuting the variables. Note that if you are expanding a symmetric expression in three variables and you see the term a^2b, then you would certainly expect to see also the terms $a^2c, b^2a, b^2c, c^2a, c^2b$.

Inequalities can also be *cyclic*, which means that they look the same after cycling the variables. Note that if you are expanding a cyclic expression in three variables and you see the term a^2b, then you would certainly expect to see also the terms b^2c and c^2a, but not necessarily b^2a, c^2b and a^2c.

We can take advantage of these facts when expanding algebraic expressions by using *symmetric sum notation* and *cyclic sum notation*. The following examples should give you some idea of how it works for three variables, although you can use this notation for more variables.

$$\begin{aligned}\sum_{\text{sym}} a^2b &= a^2b^1c^0 + a^2c^1b^0 + b^2c^1a^0 + b^2a^1c^0 + c^2a^1b^0 + c^2b^1a^0\\ &= a^2b + a^2c + b^2c + b^2a + c^2a + c^2b\end{aligned}$$

$$\begin{aligned}\sum_{\text{cyc}} a^2b &= a^2b^1c^0 + b^2c^1a^0 + c^2a^1b^0\\ &= a^2b + b^2c + c^2a\end{aligned}$$

$$\begin{aligned}\sum_{\text{sym}} a^2 &= a^2b^0c^0 + a^2c^0b^0 + b^2c^0a^0 + b^2a^0c^0 + c^2a^0b^0 + c^2b^0a^0\\ &= 2a^2 + 2b^2 + 2c^2\end{aligned}$$

$$\begin{aligned}\sum_{\text{cyc}} a^2 &= a^2b^0c^0 + b^2c^0a^0 + c^2a^0b^0\\ &= a^2 + b^2 + c^2\end{aligned}$$

Problem For positive real numbers a, b, c, prove that

$$\frac{a+b-2c}{b+c} + \frac{b+c-2a}{c+a} + \frac{c+a-2b}{a+b} \geq 0.$$

Solution It might seem like a maniacal thing to do, but we're going to multiply both sides of the inequality by $(a+b)(b+c)(c+a)$. Since the inequality is cyclic in nature, it would be foolish for us not to use cyclic sum notation. We want to prove

$$\sum_{\text{cyc}} (a+b-2c)(c+a)(a+b) \geq 0,$$

which is equivalent to

$$\sum_{\text{cyc}} a^3 + 2a^2b + b^2c + b^2a - 2c^2a - 2c^2b - a^2c \geq 0.$$

This might seem even messier than when we began, but remember that in cyclic sum notation, we have equations like

$$\sum_{\text{cyc}} a^2b = \sum_{\text{cyc}} b^2c = \sum_{\text{cyc}} c^2a = a^2b + b^2c + c^2a.$$

So, this observation allows us to cancel a lot of the terms until we are left with something which almost looks nice.

$$\sum_{\text{cyc}} a^3 + a^2b - 2a^2c \geq 0$$

It would be fantastic if we could prove that the sum $a^3 + a^2b - 2a^2c$ was non-negative, but this fact just isn't true for arbitrary positive real numbers a, b, c.

However, remember that in cyclic sum notation, the term a^2b is the same as the term c^2a. Plugging this into the inequality and factorising gives

$$\sum_{\text{cyc}} a^3 + c^2a - 2a^2c = \sum_{\text{cyc}} a(a-c)^2 \geq 0,$$

which is now obviously true, using the fact that squares are non-negative. □

11.10 Homogeneous inequalities

Many inequalities that we've seen are homogeneous, meaning that all terms have the same degree, in some sense. For example, the expression $x^3+x^2y-y^3$ is homogeneous of degree three, while the expression $x^3 + y$ is not. More precisely, we say that an inequality is *homogeneous* of degree n if multiplying all of the variables by any positive constant λ has the effect of multiplying the entire inequality by λ^n. Note this effectively leaves the inequality to be proven unchanged because dividing by λ^n returns the original inequality. With homogeneous inequalities, we are free to set the sum or product of the variables to be a particular number and just prove the inequality with that particular constraint. Hopefully, you will soon see what we mean.

Problem If a, b, c are positive, prove that

$$\left(1+\frac{a}{b}\right)\left(1+\frac{b}{c}\right)\left(1+\frac{c}{a}\right) \geq 2\left(1+\frac{a+b+c}{\sqrt[3]{abc}}\right).$$

Solution First, we note that the inequality is homogeneous, since plugging in the values ra, rb and rc gives

$$\left(1+\frac{ra}{rb}\right)\left(1+\frac{rb}{rc}\right)\left(1+\frac{rc}{ra}\right) \geq 2\left(1+\frac{ra+rb+rc}{\sqrt[3]{rarbrc}}\right),$$

which is identical to the original inequality after some cancellation. This means that if we can prove the inequality for (a, b, c), then it must hold true for (ra, rb, rc) for any positive real number r. In particular, if we can prove the inequality in the case $a + b + c = 1$, then it follows that the inequality is true for any positive value of $a + b + c$. Similarly, if we can prove the inequality for $abc = 1$, then it follows that the inequality is true for any positive value of abc.

In this particular case, fixing $abc = 1$ gets rid of the unsightly cube root for us and the inequality takes the following form

$$\left(1+\frac{a}{b}\right)\left(1+\frac{b}{c}\right)\left(1+\frac{c}{a}\right) \geq 2(1+a+b+c).$$

Expanding and simplifying leaves us with

$$a^2c + b^2c + b^2a + c^2a + c^2b + a^2b \geq 2(a+b+c).$$

The rest of the inequality can be handled by cleverly pairing the six terms on the left-hand side as shown and applying the AM–GM inequality to each pair.

$$\begin{aligned}(a^2b + a^2c) + (b^2c + b^2a) + (c^2a + c^2b) &\geq 2\sqrt{a^4bc} + 2\sqrt{b^4ca} + 2\sqrt{c^4ab} \\ &= 2(a^{\frac{3}{2}} + b^{\frac{3}{2}} + c^{\frac{3}{2}}) \\ &\geq 2(a+b+c)\end{aligned}$$

The very last step here is a simple application of the useful inequality stated in section 11.5. □

11.11 Muirhead's inequality

Muirhead's inequality, also known as *majorisation*, is a hefty weapon. Many symmetric inequalities can be smashed with this technique along with brute force calculations.

Muirhead's inequality Let $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_n)$ be two sequences of real numbers which satisfy the following inequalities.

$$\begin{gathered} a_1 \geq a_2 \geq \cdots \geq a_n \\ b_1 \geq b_2 \geq \cdots \geq b_n \end{gathered}$$

$$\begin{gathered} a_1 \geq b_1 \\ a_1 + a_2 \geq b_1 + b_2 \\ a_1 + a_2 + a_3 \geq b_1 + b_2 + b_3 \\ \vdots \\ a_1 + a_2 + \cdots + a_{n-1} \geq b_1 + b_2 + \cdots + b_{n-1} \\ a_1 + a_2 + \cdots + a_{n-1} + a_n = b_1 + b_2 + \cdots + b_{n-1} + b_n \end{gathered}$$

(Note that the last line really is an equality.)

The sequence A is said to *majorise* the sequence B and this is expressed by writing $A \succ B$.

If $A \succ B$ and $x_1, x_2, \ldots, x_n$ are non-negative real numbers, then

$$\sum_{\text{sym}} x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n} \geq \sum_{\text{sym}} x_1^{b_1} x_2^{b_2} \cdots x_n^{b_n}.$$

Problem Prove that if a, b, c are positive real numbers with product 1, then

$$a^4b + ab^4 + a^4c + ac^4 + b^4c + bc^4 \geq 2ab + 2bc + 2ac.$$

Solution Note first that since $abc = 1$ we may replace ab with a^2b^2c, bc with ab^2c^2 and ac with a^2bc^2 thus making the inequality homogeneous. Then if we write it using symmetric sum notation, the inequality takes the form

$$\sum_{\text{sym}} a^4b^1c^0 \geq \sum_{\text{sym}} a^2b^2c.$$

Since the triple $(4, 1, 0)$ majorises the triple $(2, 2, 1)$, we may invoke Muirhead's inequality to finish off the problem. □

Problem Let x, y and z be positive real numbers such that $xyz = 1$.

Prove that

$$\frac{x^5 - x^2}{x^5 + y^2 + z^2} + \frac{y^5 - y^2}{y^5 + z^2 + x^2} + \frac{z^5 - z^2}{z^5 + x^2 + y^2} \geq 0.$$

Solution We suppress the brute force part here! Suffice to say that after homogenising, clearing denominators and rearranging, the inequality takes the form

$$\sum_{\text{sym}} x^{10}yz + 4x^7y^5 + x^6y^3z^3 \geq \sum_{\text{sym}} x^8y^2z^2 + 2x^6y^5z + 2x^6y^4z^2 + x^5y^5z^2.$$

It would be nice if we could match up sequences on the LHS to majorise sequences on the RHS, but unfortunately this cannot be done. The sequence $(10,1,1)$ on the LHS majorises all of the sequences associated with the RHS. The sequence $(7,5,0)$ on the LHS majorises all the sequences on the RHS except for $(8,2,2)$. But the sequence $(6,3,3)$ on the LHS does not majorise anything on the RHS. The solution to our conundrum is to use the AM–GM on the LHS so as to smooth out the strong $(10,1,1)$ term with the weak $(6,3,3)$ term. This yields

$$x^{10}yz + x^6y^3z^3 \geq 2x^8y^2z^2.$$

It is now sufficient to prove

$$\sum_{\text{sym}} x^8y^2z^2 + 4x^7y^5 \geq \sum_{\text{sym}} 2x^6y^5z + 2x^6y^4z^2 + x^5y^5z^2.$$

We are now done because $(8,2,2)$ majorises $(5,5,2)$ while $(7,5,0)$ majorises $(6,5,1)$ and $(7,5,2)$. □

The expansion and collection of like terms which was suppressed in the above solution would require an enormous amount of accuracy, time and perseverance if performed without using symmetric sum notation. Even with this notation it is somewhat unwieldy. A useful notation that can be used is to write $[i,j,k]$ to mean $\sum_{\text{sym}} x^iy^jz^k$. Of course one must understand how to manipulate the notation. But it's not that difficult. Perhaps the hardest part is to deduce that

$$\begin{aligned}[i,j,k][p,q,r] = {} & [i+p,j+q,k+r] + [i+p,j+r,k+q] \\ & + [i+q,j+p,k+r] + [i+q,j+r,k+p] \\ & + [i+r,j+p,k+q] + [i+r,j+q,k+p],\end{aligned}$$

which represents the expansion of $\left(\sum_{\text{sym}} x^iy^jz^k\right)\left(\sum_{\text{sym}} x^py^qz^r\right)$.

Note that most expansions are simpler, such as

$$[i,j,k][p,0,0] = 2[i+p,j,k] + 2[i,j+p,k] + 2[i,j,k+p].$$

11.12 Weighted inequalities

Some inequalities have more general versions that involve weights. These include not just the AM–GM–HM, but the entire power means inequality. Furthermore, the already very general Jensen's inequality also generalises with weights.

Weighted power means inequality For positive real numbers $a_1, a_2, \ldots, a_n$, and positive real numbers $w_1, w_2, \ldots, w_n$ (called weights), define

$$M_0 = (a_1^{w_1}a_2^{w_2}\cdots a_n^{w_n})^{\frac{1}{w_1+w_2+\cdots+w_n}} \quad \text{and} \quad M_r = \left(\frac{w_1a_1^r + w_2a_2^r + \cdots + w_na_n^r}{w_1+w_2+\cdots+w_n}\right)^{\frac{1}{r}}$$

for every non-zero real number r.

If $r > s$, then $M_r \geq M_s$, and equality occurs if and only if $a_1 = a_2 = \cdots = a_n$.

Weighted Jensen's inequality If the real numbers $x_1, x_2, \ldots, x_n$ lie on an interval where the function f is convex, and the weights $w_1, w_2, \ldots, w_n$ are any positive real numbers, then

$$\frac{w_1 f(x_1) + w_2 f(x_2) + \cdots + w_n f(x_n)}{w_1 + w_2 + \cdots + w_n} \geq f\left(\frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}\right).$$

If they lie on an interval where the function f is concave, then the inequality is reversed.

Note that in both weighted inequalities, choosing $w_1 = w_2 = \cdots = w_n = 1$ recovers the ordinary versions of the inequalities.

Choosing the weights carefully can lead to *very* powerful results.

Problem Prove the Cauchy–Schwarz inequality from the weighted AM–HM inequality, where the variables in the Cauchy–Schwarz inequality are positive real numbers.

Solution We prove only the version for two sets of three variables but the proof generalises quite easily for two sets of n variables. The Cauchy–Schwarz inequality for two sets of three variables is

$$(a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) \geq (a_1 b_1 + a_2 b_2 + a_3 b_3)^2.$$

The weighted AM–HM inequality for three variables x_1, x_2, x_3, and corresponding weights w_1, w_2, w_3, is

$$\frac{w_1 x_1 + w_2 x_2 + w_3 x_3}{w_1 + w_2 + w_3} \geq \frac{w_1 + w_2 + w_3}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \frac{w_3}{x_3}},$$

which rearranges as

$$(w_1 x_1 + w_2 x_2 + w_3 x_3)\left(\frac{w_1}{x_1} + \frac{w_2}{x_2} + \frac{w_3}{x_3}\right) \geq (w_1 + w_2 + w_3)^2.$$

This looks very similar to the Cauchy–Schwarz inequality. In fact it is identical if we put

$$w_i = a_i b_i \quad \text{and} \quad x_i = \frac{a_i}{b_i}.$$ □

Note that the change of variables at the end of the solution is reversible via

$$a_i = \sqrt{w_i x_i} \quad \text{and} \quad b_i = \frac{\sqrt{w_i}}{\sqrt{x_i}}.$$

This demonstrates that the weighted AM–HM inequality and the Cauchy–Schwarz inequality are in fact the same inequality but stated in different variables!

Using weighted means it is possible to prove the more general Hölder's inequality.

Hölder's inequality Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be two sets of non-negative real numbers and let p and q be positive real numbers satisfying $\frac{1}{p} + \frac{1}{q} = 1$, then

$$(a_1^p + a_2^p + \cdots + a_n^p)^{\frac{1}{p}} (b_1^q + b_2^q + \cdots + b_n^q)^{\frac{1}{q}} \geq a_1 b_1 + a_2 b_2 + \cdots + a_n b_n.$$

In fact this generalises again to more than two sets of variables. The full version is a bit unwieldy and would rarely be used, so we just state the version for three sets of variables from which the generalisation should be apparent.

Hölder's inequality for three sets of variables Let $a_1, a_2, \ldots, a_n$, $b_1, b_2, \ldots, b_n$ and $c_1, c_2, \ldots, c_n$ be three sequences of n non-negative real numbers and let p, q and r be positive real numbers satisfying $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$, then

$$\left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} b_i^q\right)^{\frac{1}{q}} \left(\sum_{i=1}^{n} c_i^r\right)^{\frac{1}{r}} \geq \sum_{i=1}^{n} a_i b_i c_i.$$

Here is one more example of how a difficult inequality can be proved using weighted means.

Problem Prove that if a, b, c are positive real numbers, then

$$\frac{a}{\sqrt{a^2+8bc}} + \frac{b}{\sqrt{b^2+8ac}} + \frac{c}{\sqrt{c^2+8ab}} \geq 1.$$

Solution Apply Jensen's inequality to the convex function

$$f(x) = \frac{1}{\sqrt{x}}$$

and weights a, b, c in the following way.

$$\frac{af(a^2+8bc) + bf(b^2+8ac) + cf(c^2+8ab)}{a+b+c} \geq f\left(\frac{a(a^2+8bc) + b(b^2+8ac) + c(c^2+8ab)}{a+b+c}\right)$$

This may be rewritten as

$$\frac{a}{\sqrt{a^2+8bc}} + \frac{b}{\sqrt{b^2+8ac}} + \frac{c}{\sqrt{c^2+8ab}} \geq \frac{(a+b+c)^{\frac{3}{2}}}{\sqrt{a^3+b^3+c^3+24abc}}.$$

Thus it suffices to prove

$$(a+b+c)^3 \geq a^3+b^3+c^3+24abc.$$

However, this is very easy to prove after the LHS is expanded. We leave the proof of this last part to the reader. □

12

Geometric inequalities

Geometric inequalities, even simple looking ones, can be fiendishly difficult. Of course, it helps to have some mastery of geometry as well as some mastery of inequalities. We will look at a whole new bag of tricks for solving geometric inequalities in this chapter.

12.0 Problems

1. If $ABCD$ is a convex quadrilateral, which point P minimises the sum

$$PA + PB + PC + PD?$$

2. If the point P lies inside triangle ABC, prove that

$$AP + BP < AC + BC.$$

3. Let ABC be a triangle and let X, Y and Z be the midpoints of BC, CA and AB, respectively.

   Prove that

$$\frac{3}{4}(AB + BC + CA) < AX + BY + CZ < AB + BC + CA.$$

4. Two points A and B lie on different sides of a line ℓ.

   (a) Determine the point X on ℓ which minimises the difference between the lengths of AX and BX.

   (b) Determine the point X on ℓ which maximises the difference between the lengths of AX and BX.

5. Of all the triangles with a given base and a given area, which one has the smallest product of its side lengths?

6. An ant starts on the boundary of a circular disc with radius one metre and walks in a straight line. Every now and then, it turns left by $60°$ or right by $60°$, alternating each time. When the ant reaches the boundary of the disc again, it decides to stop for a rest.

   What is the maximum distance that the ant could have travelled?

7. Two planets A and B lie in space, near to a long, straight asteroid belt which can be considered as a straight line.

   What is the shortest path from A to B via the asteroid belt?

8. Let A and B be two points on a circle.

   (a) Determine the point M on the circle which maximises the value of $AM^2 + BM^2$.

   (b) Determine the point M on the circle which minimises the value of $AM^2 + BM^2$.

9. Consider a point M inside a given angle.

   Which line through M cuts off a triangle of minimal area from the angle?

10. Determine the point M inside acute triangle ABC which minimises the value of

$$AM \times BC + BM \times AC + CM \times AB.$$

11. Let triangle ABC have orthocentre H and circumradius R.

    Prove that

$$AH + BH + CH \leq 3R.$$

12. Let P, Q, R and S be points on the sides AB, BC, CD and DA of a parallelogram $ABCD$, respectively.

    Prove that

$$PQ + QR + RS + SP \geq 2AC,$$

    where AC is the shorter diagonal of the parallelogram.

13. Let ABC be a triangle. Let K, L and M be points on BC, AC and AB, respectively. Let X, Y and Z be points on LM, MK and KL, respectively. Let E_1, E_2, E_3, E_4, E_5, E_6 and E denote the areas of triangles AMY, CKY, BKZ, ALZ, BMX, CLX and ABC, respectively.

    Show that

$$E \geq 8\sqrt[6]{E_1E_2E_3E_4E_5E_6}.$$

14. A bee flies for four metres, ending where it began.

    Prove that its path can be enclosed in a sphere of radius one metre.

15. Let M and N be the midpoints of sides AD and BC of the convex quadrilateral $ABCD$.

    Prove that

$$2MN \leq AB + CD,$$

    with equality if and only if AB is parallel to CD.

16. Consider a quadrilateral of area A and with consecutive side lengths a, b, c and d.

    Prove that

$$ac + bd \geq 2A,$$

    with equality if and only if the quadrilateral is cyclic and its diagonals are perpendicular.

17. Consider the hexagon formed in a triangle by drawing the three tangents to the incircle which are parallel to the sides of the triangle.

    Prove that the perimeter of the hexagon is less than or equal to $\frac{2}{3}$ times the perimeter of the triangle.

18. In the plane we are given 5 distinct points A, B, C, P and Q, no three of which are collinear.

    Prove that

    $$AB + BC + CA + PQ < AP + AQ + BP + BQ + CP + CQ.$$

19. Let $ABCDEF$ be a convex hexagon satisfying

    $$AB = BC, \quad CD = DE \quad \text{and} \quad EF = FA.$$

    Prove that

    $$\frac{BC}{BE} + \frac{DE}{DA} + \frac{FA}{FC} \geq \frac{3}{2},$$

    and determine when equality occurs.

20. Let P be a point inside triangle ABC.

    Prove that one of the three angles $\angle PAB$, $\angle PBC$, $\angle PCA$ is at most 30°.

21. In an acute triangle ABC, let O be the circumcentre and P be the foot of the altitude from A. Suppose that

    $$\angle BCA \geq \angle ABC + 30^\circ.$$

    Prove that

    $$\angle CAB + \angle COP < 90^\circ.$$

22. Let $ABCDEF$ be a convex hexagon whose opposite sides are parallel. Let R_A, R_C and R_E be the radii of circles FAB, BCD and DEF, respectively.

    If P is the perimeter of the hexagon, prove that

    $$R_A + R_C + R_E \geq \frac{P}{2}.$$

12.1 Triangle inequality

Many geometric inequalities boil down to the well-known fact that the shortest path between two points is given by a straight line segment. For example, it will usually be shorter to travel in a straight line from A to C than to travel from A to B and then from B to C. This is precisely the content of the triangle inequality.

Triangle inequality If A, B and C are points, then

$$AC \leq AB + BC,$$

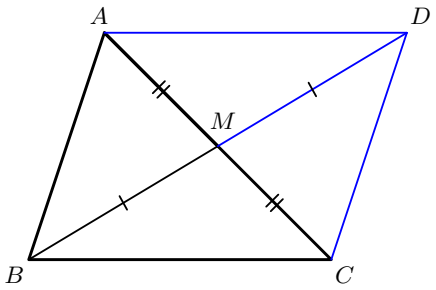
with equality if and only if B lies on the line segment AC.

Problem If M is the midpoint of the side AC in triangle[1] ABC, prove that

$$2BM < AB + BC.$$

Solution The form of this geometric inequality suggests that we should try to construct a triangle whose side lengths are AB, BC and $2BM$. The natural way to construct a line segment of length $2BM$ in our diagram is to extend BM to a point D, where $BM = DM$.

Alternatively, we can describe D as the unique point in the plane which makes $ABCD$ a parallelogram.



Applying the triangle inequality in triangle ABD implies that

$$\begin{aligned} & BD \leq AB + AD \\ \Rightarrow \quad & 2BM \leq AB + BC. \end{aligned}$$

Equality occurs if and only if A lies on the line segment BD, which is impossible since ABC is a triangle. □

12.2 Reflection principle

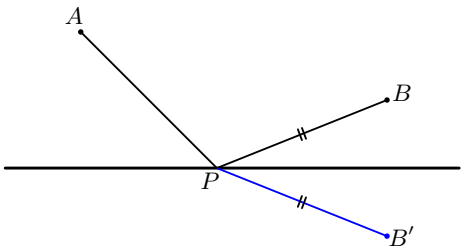
The following problem will give us our first taste of the reflection principle, a beautiful technique which can be used to solve many geometric inequalities. It's particularly useful for problems in which one has to minimise a sum of distinct lengths.

Problem Two towns A and B lie on the same side of a long, straight river.

What is the shortest path from A to B via the river?

[1] In most of our problems, including this one, *triangle* refers to a *non-degenerate triangle*, that is, one in which the three vertices are not collinear.

Solution It's clear that such a shortest path must consist of two straight line segments, AP and PB, where P is some point on the river. But where should we put the point P? The trick is to imagine a ghost town B', located at the reflection of town B through the river, as shown in the diagram.



Then for any point P on the river, the distance PB is exactly the same as the distance PB'. By the triangle inequality, we have

$$AP + PB = AP + PB' \geq AB',$$

with equality if and only if P lies on the line segment AB'.

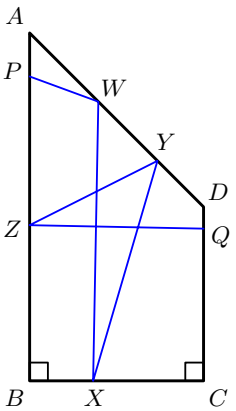
So we should put the point P at the intersection of the line segment AB' with the river. □

One thing you might notice from this solution is that, in order to obtain the shortest path, the line segments AP and PB should meet the river at equal angles. This is analogous to the physical principle which states that when a beam of light is reflected by a mirror, the angle of incidence equals the angle of reflection.

Let's now try to use the reflection principle to solve a more substantial problem.

Problem Consider the quadrilateral $ABCD$ with $AB = 16$, $BC = 8$, $CD = 8$ and $\angle ABC = \angle BCD = 90°$. Let P be the point on AB such that $AP = 2$ and let Q be the point on CD such that $DQ = 3$.

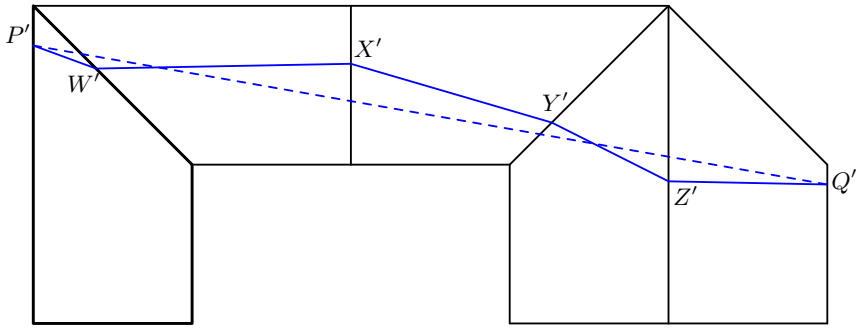
Find the length of the shortest path which begins at P, meets the side DA at a point W, then meets the side BC at a point X, then meets the side DA at a point Y, then meets the side AB at a point Z, and then finally ends at Q.



Solution It might look like a crazy problem, but you'll see just how easy it is with the help of the reflection principle. Perhaps surprisingly, we will reflect not once, not twice, not thrice, but four times!

- First take the quadrilateral and reflect it through its side AD.
- Then take the new quadrilateral and reflect it through its side BC.
- Then take the new quadrilateral and reflect it through its side AD.
- Finally, take the new quadrilateral and reflect it through its side AB.

The end result should be a diagram consisting of the original quadrilateral on the far left, and four reflected copies.



As you can see, any path

$$P \to W \to X \to Y \to Z \to Q$$

can be copied onto our new diagram to obtain a path

$$P' \to W' \to X' \to Y' \to Z' \to Q'.$$

Furthermore, since reflections keep lengths the same, we can be sure that $PW = P'W'$, $WX = W'X'$, $XY = X'Y'$, $YZ = Y'Z'$ and $ZQ = Z'Q'$. Of course, in order to minimise the length of the path from P' to Q' on our new diagram, we should simply take the straight line segment $P'Q'$. We can then copy this back on to our original quadrilateral to obtain the desired shortest path.

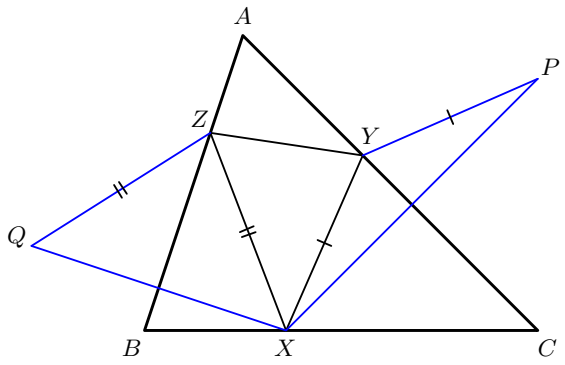
As for the path's length, it can be calculated using Pythagoras' theorem, since we know that Q' is 40 units across and 9 units down from P'. So the length of the shortest path with the desired properties is $\sqrt{40^2 + 9^2} = 41$. □

We'll use the reflection principle to full effect in the solution of the following interesting problem, first posed by Fagnano way back in the 18th century.

