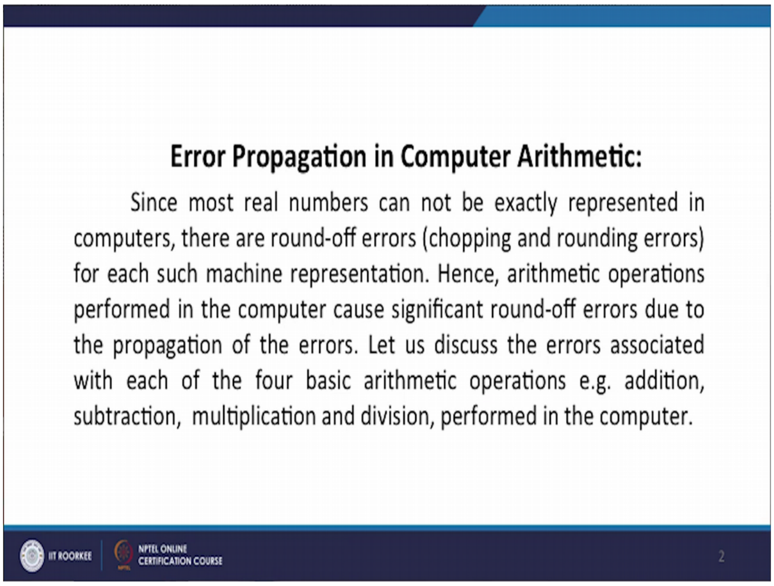
Numerical Linear Algebra

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

Lecture - 23
Error propagation in computer arithmetic

Hello friends, welcome to my lecture on Error Propagation in computer arithmetic.

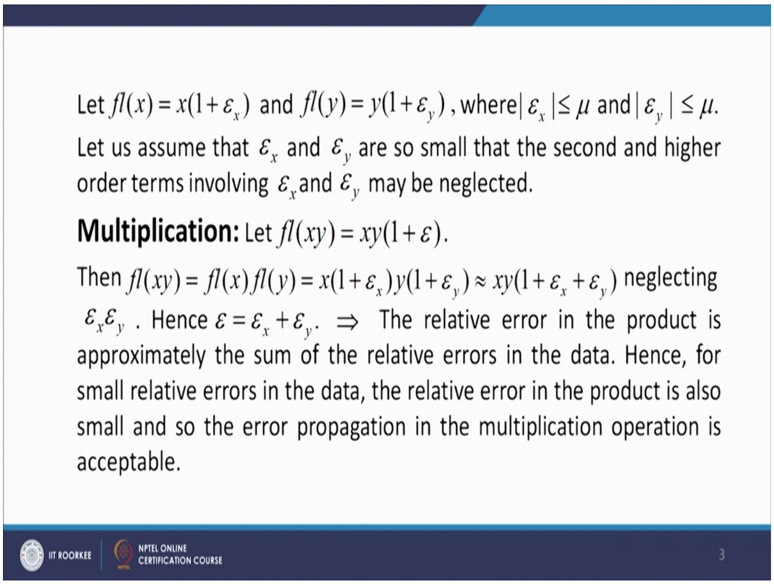
(Refer Slide Time: 00:24)



Since most real numbers cannot be exactly represented in computers, there are round off errors which are caused due to chopping and rounding errors; for each such machine representation. Hence arithmetic operations performed in the computer cause significant round off errors due to the propagation of the errors.

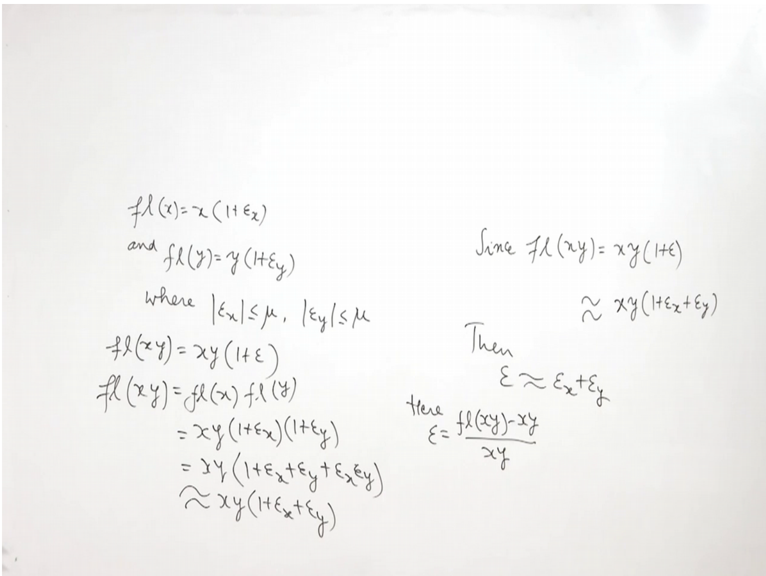
Let us discuss the errors associated with each of the four basic arithmetic operations namely addition, subtraction, multiplication and division which are performed, in the computer.

