STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Engineering

Indian Institute of Science, Bangalore

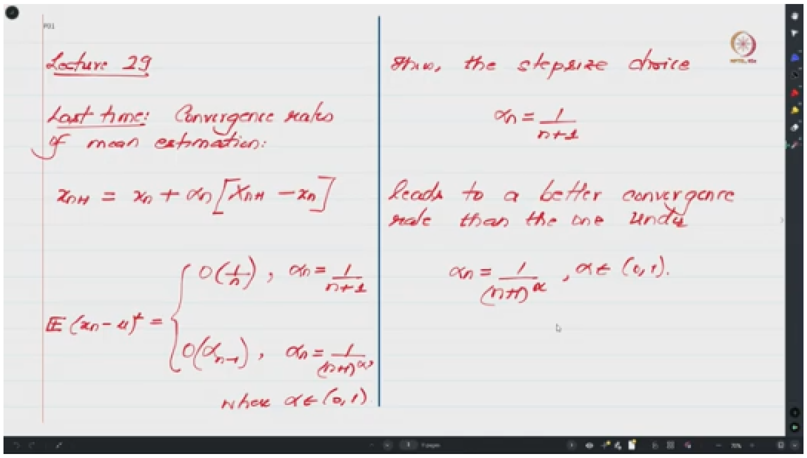
Week 7

Lecture 29

Lower Bounds and the Minimax Risk in Estimation

Hello and Namaste everyone. Welcome to lecture 29 of this NPTEL course on Stochastic Approximation. So, you must have seen over the past two lectures, we have looked at the convergence rates of stochastic approximation algorithms. In particular, we have focused on this specific category of linear stochastic approximation, and more specifically, we have looked at this mean estimation problem. So, we are focusing on this very simple algorithm so that we can understand the effect of step size on the convergence rates of such algorithms, right? And later on, I will briefly say how this knowledge actually helps us conjecture some convergence rates for more general stochastic approximation algorithms.

And in the first two lectures, at a very broad level, we discussed convergence rates of this mean estimation algorithm for two step size choices: one is of the form 1 over n plus 1, which decays very fast, and the other is of the form 1 over n plus 1 to the power alpha, where alpha is strictly less than 1, which means that the step size gradually decays. Then we said or compared the convergence rates in either of these two cases and concluded that the convergence rate with the 1 over n plus 1 is the best compared to the two. So, in today's class, what we will do is ask ourselves: can we do better than this 1 over n convergence rate that we obtained by choosing the step size of the form 1 over n plus 1? With this idea in place, let us do a formal review of what we have done so far. So, this week, we have been looking at the convergence rates of the mean estimation algorithm.



Lecture 29

Last time: Convergence rates of mean estimation:

$$x_{n+1} = x_n + \alpha_n [X_{n+1} - x_n]$$

$$\mathbb{E}(x_n - \mu)^2 = \begin{cases} O\left(\frac{1}{n}\right), & \alpha_n = \frac{1}{n+1} \\ O(\alpha_{n-1}), & \alpha_n = \frac{1}{(n+1)^\alpha} \end{cases}$$

where $\alpha \in (0,1)$

Show, the stepsize choice

$$\alpha_n = \frac{1}{n+1}$$

leads to a better convergence rate than the one under

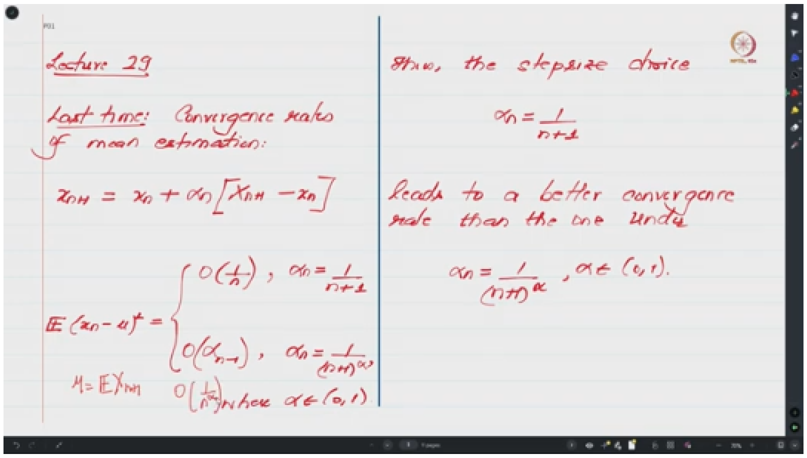
$$\alpha_n = \frac{1}{(n+1)^\alpha}, \alpha \in (0,1).$$