Problem Let ABC be an acute triangle and let X, Y and Z be points on the sides BC, CA and AB, respectively.

Where should X, Y and Z be located so that the perimeter of triangle XYZ is minimal?

Solution First, we'll minimise the perimeter of triangle XYZ, where X is some fixed point on BC. The trick is to reflect X in the sides AC and AB to obtain the points P and Q, respectively.



Then by the triangle inequality,

$$XY + YZ + ZX = PY + YZ + ZQ \geq PQ,$$

with equality if and only if Y lies on the intersection of AC and PQ, while Z lies on the intersection of AB and PQ. So for a fixed point X on BC, we can find Y and Z which minimises the perimeter of triangle XYZ. In fact, we know that the minimal perimeter in this case is precisely the length of PQ.

Now observe that, since P and Q were obtained by reflection we have

$$AP = AQ = AX.$$

Furthermore,

$$\angle PAQ = \angle PAX + \angle QAX = 2\angle CAX + 2\angle BAX = 2\angle BAC.$$

So, triangle APQ is always an isosceles triangle whose angle at A is equal to $2\angle BAC$.

Since $\angle BAC$ is fixed, the length PQ is minimised precisely when $AP = AQ = AX$ is minimised. This means that we should take X to be the foot of the altitude from A.

By a similar argument, Y should be the foot of the altitude from B, and Z should be the foot of the altitude from C. □

When stating Fagnano's problem, we were careful to mention that ABC is an acute triangle. What happens when it's right-angled or obtuse?

12.3 Transformations

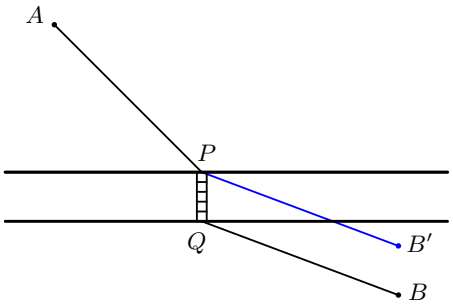
We've already seen that reflections are particularly useful for solving geometric inequalities, but they're certainly not the only transformations which come in handy. In general, transformations which preserve lengths, known as isometries, may also be useful. In fact, a common approach is to use an isometry to create a situation in which the triangle inequality can be employed.

Problem Two towns A and B lie on different sides of a long, straight river with parallel banks.

Where is the best place to build a bridge, perpendicular to the banks of the river, such that the path from A to B via the bridge is shortest?

Solution Suppose that the bridge touches the north bank at P and the south bank at Q, so that the width of the river is the distance PQ. Now consider the translation which takes B towards the river by the distance PQ. Then the path from A to B via the bridge has length

$$AP + PQ + QB = (AP + PB') + PQ.$$



By the triangle inequality, we know that

$$AP + PB' \geq AB',$$

with equality if and only if P lies on the line segment AB'.

Therefore, to minimise the path from A to B via the bridge, we should place the point P at the intersection of AB' with the north bank of the river. □

12.4 Trigonometry

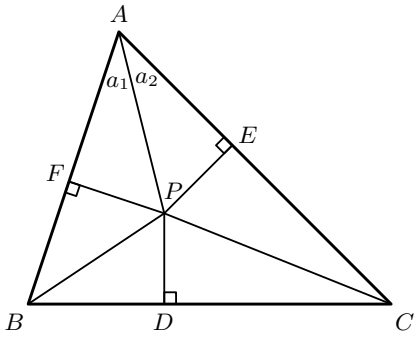
Trigonometry is often useful for solving geometric inequalities, because we can use various trigonometric identities as well as the fact that $\sin\theta \leq 1$ and $\cos\theta \leq 1$ for all angles θ. Here is a nice illustrative example.

Problem Let P be a point inside triangle ABC and let the feet of the perpendiculars from P to the sides BC, CA, AB be D, E, F, respectively.

Find the point P which maximises the value of

$$\frac{PD \cdot PE \cdot PF}{PA \cdot PB \cdot PC}.$$

Solution Consider the angles a_1 and a_2 labelled in the following diagram.



Using some trigonometric trickery, we have the following.

$$\begin{aligned}\frac{PF}{PA}\cdot\frac{PE}{PA} &= \sin a_1 \sin a_2 \\ &= \frac{1}{2}\cos(a_1 - a_2) - \frac{1}{2}\cos(a_1 + a_2) \quad \text{(product to sum)} \\ &\leq \frac{1}{2} - \frac{1}{2}\cos A \\ &= \sin^2\frac{A}{2} \qquad\qquad\qquad\qquad \text{(double angle formula)}\end{aligned}$$

Equality holds here if and only if $a_1 = a_2$ or, in other words, when P lies on the angle bisector from A. Writing down the other two analogous inequalities, multiplying them together, and taking the square root gives

$$\frac{PD\cdot PE\cdot PF}{PA\cdot PB\cdot PC} \leq \sin\frac{A}{2}\cdot\sin\frac{B}{2}\cdot\sin\frac{C}{2},$$

with equality if and only if P lies on all three angle bisectors.

So the maximum value is achieved when P is the incentre of the triangle. □

12.5 Parametrisation

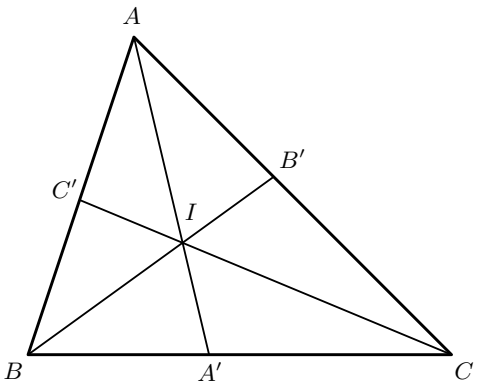
When solving a geometric inequality, it's sometimes possible to write the expression that you're trying to minimise or maximise in terms of certain parameters, such as angles or lengths. This can often help to turn a geometric inequality into an algebraic inequality.

Problem Let I be the incentre of triangle ABC and let the angle bisectors from A, B and C meet the opposite sides at A', B' and C', respectively.

Prove that

$$\frac{AI\cdot BI\cdot CI}{AA'\cdot BB'\cdot CC'} \leq \frac{8}{27}.$$

Solution The diagram is brimming with angle bisectors while the inequality is brimming with ratios, which suggests that we might be able to use the angle bisector theorem to our advantage.



A natural approach is to try to rewrite the expression $\frac{AI}{AA'}$ in terms of some nicer lengths. So let's aim to write it in terms of $a = BC$, $b = CA$ and $c = AB$, the side lengths of the triangle. We can't get the expression $\frac{AI}{AA'}$ directly from the angle bisector theorem. However, we can get the related expression $\frac{AI}{IA'}$ from using the angle bisector theorem in triangle ABA' or ACA'. The end result is

$$\frac{AI}{IA'} = \frac{AB}{BA'} = \frac{AC}{CA'}.$$

Now we use the simple fact[2] that if $\frac{x}{y} = \frac{z}{w}$, then $\frac{x+z}{y+w}$ is also equal to them.

$$\frac{AI}{IA'} = \frac{AB + AC}{BA' + CA'} = \frac{b+c}{a} \quad \Rightarrow \quad \frac{IA'}{AI} = \frac{a}{b+c}$$

Now we have all the information we need to determine $\frac{AI}{AA'}$.

$$\frac{AA'}{AI} = \frac{IA' + AI}{AI} = \frac{IA'}{AI} + 1 = \frac{a+b+c}{b+c} \quad \Rightarrow \quad \frac{AI}{AA'} = \frac{b+c}{a+b+c}$$

This concludes the parametrisation of the geometric inequality in terms of the side lengths of the triangle. We have now left geometry behind and all that remains is to prove the algebraic inequality

$$\left(\frac{b+c}{a+b+c}\right)\left(\frac{c+a}{a+b+c}\right)\left(\frac{a+b}{a+b+c}\right) \leq \frac{8}{27},$$

where the only restriction on a, b, c is that they are positive real numbers which obey the triangle inequality. Applying the AM–GM in the obvious way, we obtain

$$\begin{aligned}\left(\frac{b+c}{a+b+c}\right)\left(\frac{c+a}{a+b+c}\right)\left(\frac{a+b}{a+b+c}\right) &\leq \left(\frac{\frac{b+c}{a+b+c} + \frac{c+a}{a+b+c} + \frac{a+b}{a+b+c}}{3}\right)^3 \\ &= \frac{8}{27}.\end{aligned}$$

□

The inequality is still true if the point I is allowed to be *any* point inside the triangle. However, the proof of this requires a different, but not difficult, approach. See if you can prove it!

12.6 Ptolemy's inequality

All of the geometric inequalities that we've solved so far involve fairly well known facts from geometry or inequalities, along with some tricky techniques. The following theorem is less well known but we'll need it in our toolkit.

Ptolemy's inequality For any four points A, B, C and D, we have

$$AB \cdot CD + BC \cdot DA \geq AC \cdot BD.$$

If the four points do not lie on a line, then equality occurs if and only if $ABCD$ is cyclic with the points A, B, C and D lying in that order around the circle.

Unfortunately, Ptolemy's inequality is of no use to us if we don't know when and where to apply it. So let's take a look at an example.

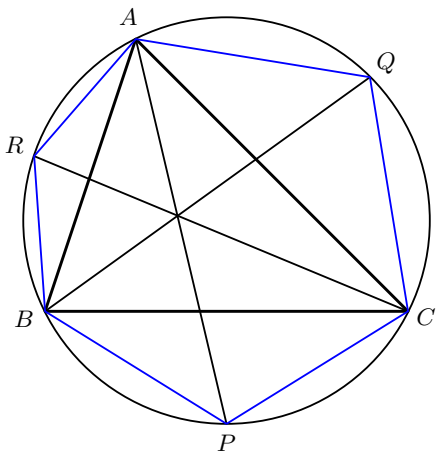
[2] This simple fact is called *addendo*.

Problem In triangle ABC, the angle bisector from A meets the circumcircle again at P. Similarly, the angle bisector from B meets the circumcircle again at Q and the angle bisector from C meets the circumcircle again at R.

Prove that

$$AB + BC + CA < AP + BQ + CR.$$

Solution



Observe that P bisects the arc between B and C on the circumcircle since AP is the angle bisector from A. In particular, this means that we have

$$\angle PBC = \angle PCB = \frac{\angle A}{2} \quad \Rightarrow \quad PB = PC.$$

Now points A, B, P and C are cyclic in that order, so we may use the equality form of Ptolemy's inequality to find

$$AB \cdot PC + PB \cdot AC = AP \cdot BC.$$

Since $PB = PC$ we may divide both sides by $PB \cdot AP$ to find

$$\frac{AB + AC}{AP} = \frac{BC}{PB} = 2\cos\frac{\angle A}{2},$$

where the last equality is seen by dropping a perpendicular from P to BC.

Since $\cos\frac{\angle A}{2} < 1$ we have

$$AB + AC < 2AP.$$

Now add this to the two analogous inequalities

$$BC + BA < 2BQ \quad \text{and} \quad CA + CB < 2CR,$$

and divide by 2 to obtain the desired result. □

It seems that Ptolemy's inequality is a good idea when there are isosceles triangles around. In the following problem, which was posed by Fermat, we'll see that it's an even better idea when there are equilateral triangles.

Problem Let ABC be a triangle with all angles less than $120°$.

Find a point P such that

$$PA + PB + PC$$

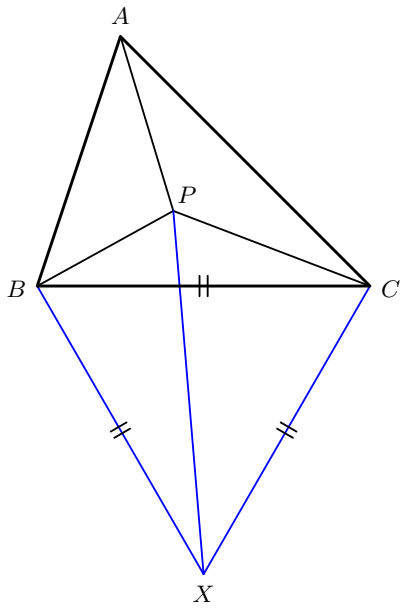
is minimal.[3]

Solution The first thing we're going to do is find a bound for $PB + PC$. And the way we're going to do that is to use Ptolemy's inequality with an equilateral triangle. But where is the equilateral triangle, you might be wondering? Well, let's just construct the point X so that triangle BCX is equilateral and external to our original triangle. Now using Ptolemy's inequality in $BPCX$ gives

$$BP \cdot CX + PC \cdot XB \geq BC \cdot PX.$$

Cancelling out the terms $CX = XB = BC$ gives us the inequality

$$PB + PC \geq PX.$$



So we've learnt that using an equilateral triangle gives us nice cancellation in Ptolemy's inequality. We now have

$$PA + PB + PC \geq PA + PX$$

and you just can't help but use the triangle inequality on this last expression. That gives

$$PA + PB + PC \geq AX.$$

But can we actually achieve equality? Yes we can! For equality to occur in Ptolemy's inequality, we need the point P to lie on the circumcircle of triangle BCX on the minor arc BC. For equality to occur in the triangle inequality, we need P to lie on the line segment AX. So P should be the intersection of the circumcircle of triangle BCX and the line segment AX. In fact, the minimal value of $PA + PB + PC$ will simply be the length of AX. □

[3] A slightly weaker version of this problem was solved in section 6.8.

Note that the conditions on P imply that $\angle BPC = 120°$. But there was nothing special about the side BC. We could just as well have constructed our equilateral triangle on CA or AB in which case we would have arrived at the result $CPA = 120°$ or $APB = 120°$.

So it turns out that P is the unique point in triangle ABC such that

$$\angle BPC = \angle CPA = \angle APB = 120°.$$

This is known as the *Fermat point*.

One other thing worth wondering about is why we required triangle ABC to have all angles less than 120°. Please analyse this for yourself.

12.7 Locus and tangency

Recall from chapter 6 that a locus is often thought of as a curve traced out by a point following some rule.

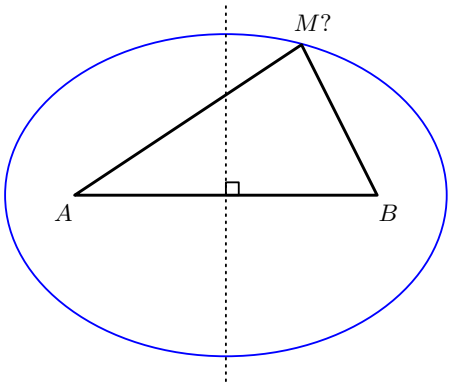
Tangency occurs when two curves just touch each other. What do locus and tangency have to do with geometric inequalities? Read on to find out!

Useful facts Let A and B be two points in the plane.

- The locus of points M such that $\angle AMB$ is constant is the union of two circular arcs symmetric about AB.
- The locus of points M such that $AM^2 + BM^2$ is constant is a circle whose centre is the midpoint of AB.
- The locus of points M such that $AM + BM$ is constant is an ellipse whose foci are A and B.

Problem Of all the triangles with a given base and a given perimeter, which one has the greatest area?

Solution



Suppose that the given base is AB and that the given perimeter is p. Consider the ellipse whose foci are A and B such that, for every point M on the ellipse,

$$AB + AM + BM = p.$$

This is all valid because p and AB are constants.

We would like to find the point M on the ellipse which maximises the area of triangle ABM. In other words, we would like to find the point M on the ellipse that is furthest away from the line AB. This occurs precisely when M lies on the perpendicular bisector of AB and triangle ABM is isosceles with base AB. □

In the previous problem, the notion of locus was crucial to the solution. In the next, we'll see how the notion of tangency also comes into play.

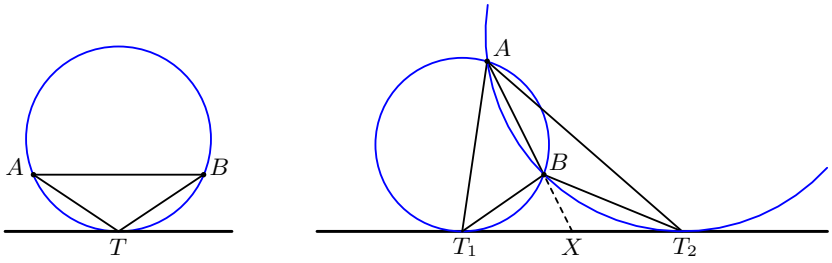
Problem If A and B are two points on the same side of a line ℓ, determine all points P on ℓ such that $\angle APB$ is maximised.

Solution First, we consider the case when AB is parallel to ℓ. Then there is a unique circle which passes through A and B and is tangent to ℓ at a point T.

Every point P on ℓ apart from T lies on the same side of AB and outside the circle. Therefore, for every point P on ℓ, we have the inequality

$$\angle APB \leq \angle ATB,$$

with equality if and only if $P = T$.



Now, consider the case when the line through AB meets ℓ at some point X. Then there are two circles which pass through A and B and are tangent to ℓ at points, which we'll call T_1 and T_2, one on each side of X.

Every point P on ℓ apart from T_1, but on the same side of X as T_1, lies on the same side of AB and outside the circle passing through T_1. Therefore, for every point P on ℓ on the same side of X as T_1, we have the inequality

$$\angle APB \leq AT_1B,$$

with equality if and only if $P = T_1$. Similarly, for every point P on ℓ on the same side of X as T_2, we have the inequality

$$\angle APB \leq AT_2B,$$

with equality if and only if $P = T_2$.

It is possible that $\angle AT_1B = \angle AT_2B$, in which case *both* points T_1 and T_2 are solutions to the problem. If r_1 and r_2 are the radii of circles ABT_1 and ABT_2, respectively, then by the extended sine rule we have

$$2r_1 = \frac{AB}{\sin \angle AT_1B} = \frac{AB}{\sin \angle AT_2B} = 2r_2.$$

This implies that the two circles have equal radius and so AB is perpendicular to ℓ. Conversely, if AB is perpendicular to ℓ, then by symmetry both T_1 and T_2 are solutions to the problem.

On the other hand, if we assume without loss of generality that $\angle AT_2B < AT_1B$, then for every point P on ℓ, we have the inequality $\angle APB < AT_1B$, with equality if and only if $P = T_1$. □

12.8 Isoperimetric inequalities

Given a rope, what is the maximum area that you can enclose? Perhaps your intuition tells you that the largest possible area is enclosed by a circular boundary, but how can you be sure? Your intuition is indeed correct, although trying to construct a rigorous proof is an extremely difficult task. We simply state this isoperimetric, literally 'same perimeter', inequality below without proof, along with a variant of it pertaining to polygons.

Isoperimetric inequalities

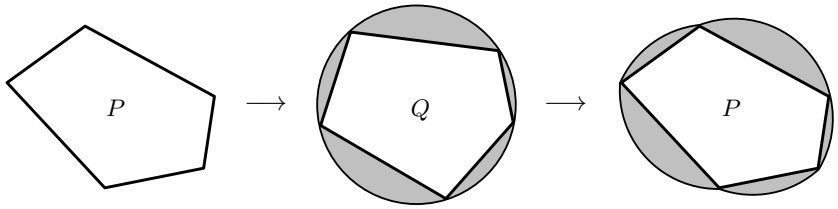
- Among all plane figures of a given perimeter, the circle has the maximum area. Equivalently, among all plane figures of a given area, the circle has the minimum perimeter.
- Among all n-gons of a given perimeter, the regular n-gon has the maximum area. Equivalently, among all n-gons of a given area, the regular n-gon has the minimum perimeter.

Problem Prove that among all polygons with given side lengths in a given order, the cyclic one has the maximal area.

Solution Implicit in the statement of this problem is the fact that if someone gives you a polygon, then you can always construct a cyclic polygon with the same side lengths in the same order, and this polygon is unique. We won't discuss a proof of this fact here, but you should certainly think about why it is true.

First, if the polygon were not convex, then turning a concave part of the polygon inside out would yield a polygon of greater area.[4] Thus we may assume that the polygon is convex.

Now suppose that P is a convex polygon with given side lengths in a given order which cannot be circumscribed by a circle. Let Q be the unique cyclic polygon with the same side lengths in the same order. The circumcircle of Q is naturally divided into the polygon itself as well as a number of segments. Now consider the shape obtained by gluing those segments onto the corresponding sides of P.



[4] This is *very* sloppy, because the new polygon might cross itself. See if you can iron out the details of this.

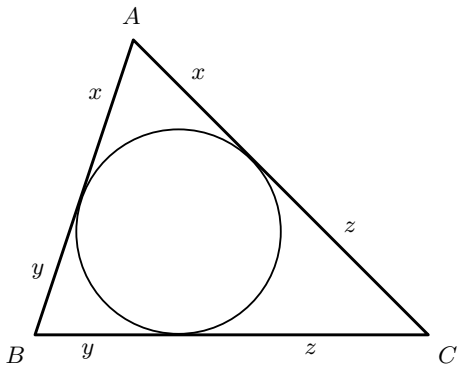
This shape[5] has the same perimeter as the circumcircle of Q. So, by the isoperimetric inequality, its area is smaller than the area of the circumcircle of Q. However, after removing the segments from each shape, we obtain the fact that the area of P is less than the area of Q. So we can conclude that, among all convex polygons with given side lengths in a given order, the cyclic one has the maximal area. □

12.9 Incircle substitution

Incircle substitution Suppose that triangle ABC has sides of length $BC = a$, $AC = b$ and $AB = c$. Let the incircle of triangle ABC be tangent to BC at X, AC at Y and AB at Z.
The *incircle substitution* is given by putting

$$a = y + z, \quad b = x + z, \quad \text{and} \quad c = x + y,$$

where $x = AY = AZ$, $y = BX = BZ$ and $z = CX = CY$.



Problem If a, b, c are the side lengths of a triangle, prove that

$$abc \geq (b + c - a)(c + a - b)(a + b - c).$$

Solution Using the incircle substitution, the inequality transforms to something which looks much nicer, that is,

$$(y + z)(z + x)(x + y) \geq 8xyz.$$

This follows from multiplying the three inequalities

$$y + z \geq 2\sqrt{yz}, \quad z + x \geq 2\sqrt{zx} \quad \text{and} \quad x + y \geq 2\sqrt{xy}$$

together. Each of these can be proved by using the AM–GM inequality or the fact that squares are non-negative. □

12.10 Triangle formulas

Sometimes knowing lots of triangle formulas can help turn a geometric inequality into an algebraic inequality.

[5] You should think about why the circular arcs around P do not intersect each other.

Problem Prove that $R \geq 2r$, where R and r are respectively the circumradius and inradius of a triangle.

Solution Perhaps surprisingly, one solution to this problem arises from considering the area A of the triangle. In particular, we have the three formulas

$$A = \frac{abc}{4R}, \quad A = rs, \quad \text{and} \quad A = \sqrt{s(s-a)(s-b)(s-c)}.$$

Here, a, b and c are the side lengths of the triangle and $s = \frac{a+b+c}{2}$ denotes the semiperimeter. If you've never seen these before, then now is definitely the time to try to prove them. The third one is particularly interesting and is often referred to as Heron's formula. From these, we can deduce the equation

$$\frac{abcrs}{4R} = s(s-a)(s-b)(s-c),$$

from which

$$\frac{r}{4R} = \frac{(s-a)(s-b)(s-c)}{abc}.$$

However, the result of the previous problem implies that

$$8abc \geq (s-a)(s-b)(s-c).$$

Thus, it follows that

$$\frac{r}{4R} \leq \frac{1}{8},$$

which simplifies to give $R \geq 2r$. □

This result is often known as Euler's inequality because it follows immediately from the following interesting result.

Euler's theorem in geometry The distance d between the circumcentre and the incentre of a triangle is given by

$$d = \sqrt{R^2 - 2Rr}.$$

You might like to prove this result by yourself, thereby providing an alternative proof of Euler's inequality. As a treat, we'll finish with the most elegant proof of this chapter—another proof of Euler's inequality.

Solution We start with a simple observation: the smallest circle which touches or crosses all three sides of a triangle is the incircle.

Recall that the midpoints of the sides of a triangle form what is known as the medial triangle. It is always similar to the original triangle and exactly half the size. Therefore, the circumcircle of the medial triangle[6] has radius $\frac{R}{2}$.

Observe that the circumcircle of the medial triangle passes through the midpoints of all three sides of the original triangle. So, from our earlier simple observation, the circumcircle of the medial triangle is at least as large as the incircle.[7]

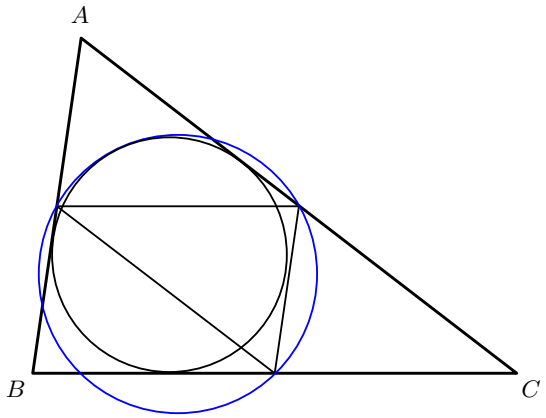
Hence, we may conclude that

$$\frac{R}{2} \geq r,$$

which can be equivalently written as $R \geq 2r$. □

[6] This is also the nine-point circle.

[7] In fact, as a carefully drawn diagram suggests, the incircle is always internally tangent to the nine-point circle. But this is more difficult to prove.


A
B
C

13

Combinatorics

In a very loose sense, combinatorics is the area of mathematics concerned with counting. By this, we don't mean calling out the numbers $1, 2, 3, \ldots$ or anything like that. Combinatorics is rather about counting in a clever way, or counting without counting. Actually, this isn't a particularly good description of combinatorics either. Pretty much anything that doesn't quite fit into another mathematical category tends to be described as combinatorial. This makes combinatorics a treasure trove of mathematical delights!

13.0 Problems

1. How many whole numbers from 1 to 2003 are not divisible by 2, 7 or 11?

2. A number is said to be *ordered* if each digit is greater than or equal to the digit on its left.

   How many three-digit numbers are ordered?

3. How many eight-digit phone numbers begin with the number 9, are even, and have the third, fourth and fifth digits in decreasing order?

4. How many odd numbers are there in the 2003rd row of Pascal's triangle?

5. The number 3 can be expressed as a sum of one or more positive integers (taking order into account) in four ways:

   $$3 = 1 + 2 = 2 + 1 = 1 + 1 + 1.$$

   How many ways can a positive integer n be so expressed?

6. In how many ways can a person obtain a score of 6 in a test containing 5 questions, each one marked out of 7?

7. How many ways are there to place two kings on a chessboard so that no two attack each other?

8. A spider has one sock and one shoe for each of its eight legs.

   In how many different orders can the spider put on its socks and shoes, assuming that on each leg the sock must be put on before the shoe?

9. Prove that
$$\sum_{k=m}^{n} \binom{n}{k}\binom{k}{m} = 2^{n-m}\binom{n}{m}.$$

10. Prove in at least two different ways that
$$\binom{n}{0}^2 + \binom{n}{1}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}.$$

11. Prove in at least two different ways that if m, n are positive integers and $1 \le k \le n$, then
$$\binom{n}{0}\binom{m}{k} + \binom{n}{1}\binom{m}{k-1} + \cdots + \binom{n}{k}\binom{m}{0} = \binom{m+n}{k}.$$
(This is a generalisation of the previous problem.)

12. Find the number of k-tuples $(S_1, S_2, \ldots, S_k)$ satisfying
$$S_1 \subseteq S_2 \subseteq \cdots \subseteq S_k \subseteq \{1, 2, \ldots, n\}.$$

13. Let there be given nine lattice points in 3-dimensional space.
Show that there is a lattice point on the interior of one of the line segments joining two of these points.

