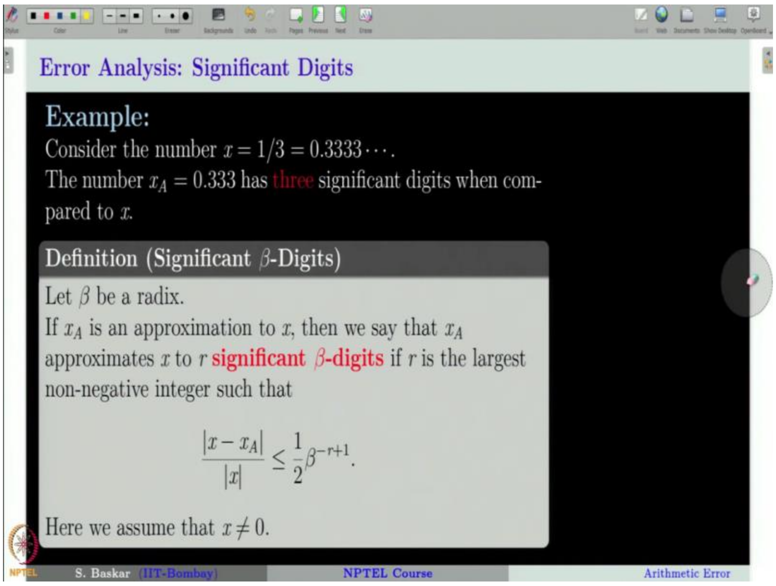
Numerical Analysis
Prof. S. Baskar
Department of Mathematics
Indian Institute of Technology – Bombay

Lecture – 05
Arithmetic Error: Loss of Significance

Hi in this lecture we will continue our discussion on floating point approximation of a number which we did from the last class. In this class, we will understand what is meant by significant digits and what is the danger in losing significant digits in our calculation.

(Refer Slide Time: 00:39)



Error Analysis: Significant Digits

Example:
Consider the number $x = 1/3 = 0.3333\cdots$.
The number $x_A = 0.333$ has three significant digits when compared to x.

Definition (Significant β-Digits)
Let β be a radix.
If x_A is an approximation to x, then we say that x_A approximates x to r **significant β-digits** if r is the largest non-negative integer such that

$$\frac{|x - x_A|}{|x|} \leq \frac{1}{2}\beta^{-r+1}.$$

Here we assume that $x \neq 0$.

Let us take the number $\frac{1}{3}$ it is given by 0.333 and so on it goes on up to infinity and you take the number x_A = 0.333. We can say that the number x_A approximate x with three significant digits. So, this is kind of intuitively clear what we meant by saying x_A is approximating x with three significant digits because you can simply count the number of digits which are matching with the digits in the original number.

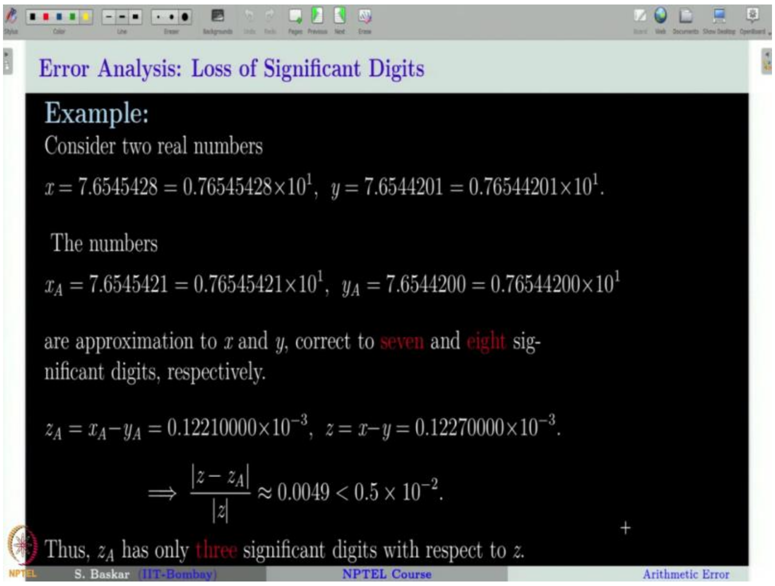
The first position of the number x_A is 3 which is also there in this. Of course, you have to forget what is there before the decimal point here. You just have to think that these are written in the floating-point representation and then try to see the number of significant digits that is the number of digits which are matching with these two numbers. So, the first three digits and the first digits here are matching.

