Speed Concepts in Mathematics

Patrick D. Bangert*

February 6, 2005

1 Introduction

The word "speed" generally indicates a measure of rapidity in approaching some goal. Whether it is the velocity of a human being approaching a well-earned diner, the approach of an approximation to the true state of the matter or the time taken by a computer to work out its next chess move, we are dealing with speed in the context of mathematics. Speed inherently involves comparing the speed of something to either the speed of something else or a standard. Comparing to a standard raises issues of measurement and quantification of the concept speed and we are now firmly in mathematical territory.

The physical speed of things can be modelled using mathematics which can be used, in turn, to predict events in the future. Apart from this use of mathematics as a tool, speed figures largely within mathematics itself. Quantities often have to be approximated by other, simpler quantities and this approximation raises questions of how quickly the approximation is good enough. Limitation of resources and time force the mathematician to restrict himself to the use of computer programs that will finish quickly but accomplish as much as possible in that time. This leads directly to the foremost problem of modern mathematics that asks if a certain salesman's journey can be scheduled by a computer program whose execution time is bounded by a polynomial function in the number of places he must visit. Let us make haste and delve into the details with patience.

2 Physical Speed Modelled by Mathematics

How does the speedometer of a car work? It counts the number of turns that the wheel of the car makes in one second. This is then translated into units of kilometers per hour by noting the circumference of the wheel. The important point is that the speedometer needs to count for a little while and can only then report its findings: Measuring speed takes time and cannot be done instantaneously. Let us follow the calculation. Suppose the speedometer counts $n = 10$ turns of a wheel of diameter 0.7 meters in one second. Then the speed is

$$\text{speed} = \frac{\text{distance moved}}{\text{time taken}} = \frac{0.7\pi n}{1}\ \text{m/s} = \frac{0.7 \cdot 360\pi n}{1000}\ \text{km/h} \approx 39.6\ \text{km/h} \tag{1}$$

We usually think of speed as an instantaneous quantity; the speed *now*. Indeed, one has a speed at the moment but to measure the speed one needs some interval of time. The reported speed is therefore an average speed over that interval of time.

To make this more apparent, let us graphically display our work as we proceed (see figure 1 for this discussion). On the vertical axis we graph the position of the car as measured in meters and on the horizontal axis we graph time in seconds. If we wish to determine the speed of the car four seconds after it starts, we might agree to measure the distance moved between the two second and six second marks. Said and done; these two points on the graph give us a straight line that intersects the real curve of position versus time in two places, namely those at which we took our measurements. The speed is now, as before, the distance moved divided by the time taken. In graphical terminology this is the slope of the line. In this particular case, the "true" function of

*International University Bremen, P.O. Box 750 561, 28725 Bremen, Germany, p.bangert@iu-bremen.de

distance d versus time t is given by the function $d(t) = 40 + 0.1t^4$. The measurement just done would yield a slope of 32 meters per second.

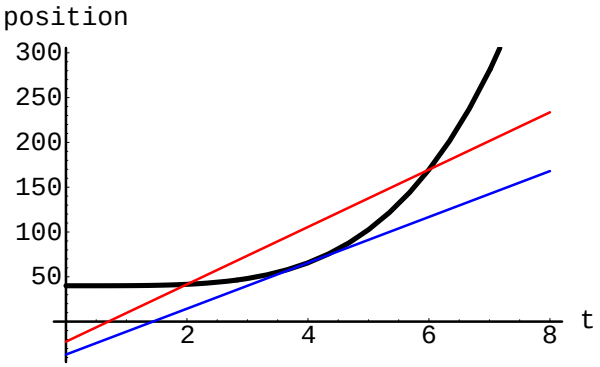

Figure 1: The position of a car is graphed in relation to the time after the start and displayed in the upward curving line with function $d(t) = 40 + 0.1t^4$. The two straight lines represent the average speed between two and six seconds and the instantaneous speed at four seconds.