14. In a circus, there are n clowns who dress and paint themselves up using a selection of 12 distinct colours. Each clown is required to use at least five different colours. One day the ringmaster of the circus demands that no two clowns have exactly the same set of colours, and no more than 20 clowns may use any one particular colour.
Find the largest number n so that the ringmaster's demand can be met.

15. In a maths competition, $2n$ students take part. Each student submits a different problem and all $2n$ problems are collected, shuffled, and then handed back to each participant. The distribution is considered *fair* if there are n participants receiving the problems proposed by the other n participants.
Prove that the number of ways in which the problems can be distributed in a fair way is a perfect square.

16. An $n \times n$ matrix which has entries coming from the set $S = \{1, 2, \ldots, 2n-1\}$ is called a *silver* matrix if, for each $i = 1, 2, \ldots, n$, the ith row and the ith column together contain all the elements of S.
Prove the following.
    (a) There is no silver matrix for $n = 1997$.
    (b) Silver matrices exist for infinitely many values of n.

17. An 8×8 square is divided into 64 unit squares, and must be covered by 64 black and 64 white isosceles right-angled triangles whose side lengths are 1, 1 and $\sqrt{2}$, respectively. Each unit square must be covered by 2 triangles. A covering is said to be *awesome* if any two triangles sharing a common side are of distinct colours.
How many different awesome coverings are there?

18. A chess master who has 11 weeks to prepare for a tournament decides to play at least one game every day, but in order to conserve his energy he decides not to play more than 12 games in any seven-day period.

    Show there are consecutive days during which he plays *exactly* 21 games.

19. Let $S = \{0000000, 0000001, \ldots, 1111111\}$ be the set of all binary sequences of length 7. The *distance* between two elements $s_1, s_2 \in S$ is the number of places in which s_1 and s_2 differ.

    Show that if $T \subset S$ and $|T| > 16$, then T contains two elements whose distance is at most 2.

20. A rectangular chessboard has 5 rows and 2008 columns. Each square is painted either red or blue.

    Determine the largest integer N which guarantees that, no matter how the board is painted, there are two rows which have matching colours in at least N columns.

21. Let $n > 0$ be an integer. We are given a balance scale and n weights of masses $2^0, 2^1, \ldots, 2^{n-1}$, respectively. We are to place each of the n weights on the balance scale, one after another, in such a way that the right pan is never heavier than the left pan. At each step we choose one of the weights that has not yet been placed on the balance scale, and place it on either the left pan or the right pan, until all of the weights have been placed.

    Determine the number of ways in which this can be done.

22. Up until now the national library of the small city state of Sepharia has had n shelves, each shelf holding at least one book. The library recently bought k new shelves. The books will be rearranged, and the librarian has announced that each of the now $n + k$ shelves will hold at least one book. Call a book *privileged* if the shelf on which it will stand in the new arrangement is to hold fewer books than the shelf where it was previously located.

    Prove there are at least $k + 1$ privileged books in the national library of Sepharia.

23. Let n and k be positive integers with $k \geq n$ and $k - n$ an even number. Let $2n$ lamps labelled $1, 2, \ldots, 2n$ be given, each of which can be either on or off. Initially all the lamps are off.

    We consider sequences of steps: at each step one of the lamps is switched from on to off or from off to on.

    Let N be the number of such sequences consisting of k steps and resulting in the state where lamps 1 through n are all on, and lamps $n + 1$ through $2n$ are all off.

    Let M be the number of such sequences consisting of k steps resulting in the state where lamps 1 through n are all on, and lamps $n + 1$ through $2n$ are all off, but where none of the lamps $n + 1$ through $2n$ is ever switched on.

    Determine the ratio $\frac{N}{M}$.

24. A group of 21 girls and 21 boys took part in a mathematical contest. Suppose that

    (i) each contestant solved at most six problems, and

    (ii) for each girl and each boy, at least one problem was solved by both of them.

    Prove that there was a problem that was solved by at least three girls and at least three boys.

13.1 Addition and multiplication

Suppose that the Irish restaurant chain O'Donald's has five types of burger and three types of drink. If you would like a burger *or* a drink, then you have $5 + 3 = 8$ choices. On the other hand, if you would like a burger *and* a drink, then you have $5 \times 3 = 15$ choices.

Many combinatorics problems boil down to applying these addition and multiplication principles in various rather ingenious ways. To make sure that you have some familiarity with these ideas, consider the following basic examples of combinatorial questions. If necessary, you should read over these carefully until you fully understand what is going on—especially because what goes on here, though basic, is extremely useful.

Problem In how many ways can n people stand in a line?

Solution There are n choices of position for the first person to stand in the line. After the first position has been chosen, there are $n-1$ positions remaining for the second person. And after the second position has been chosen, there are $n-2$ positions remaining for the third person, and so on.

Therefore, the number of ways to choose the order in which n people stand in a line is

$$n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1.$$

You probably already know that this number is written as $n!$ and is called 'n factorial'. □

Problem How many three-digit numbers contain no zeros or nines as digits?

Solution There are 8 choices for the first digit, 8 choices for the second digit and 8 choices for the third digit, that is, any number from 1 to 8.

Hence, there are $8^3 = 512$ such numbers. □

Problem Each letter in Morse code is a sequence of at most four dots and dashes. How many letters are possible?

Solution There are $2^1 = 2$ letters with 1 signal, $2^2 = 4$ letters with 2 signals, $2^3 = 8$ letters with 3 signals, and $2^4 = 16$ letters with 4 signals.

Hence, the answer is $2 + 4 + 8 + 16 = 30$. □

So Morse code only just copes with the 26 letters of the English alphabet!

13.2 Subtraction and division

Sometimes the best way to solve a counting problem is to count too much—a technique often known as *overcounting*. If you have counted k too many objects, then the answer is obtained by subtracting k. On the other hand, if you have counted each object k times, then the answer is obtained by dividing by k.

Many combinatorics problems boil down to applying these subtraction and division principles in conjunction with the addition and multiplication principles we discussed in the previous section.

Problem How many five-digit positive integers have at least two digits the same?

Solution The number of five-digit positive integers is $99999 - 10000 + 1 = 90000$.

The number of five-digit positive integers with no two digits the same is $9\times9\times8\times7\times6 = 27216$. This is because there are 9 ways to choose the first digit (because it cannot be zero), 9 ways of choosing the second digit because it cannot be the same as the first, 8 ways of choosing the third digit because it cannot be equal to either of the first two digits, and so on.

Hence, the answer is $90000 - 27216 = 62784$. □

Problem In how many ways can you choose a four-flavour combination from 10 ice-cream flavours?

Solution Well, if you actually cared about the order in which you chose them, then the answer would simply be $10\times9\times8\times7$. But since the order doesn't matter, we have overcounted each combination by a factor of $4 \times 3 \times 2 \times 1$—the number of ways of reordering the four flavours that we have chosen.

Therefore, the answer is

$$\frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} = 210.$$

This can be conveniently expressed using factorial notation as

$$\binom{10}{4} = \frac{10!}{4!\,6!}.$$ □

Problem In how many ways is it possible to arrange the letters of the word *recurrence*?

Solution There are 10! ways to rearrange the letters, but again we have overcounted. The three occurrences of the letter r can be rearranged in 3! ways among themselves. In general, if a letter occurs $k!$ times, then there are $k!$ ways of rearranging the occurrences of that letter among themselves. Since the letters r and e occur three times, the letter c twice and the letters u and n once, the answer is

$$\binom{10}{3,3,2,1,1} = \frac{10!}{3!\,3!\,2!\,1!\,1!} = 50400.$$ □

13.3 Binomial identities

A lot of fun can be had playing around with binomial coefficients. These lie in the intersection of algebra, combinatorics, polynomials and even calculus, giving rise to many problems with multiple solutions from different areas of mathematics.

Recall the following basic identities.

- **Symmetry**
$$\binom{n}{k} = \binom{n}{n-k}$$
- **Binomial theorem**
$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$
- **Addition formula**
$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

- **In-and-out formula**

$$n\binom{n-1}{k-1} = k\binom{n}{k}$$

Problem For every positive integer n, prove that

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^n \binom{n}{n} = 0.$$

Solution We provide two very distinct proofs of this fact. In the next section we provide a third proof! This is a rather common feature of binomial identities and it pays to be comfortable with the various types of proof.

- Method 1: Pascal's triangle
  Recall that we can obtain the nth row of Pascal's triangle from the $(n-1)$th row by adding the two terms above each element of the nth row as per the addition formula above. In particular, with the convention that $\binom{n}{k} = 0$ if $k < 0$ or $k > n$, we have

$$\begin{aligned}
\binom{n}{0} &= \binom{n-1}{-1} + \binom{n-1}{0} \\
\binom{n}{2} &= \binom{n-1}{1} + \binom{n-1}{2} \\
\binom{n}{4} &= \binom{n-1}{3} + \binom{n-1}{4} \\
&\vdots
\end{aligned}$$

  and so $\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots$ is equal to the sum of all the binomial coefficients on the $(n-1)$th row of Pascal's triangle. Similarly,

$$\begin{aligned}
\binom{n}{1} &= \binom{n-1}{0} + \binom{n-1}{1} \\
\binom{n}{3} &= \binom{n-1}{2} + \binom{n-1}{3} \\
\binom{n}{5} &= \binom{n-1}{4} + \binom{n-1}{5} \\
&\vdots
\end{aligned}$$

  and so $\binom{n}{1} + \binom{n}{3} + \binom{n}{5} + \cdots$ is also equal to the sum of all the binomial coefficients on the $(n-1)$th row of Pascal's triangle. Therefore,

$$\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots = \binom{n}{1} + \binom{n}{3} + \binom{n}{5} + \cdots,$$

  which is equivalent to what we wanted to prove. □

- Method 2: Algebra
  The *binomial theorem* states that

$$(x+y)^n = \binom{n}{0}x^0y^n + \binom{n}{1}x^1y^{n-1} + \binom{n}{2}x^2y^{n-2} + \cdots + \binom{n}{n}x^ny^0.$$

  All we need to do now is substitute $x = -1$ and $y = 1$ to obtain the desired result. □

Problem For every positive integer n, prove that

$$\binom{n}{1} + 2\binom{n}{2} + 3\binom{n}{3} + \cdots + n\binom{n}{n} = n2^{n-1}.$$

Solution

- Method 1: Algebra
  We use the in-and-out formula to write $k\binom{n}{k} = n\binom{n-1}{k-1}$. Then the left-hand side becomes

  $$n\binom{n-1}{0} + n\binom{n-1}{1} + \cdots + n\binom{n-1}{n-1} = n2^{n-1},$$

  where $\displaystyle\sum_{k=0}^{n-1}\binom{n-1}{k} = (1+1)^{n-1}$ is a consequence of the binomial theorem.
- Method 2: Calculus
  Substituting $y = 1$ into the binomial formula, we obtain

  $$(x+1)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + \binom{n}{3}x^3 + \cdots + \binom{n}{n}x^n.$$

  Now look at what happens when we differentiate both sides.

  $$n(x+1)^{n-1} = \binom{n}{1} + 2\binom{n}{2}x + 3\binom{n}{3}x^2 + \cdots + n\binom{n}{n}x^{n-1}$$

  All we need to do now is substitute $x = 1$ to obtain the desired result. □

13.4 Bijections

Another way to prove combinatorial identities is to look for combinatorial interpretations for both sides of the identity. This is really a good technique and a useful skill to have. Here we discuss two problems we considered earlier but present bijective solutions.

Problem For every positive integer n, prove that

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^n\binom{n}{n} = 0.$$

Solution The identity is equivalent to the fact that, for every positive integer n,

$$\sum_{k \text{ odd}}\binom{n}{k} = \sum_{k \text{ even}}\binom{n}{k}.$$

If S is a set with n elements, the left-hand side counts the number of subsets of S with an odd number of elements while the right-hand side counts the number of subsets of S with an even number of elements.

Now that we have a combinatorial interpretation for both sides of the equation, it makes sense to look for a one-to-one correspondence—a *bijection* to use more precise terminology—between the objects described by the left-hand side and the objects described by the right-hand side.

In other words, we want to find a rule for turning a subset of S with an odd number of elements into a subset of S with an even number of elements, and vice versa.

We start by picking some fixed $s \in S$. The rule is to add the element s to your set if it doesn't already contain it, and to remove the element s from your set if it does already contain it. Since we are either adding or subtracting a single element, this rule certainly turns a subset of S with an odd number of elements into a subset of S with an even number of elements, and vice versa. It is a bijection because the rule can be reversed. In fact, the rule is its own inverse![1] This completes our combinatorial proof. □

Problem For every positive integer n, prove that

$$\binom{n}{1} + 2\binom{n}{2} + 3\binom{n}{3} + \cdots + n\binom{n}{n} = n2^{n-1}.$$

Solution The idea is to find a set of objects which can be counted in two ways—one of which produces the left-hand side while the other produces the right-hand side. In this particular case, the form of the left-hand side gives a strong hint as to what that set might be.

Each term of the left-hand side is of the form $k\binom{n}{k}$ which suggests that, from a set of n people, we would like to choose a committee of k people. This can be done in $\binom{n}{k}$ ways. Furthermore, we would like to choose a president from this committee. This can be done in k ways. As k ranges from 1 up to n, we are choosing committees of every possible size. So what we have shown is that the left-hand side counts the number of ways to choose a committee, with one person designated the president, from a set of n people.

Of course, all that remains is to show that the number of ways to choose a committee, with one person designated the president, from a set of n people also happens to be $n2^{n-1}$. Whereas we earlier counted by choosing the committee first and then the president, we will now count by choosing the president first and then the remainder of the committee. But this is easy, because there are n choices for the president and for each of the remaining $n-1$ people, there are two choices. Each person is either in the committee or is not. Thus the number of ways of choosing a committee with president is $n2^{n-1}$. This completes our combinatorial proof. □

13.5 The supermarket principle

This is a common trick in counting problems. Although it is a well-defined technique, it does not seem to have a well-defined name but is another example of bijective combinatorics. We shall call it the *supermarket principle* because of the following problem.

Problem The supermarket has an unlimited supply of apple tarts, chocolate muffins and cheesecakes.

How many ways are there to buy seven cakes from the supermarket?

Solution You buy your seven cakes and take them to the checkout, where you put them on the conveyor belt. Just to make things easy, you place all the apple tarts together first, then the chocolate muffins, and then the cheesecakes. To make things really easy for the stressed-out checkout attendant, you place some of those conveyor belt dividers between each type of cake.

Now there will be nine objects on the conveyor belt: seven cakes and two dividers. (If you buy no apple tarts, for example, the first object will be a divider.) Choosing which of these

[1] As we saw in section 10.7, a rule or a function which is its own inverse is called an *involution*.

nine objects are the dividers uniquely determines what cakes you have bought. So the answer is $\binom{9}{2} = 36$. □

In this question, we established a bijective correspondence.

$$\left\{ \begin{array}{c} \text{ways of buying 7} \\ \text{cakes of 3 types} \end{array} \right\} \leftrightarrow \left\{ \begin{array}{c} \text{ways of placing 2} \\ \text{dividers among 7 cakes} \end{array} \right\}$$

You should convince yourself that this is actually correct. That is, for each way of buying the prescribed cakes, there is a way of placing the dividers; and conversely, for each way of placing the dividers, there is a way of buying cakes so that the dividers are placed there. Both of these are quite obvious once you understand what is going on, but often this is difficult to express!

Now for a less supermarket-oriented and slightly more subtle example. There are certainly other solutions to it, but perhaps this is the most elegant.

Problem Find the number of three-digit numbers whose digit sum is 10.

Solution Represent a number by a collection of 'units' (1s) and 'dividers' (Δs). For example, 325 is represented by

$$1\,1\,1\,\Delta\,1\,1\,\Delta\,1\,1\,1\,1\,1.$$

Now every such three-digit number corresponds to a way of placing two dividers among ten 1s. However *the correspondence is not bijective!* This is because the 'conveyor belt'

$$\Delta\,1\,\Delta\,1\,1\,1\,1\,1\,1\,1\,1\,1$$

would correspond to the number 019, which is not a three-digit number. You can't have a divider in the first place. But provided that you start with a 1, the correspondence *is* bijective. (You should make sure of this.)

Ignoring the initial 1, the answer is the number of ways of placing two dividers amongst the remaining nine 1s. There are 11 objects in total and so the number of ways of nominating two of them to be dividers is simply $\binom{11}{2} = 55$. □

13.6 Pigeonhole principle

Although the pigeonhole principle was introduced in sections 1.9 and 1.10, it is a powerful combinatorial tactic which we showcase here.

Problem Given 27 distinct odd positive integers less than 100, prove there is a pair of them whose sum is 102.

Solution Consider the following sets of odd positive integers.

$$\{1, 101\}, \{3, 99\}, \ldots, \{49, 53\}, \{51\}$$

Note that each of the first 25 sets consists of two numbers whose sum is 102, while the 26th set consists of a single number. Furthermore, the 26 sets form a partition of all the odd positive integers less than 100. So given 27 numbers (pigeons), some set (pigeonhole) has two different numbers (pigeons) chosen (living) from (in) it! Those two numbers add to 102. □

Problem The prime factorisations of $r+1$ positive integers $a_1, \ldots, a_{r+1}$ together involve only r primes $p_1, \ldots, p_r$.

Prove that there is a non-empty subset of these integers whose product is a perfect square.

Solution Recall that a positive integer is a perfect square if and only if the exponent of every prime in its prime factorisation is even. So given a product of powers of these primes $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$, define the *parity pattern* of n to be $(\alpha_1, \ldots, \alpha_r)$, where the α_i are considered modulo 2.

Clearly when you multiply two numbers, you add their parity patterns term by term modulo 2, and if you can divide, then you subtract their parity patterns modulo 2. We just need to show there is a non-empty subset of our $r+1$ integers whose product has parity pattern $(0, \ldots, 0)$.

So, consider all 2^{r+1} subsets of our $r+1$ integers, and consider their parity patterns. But there are only 2^r possible parity patterns. By the pigeonhole principle there are two distinct subsets X and Y of $\{a_1, \ldots, a_{r+1}\}$ whose products have the same parity pattern $\mathbf{v}$.

If X and Y were disjoint, then the product of the elements of $X \cup Y$ would have parity pattern $2\mathbf{v} = (0, \ldots, 0)$, as required.

If X and Y overlap, let $D = X \cap Y$. We claim that the product of the elements of $X \cup Y \setminus D$ is a perfect square. Indeed consider the equality,

$$\left(\prod_{x\in X} x\right)\left(\prod_{y\in Y} y\right) = \left(\prod_{x\in X\setminus D} x\right)\left(\prod_{y\in Y\setminus D} y\right)\left(\prod_{d\in D} d^2\right).$$

The LHS has parity pattern $(0, \ldots, 0)$ because $\prod_{x\in X} x$ and $\prod_{y\in Y} y$ have the same parity pattern. Also $\prod_{d\in D} d^2$ has parity pattern $(0, \ldots, 0)$. It follows that

$$\left(\prod_{x\in X\setminus D} x\right)\left(\prod_{y\in Y\setminus D} y\right)$$

also has parity pattern $(0, \ldots, 0)$ and so the product of the elements of $X \cup Y \setminus D$ is a perfect square, as required. □

13.7 Principle of inclusion–exclusion

Here is a basic idea of set theory.[2]

$$|X \cup Y| = |X| + |Y| - |X \cap Y|$$
$$|X \cup Y \cup Z| = |X| + |Y| + |Z| - |X \cap Y| - |Y \cap Z| - |Z \cap X| + |X \cap Y \cap Z|$$

This can be generalised as per the following theorem.

[2] Why not draw the corresponding Venn diagrams to see what is going on?

Principle of inclusion–exclusion

$$|X_1 \cup \cdots \cup X_n| = \sum_{\{i_1,\ldots,i_k\} \subseteq \{1,\ldots,n\}} (-1)^{k-1} \, |X_{i_1} \cap \cdots \cap X_{i_k}| \, .$$

It can also be inverted, since

$$(S \setminus X_1) \cup \cdots \cup (S \setminus X_n) = S \setminus (X_1 \cap \cdots \cap X_n)$$

and

$$(S \setminus X_1) \cap \cdots \cap (S \setminus X_n) = S \setminus (X_1 \cup \cdots \cup X_n),$$

which yield

$$|X_1 \cap \cdots \cap X_n| = \sum_{\{i_1,\ldots,i_k\} \subseteq \{1,\ldots,n\}} (-1)^{k-1} \, |X_{i_1} \cup \cdots \cup X_{i_k}| \, .$$

Problem An absent-minded postman has n letters to deliver to n different addresses.

In how many ways can he deliver the mail, one letter to each address, so that no letter is delivered to its correct address?

Solution This problem really involves permutations, which we consider to be bijections from the set $\{1, 2, \ldots, n\}$ to itself or equivalently, arrangements of the numbers $1, 2, \ldots, n$. What we are seeking is the number D_n of derangements of the set $\{1, 2, \ldots, n\}$. A *derangement* is a permutation of a set which leaves no element fixed.

For $i = 1, 2, \ldots, n$, let X_i denote the set of permutations which fix the number i. The number of elements in X_i is easy to count and it is simply $(n-1)!$. Furthermore, we know that $X_{i_1} \cap X_{i_2} \cap \cdots \cap X_{i_k}$ is simply the set which fixes the number i_1, fixes the number i_2, and so on, up to the number i_k. This is a set which is also easy to count and its number of elements is simply $(n-k)!$. Note that the total number of ways of choosing k of the X_i is equal to $\binom{n}{k}$.

What does the set $X_1 \cup X_2 \cup \cdots \cup X_n$ represent? It simply represents the number of permutations that fix 1 or 2 or 3, and so on, up to n. In other words, it is the set of permutations that fix at least one of $1, 2, \ldots, n$. So the number of elements it contains is simply $n! - D_n$. Therefore,

$$\begin{aligned}
|X_1 \cup \cdots \cup X_n| &= \sum_{\{i_1,\ldots,i_k\} \subseteq \{1,\ldots,n\}} (-1)^{k-1} \, |X_{i_1} \cap \cdots \cap X_{i_k}| \\
\Rightarrow \qquad n! - D_n &= \sum_{k=1}^{n} (-1)^{k-1} (n-k)! \binom{n}{k} \\
\Rightarrow \qquad D_n &= n! - \sum_{k=1}^{n} (-1)^{k-1} \frac{n!}{k!} \\
&= n! \left(\frac{1}{0!} - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!} \right) .
\end{aligned}$$

□

13.8 Double counting

Double counting is short for counting the same thing in two different ways.

Problem At a party each person knew exactly 22 others. For any pair of people X and Y who knew one another, there was no other person at the party whom they both knew. For any pair of people X and Y, who did not know one another, there were exactly 6 other people whom they both knew.

How many people were at the party?

Solution This problem has an obvious graph theory interpretation, where vertices represent people and edges represent a mutual acquaintanceship between two persons. Define a *vee* to be a triple of persons such that exactly two of the three pairs of acquaintances know each other. We count the number of vees in two different ways.[3]

Suppose there are n people at the party. Concentrating on vertices we see that each vertex contributes $\binom{22}{2} = 231$ vees since each vertex has 22 edges emanating from it. Thus the total number of vees is $231n$.

On the other hand adding the degrees of each vertex gives $22n$, but this overcounts the number of edges by a factor of two. Therefore, the total number of edges is $11n$. This means that the total number of pairs of vertices not connected by an edge is

$$\binom{n}{2} - 11n.$$

Each such non-edge makes a vee with 6 other vertices. Thus the total number of vees is

$$6\left(\binom{n}{2} - 11n\right).$$

Equating the two expressions for the number of vees yields

$$231n = 6\left(\binom{n}{2} - 11n\right).$$

This is easily solved for n and yields $n = 100$. □

Problem Let $p_n(k)$ be the number of permutations of the set $\{1, 2, \ldots, n\}$ which have exactly k fixed points.

Prove that

$$\sum_{k=0}^{n} k p_n(k) = n!.$$

Solution We find a combinatorial interpretation for the sum. We are counting each permutation, but counting each with multiplicity equal to the number of its fixed points. So permutations without any fixed points, that is, derangements, are not counted at all; permutations with one fixed point are counted once; those with two fixed points are counted twice; and so on. Consider a permutation as the numbers 1 to n, written in some order.

For each permutation with one or more fixed points, colour in one of the fixed points as our 'favourite'. Then the set of 'permutations with favourite fixed points' has precisely $\sum_{k=0}^{n} k p_n(k)$ elements. This is a combinatorial interpretation for the sum.

For example, here are the 'permutations with favourite fixed points' for the case $n = 3$.

$$\mathbf{1}23, \quad 1\mathbf{2}3, \quad 12\mathbf{3}, \quad \mathbf{1}32, \quad 2\mathbf{1}3 \quad \text{and} \quad 3\mathbf{2}1.$$

[3] In the graph theory interpretation a vee actually looks like the letter V.

How many of these objects are there? Well, with **1** as the favourite fixed point, there are $(n-1)!$ permutations—one for each permutation of the other objects. Similarly, with **2** as the favourite fixed point, there are also $(n-1)!$ permutations. There are $(n-1)!$ permutations for each individual favourite fixed point.

This gives $n \times (n-1)! = n!$ of them overall.[4] □

13.9 Injections

Problem A permutation $(x_1, x_2, \ldots, x_{2n})$ of the set $\{1, 2, \ldots, 2n\}$, where n is a positive integer, is said to be *good* if

$$|x_i - x_{i+1}| = n$$

for at least one i in $\{1, 2, \ldots, 2n-1\}$, and is said to be *bad* otherwise.

Show that, for each n, there are more good permutations than bad permutations.

Solution Since we want to show there are *more* of one thing than another, we don't construct a bijection, but an *injection*! We find a map from bad permutations to good permutations which is injective but not surjective: this will show there are more good permutations.

First think about what a good permutation means. For given $x \in \{1, \ldots, 2n\}$, there is only one y in the set for which $|x - y| = n$. So a permutation is good if and only if we see at least one of the pairs

$$(1, n+1), (2, n+2), \ldots, (n, 2n)$$

occurring as adjacent numbers in the permutation (in any order). A bad permutation is one where we see none of these pairs of numbers adjacent. In what follows, it will be convenient to define the notation $i * n$ as follows.

$$i * n = \begin{cases} i + n & \text{if } 1 \leq i \leq n \\ i - n & \text{if } n+1 \leq i \leq 2n \end{cases}$$

Now take any bad permutation. Here is how to make a good one out of it. Suppose $x_1 = i$. Then $i * n$ can't occur as x_2, but must occur later on. Thus the permutation looks like

$$(i, A, i * n, B),$$

where A is a non-empty sequence, and B is another (possibly empty) sequence, neither of which contains a pair of the form $(j, j * n)$. We make our permutation good by putting i and $i * n$ together to make

$$(A, i, i * n, B).$$

Thus we have a map ϕ from the set of bad permutations to the set of good permutations given by

$$\phi(i, A, i * n, B) = (A, i, i * n, B).$$

We first show our map ϕ is not surjective. Any good permutation in the image of ϕ is of the form $(A, i, i * n, B)$, where A is non-empty and the only adjacent pair of the form $(j, j * n)$ occurs if $j = i$. Clearly there are more good permutations than these. For example, any permutation of the form $(1, n+1, \ldots)$ is a good permutation that is not in the image of ϕ.

Now we show ϕ is injective. Suppose that $\phi(i_1, A_1, i_1 * n, B_1) = \phi(i_2, A_2, i_2 * n, B_2)$. Then from the definition of ϕ we have

$$(A_1, i_1, i_1 * n, B_1) = (A_2, i_2, i_2 * n, B_2.)$$

[4] There are many other solutions to this problem. See if you can find any others.