(Refer Slide Time: 00:57)



Let us say the floating point representation of x we denote by f l x so, let f l x be equal to x into 1 plus epsilon x and f y equal to y into 1 plus epsilon y.

(Refer Slide Time: 01:13)



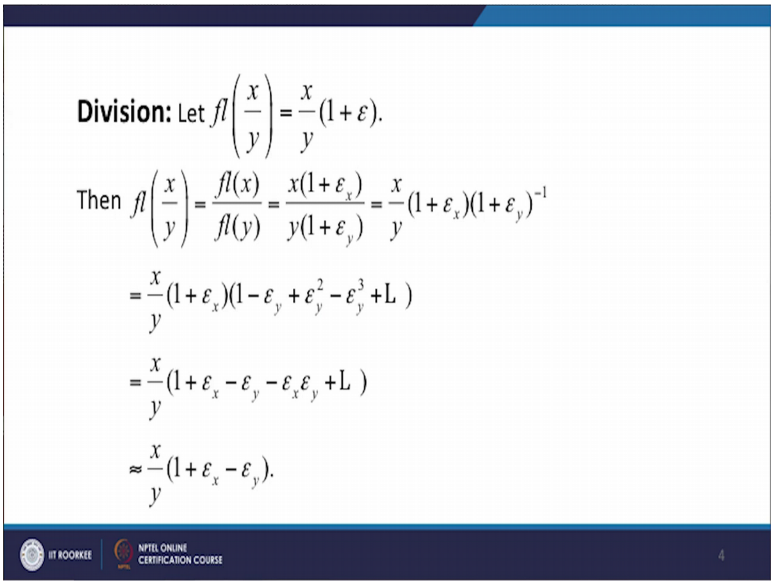
So, f l x equal to x into 1 plus epsilon x and f l y equal to y into 1 plus epsilon y where epsilon x and epsilon y are less than or equal to mu; the machine precision mu denotes the machine precision. So, we further we assume that epsilon x and epsilon y are so, small that their second and higher order terms involving epsilon x and epsilon y may be neglected. Let us first consider the case of multiplication so, in multiplication let us say we have two numbers x and y; the floating point representation of x into y will then be written as x y into 1 plus epsilon.

So, f l x y we can write as x y into 1 plus epsilon then because f l x y is equal to f l x into f l y, let us put the values of f l x and f l y, we shall get f l x y equal to x into 1 plus epsilon x f l x y equal to f l x into f l y. So, substituting the values of f l x and f l y what we get; x y into 1 plus epsilon x into 1 plus epsilon y. Let us multiply a 1 plus epsilon x and 1 plus epsilon y what we get? 1 plus epsilon x plus epsilon y plus epsilon x into epsilon y. Now, we have assumed that epsilon x and epsilon y are so, small that the terms of second and higher order terms involving epsilon x and epsilon y may be neglected.

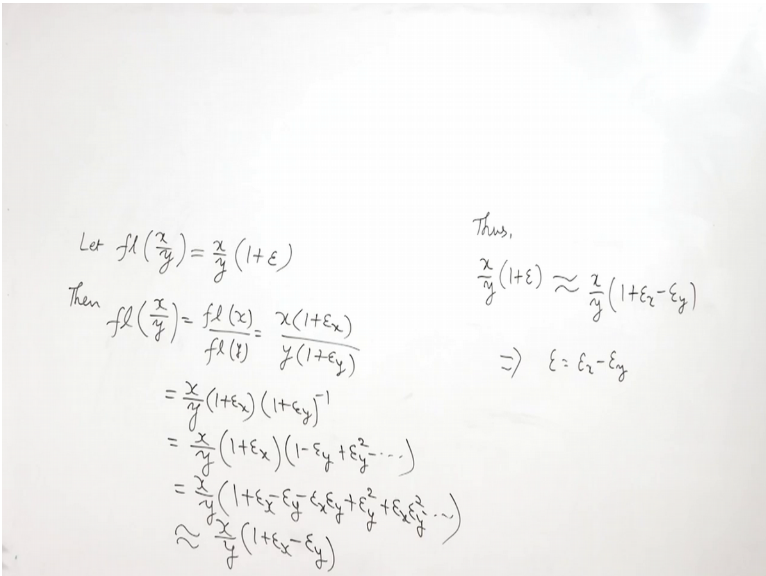
So, this is the second order term epsilon x into epsilon y so, let us neglect this cut on then we can say this is approximately x y into 1 plus epsilon x plus epsilon y. And now, f l x y is equal to x y into 1 plus epsilon. Since f l x y is equal to x y into 1 plus epsilon and it is approximately 1 plus epsilon, it is x into y 1 plus epsilon x plus epsilon y. We can say that the error the epsilon is approximately equal to epsilon x plus epsilon y so, we have. So the relative error now we can say that what is epsilon? Epsilon is here f l x y, here epsilon is equal to f l x y minus x y upon x y.

