Solving Mathematical Problems: A Personal Perspective

Terence Tao

Department of Mathematics, UCLA, Los Angeles, CA 90095

E-mail address: tao@math.ucla.edu

Dedicated to all my mentors, who taught me the meaning (and joy) of mathematics.

Contents

Foreword to the first edition

This is Terry Tao's first book. The manuscript was prepared early in 1991, when Terry was 15 years of age. We, at Deakin University, commisioned Terry to write a book on mathematical problem solving which would be suitable for use in a Deakin University course taken mainly by practising school teachers.

The brief given to Terry was to write a book that would be at least partly comprehensible to those who did not have high formal mathematical qualifications, yet would enable all readers, whatever their mathematical backgrounds, to appreciate the beauty of elegant problem-solving stragies. The outcome of Terry's effort is a work which, we are confident, will inspire teachers and students of mathematics at all levels to reflect on the obvious youthful zest, joy, yet dogged determiniation to achieve an excellent result, that characterise Terry's responses to challenging mathematical problems.

Since it was to be Terry's first book, we wanted it to be a work which, in the future, he would regard as something special. Given Terry's mathematical precocity we realised, of course, that it was likely that the book would find its place on many school and college library shelves around the world, and we wanted it to stand as a vibrant testimony to how an outstanding mathematical mind went about solving challenging mathematical problems.

Clearly, the instructions we gave Terry defined a highly problematic task. How could anyone write a book that revealed deep (yet, somewhat paradoxically, apparently simple) mathematical insights but which was simultaneously capable of being appreciated (if not fully understood) by persons without large and formal mathematical backgrounds? You, the reader, will be the judge of how well this problem is solved in this book.

The Author

The interested reader is referred to published articles on Terry Tao (Clements 1984; Gross 1986) for more complete biographical details than can be provided here. Terry was born in Adelaide in July 1975, the eldest son of Billy (a pediatrician) and Grace (an honours graduate in physics and mathematics). His parents recognized quite early that he had mathematical talent and in 1983, aged 7, he was allowed to study mathematics at a local high school. At the end of 1983 he passed the

South Australian matriculation Mathematics 1 and Mathematics 2 examinations with scores of 90% and 85% respectively, and in 1984, aged 8 years, he scored 760 on the mathematical portion of the College Board (USA) Scholastic Aptitude Test (SAT-M), a result higher than any that had been achieved by a North American child of the same age, and one that was above 99th percentile for college-bound, 12th-graders in the United States.

In 1986, 1987, and 1988, Terry obtained bronze, silver, and gold medals, respectively, for Australia in the International Mathematics Olympiad. He obtained the 'gold' during the month he turned 13, and is easily the youngest gold medal winner form any country in the history of the Olympiads. In 1989 he enrolled as a full-time student at Flinders University (in Adelaide), and in December 1990 he completed his BSc degree at Flinders, receiving a special letter of commendation from the Chancellor. In 1991 he completed a BSc honours degree in mathematics, at Flinders, and in 1992, aged 17, he commenced PhD studies in mathematics at Princeton University in the United States.

In 1986 Miraca Gross said of Terry, 'He is a delightful young boy who is aware that he is different but displays no conceit about his remarkable gifts and has an unusual ability to relate to a wide range of people, from children younger than himself to the university faculty members' (Gross 1986, p. 5). As someone who as watched Terry's development over the years, I can say that the same comment still applies today. Throughout the pages of this book you will discover an impish, yet subtle sense of humour that interacts, in an intguiging way, with an obvious and overwhelming desire to achieve the best possible solution. Julian Stanley, the Johns Hopkins University professor who, for many years, has studied mathematically precocious youngsters in the United States, was moved to write that sometimes he thought that he and his colleagues 'were learning more from Terry than he and his parents were learning from us' (Stanley 1986, p. 11). Stanley's comments are relevant in a broader educational context - this book has something to teach us all. I believe that all persons interested in mathematics, and even many who do not profess such an interest, will, by reading this book, be challenged to reflect on many general educational issues - not least of which is the question of what our schools are doing, and what they could be doing, to meet the interests and needs of those with special gifts.

References

(1) Clements, M.A. (1984), *Terence Tao*, Educational Studies in Mathematics **13**, 213–238.
(2) Gross, M. (1986), *Radical acceleration in Australia: Terence Tao*, G/C/T **9**(1), 2–9.
(3) Staney, J.C. (1986), *Insights*, G/C/T **9**(1), 10–11.

M.A. (Ken) Clements
Faculty of Education
Deakin University
December 1991

Preface to the first edition

> Proclus , an ancient Greek philosopher, said:
>
> *This therefore, is mathematics: she reminds you of the invisible forms of the soul; she gives life to her own discoveries; she awakens the mind and purifies the intellect; she brings to light our intrinsic ideas; she abolishes oblivion and ignorance which are ours by birth* . . .

But I just like mathematics because it's fun.

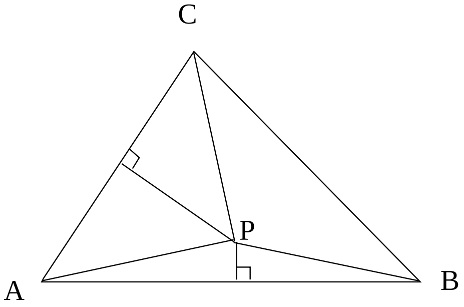
Mathematical problems, or puzzles, are important to real mathematics (like solving real-life problems), just as fables, stories and anecdotes are important to the young in understanding real life. Mathematical problems are "sanitized" mathematics, where an elegant solution has already been found (by someone else, of course), the question is stripped of all superfluousness and posed in an interesting and (hopefully) thought-provoking way. If mathematics is likened to prospecting for gold, solving a good mathematical problem is akin to a "hide-and-seek" course in gold-prospecting: you are given a nugget to find, and you know what it looks like, that it is out there somewhere, that it is not too hard to reach, that unearthing it is within your capabilities, and you have conveniently been given the right equipment (i.e. data) to get it. It may be hidden in a cunning place, but it will require ingenuity rather than digging to reach it.

In this book I shall solve selected problems from various levels and branches of mathematics. Starred problems (*) indicate an additional level of difficulty, either because some higher mathematics or some clever thinking are required; double-starred questions (**) are similar, but to a greater degree. Some problems have additional exercises at the end that can be solved in a similar manner or involve a similar piece of mathematics. While solving these problems, I will try to demonstrate some tricks of the trade when problem-solving. Two of the main weapons - experience and knowledge - are not easy to put into a book: they have to be acquired over time. But there are many simpler tricks that take less time to learn. There are ways of looking at a problem that make it easier to find a feasible attack plan. There are systematic ways of reducing a problem into successively simpler sub-problems. But, on the other hand, solving the problem is not everything. To return to the gold nugget analogy, strip-mining the neighbourhood with bulldozers is clumsier than doing a careful survey, a bit of geology, and a small amount of digging. A solution should be relatively short, understandable, and hopefully have a touch of elegance. It should also be fun to discover. Transforming a nice, short little geometry question into a ravening monster of an equation by textbook coordinate geometry doesn't have the same taste of victory as a two-line vector solution.

As an example of elegance, here is a standard result in Euclidean geometry:

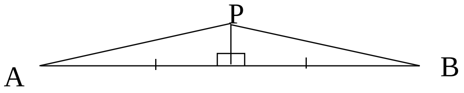
$\boxed{\text{Show that the perpendicular bisectors of a triangle are concurrent.}}$

This neat little one-liner *could* be attacked by coordinate geometry. Try to do so for a few minutes (hours?), then look at this solution:



PROOF. Call the triangle ABC. Now let P be the intersection of the perpendicular bisectors of AB and AC. Because P is on the AB bisector, $|AP| = |PB|$. Because P is on the AC bisector, $|AP| = |PC|$. Combining the two, $|BP| = |PC|$. But this means that P has to be on the BC bisector. Hence all three bisectors are concurrent. (Incidentally, P is the circumcentre of ABC.) □

The following reduced diagram shows why $|AP| = |PB|$ if P is on the AB perpendicular bisector: congruent triangles will pull it off nicely.



This kind of solution - and the strange way that obvious facts mesh to form a not-so-obvious fact - is part of the beauty of mathematics. I hope that you too will appreciate this beauty.

Acknowledgements

Thanks to Peter O'Halloran, Vern Treilibs, and Lenny Ng for their contributions of problems and advice.

Special thanks to Basil Rennie for corrections and ingenious shortcuts in solutions, and finally thanks to my family for support, encouragement, spelling corrections, and put-downs when I was behind schedule.

Almost all of the problems in this book come from published collections of problem sets for mathematics competitions. These are sourced in the texts, with full details given in the reference section of the book. I also used a small handful of problems from friends or from various mathematical publications; these have no source listed.

Preface to the second edition

This book was written fifteen years ago; literally half a lifetime ago, for me. In the intervening years, I have left home, moved to a different country, gone to graduate school, taught classes, written research papers, advised graduate students, married my wife, and had a son. Clearly, my perspective on life and on mathematics is different now than it was when I was fifteen; I have not been involved in problem-solving competitions for a very long time now, and if I were to write a book now on the subject it would be very different from the one you are reading here.

