

Data Analysis and Decision Making -I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 51
Loss function

Welcome back my dear friends a very good morning good afternoon good evening to all of you. And you know this is the DADM - I which is Data Analysis and Decision Making - I course under NPTEL MOOC. And the total duration for the course is 30 hours which is 60 lectures each lecture being from half an hour. And we are starting the 51st lecture which is that means, another 2 week is left we have completed 10 weeks. And each week as you know there are 5 lectures each being for half an hour duration. And each week after the first set of 5 lectures then again 10, then again 15 you do take one assignments related to the week of classes set of classes which has been done. And obviously, they would be one end sem or end final examination for as usual for this course. And my name is Raghu Nandan Sengupta from the IME department IIT Kanpur. Now we were discussing utility theory decisions.

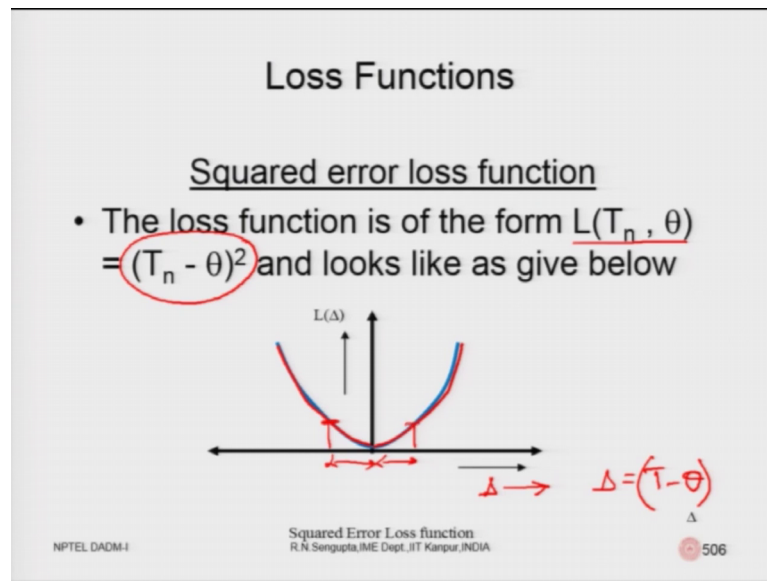
And I did discuss that considering the safety first principle there are three norms. This was the last part of the 50th lecture. Where you we they are trying to basically find out the minimum area between R P and R L. RL is some fixed value which you have it can be risk free interest rate, it can be some returns which you have put for your decision or whatever the fixed criterias. And RP I am using as a so called random variable denoting the so called some of the decisions which you are going to take your investing money taking a decision to for any engineering design or any mechanical design whatever it is your decision is based on the fact that it will be less than some cost all within the stipulated error or it would be within the stipulated deadline whatever it is. Now we will consider and then obviously, in the second part of the safety first principle we consider maximizing the expected value of the total return.

And in the other case you will try to basically push forward the RL the stipulated value which is there I am sorry ill just close it which is there such that the overall distribution moves on to the right. Now then we considered that if considering different RL values would be there how you can rotate the distribution or you can shift parallely the. So,

called best fit line in order to find out the R P star which will give you the. So, called best set of returns or the best set of expected value or minimum set of variances whatever is you are looking at.

So, today in the 51st lecture will start something to do with the loss functions and their implications and then go into the discussions of that how loss functions can be used in multivariate analysis.

(Refer Slide Time: 03:21)



Now generally when we loss of talk about loss functions we talk about some sort of penalty.

So, whenever I am talking about the penalty you penalty need not be cost it can be penalty over and up of the stipulated value which you want to find out over and up of the mean value which you want to find out it can be over and up of the median value which you want to find out over and up of the expected value or loss you want to find out they can be different connotations when I am talking about the loss with respect to a stipulated value.

So, generally will consider that for any distribution whether it is univariate whether it is multivariate we have theta which is the parameter from the distribution say for example, for the normal distribution it will be mean and variance. So, that is there are two parameters. For the exponential considering a value is not there would be one parameter.

Consider now you have basically the multiple linear regression or you have the factor analysis method or trinomial distribution, multinomial distribution, (Refer Time: 04:26) distribution whatever it is.