So, the relative error because of epsilon equal to epsilon x plus epsilon y; we can say that the relative error in the product is approximately the sum of the relative errors in the data. Epsilon x and epsilon y denote the relative errors in x and y. So, for the small relative errors in the data that is when epsilon x and epsilon y are small we can say that the relative error in the product is also small and therefore, the error propagation in the multiplication operation is acceptable.

(Refer Slide Time: 05:26)



(Refer Slide Time: 05:34)



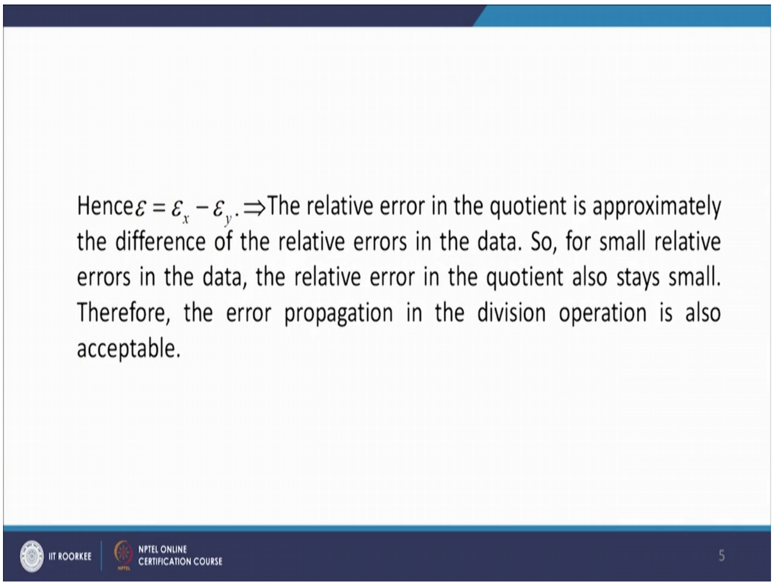
Let us now consider the case of division so, in division operation let us assume that let f l x y x over y be equal to x over y into 1 plus epsilon then f l x y x over y equal to f l x over f l y is equal to x into 1 plus epsilon x divided by y into 1 plus y.

Now, since epsilon x and epsilon y are very small what we can do I can write it as x over y 1 plus epsilon x into 1 plus epsilon y rise to the power minus 1. Since epsilon y is very small mode of epsilon y is less than 1 so, we can expand this 1 plus epsilon y to the power minus 1 binomial expansion and we have x over y 1 plus epsilon x into 1 minus epsilon y plus epsilon y square and so, on.

Now, when we multiply this infinite series to 1 plus epsilon x what we have 1 plus epsilon x minus epsilon y minus epsilon x epsilon y plus epsilon y square plus epsilon x epsilon y square and so, on. Now, since the second order terms and higher order terms containing epsilon x and epsilon y are to be neglected this is approximately equal to x over y into 1 plus epsilon x plus minus epsilon y.

So, thus x over y into 1 plus epsilon is approximately x over y 1 plus epsilon x minus epsilon y and so, what we get is epsilon is equal to approximately epsilon x minus epsilon y and from this equation epsilon equal to epsilon x minus epsilon y.

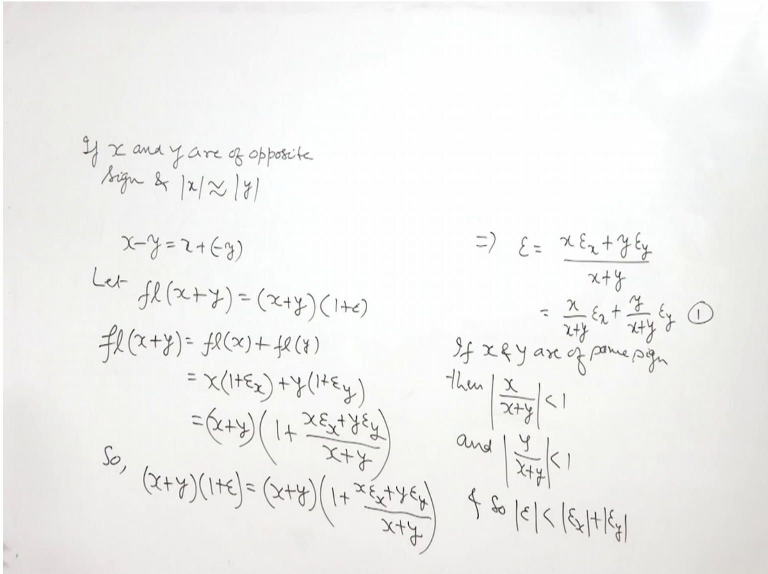
(Refer Slide Time: 08:16)