We may try this again with a smaller interval of time (still centered on four seconds) and would find a slower speed. If the interval is shortened so much that it is, for all practical purposes, of zero length, we will have found the instantaneous speed which, in this case, is 25.6 meters per second. Thus, the instantaneous speed differs from that calculated in equation 1 and is given by

$$\text{inst. speed} = \text{Limit as time taken goes to zero of} \left(\frac{\text{distance moved}}{\text{time taken}} \right) \tag{2}$$

The idea of a limit is the one that makes all the difference. In our example, the two quantities of distance moved and time taken change together; if we measure over more time, the car will have moved further. As the time interval gets smaller and smaller, we average over smaller intervals and the averaging gets closer and closer to the instantaneous value. Graphically, the two points in which the straight line of the averaging process intersects the true curve get closer together until, *in the limit*, they become the same point. This last straight line is called a *tangent* line to the curve as it touches the curve on just a single point and its slope is equal to the slope of the curve at this same point. Figuring out particular tangent lines to curves and vice-versa is the topic of a large branch of mathematics called *analysis* which is more commonly known as *calculus*[1].

The concept of limit may still be a little woolly and thus it will be explained further in the next section in the context of a concept of speed that is internal to mathematics in contrast to the car example that simply used mathematics as a tool.

3 Convergence to a Limit

3.1 Limits of Sequences

In the last section, we met the idea of a limit. This is an important idea and we need to explore it further to meet the idea of convergence speed which is a measurement of how quickly the limit is achieved. To start with, consider the sequence of numbers

$$\frac{1}{2}, \ \frac{2}{3}, \ \frac{3}{4}, \ \frac{4}{5}, \ \cdots, \ \left(1 - \frac{1}{n}\right), \ \cdots \tag{3}$$

[1]Many mathematical analysts would complain at this statement but behind all the complicated words and technicalities lies essentially the simple point of getting tangent lines. I paraphrase Ian Stewart, a foremost popularizer of mathematics and mathematics professor, from a lecture he gave in London in 1998: "Once I was asked, 'Isn't mathematics all about doing great big sums?' At first I was offended and tried to explain the breadth of the field but upon further reflection noticed that for the most part she was right. Mathematicians don't like to admit it but often they construct complicated methods of answering simple questions. In their defense however, these simple questions do not (and can not) have simple answers and so they are justified in doing what they are doing but they should not be afraid to admit it and demystify what they do."

This sequence of numbers is usually denoted $\{1 - \frac{1}{n}\}$ where n is a variable that is understood to start with 2 and increase by 1 each time *ad infinitum*. Defined in this way the sequence is an infinite sequence; it has an infinite number of terms. What happens when n gets large? The fraction $\frac{1}{n}$ gets small and thus the number $1 - \frac{1}{n}$ gets ever closer to 1 although it will not actually ever reach exactly 1. This is what we mean by a limit: The value of the number $1 - \frac{1}{n}$ can be made as close to its *limit* as we like by simply choosing n large enough. For example, if an accuracy of $1/5$ is good enough for me, n need only be 5 and the value is equal to 1 within a margin of $1/5$. If my error margin is tighter, such as 0.001, then n must be 1000.

The question of how quickly a sequence converges to its limit is answered by giving the number of terms (that is n) that one must take before the achieved value is within a given margin of the limit. Why this is useful, we will get to shortly and how this can be quantified more respectably will be shown in section 4.