You will have more than one parameters to estimate and that set of parameters we do not if it is one it is a scalar theta. If it is a vector it is a vector theta. Will also consider that given the population whether univariate or multivariate. We pick up a set of observations. As a mentioned that we pick up small n number of observations based on a small number of n number of observation we try to estimate the parameter.

And if we remember we have discussed that finding on the parameters basically intense stood important criteria which should be fulfilled which is basically unbiasedness and consistency.

So, we will come to that consistency and unbiasedness later on. So, considering that we pick up set of observations. We consider the metric from the sample is given by T which is a made to a statistic with a suffix n where n basically denotes the sample size. Later on I will vomit the n in order to make it very general. So, what we need to find out is basically that if T_n is able to estimate theta at its exact value or a basically over estimates or a underestimates would be given by the fact that there is some loss. In the squared error loss we basically consider the difference between the estimated value of theta and theta is basically given $T_n - \theta$ or $T_n - \theta$ whole square.

So, as that further more or less the estimated value is with respect to the parameter more it will be penalized, but the problem is that. If you remember I did mention very fleetingly that when you are doing the multiple linear regression there also the the issue was basically you take the error, square the error sum it up and then differentiate the errors sum of the errors with respect to the parameters, which are basically beta not or alpha whatever it is.

Then you basically differentiate with beta 1 beta 2 beta 3 put each of them is equal to 0 find out those estimated values of alpha by alpha hat beta by beta hat. And that basically proceed to find out that how good or bad the estimation or the forecasting values of y 's are y is basically the predicted value \hat{y} would be the predictive value for y which is basically the dependent variable which you want to study using the independent variables x_1 to x_P . P means the number of such random variables which are the a

separate random variables. So, when we are talking about the squared error if you notice down here.

So, this is basically the loss function which we denote this is a functional form and we take the squared errors as in this form which means that if I am measuring θ along this direction and L of θ λ sorry L of λ . So, θ is basically t minus θ this λ which is the difference in the actual value or the estimated value or the estimated value actual value whichever you denote. And is a squared error means on the positive side you also increases quadratically on the negative side also it increases quadratically.

So, if I go some distance on to the right and some distance on to the left the value of the loss is basically equally penalized. So, that is why it is known as squared error loss and it is equally penalized loss function. Now you may be interested why it is so, why it is consider this and what advantage this type of loss function gave. So, if you remember the squared error loss which is there T_n minus θ whole square. This is basically if you look at it carefully the function is basically the random variable T_n , T_n is a random variable remember because as I keep changing T and sample size T_n also changes the statistic. And the θ is the expected value for there for this T_n based on the unbiased property which you already discussed.

So, if I want to square it and then try to minimize it that basically gives me the same picture back which is basically trying to find out the minimum of the variance. So, that is why squared error loss is very important to study in the sense it gives us very good results.

(Refer Slide Time: 08:51)

Loss Functions (contd..)

- Most widely used loss function and is used in estimation problems when unbiased estimators of θ are considered, since the risk, $R(T_n, \theta) (= E[L(T_n, \theta)] = E[(T_n - \theta)^2])$, is the *mean square error* (MSE) of T_n , which reduces to the variance of T_n subject to unbiasedness
- The corresponding optimal estimator, if it exists, is called the *minimum variance unbiased (MVU) estimator*
- Another reason for the popularity of SEL is due to its relationship to the classical least square theory. Also, for most analyses SEL makes calculation relatively straight forward.

NPTEL DADM-I R.N.Sengupta, IIT Kanpur, INDIA 507

Now continuing the squared error loss most widely used lost function and it is used in estimation problems where unbiased estimation of theta which theta is basically the parameter which you want to estimate. It can be a scalar it can be a vector. Since the risk now the risk is basically the functional form the expected value of the loss because loss is that in the T_n is a random variable, T_n minus theta is also is basically what you are trying to find out is the difference between the random variable and it is expected value. If I find out the expected value of this difference it gives me the expected value of the variance of the T_n which is the variance of the parameters estimate which you want to find out and we want to basically minimize the risk.

So, if in case if it is for the squared error loss if you using the normal distribution with the parameter is new. And the statistic is basically sample mean then the sample mean minus new whole square the expected value of that would basically a square of the expected value of that would basically in the variance of that T_n which is sigma square by n for the normal case.