Mathematics is a multifaceted subject, and our experience and appreciation of it changes with time and experience. As a primary school student, I was drawn to mathematics by the abstract beauty of formal manipulation, and the remarkable ability to repeatedly use simple rules to achieve non-trivial answers. As a high-school student, competing in mathematics competitions, I enjoyed mathematics as a sport, taking cleverly designed mathematical puzzle problems (such as those in this book) and searching for the right "trick" that would unlock each one. As an undergraduate, I was awed by my first glimpses of the rich, deep, and fascinating theories and structures which lie at the core of modern mathematics today. As a graduate student, I learnt the pride of having one's own research project, and the unique satisfaction that comes from creating an original argument that resolved a previously open question. Upon starting my career as a professional research mathematician, I began to see the intuition and motivation that lay behind the theories and problems of modern mathematics, and was delighted when realizing how even very complex and deep results are often at heart be guided by very simple, even common-sensical, principles. The "Aha!" experience of grasping one of these principles, and suddenly seeing how it illuminates and informs a large body of mathematics, is a truly remarkable one. And there are yet more aspects of mathematics to discover; it is only recently for me that I have grasped enough fields of mathematics to begin to get a sense of the endeavour of modern mathematics as a unified subject, and how it connects to the sciences and other disciplines.

As I wrote this book before my professional mathematics career, many of these insights and experiences were not available to me, and so the writing here is when I wrote this book, and so in many places the exposition has a certain innocence, or even naivete. I have been reluctant to tamper too much with this, as my younger self was almost certainly more attuned to the world of the high-school problem solver than I am now. However, I have made a number of organizational changes, arranging the material into what I believe is a more logical order, and editing those parts of the text which were inaccurate, badly worded, confusing, or unfocused. I have also added some more exercises. In some places, the text is a bit dated (Fermat's last theorem, for instance, has now been proved rigourously), and I now realize that several of the problems here could be handled more quickly and cleanly by more "high-tech" mathematical tools; but the point of this text is not to present the slickest solution to a problem or to provide the most up-to-date survey of results, but rather to show how one approaches a mathematical problem for the first time, and how the painstaking, systematic experience of trying some ideas, eliminating others, and steadily manipulating the problem can lead, ultimately, to a satisfying solution.

I am greatly indebted to Tony Gardiner for encouraging and supporting the reprinting of this book, and to my parents for all their support over the years. I am also touched by all the friends and acquaintances I have met over the years who had read the first edition of the book. Last, but not least, I owe a special debt to my parents and the Flinders Medical Centre computer support unit for retrieving a fifteen-year old electronic copy of this book from our venerable Macintosh Plus computer!

Terence Tao
Department of Mathematics,
University of California, Los Angeles
December 2005

CHAPTER 1

Strategies in problem solving

The journey of a thousand miles begins with one step. - Lao Tzu

Like and unlike the proverb above, the solution to a problem begins (and continues, and ends) with simple, logical steps. But as long as one steps in a firm, clear direction, with long strides and sharp vision, one would need far, far less than the millions of steps needed to journey a thousand miles. And mathematics, being abstract, has no physical constraints; one can always restart from scratch, try new avenues of attack, or backtrack at an instant's notice. One does not always have these luxuries in other forms of problem-solving (e.g. trying to go home if you are lost).

Of course, this does not necessarily make it easy; if it was easy, then this book would be substantially shorter. But it makes it possible.

There are several general strategies and perspectives to solve a problem correctly; (Polya, 1948) is a classic reference for many of these. Some of these strategies are discussed below, together with a brief illustration of how each strategy can be used on the following problem:

> PROBLEM 1.1. A triangle has its lengths in an arithmetic progression, with difference d. The area of the triangle is t. Find the lengths and angles of the triangle.

Understand the problem. What kind of problem is it? There are three main types of problems:

- "Show that ..." or "Evaluate ..." questions, in which a certain statement has to be proved true, or a certain expression has to be worked out;
- "Find a ..." or "Find all ..." questions, which requires one to find something (or everything) that satisfies certain requirements; and
- "Is there a ..." questions, which either require you to prove a statement or provide a counterexample (and thus is one of the previous two types of problem).

The type of problem is important because it determines the basic method of approach. "Show that ..." or "Evaluate ..." problems start with given data and the objective is to deduce some statement or find the value of an expression; this type of problem is generally easier than the other two types because there is a clearly visible objective, one that can be deliberately approached. "Find a . . . " questions are more hit-and-miss; generally one has to guess one answer that nearly works, and then tweak it a bit to make it more correct; or alternatively one can alter the requirements that the object-to-find must satisfy, so that they are easier to satisfy. "Is there a . . . " problems are typically the hardest, because one first must make a decision on whether an object exists or not, and provide a proof on one hand, or a counter-example on the other.

Of course, not all questions fall into these neat categories; but the general format of any question will still show the basic idea to pursue when solving a problem. For example, if one tries to solve the problem "Find a hotel in this city to sleep in for the night", one should alter the requirements to, say "Find a vacant hotel within 5 kilometres with a room that costs less than 100$ a night" and then use pure elimination. This is a better strategy than proving that such a hotel does or does not exist, and is probably a better strategy than picking any handy hotel and trying to prove that one can sleep in it.

In Problem 1.1 question, we have an "Evaluate. . . " type of problem. We need to find several unknowns, given other variables. This suggests an algebraic solution rather than a geometric one, with a lot of equations connecting d, t, and the sides and angles of the triangle, and eventually solving for our unknowns.

Understand the data. What is given in the problem? Usually, a question talks about a number of objects which satisfy some special requirements. To understand the data, one needs to see how the objects and requirements react to each other. This is important in focusing attention on the proper techniques and notation to handle the problem. For example, in our sample question, our data are a triangle, the area of the triangle, and the fact that the sides are in an arithmetic progression with separation d. Because we have a triangle, and are considering the sides and area of it, we would need theorems relating sides, angles, and areas to tackle the question: the sine rule, cosine rule, and the area formulas, for example. Also, we are dealing with an arithmetic progression, so we would need some notation to account for that; for example, the side lengths could be a, $a + d$, and $a + 2d$.

Understand the objective. What do we want? One may need to find an object, prove a statement, determine the existence of a object with special properties, or whatever. Like the flip side of this strategy, "Understand the data", knowing the objective helps focus attention on the best weapons to use. Knowing the objective also helps in creating tactical goals which we know will bring us closer to solving the question. Our example question has the objective of "Find all the sides and angles of the triangle". This means, as mentioned before, that we will need theorems and results concerning sides and angles. It also gives us the tactical goal of "find equations involving the sides and angles of the triangle".

Select good notation. Now that we have our data and objective, we must represent it in an efficient way, so that the data and objective are both represented as simply as possible. This usually involves the thoughts of the past two strategies. In our sample question, we are already thinking of equations involving d, t, and the sides and angles of the triangle. We need to express the sides and angles in terms of variables: one could choose the sides to be a, b, and c, while the angles could be denoted α, β, γ. But we can use the data to simplify the notation: we know that the sides are in arithmetic progression, so instead of a, b, and c, we can have a, $a + d$, and $a + 2d$ instead. But the notation can be even better if we make it more symmetrical, by making the side lengths $b - d$, b, and $b + d$. The only slight drawback to this notation is that b is forced to be larger than d. But on further thought, we see that this is actually not a restriction; in fact the knowledge that b¿d is an extra piece of data for us. We can also trim the notation more, by labelling the angles α, β, and $180^\circ - \alpha - \beta$, but this is ugly and unsymmetrical - it is probably better to keep the old notation, but bearing in mind that $\alpha + \beta + \gamma = 180^\circ$.

Write down what you know in the notation selected; draw a diagram. Putting everything down on paper helps in three ways:

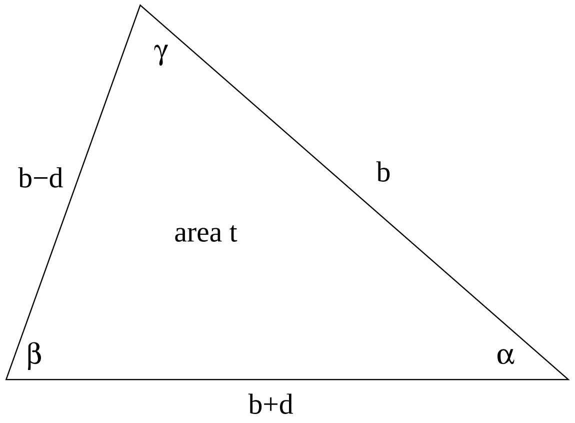
(a) You have an easy reference later on;
(b) The paper is a good thing to stare at when you're stuck; and
(c) The physical act of writing down of what you know can trigger new inspirations and connections.

Be careful, though, of writing superfluous material, and do not overload your paper with minutiae; one compromise is to highlight those facts which you think will be most useful, and put more questionable, redundant, or crazy ideas in another part of your scratch paper. Here are some equations and inequalities one can extract from our example question:

- (Physical constraints) $\alpha, \beta, \gamma, t > 0$ and $b \geq d$; we can also assume $d \geq 0$ without loss of generality;
- (Sum of angles in a triangle) $\alpha + \beta + \gamma = 180^\circ$;
- (Sine rule) $(b - d)/\sin\alpha = b/\sin\beta = (b + d)/\sin\gamma$;
- (Cosine rule) $b^2 = (b - d)^2 + (b + d)^2 - 2(b - d)(b + d)\cos\beta$, etc.;
- (Area formula) $t = \frac{1}{2}(b - d)b\sin\gamma = \frac{1}{2}(b - d)(b + d)\sin\beta = \frac{1}{2}b(b + d)\sin\alpha$;
- (Heron's formula) $t^2 = s(s - b + d)(s - b)(s - b - d)$, where $s = ((b - d) + b + (b + d))/2$ is the semiperimeter;
- (Triangle inequality) $b + d \leq b + (b - d)$.

Many of these facts may prove to be useless or distracting. But we can already identify some useful ones. The equalities are likely to be more useful than the inequalities, since our objective (and data) comes in the form of equalities. And Heron's formula looks especially promising, because the semiperimeter simplifies to $s = 3b/2$. So we can highlight "Heron's formula" as being likely to be useful.

