Elementary Numerical Analysis

Prof. Rekha P. Kulkarni
Department of Mathematics
Indian Institute of Technology, Bombay

Module No.# 01

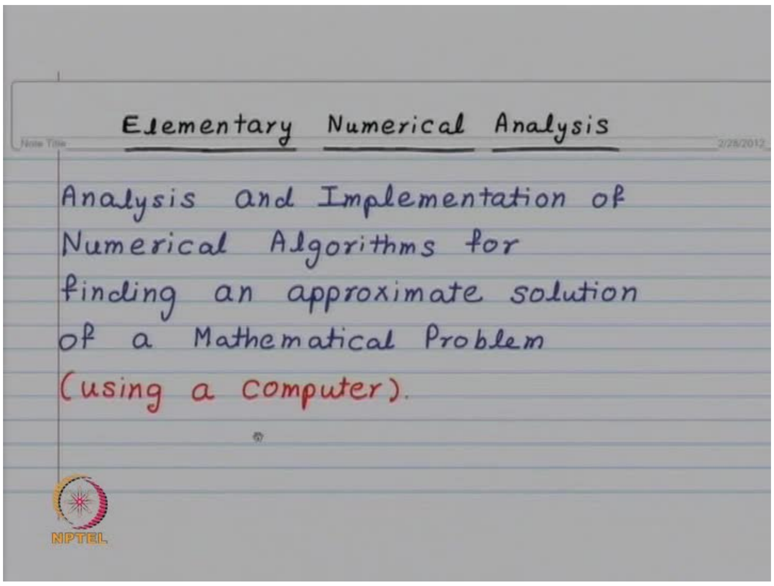
Lecture No. # 01

Introduction

In numerical analysis, we are mainly interested in implementation and analysis of numerical algorithms for finding an approximate solution to a mathematical problem. The solution in general is approximate because we use computer, and when you are using computer, you have got only finite precision and you have only finite number of steps - you perform only finite number of steps.

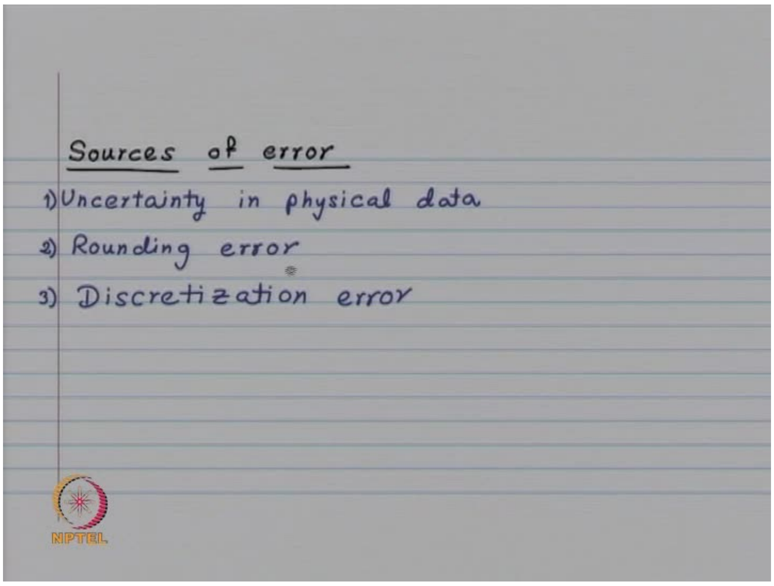
Also, the method which you are using that itself may be iterative method or even if it is a direct method then, because of the finite precision that is going to be an approximate solution. So, when one looks at the sources of error which are in the numerical analysis or in numerical algorithms then, they can be because of the uncertainty in the data or it will be rounding off errors that is because of the finite precision or there can be discretization error.

(Refer Slide Time: 01:48)

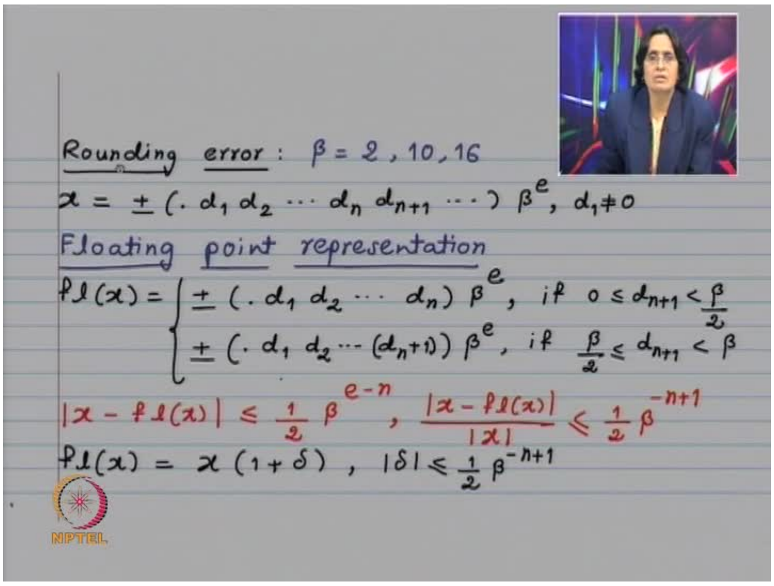


So, we are interested in this analysis and implementation of numerical algorithms. And main idea is using a computer one finds an approximate solution.

(Refer Slide Time: 02:01)



(Refer Slide Time: 02:09)



And these are the sources of datasources of error. So, let us look at first the rounding off error. You have a number, let us look at real number, it has this representation to the base beta, the base beta is either 2 or 10 or 16, so these are general values of beta, which are used on computer.

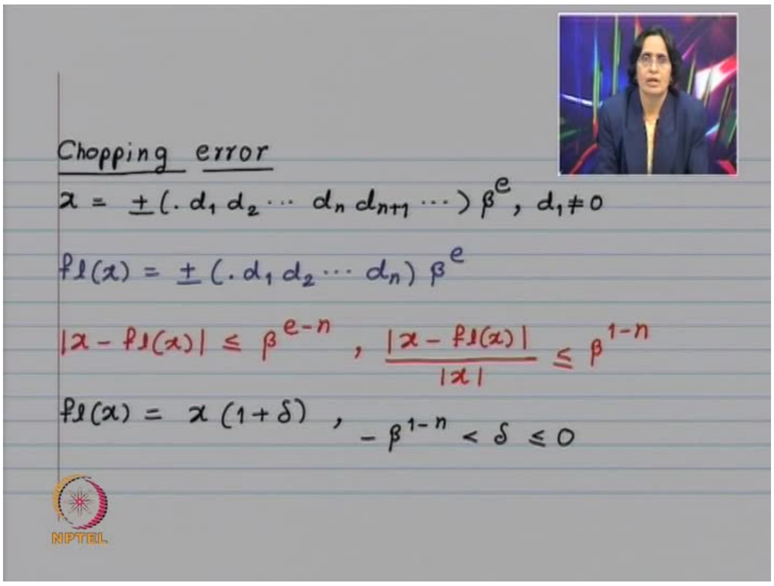
So, real number x, has this it is of this form, where as a normalization one assumes that d 1 is not equal to 0. It is always possible to do such a thing for a non-zero number, because you can adjust the exponent. Now, floating point representation of this is, as I said, we have only finite number of digits available on computer.

