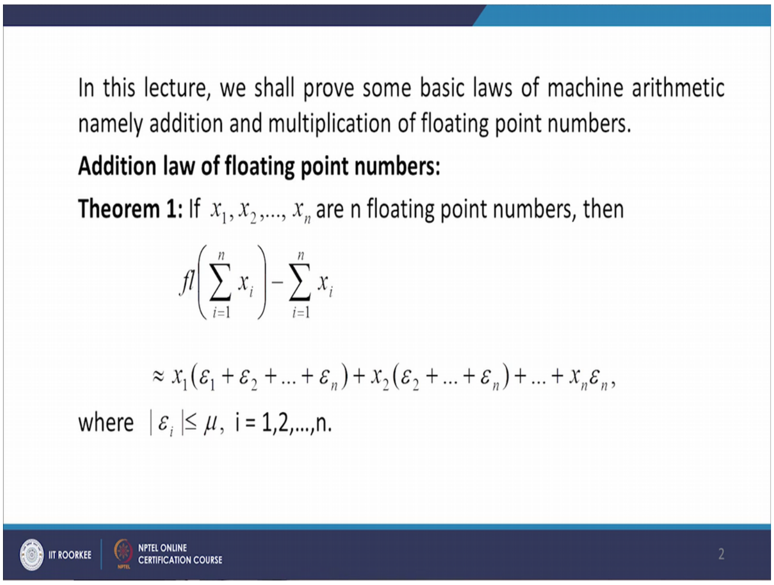
Numerical Linear Algebra

Dr. P. N. Agrawal
Department of Mathematics
Indian Institute of Technology, Roorkee

Lecture - 24
Addition and Multiplication of Floating Point Numbers

Hello friends, welcome to my lecture on Addition and Multiplication of Floating Point Numbers.

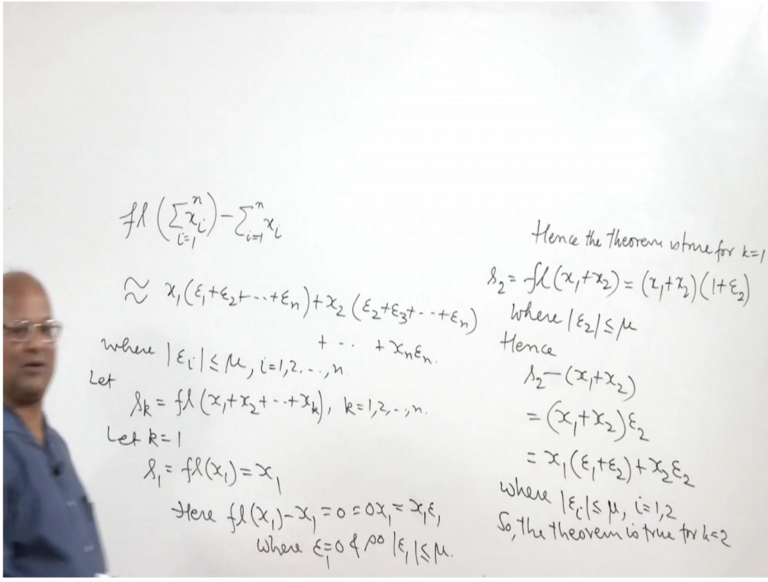
(Refer Slide Time: 00:27)



In this lecture we shall prove some basic laws of machine arithmetic namely addition and multiplication of floating point numbers. Suppose x 1, x 2, x n are n floating point numbers then we shall show that floating point representation of sigma i equal to 1 to n x i minus sigma i equal to 1 to n x i is approximately equal to x 1 into epsilon 1 plus epsilon 2 and so on, epsilon n plus x 2 into epsilon 2 plus epsilon 3 and so on, epsilon n and so on x n into epsilon n there epsilon i are the errors relative errors in the floating point representation of x i.