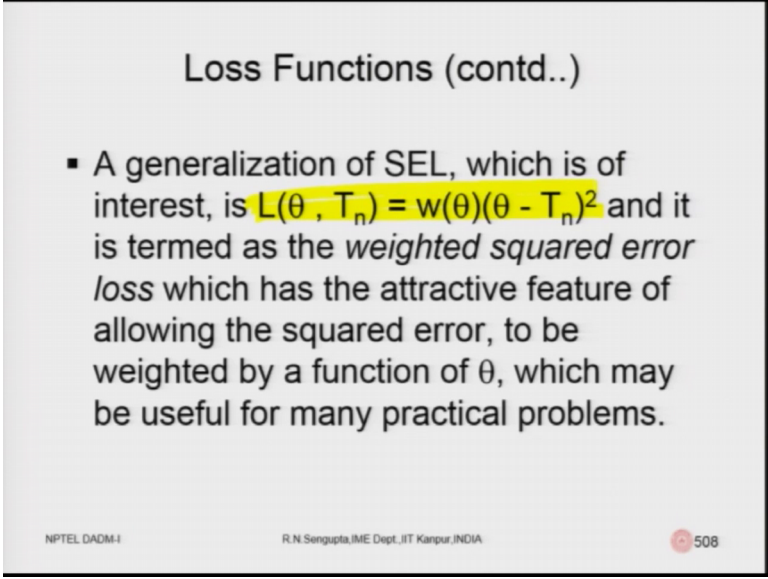
So, the risk which is $R_n(\theta)$ is given by except expected value of the loss which is the expected value of the squared error loss. Is the mean square error of T_n which reduces to the variance as I was telling of T_n subject to unbiasedness.

So, if it is unbiased which means expected value of T_n is theta, then difference between T_n minus is accepted expected value which is T_n minus theta whole square expected

value of that again basically becomes the variance. So, this is very important to note what I have been talking about why the squared error loss is important. Now the second bullet point states the corresponding optimum estimated if it exists for theta which is T_n is called the minimum variance unbiased estimator because the variance is also minimum and is also unbiased because the expected value of T_n theta which is UMVU VUE as per the concept of basic statistics.

Uniform minimum variance unbiased estimator another reason for the popularity of squared error loss is due to the relationship to the classical D square theory. Also for most analysis squared error loss makes calculation relative easy as I said and the straight forward and theoretically very nice. So, we get good results even though practically may not it was very nice result, but they are (Refer Time: 11:40) able to give us very good results based on which you can find out lot of things based for the estimation problem.

(Refer Slide Time: 11:47)



Loss Functions (contd..)

- A generalization of SEL, which is of interest, is $L(\theta, T_n) = w(\theta)(\theta - T_n)^2$ and it is termed as the *weighted squared error loss* which has the attractive feature of allowing the squared error, to be weighted by a function of θ , which may be useful for many practical problems.

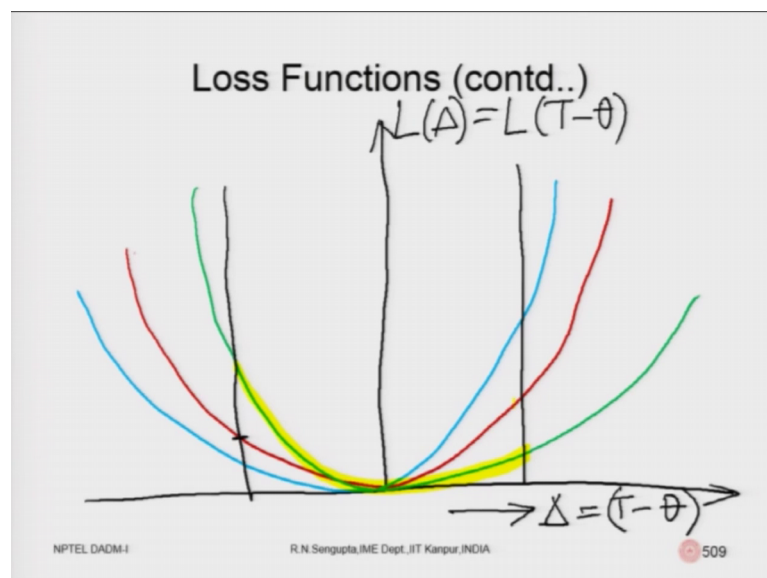
NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 508

So, to continue loss functions, [vocalised-noise] discussions a generalization of the squared error loss which is of interest to many of us if, which is used for our studies is known as the weighted squared error loss. Now if you remember for the squared error loss what we are doing is that we are trying to find out the difference with the estimated value and the parameter value squaring it up and submitting it up. Summing it up in case we need to find out and then find out the expected value which is the risk.