Therefore, we can say that there are three significant digits in the number x_A when compared to x. Let us put this idea in the mathematical form and give the definition of *r* significant digits. Let us assume that our radix that is the base is β and we say that the number x_A is an approximation to x with *r* significant β digits if *r* is the largest non- negative integer such that the relative error $\frac{|x-x_A|}{|x|} \leq \frac{1}{2}\beta^{-r+1}$.

Of course, here *x* should not be equal to 0 that is always assumed whenever we are talking about the relative error. So, if such a thing happens then this *r* which is the largest non-negative integer that we have taken and see that this inequality holds then that r is called the *r* significant digits present in the number x_A when compared to *x*. Intuitively, it-will be like you just match these two numbers.

And see how many leading non zero numbers are matching with these two numbers. However, due to some rounding problem it may not always work this counting idea. Therefore, it is always better to check this definition correctly in order to see how many significant digits are there in x_A when compared to *x*. Next, we will see what is meant by loss of significance and what is the danger in losing the significant digits.

(Refer Slide Time: 04:01)



Error Analysis: Loss of Significant Digits

Example:

Consider two real numbers

$x = 7.6545428 = 0.76545428\times10^1, \;\; y = 7.6544201 = 0.76544201\times10^1.$

The numbers

$x_A = 7.6545421 = 0.76545421\times10^1, \;\; y_A = 7.6544200 = 0.76544200\times10^1$

are approximation to x and y, correct to seven and eight significant digits, respectively.

$z_A = x_A - y_A = 0.12210000\times10^{-3}, \;\; z = x - y = 0.12270000\times10^{-3}.$

$$\Longrightarrow \frac{|z-z_A|}{|z|} \approx 0.0049 < 0.5\times10^{-2}.$$

Thus, z_A has only three significant digits with respect to z.

S. Baskar (IIT-Bombay) NPTEL Course Arithmetic Error

For that we will take an example. Consider two numbers *x* which is given like this and *y* which is given like this and I am writing it in the floating-point representation. Of course, I have to write a $(-1)^0$ here, but I am just leaving them and now for this instead of taking the exact values *x* and *y*, I am taking their approximate value x_A and y_A somehow approximated.

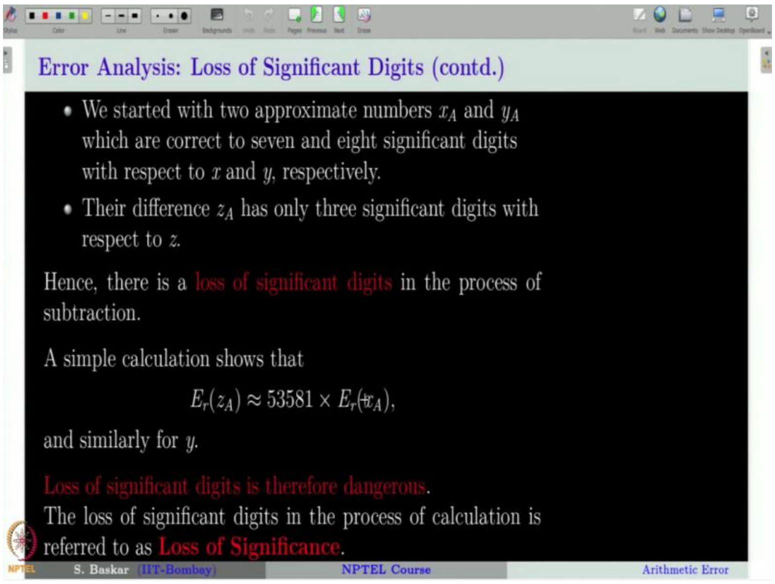
And then I want to perform the calculation with x_A and y_A instead of *x* and *y*. Before going into this you can see that *x* and *y* are correct to 7 and 8 significant digits respectively that is x_A has 7 significant digits when compared to *x* and y_A has 8 significant digits when compared to *y*. Now, we want to perform the subtraction between *x* and *y* instead of doing that you are doing the subtraction between x_A and y_A. The answer is this.