However from our earlier observation the only adjacent pair of the form $(j, j * n)$ on the LHS occurs at $j = i_1$. Similarly, the only adjacent pair of the form $(j, j * n)$ on the RHS occurs at $j = i_2$. This allows us to deduce that $i_1 = i_2$. Then it follows that $A_1 = A_2$ and $B_1 = B_2$. Therefore, ϕ is injective. □

13.10 Recursion

As a first example of the technique of *recursion*, we solve the problem about the absent-minded postman that we saw in section 13.7.

Problem An absent-minded postman has n letters to deliver to n different addresses.

In how many ways can he deliver the mail, one letter to each address, so that no letter is delivered to its correct address?

Solution Suppose that D_n is the number of derangements of n objects. The trick here is to relate D_n to some of the previous values $D_1, D_2, \ldots, D_{n-1}$.

- We will take a derangement of $n-1$ letters and construct $n-1$ distinct derangements of n letters. Suppose that in our derangement of $n-1$ letters, letter 1 is delivered to address j. Then we adjust this by delivering letter 1 to address n and letter n to address j. This gives a derangement of n letters. In general, we could take any $i \in \{1, 2, \ldots, n-1\}$ and suppose that in our derangement of $n-1$ letters, letter i is delivered to address j. Then we adjust this by delivering letter i to address n and letter n to address j. This gives a derangement of n letters.

  So given a derangement of $n-1$ letters and a number $i \in \{1, 2, \ldots, n-1\}$, we have constructed a derangement of n letters. You can check that these are indeed all different. Thus we have constructed

  $$(n-1)D_{n-1}$$

  derangements of n letters.

- Unfortunately, we have not accounted for every possible derangement of n letters. The ones we are missing are precisely those where letter n is delivered to address j and letter j is delivered to address n for some $j \in \{1, 2, \ldots, n-1\}$. But in this case, notice that the remaining letters and addresses form a derangement on $n-2$ letters. Therefore, for each j there are D_{n-2} ways in which this can happen. But since we can take any $j \in \{1, 2, \ldots, n-1\}$ there are

  $$(n-1)D_{n-2}$$

  cases in total, where the letter n has swapped addresses with another letter.

Putting our two pieces of information together we have

$$D_n = (n-1)(D_{n-1} + D_{n-2}).$$

This is a nice simple recursion for D_n which we can manipulate as follows.

$$D_n - nD_{n-1} = -(D_{n-1} - (n-1)D_{n-2})$$

Letting

$$E_n = D_n - nD_{n-1},$$

for $n = 1, 2, \ldots$, we see that

$$E_n = -E_{n-1} = E_{n-2} = \cdots = (-1)^{n-2}E_2 = (-1)^{n-2}(D_2 - 2D_1) = (-1)^n.$$

Here we have used $D_2 = 1$ and $D_1 = 0$.

Therefore,

$$D_n - nD_{n-1} = (-1)^n,$$

and so,

$$\frac{D_n}{n!} - \frac{D_{n-1}}{(n-1)!} = \frac{(-1)^n}{n!}.$$

Since this is true for all integers $n \geq 1$ we also have the same equation for n replaced successively by $n-1, n-2, \ldots, 1$. Upon summing these equations, all the middle terms on the left-hand side cancel out leaving us with

$$\frac{D_n}{n!} - \frac{D_0}{0!} = \frac{(-1)^n}{n!} + \frac{(-1)^{n-1}}{(n-1)!} + \cdots + \frac{(-1)^1}{1!},$$

which is equivalent to the result we obtained earlier in section 13.7. □

Problem In town A there are n girls and n boys such that each girl knows each boy. In town B there are n girls and $2n-1$ boys such that girl k knows boys $1, 2, 3, \ldots, 2k-1$, and only these boys. Let $A(n, r)$ denote the number of different ways in which r girls from town A can dance with r boys from town A, forming r pairs where the girl knows the boy. Similarly, let $B(n, r)$ denote the number of different ways in which r girls from town B can dance with r boys from town B, forming r pairs where the girl knows the boy.

Prove that $A(n, r) = B(n, r)$, for $r = 1, 2, \ldots, n$.

Solution We can calculate the number $A(n, r)$ very easily. The number of ways of choosing r girls is $\binom{n}{r}$ and the number of ways of choosing r boys is $\binom{n}{r}$. The number of ways of pairing them up is $r!$. Therefore, we have

$$A(n, r) = \binom{n}{r}^2 r! = \frac{n!^2}{(n-r)!^2\, r!}.$$

To show that

$$B(n, r) = \frac{n!^2}{(n-r)!^2\, r!}$$

directly is a very difficult task indeed. However, what we can do is relate $B(n, r)$ to the values $B(n-1, r-1)$ and $B(n-1, r)$ in the following way.

We establish a recurrence relation for $B(n, r)$. Let $n \geq 2$ and $2 \leq r \leq n$. There are two cases for a desired selection of r pairs of girls and boys.

- Case 1: Girl n is dancing.
  The remaining $r-1$ girls can choose their partners in $B(n-1, r-1)$ ways while girl n can choose her partner from any of the unchosen $2n-r$ boys.

  This contributes $(2n-r)B(n-1, r-1)$ to the value of $B(n, r)$.

- Case 2: Girl n is not dancing.
  This case is easy because there are simply $B(n-1, r)$ possible choices.

So for every $n \geq 2$ and $2 \leq r \leq n$, we have the recursion

$$B(n,r) = (2n-r)B(n-1,r-1) + B(n-1,r),$$

with initial conditions

$$\begin{aligned} B(n,r) &= 0, \quad \text{if } r > n \\ B(n,1) &= 1+3+5+\cdots+(2n-1) = n^2. \end{aligned}$$

It is directly verified that the numbers $A(n,r)$ satisfy the same initial conditions and recurrence relations, from which it follows that $A(n,r) = B(n,r)$ for all n and $r \leq n$. We leave it to the reader to do the algebra that verifies this. □

13.11 Double counting via tables

Often a problem can be interpreted in terms of a table. The information is somehow recorded in the cells of the table. This gives a really clear way of understanding what is happening. Then by examining the situation in the table horizontally on the one hand and vertically on the other hand, we have a natural way to do some double counting.

Problem In a competition, there are a contestants and b judges, where $b \geq 3$ is an odd integer. Each judge rates each contestant as either *pass* or *fail*. Suppose k is a number such that, for any two judges, their ratings coincide for at most k contestants.

Prove that

$$\frac{k}{a} \geq \frac{b-1}{2b}.$$

Solution It is advantageous to organise the information of this problem in a table. We do so by having one row for each contestant and one column for each judge. We place a 1 or a 0 in a cell if the judge corresponding to that column has given the contestant corresponding to that row a pass or a fail, respectively.

The tactic now is to find something in the table which we can double count. What we will count is the number of matches between two judges' ratings. More precisely, we are counting the number of instances where two different judges J_x and J_y agree on the result of contestant C. Let the number of these instances be N.

	J_1	J_2	J_3	J_4	J_5	J_6	J_7
C_1	0	1	0	0	0	0	1
C_2	1	0	1	0	1	0	1
C_3	1	1	1	0	0	0	0
C_4	1	1	0	0	1	1	0
C_5	0	1	0	1	0	1	1

- *Think vertically.*
  For each pair of judges, we know that they agree in at most k places. Since there are $\binom{b}{2}$ pairs of judges, we conclude that

$$N \leq k\binom{b}{2}.$$

- *Think horizontally.*
  Suppose a contestant has received m passes and n fails, where $m+n=b$. The m judges who awarded the contestant a pass contribute $\binom{m}{2}$ agreements, while the n judges who awarded the contestant a fail contribute $\binom{n}{2}$ agreements.

  Therefore, looking along a row, the number of agreements is

$$\begin{aligned}\binom{m}{2}+\binom{n}{2} &= \frac{m(m-1)}{2}+\frac{n(n-1)}{2}\\ &= \frac{m^2+n^2}{2}-\frac{b}{2}.\end{aligned}$$

  Since $m+n=b$ is constant, m^2+n^2 is minimised when m and n are as close to each other as possible. Since b is odd this occurs when $\{m,n\}=\{\frac{b-1}{2},\frac{b+1}{2}\}$. So we have

$$\begin{aligned}\frac{m^2+n^2}{2}-\frac{b}{2} &\geq \frac{1}{2}\left[\left(\frac{b-1}{2}\right)^2+\left(\frac{b+1}{2}\right)^2\right]-\frac{b}{2}\\ &= \left(\frac{b-1}{2}\right)^2.\end{aligned}$$

  Summing over the number of rows, we obtain

$$N \geq a\left(\frac{b-1}{2}\right)^2.$$

All that remains is to put these two pieces of information together. What we have shown is that

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

It follows that

$$\begin{aligned}\frac{k}{a} &\geq \left(\frac{b-1}{4}\right)^2 \Big/ \binom{b}{2}\\ &= \frac{b-1}{2b}.\end{aligned}$$

□

13.12 Combinatorial reciprocal principle

The *what* principle, I hear you say? Yes, this is another principle to which we gave a name, because it seemed really interesting at the time, and because nobody else seems to have named it. Here's the principle: if you have a set S of objects falling into k different categories, then

$$\sum_{x\in S}\frac{1}{\text{number of objects in the same category as } x\text{, including } x} = k.$$

To help you understand what this is saying, write down some examples and then try to prove the principle. Once it is well understood, this principle can solve some otherwise difficult, or even apparently unapproachable, problems.

Problem Students from 13 different countries participated in the 491st International Mathematics Bonanza. Each student belonged to one of five different age groups.

Prove that there were at least nine participants in the Bonanza who had more fellow participants in his or her age group than fellow participants from his or her own country.

Solution Given a student x, let A_x denote the number of students (including x) in the same age group as x, and let C_x denote the number of students from the same country. So by the combinatorial reciprocal principle,

$$\sum_x \frac{1}{A_x} = 5 \quad \text{and} \quad \sum_x \frac{1}{C_x} = 13.$$

Thus

$$\sum_x \frac{1}{C_x} - \frac{1}{A_x} = 8.$$

Since A_x and C_x are positive integers, we have

$$\frac{1}{C_x} - \frac{1}{A_x} < 1$$

for each x.

So the number of x for which $\frac{1}{C_x} - \frac{1}{A_x} > 0$ is more than 8. That is, there are at least 9 students for which $\frac{1}{C_x} > \frac{1}{A_x}$, that is, $A_x > C_x$. Thus we are done. □

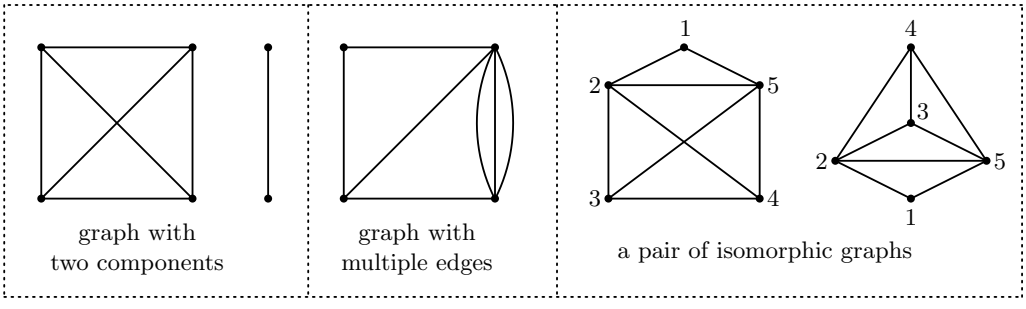
14

Graph theory

Graph theory is a truly remarkable area of mathematics. In one sense, the concept of a graph is so easy that every human being has some innate understanding of it. On the other hand, graph theory offers some notoriously difficult problems and is the site of much active mathematical research today. In between the very simple and the very complicated lies a vast region of interesting and amazing results, some of which we'll soon see.

When we use the word *graph* in this chapter, we don't mean a complicated representation of a function on a set of coordinate axes or anything like that. We simply mean a diagram obtained by joining dots.

In the usual graph theory terminology, the dots are called *vertices* while the lines connecting them are called *edges*. The edges of a graph merely represent relationships between the vertices. In particular, we usually don't care how a graph is drawn in the plane. In fact, we consider two graphs G and H to be the same—the technical word is *isomorphic*—if there is a one-to-one correspondence between their vertices, such that two vertices are connected by an edge in G if and only if the two corresponding vertices are connected by an edge in H. Note that a graph is allowed to consist of more than one connected component.



graph with two components

graph with multiple edges

a pair of isomorphic graphs

For most of our purposes, we will assume that a graph is finite and has no loops (a loop is an edge connecting a vertex with itself) or multiple edges (two or more edges connecting the same pair of vertices). However, there are situations where considering loops and multiple edges is useful.

14.0 Problems

1. Remembering that we view two graphs as being the same if they are isomorphic, how many graphs are there with at most five vertices? How many of these are planar?[1]

2. At a party with 100 people, in any set of four there is at least one person who is mutually acquainted with the other three.

   Given that there are three people who are mutually unacquainted with each other, prove that the remaining 97 people must know everyone at the party.

3. In a country with n towns, there are some one-way roads connecting pairs of towns. It's known that for any two towns you can drive from one of them to the other one.

   Prove that there is a town from which it's possible to drive to any other town.

4. Show that if every vertex in a graph has degree k, then the graph contains a path of length k. (See section 14.7 for the definition of a path.)

5. Consider a $3 \times 3 \times 3$ cube made from 27 unit cubes. These smaller cubes are rooms which have doors on each of their faces.

   Is it possible to start in the central room and visit every other room exactly once without ever leaving the large cube?

6. Consider a tennis tournament in which each of n people played against every other person exactly once.

   (a) Show that it is possible to label the players $P_1, P_2, \ldots, P_n$ in such a way that P_1 beat P_2, P_2 beat P_3, and so on to P_{n-1} beat P_n.

   (b) Suppose it is known that there is no loop of players $Q_1, Q_2, \ldots, Q_k$ such that Q_1 beat Q_2, Q_2 beat Q_3, and so on up Q_{k-1} beat Q_k and Q_k beat Q_1, for some $k \geq 3$. Prove that there is a unique way to rank the players so that each player beat everyone below them.

7. A *snozzberry* has the shape of a convex polyhedron such that three faces meet at every vertex, and each face is either a pentagon or a hexagon. Each pentagon is surrounded by five hexagons while each hexagon is surrounded by three hexagons and three pentagons.

   How many faces does a snozzberry have?

8. My wife and I were invited to a dinner party attended by four other couples, making a total of 10 people. A certain amount of handshaking took place subject to two conditions: no one shook his or her own hand and no couple shook hands with each other. Afterwards, I became curious and asked everybody else at the party how many people they shook hands with.

   Given that I received nine different answers, how many hands did I shake?

9. An n-domino consists of two squares which share an edge, where each square is labelled by a number from 1 up to n.

   (a) Prove that there are $\binom{n+1}{2}$ dominoes.

   (b) For which values of n is it possible to arrange the dominoes in a line so that adjacent halves of neighbouring dominoes have the same label?

10. Prove that a graph is bipartite if and only if it has no cycles of odd length.

[1] Some of the problems in this section use terminology that is defined later in the chapter.

11. If G is a graph with V vertices and E edges, prove that the following statements are equivalent to each other.

    (i) The graph G is a tree.

    (ii) The graph G is connected, and the removal of any edge leaves G disconnected.

    (iii) Any two vertices of G are connected by exactly one path.

    (iv) There are no cycles in G, but the addition of any edge creates a cycle.

12. A convex polyhedron with an even number of edges is given.

    Prove that an arrow can be placed on each edge so that each vertex is pointed to by an even number of arrows.

13. Prove that the complete bipartite graph $K_{3,3}$ is not planar.

14. Seventeen people correspond by mail with one another, each one with all the rest. In their letters only three different topics are discussed and each pair of people deals with only one of these topics.

    Prove that there are at least three people who write to each other about the same topic.

15. (a) During a meeting of $2n$ people, more than n^2 handshakes took place, with no pair of people shaking hands more than once.

        (i) Prove that there must have been three people who all shook hands with each other.

        (ii) Is the problem still true if exactly n^2 handshakes took place?

    (b) What is the maximal number of edges that a graph on n vertices can have such that the graph contains no triangle?

16. (a) Prove that, at any party with nine people, there must exist four mutual friends or three mutual strangers.

    (b) Show that this is not true for a party with eight people.

17. A *Platonic solid* is a convex polyhedron in which each vertex is surrounded by the same number of congruent regular polygons.

    Prove that there are exactly five Platonic solids and determine the number of vertices, edges and faces of each.

18. Consider a planar graph whose faces, including the infinitely large one, are all triangles.

    If each vertex is coloured red, green or blue, prove that the number of faces whose vertices are all different colours is even.

19. There are 1985 participants at an international meeting. In each set of three participants, there are at least two who speak the same language.

    Given that no one speaks more than five languages, prove that there are at least 200 participants who speak the same language.

20. Each edge of the complete graph K_9 is coloured either blue, or red, or left uncoloured.

    Find the smallest value of n such that whenever n edges are coloured, there necessarily exists a monochromatic triangle.

21. A prism with pentagons $A_1A_2A_3A_4A_5$ and $B_1B_2B_3B_4B_5$ as top and bottom faces is given. Each of the sides of the pentagons and each of the line segments A_iB_j, for all $i, j = 1, 2, 3, 4, 5$, is coloured either red or green. It is also known that each triangle whose vertices are vertices of the prism and whose sides have all been coloured, has two sides of a different colour.

    Show that all 10 sides of the top and bottom faces are the same colour.

22. In a certain country there are n towns, where $n \geq 4$. Initially there are no roads connecting the towns. A road may be built between towns A and B if there exist two other towns X and Y such that there is no road between towns A and X, there is no road between towns X and Y and there is no road between towns Y and B.

    What is the maximum number of roads that can be built?

23. At a school some students are friends. A *friendship* is a set of two students who are friends with each other. A *trio* is a set of three students who are all friends with each other.

    Prove that
    $$t^2 \leq \frac{2}{9}f^3,$$
    where f is the number of friendships and t is the number of trios.

24. Among a group of 120 people, some pairs are friends. A *weak quartet* is a set of four people containing exactly one pair of friends.

    What is the maximum possible number of weak quartets?

14.1 Degree

Let's start with a few fundamental definitions in graph theory. If one of the endpoints of the edge e is the vertex v, then we say that e and v are *incident*. If two vertices u and v are incident to the same edge, then we say that u and v are *adjacent*. Now define the *degree* of a vertex v to be the number of edges incident to v and denote it by $\deg(v)$.

Problem Show that at any party, there are always at least two people with exactly the same number of friends at the party.

This is the first of many party problems which we will examine. In fact, you can think of every graph theory problem as a party problem in disguise. Vertices represent people, while edges represent mutual acquaintance. For the time being, we won't deal with the case of celebrities or forgetful people, that is, where A knows B but B doesn't know A. Furthermore, we don't count people as knowing themselves, so that the graph has no loops, and we don't allow people to know each other twice over, so that the graph has no multiple edges. The degree of a vertex represents the number of people at the party that a person knows so, in some sense, degree is a popularity index!

Throughout the chapter, we will switch between party language and graph language, depending on the circumstances. This particular problem can be translated into graph theory terminology in the following way.

> Show that in any graph, there exist two vertices with the same degree.

Solution With the goal of obtaining a contradiction, suppose that each person knows a different number of people at the party.

If there are n partygoers, then they can know $0, 1, 2, \ldots, n-1$ people, and these are the only possibilities. Since there are n people and n possibilities for the number of people they know, there must be one person who knows 0 people, one person who knows 1 person, and so on, up to one person who knows all $n-1$ other people at the party. However, it's impossible for there to be two people at the party, one who knows no one else and one who knows everybody else. This is the desired contradiction, so we can conclude that there exist two people who know the same number of people at the party. □

At any party, if you ask everyone in the room, including yourself, how many hands they shook, and add up all of the answers, then you will always end up with an even number. In fact, the sum will be twice the number of handshakes that have occurred during the party. We can state this in the language of graph theory in the following way.

Handshaking lemma In any graph, the sum of the degrees of all the vertices is equal to twice the number of edges.

This is true because each edge contributes two to the sum of the degrees, one for each vertex incident to it. A simple corollary of the handshaking lemma is the fact that the number of vertices of odd degree in any graph must be even.

Problem Is it possible to build a house with exactly eight rooms, each with three doors, and such that exactly three of the house's doors lead outside?

Solution If you could build such a house, then you could construct a corresponding graph in the following way. Let there be nine vertices, one for each room and one to represent the

outside of the house. Place an edge between two vertices if there is a door between the two corresponding areas. The conditions of the problem assert that every vertex of our graph has degree three.

Since there are nine vertices, the sum of the degrees is $9 \times 3 = 27$. But the handshaking lemma tells us that there should be $13\frac{1}{2}$ edges in our graph, which is clearly impossible! □

14.2 Directed graphs

There are many graph theory problems which involve tournaments, one-way roads and the like. These are easily represented by a *directed graph*. This is a graph in which every edge has a direction, usually indicated by an arrow. For instance, in the case of a tournament, players can be represented by vertices and matches can be represented by directed edges, where the direction of an edge points from the winner to the loser.

Each vertex of a directed graph has both an *indegree*, which counts the number of incoming edges, as well as an *outdegree*, which counts the number of outgoing edges. We will denote the indegree and outdegree of a vertex v by $\text{indeg}(v)$ and $\text{outdeg}(v)$, respectively. Sometimes, we may want to refer to the total degree of a vertex, which is simply $\text{indeg}(v) + \text{outdeg}(v)$.

Problem Consider a squash tournament in which each of n people plays against every other person exactly once. Let L_k and W_k be the number of losses and wins of the kth player, respectively.

Prove that

$$L_1^2 + L_2^2 + \cdots + L_n^2 = W_1^2 + W_2^2 + \cdots + W_n^2.$$

Solution We rearrange the given equation and seek to prove that

$$(L_1 + W_1)(L_1 - W_1) + (L_2 + W_2)(L_2 - W_2) + \cdots + (L_n + W_n)(L_n - W_n) = 0.$$

In the directed graph which represents this tournament, the total degree for the kth player is

$$\text{indeg}(v) + \text{outdeg}(v) = L_k + W_k = n - 1.$$

Since this is a constant, we can divide the previous equation through by $n - 1$ to obtain

$$(L_1 - W_1) + (L_2 - W_2) + \cdots + (L_n - W_n) = 0.$$

So our aim is to prove that

$$L_1 + L_2 + \cdots + L_n = W_1 + W_2 + \cdots + W_n.$$

But this is now quite simple, since the left-hand side represents the sum of the indegrees, which is equal to the number of edges in the graph. Similarly, the right-hand side represents the sum of the outdegrees, which is also equal to the number of edges in the graph. □

14.3 Connected graphs, cycles and trees

There's quite a bit of terminology in graph theory, but thankfully, the vast majority of it seems to make perfect sense.

- A *connected graph* is one in which it's possible to walk between any two vertices along the edges. In other words, the entire graph consists of only one piece.
- A *cycle* is a sequence of distinct vertices $v_1, v_2, \ldots, v_n$ with v_1 adjacent to v_2, v_2 adjacent to v_3, and so on, with v_n adjacent to v_1.
- A *tree* is a connected graph which has no cycles.

Problem If G is a graph with V vertices and E edges, prove that G is a tree if and only if it is connected and $E = V - 1$.

Solution First, we will prove that if G is a tree, then G is connected and $E = V - 1$. Of course, the fact that G is connected follows immediately from the definition of a tree, so it's only necessary to prove that $E = V - 1$.

We will proceed by induction on the number of vertices.

The base case is rather trivial, since you can see for yourself that the statement is true if G has 1 or 2 vertices.

Suppose now that the statement holds for all trees with n vertices and consider an arbitrary tree with $n + 1$ vertices. The main idea is to prove that there exists a vertex with degree 1.

Start at a random vertex and start walking along a path which never visits the same vertex twice. Sooner or later, you must get stuck in one of two ways: either you return to a vertex you already visited, or you reach a dead end. The former case implies that G contains a cycle, which is a contradiction. The latter case gives us the vertex of degree 1 that we are looking for. Now we simply remove this vertex and the single edge incident to it. This leaves us with a tree with n vertices and, by the inductive hypothesis, $n - 1$ edges. Hence, our original tree must have $n + 1$ vertices and n edges and so satisfies $E = V - 1$.

Now we aim to prove the converse, that if G is connected and $E = V - 1$, then G is a tree. Of course, the fact that G is connected is immediate, so it's only necessary to prove that G has no cycles.

Again, we will proceed by induction on the number of vertices.

The base case is rather trivial, since you can see for yourself that the statement is true if G has 1 or 2 vertices.

Suppose now that the statement holds for all G with n vertices and consider an arbitrary graph with $n + 1$ vertices which is connected and satisfies $E = V - 1$. Once again, the main idea is to prove that there exists a vertex with degree 1.

From the handshaking lemma, we can deduce that if every vertex in the graph has degree at least 2, then $V \leq E$, which is a contradiction. Furthermore, since G is connected, there are no vertices with degree 0. This gives us the vertex of degree 1 that we are looking for. Now we simply remove this vertex and the single edge incident to it. This leaves us with a graph which is connected with n vertices and $n - 1$ edges. By the inductive hypothesis, such a graph has no cycles. Furthermore, adding in a new edge incident to a vertex with degree 1 cannot create any cycles. In other words, G is a tree. □

The solution to the previous problem should give you some ideas which you can use to prove the following important results.

Theorem Consider a graph with V vertices and E edges.

- If $E \leq V - 2$, then the graph is not connected.
- If $E \geq V$, then the graph contains a cycle.

14.4 Complete graphs and bipartite graphs

There are certain types of graph which seem to pop up in all sorts of problems.

- The *complete graph* K_n is the graph consisting of n vertices with an edge between every pair of vertices.
- A *bipartite graph* is one whose vertices can be coloured black and white such that every edge is incident to one black vertex and one white vertex.
- The *complete bipartite graph* $K_{m,n}$ is the graph consisting of m black vertices and n white vertices with an edge between each black and each white vertex.

The best way to remember jargon like this is to put it to good use!

Problem In the parliament of Frenemia each member is friends with exactly one other member and enemies with exactly one other member.

Prove that the members of parliament can be divided into two chambers so that no chamber contains a pair of mutual friends or a pair of mutual enemies.

Solution As is often the case, our first step will be to phrase this in graph theory language.

> In a graph, the edges are coloured red and blue in such a way that each vertex is incident to one red edge and one blue edge. Prove that the graph is bipartite.

We will give a procedure for colouring the vertices black and white such that every edge is incident to one black vertex and one white vertex.

If you try drawing some graphs which satisfy the conditions of the problem, then you'll find that they all consist of a bunch of cycles, each one with an even number of vertices.

With this in the back of our minds, let's start by taking any vertex and colouring it black. Now take another vertex adjacent to it and colour it white. And take another vertex adjacent to this one and colour it black. Continue walking around the graph, alternately colouring vertices black and white until you get stuck. Clearly, this can only happen in one of two ways: either you reach a dead end, or you meet a vertex which has already been coloured. The former case actually never arises, since every vertex has degree two while a dead end is a vertex of degree one. In the second case—and you should carefully think about why this is so—we must return to the vertex at which we started.

In summary, we have traversed a cycle, alternately colouring vertices black and white. Now a problem arises if the cycle has an odd number of vertices. But we've been told that the edges alternate between red and blue, so there must be an even number of vertices along the cycle. At this stage, we have successfully managed to colour some of the vertices black and white. If we have coloured every vertex, then we are done, but if not, then we simply repeat the process, starting at an uncoloured vertex.

