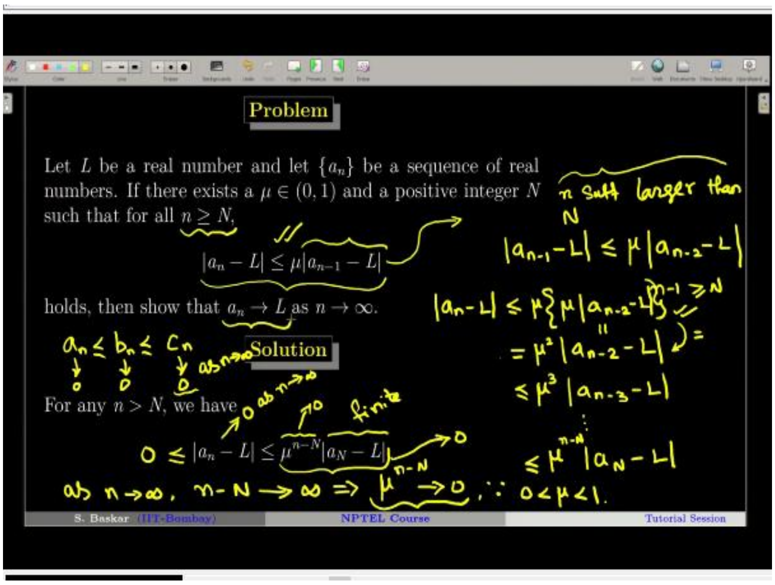
Numerical Analysis

Prof. S. Baskar
Department of Mathematics
Indian Institute of Technology-Bombay

Lecture-07
Tutorial 1: Mathematical Preliminaries, Arithmetic Error

Hi, this is our first tutorial session. In this session we will try to solve some important problems given in our notes especially from the mathematical preliminaries and also from the chapter on arithmetic errors.

(Refer Slide Time: 00:39)



Problem

Let L be a real number and let $\{a_n\}$ be a sequence of real numbers. If there exists a $\mu \in (0,1)$ and a positive integer N such that for all $\underbrace{n \geq N}$,

$$\underbrace{|a_n - L| \leq \mu \overbrace{|a_{n-1} - L|}}$$

holds, then show that $\underbrace{a_n \to L}$ as $n \to \infty$.

$\overbrace{n \text{ suff larger than } N}$

$|a_{n-1} - L| \leq \mu |a_{n-2} - L|$

$|a_n - L| \leq \mu \{\mu |a_{n-2} - L|\}$ $n-1 \geq N$

$= \mu^2 |a_{n-2} - L|$

$\leq \mu^3 |a_{n-3} - L|$

$\vdots$

$\leq \mu^{n-N} |a_N - L|$

$a_n \leq b_n \leq c_n$, $a_n \to 0$, $b_n \to 0$, $c_n \to 0$ as $n \to \infty$

Solution

For any $n > N$, we have

$$0 \leq |a_n - L| \leq \underbrace{\overbrace{\mu^{n-N}}^{\to 0}\overbrace{|a_N - L|}^{\text{finite}}}_{\to 0}$$

$\to 0$ as $n \to \infty$

as $n \to \infty$, $n - N \to \infty \Rightarrow \underbrace{\mu^{n-N} \to 0}$, $\because 0 < \mu < 1$.

S. Baskar (IIT-Bombay) NPTEL Course Tutorial Session

Let us see how to solve this problem. Here we are given that a sequence $\{a_n\}$ is going to satisfy some condition. What is that condition? $|a_n - L|$ is less than or equal to some constant $\mu|a_{n-1} - L|$. And this happens for sufficiently large N that is what the assumption says, not first few terms but after certain terms say for some positive integer N all the terms a_n's for $n \geq N$ we have this condition.

And further the question also says that the constant μ is something lies between 0 and 1. That is very important. Now we are asked to prove that the sequence a_n converges to L as n tends to infinity. That is what the question is. Well, let us see how to prove this result. To prove this result first we have to obtain this inequality. Let us see how to obtain this inequality, it is not very difficult.

It goes with an idea which is quite often used in our course. Therefore, it is important for us to understand and keep this in mind, the idea goes like this. See we are given this inequality. Now the same inequality can also be applied to the right-hand side. For that let us take n sufficiently large then you can apply the same inequality for $a_{n-1} - L$. That will be less than or equal to $\mu|a_{n-2} - L|$.

In this you have to also make sure that $n - 1$ is also greater than or equal to N. that is why we have taken n very large when compared to N and now once you have this you can see that $|a_n - L| \leq \mu\mu|a_{n-2} - L|$ and that can be written as $\mu^2|a_{n-2} - L|$. Remember, I am putting equal to sign here because this term that is the right-hand side term is equal to this term.

It does not mean the left-hand side is equal to this. Often students get confused whether this is equal to this? No, this is still less than or equal to this we are writing only because the right-hand side is equal to the second step of the right answer. And now what you can do is you can again apply the same inequality for now $|a_{n-2} - L|$ and that will give us less than or equal to $\mu^3|a_{n-2} - L|$.

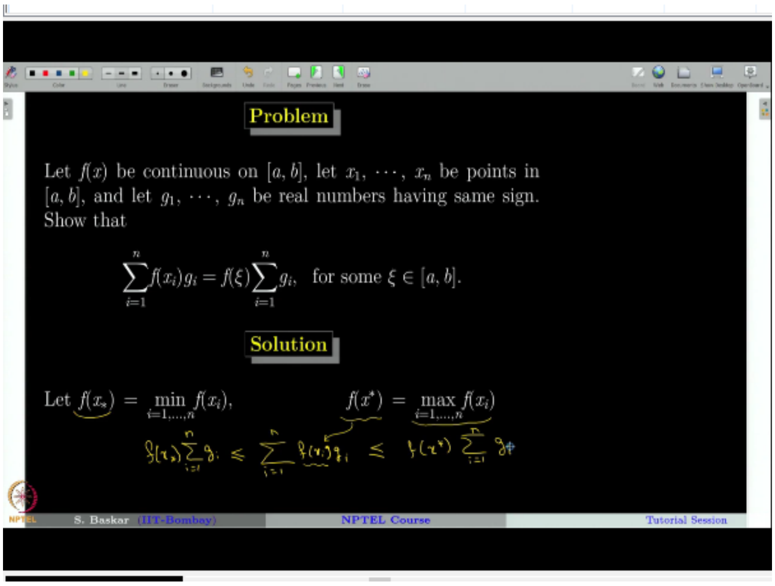
Like that now you can keep on going up to what term we can go well, we are given permission to use this inequality only up to $n = N$. Therefore, we can go only up to $|a_n - L|$ here. So, when you go up to $|a_n - L|$ your μ will have $n - N$ on the power. Therefore, this will be less than or equal to this. That is how we landed up with this inequality. Once you establish this inequality the conclusion comes almost trivially because μ is less than 1.