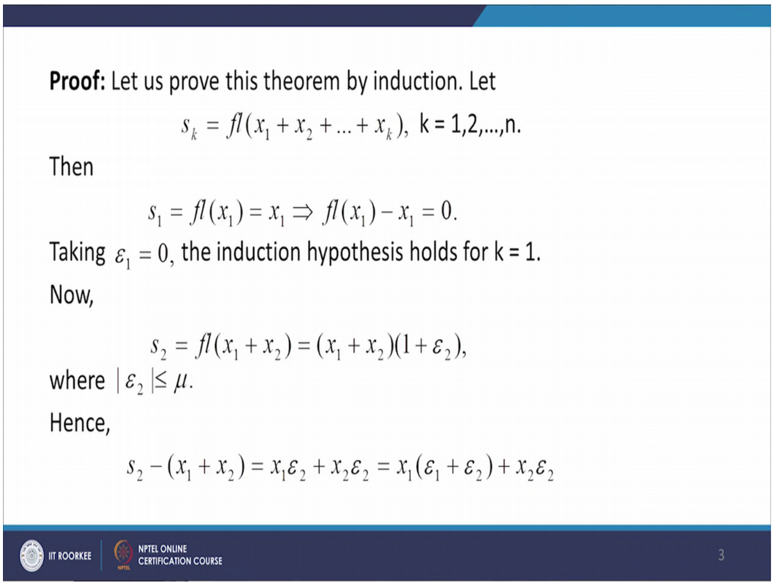
So, we know that mu is machine precision. So, mod of epsilon i is less than or equal to mu for all values of i from 1 to n. So, what we have to show is this floating point representation of sigma x i i equal to 1 to n minus sigma i equal to 1 to n x i where x 1, x 2, x n are n floating point numbers.

(Refer Slide Time: 01:23)



So, then when you find the floating point representation of sigma x i i equal to 1 to n and subtract from it sigma i equal to 1 to n x i the error involved is given as x 1 times epsilon 1 plus epsilon 2 and so on epsilon n, then x 2 times epsilon 2, plus epsilon 3 and so on epsilon n. And then we have similarly x n epsilon n, where mod of epsilon i less than or equal to mu, mu is the machine precision and i is equal to 1, 2 and so on up to n.

(Refer Slide Time: 02:28)

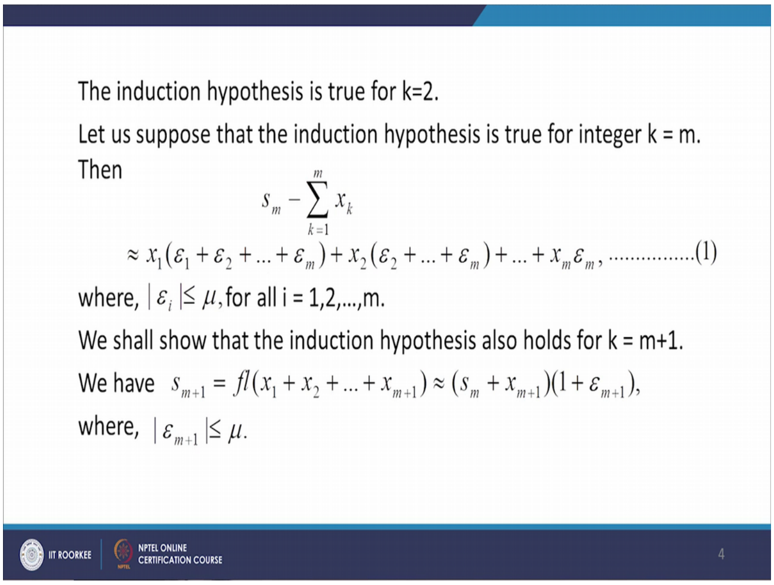


So, let us assume that let x k denote the floating point representation of x 1 plus x 2 and so on x k, where k takes values from 1 to n. Then what we will do? We shall prove this theorem by induction on k. So, when we take k equal to 1, let us take k equal to 1 then we see that s 1 is equal to fl x 1. Now, floating point representation of x 1 is equal to x 1 because x 1 is floating point representation, so x 1 x 2 x n are floating point numbers. So, this is fl x 1 equal to x 1 and therefore, we can say that now here you take x n equal to 1. So, fl x 1 minus x 1 is approximately equal to x 1 into epsilon 1 that we have to prove.