This recipe will eventually colour every vertex in such a way that each edge is incident to one black vertex and one white vertex, so the graph is certainly bipartite. □

14.5 Pigeonhole principle

Let's return to the graph theory party scene! We will assume that at a graph theory party, every pair of people either know each other, in which case we call them friends, or they do

not know each other, in which case we call them strangers. We will use vertices to represent people, red edges to represent friends and blue edges to represent strangers. Therefore, every party is simply a complete graph with each edge coloured red or blue.

Problem Prove that, at any party with six people, there must exist three mutual friends or three mutual strangers.

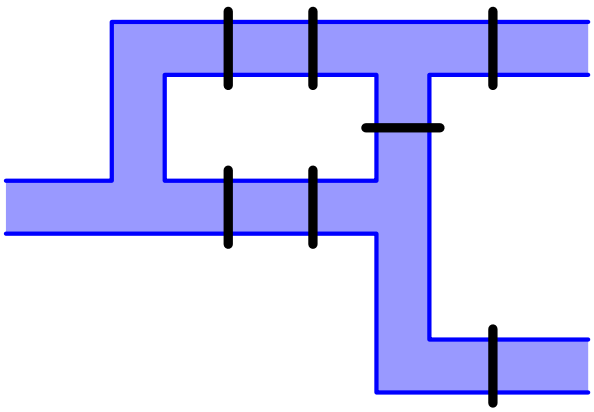
Solution In graph theory terminology, this problem translates to the following.

> Prove that given a complete graph on six vertices with each edge coloured red or blue, there exists a monochromatic[2] triangle.

Consider a random partygoer A. Of the five edges incident to A, the pigeonhole principle guarantees that at least three of them are the same colour. Without loss of generality, let these edges be red and let them join A to the party people B, C and D. If BC is red, then triangle ABC is red. If CD is red, then triangle ACD is red. If DB is red, then triangle ADB is red. So to avoid a red triangle, the edges BC, CD and DB must all be blue, which forces triangle BCD to be blue. So we simply cannot avoid having a monochromatic triangle. □

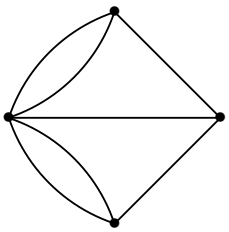
14.6 Euler trails

In the eighteenth century, the inhabitants of Königsberg, now known as Kaliningrad, liked to walk along the Pregel River. They would use the city's seven bridges to cross over to one of the two islands in the river or to the other bank of the river. And they often wondered whether or not it was possible to design a walking tour which crossed each bridge exactly once. Here is a map of the situation.



The people tried in vain for many years until the great mathematician Leonhard Euler proved that it was impossible. His first move was to reduce the map to a graph, with each vertex representing a land mass and each edge a bridge as in the following diagram.

[2]The word *monochromatic* simply means one-coloured.



In this section only, we will allow a graph to have loops and multiple edges. So what we would like to know is whether it's possible to walk around the graph, traversing every edge exactly once. Such a walk is known as an *Euler trail*.

Problem Prove that the Königsberg graph has no Euler trail.

Solution Let's suppose that the Königsberg graph has an Euler trail and hope for a contradiction.

We'll start with the obvious fact that it must start somewhere and end somewhere. Between the start and end, each time we visit a vertex, we must have walked along two edges incident to it, one going in and one coming out. Thus every vertex other than the start and end must have even degree. So if there exists an Euler trail, then there can be at most two vertices of odd degree. But you can see for yourself that the graph has four vertices of odd degree, a contradiction. □

This proof shows that, if a graph has an Euler trail, then it must have at most two vertices of odd degree. However, by the handshaking lemma, we know that any graph has an even number of vertices of odd degree. So for a graph to have an Euler trail, the number of vertices of odd degree must be either 0 or 2. It turns out that the converse of this statement is true as long as the graph is connected.

Problem Prove that a graph has an Euler trail if and only if it is connected and has 0 or 2 vertices of odd degree.

Nailing down the details of this proof requires quite a bit of work. We will only present the main ideas of the proof and leave it up to you, dear reader, to sort out all of the details!

Solution Suppose that a connected graph has 0 or 2 vertices of odd degree. If there are 2 vertices of odd degree, call one of them S and one of them F. If there are 0 vertices of odd degree, pick any vertex and call it both S and F. We will prove that there exists an Euler trail which starts at S and finishes at F.

We will proceed by strong induction on the number of edges.

For a connected graph with one edge, it's easy to find an Euler trail!

Now assume that we have a graph with E edges and that the result is true for any graph with fewer than E edges. The idea is to start at S and commence walking randomly, never traversing the same edge twice. Since there are only finitely many edges, sooner or later you will get stuck somewhere. It's impossible to get stuck at a vertex with even degree, because each time we can walk in, there will always be an edge along which we can walk out. In fact, it turns out (and you should think about why) that the only place we can get stuck is at F.

If our walk so far has traversed every edge of the graph, then we are done! Otherwise, we've missed some of the edges and these remaining edges must form a number of smaller connected

graphs $G_1, G_2, \ldots, G_n$. Each of these components G_k must have all its vertices of even degree (think carefully about why this must be so). And each such G_k intersects our previously constructed walk at some vertex v_k. Furthermore, each G_k has fewer than E edges, so the inductive hypothesis guarantees that there is an Euler trail for each G_k which starts and finishes at v_k.

Now we modify our original walk as follows. Start the same way but for each k whenever you reach one of the vertices v_k, take a detour along the Euler trail for G_k. Our new walk now traverses all the edges from the original walk, along with all the edges from the G_k. So it's an Euler trail for the original graph. □

That was quite a difficult argument, so take some time to think about it carefully. The idea of the random walk which cannot be stopped turns out to be a very useful one in graph theory.

14.7 Paths

A *walk* is any sequence of vertices $v_1, v_2, \ldots, v_n$ with v_1 adjacent to v_2, v_2 adjacent to v_3, and so on, with v_{n-1} adjacent to v_n.

A *path* is any walk with distinct vertices.

Problem In Eulerland, there are 100 cities and two airlines, Air Gauss and Air Jordan. For any two cities in Eulerland, exactly one of the companies provides direct flights in both directions between them. It's known that there are two cities a and b such that it is impossible to travel from a to b using only Air Jordan flights.

Prove that it's possible to travel between any two cities in Eulerland using only Air Gauss flights.

Solution We can use vertices to represent cities and edges to represent flights, but how can we represent the two airlines? Simple! We use red edges to represent Air Gauss and blue edges to represent Air Jordan. It turns out that a whole variety of graph theory problems involve colouring things in. In graph theory language, the problem translates to the following.

> Suppose that each edge of the complete graph on 100 vertices is coloured red or blue. It's known that there are two vertices a and b such that there is no red path from a to b.
>
> Prove that there is a blue path between any two vertices.

Let A denote the set of all vertices which are connected to a by a red path, including a itself. Similarly, let B denote the set of all vertices which are connected to b by a red path, including b itself. And finally, let C denote the set of remaining vertices. (It is possible that C is empty.) We have the following observations.

- Every edge between a vertex in A and a vertex in B must be blue—for otherwise, there would be a red path from a to b.
- Every edge between a vertex in A and a vertex in C must be blue—for otherwise, that vertex in C could be reached by a red path from a.
- Every edge between a vertex in B and a vertex in C must be blue—for otherwise, that vertex in C could be reached by a red path from b.

We will now prove that for two arbitrary vertices u and v, there exists a blue path between them. We've already established that there exists a blue edge, and hence a blue path, between any two vertices which are in different groups. So all that remains is to consider when u and v lie in the same group.

- If u and v lie in A, consider the blue path $u \to b \to v$.
- If u and v lie in B, consider the blue path $u \to a \to v$.
- If u and v lie in C, consider the blue path $u \to a \to v$.

Therefore, in all possible cases, there is a blue path between u and v. □

14.8 Extremal principle

In chapter 1, we discussed the many advantages of using the extremal principle, that is, considering the minimum or maximum of some value. The extremal principle is very useful in graph theory as you can see for yourself in the following problems!

Problem In the country of König, it's possible to travel by plane between any two of the cities, although you might have to take several flights, stopping at other intermediate cities along the way. By a journey, we mean a sequence of flights which never visits the same city twice. Let m be the maximum possible number of flights on a journey between two cities in König.

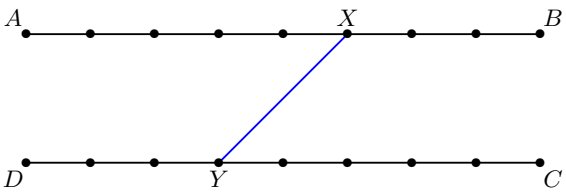
Prove that any two journeys of length m must have at least one city in common.

Solution Clearly, we can interpret cities as vertices, flights as edges, and journeys as paths. If we define the *length* of a path to be the number of edges that it traverses, then the problem can be restated as follows.

> If the longest path in a connected graph has length m, prove that any two paths of length m in the graph must share a vertex.

To obtain a contradiction, assume there exist two paths P_1 and P_2 of length m in the graph that don't share a vertex. Let P_1 join vertex A to vertex B and let P_2 join vertex C to vertex D. The aim is to find a path with length greater than m, which will then contradict the fact that m is the maximum length of a path.

How might we construct such a path? Well, since the graph is connected, there must be a path from a vertex on P_1 to a vertex on P_2. In fact, let X be a vertex on P_1 and Y a vertex on P_2 such that there exists a path from X to Y of minimal length. Using such a path of minimal length ensures that it is a path which cannot pass through any other vertices of P_1 or P_2, because otherwise it would not be of minimal length.



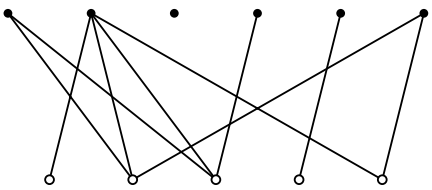
Note that either the path from A to X or the path from B to X has length at least $\frac{m}{2}$. Without loss of generality, let it be from A to X. Similarly, either the path from C to Y or the path from D to Y has length at least $\frac{m}{2}$. Without loss of generality, let it be from C to Y. Now the path from A to X to Y to C passes through no vertex twice, and has length strictly greater than m. However, this contradicts the fact that m is the maximum length of a path. Therefore, any two paths of length m in the graph must share a vertex. □

Problem There are several boys and girls at a party, where each girl dances with at least one boy but no boy dances with every girl.

Prove that there exist two boys B_1 and B_2 and two girls G_1 and G_2 such that B_1 dances with G_1, B_2 dances with G_2, B_1 does not dance with G_2 and B_2 does not dance with G_1.

It's quite difficult to attack this problem directly. As we'll soon see, the best approach is to consider the boy who dances with the most girls. Like some other proofs in this book, this may appear as surprising as a magician pulling a rabbit out of a hat! But the point, dear reader, is for you to carefully read over the proof and ask how you could have thought of it yourself. What features of the problem lead you to think about the extremal principle? And why would you apply it in this particular way? If you understand the answers to these questions, then you will soon be ready to start pulling rabbits out of your own hat!

Solution Of course, vertices will represent people while edges represent couples who dance together. It's useful to use black vertices for the boys and schematically place them at the top of the diagram and to use white vertices for the girls and schematically place them at the bottom of the diagram.



We'll start by considering one of the boys, Max say, who dances with the maximum number of girls. Now there is a girl, Anne say, who doesn't dance with Max. And Anne must dance with some boy, Bob say, who isn't Max. So if one of Max's dance partners, of whom there are maximally many, does not dance with Bob, then we're done!

But if Bob danced with all of Max's dance partners, as well as with Anne, then Bob would have danced with more girls than Max. This contradiction means that there must be some girl, Carly say, who dances with Max but not Bob.

So we can take B_1, B_2, G_1 and G_2 to be Max, Bob, Carly and Anne, respectively. □

14.9 Count and count again

The handshaking lemma asserts that, in any graph, the sum of the degrees of all the vertices is equal to twice the number of edges. One way to prove this is to consider the number of endpoints of edges. Counting these in one way, we obtain the sum of the degrees of all the vertices and counting these in another way, we obtain twice the number of edges.

We have already witnessed the virtues of double counting in section 13.8 and here, we'll see that it can be particularly useful when dealing with graphs.

Problem In a senate, there are 30 senators and each pair of them are either mutual friends or mutual enemies. Each senator has exactly 6 enemies and every group of 3 senators forms a commission.

Find the total number of commissions whose members are either all mutual friends or all mutual enemies.

Solution This problem can be paraphrased purely in graph theory terminology.

> Each edge of the complete graph K_{30} is coloured either red or blue. Each vertex is incident to exactly 6 blue edges.
>
> Find the total number of monochromatic triangles.

Let X be the number of monochromatic triangles. We define a *vee* to be a pair of edges which are incident to the same vertex. The trick here is to count *colourful vees*—that is, those consisting of one red edge and one blue edge—in two different ways.

- If we concentrate on one particular vertex, we see that it's incident to 6 blue edges and 23 red edges. This means that there are $6 \times 23 = 138$ colourful vees formed from the edges incident to one particular vertex.

  Since there are 30 vertices altogether, the total number of colourful vees is

  $$30 \times 138 = 4140.$$

- Note that there are $\binom{30}{3} = 4060$ triangles in our graph. Since X of these are monochromatic, we know that $4060 - X$ of these are not. But in a monochromatic triangle, there are no colourful vees, while in a triangle which is not monochromatic, there are two colourful vees.

  Therefore, the total number of colourful vees is

  $$2(4060 - X) = 8120 - 2X.$$

Now we simply need to equate these two results and we end up with

$$\begin{aligned} & 4140 = 8120 - 2X \\ \Rightarrow \quad & X = 1990. \end{aligned}$$

□

14.10 Planar graphs

When trying to solve graph theory problems, obviously you draw graphs. Unfortunately, the edges sometimes cross and that just looks plain ugly. It would be nice to draw our graphs without any edges crossing and, when this is possible, we call the graph *planar*.

A planar graph drawn without edges crossing will always divide the plane into regions, one of which is infinitely large, which we call *faces*. One of the most useful theorems concerning planar graphs is the following.

Euler's formula For a connected planar graph with V vertices, E edges and F faces, $V - E + F = 2$. When using Euler's formula, we always include the infinitely large face on the outside.

Problem Suppose a planar graph with E edges divides the plane into F faces.

Prove that $3F \leq 2E$.

Solution Let's count the number E of edges. Imagine cutting the planar graph along its edges. We get a collection of polygons and one figure which might be called an 'anti-polygon' corresponding to the outside face. Since each original edge splits into two edges, the total number of newly formed edges is $2E$. However, each of our F newly formed polygons, including the outside anti-polygon, has at least three edges. Thus the number of newly formed edges is at least $3F$. It follows that $3F \leq 2E$. □

Problem Prove that K_5, the complete graph on five vertices, is not planar.

Solution To obtain a contradiction, let's suppose that K_5 is planar.

Given that $V = 5$ and $E = 10$, Euler's formula tells us that if we could draw K_5 in the plane without edges crossing, then we would have $F = 7$. But from the previous problem, $2E \geq 3F$ and so $20 \geq 21$. Contradictions don't come more blatantly than that! □

Problem If a graph is planar, prove that it has at least one vertex of degree less than or equal to five.

Solution In the hope of finding a contradiction, let's start with the assumption that there exists a planar graph, all of whose vertices have degree at least six.

Suppose that this graph has V vertices, E edges and divides the plane into F faces. Then the sum of the degrees is at least $6V$, so the handshaking lemma asserts that $6V \leq 2E$ or equivalently,

$$V \leq \frac{E}{3}.$$

From the previous problem, we know that $F \leq \frac{2E}{3}$.

Substituting these two inequalities into Euler's formula, we obtain

$$2 = V - E + F \leq \frac{E}{3} - E + \frac{2E}{3} = 0,$$

an obvious contradiction. □

14.11 Polyhedra

If you take a bunch of polygons and glue them together so that no side is left unglued, then the resulting object is usually called a *polyhedron*. Typical examples include the tetrahedron, the square pyramid, the cube and the soccer ball with its pentagonal and hexagonal patches. The corners of the polygons are called *vertices*, the sides of the polygons are called *edges* and the polygons themselves are called *faces*. We say that a polyhedron is *convex* if, for each plane which lies along a face, the polyhedron lies on one side of that plane. So, for example, the cube is convex while the polyhedron formed by gluing three congruent cubes together to form an L shape is not.

The fact that every polyhedron is a graph is a rather simple statement. This is because if you have a polyhedron and simply ignore the faces, then what you have left over is just a bunch of vertices connected in pairs by edges, or in other words, a graph.

A more interesting statement is the following fact. Every convex polyhedron corresponds to a planar graph. So why is this true? Well, suppose that your polyhedron is made from some sort of rubbery material, like a balloon. If you pop the balloon by removing one of the faces, then what remains is a rubbery sheet with the vertices and edges still drawn on it. Now just stretch this out flat onto a table and there you have your planar graph. Note that this planar graph has the same number of vertices, edges and faces as the original polyhedron. The face that we removed from the polyhedron now corresponds to the infinitely large face of the planar graph. In particular, Euler's formula applies to all convex polyhedra.

Problem Suppose that you have a convex polyhedron and you are told that each face is a quadrilateral or a hexagon and that three faces meet at every vertex. Furthermore, every quadrilateral face shares an edge with four hexagonal faces, while every hexagonal face shares an edge with three quadrilateral faces and three hexagonal faces.

Deduce the number of quadrilateral faces and the number of hexagonal faces of the polyhedron.

Solution Let V be the number of vertices, E the number of edges, Q the number of quadrilateral faces, and H the number of hexagonal faces of the polyhedron. We will deduce these four values by writing down four equations that they satisfy and solving them.

- The first equation comes from applying the handshaking lemma to the polyhedron. Since three faces meet at every vertex, every vertex of the polyhedron has degree three. Therefore, the sum of the degrees is simply $3V$ and we obtain the equation
$$3V = 2E. \tag{1}$$
- The second equation comes from double counting the edges via face contributions. Each quadrilateral contributes four edges and each hexagon contributes six edges. But each edge has been double counted due to the fact that it is adjacent to two faces so we find that $4Q + 6H = 2E$, or equivalently,
$$E = 2Q + 3H. \tag{2}$$
- The third equation comes from a clever double counting argument. The trick here is to count the number of times a quadrilateral face shares an edge with a hexagonal face. This happens four times for each quadrilateral face. In other words, the number is $4Q$. Arguing in a different way, we can say that this happens three times for each hexagonal face. In other words, the number is $3H$. Of course, these numbers must be the same, so we must have
$$4Q = 3H. \tag{3}$$
- The fourth equation is simply Euler's formula. Since $F = Q + H$ we may write
$$V - E + Q + H = 2. \tag{4}$$

It is not hard to solve equations (1)–(4). For example, using (3) we have $H = \frac{4}{3}Q$. Put this into (2) to find $E = 6Q$. Put this into (1) to find $V = 4Q$. Finally put everything into (4) to obtain $Q = 6$. We may then go on to find $H = 8$, $E = 36$ and $V = 24$. □

14.12 Graph theory and inequalities

This nice intersection of mathematical topics doesn't appear too often, but is very interesting indeed.

Problem Given a graph with n vertices and E edges, prove that the number of triangles is at least

$$\frac{4E^2}{3n} - \frac{nE}{3}.$$

Solution The basic idea of the solution is rather simple and elegant. Start by labelling the vertices $v_1, v_2, \ldots, v_n$, and suppose that they have degrees $d_1, d_2, \ldots, d_n$, respectively.

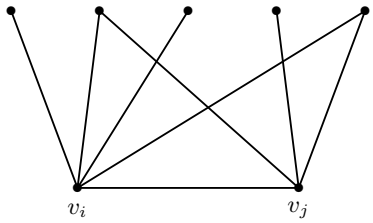
Next take two vertices v_i and v_j which are connected by an edge. Then there are $d_i - 1$ edges which emanate from v_i to the rest of the graph and $d_j - 1$ edges which emanate from v_j to the rest of the graph. Since there are $n - 2$ other vertices, then by the pigeonhole principle a triangle will be created if

$$(d_i - 1) + (d_j - 1) > n - 2.$$

In fact, we can guarantee that the edge between v_i and v_j will be involved in at least

$$(d_i - 1) + (d_j - 1) - (n - 2) = d_i + d_j - n$$

triangles.



If we sum over all possible edges, then we will count each triangle at most three times. So, letting T denote the total number of triangles, we have the inequality

$$3T \geq \sum_{\text{edges}} (d_i + d_j - n) = \sum_{\text{edges}} (d_i + d_j) - nE.$$

Since we are aiming for the result $3T \geq \frac{4E^2}{n} - nE$, all that remains to be proved is the inequality

$$\sum_{\text{edges}} (d_i + d_j) \geq \frac{4E^2}{n}.$$

Note that in this sum over edges, the number d_i appears once for each edge that is incident to the vertex v_i. That is, d_i occurs in the sum d_i times. So we can write this inequality equivalently as

$$\sum_{i=1}^{n} d_i^2 \geq \frac{4E^2}{n} = \frac{(2E)^2}{n} = \frac{1}{n}\left(\sum_{i=1}^{n} d_i\right)^2.$$

Here, we've used the handshaking lemma to express $2E$ as the sum of the degrees.

We've finally arrived at a point where the problem is reduced to algebra. One way to finish off the problem is to invoke the power means inequality[3], the QM–AM in particular.

From the power means inequality, we know

$$\sqrt{\frac{d_1^2 + d_2^2 + \cdots + d_n^2}{n}} \geq \frac{d_1 + d_2 + \cdots + d_n}{n}$$

and this is exactly what we need to complete the proof. □

[3]If you're unsure of what this means, then you might like to consult section 11.5.

15

Games and invariants

Games? Who said games? We always seem ready to play a game. Why? Because games are fun! Games often turn up in mathematics problems and if there is one important point you should remember, it is this:

Play the game!

The reason is that you gain insight into what the problem is all about. In fact you might need to play the game for quite a while before you notice something that just might be the key element in solving the problem.

15.0 Problems

1. Amy and Ben play the game of *misère noughts and crosses* on a 3×3 square array. On Amy's turn, she can place an **X** in any vacant square, while on Ben's turn, he can place an **O** in any vacant square. The players take turns to place their symbol, with Amy going first. Any player who gets three in a row (horizontally, vertically or diagonally) immediately loses the game. The game is considered drawn if there is no winner after all squares have been filled.

   (a) Which player, if any, has a winning strategy?

   (b) Answer the same question if the game is played on a three-dimensional $3 \times 3 \times 3$ cubic array.

2. Initially, there are n coins on a table and two players take turns to remove a number of them. The number of coins removed must belong to the set S and a player wins by removing the last coin.

   Find the winning and losing positions in the following cases.

   (a) $S = \{1, 2, 3, 4, 5, 6\}$

   (b) $S = \{1, 2, 3, \ldots, k\}$

   (c) $S = \{1, 3, 4\}$

   (d) $S = \{1, 3, 8\}$

   (e) $S = \{1, 2, 4, 8, 16, 32, 64, 128, \ldots\}$

   (f) $S = \{p \mid p \text{ is a prime or equal to } 1\}$

3. A 16×16 square grid is constructed from sixteen 4×4 smaller square grids called *boxes*. The game of *Ukodus* is played on the 16×16 grid as follows. Two players write, in turn, numbers from the set $\{1, 2, \ldots, 16\}$ in different squares. The numbers in each row, column and box of the 16×16 grid must be different. The loser is the one who is not able to write a number.

   Which player has a winning strategy?

4. At the start of a game, the numbers 1 and 2 are each written 10 times on a blackboard. Two players take turns to erase two of the numbers, replacing them with a 1 if they are different and with a 2 if they are the same. The first player wins if the last number on the board is 1, while the second player wins if it is 2.

   Which player has a winning strategy?

5. Sixty-three squares on an 8×8 chessboard are tiled with 3×1 rectangles.

   What are the possible locations for the square which remains untiled?

6. (a) Show that it's possible to tile an $m \times n$ chessboard with 4×1 rectangles if and only if 4 divides m or 4 divides n.

   (b) Show that it's possible to tile an $m \times n$ chessboard with $k \times 1$ rectangles if and only if k divides m or k divides n.

7. On the island of Trichroma, there are 10 blue, 15 red and 20 yellow chameleons. If two chameleons of different colours meet, they both simultaneously change to the third colour.

   Is it possible for all of the chameleons to eventually be the same colour?

8. Lex plays a game on an 8×8 chessboard which contains an **X** in the bottom-left corner. All other squares contain an **O**. Any three consecutive squares in a row or column is called a *playable set*, if either

   (i) it contains exactly two squares with an **X**, one of which must be the middle square of the playable set, or

   (ii) it contains exactly two squares with an **O**, one of which must be the middle square of the playable set.

   At any stage Lex is permitted to choose any playable set and simultaneously change every **X** into an **O**, and every **O** into an **X**.

   After a finite number of such changes, Lex has reached a configuration with exactly one square containing an **X**.

   Which squares could it be?

9. Two players start with the number 1 and take turns to multiply it by an integer from 2 to 9. The winner is the first player to obtain a number greater than or equal to 1000.

   Which player has a winning strategy?

10. A coin is placed in each square of a 4×4 grid so that all are showing heads apart from the coin in the first row and second column. You are allowed to flip all of the coins along a row, along a column, or along a line parallel to one of the diagonals. In particular, you are allowed to flip the coin in any corner square.

    Prove that it's impossible for all of the coins to show heads.

11. A positive integer is written in each square of an $m \times n$ chessboard. You are allowed to add or subtract the same integer from any two squares which share a common side, as long as the resulting numbers are both non-negative.

    When is it possible to reduce all of the numbers to zero?

12. Is it possible to fill a $7 \times 9 \times 11$ box with $3 \times 3 \times 1$ boxes?

13. The numbers 2, 3 and 6 are written on a blackboard. You are allowed to replace two of them, say a and b, with the numbers

$$\frac{3a}{5} + \frac{4b}{5} \quad \text{and} \quad \frac{4a}{5} - \frac{3b}{5}.$$

    Prove that it is impossible for a number greater than 7 to appear on the blackboard.

14. A class of students is lined up in order of height with the tallest person at the front. At any stage a student who is not in the first or second position is permitted to move two places towards the front.

    For what size class is it possible that the students could end up in the reverse order of height?

15. In the game of *Fifteen* we have 15 unit squares, numbered from 1 to 15. They are arranged in order in a 4×4 square array leaving a space of one unit square in the bottom-right corner. Any square adjacent to the space is free to slide into the space.

    (a) By using these sliding moves is it possible to swap the positions of the squares labelled 14 and 15, as shown in the diagram below?

    (b) How many different positions are achievable in this game?

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

$\xrightarrow{?}$

1	2	3	4
5	6	7	8
9	10	11	12
13	15	14	

16. A real number is written in each square of an $m \times n$ grid. If the sum of the numbers in a row or column is negative, then you are allowed to switch the signs of the numbers in that row or column.

    Prove that the sum of the numbers in each row and column will eventually be non-negative no matter in what order the rows and columns are chosen.

17. Fifty coins of various denominations lie in a row. Ollie picks up a coin from one end of the row, then Ellie picks up a coin from one end of the row of remaining coins. They alternate in this way until they each have 25 coins.

    Prove that Ollie can guarantee to win at least as much money as Ellie.

18. A deck of n cards labelled $1, 2, \ldots, n$ is shuffled. If the top card is labelled k, then the order of the top k cards is reversed.

    Prove that after finitely many of these shuffles, the card labelled 1 will eventually be on top of the deck.