We can conclude that the relative error in the quotient the relative error in the quotient epsilon is approximately the difference of the relative errors in the data, the difference of the relative error of x and the relative error of y and thus for small relative errors in the data that is when epsilon x and epsilon y are small the relative error in the quotient also stage small and so, the error propagation in the division operation is also acceptable.

Now, let us consider the case of addition and subtraction.

(Refer Slide Time: 09:11)

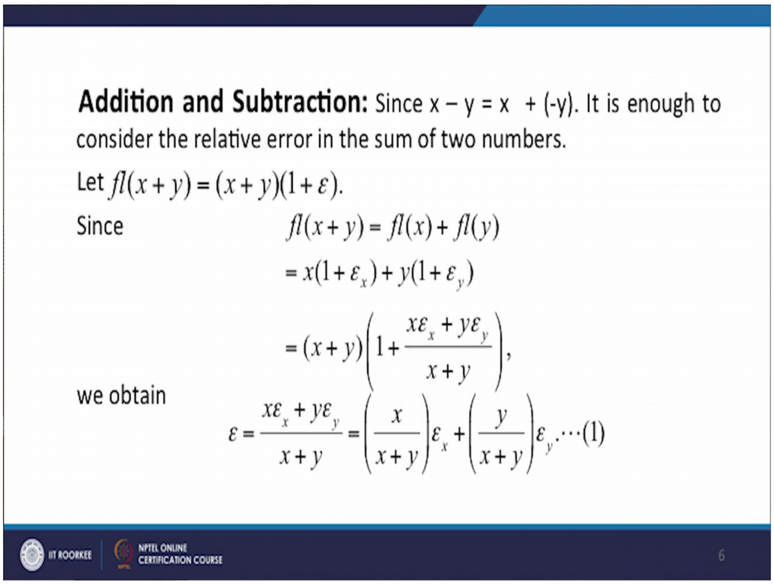


So, first we observe that if we have two numbers x and y then x minus y can be regarded as x plus minus y and therefore, we can say that it is enough to consider the relative error in the sum of two numbers x minus y is sum of x and minus y.

So, just let us consider the relative error in the sum of two numbers let us assume that f l x plus y is equal to x plus y into 1 plus epsilon , where epsilon is the relative error in the sum of x and y. So, then f l x plus y equal to f l x plus f l y gives f l x plus y equal to f l x plus f l y this is equal to x times 1 plus epsilon x plus y times 1 plus epsilon y, what we get is x plus y if we take common we get 1 plus x epsilon x plus y epsilon y divided by x plus y.

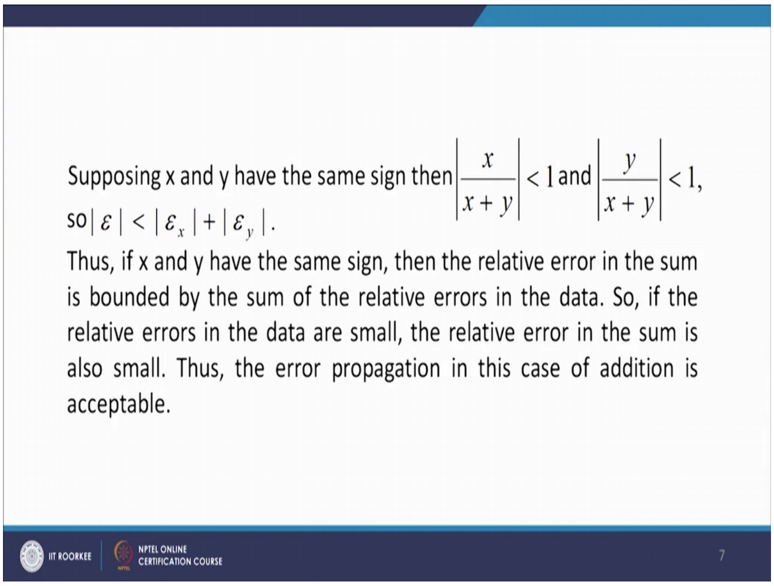
Now, f l x plus y equal to x plus y into 1 plus epsilon so, x plus y into 1 plus epsilon is equal to x plus y into 1 plus x epsilon x plus y epsilon y divided by x plus y. From this equation we see that epsilon is equal to x epsilon x plus y epsilon y divided by x plus y or we can write it as x over x plus y into epsilon x plus y over x plus y into epsilon y.