So, what do we get here? So, here this equal to 0, this 0 can be regarded as 0 into x 1. So, we can take epsilon 1 equal to 0. So, this equal to x 1 into epsilon 1, where epsilon 1 is equal to 0 and so mod of epsilon 1 is less than or equal to mu. So, the we can say that fl x 1 minus x 1 is equal to fl x 1 minus x 1 equal to x 1 into epsilon 1, where mod of epsilon 1 is less than or equal to mu. So, the theorem holds true for n equal to 1, hence for k equal to 1. If you take in s and k, s k k equal to 1 the theorem is true.

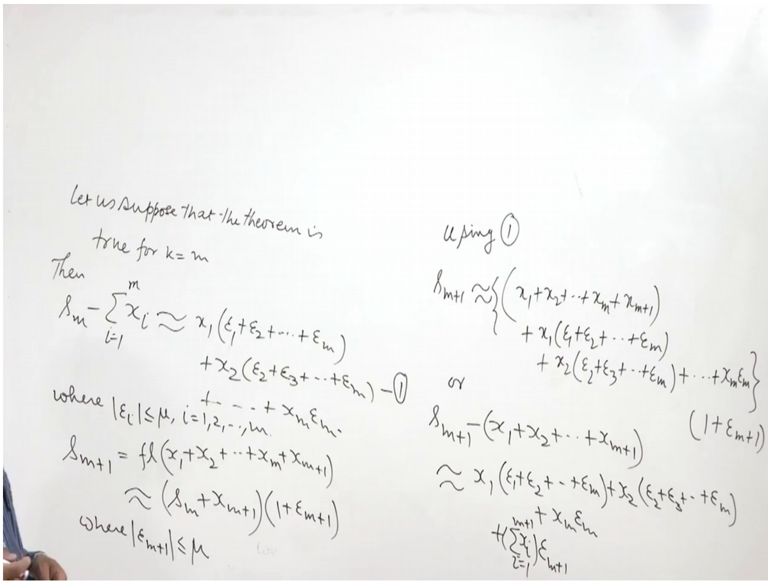
Now, we can also show that the term is true for k equal to 2, fl of x 1 plus x 2, we can write as x 1 plus x 2 into 1 plus epsilon. So, then we shall see that 1 plus epsilon 2. So, then what we will see is that where mod of epsilon 2 it less than or equal to mu hence s 2 minus s 1 plus s 2. This is s 2, s 2 minus x 1 plus x, x 2 is equal to x 1 plus x 2 into epsilon 2 from s 2 when we subtract x 1 plus x 2 we get x 1 plus x 2 into epsilon 2 and so what we can says that we can write it as x 1 times because epsilon is 1 is equal to 0. So, x 1 epsilon 2 plus x 2 epsilon 2, we can add the term x 1 epsilon 1 and write it as x 1 times epsilon 1 plus epsilon 2 plus x 2 epsilon 2, where mod of epsilon 1 is less than or equal to mu and mod of epsilon 2 is also less than or equal to mu and so the theorem is true for k equal to 2. So, the theorem is true for k equal to 2.

(Refer Slide Time: 07:12)



Now, let us suppose that the theorem is true for certain value k equal to m. So, let us suppose that k equal to m and n for all values of k less than or equal to m minus 1.

(Refer Slide Time: 07:21)



So, then s m is minus s m that is floating point representation of x 1 plus x 2 and so on s m plus sigma i equal to 1 to m x i is approximately x 1 into epsilon 1 plus epsilon 2 and so on. Epsilon m plus x 2 times epsilon 2, epsilon 3 and epsilon m and then x m epsilon m by s by our induction hypothesis where mod of epsilon i less than or equal to mu for all I equal to 1 2 and so on up to m.