We can of course also draw a picture. This is often quite helpful for geometry questions, though in this case the picture doesn't seem to add much:



Modify the problem slightly. There are many ways to vary a problem into one which may be easier to deal with:

(a) Consider a special case of the problem, such as extreme or degenerate cases.
(b) Solve a simplified version of the problem.
(c) Formulate a conjecture which would imply the problem, and try to prove that first.
(d) Derive some consequence of the problem, and try to prove that first.
(e) Examine solutions of similar problems.
(f) Generalize the problem.

This is useful when you can't even start a problem, because solving for a simpler related problem sometimes reveals the way to go on the main problem. Similarly, considering extreme cases and solving the problem with additional assumptions can also shed light on the general solution. But be warned that special cases are, by their nature, special, and some elegant technique could conceivably apply to them and yet have absolutely no utility in solving the general case. This tends to happen when the special case is *too* special. Start with modest assumptions first, because then you are sticking as closely as possible to the spirit of the problem.

In Problem 1.1, we can try a special case such as $d = 0$. In this case we need to find the side length of an equilateral triangle of area t. In this case, it is a standard matter to compute the answer, which is $b = 2t^{1/2}/3^{1/4}$. This gives us no clues on the general problem, except perhaps as a check. This strategy is not really appropiate for this calculation-type question, athough it can sometimes suggest what the general answer would be (in particular, we should now expect some square roots and fourth roots in the final answer). Consideration of similar problems draws little as well, except one gets further evidence that a gung-ho algebraic attack is what is needed.

Modify the problem significantly. In this more aggressive type of strategy, we perform major modifications to a problem such as removing data, swapping the data with the objective, or negating the objective (e.g. trying to disprove a statement rather than prove it). Basically, we try to push the problem until it breaks, and then try to identify where the breakdown occurred; this identifies the weak spots of the problem, as well as where the main difficulty will lie. These exercises can also help in getting an instinctive feel of what will "work", and what will probably fail.

In regard to our particular question, one could replace the triangle with a quadrilateral, circle, etc. Not much help there: the problem just gets more complicated. But on the other hand, one can see that one doesn't really need a triangle in the question, but just the dimensions of the triangle. We don't really need to know the position of the triangle. So here is further confirmation that we should concentrate on the sides and angles (i.e. $a, b, c, \alpha, \beta, \gamma$) and not on coordinate geometry, or similar approaches.

We could omit some objectives; for example instead of working out all the sides and angles we could work out just the sides, for example. But then one can notice that by the cosine and sine rules, the angles of the triangle will be determined anyway. So it is only neccesary to solve for the sides. But we know that the sides have lengths $b - d$, b, and $b + d$, so we only need to find what b is to finish the problem.

We can also omit some data, like the arithmetic difference d, but then we seem to have several possible solutions, and not enough data to solve the problem. Similarly, omitting the area t will not leave enough data to clinch a solution. (Sometimes one can *partially* omit data, for instance by only specifying that the area is larger or smaller than some threshold t_0; but this is getting complicated. Stick with the simple options first.) Reversal of the problem (swapping data with objective) leads to some interesting ideas though. Suppose you had a triangle with arithmetic difference d, and you wanted to shrink it (or whatever) until the area becomes t. One could imagine our triangle shrinking and deforming, while preserving the arithmetic difference of the sides. Similarly, one could consider all triangles with a fixed area, and mold the triangle into one with the sides in the correct arithmetic progression. These ideas could work in the long run: but I will solve this question by another approach. Don't forget, though, that a question can be solved in more than one way, and no particular way can really be judged the absolute best.

Prove results about our question. Data is there to be used, so one should pick up the data and play with it. Can it produce more meaningful data? Also, proving small results could be beneficial later on, when trying to prove the main result or to find the answer. However small the result, don't forget it - it could have bearing later on. Besides, it gives you something to do if you're stuck.

In a "Evaluate..." problem like the triangle question, this tactic is not as useful. But one can try. For example, our tactical goal is to solve for b. This depends on the two parameters d and t. In other words, b is really a function: $b = b(d, t)$.

(If this notation looks out of place in a geometry question, then that is only because geometry tends to ignore the functional dependence of objects. For example, Heron's formula gives an explicit form for the area A in terms of the sides a, b, and c: in other words, it expresses the function $A(a, b, c)$.) Now we can prove some mini-results about this function $b(d, t)$, such as $b(d, t) = b(-d, t)$ (because an arithmetic progression has an equivalent arithmetic progression with inverted arithmetic difference), or $b(kd, k^2t) = kb(d, t)$ (this is done by dilating the triangle that satisfies $b = b(d, t)$ by k). We could even try differentiate b with respect to d or t. These tactics are yet another way of attacking the problem, and I will leave these ideas for you.

Simplify, exploit data, and reach tactical goals. Now we have set up notation and have a few equations, we should seriously look at attaining our tactical goals that we have established. In simple problems, there are usually standard ways of doing this. (For example, algebraic simplification is usually discussed thoroughly in high-school level textbooks.) Generally, this part is the longest and most difficult part of the problem: however, once can avoid getting lost if one remembers the relevant theorems, the data and how they can be used, and most importantly the objective. It is also a good idea to not apply any given technique or method blindly, but to think ahead and see where one could hope such a technique to take one; this can allow one to save enormous amounts of time by eliminating unprofitable directions of inquiry before sinking lots of effort into them, and conversely to give the most promising directions priority.

In Problem 1.1, we are already concentrating on Heron's formula. We can use this to attain our tactical goal of solving for b. After all, we have already noted that the sine and cosine rules can determine α, β, γ once b is known. As further evidence that this is going to be a step forward, note that Herons formula involves d and t - in essence, it uses all our data (we have already incorporated the fact about the sides being in arithmetic progression into our notation). Anyway, Herons formula in terms of d, t, b becomes

$$t^2 = \frac{3b}{2}(\frac{3b}{2} - b + d)(\frac{3b}{2} - b)(\frac{3b}{2} - b - d)$$

which we can simplify to

$$t^2 = \frac{3b^2(b-2d)(b+2d)}{16} = \frac{3b^2(b^2-4d^2)}{16}.$$

Now we have to solve for b. The right hand side is a polynomial in b (treating d and t as constants), and in fact it is a quadratic in b^2. Now quadratics can be solved easily: if we put clear denominators and put everything on the left-hand side we get

$$3b^4 - 12d^2b^2 - 16t^2 = 0$$

so, using the quadratic formula,

$$b^2 = \frac{12d^2 \pm \sqrt{144d^4 + 196t^2}}{6} = 2d^2 \pm \sqrt{4d^2 + \frac{16}{3}t^2}.$$

Because b has to be positive, we get

$$b = \sqrt{2d^2 + \sqrt{4d^4 + \frac{16}{3}t^2}};$$

as a check, we can verify that when $d = 0$ this agrees with our previous computation of $b = 2t^{1/2}/3^{1/4}$. Once we ompute the sides $b - d, b, b + d$, the evaluation of the angles α, β, γ then follows from the cosine laws, and we are done!

CHAPTER 2

Examples in number theory

> *There is divinity in odd numbers, either in nativity, chance, or death.* William Shakespeare, “The Merry Wives of Windsor”.

Number theory may not neccesarily be divine, but it still has an aura of mystique about it. Unlike algebra, which has as its backbone the laws of manipulating equations, number theory seems to derive its results from a source unknown. Take for example *Lagrange's theorem* (first conjectured by Fermat) that every positive integer is a sum of four perfect squares (e.g. $30 = 4^2 + 3^2 + 2^2 + 1^2$). Algebraically, we are talking about an extremely simple equation: but because we are restricted to the integers, the rules of algebra fail. The result is is infuriatingly innocent-looking and experimentation shows that it does seem to work, but offers no explanation why. Indeed, Lagrange's theorem cannot be easily proved by the elementary means covered in this book: an excursion into Gaussian integers or something similar is needed.

Other problems, though, are not as deep. Here are some simple examples, all involving a natural number n:

(a) n always has the same last digit as its fifth power n^5.
(b) n is a multiple of 9 if and only if the sum of its digits is a multiple of 9.
(c) (Wilson's theorem) $(n-1)!+1$ is a multiple of n if and only if n is a prime number.
(d) If k is a positive odd number, then $1^k + 2^k + \ldots + n^k$ is divisible by $n+1$.
(e) There are exactly four numbers that are n digits long (allowing for padding by zeroes) and which are exactly the same last digits as their square. For instance, the four three-digit numbers with this property are 000, 001, 625, and 876.

These statements can all be proved by elementary number theory; all revolve around the basic idea of *modular arithmetic*, which gives you the power of algebra but limited to a finite number of integers. Incidentally, trying to solve the last assertion (e) can eventually lead to the notion of *p-adics*, which is sort of an infinite-dimensional form of modular arithmetic.

Basic number theory is a pleasant backwater of mathematics. But the applications that stem from the basic concepts of integers and divisibility are amazingly diverse and powerful. The concept of divisibility leads naturally to that of *primes*, which moves into the detailed nature of factorisation and then to one of the jewels of mathematics in the last part of the previous century: the prime number theorem, which can predict the number of primes less than a given number to a good degree of accuracy. Meanwhile, the concept of integer operations lends itself to modular arithmetic, which can be generalized from a subset of the integers to the algebra of finite groups, rings, and fields, and leads to algebraic number theory, when the concept of "number" is expanded into irrational surds, splitting fields, and complex numbers. Number theory is a fundamental cornerstone which supports a sizeable chunk of mathematics. And, of course, it's fun too.

Before we begin looking at problems, let's review some basic notation. A *natural number* is a positive integer (we will not consider 0 a natural number). The set of natural numbers will be denoted $\mathbf{N}$. A *prime number* is a natural number with exactly two factors: itself and 1; we do not consider 1 to be prime. Two natural numbers m and n are *coprime* if their only common factor is 1.