By that, I mean an update rule of the form Xn plus 1 equals Xn plus alpha n times capital Xn plus 1 minus little xn. And we showed that the mean squared error, which is defined in the following way—that is, the expected value of little xn minus mu the whole square, where mu is the expected value of your Xn plus 1 random variable. And we have assumed that all random variables are identically distributed, and hence mu is the mean for Xn for every n. And we showed that the error in this sense is O when the step size is of the form 1 over n plus 1, and it is alpha n minus 1 when alpha n is of the form 1 over n plus 1 to the power alpha.

$$x_{n+1} = x_n + \alpha_n \left[X_{n+1} - x_n\right]$$

$$E\left(x_n - \mu\right)^2 = \{0\left(\frac{1}{n}\right), \qquad \alpha_n = \frac{1}{n+1}\, 0\left(\alpha_{n-1}\right), \quad \alpha_n = \frac{1}{(n+1)\alpha}$$

So, which implies that this convergence rate is of the form 1 over n to the power alpha. So, because alpha is strictly less than 1, one can conclude that this convergence rate is better than this convergence rate.



Lecture 29

Last time: Convergence rates of mean estimation:

$$x_{n+1} = x_n + \alpha_n [X_{n+1} - x_n]$$

$$\mathbb{E}(x_n - \mu)^2 = \begin{cases} O(\frac{1}{n}), & \alpha_n = \frac{1}{n+1} \\ O(\alpha_{n-1}), & \alpha_n = \frac{1}{(n+1)^\alpha} \end{cases}$$

$\mu = \mathbb{E}X_{n+1}$ $O(\frac{1}{n^\alpha})$ where $\alpha \in (0,1)$

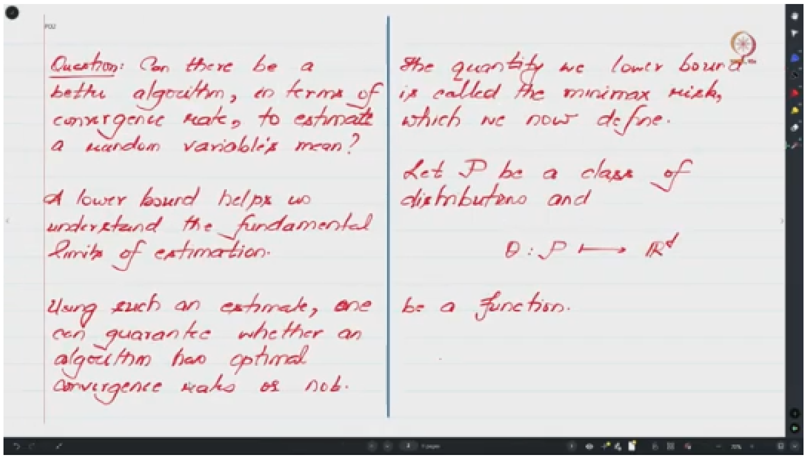
Thus, the stepsize choice

$$\alpha_n = \frac{1}{n+1}$$

leads to a better convergence rate than the one under

$$\alpha_n = \frac{1}{(n+1)^\alpha}, \alpha \in (0,1).$$

So, one can then ask, you know, the simple averaging algorithm is obtained when you take Alpha n equals 1 over n plus 1, and by simple averaging, I mean that, you know, when alpha n is 1 over n plus 1, your little xn is like x1 plus dot dot dot plus capital Xn by n, right? So, this is like a simple averaging algorithm. Right, and our, you know, discussions in the past over the past two lectures show that such a simple averaging algorithm actually achieves this 1 over n convergence rate, right. So, then one can ask, is there a better algorithm in terms of convergence rate to estimate a random variable's mean, right? And to answer such a question, a lower bound helps.



Question: Can there be a better algorithm, in terms of convergence rate, to estimate a random variable's mean?

A lower bound helps us understand the fundamental limits of estimation.

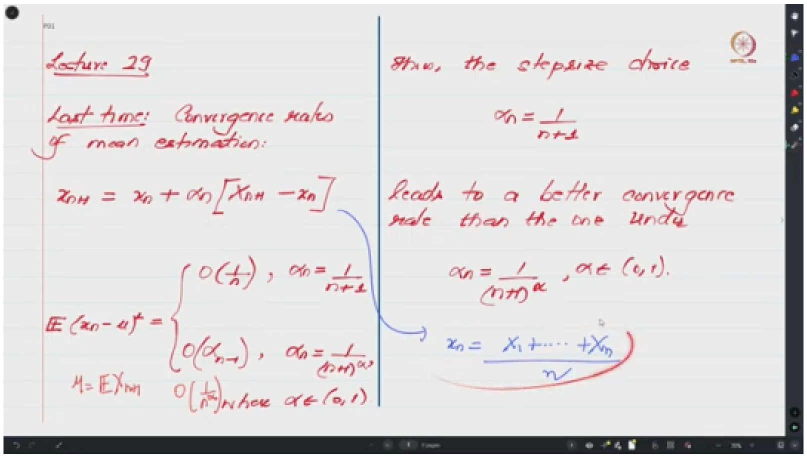
Using such an estimate, one can guarantee whether an algorithm has optimal convergence rate or not.