Now, we shall show that the induction hypothesis also holds for k equal to m plus 1. So, let us see what is s m plus 1? s m plus 1 equal to floating point representation of x 1 plus x 2 and so on this fl x 1 plus x 2 and so on x m plus x m plus 1 which is equal approximately s m plus x m plus 1 into 1 plus epsilon m plus 1, now where mod of epsilon m plus 1 is less than or equal to m. So, epsilon m plus 1 is the error in the computation of x 1 plus x 2 plus x m plus x m plus 1.

Now, so let us use what is given to us. So, using 1, using 1 is s m minus sigma k equal to 1 to m x k yeah, i equal to 1 to m x i is approximately x 1 times epsilon 1 plus epsilon 2 and so on epsilon m plus x 2 times epsilon 2 plus and so on epsilon m plus x m epsilon m. So, let us use this and then we can write that s m plus 1. So, using 1, this is 1, s m plus 1 is approximately sigma i equal to 1 to m x i. So, we have x 1 plus x 2 and so on x m plus x m plus 1 plus x 1 into epsilon 1 plus epsilon 2 and so on epsilon m plus x 2 times epsilon 2 plus epsilon 3 and so on epsilon m plus and so on x m epsilon m and this multiplied by 1 plus epsilon m plus 1.

So, this is 1 plus epsilon m plus 1 is multiplied to the whole thing. So, what we will get is. Now, this is or we can say s m plus 1, x plus 1 plus x plus x 2 and so on x m plus 1 see we are what we are doing we are multiplying by one first when we are multiplying by 1, here we get x 1 plus x 2 plus x m plus x m plus 1 and then this whole thing. So, that x 1 plus x 2 plus x m plus 1 I am bringing to the left side this is approximately equal to x 1 times epsilon 1 plus epsilon 2 and so on epsilon m plus x 2 times epsilon 2 ok.

So, when we multiply by 1 we get the whole thing and then we multiply by epsilon m plus 1. So, what we get is this whole thing multiplied by epsilon m plus 1. So, sigma x i, i equal to 1 to m plus 1 multiplied by x m epsilon m plus 1 and then we will get x 1 times epsilon 1 into epsilon m plus 1 epsilon 2 into epsilon m plus 1 which are terms of second order. So, we can neglect them because epsilon is are too small and therefore, we have assume that they are second and higher order terms can be neglected. So, what we get here now? s m plus 1 minus sigma i equal to 1 to m plus 1 x i is approximately equal to; now here we shall have x 1 epsilon m plus 1 that x 1 epsilon m plus 1 we can observe here and write x 1 times.

(Refer Slide Time: 14:22)

$$s_{m+1} - \sum_{i=1}^{m+1} x_i \approx x_1(\epsilon_1+\epsilon_2+\cdots+\epsilon_{m+1}) + x_2(\epsilon_2+\epsilon_3+\cdots+\epsilon_{m+1}) + \cdots + x_m(\epsilon_m+\epsilon_{m+1}) + x_{m+1}\epsilon_{m+1}$$

where $|\epsilon_i| \le \mu, \forall i = 1,2,\ldots,m+1$

So, the induction hypothesis holds true for $k = m+1$.

using (1)

$$s_{m+1} \approx \{(x_1+x_2+\cdots+x_m+x_{m+1}) + x_1(\epsilon_1+\epsilon_2+\cdots+\epsilon_m) + x_2(\epsilon_2+\epsilon_3+\cdots+\epsilon_m)+\cdots+x_m\epsilon_m\}(1+\epsilon_{m+1})$$