The notation "$x = y \pmod{n}$", which we read as "x equals y modulo n", means that x and y differ by a multiple of n, thus for instance $15 = 65 \pmod{10}$. The notation "(mod n)" signifies that we are working in a *modular arithmetic* where the *modulus* n has been identified with 0; thus for instance modular arithmetic (mod 10) is the arithmetic in which $10 = 0$. Thus for instance we have $65 = 15 + 10 + 10 + 10 + 10 + 10 = 15 + 0 + 0 + 0 + 0 + 0 = 15 \pmod{10}$. Modular arithmetic differs also from standard arithmetic in that inequalities do not exist, and that all numbers are integers. For example, $7/2 \neq 3.5 \pmod{5}$, but rather $7/2 = 12/2 = 6 \pmod{5}$ because $7 = 12 (\mathrm{mod}\ 5)$. It may seem strange to divide in this round-about way, but in fact one can find that there is no real contradiction, although some divisions are illegal, just as division-by-zero is illegal within the traditional field of real numbers. As a general rule, division is OK if the denominator is coprime with the modulus n.

Digits

We mentioned above that one can learn something about a number (in particular, whether it is divisible by 9) by summing all its digits. In higher mathematics, it turns out that this operation is not particularly important, but it is quite popular in recreational mathematics and has even has been given mystical connotations by some! Certainly, digit summing appears fairly often in mathematics competition problems, such as this one.

Problem 2.1 (Taylor 1989, p. 7). Show that among any 18 consecutive 3-digit numbers there is at least one which is divisible by the sum of its digits.

This is a finite problem: there are only 900 or so 3-digit numbers, so theoretically we could evaluate the problem manually. But let's see if we can save ourselves some work. First of all, the objective looks a little weird: we want the a number to be divisible by the sum of digits. Let's first write down the objective as a mathematical formula, so that we can manipulate it more easily. A 3-digit number can be written in the form abc_{10} where a, b, c are the digits; we are writing abc_{10} to avoid confusion with abc; note that $abc_{10} = 100a+10b+c$, but $abc = a\times b\times c$. If we use the standard notation $a|b$ to denote the statement that a divides b, we now want to solve

$$(a+b+c)|abc_{10} \tag{1}$$

where abc_{10} are the digits of one of the 18 given consecutive numbers. Can we reduce, simplify, or somehow make usable this equation? It is possible, but it is not simplifiable to anything halfway decent (e.g. a useful equation connecting a, b, and c directly). In fact (1) is a horrendous thing to manipulate, even after one substitutes $100a + 10b + c$ for abc_{10}. Take a look at the solutions abc_{10} of (1):

$$100, 102, 108, 110, 111, 112, 114, 117, 120, 126, \ldots, 990, 999$$

They seem to be haphazard and random. However, they do seem to occur often enough so that any run of 18 consecutive numbers should have one. And what is the significance of the 18 anyway? Assuming it is not a red herring, (perhaps only 13 consecutive numbers are needed, but the 18 is there to throw you off the track) why have 18? It may occur to some that the sums of digits of a number are rather related to the number 9 (e.g. any number has the same remainder as its digit sum upon dividing by 9) and 18 is related to 9, so there could be a vague connection. Still, consecutive numbers and divisibility don't mix. It seems that we have to reformulate the question or propose a related one to have a hope of solving it.

Now that we are on the lookout for anything related to the number 9, we should notice that most numbers which actually do satisfy (1) are multiples of 9, or at least multiples of 3. In fact there are only three exceptions on the list above (100, 110 and 112), and practically all of the multiples of 9 satisfy (1). So instead of trying to prove

> For any 18 consecutive numbers, at least one solves (1).

directly, we could try something like

> For any 18 consecutive numbers, there is a multiple of 9 which solves (1).

This route seems to "break the ice" between our data (18 consecutive numbers) and the objective (A number satisfying (1)) because 18 consecutive numbers always contain a multiple of 9 (in fact they contain two such multiples), and from numerical evidence, and the heuristic properties of the number 9, it seems that multiples of 9 satisfy (1). This "stepping stone" approach is the best way to reconcile two unfriendly statements.

Now this particular stepping stone (considering multiples of 9) does work, but a bit of extra work is needed to cover all the cases. It is actually better to use multiples of 18:

$$\boxed{\text{18 consecutive numbers}} \implies \boxed{\text{a multiple of 18}} \implies \boxed{\text{a solution to (1)}}$$

The reasons for this change are twofold:

- 18 consecutive numbers will always contain exactly one multiple of 18, but they would contain two multiples of 9. It seems neater, and more appropriate, to use multiples of 18 than to use multiples of 9. After all, if we used multiples of 9 to solve the problem, the question would only need 9 consecutive numbers instead of 18.
- It should be easier to prove (1) for multiples of 18 than for multiples of 9, since multiples of 18 are nothing more than a special case of multiples of 9. Indeed, it turns out that multiples of 9 don't always work (consider for instance 909), but multiples of 18 will, as we shall see.

Anyway, experimentation shows that multiples of 18 seem to work. But why? Take, for example, 216, which is a multiple of 18. The sum of digits is 9, and 9 divides 216 because 18 divides 216. To consider another example: 882 is a multiple of 18, and the sum of digits is 18. Hence 882 is obviously divisible by its digit sum. Messing around with a few more examples shows that the sum of digits of a multiple of 18 is always 9 or 18, which divides the original number almost by default. And with these guesses a proof soon follows:

Proof. Within the 18 consecutive numbers, one must be a multiple of 18, say abc_{10}. Because abc_{10} is a multiple of 9 as well, $a + b + c$ must be a multiple of 9. (Remember the divisibility rule for 9? a number is divisible by 9 if and only if its digit sum is divisible by 9). Because $a + b + c$ ranges between 1and 27, $a + b + c$ must be 9, 18, or 27. 27 only occurs when $abc = 999$, but that is not a multiple of 18. Hence $a + b + c$ is 9 or 18, and so $a + b + c|18$. But $18|abc_{10}$ by definition, so $a + b + c|abc_{10}$, as desired. □

Remember that with questions involving things like digits, a direct approach is not usually the answer. A cumbersome formula should be simplified into something more manageable. In this case, the phrase "one number out of any 18 consecutive numbers must be" was replaced by "any multiple of 18 must be" which was weaker, but simpler and more relevant to the question (which was related to divisibility). It turned out to be a good guess, though. And remember that with finite problems, the strategies are not like those in higher mathematics. For example, the formula

$$a + b + c|abc_{10}$$

was not treated like typical mathematics (e.g. application of modular arithmetic), but instead we placed bounds on $a + b + c$ (9, 18, or 27) due to the fact that all numbers had only three digits, leaving us with the much simpler

$$9|abc_{10}, \quad 18|abc_{10}, \quad \text{ or } 27|abc_{10}.$$

Indeed, we never even had to expand out abc_{10} algebraically as $100a+10b+c$; while that may have seemed like the logical first step, it turns out that is sort of a red herring and does not make the problem any clearer to solve.

A final remark: 18 consecutive numbers are the least number needed to insure one of them satisfies (1). 17 numbers won't work; consider for instance the sequence from 559 to 575. (I used a computer for that, not some tricky mathematics.)

Exercise 2.1. In a parlour game, the "magician" asks one of the participants to think of a three-digit number abc_{10}. Then the magician asks the participant to add the five numbers acb_{10}, bac_{10}, bca_{10}, cab_{10} and cba_{10}, and reveal their sum. Suppose the sum was 3194. What was abc_{10} originally? (Hint: Get a better expression for the sum of the five numbers, something more mathematical. Then use modular arithmetic to get some bounds on a, b, and c.)

Problem 2.2 (Taylor 1989, p. 37). Is there a power of 2 such that its digits could be re-arranged and made into another power of 2? (No zeroes are allowed in the leading digit: e.g. 0032 is not allowed.)

This seems like an unsolvable combination: powers of 2, and digit re-arranging. This is because

(a) digit re-arrangement has so many possibilities, and
(b) it is not easy to determine individual digits of a power of two.

This probably means that something sneaky is needed.

The first sneaky thing to be done is to guess the answer. Circumstantial evidence (this problem is from a mathematics competition) suggests that this is not a trial-and-error question, and so the answer should probably be "no". (On the other hand, some exceptionally ingenious construction could pull off a clever rearrangement of digits - but such a construction is probably not easy to find. Guess the easy options first. If you are right, you have saved a lot of time by not pursuing the hard ways. If you are wrong, you were doomed to a long haul anyway. This does not mean that you should forget about a promising but hard way to solve the problem: but rather, to take a sensible look around before plunging into deep water.)

Like in Problem 2.1, the digits are really sort of a red herring. In Problem 2.1, we only wanted to know two things about the sum of the digits: firstly, a divisibility conditionn, and secondly a size restriction. We didn't want to introduce all the complications of an exact equation. It will probably be much the same here: we have to simplify the problem by generalising the digit-switching process. From a purely logical viewpoint, we are worse off because we have to prove more: but in terms of clarity and simplicity we are gaining ground. (Why burden yourself with data that cannot be used? It will just be a distraction.)

So, we now have to pick out the main properties of powers of 2 and digit-switching - hopefully, we will find properties of one that are incompatible with the other. Now let's tackle powers of two first; they are easier to handle. Here they are:

$$1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, 65536, \ldots$$

Well, there is not very much you can say about the digits here. The last digit of a power of 2 is obviously even (except for the number 1), but the other digits are quite random-looking. Suppose you took the number 4096, for instance. An odd digit, a few even digits and even a 0 digit here. What's stopping it being re-arranged into another power of 2? Could it be re-arranged into $2^{4256} = 1523...936$, for instance? "Of course not!" you would say. Why? "Because it's far too big!". So, does size count? "Yes - There would be thousands of digits in 2^{4256}, and only four digits in 4096." Aha - so rearranging digits cannot change the total number of digits. (Write down any facts which could be of use to your problem, even if they are simple - do not assume that "obvious" facts will always spring to mind when needed. Even shallowly-dug gold has to be searched - and held on to.)