Now you take $n \rightarrow \infty$ and see what happens. When you take $n \rightarrow \infty$ you can see that $n - N$ will also tend to infinity and that implies $\mu^{n-N} \rightarrow 0$ why since $0 < \mu < 1$. Because of that we have this property. Now you can see that this term is going to 0 and this term is a finite quantity and therefore the entire term will go to 0.

On the left-hand side you can see that we have taken modulus for this. Therefore, this is surely greater than or equal to 0. Now if you recall in the sandwich theorem if you have two sequences $\{a_n\}$ and $\{c_n\}$ and you know that $\{a_n\} \leq \{b_n\} \leq \{c_n\}$. And if this goes to say 0 and this also goes to 0 as $n \rightarrow \infty$, then this will also go to 0. In the sandwich theorem we have stated it with any limit.

Here we are using it for that limit as 0 and you see by using sandwich theorem therefore you can see that this term also goes to 0 as $n \to \infty$. That is equivalent to saying that $a_n \to L$. Now let us go on to the next problem.

(Refer Slide Time: 07:13)



Problem

Let $f(x)$ be continuous on $[a, b]$, let $x_1, \cdots, x_n$ be points in $[a, b]$, and let $g_1, \cdots, g_n$ be real numbers having same sign. Show that

$$\sum_{i=1}^{n} f(x_i) g_i = f(\xi) \sum_{i=1}^{n} g_i, \quad \text{for some } \xi \in [a, b].$$

Solution

Let $\underbrace{f(x_*)}$ $= \min_{i=1,\ldots,n} f(x_i)$, $\qquad \underbrace{f(x^*)} = \underbrace{\max_{i=1,\ldots,n} f(x_i)}$

$$f(x_*) \sum_{i=1}^{n} g_i \leq \sum_{i=1}^{n} \underbrace{f(x_i)} g_i \leq f(x^*) \sum_{i=1}^{n} g_i$$

S. Baskar (IIT-Bombay) NPTEL Course Tutorial Session

The next problem is a discrete version of the second mean value theorem for integration. The problem says that you have a function f which is continuous on the interval $[a, b]$ and you are also given some n points taken from the interval $[a, b]$ and also you are given some values g_1, g_2 up to g_n. They are given to be real numbers and more importantly they are of one sign. It means either everybody is positive or all these g's are negative.

There is no sign change among these g's. That is what it means. Then we have to show that

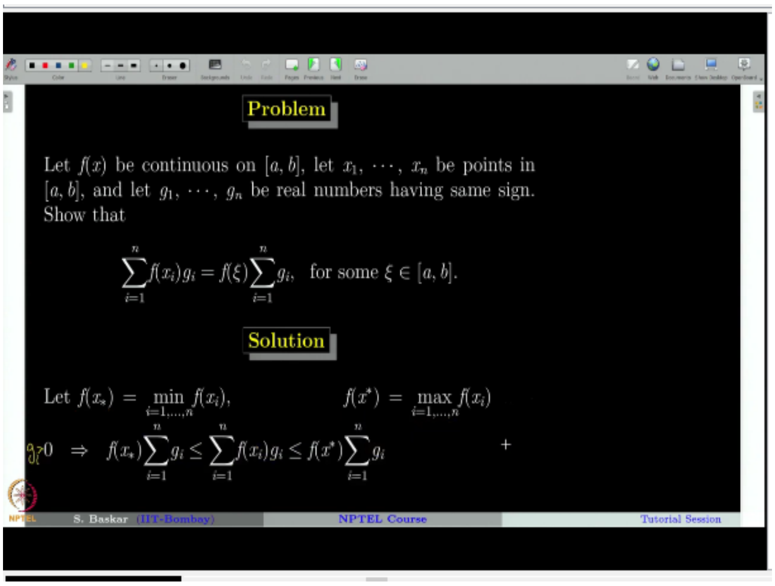
$$\sum_{i=1}^{n} f(x_i) g_i = f(\xi) \sum_{i=1}^{n} g_i$$

Where the ξ is some number lying between $[a, b]$. That is the problem how to prove this? This is a very simple application of the intermediate value theorem. What you do is, just take the minimum over all the values of f at the points x_i.

Call it as $f(x_*)$ and similarly take the maximum of all these numbers, call this as $f(x^*)$. Now what are we going to do with that? Well we have the left-hand side $\sum_{i=1}^{n} f(x_i) g_i$. Now if you replace all these terms by the minimum then you will have this is less than or equal to $f(x_*)$, it will come out of this sum because now it is independent of i, into $\sum_{i=1}^{n} g_i$.

Similarly if you replace all this by the maximum then it will be less than or equal to $f(x^*)$ and since this is independent of i, this term will come out of each term and you will have $\sum_{i=1}^{n} g_i$. That is very simple to understand.