19. Two players play a game involving a knight on an 8×8 chessboard. The first player places the knight on the board and the second player makes a knight's move.[1] The two players then take turns to make a knight's move, but may not place the piece on a square it has already visited.

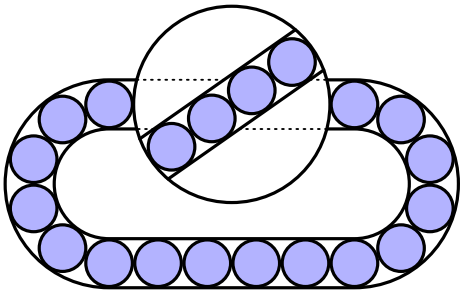
    If the player who is unable to move is considered the loser, which player has a winning strategy?

20. A game consists of circular discs numbered from 1 to 20 that are allowed to slide freely along a track, which is a closed loop. Furthermore, there is a section of the track on which exactly four discs can fit, which can be swivelled around its centre so that the order of the four discs is reversed.

    (a) From any given initial configuration, is it possible to move the discs into a configuration where all 20 numbers are in order clockwise around the track?

    (b) How many different positions are achievable in this game?

    (c) Answer the same questions if there are 21 discs instead.



21. There are three amoebas sitting at the points $(0, 0)$, $(0, 1)$ and $(1, 0)$ of the coordinate plane. Every now and then an amoeba splits into two separate amoebas, one of which will move one unit upwards, while the other moves one unit to the right. They do this in such a way that no two amoebas ever sit at the same point.

    Is it possible for the amoebas to split in such a way that the points $(0, 0)$, $(0, 1)$ and $(1, 0)$ will eventually be unoccupied?

22. The game of *Domination* is played on an $m \times n$ chessboard. Two players take turns to place a domino so that it covers two adjacent squares of the chessboard. Dominoes are not allowed to overlap and the player who cannot move loses.

    (a) Which player has a winning strategy if Domination is played on a 3×3 board?

    (b) Which player has a winning strategy if Domination is played on an $m \times n$ board, where m and n are even?

    (c) Which player has a winning strategy if Domination is played on an $m \times n$ board, where m is odd and n is even?

23. Two people play a game involving n coins on a table. The first player takes at least one, but not all, of the coins. The players then take turns to take at least one coin, but no more than was taken on the previous move. The player who takes the last coin is considered the winner.

    For which values of n does the second player have a winning strategy?

[1] A chess knight moves either two squares horizontally and one square vertically or two squares vertically and one square horizontally.

24. *Chomp* is played with an $m \times n$ grid of chocolate. Two players take turns to eat a square of chocolate, along with every square which is above and to the right of it. Unfortunately the bottom-left square is poisonous, so that the player who is forced to eat it is considered the loser.

    (a) Determine a winning strategy for the first player on a $2 \times n$ grid.

    (b) Determine a winning strategy for the first player on an $n \times n$ grid.

    (c) Show that there exists a winning strategy for the first player on any size grid apart from 1×1.

25. Initially, the number 2 is written on a blackboard. Two players take turns to erase the number N from the blackboard and replace it with the number $N + d$, where d is one of the divisors of N satisfying $0 < d < N$.

    (a) If the player who first writes a number greater than 12345 loses the game, which player has a winning strategy?

    (b) If the player who first writes a number greater than 123456 loses the game, which player has a winning strategy?

26. Write the numbers 1, 4, 5, 8, 6, 2, 3, 9 and 7 in that order, around a circle. A move consists of changing three consecutive numbers (a, b, c) to $(a + 1, b - 2, c + 1)$.

    Can you change all of the numbers to 5 using such moves?

27. Ten girls sitting around a circular table are playing a game with N cards. Initially, one girl is holding all of the cards. Each minute, if there is at least one girl holding at least two cards, one of them must pass a card to each of her two neighbours. The game ends if no girl is holding more than one card.

    (a) Prove that if $N \geq 10$, then it is impossible for the game to end.

    (b) Prove that if $N < 10$, then the game must eventually end.

28. Consider a chocolate bar in the shape of an equilateral triangle, with sides of length n, divided by grid lines into equilateral triangles of side length 1. Two players take turns to break off a triangular piece along one of the grid lines and pass the remaining block of chocolate to the other player. A player who is unable to move or who leaves an equilateral triangle of side length 1 is declared the loser.

    For which values of n does the second player have a winning strategy?

29. On an infinite grid, n^2 pieces are arranged in an $n \times n$ block of squares, one piece per square. A move consists of jumping a piece horizontally or vertically over a piece in an adjacent occupied square to an unoccupied square immediately beyond. The piece which has been jumped over is then removed.

    Find those values of n for which this game can end with only one piece remaining.

30. Consider the unit squares formed by the integer lattice points in the x-y plane. At n squares north of the x-axis is a princess whom her loyal subjects wish to rescue. In each square below the x-axis is one pawn. (These are the princess's loyal subjects.) But the pawns can only move by jumping either horizontally or vertically over another immediately adjacent pawn to a vacant square, after which the pawn that has been jumped over is removed.

    For which n is it possible that one of the princess's loyal subjects can reach her by landing on her square?

31. Amy and Ben play a game of 11-in-a-row on an infinite two-dimensional square array. They take turns to choose a vacant square and mark it. Amy goes first and marks squares with an **X**, while Ben marks squares with an **O**. A player wins by being the first to mark 11 consecutive squares vertically, horizontally or diagonally.

    (a) Show that Amy can prevent Ben from winning.

    (b) Show that Ben can prevent Amy from winning.

    (c) Show that (a) and (b) are still true if they are playing 9-in-a-row instead.[2]

[2]For the general case of n-in-a-row, it is known that the first player can force a win for $n \leq 5$ and that the second player can force a draw for $n \geq 8$. As of September 2014, the status for $n = 6, 7$ is unknown.

15.1 Number invariants

There are lots of fun problems which look something like this.

> You are given the configuration A and a set of legal moves which change the configuration. Can you use these moves to end up with the configuration B?

If the answer is *yes*, then you could prove this by simply demonstrating a sequence of legal moves which takes configuration A to configuration B. But if the answer is *no*, then you have to be much trickier. More often than not, you will need to use the idea of an invariant. An *invariant* is something—for example, a number—which we can associate to every configuration such that performing a legal move doesn't change it. So if the value of the invariant for configuration A differs from its value for configuration B, then the task is impossible to achieve, no matter how hard you try. All of this probably sounds quite cryptic and will remain so until we see some examples in action.

Problem Given some numbers, we may choose two of them, say a and b, and replace them with the single number $a + b$.

Prove that if we start with the numbers

$$1, 2, 3, \ldots, 100$$

and apply the operation 99 times, we always end up with the same final number.

Solution It should be obvious that, after applying the operation 99 times, we will be left with only one number. And it should seem intuitively clear that, no matter what order we choose to perform the additions, the final number will be the sum of the original numbers. This is certainly true, but we can state it in the language of invariants by associating to the numbers $a_1, a_2, \ldots, a_n$ the sum

$$I = a_1 + a_2 + \cdots + a_n.$$

The number I is an invariant because it doesn't change when we apply the operation. Therefore, it must be the same for both the initial and final configurations. Indeed, we have

$$1 + 2 + \cdots + 100 = 5050,$$

so the final number will always be 5050. □

Problem Given some numbers, we may choose two of them, say a and b, and replace them with the single number $ab + a + b$.

Prove that if we start with the numbers

$$1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{100}$$

and apply the operation 99 times, we always end up with the same final number.

Solution Again, it should be obvious that, after applying the operation 99 times, we will be left with only one number. Although this problem is more difficult than the previous one, it's still easily solved once you stumble upon the correct invariant. We simply associate to the numbers $a_1, a_2, \ldots, a_n$ the value

$$I = (a_1 + 1)(a_2 + 1) \cdots (a_n + 1).$$

Removing the numbers a and b from our list divides the value of I by

$$(a+1)(b+1).$$

However, adding the number $ab+a+b$ to our list multiplies the value of I by

$$ab+a+b+1=(a+1)(b+1).$$

So, the number I is an invariant because it doesn't change when we apply the operation.

The value of I for the initial configuration can be determined using the following veritable feast of cancellation.

$$I=\left(\frac{1}{1}+1\right)\left(\frac{1}{2}+1\right)\left(\frac{1}{3}+1\right)\cdots\left(\frac{1}{100}+1\right)=\frac{\cancel{2}}{1}\cdot\frac{\cancel{3}}{\cancel{2}}\cdot\frac{\cancel{4}}{\cancel{3}}\cdots\frac{101}{1\cancel{00}}=101$$

Since the invariant must be the same for both the initial and final configurations, the final number will always be 100. □

Of course, the difficulty in this solution lies in finding the actual invariant to use. A good problem will often leave clues which can guide you in the right direction. For example, an observation that might lead you to discover the correct invariant for this problem is the fact that $ab+a+b$ almost factorises as $(a+1)(b+1)-1$.

15.2 Parity

Another extremely useful invariant is *parity*, which essentially means oddness or evenness. You should always keep your eyes open for a parity argument.

Problem Given some numbers, we may choose two of them, say a and b, and replace them with the difference $|a-b|$. Suppose that we start with the numbers

$$1,2,3,\ldots,2n,$$

for some odd positive integer n.

If we apply the operation $2n-1$ times, show that we always end up with an odd number.

Solution Since we always end up with an odd number, a likely candidate for an invariant is the parity of the sum of the numbers. Initially, we have

$$1+2+3+\cdots+2n=\frac{2n(2n+1)}{2}=n(2n+1),$$

which is odd.

To show that the parity of the sum of the numbers doesn't change after each operation, suppose that we remove the numbers a and b, where we assume without loss of generality that $a\geq b$. Then the sum of the numbers would change by

$$-a-b+(a-b)=-2b,$$

which is an even number.

So the parity of the sum of the numbers is indeed an invariant. This shows that the final number must be odd. □

Problem You are given an 8×8 chessboard where the squares are coloured black and white in the usual way. You are allowed to switch the colours of all the squares in a row or column.

Can you end up with exactly one black square on the board?

Solution The first thing you should do is take out some pen and paper, draw a chessboard, and try to obtain exactly one black square. Playing around with the problem in this way should convince you, sooner or later, that you probably cannot do it. In fact, you might even be able to conjecture that it's impossible to obtain an odd number of black squares. So let's try to show that the parity of the number of black squares is an invariant.

We note that if a row or column has x black squares, then after switching the colours, it will have $8 - x$ black squares. So the change in the number of black squares is

$$-x + (8 - x) = 8 - 2x,$$

obviously an even number.

Therefore, the parity of the number of black squares is indeed an invariant and, since there are initially 32 black squares, it's impossible to end up with exactly one black square on the board. □

15.3 Modular arithmetic invariants

Parity is the same thing as considering an integer modulo 2. But in the world of invariants, sometimes modulo 2 just isn't good enough and you might need to consider integers modulo other numbers.

Problem A magic lolly machine has the property that if two of Stephen's Stupendous Smarties are inserted, then three of Justin's Jumbo Jaffas come out. Also, if three of Stephen's Stupendous Smarties are inserted, then two of Justin's Jumbo Jaffas come out. The reverse also occurs, that is, if two of Justin's Jumbo Jaffas are inserted, then three of Stephen's Stupendous Smarties come out. And if three of Justin's Jumbo Jaffas are inserted, then two of Stephen's Stupendous Smarties come out.

(a) If I want to turn two Jaffas into exactly 61 Jaffas, what is the minimum number of Smarties that I also end up with?

(b) Can I turn one Jaffa and one Smartie into 10 Jaffas and no Smarties?

Solution

(a) The first thing you should do is take out some pen and paper, pretend you have a magic lolly machine, and work out how many Jaffas and Smarties you can obtain. For example, you might come up with the following possibilities, created from just two Jaffas.

Jaffas	2	0	3	1	7	1	13	1	28	0	63	61
Smarties	0	3	1	4	0	9	1	19	1	43	1	4
Difference	2	−3	2	−3	7	−8	12	−18	27	−43	62	57

It seems that if I want to turn two Jaffas into 61 Jaffas, then I might have to create at least four Smarties.

If we're looking for an invariant, we could think about the sum of the number of Jaffas and Smarties, but this doesn't seem to be very useful.

The *difference*, on the other hand, is very interesting indeed. In fact, it seems that

$$J - S \pmod 5$$

is an invariant, where J is the number of Jaffas and S is the number of Smarties.

Let's prove this by considering the effect of one operation of the magic lolly machine. The pair (J, S) can become any of

$$(J-2, S+3), \quad (J-3, S+2), \quad (J+3, S-2) \quad \text{or} \quad (J+2, S-3).$$

In all of these cases, the difference changes from $J - S$ to $J - S \pm 5$, so we can now conclude that $J - S \pmod 5$ is indeed an invariant.

Initially, we only have two Jaffas and $J - S \equiv 2 \pmod 5$. So to end up with 61 Jaffas, we need at least four Smarties in order to maintain $J - S \equiv 2 \pmod 5$.

But be careful! While the invariant allows us to decide that certain (J, S) combinations are *not* possible, it doesn't necessarily allow us to decide which ones actually *are* possible. So we still need to show that one can obtain 61 Jaffas and four Smarties, but this we have already accomplished in the table above. □

(b) Consider the problem of turning one Jaffa and one Smartie into 10 Jaffas and no Smarties. Our invariant does not exclude this as a possibility, since $J - S \equiv 0 \pmod 5$ for both cases. However, observing that we need at least two Jaffas or at least two Smarties to use the magic lolly machine tells us that this task is impossible. □

15.4 Colouring invariants

You may have come across the following classic puzzle before. The idea behind its beautifully simple solution can be generalised to solve far more difficult problems.

Problem Consider an 8×8 chessboard, where the top-right and bottom-left squares have been removed.

Is it possible to tile this mutilated chessboard with 2×1 rectangles?

Solution The first thing you should do is take out some pen and paper, draw a mutilated chessboard, and try to tile it with 2×1 rectangles. However, I can tell you right now that you will fail, not because your tiling skills are poor, but because the task is impossible!

Perhaps surprisingly, the key to this problem is the standard black-and-white colouring of the chessboard. This is because a 2×1 rectangle will always occupy two adjacent squares on the chessboard and hence, cover one black square and one white square. Therefore, any part of the chessboard that can be tiled with 2×1 rectangles must have the same number of black and white squares.

Now we note that the standard chessboard has 32 squares of each colour, while the mutilated chessboard is obtained by removing two squares of the same colour. Since there are now 30 squares of one colour remaining and 32 squares of the other colour, it is impossible to tile the mutilated chessboard with 2×1 rectangles. □

We were lucky in this problem, because the standard 8×8 chessboard came with a colouring which helped our cause, free of charge. But sometimes, as in the next problem, you have to invent your own colouring.

Problem Show that a 10×10 chessboard cannot be tiled with 4×1 rectangles.

Solution The tactic is to find a colouring of the chessboard such that any 4×1 rectangle on the board occupies one square of each colour. Of course, this means that we require four colours, which we will call 0, 1, 2 and 3. Working along the bottom row of the chessboard, we may as well label the first four squares 0, 1, 2 and 3, in that order. After that, every square in the row must be coloured according to the repeating pattern $0, 1, 2, 3, 0, 1, 2, 3$, and so on. If we apply the same argument along the columns, we might end up with the following colouring.

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

We call this a modulo 4 colouring, because if we label the rows and columns $0, 1, 2, \ldots$, then the square in row i and column j is coloured $i + j$ modulo 4.

This colouring certainly obeys the rule that a 4×1 rectangle on the board always occupies one square of each colour. Of course, we're hoping that there are not the same number of squares of each colour. One way to verify this is to simply count them and you would indeed find that this is true. However, that is rather pedestrian, so let's use a slicker, more stylish, approach.

We simply note that it is quite easy to demonstrate a tiling of the entire board except for the 2×2 square in the top-right corner. The tiled part of the board must certainly contain the same number of squares of each colour, otherwise we wouldn't have been able to tile it. However, the remaining part of the board does not because there is one square coloured 0, two squares coloured 1, one square coloured 2 and no squares coloured 3. Hence there cannot be the same number of squares of each colour on the entire chessboard. We conclude that a 10×10 chessboard cannot be tiled with 4×1 rectangles. □

For other problems, you might need to use a modulo n colouring for some other positive integer n. Or something completely different might be needed! Using the notation (i, j) to represent the square in row i and column j, other useful colourings include the following.

- Colour (i, j) according to $i \pmod 2$. This yields a striped pattern.
- Colour (i, j) according to $(i \pmod 2, j \pmod 2)$. This uses four colours but is different from the modulo 4 colouring used in the above solution.

15.5 Monovariants

An invariant is something which doesn't change when you perform a particular move. On the other hand, a *monovariant* is a value which always gets larger or always gets smaller when you perform a particular move. For example, if you keep spending money without ever earning any, then you will never again have as much money as when you started spending. Here the monovariant is obviously the amount of money that you have. This idea is crucial to solving many problems, including the following.

Problem Given some numbers, we may choose two of them, say a and b, and replace them with the numbers

$$a + \frac{b}{2} \quad \text{and} \quad b - \frac{a}{2}.$$

If we start with a set of non-zero numbers S and keep applying the operation, show that we can never again obtain the set S.

Solution Let the numbers be $a_1, a_2, \ldots, a_n$ and consider the sum of squares

$$M = a_1^2 + a_2^2 + \cdots + a_n^2.$$

We will determine the change in M after we replace two of the numbers, say a and b.

$$\text{change in } M = \left(a + \frac{b}{2}\right)^2 + \left(b - \frac{a}{2}\right)^2 - a^2 - b^2 = \frac{a^2}{4} + \frac{b^2}{4} \geq 0$$

So M is not an invariant but a monovariant, because it never decreases. In fact, since S consists of non-zero numbers, M is guaranteed to increase after the first operation. So the same value of M can never again be obtained by applying the operation. Therefore, we can never again obtain the set S. □

For some reason, squares often feature in monovariant problems, as they did in the previous one. Next, we will see how the idea of a monovariant can help when it doesn't seem like it should.

Problem A unit fraction is a number of the form $\frac{1}{n}$, where n is a positive integer.

Prove that every rational number between 0 and 1 can be expressed as a sum of finitely many distinct unit fractions.

Solution We will show that this can be achieved using the greedy algorithm.[3] You should convince yourself that the solution to the problem follows immediately once we have proven the following statement.

> Given a rational number $0 < r < 1$, subtract the largest unit fraction less than or equal to r. Prove that, if we continue to do this, we must eventually reach the number 0 after finitely many subtractions. Furthermore, we never subtract the same unit fraction more than once.

[3] That is, take as much as you can at each stage.

The idea is to show that the numerator of the rational number is a monovariant which decreases until we reach 0.

Write the number r as

$$r = \frac{a}{b},$$

where a and b are relatively prime positive integers.

Let the largest unit fraction less than or equal to r be $\frac{1}{m}$, so that

$$\frac{1}{m} \leq \frac{a}{b} < \frac{1}{m-1}.$$

Therefore, after one step, we have the fraction

$$\frac{a}{b} - \frac{1}{m} = \frac{am-b}{bm}.$$

However, the inequality $\frac{a}{b} < \frac{1}{m-1}$ implies that $am - b < a$.

This means that the numerator of the fraction strictly decreases after each step[4], until we eventually reach the number 0.

It should be clear that we never subtract the same unit fraction more than once because if $\frac{1}{m}$ is the largest unit fraction less than or equal to r, then $r - \frac{1}{m}$ is too small to be able to subtract $\frac{1}{m}$ again. □

15.6 Invariants as cost

In our increasingly capitalistic world, we regularly think about money. So it's sometimes useful to think of an invariant as the cost of something, like we do in the following example.

Problem Initially, there is a pawn placed in each square in the bottom four rows of an 8×8 chessboard. If two pawns are in adjacent squares of the same row, you are allowed to remove them and add a pawn in the row above.

Is it possible to place a pawn in the top row of the chessboard?

Solution Let's number the rows in order so that 1 is the lowest row while 8 is the highest row. The idea behind this problem is as follows: since two pawns in row R can become one pawn in row $R+1$, we should consider the cost of a pawn in row $R+1$ to be twice the cost of a pawn in row R.

So suppose that pawns in row 1 cost \$1, pawns in row 2 cost \$2, pawns in row 3 cost \$4, and so on, so that pawns in row 8 cost \$128. The total value of the pawns initially on the chessboard is

$$8 \times (\$1 + \$2 + \$4 + \$8) = \$120.$$

Since this cost is invariant, it's impossible to place a pawn in the top row of the chessboard, which would cost \$128. □

[4]There is, of course, the possibility that $\frac{am-b}{bm}$ is not in lowest terms. However, in this case, cancelling the common factor decreases the numerator even further.

15.7 Permutation parity

A *permutation* is just a rearrangement of objects. More formally a permutation of a set S is a bijection $f\colon S \to S$.

For finite sets there is a convenient two-line notation that represents a permutation. For example, if $S = \{1, 2, 3, 4\}$, then the notation

$$\begin{pmatrix} 1\,2\,3\,4 \\ 2\,3\,1\,4 \end{pmatrix}$$

represents the permutation $f(1) = 2$, $f(2) = 3$, $f(3) = 1$ and $f(4) = 4$.

A shorter one-line notation is just to write

$$(2\,3\,1\,4)$$

to mean the same thing.

We can combine permutations by applying them successively. Thus if we applied $(2\,3\,1\,4)$ and then followed this by $(1\,2\,4\,3)$, the net result is $(2\,4\,1\,3)$. This is written as

$$(1\,2\,4\,3)(2\,3\,1\,4) = (2\,4\,1\,3) \quad \text{or} \quad \begin{pmatrix} 1\,2\,3\,4 \\ 1\,2\,4\,3 \end{pmatrix}\begin{pmatrix} 1\,2\,3\,4 \\ 2\,3\,1\,4 \end{pmatrix} = \begin{pmatrix} 1\,2\,3\,4 \\ 2\,4\,1\,3 \end{pmatrix}.$$

Note that because permutations are functions, we write the order of composition of permutations from right-to-left, as shown in the above example.

It turns out that permutations have a parity invariant which is highly useful.

Parity of a permutation The *parity* of a permutation is defined as the parity of the number of inversions in the permutation.

A pair $i < j$ is said to be an *inversion* if j occurs before i in the permutation. For example, the permutation $(2\,3\,1\,4)$ has inversions $(2, 1)$ and $(3, 1)$. Thus $(2\,3\,1\,4)$ is an *even* permutation. The permutation $(1\,2\,4\,3)$ has $(4, 3)$ as its only inversion and so it is an *odd* permutation. The composition $(2\,4\,1\,3)$ has three inversions so it is an odd permutation.

The important thing you need to know about permutation parity is the following.

Parity of compositions of permutations When composing two permutations, their parities behave in the same way as ordinary parity does for addition.

In particular we have the following.

- The composition of two even permutations is an even permutation.
- The composition of two odd permutations is an even permutation.
- The composition of an even permutation and an odd permutation is an odd permutation.

These can be proved via the following exercises which we leave for you to do.

- A *transposition* is a permutation which swaps two elements. Prove that applying a transposition to a permutation changes its parity.
- Prove that any permutation can be written as a composition of transpositions.

- Thus conclude that the parity of a permutation is simply the parity of the number of transpositions which can be used to express it.[5]

The following illustrates how permutations and parity may be used to solve a problem quickly.

Problem Andrew, Brenda and Chris are competing in a three-person race. At some point in the race Andrew is winning, with Brenda coming second and Chris third. From here on until the end of the race it is noted that there are 36 times when their relative order changes. The race ends with Brenda finishing third.

If at no point in time did all three draw level, who won the race?

Solution The original order may be notated as [A,B,C] (Andrew first, Brenda second, Chris third). Each change in the order is a transposition which is an odd permutation. Since there are 36 such permutations, the net result is an even permutation of [A,B,C].

We are given that B comes last, so the order is either [A,C,B] or [C,A,B]. But [A,C,B] is an odd permutation, whereas [C,A,B] is an even permutation. Thus the final order must be Chris first, Andrew second and Brenda third. □

15.8 Combinatorial games

For the remainder of this chapter, we consider combinatorial games, which are games that satisfy the following conditions.

- There are two players who take turns to move.
- There is no luck involved.
- The game always ends after a finite number of moves.
- Each player's moves are known to the other player.
- Either one player wins and the other loses or the game results in a draw.

The most important result concerning combinatorial games is the following.

Fundamental theorem of combinatorial games In a combinatorial game, either there exists a winning strategy for one of the players or both of them can force a draw.

Problem There are two piles on a table, one containing 1000 coins and the other containing 1001 coins. Two players take turns to remove at least one coin from one of the piles. The winner is the player who removes the last coin from the table. Since a draw is impossible, the fundamental theorem of combinatorial games guarantees that there exists a winning strategy for one of the players.

Which player has a winning strategy?

Solution The easiest way to determine which player has a winning strategy is simply to come up with a winning strategy which works!

[5] This also proves that no matter how we write a permutation as a composition of transpositions (and there are many different ways of doing this) the parity of the number of transpositions is always the same. Specifically, it is equal to the parity of the permutation.

In this game, the first player has a winning strategy which we can describe as follows: take coins from the larger pile to leave two piles of the same size.

For this strategy to work, the first player must always see two piles of different sizes on their move. This is certainly true for their first move since the piles contain 1000 and 1001 coins. Subsequently, the second player is always faced with two piles of the same size which they must turn into two piles of different sizes. So the first player continues to see two piles of different sizes when it is their turn to move.

But we still have to show why the second player has no chance of winning against this strategy. The reason for this is that, as we have already mentioned, the second player always leaves two piles of different sizes. However, to win the game, you must leave two piles of equal size, both with zero coins. Since the second player can never win the game as long as the first player follows this plan, what we have described is a winning strategy for the first player. □

15.9 Position analysis

Often, a winning strategy doesn't present itself very easily but can still be found by using a technique called *position analysis*. If you are about to move and you have a strategy which allows you to win, then we call the current position of the game a *winning position*. On the other hand, if you are about to move and you do not have a strategy which allows you to win, then we call the current position of the game a *losing position*. One corollary of the fundamental theorem of combinatorial games is the following extremely useful result.

Theorem In a combinatorial game with no draws, every position of the game can be categorised as winning or losing. Furthermore, it must be the case that

- from a winning position, it is *possible* to move to a losing position, and
- from a losing position, it is *impossible* to move to a winning position.

The idea of position analysis is to use this result to analyse enough small cases until we see can some general patterns. If all goes well, then we should be able to describe all of the winning positions, all of the losing positions, and perhaps even a winning strategy.

Problem Initially, there are n coins on a table and two players take turns to remove 1, 2, 3 or 4 coins. A player wins if he removes the last coin.

Find the winning and losing positions.

Solution In this problem, we can describe the position of the game by the number of coins on the table. It is clear that 1, 2, 3 and 4 are all winning positions because we can simply remove all of the coins on the table if it is our move.

But what happens when there are 5 coins on the table? Well, the only possibilities are for us to leave 1, 2, 3 or 4 coins on the table, thereby leaving our opponent in a winning position. So 5 must be a losing position.

But if 5 is a losing position, then 6, 7, 8 and 9 must all be winning positions. That is because from these positions, we can remove 1, 2, 3 or 4 coins to leave 5 coins on the table, thereby leaving our opponent in a losing position.

Continuing this argument gives us the following table.

Winning	1, 2, 3, 4, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18, 19, ...
Losing	5, 10, 15, 20, ...

At this stage, it seems like a safe bet that the losing positions are simply the multiples of 5. To prove that this is true, all we need to do is demonstrate that

- from a non-multiple of 5, it is possible to move to a multiple of 5, and
- from a multiple of 5, it is impossible to move to a multiple of 5.