So, you look at value of d n plus 1, when the base is beta, all these digits they are going lie between 0 to beta, they are going to be strictly less than beta. So, if d n plus 1, if it is less than beta by 2, then one writes this as point d 1, d 2, d n into beta raise to e, or if d n plus 1 lies between beta by 2 and beta, then 1 increases the value of d n to d n plus 1, so this is when you are using rounding off.

Now, it is easy to see that x is our number, floating point of x, that is, the floating point representation, that is going to be less than 1 by 2 into beta raise to e minus n. Now, this is absolute error, the more relevant error is to consider relative error, that is, modulus of x minus floating point of x divided by mod x, this will lie between or this will be less than or equal to 1 by 2 into beta raise to minus n plus 1.

This same thing one can represent as floating point of x is equal to x into 1 plus delta and modulus of delta is going to be less than or equal to 1 by 2 into beta raise to minus n plus 1. So, this is there is a error, x was our real number, when you represent it on computer, you are going to consider floating point representation and then, you introduce a error. That relative error is less than or equal to 1 by 2 beta raise to one minus n, but there is a error.

(Refer Slide Time: 05:24)



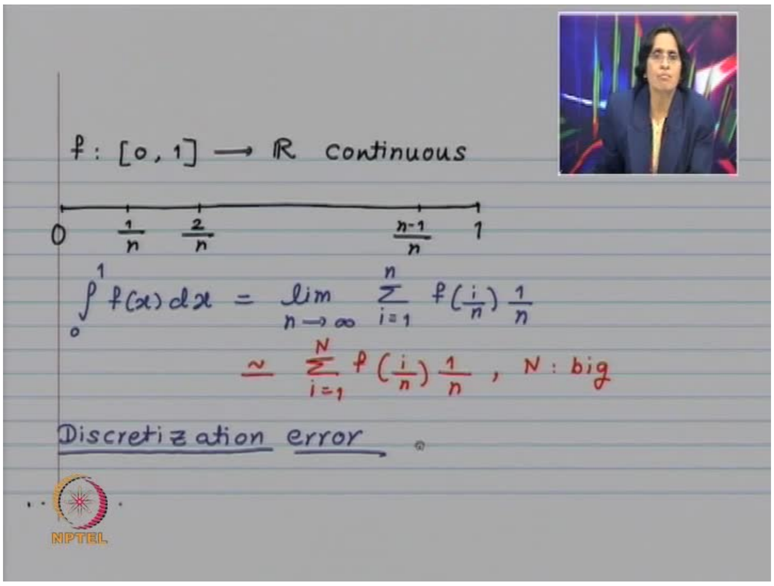
And this error, you are going to perform various operations, so it will keep accumulating. Now, instead of rounding off on some computers what is done is, chopping off. So, the chopping is This is our number, so irrespective of value of d n plus 1; d n plus 1 will be such that 0 less than or is equal to d n plus 1 less than beta.

So, you consider floating point of x to be… just forget this d n plus 1 on words, so it is plus or minus that is the sign point d 1, d 2, d n beta raise to e. And in this case, the relative error is less than or equal to beta raise to 1 minus n. When it is rounded off, it was 1 by 2 beta raise 1 minus n, so they are similar. So, in this case, floating point of x is going to be x into 1 plus delta, this delta is going to be always negative in this case and it will be bigger than minus beta raise to 1 minus n and less than or equal to 0.

So, this is about the finite precision that on a computer you can represent a number by using only finite number of digits. Another source of error that is the discretization error, when one defines mathematically a quantity, for example, when one considers integral a to be f x dx.

If f is a continuous then, one defines it as limit of Riemann sums, so when one considers limit of Riemann sums, it is limit as n tends to infinity, where n is number of intervals - sub intervals of our interval a b.

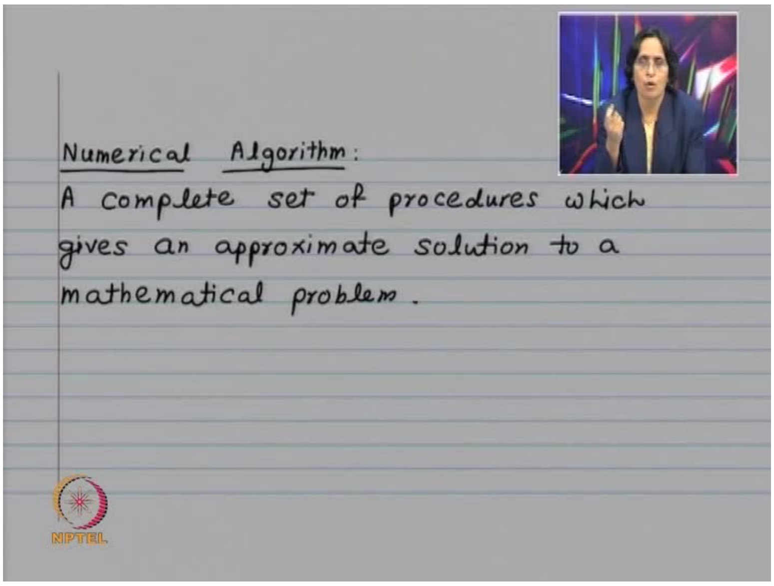
(Refer Slide Time: 07:52)



Now, mathematically, this definition is fine, but when I bound to calculate integral a to b f x dx using a computer then, i can have only finite number of steps at my disposal and then that will introduce discretization error. So, when you replace infinite process by a finite process which you have to when you are using computer that introduces the discretization error. So, you have f to be a continuous function, say, defined on interval 0 to 1, all continuous functions are integrable, you subdivide the interval 0 to 1 into n equal parts, then integral 0 to 1 f x dx is limit n tending to infinity summation i goes from 1 to n f of i by n into 1 by n, value of our function f at nodes 1 by n 2 by n and 1 multiplied by length of the sub interval that is 1 by n.

Now, when you are not using computer, then for some special functions one can evaluate this limit, but what we want is, given function our computer should tell us what is integral 0 to 1 f x dx. So, this limit will be approximately equal to summation i going from 1 to N f of i by n into 1 by n, one chooses N to be big enough. So, that will depend on how big accuracy or how accurate the solution new one, so this is going to be the discretization error, so these are the two main sources of error.

(Refer Slide Time: 09:29)



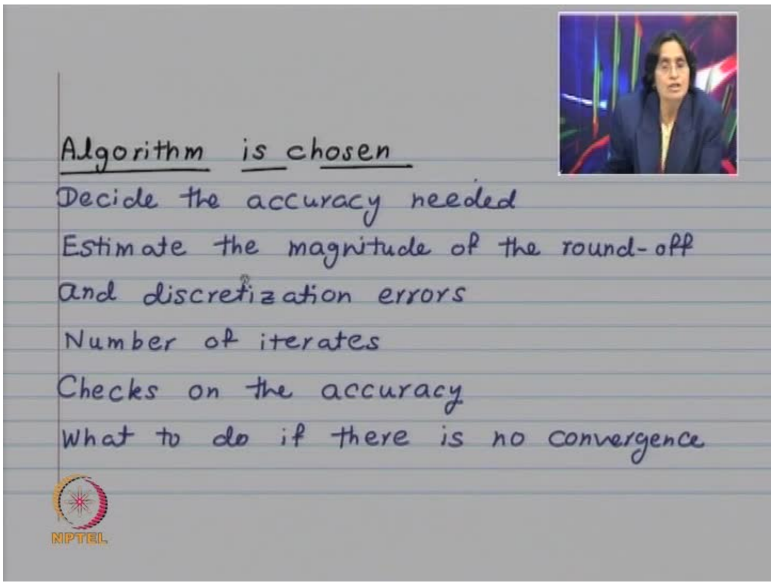
You have rounding off or truncation error and you have discretization error. Now, a numerical algorithm that is nothing but it is a complete and unambiguous set of procedures which gives an approximate solution to a mathematical problem. Now, there are various mathematical problems, one problem was finding integral a to b f x dx or it can be finding the derivative value or it can be solution of a differential equation and so on.