(Refer Slide Time: 09:43)



Problem

Let $f(x)$ be continuous on $[a,b]$, let $x_1, \cdots, x_n$ be points in $[a,b]$, and let $g_1, \cdots, g_n$ be real numbers having same sign. Show that

$$\sum_{i=1}^{n} f(x_i) g_i = f(\xi) \sum_{i=1}^{n} g_i, \quad \text{for some } \xi \in [a,b].$$

Solution

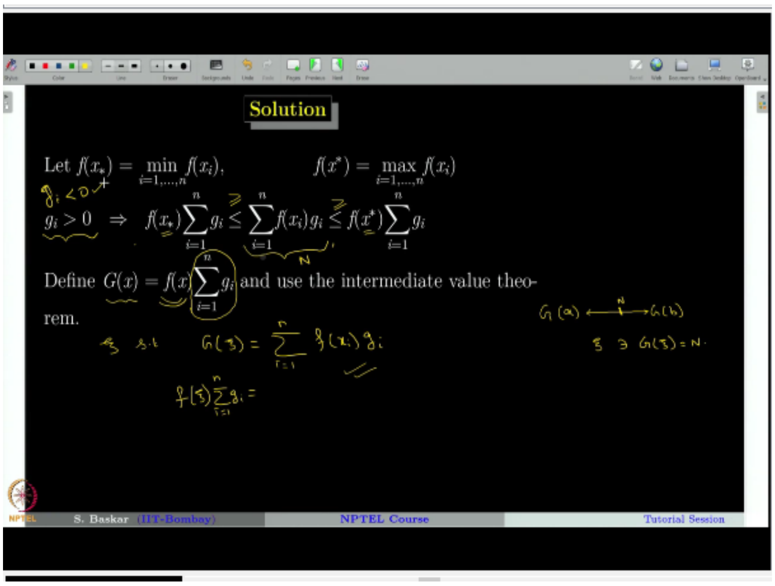
Let $f(x_*) = \min_{i=1,\ldots,n} f(x_i)$, $\qquad f(x^*) = \max_{i=1,\ldots,n} f(x_i)$

$$g_i > 0 \Rightarrow f(x_*) \sum_{i=1}^{n} g_i \le \sum_{i=1}^{n} f(x_i) g_i \le f(x^*) \sum_{i=1}^{n} g_i$$

S. Baskar (IIT-Bombay) NPTEL Course Tutorial Session

So, therefore we have this.

(Refer Slide Time: 09:42)



Solution

Let $f(x_*) = \min_{i=1,\ldots,n} f(x_i)$, $\qquad f(x^*) = \max_{i=1,\ldots,n} f(x_i)$

$g_i < 0$ ✓+

$$g_i > 0 \Rightarrow f(x_*) \sum_{i=1}^{n} g_i \overset{\geq}{\le} \underbrace{\sum_{i=1}^{n} f(x_i) g_i}_{N} \overset{\geq}{\le} f(x^*) \sum_{i=1}^{n} g_i$$

Define $G(x) = f(x) \sum_{i=1}^{n} g_i$ and use the intermediate value theorem.

ξ s.t. $\quad G(\xi) = \sum_{i=1}^{n} f(x_i) g_i$

$f(\xi) \sum_{i=1}^{n} g_i =$

$G(a) \longleftarrow \overset{N}{+} \longrightarrow G(b)$

$\xi \ni G(\xi) = N.$

S. Baskar (IIT-Bombay) NPTEL Course Tutorial Session

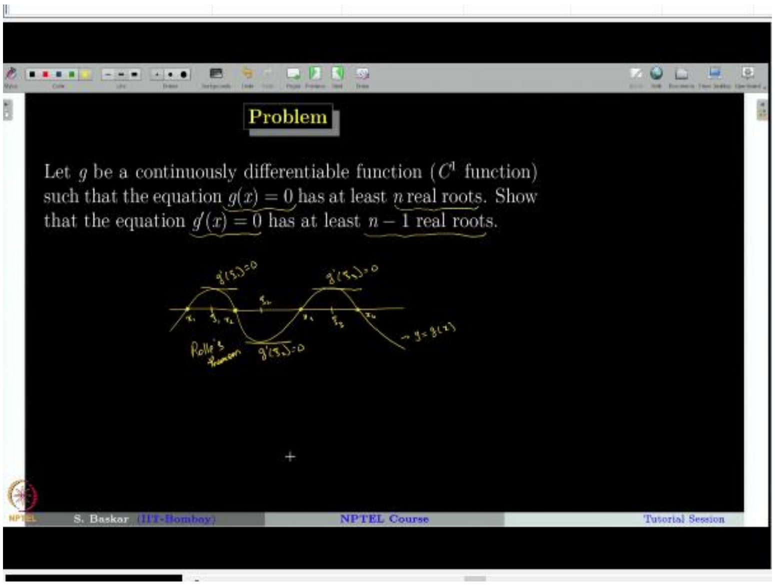
Remember for this we have used the fact that g is of one sign and we assume that g is greater than 0. Otherwise, if all g's are less than 0 you will have a reverse inequality. That is all. Otherwise the idea of the proof goes exactly same as in the case of g_i's to be greater than 0. So, here itself we have to assume that g_i's are greater than 0 and then we get this inequality.

Now what we will do is, we will take these terms and define a function $G(x)$ is equal to instead of x_* and $f(x^*)$ we will take it as $f(x)\sum_{i=1}^{n} g_i$. You can see that g is a continuous function why because f is given to be a continuous function and this is something which is a fixed number positive number as per our assumption.

Therefore $G(x)$ is constant times a continuous function therefore G itself is a continuous function. Now if you recall you have a continuous function g and you have G of say a and $G(b)$. Then given any number between these two numbers say n, you can find the ξ such that $G(\xi) = N$. That is what the intermediate value theorem says. So, we will just use this now to get a ξ such that.

Now you see you take this as your n, $G(\xi)$ is equal to this $\sum_{i=1}^{n} f(x_i)g_i$. So, that is what precisely we wanted to show that this term can be written as $f(\xi)\sum_{\text{i}=1}^{\text{n}} g_i$. So, that is what we wanted to show here and that comes directly once if you write this inequality it comes directly from the intermediate value theorem. Similarly, if $g_i < 0$ then as I told the inequality will reverse that is all. The same idea for conclusion will go through even in this case also.

(Refer Slide Time: 12:52)



Problem

Let g be a continuously differentiable function (C^1 function) such that the equation $\underbrace{g(x) = 0}$ has at least $\underbrace{n \text{ real roots}}$. Show that the equation $\underbrace{g'(x) = 0}$ has at least $\underbrace{n - 1 \text{ real roots}}$.