Both of these statements are obvious! For the first, we use the fact that a non-multiple of 5 must be of the form $5k+1, 5k+2, 5k+3$ or $5k+4$ and subtracting 1, 2, 3 or 4 from each of these, respectively, leaves a multiple of 5.

For the second, we note that the difference between two multiples of 5 is still a multiple of 5 and certainly cannot be 1, 2, 3 or 4. □

Note that not only have we determined the winning and losing positions, we have also uncovered a simple winning strategy: always reduce the number of coins to a multiple of 5.

15.10 The copycat strategy

The notion of symmetry is a most powerful one in mathematics. Here, we will see how to use it to beat your friends at certain games, simply by being a copycat.

Problem Next to a square table is a pile of circular coins, all the same size. Two players take turns putting a coin on the table so that it doesn't touch any other coin. The player who cannot do so loses the game.

Show that the first player can always win.

Solution The first player can win by placing the first coin at the centre of the table. Now, wherever the second player places their coin, the first player simply copies them by placing their coin symmetrically opposite.

After each of the first player's moves, the configuration of coins is symmetric under a 180° rotation of the table. Thus, whenever the second player can place a coin on the table, the first player can also do so by placing their coin symmetrically opposite.[6] So by using this copycat strategy, the first player can always win. □

15.11 Pairing strategies

Many games involve the movement of pieces into various positions. Sometimes, a strategy can be given by pairing up the positions so that wherever the first player can go, the second player should go in the position that it is paired with.

Problem Alice and Bob play a game on a large[7] grid where they take turns to choose a square and mark it. Alice moves first and marks squares with an **X** while Bob marks squares

[6] It is possible that a coin may overlap a symmetrically opposite coin. But this occurs only if the coin lies over the centre of the table. This is ruled out by the first player's very first move.

[7] A 100×200 grid is a nice large grid for the purposes of this problem.

with an **O**. They play until one of the players marks a row or a column of five consecutive squares, and this player wins the game. If no player marks a row or column of five consecutive squares, then the game is declared a draw.

Show that Bob can prevent Alice from winning.

Solution Label the board as shown in the diagram, repeating the pattern in all directions.

	1	2	3	3	1	2	3	3	
	1	2	4	4	1	2	4	4	
	3	3	1	2	3	3	1	2	
	4	4	1	2	4	4	1	2	
	1	2	3	3	1	2	3	3	
	1	2	4	4	1	2	4	4	
	3	3	1	2	3	3	1	2	
	4	4	1	2	4	4	1	2	

Each square is paired with the neighbouring square that contains the same label.

Suppose that whenever Alice plays, Bob plays in the neighbouring square with the same label. In this way, Alice can never occupy both squares of such a pair.

But the pairing was chosen very carefully so that any block of five consecutive squares in a row or column contains such a pair. Needless to say, you should check that this is true for yourself. Hence, Alice can never mark a row or a column of five consecutive squares. □

Since Bob can prevent Alice from winning, the fundamental theorem of combinatorial games states that he must have a winning strategy or both players can force a draw. In the next section, we will use a sneaky technique to show that the latter is the case.

15.12 Strategy stealing

In the previous problem we showed that Bob could prevent Alice from winning. But as unlikely as it seems, could Bob force a win for himself? This is where strategy stealing shows its usefulness.

Problem Alice and Bob play a game on a large grid where they take turns to choose a square and mark it. Alice moves first and marks squares with an **X** while Bob marks squares with an **O**. They play until one of the players marks a row or a column of five consecutive squares, and this player wins the game. If no player marks a row or column of five consecutive squares, then the game is declared a draw.

Show that Alice can prevent Bob from winning.

Solution The main idea is that Alice's extra move at the start of the game can never hurt her chances of winning. However, this intuition doesn't constitute a proof in itself, but can be turned into one by using the concept of strategy stealing.

In order to obtain a contradiction, suppose that Bob has a winning strategy. In other words, it is possible to write a book which describes how Bob can win, no matter how Alice plays. Suppose now that Alice manages to steal this second player's strategy book and plays as follows. She simply places her first move randomly on the grid, ignores the fact that she has moved, and then pretends that she is the second player. She continues to do this until the book tells her to move in the square where she moved first. Since she has already done this, she can use this move to mark a different random square on the grid. Continuing in this way, we see that any book which provides a winning strategy for the second player can be used to provide a winning strategy for the first player! This blatantly contradicts the fundamental theorem of combinatorial games. So we must conclude that there is no winning strategy for the second player at all. In other words, Alice can prevent Bob from winning. □

Strategy stealing is brimming with trickery, so have a read through the above solution again until it's well understood. After that, you may have a look at the following example of strategy stealing at its best.

Problem In the game *Double Chess*, the rules of chess are changed so that White and Black alternately make two legal moves at a time.

Show that Black doesn't have a winning strategy.

Solution In order to obtain a contradiction, let us suppose to the contrary that Black does have a winning strategy. Then it must be the case that whatever White does on the first move, Black can win from the resulting position.

So what would happen if White started by moving a knight out and then back to the square that it was originally on? At this stage, the board looks exactly the same as it did initially, but it's now Black's turn to move. Remember that, by assumption, Black can force a win from this position. But if that is the case, couldn't White have just mirrored Black's strategy from the very beginning of the game? Of course this is possible and so White also seems to have a winning strategy. This contradicts the fundamental theorem of combinatorial games, because White and Black cannot both have winning strategies. So our original assumption must have been wrong and we conclude that Black doesn't have a winning strategy. □

16

Combinatorial geometry

Combinatorial geometry is an exotic hybrid. Usually a combinatorial problem is posed in a geometric setting. The usual ideas in combinatorics are often insufficient to take into account extra constraints that come from the geometric setting. But in this chapter we will see ideas that can be used for such problems.

16.0 Problems

1. Several points lie in the plane such that the area of the triangle formed by any three of them is no more than 1.

   Show that all the points lie in a triangle of area no more than 4.

2. Let K be a set of points in $\mathbb{R}^3$ such that every triangle with vertices in K has a side of length at most 1.

   Prove that there exist two spheres S_1 and S_2, both of radius 1, such that $K \subseteq S_1 \cup S_2$.

3. Let b, r be positive integers. Suppose we are given $2r$ red points and $2b$ blue points in the plane such that no three points are collinear.

   (a) Show there exists a line in the plane with exactly b blue points and r red points on each side.

   (b) Can you generalise this question to three dimensions?

4. Each point on the perimeter of an equilateral triangle is coloured either black or white.

   Is it always possible to find three points of the same colour which are also the vertices of a right-angled triangle?

5. Prove that any convex polygon of area 1 lies inside a rectangle of area 2.

6. Show that for all $n \geq 4$ there exists a convex hexagon which can be dissected into n congruent triangles.

7. (a) Show that for all $n \geq 4$, any cyclic quadrilateral can be dissected into n cyclic quadrilaterals.

   (b) What if the original quadrilateral is not cyclic?

8. A *Platonic solid* is a convex polyhedron in which each vertex is surrounded by the same number of congruent regular polygons.

   Prove that there are exactly five Platonic solids and determine the number of vertices, edges and faces of each.

9. (a) Let S be a set of $n \geq 3$ convex sets in the plane with empty intersection. Show that there exist three members of S with empty intersection.

   (b) Let S be a set of $n \geq 4$ convex sets in space with empty intersection. Show that there exist four members of S with empty intersection.

   (These results are known as *Helly's theorem*.)

10. There exist rectangles which can be dissected into squares of different sizes.

    Can you find an example of this using just nine squares?[1]

11. Show that it is not possible to dissect a cube into finitely many cubes of different sizes.

12. (a) Show that any polygon of n sides can be cut into triangles.[2]

    (b) Show that this can be done by cutting along just $n-3$ diagonals of the polygon resulting in $n-2$ triangles.

    (c) There are examples where we can get away with fewer than $n-2$ triangles. What is the least possible number of triangles?

    (d) Show that any polyhedron can be dissected into tetrahedra.

13. For any set of points S, we say that S *admits* distance d, if there are two points in S such that the distance between them is d.

    (a) We wish to colour every point of the plane with finitely many colours in such a way that no colour admits distance 1. Let χ be the minimal such number of colours for which this is possible.

    Show that $3 \leq \chi \leq 9$.

    (b) Can you show that $4 \leq \chi \leq 7$?[3]

    (c) If we only colour the rational points of the plane[4], show that $\chi = 2$.

14. All points in space are coloured in one of three colours.

    Prove that one of these colours admits all distances.

15. We are given n points on a circle. We join each pair of such points with a line segment forming $\binom{n}{2}$ chords. These chords divide the circle into regions.

    Find a sharp upper bound for the number of regions formed.

16. A cube of side length 30 contains 999 blue points in its interior, no four of which are coplanar. Consider also the eight vertices of the cube which are coloured red.

    Among the 1007 coloured points considered, prove that four of them, including at least one blue point, form the vertices of a tetrahedron whose volume is less than 9.

[1] This is called *squaring* a rectangle. It is also possible to square a square, but doing this requires at least 21 squares.

[2] It is the non-convex case that is particularly tricky!

[3] The number χ is known as the *chromatic number of the plane*. As of September 2014, it is still an open problem to determine the exact value of χ. The bound given here is the best known so far. As for the corresponding problem in three-dimensional space, it is known that $6 \leq \chi \leq 15$.

[4] One can also prove that $\chi = 2$ for the rational points of three-dimensional space.

17. Determine all integers $n \geq 4$ for which there exist n points $A_1, A_2, \ldots, A_n$ in the plane and real numbers $r_1, r_2, \ldots, r_n$ satisfying the following two conditions.

    (i) No three of the points $A_1, A_2, \ldots, A_n$ lie on a line.

    (ii) For each triple $1 \leq i < j < k \leq n$ we have

    $$\text{Area}(\triangle A_i A_j A_k) = r_i + r_j + r_k.$$

18. A square $ABCD$ is given. A *triangulation* of the square is a partition of the square into triangles such that any two triangles are either disjoint, share only a common vertex, or share only a common side. A *good triangulation* of the square is a triangulation in which all the triangles are acute.

    (a) Give an example of a good triangulation of the square.

    (b) What is the minimal number of triangles required for a good triangulation?

19. Find all possible sets S of $n \geq 3$ points in the plane such that every perpendicular bisector of every pair of distinct points of S is an axis of symmetry of S.

20. Each point of the coordinate plane is coloured using finitely many colours. Let O denote the origin. For each point X different from O, let $C(X)$ be the circle with centre O and radius

    $$OX + \frac{\alpha(X)}{OX},$$

    where $\alpha(X)$ is the angle measured clockwise in radians that OX makes with the positive x-axis.

    Prove there exists a point Y such that $\alpha(Y) > 0$, and the colour of Y appears on the circle $C(Y)$.

21. For any set S of five points in the plane, no three of which are collinear, let $M(S)$ and $m(S)$ denote the largest and smallest areas, respectively, of triangles determined by three points from S.

    What is the minimum possible value of $M(S)/m(S)$?

22. One is given a finite set of points in the plane, each point having integer coordinates.

    Is it always possible to colour some of the points in the set red and the remaining points white, in such a way that for any straight line L parallel to either one of the coordinate axes the difference (in absolute value) between the numbers of white points and red points on L is not greater than 1?

23. Each point in a square is coloured in one of 2005 colours.

    Prove that there exists a rectangle with vertices of the same colour.

24. Determine whether or not there exist two disjoint infinite sets A and B of points in the plane satisfying the following two conditions.

    (i) No three points of $A \cup B$ are collinear and the distance between any two is at least 1.

    (ii) There is a point of A in any triangle whose vertices are in B, and there is a point of B in any triangle whose vertices are in A.

25. We are given $n \geq 2$ distinct lines in the plane such that no two lines are parallel and such that the lines are not all concurrent through a single point.

    Prove that there exists a point through which exactly two of the lines pass.[5]

26. Let n, k be positive integers and let S be a set of n points in the plane for which no three points of S are collinear, and for every point P of S there are at least k points of S equidistant from P.

    Prove that

$$k < \frac{1}{2} + \sqrt{2n}.$$

27. Consider a square of side length a positive integer n. Suppose that there are $(n+1)^2$ points in the interior of the square.

    Show that three of these points define a (possibly degenerate) triangle of area at most $\frac{1}{2}$.

[5] This is the dual of *Sylvester's theorem.*

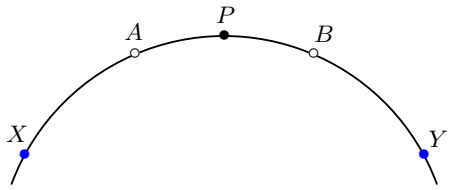
16.1 Proof by contradiction

This is one of the most basic techniques in mathematics in general. Recall from section 1.5 that the method is to show that if the proposition is false, then it follows that some nonsense is true. Of course this is a situation we cannot tolerate and so the proposition must be true.

Problem Is it possible to colour every point on a circle using the two colours white and blue so that there is no isosceles triangle whose vertices all have the same colour?

Solution After experimenting for a while you may be convinced that it is not possible. So assume for the sake of contradiction that it is possible.

Then certainly we can pick two points A and B of the same colour, say white. If X is the point on the circle such that the arc lengths XA and AB are equal, then triangle AXB is isosceles and so X must be blue. Similarly the point Y is blue where Y satisfies $AB = BY$.



Consider now the point P, say, on the minor arc of the circle halfway between A and B. Since $AP = PB$, P cannot be white. Furthermore, since $XP = YP$, point P cannot be blue.

Thus P cannot be any colour, which is a contradiction. □

This proof is not quite complete because it may occur that the points A, B, X, Y and P are not all distinct. For instance, if ABX forms an equilateral triangle, then $X = Y$.

We leave it for you to think about how to overcome this problem.

Actually, there is a really short solution to this problem as follows.

Solution Let $ABCDE$ be any regular pentagon inscribed in the circle. Then by the pigeonhole principle, three of its vertices must be the same colour.

However any three points of a regular pentagon form an isosceles triangle! □

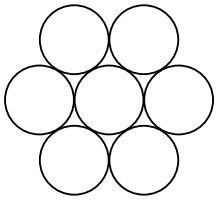
16.2 Extremal principle

Sometimes picking out something that is extremal in some sense can be just what needs to be focused on to solve a problem completely.

Problem We are given a set of discs in the plane with pairwise disjoint interiors. Each disc is tangent to at least six other discs of the family.

Prove that there are infinitely many discs in the set.

Solution Assume that the family is finite. Then there is a disc, D say, of minimal radius r. Thus there are at least six discs around D and of radius at least r. However there is only room for there to be exactly six discs, all of radius equal to r around D.



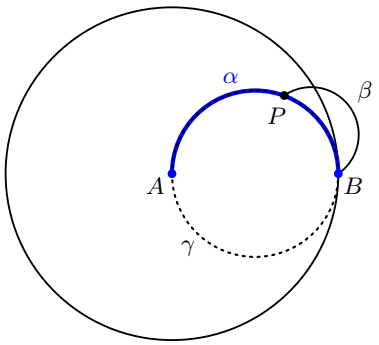
We may apply the same argument to each of these discs, thus generating infinitely many discs of radius r at ever increasing distances from D. This is a contradiction. □

Problem A closed[6] and bounded[7] shape S in the plane has the property that any two points of S can be connected by a semicircular arc which lies completely in S.

Find all possibilities for the figure S.

Solution Since S is closed and bounded, there exist points $A, B \in S$ such that the distance $d = AB$ is maximal.[8] Let α be the semicircular arc lying in S which joins A to B and let γ be the full circle defined by α.

Let P be any point on α. Then we know that there is a semicircular arc β lying in S which joins B to P. If β lies outside of γ, then there exists a point on β, whose distance from A is greater than d.[9] This is a contradiction.



Thus β lies inside γ. This is true for all points $P \in \alpha$. Since the set of such β covers the entire interior of γ, we see that the interior of γ is a subset of S.

Since S is closed, it follows that the boundary of γ is also a subset of S. Finally, since any point lying outside of γ would give rise to a distance in S greater than d, we conclude that S is in fact the closed disc described by γ. □

[6] We say that a set is *closed* if it contains its boundary. For example, the solid disc $\{(x, y) \mid x^2 + y^2 \leq 1\}$ is closed. So is the unit circle $\{(x, y) \mid x^2 + y^2 = 1\}$. But the disc $\{(x, y) \mid x^2 + y^2 < 1\}$ is not closed because it is missing some (in fact all) of its boundary, namely the unit circle.

[7] We say that a set is *bounded* if there is a real number $R > 0$ such that the distance between the origin and any point of the shape is at most R. The three examples in the previous footnote are all bounded because all points of those shapes lie within distance 1 of the origin.

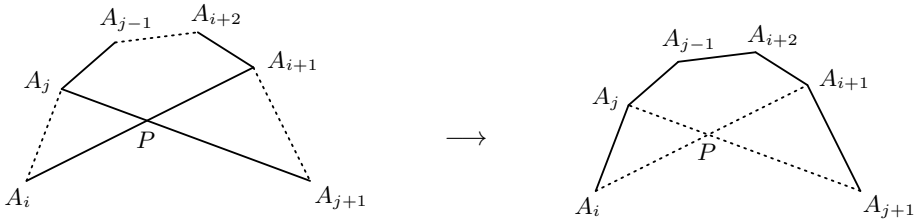
[8] The existence of two such points is a consequence of the fact that a continuous function on a compact (i.e. closed and bounded) set achieves a maximal value. In this case the compact set is $S \times S = \{(X, Y) \mid X, Y \in S\}$ and the function is $f \colon S \times S \to \mathbb{R}$, $f(X, Y) = XY$.

[9] For example, a point which lies very close to B on β would be such an offending point.

Problem Given n points in the plane, no three of which are collinear, show it is possible to join them up in sequence so that we have a broken line consisting of $n-1$ segments, no two of which cross each other.

Solution There does not seem to be anything obviously extremal to look at here. However, if you drew a really long path through the points, it seems likely that the path would contain lots of intersections. A shorter path probably would contain fewer intersections. The shortest path hopefully would contain none. Let us prove that this is indeed the case.

Let $A_1A_2\ldots A_n$ be a path of minimal length. Suppose that A_jA_{j+1} crosses A_iA_{i+1} $(i<j)$. Then consider the quadrilateral $A_iA_jA_{i+1}A_{j+1}$. The diagonals intersect at a point P inside the quadrilateral.



From the triangle inequality we have

$$A_iP + A_jP > A_iA_j \quad \text{and} \quad A_{i+1}P + A_{j+1}P > A_{i+1}A_{j+1}.$$

Adding these two equations yields

$$A_iA_{i+1} + A_jA_{j+1} > A_iA_j + A_{i+1}A_{j+1}.$$

Thus the path

$$A_1A_2\ldots A_iA_jA_{j-1}\ldots A_{i+2}A_{i+1}A_{j+1}A_{j+2}\ldots A_n$$

is of shorter length than $A_1A_2\ldots A_n$, which is a contradiction. □

16.3 Perturbation

Sometimes objects such as lines or points may be not quite in the position that you want them. A miniscule jiggle of the configuration can sometimes rectify this. We illustrate this with another way to approach the previous problem.

Problem Given n points in the plane, no three of which are collinear, show it is possible to join them up in sequence so that we have a broken line consisting of $n-1$ segments, no two of which cross each other.

Solution The points lie in the x-y plane. So if we label the points $P_1, P_2, \ldots, P_n$ according to increasing x-coordinate, then we could simply join P_1 to P_2, P_2 to P_3, and so on.

However, it may not be the case that the x-coordinates are all distinct. Surely we can rotate the configuration in the plane so that all x-coordinates are distinct! Indeed all we have to do is rotate the configuration so that the y-axis is not parallel to any of the lines formed by joining all $\binom{n}{2}$ pairs of points. □

16.4 Induction

More than likely, a problem for which we can build larger examples out of smaller examples can be approached by mathematical induction.

Problem Given 1002 distinct points in the plane, we join every pair of points with a line segment and colour its midpoint red.

Show that there are at least 2001 red points.

Solution We prove by the induction the more general statement that for $n \geq 2$ points there are at least $2n-3$ red points.

The result is clearly true for $n = 2$.

Suppose now that the result is true for $n = 2, 3, \ldots, m$ where $m \geq 2$. Consider an arrangement of $m+1$ points. By using a perturbation argument we may assume that all points have distinct x-coordinates. Label them as $A_1, A_2, \ldots, A_m, A_{m+1}$ by increasing x-coordinate. By the inductive assumption we have at least $2m-3$ red points from the midpoints of $A_1, A_2, \ldots, A_m$. Can we find two more red points by using A_{m+1}? Yes! The midpoints of $A_{m+1}A_{m-1}$ and $A_{m+1}A_m$ are distinct and both are to the right of all red points considered so far. Thus we have at least $2m-1 = 2(m+1)-3$ red points in all. This completes the induction. □

As an extension, can you determine where equality occurs in the above problem?

16.5 Discrete intermediate value theorem

The ordinary *intermediate value theorem* states that if $f(x)$ is a continuous function defined on some interval which achieves the value $f(a)$ somewhere and the value $f(b)$ somewhere else, then $f(x)$ achieves all values between $f(a)$ and $f(b)$. In particular, if $f(x)$ is positive somewhere and negative somewhere else, then it is necessarily zero somewhere in between.

There is a discrete analogue of this that involves sequences of integers.

Discrete intermediate value theorem If $a_1, a_2, \ldots, a_n$ is a sequence of integers with the property that $|a_i - a_{i+1}| \leq 1$ for $i = 1, 2, \ldots, n-1$ and is such that $a_i < 0$ and $a_j > 0$ for some $1 \leq i, j \leq n$, then $a_k = 0$ for some $1 \leq k \leq n$.

The following problem illustrates how useful this can be.

Problem We are given $2n+1$ blue points in the plane such that no three are collinear and no four are concyclic.

For every pair of blue points A, B, show that there exists a circle passing through A, B with $n-1$ blue points inside it, three points on its boundary and $n-1$ blue points outside it.

Solution Consider any two blue points A and B and orient the plane so that AB is vertical. For each point P on the perpendicular bisector of AB there is an associated circle Γ_P having centre P and passing through A and B. The idea is to consider what happens as P varies from far to the left of AB to far to the right of AB.

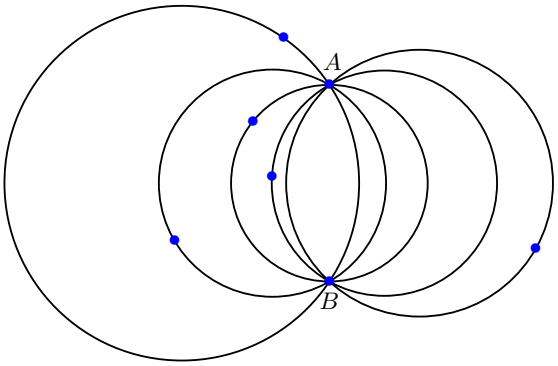
Suppose there are a blue points to the left of the line AB and b blue points to the right of the line AB. WLOG $a \geq b$. Since $a + b = 2n - 1$, a and b are of opposite parity and so we must have $a > b$.

There is a point P_{left}, such that the corresponding circle contains in its interior all of the a blue points located to the left of AB but none of the b blue points to the right of AB. There is also a point P_{right}, such that the corresponding circle contains in its interior all of the b blue points located to the right of AB but none of the a blue points to the left of AB.[10]

As P varies continuously between P_{left} and P_{right} the circle Γ_P will meet the remaining $2n-1$ blue points one by one. When Γ_P meets the ith such blue point, let a_i be given by

$$a_i = I(\Gamma_P) - O(\Gamma_P),$$

where $I(\Gamma_P)$ is the number of blue points lying inside Γ_P and $O(\Gamma_P)$ is the number of blue points lying outside Γ_P. Note that a_i is even because $I(\Gamma_P)$ and $O(\Gamma_P)$ have the same parity due to $I(\Gamma_P) + O(\Gamma_P) = 2n-2$.



If the first point that Γ_P meets is to the left of AB, then $a_1 = a-1-b$. If it is to the right of AB, then $a_1 = a-(b-1)$. Either way we have $a_1 \geq a-b-1 \geq 0$ because $a > b$. Similarly, if the last point that Γ_P meets is to the left of AB, then $a_{2n-1} = b-(a-1)$. If it is to the right of AB, then $a_{2n-1} = b-1-a$. Either way we have $a_{2n-1} \leq b-a+1 \leq 0$ because $b < a$. To summarise we have $a_1 \geq 0$ and $a_{2n-1} \leq 0$.

What happens when Γ_P goes from meeting the ith blue point to meeting the $(i+1)$th blue point?

- If both such points are to the left of AB, then $I(\Gamma_P)$ decreases by 1 while $O(\Gamma_P)$ increases by 1. Thus $a_{i+1} = a_i - 2$.
- If both points are to the right of AB, then $I(\Gamma_P)$ increases by 1 while $O(\Gamma_P)$ decreases by 1. Thus $a_{i+1} = a_i + 2$.
- If the points are on opposite sides of AB, then $I(\Gamma_P)$ and $O(\Gamma_P)$ remain unchanged. Thus $a_{i+1} = a_i$.

In all cases we have $|a_{i+1} - a_i| \leq 2$. Finally, applying the discrete intermediate value theorem to the sequence $\frac{1}{2}a_1, \frac{1}{2}a_2, \ldots, \frac{1}{2}a_{2n-1}$ guarantees that $a_i = 0$ for some i. Then the circle corresponding to this i satisfies the conclusion of the problem. □

16.6 Convex hull

Consider a wooden board with some nails hammered into it. What happens when you take a rubber band and stretch it tightly around the area occupied by the nails? You end up with a

[10] You might like to think of a reason for why P_{left} and P_{right} must exist.

polygon whose angles are all less than or equal to 180°. The area occupied by this polygon is called the *convex hull* of the nails.

The notion of convexity is defined as follows. A set S is *convex* if for any two points A and B in S, the whole line segment AB lies entirely in S. It is easily shown that the intersection of convex sets is also a convex set. The convex hull of a set T is defined to be the intersection of all convex sets containing T. It is in fact the smallest convex set containing T.

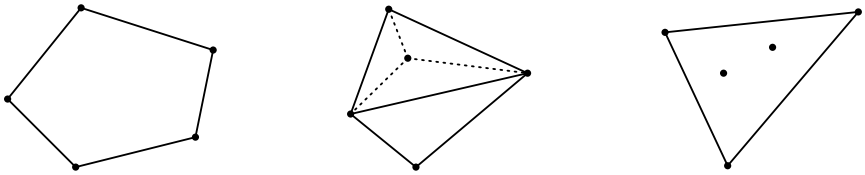
Problem Let S be a set of five distinct points in the plane.

Show there exist three points A, B and C of S such that

$$108^\circ \leq \angle ABC \leq 180^\circ.$$

Solution It is tempting to oversimplify by asserting that the five points form a pentagon and so one of its interior angles is at least 108°. Although this is true, that interior angle might also be more than 180°. This is where the convex hull becomes useful.

Consider the perimeter of the convex hull. If it contains all five points, then we may use the above argument since all interior angles are at most 180°.



If the convex hull is a quadrilateral, divide the quadrilateral into two triangles as shown. The fifth point must be inside one of those triangles. Using line segments, join that fifth point to the vertices of the triangle in which it lies. Since these segments define three angles around the fifth point, none of which exceeds 180°, and whose sum is 360°, it follows that one of those angles is between 120° and 180°.

If the convex hull is a triangle, again we find a triangle with a point (in fact two points) in its interior leading to an angle between 120° and 180°.

If the convex hull is a segment, then all points are collinear and so three of them form a 180° angle. □

16.7 Euler's formula

Euler's formula was discussed in sections 14.10 and 14.11. Recall that for convex polyhedra Euler's formula tells us

$$V - E + F = 2,$$

where V, E and F are the respective numbers of vertices, edges and faces. Recall further that this is also valid for planar graphs remembering that we count the infinitely large face on the outside as one of the faces.