(Refer Slide Time: 11:20)



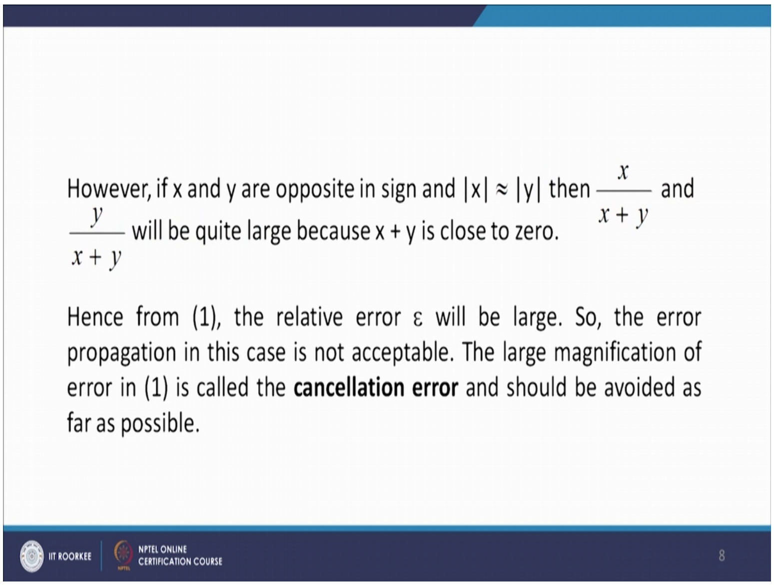
So, we get x equal to epsilon equal to x over x plus y into epsilon x plus y over epsilon x plus y into epsilon y and if we suppose that x and y have the same sign then mode of x over x plus y.

(Refer Slide Time: 11:32)



So, if x and r of same sign then we can say that mode of x over x plus y is strictly less than 1 and mode of y over x plus y is strictly less than 1 and therefore, mode of epsilon and so, mode of epsilon is less than mode of epsilon x plus mode of epsilon y. Thus, if x and y have the same sign then the relative error in the sum is bounded by the sum of the relative errors in x and y ah, if the relative error in the data are small that is x epsilon x and epsilon y are small then we can see that the relative error in the sum of x and y is also small and so, the error propagation in this case of addition is also acceptable

(Refer Slide Time: 12:47)

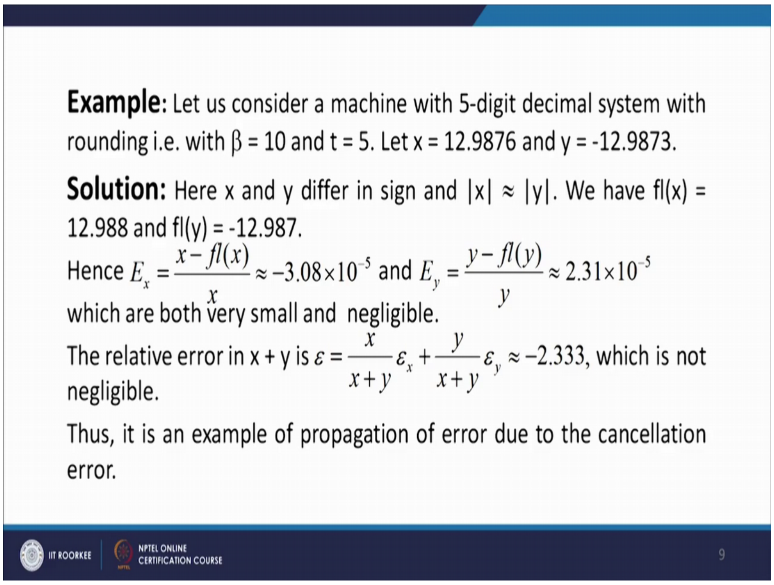


Now, if x and y are of opposite sign let then what do we have, if x and y are of opposite sign and mode of x minus is approximately equal to mode of y and mode of x and mode of y are approximately equal then x over x plus y because x plus y will be very near to 0. So, x over x plus y and y over x plus y will be quite large because x plus y is close to 0 and hence from equation 1, this equation 1 this one this equation 1.

So, from equation 1 it follows that the relative error in the computation of epsilon is also is large so, the relative epsilon will be large because epsilon is x over x plus y into epsilon x plus y over x plus y into epsilon y. So, when x over x plus y and y over x plus y are large relative error epsilon will be large so, the error propagation in this case is not acceptable. The large magnification of error in the equation 1 is called the cancellation error and should be avoided as far as possible.