S. Baskar (IIT-Bombay) NPTEL Course Tutorial Session

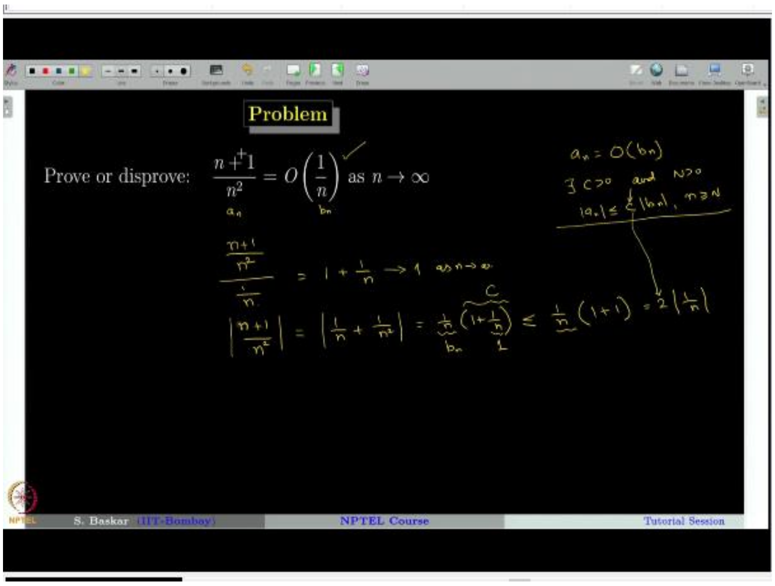
Let us go to the next problem. Again, this property is often used in our course especially if you go through the proof of Taylor's theorem and also the theorem or polynomial interpolations we will use this idea. What this problem says? You have a continuously differentiable function g and we know that the equation $g(x) = 0$. This is the equation, it has at least n real roots.

Then the equation $g'(x) = 0$ will have at least $n - 1$ real roots. This idea is used as I told in the proof of Taylor's theorem as well as in the proof of the error analysis of the polynomial interpolations. What is the idea behind this? Well, you know that the equation $g(x)$ say it has n roots, say for instance these are the roots of the equation $g(x) = 0$. Let us say this is the graph of the function $y = g(x)$ and you have x_1, x_2, x_3 and x_4 are the roots of this equation.

Then you have to show that g' has at least $n - 1$. Say for instance if this equation that is $g(x) = 0$ has 4 roots then $g'(x) = 0$ should have 3 roots. How will you get? Well you use the Roll's theorem. Roll's theorem says that in between these two points you have a point ξ such that let us call it as ξ_1 such that $g'(\xi_1) = 0$. Similarly here also you can apply the Roll's theorem between x_2 and x_3.

And you will get say ξ_2 such that $g'(\xi_2) = 0$ and similarly here also you have ξ_3 and $g'(\xi_3) = 0$. So, that is the idea, it is just a direct application of the Roll's theorem. I hope you can write the solution for this problem now.

(Refer Slide Time: 15:31)



Problem

Prove or disprove: $\underbrace{\frac{n+1}{n^2}}_{a_n} = O\Big(\underbrace{\frac{1}{n}}_{b_n}\Big)$ ✓ as $n \to \infty$

$a_n = O(b_n)$

$\exists\, c > 0$ and $N > 0$

$|a_n| \le c|b_n|,\ n \ge N$

$$\frac{\frac{n+1}{n^2}}{\frac{1}{n}} = 1 + \frac{1}{n} \to 1 \text{ as } n \to \infty$$

$$\left|\frac{n+1}{n^2}\right| = \left|\frac{1}{n} + \frac{1}{n^2}\right| = \underbrace{\frac{1}{n}}_{b_n}\overbrace{\Big(1+\underbrace{\frac{1}{n}}_{1}\Big)}^{c} \le \frac{1}{n}(1+1) = 2\left|\frac{1}{n}\right|$$

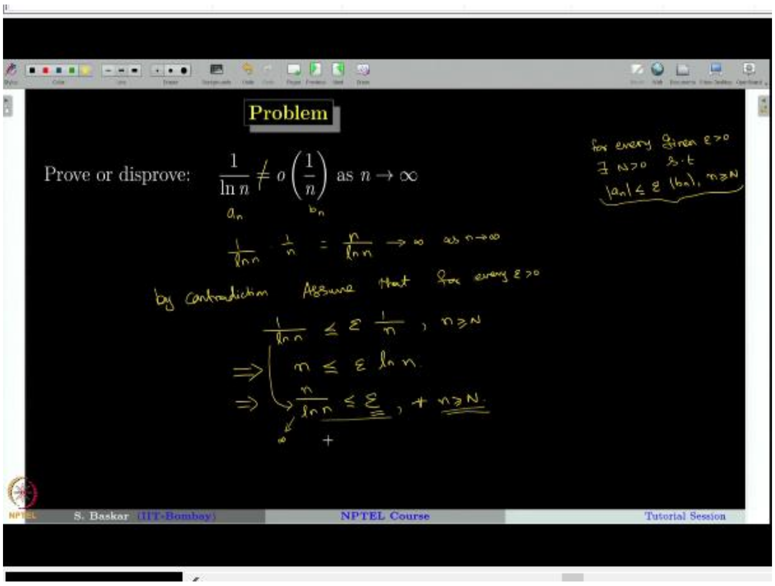
Well now let us pass on to the discussion on the order of convergence, if you recall we have introduced two notations; one is O and o. When will you say that a sequence $a_n = O(b_n)$. If you recall you should be able to find a constant c and a n such that $|a_n| \le c|b_n|$ for all sufficiently large Ns. So, that is what we have to show. In order to show that $a_n = O(b_n)$.

Here $a_n = \frac{n+1}{n^2}$ and $b_n = \frac{1}{n}$. One way is to directly take this term $\frac{n+1}{n^2}$ divided by $\frac{1}{n}$ and see that this is bounded. As n tends to infinity this should be a bounded quantity. This is obviously equal to $1 + \frac{1}{n}$ and that tends to 1 as i tends to infinity. Therefore, this is correct that this sequence is a $O\left(\frac{1}{n}\right)$. You can also directly use the definition.