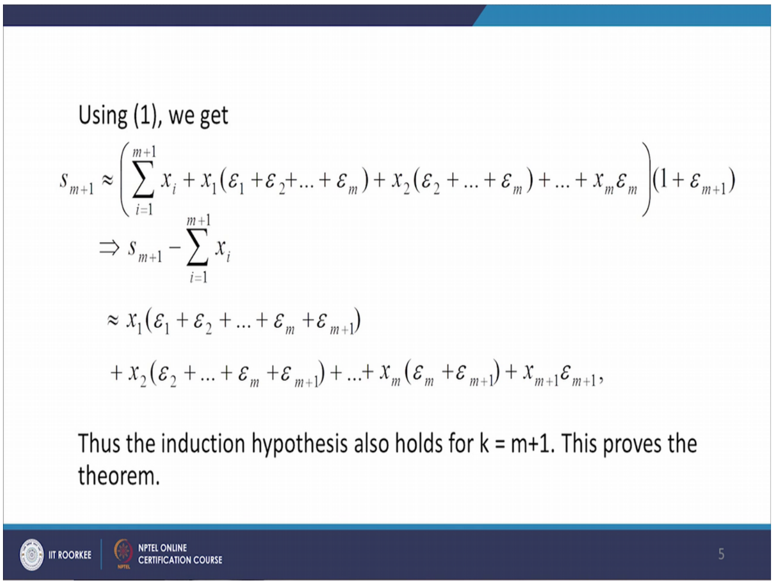
or

$$s_{m+1} - (x_1+x_2+\cdots+x_{m+1}) \approx x_1(\epsilon_1+\epsilon_2+\cdots+\epsilon_m)+x_2(\epsilon_2+\epsilon_3+\cdots+\epsilon_m) + x_m\epsilon_m + \left(\sum_{i=1}^{m+1} x_i\right)\epsilon_{m+1}$$

Then x 2 into epsilon m plus 1 that term can be brought here. So, x 2 times epsilon 2 plus epsilon 3 and so on epsilon m plus 1 and then before this term we will have x m minus one in that we can multiply we can add there x m minus 1 into epsilon m plus 1 term and here we can add x m into epsilon m plus 1. So, we shall have and so on and lastly we have this where mod of epsilon i is less than or equal to mu for all i equal to 1 2 and so on up to m plus 1.

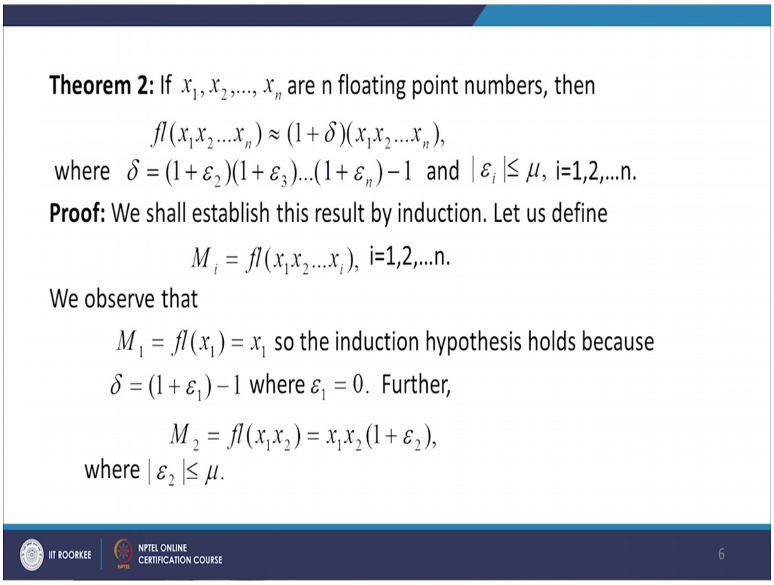
So, the induction hypothesis holds true for k equal to m plus 1 and therefore, it holds for all integers m or you can say it holds for all integers positive integers n. So, this is the proof of first theorem.

(Refer Slide Time: 16:26)