Now in case if we give weights. So, weights can be given both on the positive size and the negative size that means T_n minus θ is positive or T_n minus θ is negative based on that we can give weights.

So, this is known as the weighted squared error loss function, which has attractive features of allowing the squared errors to be weighted by a function depending on what we think is the importance of the loss function to be studied. So, let me make one slide and try to basically discuss how it would look like. This is where we what I said is the weighted squared error loss function.

(Refer Slide Time: 12:57)



So, if you consider squared errors weighted one it would look like this. So, this is x axis the difference between the estimate and the parameter.

So, if I will try to draw all the three graphs, even though it may not be possible let me try. Let me first reduce the axis as black so it will be easier for me. So, it is squared error basically took (Refer Time: 13:38) like this the squared error. If you want to give weights both on the little left or I have to use the green colour maybe it like basically bowl sort of thing which will basically be tilt it.

And this to a green was tilted on the right blue one will be tilted. So, for the same level of if you consider this line, if you consider this line and I A M A writing delta which is equal to. So, I am not using the suffix of T. So, it will be n. This is basically n of delta.

So, the height which I have let me use a highlighter. So, the height which I have corresponding to the squared error which is equal I will be used, try to use whatever my drawing skills for. So, the height of this loss function here left and the right should be equal. If I use the p use again the yellow colour, but let me erase it. If I use the blue one so, left hand side with respect to right hand side, the errors on the right hand side which is more penalized you can erase it. And I go to the green one.

So the left hand side which under estimation would be more penalized than on over estimation. So, we are able to utilize the squared error using the weights. So, in the case when the weights are more or less you will basically half the squared errors being tilted more on to the left or to the right.

(Refer Slide Time: 16:35)

Loss Functions (contd..)

- If $\theta = (\theta_1, \dots, \theta_n)$ is a vector estimated by $T = (T_1, \dots, T_n)$ and Q is $(n \times n)$ positive definite matrix, then, $L(\theta, T) = (\theta - T)'Q(\theta - T)$ is called a quadratic loss function

NPTEL DADM-I
R.N Sengupta, IIT Kanpur, INDIA
510

Now consider the loss function which is again quadratic or weighted quadratic, but based on the fact now it is basically a multiple multivariate problem.

So, consider that there are random of variables random variables basically formulates and gives us some loss function. And the random variables of the set of random variables are basically I will I am considering n here for which is basically p or k. So, if we should not be confused about that and consider theta is basically a vector consists of theta 1 theta 2 till theta n or theta p whatever it is. And the theta is basically bold in this which is a vector. Now it is corresponding the value with from a sample which is a sample statistic T which is also bold is also vector consists of the values T 1 to T n.

So, here n does not imply the sample size basically technically it is PRQ it should have been P and Q whatever. And let us consider that q is basically positive definite matrix of size n into p here n is the sample size p is the number of random variables. So, it is basically a positive definite matrix and what do we need to find out is basically a loss function which is a squared error in terms of the multivariate cases.

So, what we are doing is that we agree if we give weights Q 's are basically the weights are in all the parameter values of θ are such that the weight we give to the squares of each individual differences between the statistic and the corresponding parameters if they are equal then obviously, the q matrix the cell value should be equal in all respect. In case if q matrix the cell values are unequal.

So, they will give unequal weights to the difference between the square of T minus θ hence will basically have the unequal weighted squared errors, but in this case we will find out the square errors are corresponding to a multivariate case, we can have problems in order to solve them.

(Refer Slide Time: 19:05)

Loss Functions (contd..)

Linear loss function

- When the utility function is approximately linear the loss function will tend to be linear in nature which is of the form,
 - $L(\theta - T_n) = K_1(\theta - T_n)$ if $(\theta - T_n) \geq 0$
 $= K_2(T_n - \theta)$ if $(\theta - T_n) < 0$

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 511

Another loss function which is quite popular is known as the linear loss function and then as the words mentions it is linear in nature not quadratic as we have considered. When the utility function is approximately linear or the loss function is linear they tend to be linear in nature and it is of this form.