In that case you have to write $\left|\frac{n+1}{n^2}\right|$ and that is written as $\left|\frac{1}{n} + \frac{1}{n^2}\right|$ which is equal to $\frac{1}{n}\left(1 + \frac{1}{n}\right)$ and that can be written as less than or equal to $\frac{1}{n}$. I will just replace this by 1. I just want a constant here that is the idea and b_n is nothing but $\frac{1}{n}$. Therefore, I am having already b_n here, $\frac{1}{n}$ here. This I have to somehow freeze this n and make it a constant.

For that I am dominating $\frac{1}{n}$ by just 1, you can dominate it by 2, 3 and so on anything any finite number, but I am just dominating it by 1 and that gives me $2\left(\frac{1}{n}\right)$. Therefore in this case my c is just 2 and that shows also from the definition that this sequence is a $O\left(\frac{1}{n}\right)$.

(Refer Slide Time: 18:25)



Problem

Prove or disprove: $\underset{a_n}{\frac{1}{\ln n}} \neq o\underset{b_n}{\left(\frac{1}{n}\right)}$ as $n \to \infty$

for every given $\varepsilon > 0$ $\exists\ N > 0$ s.t $|a_n| \leq \varepsilon\,|b_n|,\ n \geq N$

$\frac{1}{\ln n} \cdot \frac{1}{n} = \frac{n}{\ln n} \to \infty$ as $n \to \infty$

by contradiction Assume that for every $\varepsilon > 0$

$\frac{1}{\ln n} \leq \varepsilon \frac{1}{n}$, $n \geq N$

$\Rightarrow$ $n \leq \varepsilon \ln n$.

$\Rightarrow$ $\underset{\infty}{\frac{n}{\ln n}} \leq \varepsilon$, $\forall$ $n \geq N$.

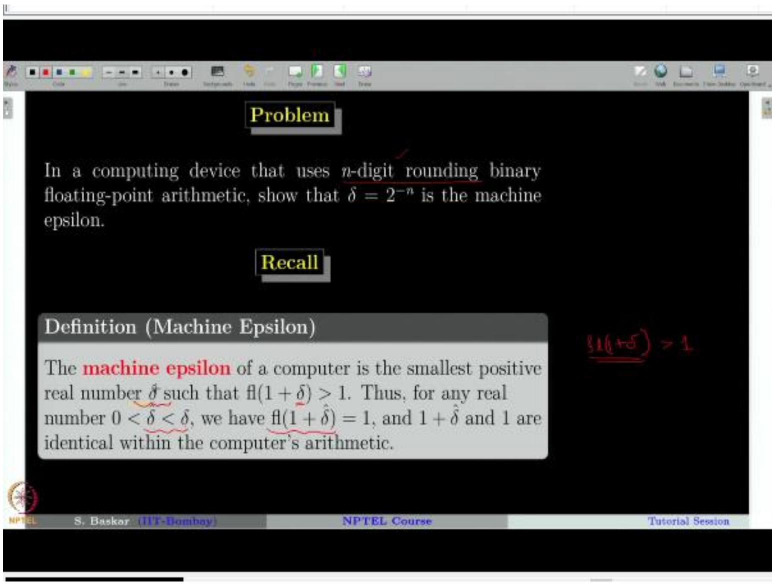
Let us take the other problem. Here we have $a_n = \frac{1}{ln\, n}$ and you want to show that this is $o\left(\frac{1}{n}\right)$. Therefore b_n is this. If you recall we have to here show that for every given $\epsilon > 0$ we have to find a N such that $|a_n| \leq \epsilon |b_n|$ for sufficiently large Ns. That is what we have to show. Now one way to see is to directly compute the limit and see this $\frac{1}{ln\, n}\frac{1}{n} = \frac{n}{ln\, n}$.

That goes to infinity as n tends to infinity. Therefore this is actually a_n is not a $o\left(\frac{1}{n}\right)$. Now if you want to see it through this definition how will you see? Well we can see by contradiction. Assume that $\frac{1}{ln\, n} \leq \epsilon \frac{1}{n}$ for some sufficiently large N and for some ϵ or rather we should write for every ϵ because for every given ϵ this should happen.

So, for every given ϵ, I will find a n such that this happen. That is what I am assuming and I will show that it leads to a contradiction. How it leads to a contradiction? This implies $n \leq \epsilon \ln \mathrm{n}$ and that implies $\frac{n}{ln\, n} \leq \epsilon$. Well, I can directly write it from here for all $n \geq N$. You can see that the right-hand side is tending to infinity as n tends to infinity.

This should hold for all n sufficiently large. So, that cannot happen because this side is going to infinity whereas this is some very small number if I take then this cannot happen. So, that is a contradiction.

(Refer Slide Time: 21:20)



Problem

In a computing device that uses n-digit rounding binary floating-point arithmetic, show that $\delta = 2^{-n}$ is the machine epsilon.

Recall

Definition (Machine Epsilon)

The **machine epsilon** of a computer is the smallest positive real number δ such that $\mathrm{fl}(1+\delta) > 1$. Thus, for any real number $0 < \hat{\delta} < \delta$, we have $\mathrm{fl}(1+\hat{\delta}) = 1$, and $1+\hat{\delta}$ and 1 are identical within the computer's arithmetic.