So, numerical algorithm, it tells us that how I should proceed, so I have a set of instructions and that gives us a method for finding an approximate solution. Now, once you choose an algorithm, what you have to do is, you have to decide the accuracy, how much accuracy you want. Then, you have to decide the number of iterations, because generally your method is going to be a iterative method or the finite number of steps.

Then, some bounding of the error, like there is going to be some error, so whether for a given problem I can know beforehand that my modulus of error is going to be less than or equal to something, or also we have to keep track of rounding off errors that you are performing. Each number is going to be represented only using finite number of digits and then, you do various operations, so the operations they are going to be addition, multiplication, subtraction and division. These are going to be the 4 operations which we will be doing.

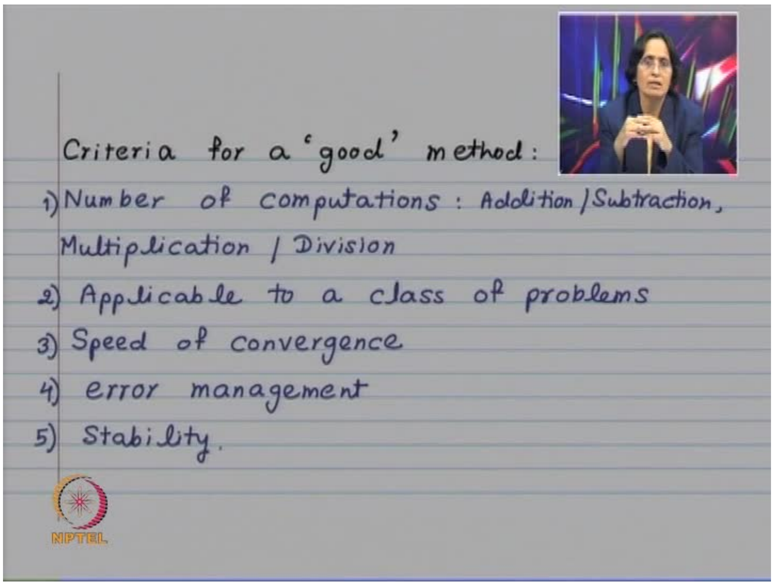
So, then the rounding off error that will go on adding, so one has to keep track of that then, also in case the method is not converging then one has to keep some, say, way out that ok you perform these many number of iterations, if it does not work then think of something again.

(Refer Slide Time: 12:00)



So, here are the main points that you decide the accuracy which is needed, then magnitude of round-off errors and discretization errors, you have to find an upper bound, the number of iterates you fix, then check on accuracy and then what to do if there is no convergence.

(Refer Slide Time: 12:40)



So, algorithm will be a set of procedures and then, these are the things one decides or these are the things which need to be taken care of. Now, there can be various methods to solve a problem, now criteria for a good method that is going to be number of computations. So, how many additions or subtractions and how many multiplications or divisions one needs to do. When one calculates the number of computations, addition and subtraction, they are considered on par, and multiplication or division they are considered on par. Our method it should be applicable to a large class of problems, it cannot be specific just to a very narrow class of problems, so applicability then how fast there is going to be convergence.

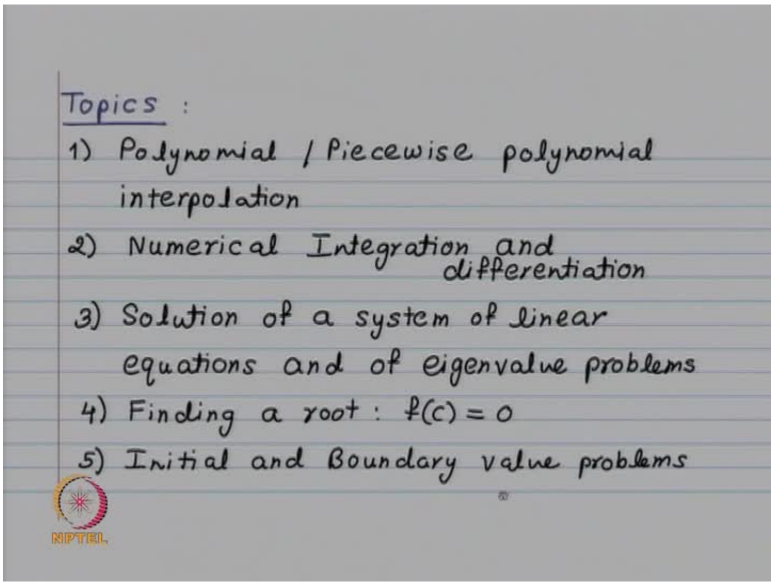
Like if you have two methods, then one considers the speed of convergence of the two methods. Of course, the faster method will be preferred, but then generally there is a tradeoff that if the method converges fast then, you are doing more work, so one has to look at the balance.

Then, about the error management, that you have got round-off error, you have got discretization error, so then for a particular method whether one can have some idea about a error. And then the fifth point is stability, so about the stability I will say little later, so these are the criteria for a good method. Now, in this course the topics which we are going to consider they are going be… first we will look at polynomial and piecewise polynomial interpolation. The topics which follow, they are going to depend on this polynomial interpolation.

So, we will first consider polynomial approximation, then the idea is, you are going to replace a continuous function by a polynomial and the polynomials, one can integrate and one can differentiate using computer. So, if you have a function f, if you find it is integral replace it by a nice function, such as polynomial and that will give raise to numerical integration, similar idea will be used for numerical differentiation.

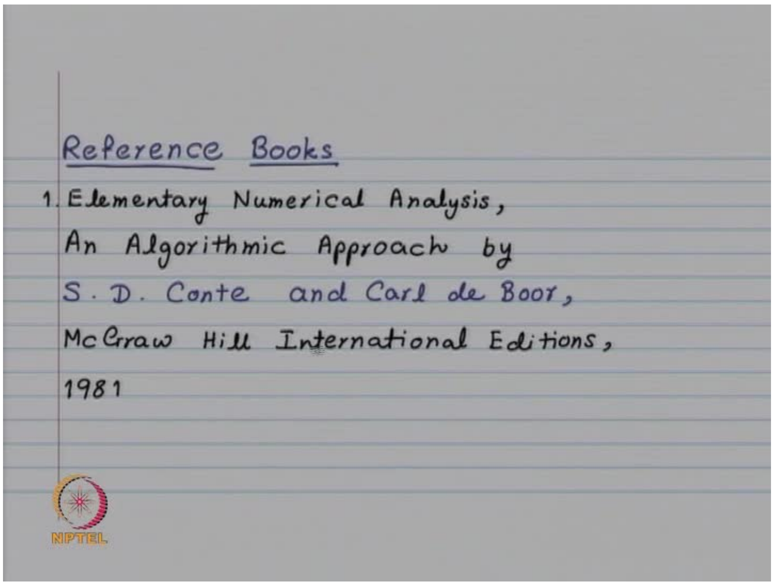
Then, we will also be considering approximate solution of system of linear equations and associated Eigen value problem. Next, we will consider the route finding method for a non-linear equation and the last topic will be the solution of boundary value problems and initial value problem.