If you would have done the same calculation with exact values you would have got this number. Let us see how many significant digits are there in z_A when compared to *z*. You can see that $\frac{|z-z_A|}{|z|}$ is approximately this number and this number can be written as $(-1)^0 \times .49 \times 10^{-2}$. So, what you do in order to find the number of significant digits you just find this number first.

And get the number and write it in the floating-point form and then you will see what is the nearest 0.5 that you can dominate this? You can see that this is less than equal to of course this is something which you can avoid writing you need not write all the time $.5 \times 10^{-2}$. So, that is what I am writing here. So, once you write it in the floating-point form, then it is very easy to get what is the nearest number with 0.5 that dominates your original number.

So, now once you get this you know that – *r* + 1 should come here and that happens to be –2 from there you can get what is your *r* and that happens to be 3 and therefore z_A has three significant digits when compared to the number *z*.

(Refer Slide Time: 07:30)



Error Analysis: Loss of Significant Digits (contd.)

- We started with two approximate numbers x_A and y_A which are correct to seven and eight significant digits with respect to x and y, respectively.
- Their difference z_A has only three significant digits with respect to z.

Hence, there is a loss of significant digits in the process of subtraction.

A simple calculation shows that

$$E_r(z_A) \approx 53581 \times E_r(x_A),$$

and similarly for y.

Loss of significant digits is therefore dangerous.
The loss of significant digits in the process of calculation is referred to as Loss of Significance.

S. Baskar (IIT-Bombay) NPTEL Course Arithmetic Error

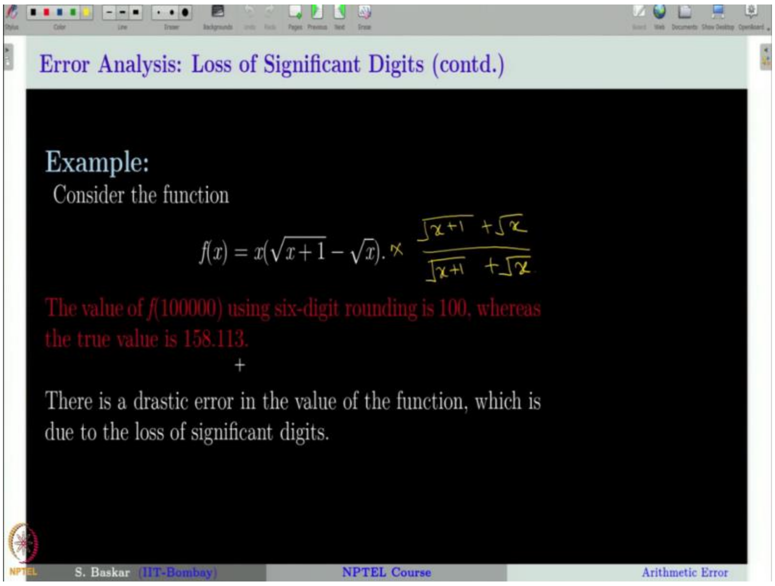
So, to summarize we started with two numbers x_A and y_A which are correct to 7 and 8 significant digits with respect to x and y respectively and their difference z_A has only three significant digits with respect to z and therefore we had a loss of significant digits in the process of subtracting this is what we understand here. So what, if you ask this question. Let us see the relative error in z_A when compared to z is 53,581 times more than the relative error in x_A when compared to x.

Therefore, the process of subtraction has amplified the error in the initial data by a factor of around 53,000 so that is the danger which is happening in this calculation and similarly you can also see how much the amplification happened when compared to the relative error in y_A and this shows that the loss of significant digits is very dangerous it can magnify the error drastically.

Suppose you just assume that you have a calculation and that calculation goes in a loop say for 1 million times and every time the calculation is done in every single loop if the error is magnified by say 50,000 times than 1 million times if you do the calculation again and again rigorously, you can imagine how much the error will be amplified in the final answer. Therefore, loss of significant digits is very dangerous in computation.