$\mathrm{fl}(1+\delta) > 1$

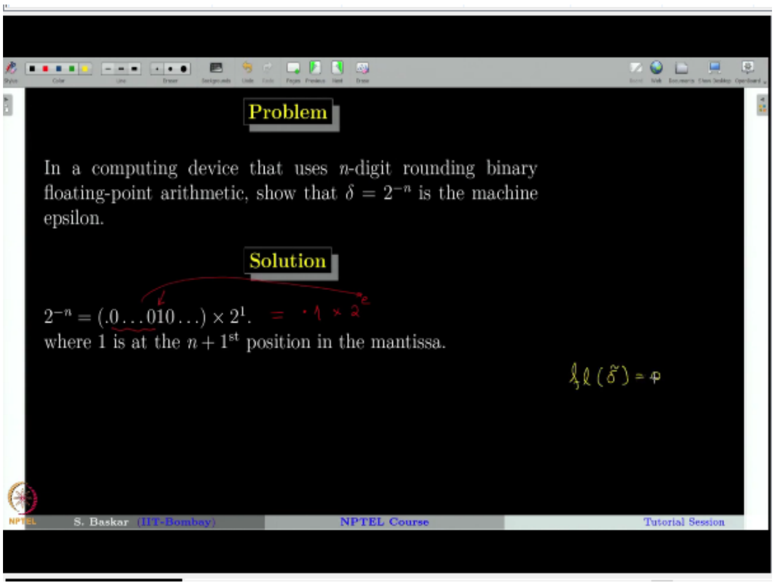
S. Baskar (IIT-Bombay) NPTEL Course Tutorial Session

Let us now solve some problems in the chapter on arithmetic errors. Here let us take this problem. The problem says that we have a computing device that uses *n-digit* rounding binary floating-point arithmetic. Show that 2^{-n} is the machine epsilon. Let us recall what is mean by machine epsilon. Machine epsilon means it is a very small number which we denote by δ.

Such that you take that δ add it with 1 and then take the floating-point approximation, whatever the floating-point approximation that you want to take well in this problem it is the n-digit rounding is the floating-point approximation. Then that delta will actually give some number which is greater than 1. However, if you take any number less than δ, whatever may be the number it is.

If you take any number less than the δ then the floating-point approximation in our problem again it is n-digit rounding should actually give 1, it should not give anything greater than 1. Only for δ it will give that is the last number for which it will recognize it as a non-zero number, anything less than that the computing device will recognize it as 0 only. That is what is called the machine epsilon.

(Refer Slide Time: 23:04)



Problem

In a computing device that uses n-digit rounding binary floating-point arithmetic, show that $\delta = 2^{-n}$ is the machine epsilon.

Solution

$2^{-n} = (.\underbrace{0\ldots}010\ldots) \times 2^1.$ $= .1 \times 2^{e}$

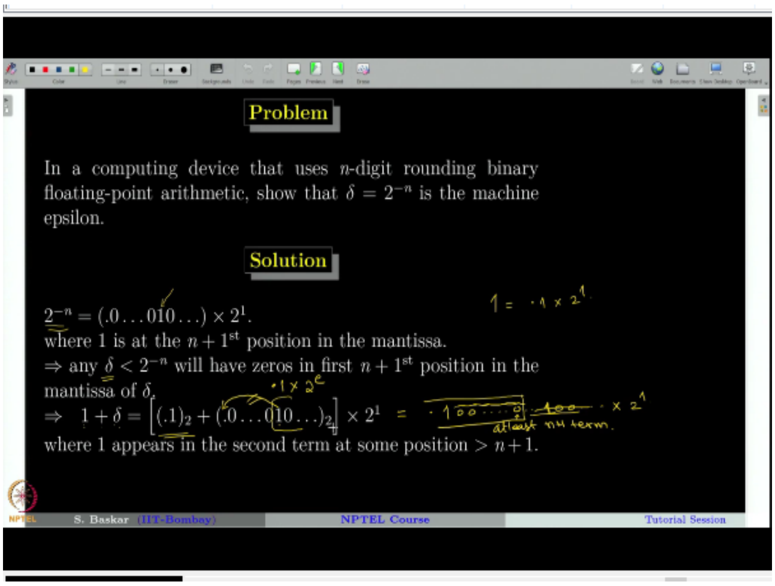
where 1 is at the $n + 1^{\text{st}}$ position in the mantissa.

$fl(\tilde{\delta}) = \varphi$

S. Baskar (IIT-Bombay) NPTEL Course Tutorial Session

Now let us see how to solve this problem. Well, if you see how to represent this number in the binary form. We can write it as .0000 up to n terms and n + first term is going to have 1 and then 0’s, into 2^1. So, if you recall its floating-point approximation is 0.1 if you write and then into 2^e. So, whatever may be the e that comes here you have to write. But I am just writing it in the direct form.

(Refer Slide Time: 23:43)



Problem

In a computing device that uses n-digit rounding binary floating-point arithmetic, show that $\delta = 2^{-n}$ is the machine epsilon.

Solution

$2^{-n} = (.0\dots010\dots) \times 2^1$.

where 1 is at the $n+1^{\text{st}}$ position in the mantissa.

$\Rightarrow$ any $\delta < 2^{-n}$ will have zeros in first $n+1^{\text{st}}$ position in the mantissa of δ.

$\Rightarrow \; 1+\delta = \left[(.1)_2 + (.0\dots010\dots)_2\right] \times 2^1$ = $.100\dots0$ ~~100~~ $\times 2^1$

where 1 appears in the second term at some position $> n+1$.

And then you can see that the 1 is sitting at the n + first position. Now any δ that you take which is less than 2^{-n} will have 0 in first n + first position, even in this position it will have 0 and then it will have 1 somewhere else. That is how δ which is less than 2^{-n} will look like. Therefore, if I add 1 + that δ what I will have, well 1 can be written as 0.1 into 2 to the power of 1.

And the δ will have first many terms 0 at least n + 1 terms are 0 and then you will have 1 somewhere after $(n + 1)$th term. Now if you see this will be equal to .1 then 0, 0, many 0s at least n + 1 terms. That is what we know because 2^{-n} has 1 at the n + first term. Therefore any δ will have 0 at the n + first term at least and have 1 somewhere else into 2^1.