(Refer Slide Time: 16:19)

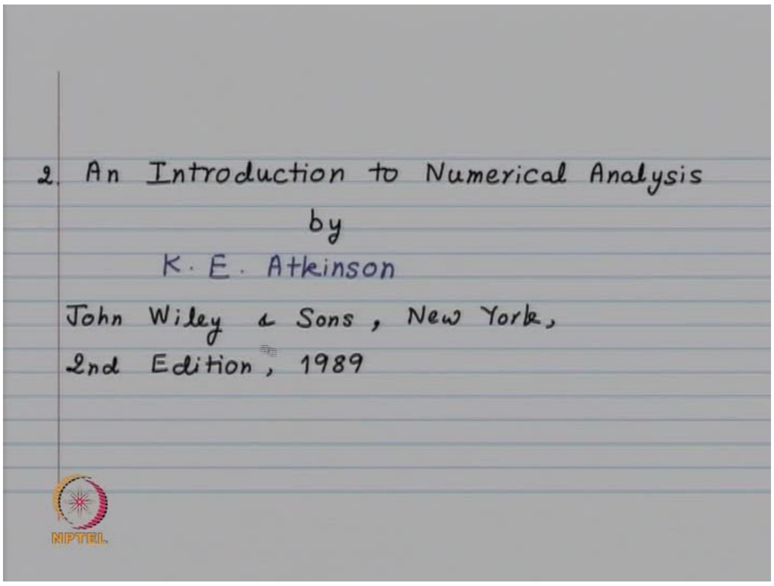


So, here is polynomial and piecewise polynomial interpolation the first topic. Based on that you will have numerical integration and differentiation, then we will consider the direct methods and iterative methods for solution of system of linear equations and for Eigen value problem, they are going to necessarily iterative methods. We will have approximate methods for finding the 0 of a function and there again we will be using polynomial interpolation.

(Refer Slide Time: 17:01)



(Refer Slide Time: 17:23)



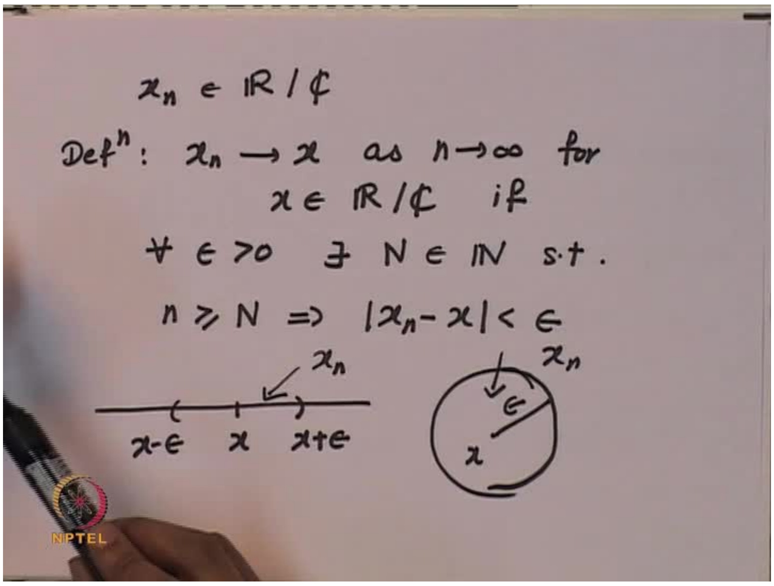
And lastly the differential equations, so in that we will be considering initial value problem and boundary value problem. The lectures will be based on mainly material from the following two reference books, so the first one is book by Conte and Carl de Boor. So, it is elementary numerical analysis and their emphasis is on algorithms it is available in the international editions and a book by K. E. Atkinson introduction to numerical analysis which is published by john wily in 1989.

So, now, we are going to look at some mathematical preliminaries. See now, there are many packages available for finding solution of various problems. The reason why one studies numerical analysis is that, you should be able to write your own algorithms or even if you are using a standard package, you should be aware of its limitation.

That you are going to use computer, so you are going to get approximate solution, so how reliable that approximate solution is. You should be able to write a adapt an algorithm which is specific to your problem, that an algorithm may be available for a general class of problem, but for your special problem, may be you can make some changes and you can have a better method. So, that is why one is interested in analysis of various algorithms.

Now, this analysis is done by using of course results from mathematics, so I am going to review some of the basic results from mathematics. What is assumed is that you have familiarity with properties of real valued functions with real variables or complex valued functions and a bit of basics of linear algebra. So, the first result which I want to recall is, about convergence of a sequence.

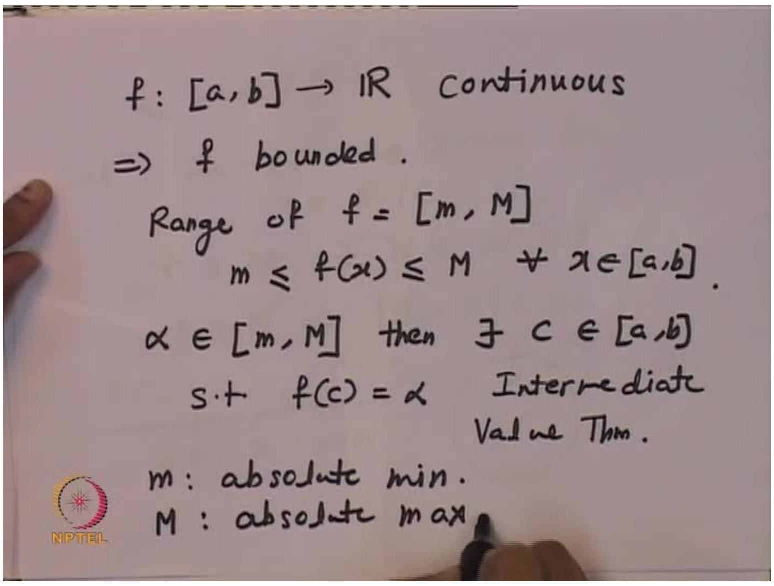
(Refer Slide Time: 19:42)



So, we are going to consider a sequence of real numbers or complex numbers and I am just going to recall the definition, so we have got x n, it a sequence of either real numbers or complex numbers, then definition x n converges to x as n tends to infinity for x belonging to R or C, if for every epsilon greater than 0.

There exists a natural number N such that whenever n is bigger than or equal to capital N, it implies that modulus of x n minus x is less than epsilon. So, we have suppose x is a real number and you are dealing with sequence of real numbers, then what we are saying is, look at the interval x minus epsilon to x plus epsilon. So, after a certain stage all the elements x n, they should be in this interval, if your x is complex then, you are looking at a disk with center x and radius epsilon. So, your x n after the certain stage they should lie in this disk with center x and radius epsilon.