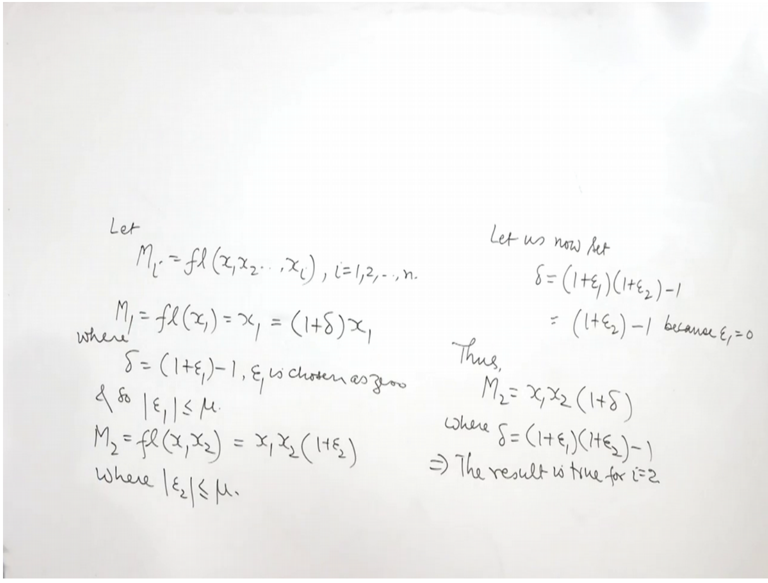
Now, let us go to the proof of the second theorem which is on multiplication.

(Refer Slide Time: 16:28)



So, let us assume that we have x 1, x 2, x n as n floating point numbers we are given n floating point numbers. Then the floating point representation of x 1 into x 2 into x n, if we find out this then this approximately equal to 1 plus delta times x 1 into x 2 into x m where delta is the error where delta is the expression 1 plus epsilon 1 into 1 plus epsilon 2 and so on 1 plus epsilon minus epsilon n minus one and mod of epsilon i is less than or equal to mu, i is equal to 1, 2, 3 and so on up to n. So, this result also we shall prove by induction. So, let us assume that m i is equal to fl x 1 x 2 x i, i is equal to 1, 2, 3 and so on up to n.

(Refer Slide Time: 17:18)

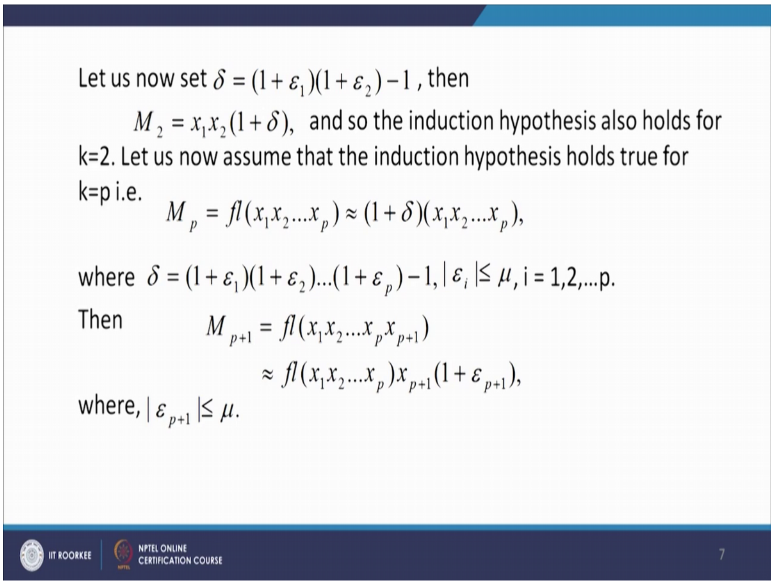


So, again let us prove that the result holds true for i equal to 1. So, M 1 is equal to fl x 1 and floating point representation of x 1 is x 1 because x 1 is a floating point number and so the induction hypothesis holds true because delta we can write as 1 plus epsilon 1 minus 1.

So, here we can write M 1 which is fl x 1, fl x 1 is equal to 1 plus delta into, so this delta where we where we take epsilon 2 the epsilon 1 to be equal to 0. So, what we do is we can write this is equal to 1 plus delta into x 1 where delta is given by 1 plus epsilon 1 minus 1 epsilon 1 is taken as 0; epsilon 1 is chosen as 0 and so mod of epsilon 1 is less than or equal to mu. So, the result holds true for i equal to 1, now let us show that the result holds true for I equal to 2. So, m 2 is equal to floating point representation of x 1 into x 2.

Now, let us say we write it as x 1 x 2 into where epsilon 2 is the error relative error in the computation of x 1 into x 2 floating point representation of x 1 into x 2 and mod of epsilon 2, here mod of epsilon 2 is less than or equal to mu.