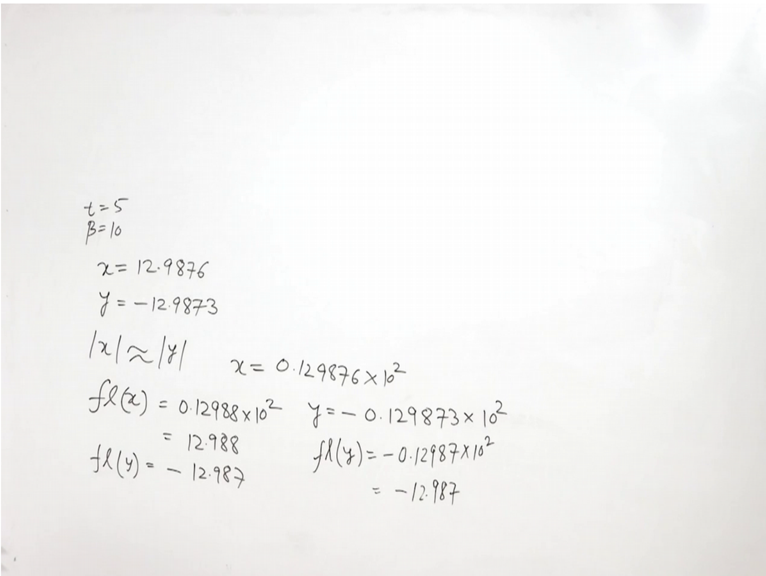
Now, let us see some cases where we will see how we can avoid the cancellation error.

(Refer Slide Time: 14:30)



So, let us consider the machine with 5 digit decimal system so, here we have t equal to 5 and beta equal to 10, beta tells us the base of the digital system and t tells us the number of significant digits that we have to retain.

(Refer Slide Time: 14:36)



So, let us consider machine with 5 digit decimal system with rounding so, beta is equal to 10 t equal to 5. Let us take x equal to 12.9876; x equal to 12.9876 and y equal to minus 12.9873 you can see that the x and y are of opposite sign and the absolute value of x is equal to 12.9876 absolute value of y is 12.9873 so, they are approximately same.

So, mode of x and mode of y are nearly equal now let us see when we find f l x the floating point representation of x so, for floating point representation of x we have to write it in the decimal system. So, x can be written as 12; x can be written as 0.129876 into 10 to the power 2.

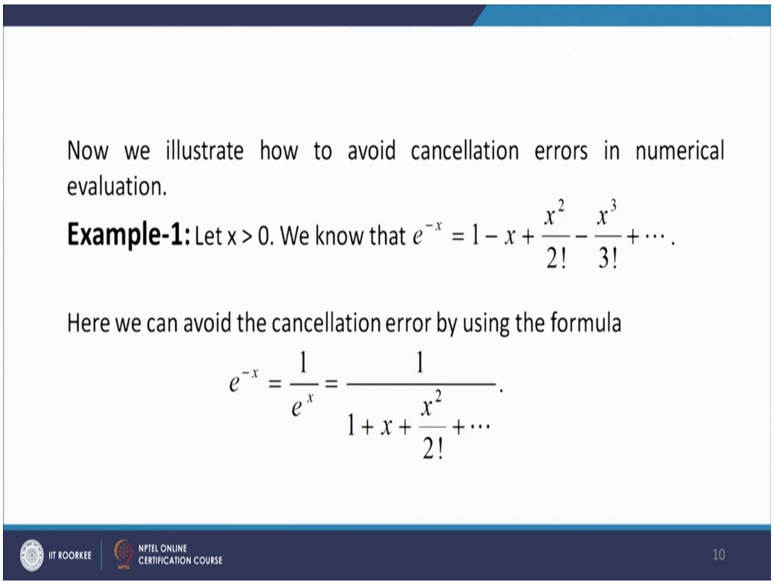
Now, we have to use rounding and we have to consider 5 significant digits so, 1, 2, 3, 4, 5 since this d t plus 1 is 6 it is more than 5 we have to increase 7 by 1 unit and therefore, we shall have f l x equal to 0.1298 8 into 10 to the power 2 or we can say this is 12.988. Similarly, y is equal to y can be written as minus 0.129873 into 10 to the power 2 and so, when we write floating point representation of y with t equal to 5; what we have minus 0 point now here 1, 2, 3, 4, 5. So, d t plus 1 is 3 which is less than beta by 2 that is 5 so, we keep this digit d t s m it we do not change so, 0.12987 into 10 to the power 2 or we get minus 12.987. So, f l x is 12.988 f l y is minus 12. 987.