(Refer Slide Time: 21:44)

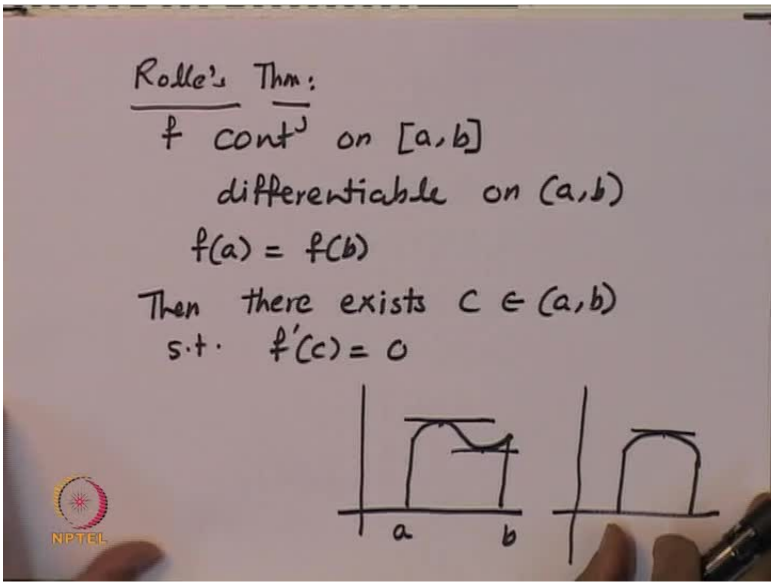


So, that is the definition of convergence of sequence of real numbers. Now, next thing I want to do is, I want to consider continuous function defined on interval a b, closed interval a b, taking real values and recall some of the basic properties of continuous function. So, our function f is defined on interval a b closed and bounded taking real values and it is a continuous function.

In that case, f is going to be bounded, range of f is going to be interval small m to capital M. So, our m is less than or equal to f x less than or equal to capital M, for all x belonging to interval a b. Now, if I take alpha to be any number between capital M to small m, then there exists c belonging to a b such that f of c is equal to alpha.

In fact, when I write range of f is equal to closed interval m to capital M, it precisely means that any number which lies between small m and capital M, that is going to be in the range, so alpha lying in this interval is equal to f of c, this is known as intermediate value theorem, and small m is known as absolute minimum and capital M is known as absolute maximum, so this is about continuous functions. Now, suppose your function f is differentiable, then the basic theorems for differentiable functions, they are Rolle's Theorem, Cauchy's mean value theorem and Taylor theorem.

(Refer Slide Time: 24:03)

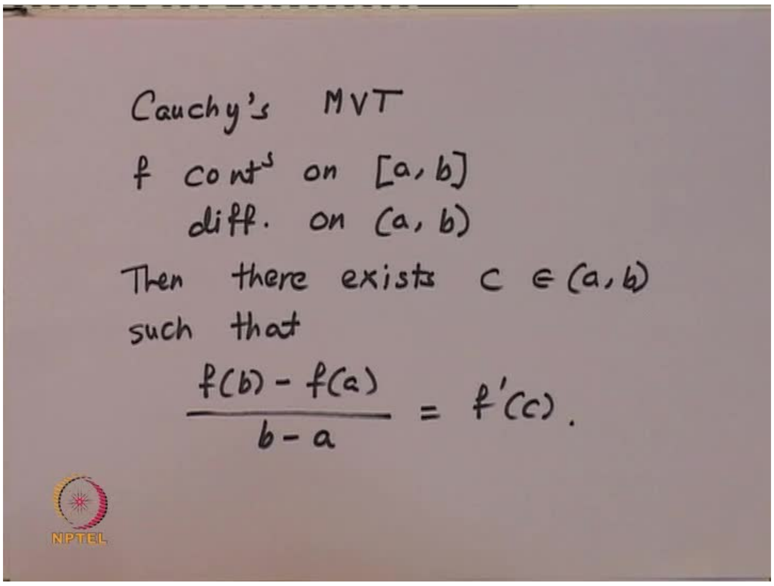


So, the Rolle's theorem is, if you have f to be continuous on closed interval a b, I am assuming f to be real valued, differentiable on open interval a b and f of a is equal to f of b. If these three conditions are satisfied then, the conclusion of the theorem is, then there exists c belonging to a b such that f dash c is equal to 0. So, graphically what it means is, suppose I have got a function defined on interval a b, such that it is continuous on closed interval a b, differentiable on open interval a b and f a is equal to f b.

So, suppose, the function graph of the function is something like this, then there exists c belonging to a b such that f dash c 0; f dash c 0 that we mean that at point c, the tangent is horizontal. So, here you have tangent to be horizontal, here we are saying then there exists c there can be more than such c. For example, here at this point also, the tangent is going to be horizontal, if I look at some different function, say, a function like this, then I am going to have a unique c such that f dash c is equal to 0.

So, it depends on your function, but what we know is, there will be definitely one c such that f dash c is equal to 0, so this is Rolle's theorem. And now, this is Cauchy's mean value theorem, it is very important and we will be using it several times during the course.

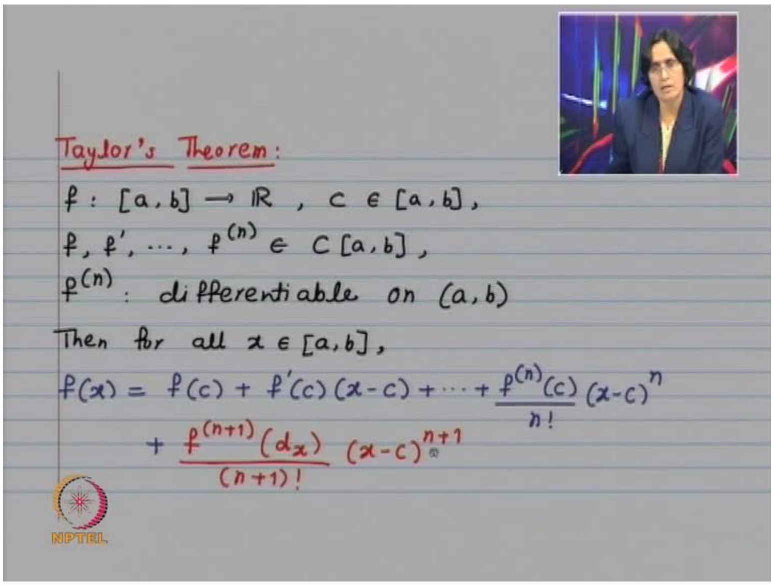
(Refer Slide Time: 26:58)



So, for the Cauchy's theorem, the conditions are again, it should be continuous on closed interval a b, differentiable on open interval a b then, when I look at the quotient f b minus f a upon b minus a that will be equal to f dash c for some c in the interval a to b. And later on, we will see the applications or where one needs this Cauchy's theorem, so we have Cauchy's mean value theorem.

So, you have f to be continuous on closed interval a b, differentiable on open interval a b, then there exists c belonging to open interval a b, such that f b minus f a upon b minus a is equal to f dash c. If f a is equal to f b as in Rolle's theorem, then f dash c will be 0 that was the conclusion of Rolle's theorem.