So, what we see is that the loss function is basically the difference between the parameter value and the estimated value; parameter value is θ estimated value corresponding to θ is T_n suffix, n is the sample size or T_n delta is basically delta is the function which basically denotes the difference between T_n and θ .

And you give weights to both overestimation and underestimation, in case if $\theta - T_n$ or $T_n - \theta$ whichever way you look is greater than 0 that is on the right hand side. Then you will give some weights which is given as K_1 and if it is on the left hand side such that $\theta - T_n$ is less than 0 you give a weight of K_2 . Now in case K_1 and K_2 are equal it is one you will basically have a 45 degree line both phase.

So, this is what I will basically demonstrate.

(Refer Slide Time: 20:21)

Loss Functions (contd..)

- The constants K_1 and K_2 are to be chosen to reflect the relative importance of overestimation and underestimation. In general, these constants are different, but when they are equal, the equivalent loss function is of the form $L(\theta, T_n) = |\theta - T_n|$, which is called *absolute error loss*

NPTEL DADM-I R.N Sengupta, IIT Kanpur, INDIA 512

So, here n come to that picture very soon; the constant K_1 and K_2 initially can be considered to be reflect the relative importance we give on the weightages corresponding to over estimation under estimation. In general this constants are different, but when they are equal the equivalent loss functions would be given by $T_n - \theta$ mod of that or $\theta - T_n$ and mod of that and we can solve the problems accordingly.

So, what we are doing is that if you remember we are done in the concept of interval estimation and that $\bar{x}_n - \mu$ we found out that in the interval estimation what was the total probability that the parameter value has been estimated by T_n and

what is the probability that T_n would live within that range that depending on the value of α which you have level of confidence.

So, it can be higher or lower and it can be both for the interval estimation later on we basically converted into a hypothesis testing for the left hand side hypothesis right hand side hypothesis and not equal to hypothesis. Now in this case if K_1 and K_2 are same. Then it is equal weightages in case it is not same they will be a unequal weightages.

(Refer Slide Time: 21:39)

Loss Functions (contd..)

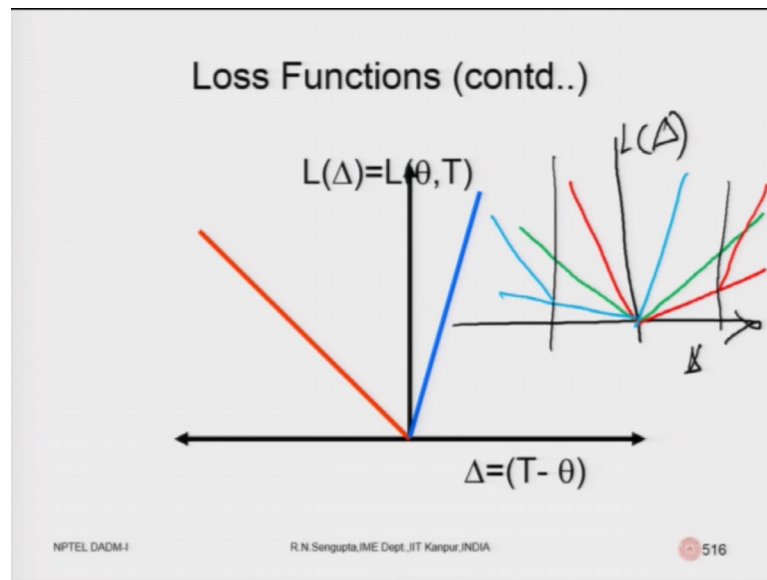
- The optimal estimator for this absolute error loss function, if it exists, is called the *minimum mean absolute error estimator*
- If K_1 and K_2 are themselves functions of θ , then the loss function is termed as the *weighted linear loss function*.

NPTEL DADM-I R.N. Sengupta, IIT Kanpur, INDIA 513

The optimum estimate for this absolute error loss function which is $T_n - \theta$ that mod of that; if it exists is called the minimum mean absolute error estimation and we can do the calculations accordingly. If K_1 and K_2 are themselves functions of θ it is possible that and the errors which is basically $T_n - \theta$ are being weighted, but those weights themselves are depend on θ .