Problem The square $ABCD$ contains n points $P_1, P_2, \ldots, P_n$ in its interior such that no three of the $n + 4$ points $A, B, C, D, P_1, \ldots, P_n$ are collinear.

(a) Show that it is possible to subdivide the square into triangles in such a way that the vertices of each triangle are among the $n+4$ given points.

(b) Show that the number of resulting triangles in (a) is always the same no matter how the subdivision is carried out.

Solution

(a) It is easy to establish using induction that a subdivision is always possible. Indeed each extra point would land inside some triangle. Subdividing this triangle into three further triangles by joining the interior point to the three vertices loses the original triangle but creates three smaller triangles. This construction yields $2n+2$ triangles. □

Note that such a proof by induction that *all* subdivisions as described in the problem lead to the same number of triangles is faulty. This is because not all such subdivisions can be constructed inductively in this way from subdivisions with fewer triangles.[11] So instead for part (b) we resort to Euler's formula along with a counting argument.

(b) Let T be the number of triangles. Then $F = T+1$ because of the infinitely large outside face that has four sides. Each edge belongs to two faces. Since each triangle has three edges and the outside face has four edges we have

$$E = \frac{3T+4}{2}.$$

Substituting this into Euler's formula where $V = n+4$ yields

$$n+4-\frac{3T+4}{2}+T+1=2.$$

Thus $T = 2n+2$. □

16.8 Pigeonhole principle

Problem Six points are given inside an equilateral triangle of area 4.

Prove that among the nine points which include the three vertices of the triangle and the six given points, three of these form a triangle of area at most 1.

Solution Divide the triangle up into four equilateral triangles of area 1. Since we have nine points in total, by the pigeonhole principle at least three of these points lie inside or on the boundary of one of these four triangles and thus define a triangle of area at most 1. □

In fact more is true! We can sharpen the result from 1 to $\frac{4}{13}$ as follows.

Suppose that we place one of the six points inside the triangle, and use this point to subdivide the triangle into three smaller triangles. Next place a second point of the six points. This will fall inside one of the three smaller triangles and we can use our second point to subdivide this smaller triangle into three even smaller triangles, making five triangles in all. Continuing in this fashion placing one point at a time until all six points are placed results in the original triangle being subdivided into 13 triangles.

Thus one of these triangles has area at most $\frac{4}{13}$.

[11] There are many examples of this. Can you find one?

16.9 Colouring

Questions about colouring points in the plane often use the pigeonhole principle.

Problem Let S be a disc. The points of S are painted in finitely many colours.

Show that for every $n \geq 3$ there exist infinitely many congruent polygons with n sides contained in S such that all of them have their vertices painted in the same single colour.

Solution This problem initially seems quite daunting. The number of colours and the value of n are allowed to be quite large. The best way to start is to examine simple cases. We begin with the simplest case where we have only two colours and our congruent polygons are triangles.

Consider any regular pentagon P, lying entirely in the interior of S. By the pigeonhole principle three of the vertices of P have the same colour. However, there are infinitely many disjoint translates of the set of five vertices of P which also lie in S. By the preceding argument each of these contains a monochromatic triangle. Thus we have infinitely many such monochromatic triangles.

But the triangles we are considering only come in one of two congruence types. Thus applying the infinite pigeonhole principle, one of these congruence types contains infinitely many monochromatic triangles.

We now have infinitely many congruent monochromatic triangles. Finally, a second application of the infinite pigeonhole principle allows us to conclude that one of the two colours contains infinitely many such congruent monochromatic triangles.

We leave it to the reader to work out how to solve the problem in its full generality as stated. For m colours, the issue is basically to find a number k so that whenever P has k vertices there are always n vertices of P of the same colour. □

A couple of comments are in order here. First, P did not have to be regular. In the case P is not regular, we have perhaps up to $\binom{5}{3} = 10$ different congruence types of triangle for each translate of P. But that is still finitely many. Second, why did we choose P to be a pentagon? The answer is that a pentagon has enough vertices so that if we colour them using two colours, a monochromatic triangle always appears. So P could have had any number of vertices greater than or equal to 5 and the proof would still work.

17

Appendices

17.1 How do complex numbers work?

'But $\sqrt{-1}$ is not a real number. It can't exist!' The number $i = \sqrt{-1}$ is not real in the sense that it is not on the real number line. However, we represent it geometrically by a point one unit directly above the real number 0.

Basic arithmetic

Computations with numbers involving i can be done by treating i as an unknown quantity, but replacing i^2 with -1 every time it occurs.[1]

Complex numbers include ones such as $2 - 3i$. This lies 3 units directly below the number 2. In fact considering all expressions of the form $a + bi$ $(a, b \in \mathbb{R})$, we fill up the entire plane. This is called the *complex plane*.

Addition and subtraction are very easy to do. For example,

$$(2 - 3i) - (1 - 7i) = 1 + 4i.$$

Multiplication is almost as easy. For example,

$$(2 - 3i)(1 - 7i) = 2 - 17i + 21i^2 = -19 - 17i,$$

since $i^2 = -1$.

Division is easy once you think of rationalising the denominator. For example,

$$\frac{2 - 3i}{1 - 7i} = \frac{(2 - 3i)(1 + 7i)}{(1 - 7i)(1 + 7i)} = \frac{2 + 11i - 21i^2}{1 - 49i^2} = \frac{23 + 11i}{50} = \frac{23}{50} + \frac{11}{50}i.$$

One of the amazing things about complex numbers is the *fundamental theorem of algebra*. This states that any non-constant polynomial with complex coefficients has at least one complex root.

[1] A similar thing is done when you are working with surds such as $\sqrt{2}$. You treat it as an unknown quantity except that you can replace $(\sqrt{2})^2$ with 2.

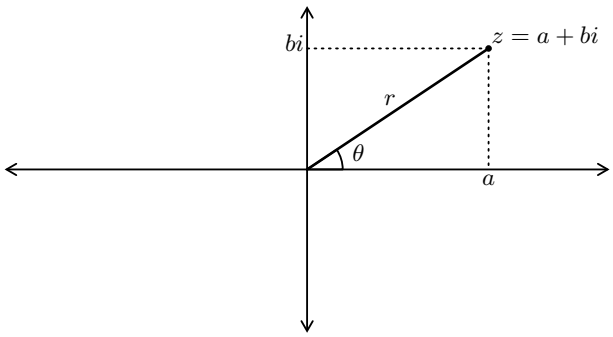
To deal with quantities like $\sqrt{-1}$ we have to go beyond the real numbers and invent i. However, for the quantities like $\sqrt{i}$ and $\sqrt{-i}$ we don't need to invent any more numbers. They are roots of the polynomials $x^2 - i = 0$ and $x^2 + i = 0$, respectively. The fundamental theorem of algebra guarantees that both these polynomials have roots. In fact you might like to check by hand that

$$\sqrt{i} = \pm\left(\frac{1+i}{\sqrt{2}}\right) \quad \text{and} \quad \sqrt{-i} = \pm\left(\frac{i-1}{\sqrt{2}}\right).$$

So far we have considered complex numbers to be of the form $a + bi$ $(a, b \in \mathbb{R})$. This is called the *Cartesian* form. The real numbers a and b basically tell us the x-coordinate and the y-coordinate of the complex number so we can plot it in the complex plane if we wish.

Polar form

There is another form that is very useful called the *polar* form. Take a non-zero complex number $z = a + bi$ in Cartesian form. Draw a line segment from the number 0 to the number z in the complex plane. Using Pythagoras' theorem, the length r of this segment is $r = \sqrt{a^2 + b^2}$ and the angle θ the segment makes when measured anticlockwise from the positive x-axis satisfies $a = r\cos\theta$ and $b = r\sin\theta$.



The elements r and θ make up the polar form and we can write

$$z = r(\cos\theta + i\sin\theta).$$

This is often abbreviated as

$$z = r\,\mathrm{cis}\,\theta.$$

The number r is called the *magnitude* and the angle θ is called the *argument* of the complex number. We use the notation $r = |z|$ and $\theta = \arg z$. Note that θ is defined modulo 360°.

A marvellous property of the polar form is that it behaves so neatly under multiplication. When two complex numbers are multiplied together, their magnitudes multiply and their arguments add. Specifically, if

$$z_1 = r_1\,\mathrm{cis}\,\theta_1 \quad \text{and} \quad z_2 = r_2\,\mathrm{cis}\,\theta_2,$$

then

$$z_1 z_2 = r_1 r_2\,\mathrm{cis}(\theta_1 + \theta_2).$$

This property is not at all obvious but it is most important. If you want to try and prove it (it's not that hard), you might find the compound angle formulas for $\sin(x+y)$ and $\cos(x+y)$ helpful.

A spectacularly amazing result is the following formula[2] due to Euler.

$$e^{i\theta} = \cos\theta + i\sin\theta$$

Here θ is measured, not in degrees, but in radians.[3] Once you know this amazing result it is very easy to prove the rule for multiplying complex numbers in polar form. In fact you can also prove the trigonometric compound angle formulas from this!

Here is an example of the power of the polar form. Suppose we are searching for complex numbers z such that

$$z^3 = 1,$$

that is, cube roots of 1. Write $z = re^{i\theta}$. Then we have

$$r^3 e^{3i\theta} = 1.$$

The polar form for 1 has $|1| = 1$ and $\arg(1) = 0$. Thus we have $r^3 = 1$ and $3\theta = 0$ and so $r = 1$ and $\theta = 0$.

However, we have missed something, namely, that the arg function is defined modulo 2π (in radians). Thus we also need to check $3\theta = \pm 2\pi, \pm 4\pi, \pm 6\pi, \ldots$. These lead to just two other values (modulo 2π) namely, $\theta = \frac{2\pi}{3}$ and $\theta = \frac{4\pi}{3}$.

So we have not one but three cube roots of 1, two of which are non-real complex numbers. If you plot all three cube roots in the complex plane, you will see that they form the vertices of an equilateral triangle inscribed in the unit circle. As a curiosity try computing their Cartesian forms.

In general if n is a positive integer, the n roots of $z^n = 1$ form the vertices of a regular n-gon inscribed in the unit circle.

[2] You might be wondering, 'How do I even compute e to the power of i?' Unfortunately, a discussion of what this even means would take us too far afield. An internet search could be most illuminating.

[3] Note that 360 degrees is equal to 2π radians.

17.2 Function notation

A few examples should suffice for you to get the hang of it. If we write

$$f\colon \mathbb{N}^+ \to \mathbb{R},$$

this means that f is a function defined for all elements of $\mathbb{N}^+$ (the set of positive integers) and taking values in the set $\mathbb{R}$ (the set of real numbers). An example of such a function is $f(x) = -x$.

The set that appears immediately after the colon ($\mathbb{N}^+$ in our example) is called the *domain* of the function.

The set that appears after the arrow ($\mathbb{R}$ in our example) is called the *codomain*.

The set of values that f actually achieves (in our case it is the set of negative integers) is called the *image*.

Note that if we tried to define the function $f\colon \mathbb{N}^+ \to \mathbb{N}^+$ satisfying $f(x) = -x$, then no such function exists! Indeed changing the domain or codomain can drastically alter things.

17.3 Directed angles

One of the most effective ways to deal with diagram dependence is through the use of *directed angles*. This often helps avoid a large number of case distinctions.

For two lines m and n, the directed angle between them is denoted by $\angle(m, n)$. It is the angle by which one may rotate m anticlockwise to obtain a line parallel to n. Directed angles have the following properties. Each of these properties is quite straightforward to prove.

- All directed angles are considered modulo $180°$. In algebraic terms this means that $\angle(m, n) = \angle(m, n) + 180°$.
- $\angle(n, m) = -\angle(m, n)$.
- For any line k we have $\angle(m, n) = \angle(m, k) + \angle(k, n)$.
- Points A, B, C are collinear if and only if $\angle(AB, AC) = 0°$.
- Points A, B, C, D are concyclic if and only if $\angle(AD, AC) = \angle(BD, BC)$.[4]
- If A, B, C are three points on a circle Γ and m is a line through A, then m is tangent to Γ if and only if $\angle(m, AB) = \angle(AC, BC)$.[5]

As an example, here is how you can prove part of the generalised pivot theorem as shown in section 6.5. We shall prove that if D, E and F lie on lines BC, AC and AB, respectively, and if P is the second intersection point of circles AEF and CDE, then P also lies on circle DBF.

Solution Since A, E, P and F are concyclic, as are C, E, P and D, we know

$$\angle(FP, FA) = \angle(EP, EA) \quad \text{and} \quad \angle(EP, EC) = \angle(DP, DC).$$

But EC and EA both define the same line, thus we may deduce that

$$\angle(FP, FA) = \angle(DP, DC).$$

It only remains to notice that FA and FB define the same line as do DC and DB. Thus

$$\angle(FP, FB) = \angle(DP, DB),$$

which means that D, B, F and P all lie on the same circle. □

Of course in reality you would have solved this question by doing an angle chase on the diagram. But in the write-up, using directed angles helps you to avoid having to deal with any case distinctions, such as what happens if P is outside of triangle ABC or if any of the points D, E or F are on the extensions of the sides of the triangle.

[4] This one is especially useful because depending on the order of A, B, C, D around the circle you either have the classic 'bow tie' configuration (angles standing on the same arc) or you have that opposite angles in a convex quadrilateral add to $180°$.

[5] This is a limiting case of the preceding property.

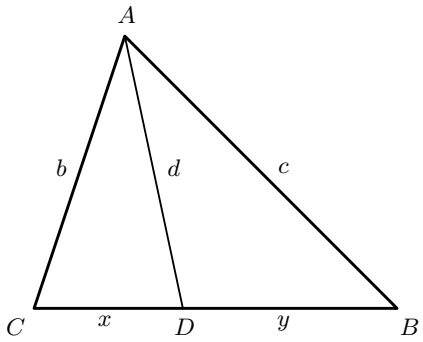
17.4 Some useful triangle formulas

In triangle ABC, let a, b, c stand for the lengths of the sides opposite A, B, C, respectively. Similarly, let α, β, γ be the measures of $\angle CAB, \angle ABC, \angle BCA$, respectively. Additionally, let $s = \frac{a+b+c}{2}$ be the semiperimeter, r the inradius, R the circumradius and $\triangle$ the area of the triangle.

Area Formulas

$$\begin{aligned}\triangle &= \frac{1}{2} \times \text{base} \times \text{height} \\ &= rs \\ &= \frac{abc}{4R} \\ &= \sqrt{s(s-a)(s-b)(s-c)} \qquad \text{(Heron's formula)} \\ &= \frac{1}{2}bc\sin\alpha.\end{aligned}$$

Formulas involving a cevian



If AD is the internal angle bisector, then we have the angle bisector theorem

$$\frac{b}{c} = \frac{x}{y}$$

as well as

$$\begin{aligned}x &= \frac{ab}{b+c} \\ y &= \frac{ac}{b+c} \\ d^2 &= bc - xy.\end{aligned}$$

If AD is the median, then we have Apollonius' theorem.

$$2d^2 + 2\left(\frac{a}{2}\right)^2 = b^2 + c^2$$

If AD is a general cevian, then we have Stewart's theorem.

$$b^2y + c^2x = a(xy + d^2)$$

Index

addendo, 112, 216
adjacent, 247
affine transformation, 134
alternate segment switch, 82
alternate segment theorem, 59
altitude, 63, 90
    extension to circumcircle, 67
AM–GM inequality, 191, 216, 222
AM–HM inequality, 194
angle bisector, 63, 76
    extension to circumcircle, 66
angle bisector theorem, 73, 215, 216, 298
angle chasing, 59
Apollonius' theorem, 298
arithmetic mean, 191, 195
associative trick, 178

base-n, 31
Beatty's theorem, 26
bicentric quadrilateral, 107
bijection, 231, 232
bijective function, 176, 177, 180
binomial identities, 229
binomial theorem, 230
bipartite graph, 250
bounded set, 286
bounding arguments, 44
butterfly theorem, 103

Carmichael number, 23, 38
case bash, 12, 44
Cauchy's functional equation, 173, 174
Cauchy–Schwarz inequality, 193, 204
centroid, 62, 91, 140
Ceva's theorem, 111, 112
    trigonometric version, 113
cevian, 66, 111, 112, 298
Chinese remainder theorem, 34, 35
chromatic number of the plane, 282
circumcentre, 63, 91
circumcircle, 66, 85, 94, 102
closed set, 286
codomain, 296
coefficient, 149
collinear points, 77, 78, 83, 86–88, 91, 93, 98, 101, 102, 108
    angle condition, 108, 297
    dilation property, 131
    Menelaus' theorem, 109
    Pascal's theorem, 122
    ratio condition, 83
colouring invariants, 270
combinatorial games, 275
    copycat strategy, 277
    pairing strategies, 277
    position analysis, 276
    strategy stealing, 278
combinatorial geometry, 281
combinatorial interpretations, 231
combinatorial reciprocal principle, 241
combinatorics, 225
    bijections, 231
    binomial identities, 229
    double counting, 235, 240
    inclusion–exclusion principle, 234
    injections, 237
    overcounting, 228

pigeonhole principle, 233
supermarket principle, 232
complete bipartite graph, 250
complete graph, 250
complete quadrilateral, 93
completely multiplicative function, 183
completing the square, 5
complex numbers, 137, 160, 293
Cartesian form, 294
comparing algebra and geometry, 137
complex plane, 293
polar form, 294
similar triangles, 144
composition of spiral symmetries, 126
concave function, 196
concurrent circles, 92, 116
concurrent lines, 88, 95, 97, 110
Ceva's theorem, 111
isogonal conjugates, 111
radical axis theorem, 118
concyclic points, 77–79, 85, 86, 90, 100, 113
angle condition, 113, 297
cyclic quadrilateral, 59, 60
power of a point, 117
Ptolemy's inequality, 216
conjugate
complex conjugate pairs, 160
surds, 27
conjugate root theorem, 160
connected graph, 249
constant term, 149
contradiction, 8, 285
contrapositive, 7
converse, 5
convex function, 196
convex hull, 289
copycat strategy, 277
cycle, 249
cyclic hexagon, 95
cyclic quadrilaterals, 59
cyclic sum notation, 200
cyclotomic polynomials, 51

dealing with digits, 25
degree
indegree, 248
of a polynomial, 149
of a vertex, 247
outdegree, 248
derangement, 235, 236
Desargues' theorem, 88
diagram dependence, 68, 108, 109, 297
difference of perfect squares, 29
dilation, 126, 131
Diophantine equation, 39
direct proof, 5
directed angle, 297
directed graph, 248
directed length, 109
discrete intermediate value theorem, 288
division algorithm, 29
polynomial, 154
divisor, 43
polynomial, 149
domain, 296
double counting, 235, 240, 255

edge, 243
Eisenstein's criterion, 151
elementary symmetric function, 157
ellipse, 119, 219
Euclid's algorithm, 29
Euler line, 91, 131
Euler phi function, 20, 36, 46
Euler trail, 251
Euler's complex exponential formula, 295
Euler's formula (graph theory), 256, 258, 290
Euler's inequality, 223
Euler's theorem (geometry), 223
Euler's theorem (number theory), 36, 46
excentre, 63, 85, 97–100, 102
excircle, 84, 98, 100
exhaustion, 12
extremal principle, 16, 254, 285

face, 256
factor
polynomial, 149
factor theorem, 154
factorisation, 43
difference of perfect powers, 163
difference of perfect squares, 29
into primes, 24
polynomials (mod p), 163
Fagnano's problem, 212
Fermat point, 120, 217, 219
Fermat's little theorem, 35
Fibonacci sequence, 2, 12, 30
fixed point, 180
floor function, 26
function
bijective, 176, 177

codomain, 296
completely multiplicative, 183
concave, 196
convex, 196
domain, 296
fixed point, 180
image, 296
injective (one-to-one), 176
inverse, 177
involution, 179
monotonic, 43
somewhere versus everywhere, 182
surjective (onto), 176
functional equation, 169
Cauchy's functional equation, 173
fundamental theorem
of algebra, 155, 160, 293
of arithmetic, 24, 183
of combinatorial games, 275
of symmetric polynomials, 157

games, 261
Gauss' lemma, 162
gcd, 29, 31, 47
gcd trick, 37, 51
generator, 37
geometric inequalities, 207
geometric mean, 191, 195
geometric transformations, 125
geometry
area, 70
combinatorial, 281
constructions, 64
diagram dependence, 108
important configurations, 75
incidence, 105
need for good diagrams, 53
plane geometry, 53
reverse reconstruction, 68
transformations, 125
trigonometry, 69
graph, 243
bipartite, 250
complete, 250
complete bipartite, 250
connected, 249
cycle, 249
directed, 248
edge, 243
Euler trail, 251
face, 256
isomorphic, 243
path, 253
planar, 256
tree, 249
vertex, 243
walk, 253
graph theory, 243
Euler's formula, 256
handshaking lemma, 247
greatest common divisor, 29, 47
greedy algorithm, 31, 272

Hölder's inequality, 204
handshaking lemma, 247, 255
harmonic mean, 194, 195
harmonic quadrilateral, 101
Helley's theorem, 282
Heron's formula, 223, 298
homogeneous inequality, 201

identity theorem, 154
if and only if, 6
image, 296
incentre, 63, 85, 97–100, 102
incidence geometry, 105
incident, 247
incircle, 63, 84, 89, 96, 98, 100, 222
incircle substitution, 198, 222
inclusion–exclusion principle, 234
indegree, 248
induction, 9, 288
inductive hypothesis, 9
strong, 11
inequalities, 187
addition and multiplication, 198
cyclic sum notation, 200
expand and conquer, 200
geometric, 207
homogeneous, 201
majorisation, 202
reverse engineering, 198
substitutions, 194, 197
symmetric sum notation, 200
useful $[i, j, k]$ notation, 203
weighted, 203
inequality
AM–GM, 191
AM–HM, 194
Cauchy–Schwarz, 193, 204
Hölder's, 204
Jensen's, 196
Muirhead's, 202
power means, 195

quadratic mean, 195
rearrangement, 191, 193
squares are non-negative, 190
triangle, 210
weighted Jensen's, 204
weighted power means, 203
infinite descent, 49
infinite pigeonhole principle, 16, 292
infinitude of primes, 8, 37
injective function, 176, 237
integer polynomial, 149
integer polynomials, 158
intermediate value theorem, 288
intersecting chords theorem, 117
invariants, 261
as cost, 273
colouring, 270
modular arithmetic, 269
monovariants, 272
number invariants, 267
parity, 268
permutation parity, 274
inverse function, 177
involution, 179, 232
irrationality of $\sqrt{2}$, 8
irreducible polynomial, 162
isogonal congugates, 111
isomorphic, 243
isoperimetric inequalities, 221

Jensen's inequality, 196

known diagram, 72, 75

Lagrange interpolation formula, 166
leading coefficient, 149
locus, 105, 118, 119, 219
logic and deduction, 5

majorisation, 202
median, 62, 94, 97
Menelaus' theorem, 109
midpoint theorem, 70, 114
mixtilinear circle, 102
modular arithmetic, 34, 46
polynomial, 45, 163, 166
modular arithmetic invariants, 269
modulo n colouring, 271
Monge's theorem, 132
monic, 50, 149
monotonicity, 43
monovariants, 272
Morley's theorem, 57
Muirhead's inequality, 202
multiplicity, 155

Newton–Gauss line, 93
nine-point centre, 91, 132
nine-point circle, 90, 114, 132, 223
number invariants, 267
number theory, 19, 39

one-to-one function, 176
onto function, 176
orthocentre, 63, 86, 90, 91
outdegree, 248
overcounting, 228

pairing strategies, 277
Pappus' theorem, 87, 122
parallel lines, 99
parametrisation, 215
parity
modulo 2 invariants, 268
permutation, 274
parity pattern, 234
Pascal's theorem, 87, 122
Pascal's triangle, 230
path, 253
perfect squares, 46
permutation
inversion, 274
parity, 274
transposition, 274
perpendicular bisector, 63, 76
perpendicularity
condition for, 81
perspective triangles, 88
perturbation, 287
pigeonhole principle, 13, 15, 16, 25, 233, 250, 291, 292
pivot theorem, 77, 116
planar graph, 256
plane geometry, 53
Platonic solid, 245, 282
polyhedron, 257
polynomial, 149
coefficient, 149
constant term, 149
cyclotomic, 51
degree, 149
division algorithm, 155
divisor, 149
factor, 149
factor theorem, 154
identity theorem, 154

integer, 149, 158
irreducible, 162
leading coefficient, 149
modular arithmetic, 45, 163, 166
monic, 50, 149
rational root theorem, 158
reducible, 162
remainder theorem, 154
root, 149, 167
symmetric, 157
unique factorisation (mod p), 164
upstairs–downstairs, 163, 165
Vieta's formulas, 156
zero, 149
polynomial modulus, 45, 166
position analysis, 276
power means inequality, 195
power of a point, 117
powers of two, 28
primes
factorisation into, 24, 183
infinitude of, 8, 37
primitive root, 37
proof, 1
by contradiction, 8, 285
by exhaustion, 12
by induction, 9
direct proof, 5
Ptolemy's inequality, 216
Pythagoras' theorem, 6

quadratic discriminant, 46, 47
quadratic equations, 46, 50
quadratic formula, 5
quadratic mean, 195
quadrilateral
bicentric, 107
complete quadrilateral, 93
conditions for incircle to exist, 89, 96
cyclic, 59
harmonic, 101
midpoints of diagonals, 93
midpoints of sides, 70, 140
Newton–Gauss line, 93

radical axis, 78, 118
bisects common tangent, 80
radical axis theorem, 78, 118
rational root theorem, 158
ratios
addendo, 112, 216
condition for collinearity, 83
rearrangement inequality, 191, 193
recursion, 238
reducible, 162
reduction of variables, 48
reflection principle, 210
remainder theorem, 154
reverse reconstruction, 68
root, 149, 167
roots of unity, 146
rotation, 126, 130, 142

similar triangles
complex numbers, 144
similar switch, 79
Simson line, 86, 109
Simson's theorem, 86, 109
spiral symmetry, 79, 125, 126, 133
complex number description, 125, 126
composition, 126
subgroups of, 126
square root, 27
squares are non-negative, 190
Stewart's theorem, 298
strategy stealing, 278
strong induction, 11
inductive hypothesis, 12
substitutions
functional equations, 175
inequalities, 194, 197
number theory, 48
polynomial, 160
supermarket principle, 232
surjective function, 176
symmedian, 94
characterisations of, 94
symmetric polynomial, 157
symmetric sum notation, 200, 202, 203
symmetry
functional equation, 179
geometry, 66

tangency, 119, 219
tangent circles, 77, 80, 94, 96, 102
telescoping, 18
Tower of Hanoi, 4
transformation geometry, 125
transformations, 125
affine, 134
dilation, 126, 131
geometric inequalities, 213
reflection, 210
rotation, 126, 130, 142

spiral symmetry, 125, 126, 133
translation, 126, 130
translation, 126, 130
tree, 249
triangle centres, 62, 85, 91, 99, 100, 102
triangle formulas, 222, 298
triangle inequality, 17, 210, 218
triangles
perspective, 88
similar, 79, 144
trigonometry, 69
geometric inequalities, 214
version of Ceva's theorem, 113
upstairs–downstairs, 163, 165

vertex, 243
Vieta jumping, 50
Vieta's formulas, 50, 156

walk, 253
weighted Jensen's inequality, 204
weighted power means inequality, 203
well-ordering, 184
WLOG, 44, 45

Zeckendorf's theorem, 12
zero, 149
zero polynomial, 154