(Refer Slide Time: 28:20)



So, now, we are going to look at Taylor's theorem, so mean value theorem will be a special case of taylor theorem. In mean value theorem, we assumed f to be once differentiable. Now suppose, f is n plus one times differentiable, then we have got the Taylor's theorem and the statement of that Taylor's theorem is f from a b to R and c is a point in interval a b, our assumption is function f, its derivative f dash and it is n th derivative, they are all continuous.

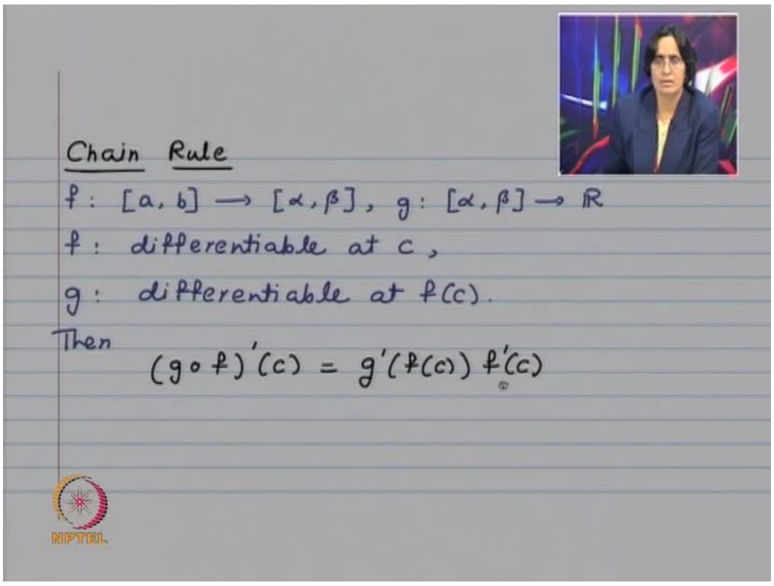
So, the first n derivatives they are continuous, and f n is differentiable on open interval a b. So, if you put n is equal to 0 then, that means f continuous and continuous on closed interval a b and differentiable on open interval a b, which was the assumption in mean value theorem.

If this happens, then for all x belonging to a b, one can write f of x is equal to f of c plus f dash c into x minus c plus n th derivative of f at c upon n factorial x minus c raise to n plus n plus first derivative evaluated at some point d x, this is going to be a point in the open interval a b divided by n plus 1 factorial into x minus c raise to n plus 1.

Our point c is fixed, it can be any point in the interval a b including the end points a and b. So, look at the expression in blue, c is fixed, so it is f of c plus f dash c into x minus c plus f n c upon n factorial x minus c raise to n, x can vary over interval a b. So, this is going to be a polynomial in x, whereas the expression in the red, that is the reminder, here, this f n plus 1 and this point d it depends on x, it also depends on c, but c is fixed.

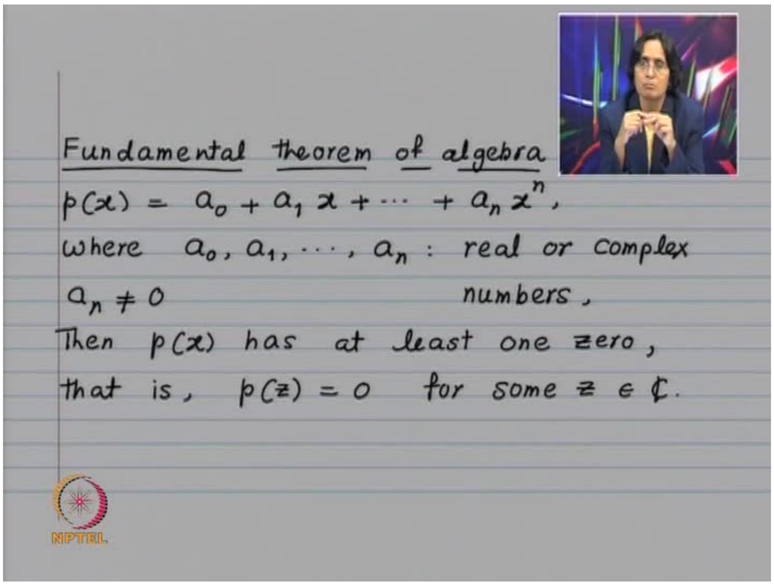
So, f n plus 1 d x into x minus c raise to n plus 1 n plus 1 factorial, so the dependence over x is here as well as here. So, you are writing effects as a polynomial and then there is going to be some remainder, so this a very basic theorem and that is the Taylor's theorem.

(Refer Slide Time: 31:18)



Now, another rule, it is about the derivative of a composite function, so suppose your function f is defined on interval a b and range is contained in interval alpha to beta. The function g is defined on alpha to beta and taking real values, so i can talk of g composed with f, if your function f is differentiable at c, where c is some point in the interval a b. And if g is differentiable at image of c by f, then the g composed with f it is derivative at c will be given by g dash of f c into f dash of c. So, these were the results about the function defined on closed interval a b and we will be assuming generally the function to be continuous.

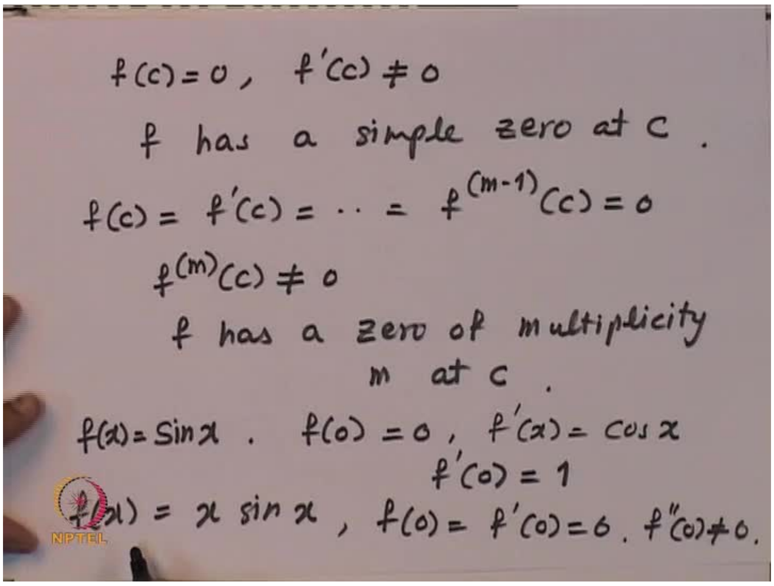
(Refer Slide Time: 32:43)



Now, the next thing I want to review, it is about the fundamental theorem of algebra. So, the theorem is, you look at a polynomial p x, which is a 0 plus a 1 x plus a n x raise to n, x can be either real or complex. The coefficients a 0, a 1, a n these also can be real or complex numbers, then let us assume that a n is not equal to 0, that means, we are looking at polynomial of degree n. Then, the fundamental theorem of algebra, it tells that p x has at least 10. Now, even if a 0, a 1, a n they are real, still what we can guarantee is that it has a root, with that root it may be complex. So, even for a real polynomial the root can be complex, so this is the fundamental theorem of algebra.