Well, with this iota of information, can we proceed with our generalizing plan? Our generalized question is now

> Is there a power of 2 such that there is another power of 2 with the same number of digits as the first power of 2?

Unfortunately, the answer to this question is quickly seen to be "yes"; 2048 and 4096, for example. We were too general. (Also, a "yes" answer to this question doesn't necessarily yield a "yes" answer to the original problem.) Again, look to Problem 2.1. Merely knowing "the sum of digits of a multiple of 18 has to be a multiple of 9" is not sufficient to solve the problem : we also needed the fact that "the sum of digits of three-digit number is at most 27". In short, we haven't found enough facts about our problem to solve it. Yet - we are still partially successful, because we have restricted the possibilities of digit rearranging. Take the number 4096 again. This can only be re-arranged into another four-digit number. And how many four-digit powers of 2 are there? only four - 1024, 2048, 4096, and 8192. This is because the powers of two keep doubling: they can't stay in the same "tax bracket" for too long. In fact, one can soon see that at most four powers of two can have the same number of digits. (The fifth consecutive power of 2 would be 16 times that of the first, and hence would have to have more digits than the first power of 2). So what this means is that for each power of 2, there are at most three other powers of two that could possibly be digit-rearrangements of the original power of two. A partial victory: only three suspects left to eliminate for each power of two, instead of the infinite number before. Perhaps with a bit of extra work we can eliminate those suspects as well.

We have said that when we switch the digits, the number you end up with has the same number of digits as the original. But the reverse is far from true, and this lone property of digit-switching will not solve the problem on its own. This means that we have generalized too far and pushed our luck too much. Let's reel ourselves in again. Something else could be preserved when we switch digits. Let's take a look at some examples - let's take 4096 again, since we've already got some experience with this number. The digit-switching possibilities are

$$4069, 4096, 4609, 4690, 4906{,}4960, 6049, 6094, 6409,$$
$$6490, 6904, 6940, 9046, 9064{,}9406, 9460, 9604, 9640.$$

What do they have in common? They have the same set of digits. That's all very well and good, but the "set of digits" is not a very useful mathematical object (not many theorems and tools use this concept). However, the *sum of digits* is a more conventional weapon. And, well, if two numbers have the same set of digits, then they have to have the same digit-sum. So we have another iota of information: digit-switching preserves the digit-sum. Combining this with our previous iota we have a new replacement question:

Is there a power of 2 such that there is another power of 2 with the same number of digits *and* the same digit-sum as the first power of 2?

Again, if this question is true, the original question is true. Now this question is a bit easier to cope with than the original, because "number of digits" and "digit-sums" are standard number-theory stuff.

With this new concept in mind, let's look at the digit-sums of the powers of 2, seeing as our new question involves them. Well, we have

Power of 2	Digit sum	Power of 2	Digit sum	Power of 2	Digit sum
1	1	256	13	65536	25
2	2	512	8	131072	14
4	4	1024	7	262144	19
8	8	2048	14	524288	29
16	7	4096	19	1048576	31
32	5	8192	20		
64	10	16384	22		
128	11	32768	26		

From this we note that

- The digit sums tend to be quite small. For instance, the digit sum of 217 is a mere 14. This is actually a small bit of bad luck, because small numbers are more likely to match than are big numbers. (If ten people each randomly pick one two-digit number, there is a sizeable (9.5%) chance of a match, but if they each pick ten-digit numbers, then there is only a one in a million chance of a match: something about as lousy as the chances of winning the lottery.) But the smallness of the numbers also aids in picking out patterns, so perhaps it is not all bad news.

- Some digit-sums match: for example, 16 and 1024. But it seems that the digit sums slowly climbs higher anyway: you would expect that a 100-digit power of 2 would have a higher digit-sum than a 10-digit one. But also remember that we are confining ourselves to powers of two with the same number of digits, so this idea will not be not much help.

The upshot of these observations is this: digit-sums have an easily appreciable macroscopic structure (slowly increasing with n; in fact it is highly probable (though not proven!) that the digit sum of 2^n is approximately $(4.5 \log_{10} 2)n \approx 1.355n$ for large n) but a lousy microscopic structure. The digits just fluctuate too much. We mentioned earlier that "set of digits" was unwieldy: now it seems that "digit-sum" is not so flash either. Is there another reduction of the problem that will leave us with something we can really work with?

Hmm. We mentioned earlier that "digit-sum" was a "conventional weapon" in mathematics. Take a look at the preceding question for instance. But the only real way digit-sums can be successfully "mainstreamed" is by considering the digit sum modulo 9. One may recall that a number is equal to its digit-sum modulo 9; for example,

$$\begin{aligned} 3297 &= 3 \times 10^3 + 2 \times 10^2 + 9 \times 10^1 + 7 \times 10^0 \pmod 9 \\ &= 3 \times 1^3 + 2 \times 1^2 + 9 \times 1^1 + 7 \times 1^1 \pmod 9 \\ &= 3 + 2 + 9 + 7 \pmod 9 \end{aligned}$$

because 10 is equal to 1 (mod 9).

So now our new modified question is as follows:

Is there a power of 2 such that there is another power of 2 with the same number of digits and the same digit-sum *modulo* 9 as the first power of 2?

Now we can use the fact that a number is equal to its digit-sum modulo 9 to rephrase this question again:

Is there a power of 2 such that there is another power of 2 with the same number of digits and the same remainder (mod 9) as the first power of 2?

Note that the pesky notions of "rearranging digits", "set of digits", and "sum of digits" have been completely eliminated, which looks promising. Now let's modify the above table of digit-sums of powers of 2 and see what we get.

	(mod 9)		(mod 9)		(mod 9)
Power of 2	Remainder	Power of 2	Remainder	Power of 2	Remainder
1	1	256	4	65536	7
2	2	512	8	131072	5
4	4	1024	7	262144	1
8	8	2048	5	524288	2
16	7	4096	1	1048576	4
32	5	8192	2		
64	1	16384	4		
128	2	32768	8		

What we have to prove is that no two powers of two have the same remainder (mod 9) and the same number of digits. Well, looking at the table, there are several powers of 2 with the same remainder: 1, 64, 4096, and 262144 for example. But none of these four have the same number of digits. Indeed, powers of 2 with the same remainder (mod 9) seem to be so separated that there is no hope of them having the same number of digits. In fact, the powers of 2 with the same remainder seem to be quite regularly spaced ... and one can quickly see that the remainders (mod 9) repeat themselves every six steps. This conjecture can be easily proved by modular arithmetic:

$$2^{n+6} = 2^n 2^6 = 2^n \times 64 = 2^n (\text{mod } 9) \text{ because } 64 = 1 \ (\text{mod } 9).$$

This result means that the remainders of the powers of 2 will repeat themselves endlessly, like a repeating decimal: $1, 2, 4, 8, 7, 5, 1, 2, 4, 8, 7, 5, 1, 2, 4, 8, 7, 5, \ldots$. This in turn means that two powers of 2 with the same digit-sum (mod 9) must be at least six steps apart. But then the powers of two cannot possibly have the same number of digits, because one would be 64 times bigger than the other, at least. So this means that there are no powers of two with the same number of digits and the same digit-sum (mod 9). We have now proved our modified question, so we can work backwards until we reach our original question, and write out the full answer:

Proof. Suppose two powers of 2 are related by digit-switching. This means that they have the same number of digits, and also have the same digit-sum (mod 9). But the digit-sums (mod 9) are periodic with a period of 6, so the two powers are at least six steps apart. But then it is impossible for them to have the same number of digits, a contradiction. □

This problem was simplified repeatedly until the more unusable and unfriendly parts of the problem were exchanged with more natural, flexible and co-operative notions. This simplification can be a bit of a hit-and-miss affair; there is always the danger of oversimplification, or mis-simplification (simplifying into a red herring). But in this question, almost anything was better than playing around with digit-switching, so simplification couldn't do much more harm. There is a chance that maneuvering and simplifying may land you into a wild goose chase, but if you're really stuck anyway, anything is worth a try.

Diophantine equations

A *Diophantine equation* is an algebraic equation (the classic one is $a^2 + b^2 = c^2$) with the constraint that all variables are integers. The usual objective is to find all solutions to the equation. Generally, there is more than one solution, even if everything is integral. These equations can be solved algebraically, but one also can use the number-theoretical methods of integer division, modular arithmetic, and integral factorisation. Here is one:

Problem 2.3 (Australian Mathematics Competition 1987, p. 15). Find all integers n such that the equation $1/a + 1/b = n/(a+b)$ is satisfied for some non-zero integer values of a and b (with $a+b \neq 0$).

This seems like a standard Diophantine equation, so we would probably begin by multiplying out the denominators, to get

$$(a+b)/ab = n/(a+b)$$

and then

$$(a+b)^2 = nab. \tag{2}$$

Now what? We could eliminate the n, and say that

$$ab|(a+b)^2$$

(using the divisibility symbol | that we used in Problem 2.1) or try to concentrate on the fact that nab is a square. These techniques are good, but they don't seem to work on this problem. The relationships of the left and right sides of (2) are not strong enough. One side is a square, the other is a product.