3.2 Approximation by Series

If the terms of a sequence are added up, we may ask the question of what value this sum tends to as more and more terms of the sequence are added to the total (if indeed there is such a limiting value). By methods we will not discuss here, it can be shown that the following three *series* (a series is the sum of the terms of a sequence of numbers) are equal to well-known mathematical functions

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \cdots \tag{4}$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots \tag{5}$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots \tag{6}$$

where the exclamation mark is called the *factorial function* and is defined like this

$$n! = 1 \cdot 2 \cdot 3 \cdots (n-2) \cdot (n-1) \cdot n \tag{7}$$

The mathematical function $\cos x$ is equal to the series only when all infinitely many terms are considered. In practise, we cannot sum an infinite number of terms and so we must ask after how many terms it has converged to within a given error margin. To visualize this better, we draw (in figure 2) the function $\cos x$ as well as the series approximating the function by three, four and five terms. After $x \approx 1.6$ we clearly see that the three term approximation becomes bad, after $x \approx 2.4$ the four term approximation also becomes unreliable and the five term approximation is fairly good for the whole interval over which we graphed the functions.

Let us say that we need to be able to compute all three functions above to an accuracy of five decimal places (error margin 0.00001) over the range from $x = -3$ to $x = 3$. A little consideration shows that nine terms for cosine and sine and 17 terms for the exponential function are necessary to achieve that accuracy. We may fairly say that cosine and sine converge equally quickly but the exponential function converges slower than them.

Considerable effort has gone into computing the decimal digits of the number π in the history of mathematics. Many methods exist for doing so and here is a simple series that will do the trick

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots \tag{8}$$

This series converges rather slowly; 200,000 terms are required for five decimal place accuracy (compare with nine for sine and cosine). Because of its slow convergence, we begin to notice the computational effort it requires to work out this number. The few terms for sine are annoying to add up by hand and appear virtually instantaneously on a computer but the terms for π are an impossibility by hand and require noticeable time even for a computer.

3.3 Numerical Simulation

In many circumstances of practical significance (weather prediction for example) a complex system needs to be modelled by quantitative means. Mathematical equations are constructed that describe

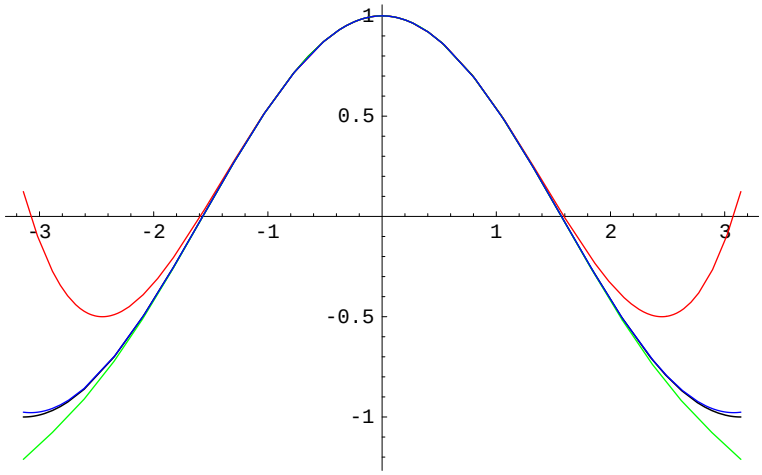

Figure 2: The cosine function with its series approximation over three, four and five terms. Enumerated from the bottom left upwards, the curves are the four-term approximation, the cosine function, the five-term approximation and the three-term approximation.

the system in question and their solution represents the answer to questions such as, "will it rain tomorrow?" Unfortunately most such equations cannot be solved using pencil and paper because they are simply too complicated. That is to say that they cannot be solved using clever functions but must necessarily be solved by numerical calculations that are too lengthy for a human being to do accurately.

Just as with the computation of sine and cosine or the value of π, the accuracy of the computer program that solves the modelling equations is a prime concern. When functions are approximated by their series (as is actually the case for real computers and the above functions) the computer must be told how many terms of the series to take and the user will just have to wait.

Numerical simulations therefore need to satisfy two conflicting constraints. Better accuracy on the one hand raises the resources required and the need for a fast answer as well as economical reasons limit the resources available. Motivated by this, mathematicians search for better methods (faster and more accurate) to solve problems. As a practical example, the movement of water around complex obstacles is so time-consuming to model accurately that it has only been in the last few years that movie animation studios have been able to convince viewers with digitally created water. However there remain features, such as the white water found on rapids in rivers, that still escapes modelling efforts even just accurate enough for movie purposes, not to speak of industrial design. The key is always the speed at which the approximation converges to the real answer.