One has to be alert in the computation to see where all you are losing the significant digits. The process of losing significant digits is often referred as loss of significance.

(Refer Slide Time: 09:49)



Error Analysis: Loss of Significant Digits (contd.)

Example:

Consider the function

$$f(x) = x(\sqrt{x+1} - \sqrt{x}) . \times \frac{\sqrt{x+1} + \sqrt{x}}{\sqrt{x+1} + \sqrt{x}}$$

The value of $f(100000)$ using six-digit rounding is 100, whereas the true value is 158.113.

There is a drastic error in the value of the function, which is due to the loss of significant digits.

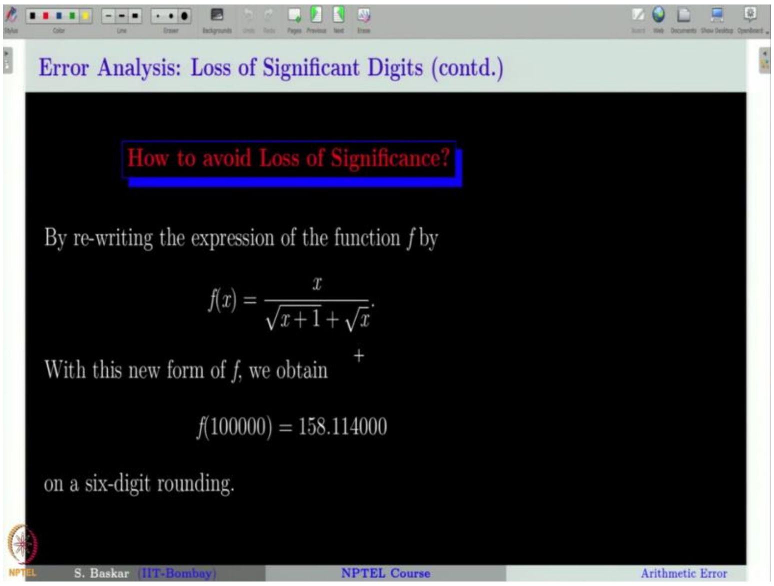
S. Baskar (IIT-Bombay) NPTEL Course Arithmetic Error

Let us take the example which we have done previously. Let us take $f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right)$

and if you recall in the last lecture we have computed the value of this function at x = 100000 using 6-digit rounding and we got the answer as 100. The exact value of this expression is actually 158.113. You can see that there is a significant error involved in this calculation with 6-digit rounding which is obviously due to loss of significance because you can see for a very large x, $\sqrt{x+1}$ and $\sqrt{x}$ they are very close to each other.

Therefore, they have lot of significant digits in them and they all get cancelled because of the subtraction and that is how the loss of significance happened in this case.

(Refer Slide Time: 11:11)



Error Analysis: Loss of Significant Digits (contd.)

How to avoid Loss of Significance?

By re-writing the expression of the function f by

$$f(x) = \frac{x}{\sqrt{x+1}+\sqrt{x}}.$$

With this new form of f, we obtain

$$f(100000) = 158.114000$$

on a six-digit rounding.

S. Baskar (IIT-Bombay) NPTEL Course Arithmetic Error

Is it possible to avoid loss of significance in our calculations? Well, this is not always possible, but a good news is that in the function which we considered previously that is this function you have a way to avoid this loss of significance. How to do that? Well, that expression can be rewritten in this form. How will you do that? Well, you can multiply and divide by $\sqrt{x+1}+\sqrt{x}$ divided by $\sqrt{x+1}+\sqrt{x}$.