One thing to keep in mind when problem-solving is to be prepared to abandon temporarily one interesting - but fruitless - approach and try a more promising one. One could try algebra to attack the problem, then re-apply number theory later if algebra failed to work. Expanding (2) and collecting terms we can get

$$a^2 + (2-n)ab + b^2 = 0,$$

and if one is brave enough to use the quadratic formula we get

$$a = \frac{b}{2}[(n-2) \pm \sqrt{(n-2)^2 - 4}].$$

This looks very messy, but actually we can turn this messiness to our advantage. We know that a, b, and n are integers, but there is a square root in the formula. Now this can only work if the term inside the square root, $(n-2)^2 - 4$, is a perfect square. But this means that 4 less than a square is a square. This is very restrictive. Because the gaps between the squares get higher than 4 after the first few squares, we only need test low numbers of n. It turns out that $(n-2)^2$ has to be 4, and hence n is either 0 or 4. Now we can work each case separately, finding either an example of each or a proof that no such example exists.

Case 1: $n = 0$**.** Feeding this back into, say, (2) we get $(a+b)^2 = 0$, and thus $a + b = 0$. But this is impossible as in our original equation we now have a $0/0$, which is illegal. Hence n cannot be 0.

Case 2: n=4. Again, (2) gives us $(a+b)^2 = 4ab$, which upon collecting terms gives $a^2 - 2ab + b^2 = 0$. Factorizing this we get $(a-b)^2 = 0$, so a must equal b. This is not a contradiction, but an example: $a = b$, $n = 4$, works when put into the original equation (2).

So our answer was $n = 4$, but it was obtained by the rather inelegant method of the quadratic formula. Using it is usually clumsy, but as it introduces a square root term, which implies that the term inside the square root must be a perfect square, it occasionally comes in useful.

Diophantine problems can get extremely difficult when one of the variables appears in the exponent; the most notorious of these is *Fermat's last theorem*, which asserts that there are no natural number solutions to $a^n + b^n = c^n$ with $n > 2$. Fortunately there are other problems involving exponents which are easier to handle.

Problem 2.4 (Taylor 1989, p.7). Find all solutions of $2^n + 7 = x^2$ where n and x are integers.

This kind of question really needs trial and error to find the right tack. With diophantine equations, the most elementary methods are modular arithmetic and factorisation. Modular arithmetic transfers the entire equation to a suitable modulus, sometimes constant (e.g. (mod 7), or (mod 16)) or sometimes variable (e.g. (mod pq)). Factorisation alters the problem into the form (factor) × (factor) = (something nice), where the right-hand side could be a constant (the best possible result), a prime, a square, or something else that has a limited choice of factors. For example, in Problem 2.3, both methods were considered early on, but discarded in favour of an algebraic approach, which is actually a factorisation technique in disguise (remember we eventually got $(n-2)^2 - 4 =$ (square)?).

Now it is best to try elementary techniques first, as it may save a lot of dashing about in circles later. One may have abandoned these methods and tried to analyze the approximate equation

$$x = \sqrt{2^n + 7} \approx 2^{n/2}$$

which can get into some serious number theory involving topics such as continued fractions, Pell's equation, and recursion relations. It can be done; but we'll look for the elegant (i.e. lazy) way out.

Obtaining a useful factorisation is next to impossible, except when n is even. Then we get a difference of two squares (a vital factorisation in Diophantine equations) like so:

$$7 = x^2 - 2^n = (x - 2^m)(x + 2^m)$$

where $m = n/2$. Then we can say that $x - 2^m$ and $x + 2^m$, being factors of 7, must be -7, -1, 1 or 7; and further breakup into cases soon shows that there are no solutions (if we assume n is even). But that is about as much as the factorisation method can tell us; it doesn't tell us where the actual solutions are and how many of them there are. (Although we do now know that n must be odd.)

The modular arithmetic approach is next. The strategy is to use the modulus to get rid of one or more of the terms. For example, we could write the equation modulo x, to obtain

$$2^n + 7 = 0 \pmod{x},$$

or maybe modulo 7, to get

$$2^n = x^2 \pmod{7}.$$

Unfortunately, these methods don't work well at all. But before we give up, there is one more modulus to try. We tried eliminating the "7" and the "x^2" terms; can we eliminate the 2^n term instead? Yes, by choosing, say, mod 2. Then we get

$$0 + 7 = x^2 (\text{mod } 2)$$

when $n > 0$, and

$$1 + 7 = x^2 (\text{mod } 2)$$

when $n = 0$. This is not too bad as we have almost eliminated the role of n completely. But it still doesn't work, as the x^2 term on the right-hand side could be 0 or 1, so we haven't really excluded any possibilities. To restrict the values of x^2, we have to choose a different modulus. With this line of thought - to restrict the values on the right-hand-side - one now thinks to try the modulus 4 instead of 2:

$$2^n + 7 = x^2 \pmod{4}.$$

In other words, we have

$$0 + 3 = x^2 \pmod{4} \text{ when } n > 1 \tag{3}$$

$$2 + 3 = x^2 \pmod{4} \text{ when } n = 1 \tag{4}$$

$$1 + 3 = x^2 \pmod{4} \text{ when } n = 0. \tag{5}$$

Because x^2 must be 0 (mod 4) or 1 (mod 4), possibility (3) is eliminated. This means n can only be 0 or 1. A quick check shows then that only $n = 1$ can work, and x must be $+3$ or -3.

The main idea, when solving Diophantine equations of the form "find all solutions", is to eliminate all but a finite number of possibilities. This is another reason why the (mod 7) and (mod x) would not work; for if they did, they would have eliminated all the cases, unlike the (mod 4) approach, which eliminated all but a handful.

Exercise 2.2. Find the largest positive integer n such that $n^3 + 100$ is divisible by $n + 10$. Hint: use (mod $n + 10$). Get rid of the n by using the fact that $n = -10(\text{mod } n + 10)$.

Sums of powers

PROBLEM 2.5 (Hajós *et al.* 1963, p. 74). Prove that for any nonnegative integer n, the number $1^n + 2^n + 3^n + 4^n$ is divisible by 5 if and only if n is not divisible by 4.

This problem looks a bit daunting at first: equations like the above may remind one of Fermat's last theorem, which is notorious for its insolvability. But our question is much milder. We wish to show that a certain number is (or is not) divisible by 5. Unless a direct factorisation is evident, we will have to use the modulus approach. (i.e. show that $1^n + 2^n + 3^n + 4^n = 0 \pmod 5$ for n not divisible by 4, and $1^n + 2^n + 3^n + 4^n \neq 0 \pmod 5$ otherwise.)

Because we are using such small numbers, we can evaluate some of the values of $1^n + 2^n + 3^n + 4^n \pmod 5$ manually. The best way to do this is to work out $1^n \pmod 5$, $2^n \pmod 5$, $3^n \pmod 5$, and $4^n \pmod 5$ individually before adding:

(mod 5)

n	1^n	2^n	3^n	4^n	$1^n + 2^n + 3^n + 4^n$
0	1	1	1	1	4
1	1	2	3	4	0
2	1	4	4	1	0
3	1	3	2	4	0
4	1	1	1	1	4
5	1	2	3	4	0
6	1	4	4	1	0
7	1	3	2	4	0
8	1	1	1	1	4

Now it is obvious that some periodicity is evident. In fact 1^n, 2^n, 3^n and 4^n are all periodic with period 4. To prove this conjecture, we can just fiddle with the definition of periodicity.

Take 3^n, for example. Saying that this is periodic with period 4 just means that

$$3^{n+4} = 3^n \pmod 5.$$

But this is easy to prove, as

$$3^{n+4} = 3^n \times 81 = 3^n \pmod 5$$

because $81 = 1 \text{mod } 5$.

Similarly we can prove 1^n,2^n, and 4^n are periodic with period 4. This means that $1^n + 2^n + 3^n + 4^n$ is periodic with period 4. This in turn implies that we only need to prove our question for $n = 0, 1, 2, 3$, because periodicity will take care of all the other cases of n. But we have already shown the question to be true in these cases (see the above table). So we are done. (By the way, there is a more elementary method available if we assume that n is odd: simply pair up and cancel terms.)

Whenever trying to prove equations involving a parameter (in this case n), periodicity is always handy, as one no longer needs to check all values of the parameter to verify the equation. Checking one period (e.g. $n = 0$, 1, 2, and 3) will be sufficient.

Incidentally, the above question can be generalized as follows:

Theorem 2.1. If p is a prime, then $1^n + 2^n + 3^n + \ldots + (p-1)^n$ is always divisible by p except when n is divisible by $(p-1)$.

This theorem is a bit more complicated, and it involves some manipulation of residue classes, as well as knowledge of generators. Here is a quick sketch for those who have studied elementary number theory in some depth:

Proof. If n is not divisible by $p-1$, and a is a generator of p, then

$$a^n \neq 1 \pmod{p}.$$

Now

$$\begin{aligned} a^n(1^n + 2^n + 3^n + \ldots + (p-1)^n) &= a^n + (2a)^n + (3a)^n + \ldots + ((p-1)a)^n \\ & 1^n + 2^n + 3^n + \ldots + (p-1)^n \pmod{p} \end{aligned}$$

because the set of residues $\{a, 2a, \ldots, (p-1)a\}$ is equal to the set of residues $\{1, 2, \ldots, p-1\}$ modulo p. So denoting our sum $1^n + 2^n + 3^n + \ldots + (p-1)^n$ by X, we have shown that

$$a^n X = X \pmod{p}$$

so X must be 0 (mod p), as desired. □

Exercise 2.3. Show that the equation $x^4 + 131 = 3y^4$ has no solutions if x and y are integers.

Now we turn to a trickier problem concerning sums of powers.

Problem 2.6 (Schklarsky *et al.*, 1962, p. 14). (**) Let k, n be natural numbers with k odd. Prove that the sum $1^k + 2^k + \ldots + n^k$ is divisible by $1 + 2 + \ldots + n$.