4 Growth

4.1 Functional Growth

Suppose we have worked out $\sin 3$ accurate to four decimal places and need to get one more decimal place, how many more terms do we need? To work this out, we notice that the next term in the series (in our case $3^{17}/17! \approx 0.00000036$) gives the error on the value computed thus far. Thus, just one more term will do the trick in this case. Indeed, for every further decimal place hereafter, we need to compute at most one more term and sometimes no further term at all. As the number of terms to compute grows more slowly (as we sometimes need no further terms for further accuracy) this is what is known as *sub-linear complexity*. The word "complexity" is used to give an idea of how quickly the effort to compute something grows with the required size of that thing which, in our case, is the number of decimal places required. A linear complexity would indicate that the effort and the required size grow in equal amounts. Here the effort grows more slowly than the size and so we call this sub-linear complexity.

On the other hand, consider working out π. If we have it accurate to four decimal places

(requiring 20,000 terms), by the same argument, we need a further 180,000 terms for just one more decimal place. The total number of terms required for five decimal place accuracy is thus ten times as large as the number of terms needed for four decimal place accuracy. Some reflection shows that this holds true for further decimal places: Every time another decimal place of accuracy is sought, the total number of terms needed is multiplied by ten. As a function, the complexity of this series is then roughly 10^n where n is the number of decimal places needed. Note that this is slightly simplified as this function does not take into account complications that happen for the first few places. Mathematicians call functions of that type an *exponential function* and so this series is said to have *exponential complexity*.

A linear function looks like $y(x) = ax + b$ where a and b are constants. In our considerations of complexity, the constants do not matter so much. Sure enough $10x$ increases faster than $2x$ but it does so at a constant multiple, namely five times as much. More important are functions of different nature and not merely different constants. An example is $y(x) = x^2$. For small values of x, the function $10x$ is bigger than x^2 but not for long; at the point $x = 10$, the function x^2 will overtake $10x$ and continue to rise faster — even at an ever increasing rate.

A function for which the variable is raised by a constant power is called a *polynomial function* such as x^2 in contrast to a function where a constant is raised to a variable power such as 2^x which is called an *exponential function*. We have already seen that both of these types of functions occur in accuracy considerations of series approximations to important functions. Let us try to answer the question which of these types of functions increases faster than the other.

In mathematics, we usually investigate lots of examples using paper and pencil, a computer and some graph paper. After some time, we arrive at a statement about the question that seems to be true of all the cases we tried. This is called a *conjecture*. At this stage we want to establish whether this is true of all possible cases and so we seek for a general argument that gives rise to this statement. If we are successful, this argument is called a *proof* and the statement turns from a conjecture into a *theorem*. The collecting of evidence is a bit tedious and we will spare the reader this work and simply jump to the end and present our theorem and its proof.

Theorem 4.1 *The function x^a increases less rapidly than the function b^x for any constants a and b.*

Proof. A proof is usually called *beautiful* or *elegant* when it is particularly clever and short; if we can use one simple idea to deduce the theorem. Here we will make no attempt at an elegant proof but rather one which is easy to understand.

As we are trying to compare two functions, it is not a bad idea to divide them. So we define the function $f(x)$ to be their ratio

$$f(x) = \frac{b^x}{x^a} \tag{9}$$

As we cannot do much with this just yet, let us take the logarithm of base b of this new function; recall that the logarithm is usually a convenient way to get rid of powers that make life complicated. Thus,

$$\log_b f(x) = \log_b \frac{b^x}{x^a} = \log_b b^x - \log_b x^a = x \log_b b - a \log_b x = x - a \log_b x \tag{10}$$