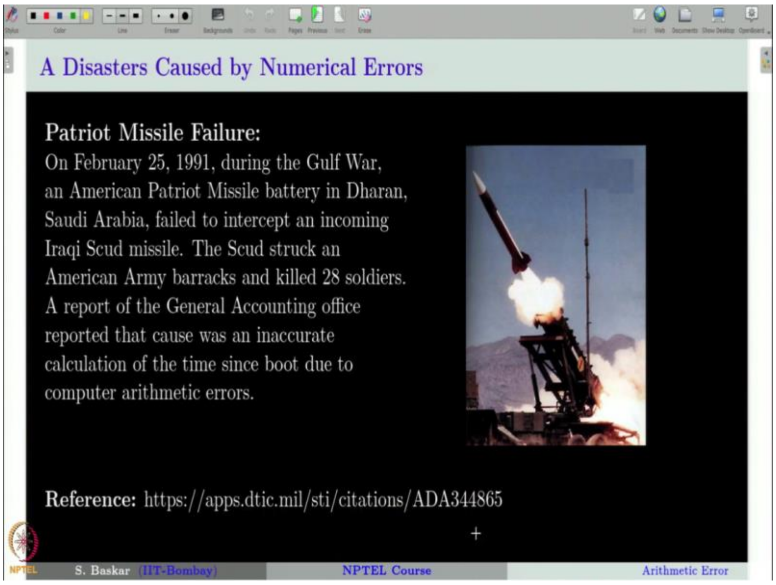
Now what happens numerator will give you just 1 and divided by this. Therefore, $f(x)$ can also be equivalently written as $\frac{x}{\sqrt{x+1}+\sqrt{x}}$. Now you go back to our previous example where we have computed the value of $f(100000)$ using 6 significant digit rounding. The same idea you follow and now calculate the value of f using this expression at the point x = 100000 using the same 6-digit rounding calculation.

Now you will get the value 158.114. I will again repeat you do the 6-digit rounding calculation just like what we did in the previous class. Now, instead of the previous expression now you use this equivalent expression for *f*. Mathematically this expression and the expression that we have taken initially that is this expression these two are one and the same mathematically, but if I use 6-digit rounding and perform the calculation and find the value of $f(100000)$ using this expression then I get the value 158.114 which is pretty close to the exact value which is 158.113.

Why it is happening you can clearly see there is no scope of losing significant digits in this expression unlike in the previous expression. Therefore, this loss of significant digits is very important in numerical analysis even if you are working with a very good method mathematically that can lead to very bad answers because of the loss of significant digits in your calculation somewhere.

And therefore, one has to be careful especially when you are implementing your method on a computer.

(Refer Slide Time: 14:21)



A Disasters Caused by Numerical Errors

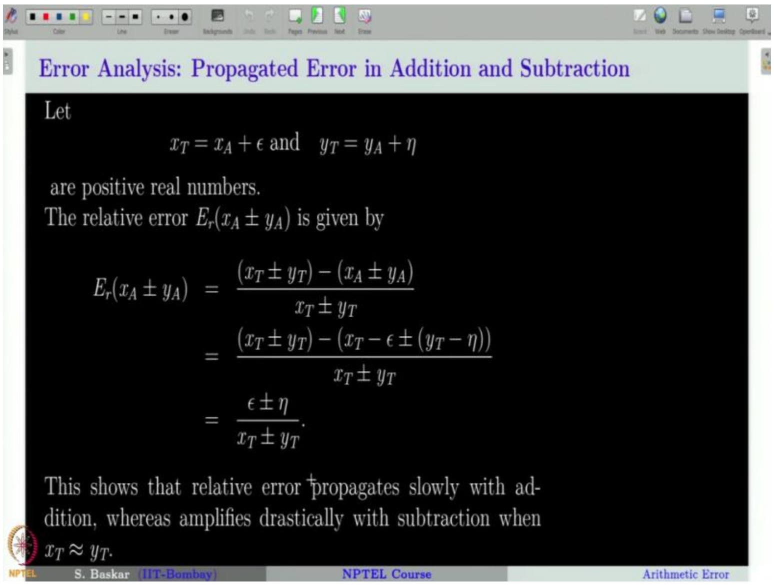
Patriot Missile Failure:
On February 25, 1991, during the Gulf War, an American Patriot Missile battery in Dharan, Saudi Arabia, failed to intercept an incoming Iraqi Scud missile. The Scud struck an American Army barracks and killed 28 soldiers. A report of the General Accounting office reported that cause was an inaccurate calculation of the time since boot due to computer arithmetic errors.

Reference: https://apps.dtic.mil/sti/citations/ADA344865

S. Baskar (IIT-Bombay) NPTEL Course Arithmetic Error

In fact, there are many disasters that happened in practical situations you can read one such disaster in this website and this disaster is due to the rounding error in a software.