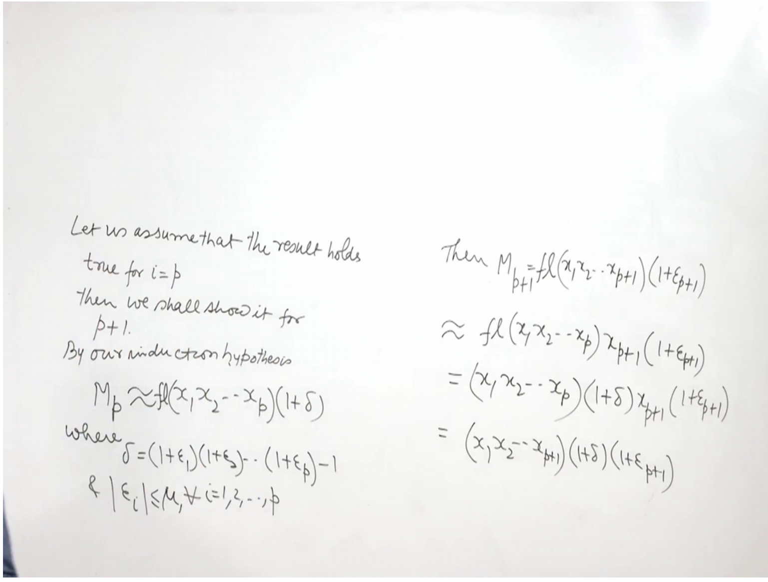
(Refer Slide Time: 20:00)



Now, let us set delta equal to remember epsilon 1 is equal to 0. So, this is actually 1 plus epsilon 2 minus 1 because epsilon 1 we have chosen as 0. So, delta is equal to 1 plus epsilon 2 minus 1 and this is therefore, equal to delta is equal to epsilon 2. So, you can also write the m 2 is equal to x 1, x 2 into 1 plus delta where delta is 1 plus epsilon 1, 1 plus epsilon 2 minus 1.

Now, let us assume that the result holds true for k equal to p here. So, thus, so the result is true for i equal to 2. Let us assume that the result holds true for i equal to p. So, let us assume that the result holds true for i equal to p then we shall show it for i plus 1 or so, p plus 1 and then we shall be able to say that it holds true for all positive integers p or it holds true for all positive integers n.

(Refer Slide Time: 21:37)

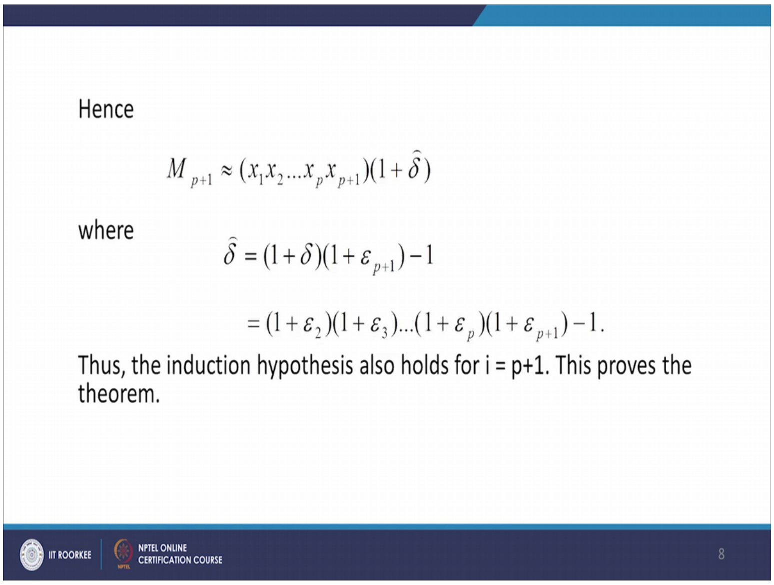


Now, so, by our hypothesis M p, M p, i equal to p. So, M p minus x 1 into x 2 and so on x p this is this is approximately equal to 1 plus, M p is equal to, floating point M p is x 1 plus x 1 x 2 and so on x p into 1 plus delta where let me write delta cap this is then where delta is equal to 1 plus epsilon 1, 1 plus epsilon 2 and so on 1 plus epsilon p minus 1 and mod of epsilon i is less than or equal to 1 for all I we have let us write delta only we need to write delta here.