We know that the logarithm is a steadily increasing function so that if $f(x)$ increases $\log_b f(x)$ will increase also. The last expression in equation 10 clearly increases to infinity as x increases and so the function $f(x)$ increases to infinity with x. As this function was the ratio of the exponential to the polynomial function, we have proven what we set out to prove: The exponential function increases more rapidly than the polynomial one for any value of the constants involved. □

4.2 In How Many Ways?

Comparing the rates of growth of different functions is relevant not only for approximating values of functions but also in determining relative freedom in certain situations. Consider the task of seating n people into n seats around a round table. The study of *combinatorics* investigates questions such as computing how many possible seating arrangements there are for this situation. In our newly discovered theorem-proof manner, let us thus prove what is known as *King Arthur's Last Theorem*:

Theorem 4.2 *There are* $(n-1)!$ *distinct seating arrangements of* n *people on a round table with* n *places.*

Proof. "Conventional wisdom has it that no position at a round table is of any particular significance. There is no 'head' of a round table. If one of the persons so seated is wearing a crown and can order the summary execution of any of the others at the drop of a thumbscrew conventional wisdom would appear to be somewhat superficial." (John Lammin, 1994)

Pick any chair to start seating people. As there is no 'head' to the table, any chair will do just peachily. Clearly you can seat any of the n people there. Then move on to the next chair. As one person is already sitting on the table, you can place any of the remaining $n-1$ people there. And so on all around the merry table until the last person has to sit in the last chair. Therefore, how many ways have we had so far? The product of the number of choices so far gives the number of different final outcomes, $n \cdot (n-1) \cdot (n-2) \cdots 1 = n!$

So far we have done nothing but counting. We have to ask ourselves now whether all of these possibilities are truly different. Suppose you have seated everyone and then ask them all to move one seat to their right. You will get a seating arrangement that is just the same as there is no distinguished seat on the table: If all chairs are the same then it does not matter where you begin counting. Distinct seating arrangements mean that at least some people will have different neighbors and so we have to divide the number of options thus far by the number of times that we can rotate a given seating arrangement around the table. This is clearly n times as there are n seats. Thus the final number of distinct seating arrangements is $(n-1)!$ □

Using the fact that the table does not have a distinguished chair means that we cannot distinguish seating arrangements by the person at the head of the table, we can only do so by their neighbors. This is a so-called *symmetry* of the situation: Rotating the seating arrangement does not change it. Noticing and making use of symmetries is a major theme in all of mathematics and yields some very powerful results.

If we have a party of five people, the number of arrangements is $4! = 24$ a number that we could conceivably list on paper and then choose the best one that suits our needs. There were however 13 knights on the round table of King Arthur which already gives 479,001,600 possible seating arrangements. Enumerating these one by one would take a very long time. In the study of questions of the form "in how many ways ...?" answers usually involve the factorial function and the number of ways rises rapidly as in the case of seating people. This rapid rise is referred to as the *combinatorial explosion* and is a major obstacle for many problems of practical significance.

One problem of practical significance is the *travelling salesman problem*: The salesman has n clients living in different cities and he must make a journey visiting each city once (and only once) returning to his point of origin. Furthermore, he must do it such that the total distance covered is the least possible in order minimize the costs to his employer.

In order to solve this problem, we use an ancient approach of mathematics called *reduction*. In reduction, we look for a problem that we have already solved that has very similar features and try to alter the old solution to give us a new one. Here this is very easy. We need to come up with a list of cities such that no city appears on it twice and the journey is a effectively a circle as the salesman comes back to his point of origin. So this is structurally identical to seating n people around a round table and thus there are $(n-1)!$ routes to choose from. Every seating arrangement of the n cities has an associated cost value, namely the total distance that needs to be covered to realize this travelling schedule. Of all these, we must just choose the one of least cost. Simple! Just list them all and choose the shortest arrangement.