(Refer Slide Time: 14:43)



Error Analysis: Propagated Error in Addition and Subtraction

Let

$$x_T = x_A + \epsilon \text{ and } \quad y_T = y_A + \eta$$

are positive real numbers.

The relative error $E_r(x_A \pm y_A)$ is given by

$$\begin{aligned} E_r(x_A \pm y_A) &= \frac{(x_T \pm y_T) - (x_A \pm y_A)}{x_T \pm y_T} \\ &= \frac{(x_T \pm y_T) - (x_T - \epsilon \pm (y_T - \eta))}{x_T \pm y_T} \\ &= \frac{\epsilon \pm \eta}{x_T \pm y_T}. \end{aligned}$$

This shows that relative error propagates slowly with addition, whereas amplifies drastically with subtraction when $x_T \approx y_T$.

S. Baskar (IIT-Bombay) NPTEL Course Arithmetic Error

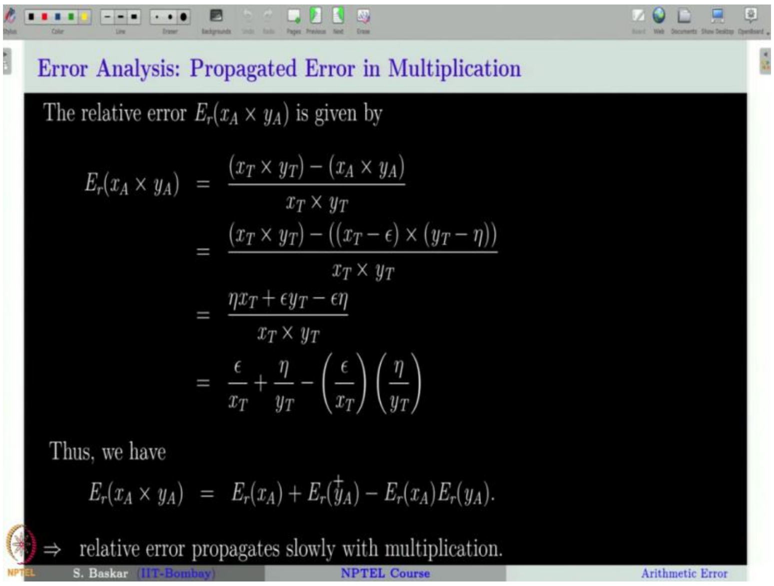
Next, let us pass on to the next topic of propagation of error in the four arithmetic operations. In this let us understand how the error from the initial numbers gets into the calculation when we use addition or subtraction or multiplication or division. Let us take as a first case the addition of two numbers. Assume that you have true values x_T and y_T . Instead of using x_T and y_T we use their approximation x_A and y_A where $x_T = x_A + \epsilon$ and $y_T = y_A + \eta$.

Now, our interest is to find $x_T + y_T$ or $x_T - y_T$. Instead of that we perform $x_A + y_A$ or $x_A - y_A$. Now the question is what is the relative error in this approximately computed value when compared to the value that is computed using true values. Therefore, let us take the definition of the relative error. The relative error is nothing, but the true value minus approximate value divided by the true value.

Now we will eliminate this x_A plus or minus y_A by the true value using these expressions that will happen to be the true value minus instead of x_A I am putting $x_T - \epsilon$ and instead of y_A I am putting $y_T - \eta$ divided by this true value. Now, we can just simplify this expression to get $\epsilon \pm \eta$ divided by the true value. Therefore, the relative error in sum or difference between two values which are approximately taken when compared to the corresponding true value is given by this expression.

If x_T and y_T are positive then addition will not make the relative error to grow drastically whereas the subtraction can make this when the true values are very close to each other. Therefore, subtraction is very dangerous when x_T and y_T are very close to each other whereas addition is not going to be dangerous at all when both x_T and y_T are positive numbers.