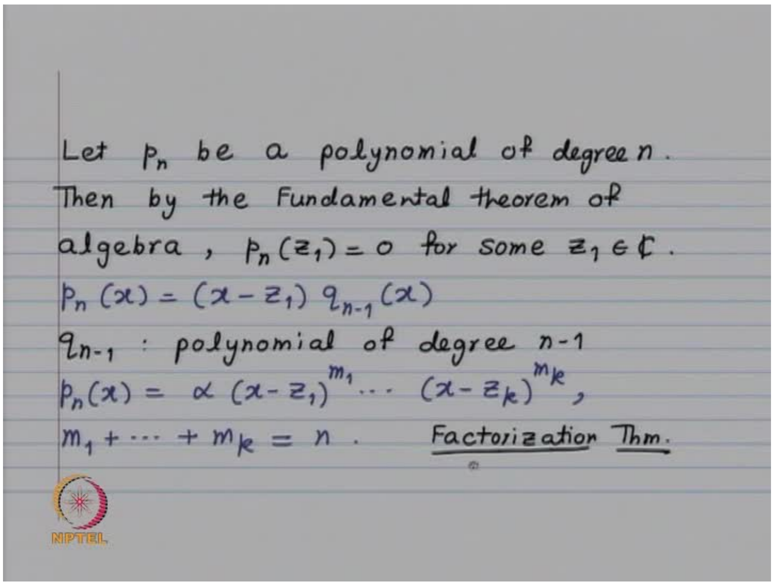
Now, from this theorem, it is easy to deduce that a polynomial of degree n is going to have exactly n roots, here we are saying that it has at least 10, and now we want to say that it has got exactly n zeros. Now, those zeros they can be repeated, so let me define what is a simple zero and what is a multiple zero.

(Refer Slide Time: 34:22)



So, the definition is, you have a function f, so if f of c is equal to equal to 0 and f dash c not equal to 0, then we say that f has a simple 0 at c. On the other hand, if you have got f of c is equal to f dash c is equal to f m minus 1 c is equal to 0 and f m c is not equal to 0, then we say that, f has a zero of multiplicity m at c. For example, look at sin x f x f x is equal to sin x, f of 0 will be 0, f dash x it is going to be equal to cos x. So, f dash at 0 is going to be equal to 1, so that means, f has a simple 0 at 0. On the other hand, if I look at f x is equal to x sin x, then f of 0 is equal to f dash of 0 is equal to 0 and f double dash 0 is not equal to 0, so this function f will have a double 0.

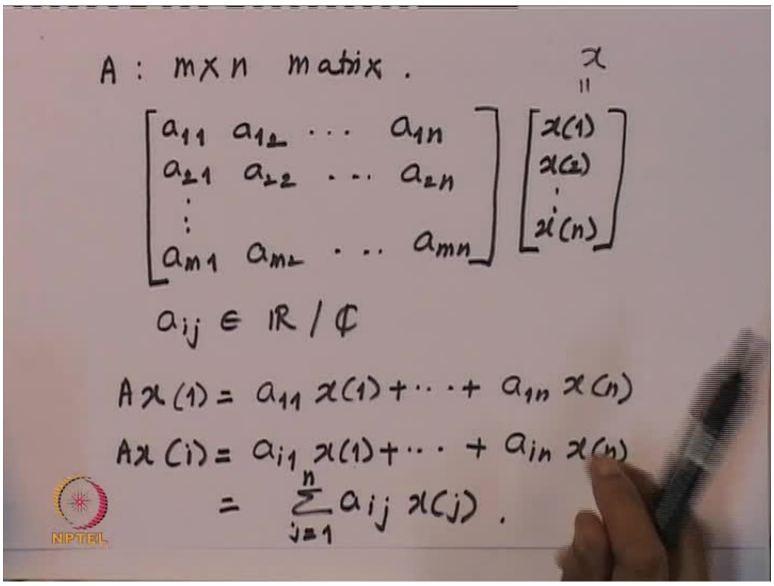
(Refer Slide Time: 36:09)



So, now, look at p n to be a polynomial of degree n, that means, it is of the form a 0 plus a 1 x plus a n x raise to n a n not equal to 0. Then, by the fundamental theorem of algebra it has one 0 or it has a root, so p n z 1 is equal to 0 for some z 0 belonging to c. Now you can factorize, so p n x is equal to x minus z 1 into q n minus 1 x. Now, this q n minus 1 it will be a polynomial of degree n minus 1, p n is a polynomial of degree n we are writing it as x minus z 1 into something, so that something has to be a polynomial of degree n minus 1. Now, q minus 1 will have a 0, so you apply fundamental theorem of algebra to q n minus 1. And then the 0 of q n minus 1, it can be either z 1 or it can be something else. So, using this argument, you get p n x to be equal to alpha into x minus z 1 raise to m 1 into x minus z k raise to m k, where the powers m 1 plus m 2 plus m k, they will add up to n and alpha is going to be a constant.

So, this is a factorization theorem and this tells you that p n has exactly n 0s, provided you count z 1 as repeated m 1 times, z k as repeated m k times, so z 1, z 2, z k are distinct 0s, but when you add the multiplicity that is going to be equal to n. So, once again, this is a basic result which we are going to need later on. Now, we looked at some properties of continuous functions, continuous and differentiable, then fundamental theorem of algebra which tells us about the roots of a polynomial. And now, I want to recall the notation and basic things or basic properties about matrices.

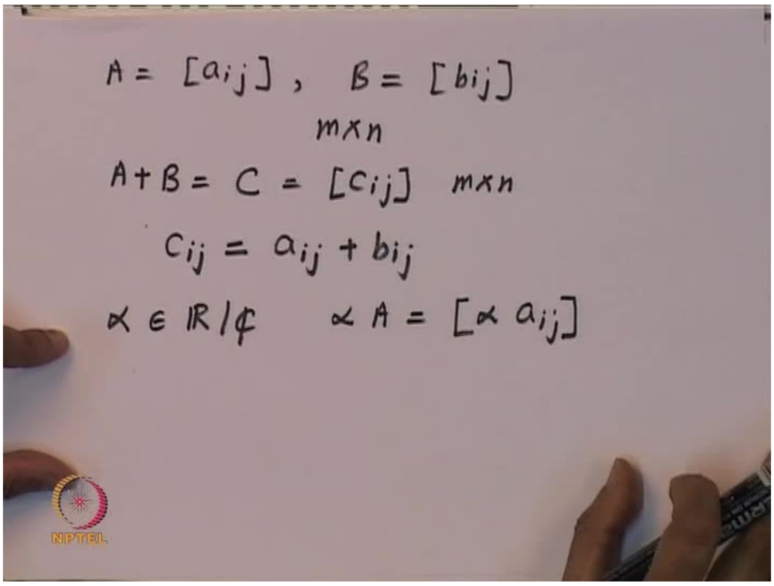
(Refer Slide Time: 38:53)



So, you were matrix A and let us consider the matrices with the entries either as real numbers or complex numbers, so a is m by n matrix, that means, it is an array with m rows, so you have got a 11, a 12 up to a 1 n, then a 21, a 22, a 2 n and a m 1, a m 2, a mn. So, here a i j they belong to either real numbers or complex numbers, i denotes the row index, j denotes the column index. So, this is a m by n matrix, if you have a vector which is say x 1, x 2 up to x n, so it is a column, vector m by n matrix and then this is our x then one defines A into x. So, this A into x it is going to be a m by 1 vector and the matrix in to vector multiplication is the first entry A x 1, the component of A x its first component is given by a 1 1 x 1 plus a 1 n x n. In general, A x its i th component will be given by, you have to look at i th row, so it is a i 1 x 1 plus a in x n.