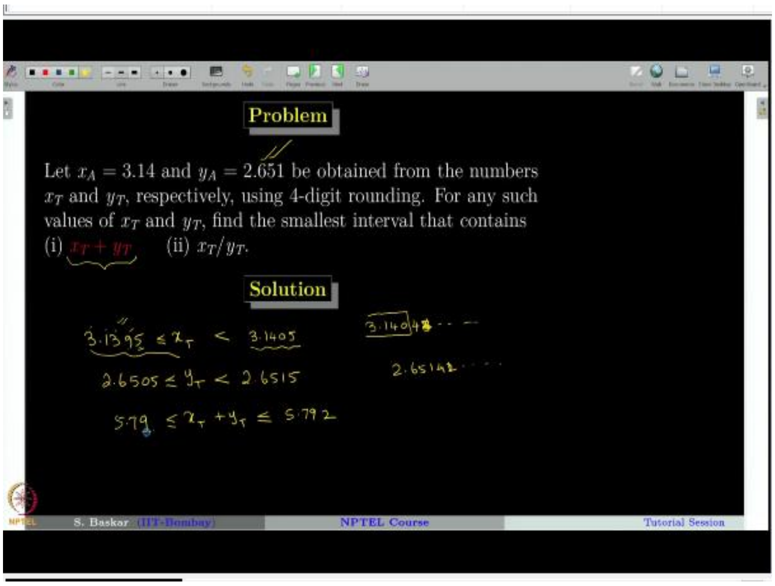
Now what I will do is, I will take the floating-point approximation of this. Thereby I will do the rounding up to here. When I do rounding up to here that all these are zeros and the non-zero term is actually truncated here and thereby the floating-point approximation of 1 + δ will remain as 1 because I have truncated up to n + first term and up to that there was no non-zero term in the representation of δ and therefore you have 1.

That shows that 2^{-n} is the machine epsilon for the computing device that uses n-digit rounding binary floating-point representation. You can also understand from here why the definition of machine epsilon demands that the floating-point approximation of 1 + δ should be equal to 1. Because if you do not have that if you just say that $\text{fl}(\delta)$ should be equal to 0.

If you say then the floating-point representation of that will actually push that non zero one to the first position and thereby when you do the n-digit rounding you will not be losing that non zero information in the floating-point level. That is why in order to lose that information, you are actually adding 1 to it here and that is how this information is lost. Otherwise, this would have written as .1 into 2 to the power of some e.

And therefore this 1 would have come to the first position and when you do the n-digit rounding of that you will not be losing this information. So, you are adding 1 to it and checking. That is the reason why we have given one in the definition of the machine epsilon here. This machine epsilon is a very important concept in computation one has to understand this.

(Refer Slide Time: 27:48)



Problem

Let $x_A = 3.14$ and $y_A = 2.651$ be obtained from the numbers x_T and y_T, respectively, using 4-digit rounding. For any such values of x_T and y_T, find the smallest interval that contains (i) $\underbrace{x_T + y_T}$ (ii) x_T/y_T.

Solution

$\underbrace{3.1395} \leq x_T < \underbrace{3.1405}$

$2.6505 \leq y_T < 2.6515$

$5.79 \leq x_T + y_T \leq 5.792$

$\boxed{3.140}4\ldots$

$2.6514\ldots$

S. Baskar (IIT-Bombay) NPTEL Course Tutorial Session

Let us take the next problem. Here we have the number x_A that is given as 3.14 and y_A is given as 2.651. These are given to us as approximate numbers generated from some true numbers x_T and y_T. We do not know what are these two numbers, all we know is that we got this approximate numbers by using 4-digit rounding. We are not given these information y_T and x_T are not given to us.

We are only given that x_A is obtained by rounding 4-digits from x_T and similarly y_A is obtained by doing 4-digit rounding of y_T. Now we want to find the smallest interval that contains this number and the second subdivision is we want to find the smallest interval that contains $\frac{x_T}{y_T}$. Remember we do not know x_T and y_T. Therefore, we just have to find an estimate of x_T and y_T in terms of both the lower bound as well as the upper bound.

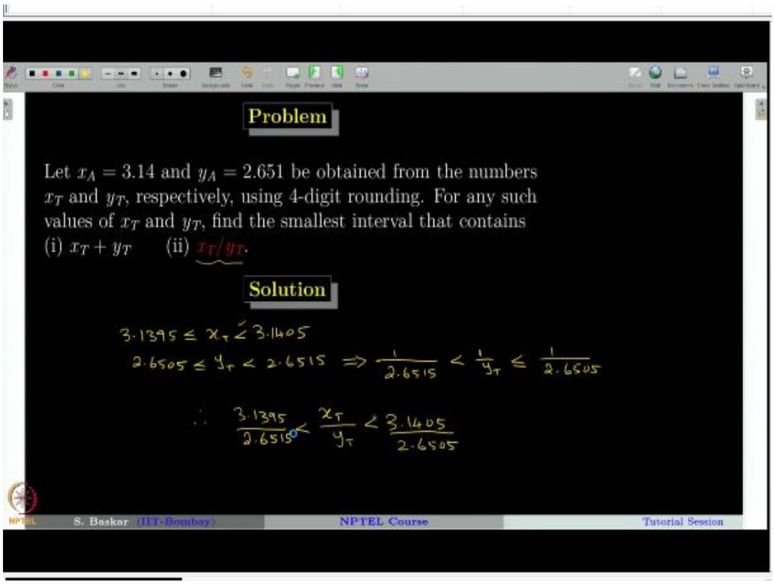
So, how to get the lower bound and upper bound of $x_T + y_T$? Let us see; for that first we have to understand what is the range for x_T and what is the range for y_T that is what we have to understand. You can see that x_T can be any number greater than or equal to 3.1395. You have to tightly find this number that should be the smallest number which when rounded should give you 3.14.

You can see that that is the smallest positive possible number that gives you 3.14 when rounded to 4-digits, 4-digits after writing the floating-point approximation. So, you have to do 4-digit rounding. It means 1, 2, 3, 4. This is 5, therefore it will be rounded to add 1 here and that will make this as 3.14. Similarly, the largest number that can lead to 3.14 is actually 3.1405.

Because you can see that 3.1404 also gives us when you do 4-digit rounding it gives 3.140 then 41, 42, 43, 44 and so on all these numbers when you round to 4-digit rounding it gives you 3.14. Therefore, x_T should be anything less than this number. That is the maximum possible that you can think and once you get this idea now it is very easy to solve this problem.