(Refer Slide Time: 17:50)



Error Analysis: Propagated Error in Multiplication

The relative error $E_r(x_A \times y_A)$ is given by

$$\begin{aligned} E_r(x_A \times y_A) &= \frac{(x_T \times y_T) - (x_A \times y_A)}{x_T \times y_T} \\ &= \frac{(x_T \times y_T) - ((x_T - \epsilon) \times (y_T - \eta))}{x_T \times y_T} \\ &= \frac{\eta x_T + \epsilon y_T - \epsilon\eta}{x_T \times y_T} \\ &= \frac{\epsilon}{x_T} + \frac{\eta}{y_T} - \left(\frac{\epsilon}{x_T}\right)\left(\frac{\eta}{y_T}\right) \end{aligned}$$

Thus, we have

$$E_r(x_A \times y_A) = E_r(x_A) + E_r(y_A) - E_r(x_A)E_r(y_A).$$

$\Rightarrow$ relative error propagates slowly with multiplication.

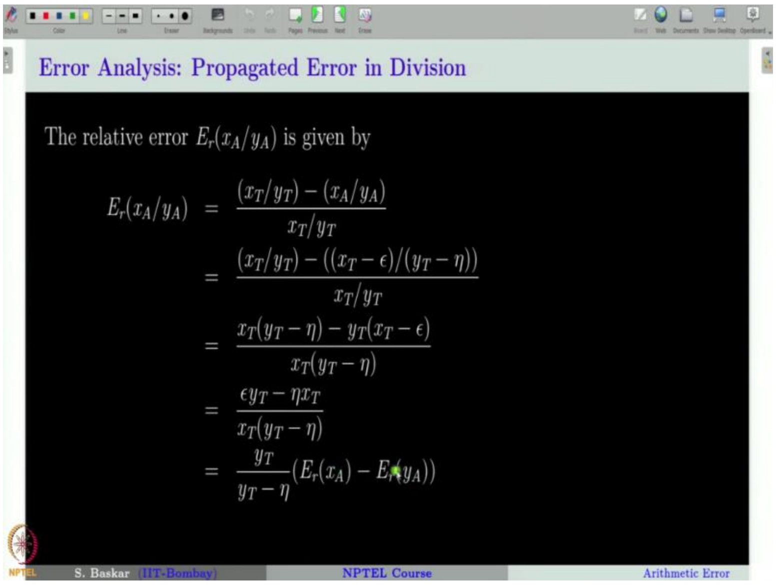
S. Baskar (IIT-Bombay) NPTEL Course Arithmetic Error

Let us try to see the propagation of relative error in multiplication. Again, what is our interest? Our interest is to find $x_T \times y_T$ instead of this we have performed $x_A \times y_A$ where x_A is an approximation to x_T and y_A is an approximation to y_T. Now our interest is to see what is the relative error in this approximate value when compared to the true value the definition is obviously this.

Now, as we did in the previous case we will replace x_A by $x_T - \epsilon$ and y_A by $y_T - \eta$ and that give us this expression and again you go to simplify this to get this expression. Now, you can write this as $\frac{\epsilon}{x_T}$ this is nothing, but the relative error in x_A when compared to x_T and this is the relative error in y_A when compared to y_T. I am just splitting this sum and writing them in individual terms and then you have this.

So, that gives a nice expression for the relative error in the product of two numbers as sum of the relative errors minus this. You can see that if $E_r(x_A)$ and $E_r(y_A)$ are very small numbers which may be true in practical then the product of those two small numbers will be much more smaller. Therefore, the relative error in product will actually amplify the error only by the sum of the relative errors in their initial data. Therefore, in this way multiplication is also not so dangerous.

(Refer Slide Time: 19:53)