Now, we can calculate epsilon x and epsilon y epsilon x is equal to x minus f l x over x and it is minus 3.08 into 10 to the power minus 5 epsilon y is y minus f l y over y which is approximately 2.31 into 10 to the power minus 5, you can see both are very small and so, can be assumed as negligible.

The relative error in x plus y which is given by epsilon equal to x over x plus y into epsilon x plus y over x plus y into epsilon y is approximately minus 2.333 which is not negligible as you can see it is minus 2.333 which is which cannot be neglected. Thus, it is an example of propagation of error due to the cancelation error x and y are of opposite sign the absolute values of x and y are nearly same ah.

So, when we calculate epsilon x and epsilon y the their values are very small, but still when we calculate epsilon which is the relative error in x plus y it turns out that it is not small it is not negligible. So, it is an example of propagation of error due to the cancellation error.

(Refer Slide Time: 19:08)

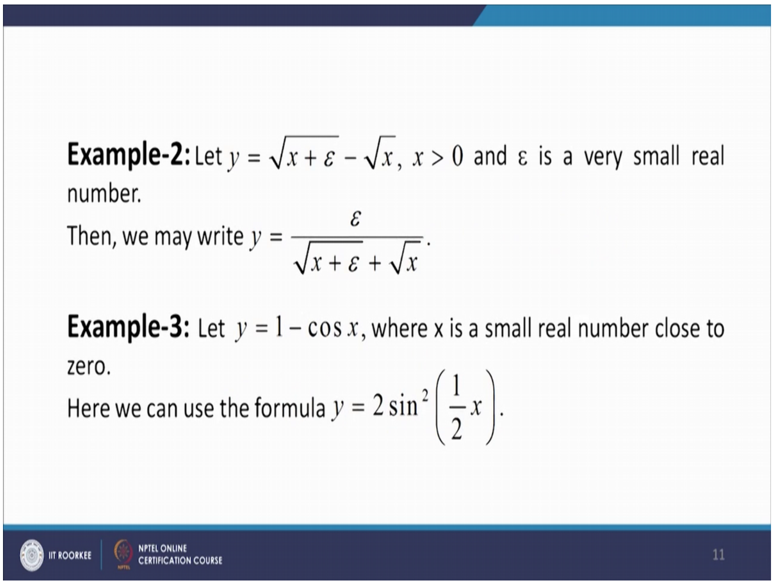


Let us illustrate how to avoid cancellation errors in numerical evolution for example, let us consider x to be greater than 0 we know that epsilon e to the power minus x is 1 minus x plus x square by 2 factorial minus x cube by 3 factorial. So, if you want to calculate the value of e to the power x for an x greater than 0 then you can see that there are negative positive terms alternately negative and positive terms.

So, there may there may there may be a situation where the two sub successive values are very close to each other so, and there will be an opposite sign in between the two. So, what we will do is here we can avoid the cancelation error by using the formula e to the power minus x equal to 1 over e to the power x because in e to the power x all terms will be positive when x is positive. So, there would there will not occur a situation where two successive terms are sums are nearly equal and they are opposite in sign.

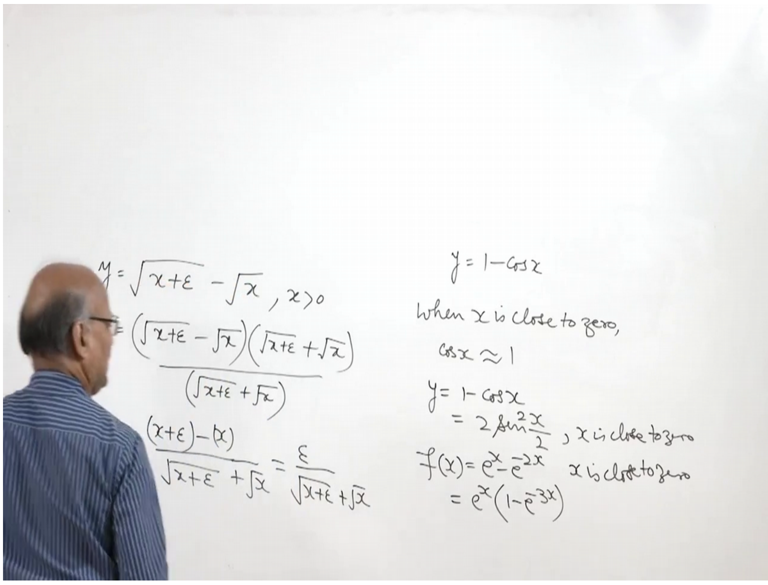
So, we can to calculate e to the power minus x we can use the formula 1 over e to the power x which is 1 over 1 plus x plus x square by 2 factorial and so, on and we can avoid the cancellation error.