4.3 How Long will it Take?

If our salesman has to visit $n = 20$ cities and we have a table giving the distance between every pair of cities, we need to list $(n-1)!$ possible schedules and compute the total distance for each one. If we can assess one million schedules per second on our computer (more than current computers could do), then it would still take over 3800 years to finish the computation. The time is given by the factorial function which can be approximated according to

$$n! \approx n^n e^{-n} \sqrt{2\pi n} \tag{11}$$

Justifying this formula is beyond the current discussion but it shows that the factorial function grows at least as fast as an exponential function, i.e. quite rapidly. The time cost of computing a travelling salesman solution is therefore exponentially rising as a function of the number of cities involved.

Clearly we need a different method of solution for this type of problem. It is not the speed of convergence that limits us here but the speed of functional growth. In both circumstances the final manifestation of the problem is in the amount of time it takes for us to compute the answer, the speed of approach to the solution as a function of the input size.

To be able to compare the difficulty of such combinatorial problems we need to reformulate them so that they all have the same structure. The usual way to do this is to turn them all into *decision problems*. For the travelling salesman problem this means asking "is there a journey requiring less than x distance to be covered?" The answer to a decision problem is always either yes or no and that gives it a simple structure such that different such problems can be compared.

We distinguish two basic types of problems. Those which can be solved by an algorithm that takes an amount of time given by a polynomial function or less are said to be in the class P of *polynomial-time problems*. The problems for which a guess can be verified to be a solution in polynomial-time are said to be in the class NP of *non-deterministic polynomial-time* (the "non-deterministic" comes from the guess). This second class is a little hard to understand. In the context of the travelling salesman problem consider proposing a route. It is then easy to check whether it is shorter than the distance given in the question by simply comparing the two numbers. Thus there are problems that are easy to solve (P) and problems whose solution is easy to verify (NP). Clearly any problem that can be solved in polynomial-time can be verified in polynomial-time, so that the class P is a subclass of NP.

Within the class NP of problems there are some that have a special property: Every problem in NP can be reduced to these problems by a polynomial-time algorithm. That is to say, if any one of these special problems can be solved, then all problems in NP can be solved using that method augmented by a translation into this special problem. Furthermore, this translation can be done in polynomial-time, i.e. relatively quickly. In short, a solution of any one of these special problems solves all problems in the class NP. As such these problems are called *NP-complete*. The travelling salesman problem is an example of an NP-complete problem. Many problems of practical significance are members of the class NP. To state it pointedly, if the travelling salesman can find his way quickly, we could solve most practical problems quickly as well.

Sadly this is not currently the case. We know that P is included in NP already and ask if P and NP are actually the same class. This would be proven if one could find a polynomial-time algorithm to solve any one of the NP-complete problems such as travelling salesman. Proving that P and NP are different means that one would have to show that it is impossible to solve an NP-complete problem in polynomial-time. To this day, no one has managed to prove either statement and so we do not know whether P is equal to NP.

The problem $P =^{?} NP$ is without question the single most researched mathematics problem in history and recently figures among the seven Millennium problems that will earn their solver one million dollars each. The solution would have dramatic consequences not only for mathematics but also for a host of disciplines that require the solutions of combinatorial problems such as travelling salesman (manufacturing, logistics, computer science, etc.). Fundamentally it is a question about speed: Can certain problems be solved faster than others?

We end with an interesting anecdote. The problem of telling whether a given integer is a prime number or not was always a very famous issue. Much effort and research over centuries have focussed on this question and particularly on the efficient computing of its answer. Until August 2002 this problem was thought to be fundamentally hard, i.e. not solvable by a polynomial-time algorithm (although it is not an NP-complete problem). Two recent bachelor graduates together with their thesis advisor at the Indian Institute of Technology Kanpur convinced the world that this problem can be solved in polynomial-time. Moreover, the method they provided is very short and simple given that this problem had taxed many of the most talented minds in mathematics since Erathostenes (284–202 BC).

Research on speed in mathematics is far from over and much remains to be done, God-speed!