Error Analysis: Propagated Error in Division

The relative error $E_r(x_A/y_A)$ is given by

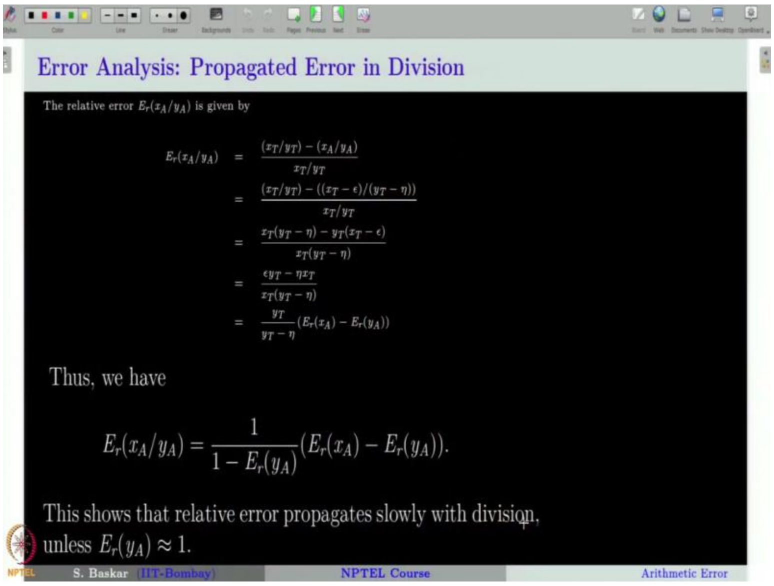
$$\begin{aligned} E_r(x_A/y_A) &= \frac{(x_T/y_T) - (x_A/y_A)}{x_T/y_T} \\ &= \frac{(x_T/y_T) - ((x_T-\epsilon)/(y_T-\eta))}{x_T/y_T} \\ &= \frac{x_T(y_T-\eta) - y_T(x_T-\epsilon)}{x_T(y_T-\eta)} \\ &= \frac{\epsilon y_T - \eta x_T}{x_T(y_T-\eta)} \\ &= \frac{y_T}{y_T-\eta}(E_r(x_A) - E_r(y_A)) \end{aligned}$$

S. Baskar (IIT-Bombay) NPTEL Course Arithmetic Error

Let us go on to division and see how division amplifies the relative error. Here you can see that the relative error definition is this and again you replace x_A by $x_T - \epsilon$ and similarly y_A and you will get this expression. Again one has to simplify this to get this expression and now you can further simplify it to get this expression. These are very simple ideas therefore you can just understand how I am simplifying from one step to the other step.

And now finally you can write this expression as $\frac{y_T}{y_T-\eta}$ into the difference between the relative errors in x_A and y_A.

(Refer Slide Time: 20:54)



Error Analysis: Propagated Error in Division

The relative error $E_r(x_A/y_A)$ is given by

$$\begin{aligned} E_r(x_A/y_A) &= \frac{(x_T/y_T) - (x_A/y_A)}{x_T/y_T} \\ &= \frac{(x_T/y_T) - ((x_T-\epsilon)/(y_T-\eta))}{x_T/y_T} \\ &= \frac{x_T(y_T-\eta) - y_T(x_T-\epsilon)}{x_T(y_T-\eta)} \\ &= \frac{\epsilon y_T - \eta x_T}{x_T(y_T-\eta)} \\ &= \frac{y_T}{y_T-\eta}(E_r(x_A) - E_r(y_A)) \end{aligned}$$

Thus, we have

$$E_r(x_A/y_A) = \frac{1}{1-E_r(y_A)}(E_r(x_A) - E_r(y_A)).$$

This shows that relative error propagates slowly with division, unless $E_r(y_A) \approx 1$.

S. Baskar (IIT-Bombay) NPTEL Course Arithmetic Error

You can write it as 1 by 1 – relative error in y_A into this. What I am doing I am just taking y_T outside and therefore I am getting $\frac{1-\eta}{y_T}$ that is the relative error in y_A. So, y_T comes out and gets cancelled with the numerator and therefore 1 by 1 – relative error in y_A into this. Now, what this expression tells us is that, if you make a error in y_A which is 100 percent error then this relative error will be infinity that is what it says.

It means if you make 100 percent error in y_A then the division will be a disaster that is what this expression says that is quite acceptable because in any circumstances we do not intend to make 100 percent error if you make 100 percent error then our answer has to be bad. So, given that we do not make such a big error division can be regarded as a nice operation. Now from these four analysis what we understand is only subtracting two positive number is going to be very dangerous in any calculation on a computer especially.

Whereas other three arithmetic operations are relatively peaceful when compared to the subtraction and our previous examples shows that subtracting two very close numbers will lead to loss of significance and that loss of significance is in general dangerous in the sense that they can amplify the error drastically. We will continue our discussion in the next class with condition number of a function.

And we will try to understand what kind of functions can be more nice to be evaluated on a computer based on their condition numbers. Thanks for your attention.