You can see y_T will lie between 2.6505 and that is less than or equal to; on the upper bound you have this is less than 2.6515 because 2.6514 will also give 2.6514 and 41, 42 and so on, any number will give this number. Once you have this now it is just a matter of adding this two number $x_T + y_T$ and that is less than or equal to 5.792 and this is less than or equal to or 5.79 roughly. So, that is what we get. Let us go to the next problem.

(Refer Slide Time: 32:18)



Problem

Let $x_A = 3.14$ and $y_A = 2.651$ be obtained from the numbers x_T and y_T, respectively, using 4-digit rounding. For any such values of x_T and y_T, find the smallest interval that contains (i) $x_T + y_T$ (ii) x_T/y_T.

Solution

$$3.1395 \le x_T < 3.1405$$

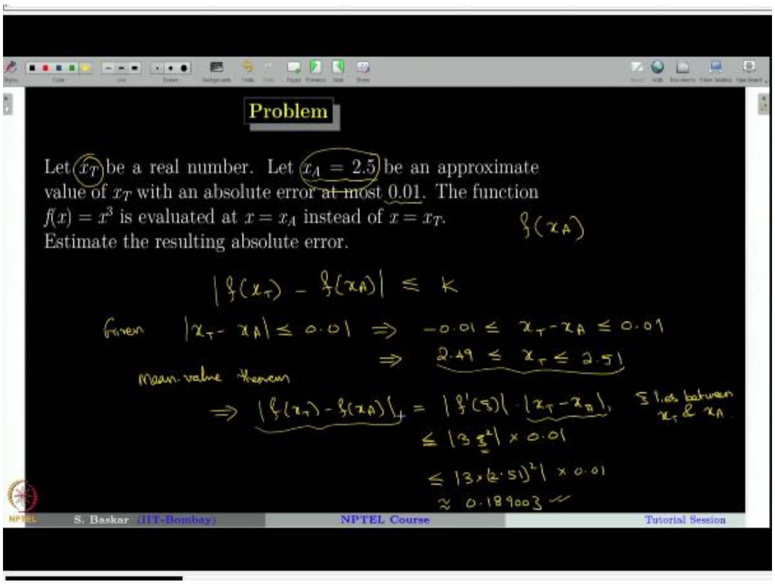
$$2.6505 \le y_T < 2.6515 \Rightarrow \frac{1}{2.6515} < \frac{1}{y_T} \le \frac{1}{2.6505}$$

$$\therefore \frac{3.1395}{2.6515} < \frac{x_T}{y_T} < \frac{3.1405}{2.6505}$$

S. Baskar (IIT-Bombay) NPTEL Course Tutorial Session

Next problem is to estimate $\frac{x_T}{y_T}$. Again, we have $3.1395 \le x_T \; < 3.1405$ and similarly $2.6505 \le y_T < 2.6515$ and that implies $\frac{1}{2.6515} < \frac{1}{y_T} \le \frac{1}{2.6505}$, I am just doing the reverse of this inequality and therefore $\frac{x_T}{y_T} < \frac{3.1405}{2.6505}$ and this side it is less than $\frac{3.1395}{2.6515}$. That is the answer for this problem.

(Refer Slide Time: 33:55)



Problem

Let x_T be a real number. Let $x_A = 2.5$ be an approximate value of x_T with an absolute error at most 0.01. The function $f(x) = x^3$ is evaluated at $x = x_A$ instead of $x = x_T$. Estimate the resulting absolute error.

$f(x_A)$

$$|f(x_T) - f(x_A)| \le K$$

Given $|x_T - x_A| \le 0.01 \Rightarrow -0.01 \le x_T - x_A \le 0.01$

$\Rightarrow 2.49 \le x_T \le 2.51$

Mean-value theorem

$\Rightarrow |f(x_T) - f(x_A)| = |f'(\xi)| \cdot |x_T - x_A|$, ξ lies between x_T & x_A

$\le |3\xi^2| \times 0.01$

$\le |3 \times (2.51)^2| \times 0.01$

≈ 0.189003

S. Baskar (IIT-Bombay) NPTEL Course Tutorial Session

Finally let us solve this problem where again we are not given what is the true value x_T, we are given the approximate value of x_T as 2.5 with an absolute error of at most 0.01. That is what is given to us. Now we want to evaluate the function $f(x) = x^3$. Actually we want to evaluate it at x_T, but unfortunately we have only x_A. Therefore we are evaluating it at $f(x_A)$.

Now what is the absolute error involved in $f(x_A)$ when compared to $f(x_T)$ is the question. So, we have to estimate $f(x_T)$, this is what we want to actually find but we actually found x_A because we are only given the approximate value. Now we want to find an estimate for this. What is mean by estimate of anything? You either have to find the upper bound. That is, you have to find some fixed number say K such that this is less than equal to K or it may also be that you have to find a lower bound say k.

Such that i is less than equal to this number, but here we will only find the upper bound. What is given to us? Given condition is $|x_T - x_{A|} \le 0.01$. That implies that $-0.01 \le |x_T - x_A| \le 0.01$. We know x_A, so just substitute that you will have $2.45 \le x_T \le 2.51$. Therefore, we have an estimate for x_T from here as lower bound as well as the upper bound.

Now let us use the mean value theorem. That gives us $|f(x_T) - f(x_A)|$ which we want to find can be written as $|f'(\xi)||x_T - x_A|$. What is ξ? ξ lies between x_T and x_A. Now what is this? This is nothing but $3\xi^2$ and this is 0.01, well if I have to substitute 0.01 for this, I should have less than or equal to. And again we know that ξ lies between x_T and x_A and we already saw that x_T lies between these two numbers.

The maximum that it can take is 2.51. Therefore, we can roughly estimate ξ to be less than or equal to $|3 \times (2.51)^2| \times 0.01$ and that is approximately 0.189003. So, this is an approximate estimate for the error involved in evaluating the function $f(x) = x^3$ at the point x_A when compared to $f(x_T)$. So, these are some of the important and interesting problems from our classes covered in week 1. With this we will finish our tutorial session 1. Thanks for your attention.