(Refer Slide Time: 20:23)



Now, similarly let us take the case of an example of 2 which were y is equal to square root of x plus epsilon minus square root of x and x is positive.

(Refer Slide Time: 20:43)



So, here y is equal to since epsilon is very small is square root of x plus epsilon and square root of x are nearly equal and so, when we subtract the 2 there will be cancellation error.

So, to avoid the cancelation error what we do we write it as we multiply and divide by x square root of x plus epsilon plus square root of x then this is a minus b into a plus b. So, we have a square minus b square so, we have x plus epsilon minus x divided by and this is equal to epsilon over ok.

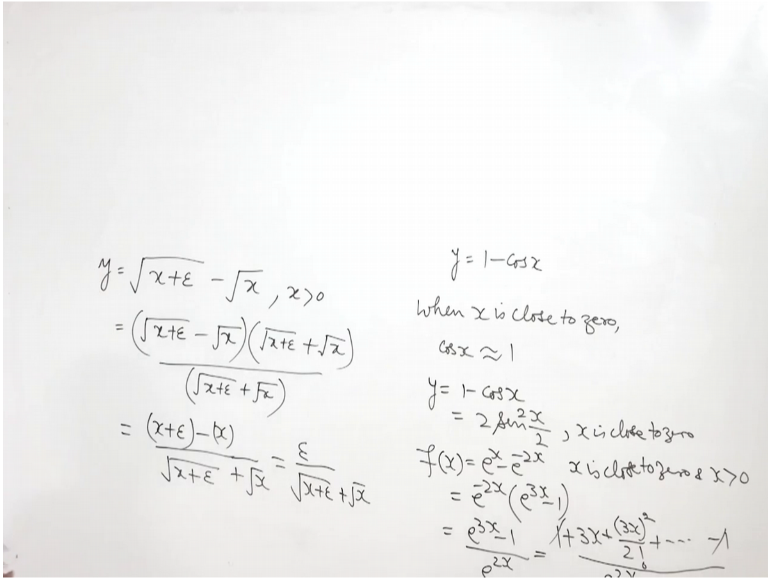
Now, we have epsilon over x square root x plus epsilon plus square root x so, which are of which we are we have to find the sum of x square root x plus epsilon and square root x and they are of same sign. So, the cancellation error can be avoided.

Now, another example we is there say y is equal to 1 minus cos x we have to calculate y equal to 1 minus cos x where x is given to be a small real number which close to 0 ok. So, when x is close to 0 cos x will be close to 1 it will be approximately one so, 1 minus cos x again we will have difference of 2 nearly equal numbers. So, what we do we can use the formula y equal to 1 minus cos x equal to 2 sin square x by 2.

We know this formula from trigonometry 1 minus cos theta equal to 2 sin square theta by 2 so, 1 minus cos x is equal to 2 sin square x by 2 here we can easily find the value for x which is x is close to 0. So, we can easily find the value because there is no cancellation error involved here.

One more example let us take say f x equal e to the power x minus e to the power minus 2 x here again let us assume that we are we have to calculate f x for values of x which are near 0 x is close to 0. So, when x is close to 0 e to the power x and e to the power minus 2 x are both close to 1 so, we will have again a case of cancellation of cancellation error here. So, what we do is we can write it in alternate manner by writing this way e to the power x we can take common so, 1 minus e to the power minus 3 x we have or we can write it this way e to the power minus 2 x we write outside.

(Refer Slide Time: 24:27)



Then this will be e to the power 3 x minus 1 and this is equal to e to the power 3 x minus 1 divided by e to the power 2 x, this will be equal to 1 plus 3 x minus 3 x whole square by 2 factorial plus 3 x by 2 whole square and so, on minus 1 divided by divided by e to the power 2 x ok.

So, this 1 and this 1 cancel x is given to be a number x is close to 0 and greater than 0 x is greater than 0. So, what we will have here all terms will be of same sign and therefore, we can calculate the sum and this e to the power 2 x also can be calculated there will be no cancellation error involved here. So, e to the power 2 e to the power x minus e to the power minus x when x is close to 0 and is greater than 0 is an example of cancellation error, but the cancellation error can be avoided by writing this f x function as e to the power 3 x minus 1 over e to the power 2 x.

This 1 which is subtracted from e to the power 3 x can be cancelled with the one occurring in e to the power 3 x and then we see that all terms are of same sign because x is positive. So, there is no cancellation error in here and e to the power 2 x is also positive for all x I mean will also contain all terms of the same sign. So, there will be no cancellation error involved here so, the cancellation error can be avoided by writing f x as e to the power 3 x minus 1 over e to the power 2 x, with that I would like to conclude my lecture.

Thank you very much for your attention.