So, which can happen that if you are trying to estimate very higher values of θ or lower values of θ the weights as you go more on to the right or more on to the left even if it is linear you give more or less weights depending on the importance issue on which you want to put. So, if K_1 and K_2 are themselves functions of θ then the loss function is termed as the weighted linear loss function.

(Refer Slide Time: 22:29)



So, the diagrams which you see here is basically along the x axis I measure delta which is the difference between T minus theta I am not putting any suffix for T and along the y axis I measure L delta which is L of the functional form of the difference between T minus theta. And what is very important to note down is that in case the diagram if it is has been made correctly if I am not mistaken. The values of K_1 K_2 if there are 1.

So, obviously, it would mean that the angle between the overestimation which is shown here in blue colour underestimation which is shown in red colour they will be equally inclined both on the first quadrant and the second quadrant; that means, on the positive signs and negative signs. In case if say for example, K_1 and K_2 are different then obviously, it would mean that K_1 at the blue line can either beta on left or right or and the red line can be either turned left or right depending on the level of importance which you want to give for over estimation and underestimation.

So, continuing with the discussion in the loss function which you see now the blue line continues in the same case as it is which is K_1 is say for example, 1 and K_2 has been penalized K_2 is basically the weight which I am considering for the underestimation problem K_2 has been increased such that for any small deviation in the negative direction the overall loss is much higher. So, in this case underestimation is more problematic or more penalized than an overestimation.

So, the picture can definitely be turned opposite. In the sense that underestimation can be equally penalized with the sense that K_1 or K_2 which you are we have considered the weights for overestimation and underestimation will consider K_2 to be 1. And in this case as you can see in the diagram overestimation is heavily penalized which means that you will give weights accordingly. So obviously, K_1 would change either on the higher side or the lower side depending on what you think is important for the weightages which you want to give for overestimation and underestimation.

Now it may so, happen that in the case and these are theoretical and practically problems which can be formulated in case say for example, you have this over estimation underestimation being equally penalize would basically have I will use the same coloured for. So, this is overestimation, this in underestimation overestimation both equally penalized. Which I have drawn separately, but I am just combining them underestimation being more penalized, overestimation being less penalized. And then we will consider overestimation being more penalized underestimation being less penalized, but it will also me may mean.

So, this actually gives in the diagram combined I should also write this is basically Δ this is L of Δ . So, it would also mean that if they are independent on θ or domain. So, in case θ also comes into the picture weightages are to be given.

So, it may happen that for some value of θ here or here both for the positive and negative sign. This overestimation being less penalized suddenly the picture may change where for a certain values of θ more than the stipulated values which you heads had. Then in that case that overestimation can be more penalized. Similarly it may happen here for the for the blue one. It is possible that then can me more penalized. So, this will depend on what type of functional form which you have.

(Refer Slide Time: 26:53)

Loss Functions (contd..)

0-1 Loss function

- This is of the form
$$L(\theta, T_n) = 1, \text{ if } |T_n - \theta| > \varepsilon,$$
$$= 0, \text{ otherwise}$$
for $0 < \varepsilon < 1$
- Here risk is $P(|T_n - \theta| > \varepsilon)$
- This refers to the large deviation probability, and the optimality of an estimator T_n is interpreted in terms of minimization of this probability or in terms of the fastest rate of decline (with n) of this probability

NPTEL DADM-I R.N.Sengupta,IIT Kanpur,INDIA 517

Now another loss function which will consider is the 0 1 loss function. So, this is the form that if the difference between the estimated value and the predicted value is within a range. So, this range this epsilon gives us some indication about the alpha value for interval estimation which you have just discussed now. And we have been discussing in the interval estimation problem. If it is equal to between that it is more than that epsilon you will give a weightages 1 otherwise you will give it a weightage of 0.

Here the risk would be given with the probability of the difference between T minus θ is being greater than epsilon. This reverse refers to the large deviation probability and the optimality of an estimate T suffix n whatever it is interpreted in terms of minimization of this probability or in terms of the fastest rate of decline with n of this probability.

So, as n increases the interval shrinks in the sense the estimated value of T_n gets closer to θ both it can be random, it can be on the left hand side on the right hand side. So, basically it approaches the difference approaches 0 which means T_n basically approaches θ . So, with this I will close this 51st lecture and continue more discussion about loss function in the 50 second and 53rd and. So, on and so forth have a nice day.

Thank you very much.