So, delta is equal to this mod of epsilon i is less than or equal to mu, for all i equal to 1, 2, 3 and so on up to p. So, then I am take p plus 1 and p plus 1 which is x 1, x 2, x p plus 1 into 1 plus epsilon p plus 1 where epsilon p plus 1 is the error. So, this gives us fl x 1, x 2, x p, x 1, x 2, x p I have not written fl here, and here also we need to write fl. So, fl x 1, x 2, x p into 1 plus delta where this is equal to this, so this into x p plus 1 into 1 plus.

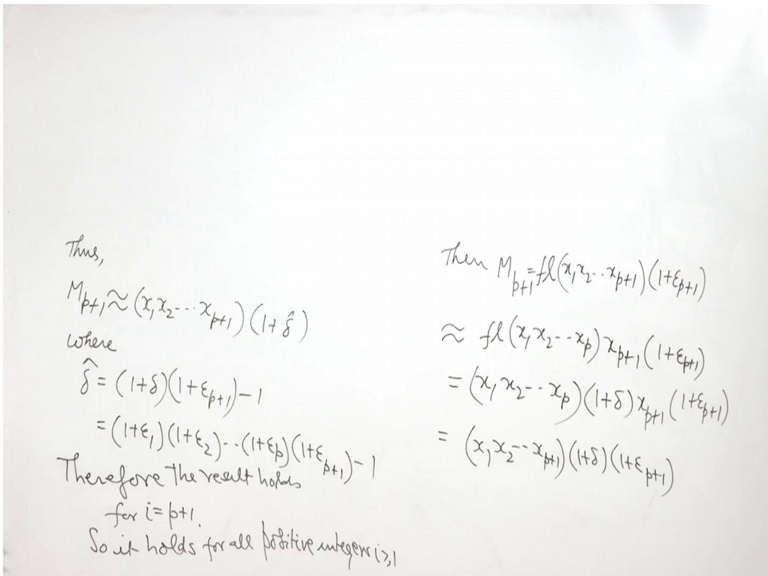
Now, what? So, this is equal to let us apply the induction.

(Refer Slide Time: 25:34)



This is, so fl x 1, x 2, x p this is equal to x 1, x 2, x p into x p in to 1 plus delta into x p plus 1 into 1 plus epsilon p plus 1 or we can write it as x 1, x 2, x p plus 1 into this. Thus we can say that M p plus 1 is approximately equal to x 1, x 2, x p plus 1 into 1 plus delta cap, where delta cap is equal to 1 plus delta into 1 plus epsilon p plus 1 minus 1.

(Refer Slide Time: 26:21)



1 plus delta we have seen, 1 plus delta is equal to 1 plus epsilon 1. So, it is 1 plus epsilon 1 into 1 plus epsilon 2 and so on 1 plus epsilon p into 1 plus epsilon p plus 1 minus 1. Now, this is approximately equal to we multiply by 1 here. And so what we will get? 1

plus epsilon 1, 1 plus epsilon 2, 1 plus epsilon p into 1 plus epsilon p plus 1 minus 1. With that we have proved the result.

So, we can write M p plus 1 as x 1, x 2, x p plus 1 into 1 plus delta where delta cap is equal to 1 plus delta into 1 plus epsilon p plus 1 minus 1, but 1 plus delta is equal to 1 plus epsilon 1 into 1 plus epsilon 2 and so on 1 plus epsilon p into 1 plus epsilon p minus 1 and therefore, the result holds for p plus 1. So, it holds for all positive integers i and therefore, it holds for all integers and equal to 1, 2, 3 and so on. So, this proves the theorem on multiplication of floating point numbers. With that I would like to conclude my lecture.

Thank you very much for your attention.