This question, by the way, is a standard exercise in Bernoulli polynomials (or some astute applications of the Remainder Theorem), an interesting portion of mathematics that has many applications. But without the sledge-hammer of Bernoulli polynomials (or the Riemann ζ function) we'll just have to use plain old number theory.

First of all, we know that $1 + 2 + \ldots + n$ can also be written in the form $n(n+1)/2$. Which form shall we use? The former is more aesthetic, but a bit useless in a divisibility question. (It is always easier if the divisor is expressed as a product, rather than a sum.) It might have been useful if there was some nice factorisation of $1^k + 2^k + .. + n^k$ which involved $1 + 2 + \ldots + n$, but there isn't (at least, not an obvious one). If there was some way to relate divisibility by $1 + 2 + \ldots + n$ to divisibility by $1 + 2 + \ldots + (n + 1)$ then induction might be a way to go, but that doesn't seem likely either. So we will try the $n(n + 1)/2$ formulation instead.

So, using modular arithmetic (which is the most flexible way to prove that one number divides another), our objective is to show that

$$1^k + 2^k + \ldots + n^k = 0 \text{ (mod } n(n + 1)/2).$$

Let us ignore for the moment the "2" in the n(n+1)/2. Then we are trying to prove something of the form

$$\text{(factor 1)} \times \text{(factor 2)}|\text{(expression)}.$$

If the two factors are coprime, then our objective is equivalent to proving both of

$$\text{(factor 1)}|\text{(expression) and (factor 2)}|\text{(expression)}$$

separately. This should be simpler to prove: it is easier to prove divisibility if the divisors are smaller. But there is an annoying "2" in the way. To deal with that we will just break up into cases, depending on whether n is even or odd[1]. The cases are quite similar and I will only do the case when n is even. In this case we can write $n = 2m$ (so as to avoid staring at messy "n/2" terms in the following equations - little housekeeping things like this help a solution run smoothly.) Replacing all the n's by $2m$'s, we have to prove

$$1^k + 2^k + \ldots + (2m)^k = 0 \text{ (mod } m(2m + 1)),$$

but since m and $2m + 1$ are coprime, this is equivalent to proving

$$1^k + 2^k + \ldots + (2m)^k = 0 \text{ (mod } 2m + 1)$$

and

$$1^k + 2^k + \ldots + (2m)^k = 0 \text{ (mod } m).$$

Let's tackle the (mod $2m$+1) part first. It is quite similar to Problem 2.5 but is a bit easier, because we know that k is odd. Using the modulus $2m + 1$, $2m$ is equivalent to -1, $2m - 1$ is equivalent to -2, and so on, so our expression $1^k + 2^k + \ldots + (2m)^k$ becomes

$$1^k + 2^k + \ldots + (m)^k + (-m)^k + \ldots + (-2)^k + (-1)^k \text{ (mod } 2m + 1).$$

We have done this so that we can do some nice cancelling. k is odd, so $(-1)^k$ is equal to -1. Therefore $(-a)^k = -a^k$. The upshot of this is that the above sum can be pairwise cancelled: 2^k and $(-2)^k$ will cancel, 3^k and $(-3)^k$ will cancel, etc, leaving 0 (mod $2m + 1$), as desired.

Now we have to do the (mod m) part: i.e. we have to show

$$1^k+2^k+3^k+\ldots+(m-1)^k+(m)^k+(m+1)^k+\ldots+(2m-1)^k+(2m)^k = 0 \text{ (mod } m).$$

[1] Another way is to multiply both sides by two, so that we now want to prove $2(1^k + 2^k + \ldots + n^k = 0$ (mod $n(n + 1)$). This ends up being more or less equivalent to the approach given below.

But we are working modulo m, so some of the above terms can be simplified. m and $2m$ are both equivalent to 0 (mod m), and $m+1$ is equivalent to 1, $m+2$ is equivalent to 2, and so on. So the above summation simplifies to

$$1^k + 2^k + 3^k + \ldots + (m-1)^k + 0^k + 1^k + \ldots + (m-1)^k + 0 \pmod{m}$$

But several terms appear twice, so recombining (and ditching the 0s) we get

$$2(1^k + 2^k + 3^k + \ldots + (m-1)^k) \pmod{m}$$

Now we can almost do the same thing as for the (mod $2m+1$) case, except there is a small hitch when m is even. If m is odd, we can reformulate the above expression as

$$2(1^k + 2^k + 3^k + \ldots + ((m-1)/2)^k + (-(m-1)/2)^k + \ldots + (-2)^k + (-1)^k) \pmod{m}$$

and do the same procedure of cancellation as before. But if m is even (so $m = 2p$, say) there is a middle term, p^k, which doesn't cancel with anything. In other words, in this case the expression does not collapse to 0 immediately, but instead cancels to

$$2p^k (\text{mod } 2p)$$

But this, of course, is equal to 0. Regardless of whether m is odd or even, we have proved that $1^k + 2^k + 3^k + \ldots + n^k$ is divisible by $n(n+1)/2$ if n is even.

Exercise 2.4. Complete the proof of the above problem by working out what happens when n is odd.

Now let's turn to a special type of "sums of powers" problem, namely sums of reciprocals.

Problem 2.7 (Schklarsky *et al.*, 1962, p. 17). Let p be a prime number greater than 3. Show that the numerator of the (reduced) fraction

$$1/1 + 1/2 + 1/3 + \ldots + 1/(p-1)$$

is divisible by p^2. For example, when p is 5, the fraction is $1/1 + 1/2 + 1/3 + 1/4 = 25/12$, and the numerator is obviously divisible by 5^2.

This question is a "Prove that" question, not a "Find a" or "Show there exists" question, so it shouldn't be completely impossible. However, we have to prove something about a numerator of a reduced fraction - not something easily dealt with! This numerator will need to be transformed into something more standard, like an algebraic expression, so that we can manipulate it better. Also, the question does not just need divisibility by a prime, it needs divisibility by the square of a prime. This is significantly harder. We would like to somehow reduce the problem to mere prime divisibility to make the problem more solvable.

So by looking at the shape of the question, we have the following objectives to keep in mind:

(a) Express the numerator as a mathematical expression, so that we can manipulate it.
(b) Aim to reduce the problem from a p^2-divisiblity problem to something simpler, perhaps a p-divisibility problem.

Let's tackle (a) first. First of all, we can get a numerator easily, but not the reduced numerator necessarily. By adding up the fractions under a common denominator we get

$$\frac{2 \times 3 \times \ldots \times (p-1) + 1 \times 3 \times \ldots \times (p-1) + \ldots + 1 \times 2 \times 3 \times \ldots \times (p-2)}{(p-1)!}$$

Now suppose that we can manage to prove that this numerator is divisible by p^2. How does this help us prove that the reduced numerator is also divisible by p^2? Well, what is the reduced numerator? It is the original numerator after some cancellation with the denominator. Can cancelling destroy the property of p^2-divisibility? Yes, if a multiple of p is cancelled. But multiples of p cannot be cancelled, because the denominator is coprime to p (p is prime, and $(p-1)!$ can be expressed as a product of numbers less than p). Aha! This means that we only need to prove that the ugly-looking numerator above is divisible by p^2. This is better than the other numerator because now we have an equation to solve:

$$\begin{aligned} 2 \times 3 \times \ldots \times (p-1) + 1 \times 3 \times \ldots \times (p-1) + \ldots \\ +1 \times 2 \times 3 \times \ldots \times (p-2) = 0 (\text{mod } p^2). \end{aligned}$$

(Again, we have switched over to modular arithmetic, which is usually the best way to show that one number divides another. However, if the question involves more than one divisibility, e.g. something involving all divisors of a certain number, other techniques are sometimes better.)

Although we have got an equation now, it is a mess. Our next task is to simplify it. What we have now on the left-hand side is an indefinite sum of indefinite products. (Indefinite just means that there are "dot dot dots" in the expression.) However, we can represent the infinite products more neatly. Each infinite product is basically the numbers from 1 to $p-1$ multiplied together, except for one number, say i, which is between 1 and $p-1$. This can be expressed more compactly as $(p-1)!/i$; it is legitimate to divide by i modulo p^2 because i is coprime to p^2. So now our objective is now to prove

$$\frac{(p-1)!}{1} + \frac{(p-1)!}{2} + \frac{(p-1)!}{3} + \ldots + \frac{(p-1)!}{p-1} = 0 \ (\text{mod } p^2).$$

We factorize this to get

$$(p-1)! \left[\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{p-1} \right] = 0 \ (\text{mod } p^2). \tag{6}$$

(Remember that we are dealing with modular arithmetic, so that a number like 1/2 will be equivalent to an integer. For example, $1/2 = 6/2 = 3$ (mod 5).)

Now look at what we have: something of the form

$$(\text{factor}) \times (\text{factor}) = 0 \pmod{p^2}.$$

If it were not for the modular arithmetic, then we could quickly say that one of the factors is 0. With modular arithmetic, we can say nearly the same thing, but we have to be careful. Luckily, the first factor, $(p-1)!$, is coprime to p^2 (because $(p-1)!$ is coprime to p) so we can divide it out. The upshot of this is that (6) is equivalent to

$$\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{p-1} = 0 \pmod{p^2}$$

(Note that this looks very similar to our original question, the only difference being that we are considering the entire fraction, not just the numerator of it. But one cannot just jump from one form to another without care. The above complications were necessary.)