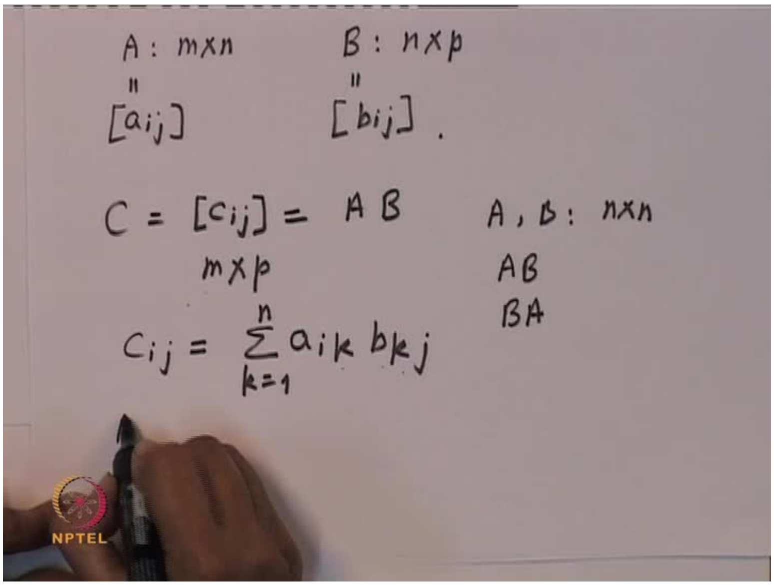
And this we will be writing in a compact notation as summation j going from 1 to n a i jx j, so we have defined a matrix in to vector multiplication. Then, we can consider the two matrices and their multiplication. If you have got two matrices of the same size m by n, then you can add the addition is done by component wise you can multiply a matrix by a scalar, so that is you multiply each entry by that scalar alpha. If you are looking at matrices, so one matrix of the order m by n another of n by p.

(Refer Slide Time: 42:02)



So, the first one has m rows and n columns, second one has n rows and p columns, then you can multiply these two matrixes. So, you have, say A is equal to a i j, b is equal to B i j, both of the size m by n, then addition is A plus B, it will be a matrix C which is again of the same size m by n and then c i j is nothing but a i j plus b i j, add the corresponding entries.

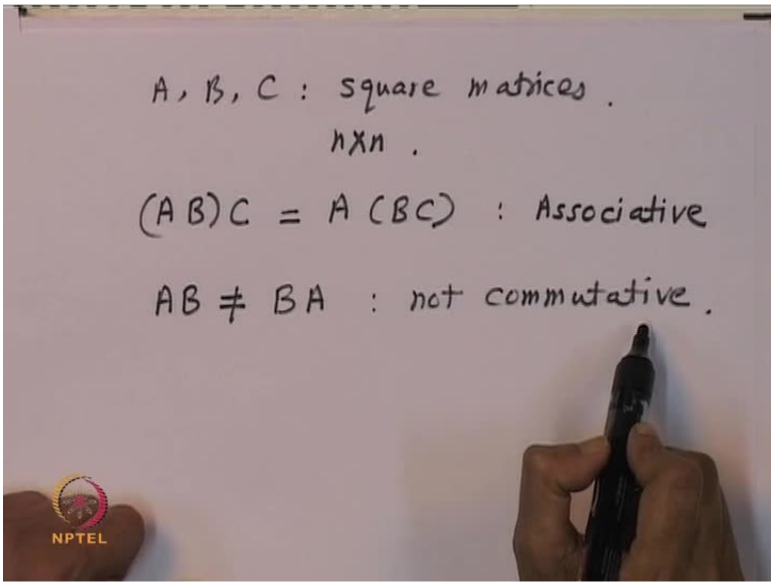
(Refer Slide Time: 43:09)



If your alpha is either a real number or a complex number, then alpha times A is going to be alpha times a i j, you multiply each entry by alpha. So, if you have got 2 matrices of the same size you can add them or you can multiply by a scalar. Now, let us look at the multiplication of 2 matrices, so if you have A to be a m by n matrix and B to b a n by p matrix, so which is given by b i j.

Then C, which is c i j, it is a multiplication of matrix A into matrix B, where the size of C is going to be given by m rows and p columns and the entries c i j they are given by summation a i k b k j, k going from 1 to n; a i k, this k refers to the column of a and this k refers to the row of b, so that is why you need A to be m by n and B to be a n by n.

(Refer Slide Time: 44:44)



If you are taking square matrices then, there will be no problem if you have got A and B to be a n by n matrix, then you a talk of A into B you can also talk of B into A. If you have got a general matrix, you can talk of A into B but not of B into A, so if you have got - say - A, B, C to be square matrices, so A, B and C these are square matrices, so all of them of the same size n by n, then I can consider A, B into C.

This is going to be same as A into B C, so this is known as associative property, that so long as you respect the order you are going to get the same result here.You are first multiplying A and B and then multiplying by C later on, here you are multiplying B and C first and then, pre multiplying by A, then the result is going to be the same. If you look at A into B, this is in general not equal to B into A.

So, it is not commutative, so this is a very important point that matrix multiplication is not commutative; it is associative, but it is not commutative. So, these where some of the basic results, which we will be using throughout the course. Now, there will be some more results, but what I am going to do is, I am going to prove or I am going to quote when we need those.

So, now, what we want to do is, suppose, you are given a function f, we want to approximate it by a nice function. Look at polynomial, so the polynomial is of that form a 0 plus a 1 x plus a n x raise to n for some n, so if I want represent this polynomial in a computer what I will need to do is, store the coefficients a 0, a 1, a n that contains all the information about my polynomial.

This is not case with a general function, suppose, I am looking at a function defined on interval A B, then I need to tell what value of f is at each point x belonging to interval A B there are infinitely many points. On the other hand, the for a polynomial I store a 0, a 1 a n, if I want to evaluate it at some point I know that it is I have to look at a 0 plus a 1 multiplied by a x plus a 2 multiplied by x square plus a n multiplied by x raise to n.

So, for storing the polynomials they are very useful in computer, also those are infinitely many times differentiable, and derivative again it is a polynomial, it is going to be a polynomial of 1 degree less, so for the derivative once again I can use computer. If I want to integrate then, I will get again a polynomial with one degree higher and if you are looking at indefinite integral there will be some constant of integration, so on a computer what you can do is, addition, subtraction, multiplication.

So, using these three numbers what you can have is, you can represent a polynomial. You can have also division, then if you use division then what you can represent is the rational functions, that means, the quotient of 2 polynomials. And you have one more thing available on computer that is the comparison. So, that will give us possibility of piecewise polynomial or piece wise rational function, but the polynomials they had a smooth functions, they are infinitely many times differentiable any continuous function can be approximated by a sequence of polynomials, so that is why we will be considering approximation of a continuous function by polynomial.

Now, what we are going to do is we are going to look at say various ways of approximating a function by a polynomial and one of that is going to be interpolating polynomial, so this we will do in our next lecture, thank you.