Now we have reduced the question to proving a rather benign-looking modular arithmetic equation. But where to go on from here? Perhaps an example will help. Let's take the same example as the one given in the question: namely, $p = 5$. We have

$$\begin{aligned}\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} &= 1 + 13 + 17 + 19 \pmod{25}\\ &= 0 \pmod{25}\end{aligned}$$

as desired. But why does this work? The numbers 1, 13, 17, and 19 seem to be random, but "magically" add up to the right amount. Perhaps it is a fluke. Let's try $p = 7$.

$$\begin{aligned}\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} &= 1 + 25 + 33 + 37 + 10 + 41 (\text{mod } 49)\\ &= 0 \pmod{49}\end{aligned}$$

This has the same "flukiness" about it. How does this work? It is not clear how everything manages to cancel out modulo p^2. Perhaps, keeping objective (b) in mind, we can prove it (mod p) first, i.e. let us first prove

$$\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{p-1} = 0 \pmod{p} \tag{7}$$

If nothing else, it will give us something to do. (Besides, if we can't solve this (mod p) problem, there is no way that we will be able to solve the (mod p^2) problem.)

It turns out that the simpler problem (7) is much easier to work out. For example, when p is 5, we have

$$\begin{aligned}\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} &= 1 + 3 + 2 + 4 \pmod{5}\\ &= 0 \pmod{5}\end{aligned}$$

while when p is 7 we have

$$\begin{aligned}\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6} \pmod 7 &= 1+4+5+2+3+6 \pmod 7\\ &= 1+2+3+4+5+6 \pmod 7\\ &= 0 \pmod 7.\end{aligned}$$

Now we have a pattern emerging: the reciprocals $1/1, 1/2, \ldots, 1/(p-1) \pmod p$ seem to cover all the residues $1, 2, \ldots, (p-1) \pmod p$ exactly once. For example, in the above equation with $p = 7$, the numbers $1+4+5+2+3+6$ re-arrange to form $1+2+3+4+5+6$, which is 0. To check a lengthier example, mod 11 yields

$$\begin{aligned}\frac{1}{1}+\frac{1}{2}+\ldots+\frac{1}{11} &= 1+6+4+3+9+2+8+7+5+10 \pmod{11}\\ &= 1+2+3+4+5+6+7+8+9+10 \pmod{11}\\ &= 0.\end{aligned}$$

This tactic, showing that the reciprocal numbers can be rearranged in this orderly fashion, works neatly for (mod p), but it doesn't generalize easily to (mod p^2). Instead of floundering around trying to fit a square block into a round hole (although it can be done if you push hard enough), it's better to find a block that is more round. So what we have to do now is find another proof of the fact that $\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{p-1} = 0 \pmod p$; one that generalizes, at least partially, to the (mod p^2) case.

Now it is time to use experience with these sorts of problems. For example, if we are fresh from solving Problem 2.6, we know that symmetry, or anti-symmetry can be exploited, especially in modular arithmetic. In the problem of proving (7) we can make the sum more anti-symmetric by replacing $p-1$ with -1, $p-2$ with -2, and so forth, to get

$$\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{p-1} = \frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{-3}+\frac{1}{-2}+\frac{1}{-1} \pmod p.$$

And now we can pair off and cancel easily (there is no "middle term" that doesn't pair off, as p is an odd prime). Can we do the same in (mod p^2)?

The answer is "sort of". When we solved the problem (mod p), we paired off 1/1 and $1/(p-1)$, 1/2 and $1/(p-2)$, and so forth. When we try the same pairing in (mod p^2), what we get now is this:

$$\begin{aligned}\frac{1}{1}+\frac{1}{2}+\ldots+\frac{1}{p-1} &= (\frac{1}{1}+\frac{1}{p-1})+(\frac{1}{2}+\frac{1}{p-2})+\ldots+(\frac{1}{(p-1)/2}+\frac{1}{(p+1)/2})\\ &= \frac{p}{1\times(p-1)}+\frac{p}{2\times(p-2)}+\ldots+\frac{p}{(p-1)/2\times(p+1)/2}\\ &= p\left[\frac{1}{1\times(p-1)}+\frac{1}{2\times(p-2)}+\ldots+\frac{1}{(p-1)/2\times(p+1)/2}\right]\\ &\quad \pmod{p^2}.\end{aligned}$$

Now this, at first, looks like a complication rather than a simplification. But we have gained a very important factor of p on the right-hand side. Now, instead of having to prove that

$$(\text{expression}) = 0 \pmod{p^2}$$

we now have to prove something like

$$(p \times \text{expression}) = 0 \pmod{p^2}$$

which is equivalent to proving something of the form

$$(\text{expression}) = 0 \pmod{p}.$$

In other words, we are now reduced to a (mod p) question instead of a (mod p^2) question. Now we have achieved objective (b) given above: reduced the question to that of a smaller modulus, which is well worth the slight increase in complexity.

And it is quickly seen that the apparent increase in expression complexity is just illusionary, as the (mod p) can get rid of a lot more terms than (mod p^2) can. Now, we only have to show that

$$\frac{1}{1 \times (p-1)} + \frac{1}{2 \times (p-2)} + \ldots + \frac{1}{(p-1)/2 \times (p+1)/2} = 0 \pmod{p}.$$

But $p-1$ is equivalent to -1 (mod p), $p-2$ is equivalent to -2 (mod p), and so forth, so the equation reduces to

$$\frac{1}{-1^2} + \frac{1}{-2^2} + \ldots + \frac{1}{-((p-1)/2)^2} = 0 \pmod{p},$$

or equivalently

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{((p-1)/2)^2} = 0 \pmod{p}.$$

This equation is not too bad, except that the series on the left-hand side ends in an obscure spot (at $1/((p-1)/2)^2$, rather than the more natural $1/(p-1)^2$, for example). But we can "double up", making use of the fact that $(-a)^2 = a^2$ to get

$$\begin{aligned}
&\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{((p-1)/2)^2} \\
&\qquad = \frac{1}{2}\left[\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{((p-1)/2)^2}\right. \\
&\qquad\quad \left. + \frac{1}{(-1)^2} + \frac{1}{(-2)^2} + \frac{1}{(-3)^2} + \ldots + \frac{1}{(-(p-1)/2)^2}\right] \pmod{p} \\
&\qquad = \frac{1}{2}\left[\frac{1}{1^2} + \ldots + \frac{1}{(p-1)^2}\right] \pmod{p}
\end{aligned}$$

So proving that $\frac{1}{1^2} + \ldots + \frac{1}{((p-1)/2)^2}$ is equal to 0 (mod p) would be equivalent to proving that $1/1^2 + \ldots + 1/(p-1)^2$ is equal to 0 (mod p). The latter is more desirable because of its more symmetrical format. (Symmetry is nice to keep - until it can be used to its full effect - while anti-symmetry, like what we did on the previous page, is nice to cancel.)

So now we only have to prove

$$\frac{1}{1^2} + \frac{1}{2^2} + \ldots + \frac{1}{(p-1)^2} = 0 \pmod{p}. \tag{8}$$

to prove the whole question. This is tactically a much better formulation than the original one involving numerators and p^2 divisibility, which is a lot stronger (hence harder to prove) than mere p-divisiblity.

So now we have achieved all our tactical goals, and reduced the question down to decent proportions. But where do we go from here? Well, the question seems very closely related to the other problem (7) that we were considering. But we are not going around in circles. Our current goal (8) will imply the original question, whereas (7) was just a side-problem, a simpler version of the question. Rather than going around in circles, we are going around in spirals, heading towards a solution. We have already proved (7): can we prove (8) by the same methods?

Well, we are in luck, because there were two methods we used to solve (7): one was the rearrangement of reciprocals, and the other was cancellation of pairs. Cancellation of pairs unfortunately does not work as well with (8) as it did with (7), mainly because of the squares in the denominators, which produce symmetry rather than anti-symmetry. But the rearrangement method is promising. Take, yet again, the example of $p = 5$ (so we can reuse some previous work):

$$\begin{aligned}\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2} &= 1^2+3^2+2^2+4^2 \pmod 5\\ &= 1^2+2^2+3^2+4^2 \pmod 5\\ &= 0\end{aligned}$$

The way it works when $p = 5$ shows the way for the general case. Based on the above examples it looks like the residue classes $1/1, 1/2, 1/3, \ldots, 1/(p-1) \pmod p$ are just a rearrangement of the numbers $1, 2, 3, \ldots, (p-1) \pmod p$; a proof of this fact will be given at the end of this discussion. Thus we can say that the numbers $1/1^2, 1/2^2, \ldots, 1/(p-1)^2$ are just rearrangements of the numbers $1^2, 2^2, 3^2, \ldots, (p-1)^2$. In other words:

$$\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\ldots+\frac{1}{(p-1)^2} = 1^2+2^2+3^2+\ldots+(p-1)^2 \pmod p.$$

This is an easier expression to deal with, because we have removed the reciprocals, which are a nuisance when trying to sum things. In fact, we can now get rid of the sum altogether, using the standard formula

$$1^2+2^2+\ldots+n^2 = \frac{n(n+1)(2n+1)}{6}$$

(which is easily proven by induction), so we have reduced (8) to just proving that

$$\frac{(p-1)p(2p-1)}{6} = 0 \pmod p.$$

And one can easily show that this is true when p is a prime greater than 3 (because $(p-1)(2p-1)/6$ is an integer in this case).

So that's it. We keep reducing the equation to simpler and simpler formulations, until it just collapses into nothing. A bit of a long haul, but sometimes it is the only way to resolve these very complicated questions: step-by-step reduction.

Now for the proof that the reciprocals $1/1, 1/2, \ldots, 1/(p-1) \pmod{p}$ are a permutation of the numbers $1, 2, \ldots, (p-1) \pmod{p}$: This is equivalent to saying that each nonzero residue (mod p) is the reciprocal of one and only one nonzero residue (mod p), which is obvious.

EXERCISE 2.5. Let $n \geq 2$ be an integer. Show that $\frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{n}$ is *not* an integer. (You will need *Bertrand's postulate* (actually a theorem), which shows that given any positive integer n there is at least one prime between n and $2n$.)