The quantity we lower bound is called the minimax risk, which we now define.

Let $\mathcal{P}$ be a class of distributions and

$$\theta : \mathcal{P} \longmapsto \mathbb{R}^d$$

be a function.



Lecture 29

Last time: Convergence rates of mean estimation:

$$x_{n+1} = x_n + \alpha_n [X_{n+1} - x_n]$$

$$\mathbb{E}(x_n - \mu)^2 = \begin{cases} O(\frac{1}{n}), & \alpha_n = \frac{1}{n+1} \\ O(\alpha_{n-1}), & \alpha_n = \frac{1}{(n+1)^\alpha} \end{cases}$$

$\mu = \mathbb{E}X_{n+1}$ $\quad O(\frac{1}{n^\alpha})$ where $\alpha \in (0,1)$

Show, the stepsize choice

$$\alpha_n = \frac{1}{n+1}$$

leads to a better convergence rate than the one under

$$\alpha_n = \frac{1}{(n+1)^\alpha}, \alpha \in (0,1).$$

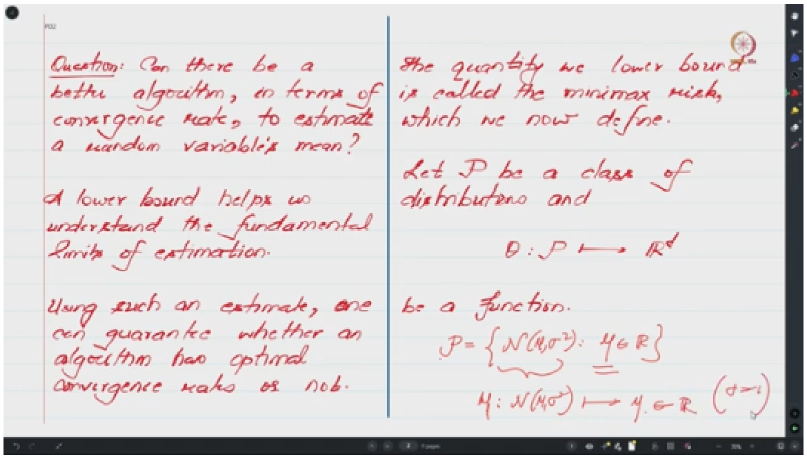
$$x_n = \frac{X_1 + \cdots + X_n}{n}$$

In particular, a lower bound helps us understand the fundamental limits of estimation, and using such an estimate, one can guarantee whether a particular algorithm has an optimal convergence rate or not. Right, and the quantity we shall, you know, obtain a lower bound for today is what we will refer to as the minimax risk. I am going to define that now, and I am also going to motivate how one comes up with this lower bound problem and how one can use that lower bound problem to, you know, say eventually that a particular algorithm has the optimal convergence rate or not. In particular, in the context of estimating a random variable's mean, we would like to know whether the simple average, which is basically, you know, taking the sum of all random variables and dividing by n, which refers to the number of samples, is this going to achieve the best possible convergence rate or not. So, towards that, we are going to define this minimax risk quantity, and to define this minimax risk notion,

let us first begin with defining calligraphic P to be the class of distributions and let be a function which, you know, whose domain is this calligraphic P, which is the class of distributions, and the output, let us say, is some RD for some D, right? So, the way to interpret this is theta takes as input one distribution and spits out a vector, right? Like, for example, your calligraphic P is could be, you know, all Gaussian distributions, right, whose variance is fixed but, let us say, mean, you know, is some number, some real number, right? For different choices of mu, we have different distributions.

$$\theta: p \mapsto R^d$$

And one can imagine mu, which is like a special instance of your theta function, which takes as input one of these distributions and, let us say, maps it to mu. Is this okay? So, mu here will be an element in R, and in this case, D is 1. So, you can think of theta as a function which takes as input some distribution and spits out some information regarding this distribution. In particular, this mu function takes the distribution as input and spits out, let us say, the mean of this distribution. Now, in this context of minimax risk, the goal is to estimate theta of P, where P is a specific element in this class of distributions. However, the challenge is that instead of being given access to P, what we are given access to is observations that are drawn independently from this unknown distribution P. So, let me elaborate again. The goal is, if there is some P, right, in this class of distributions, our goal is to find theta of P, right, but you cannot find theta of P using the knowledge of P.



Question: Can there be a better algorithm, in terms of convergence rate, to estimate a random variable's mean?

A lower bound helps us understand the fundamental limits of estimation.

Using such an estimate, one can guarantee whether an algorithm has optimal convergence rate or not.

The quantity we lower bound is called the minimax risk, which we now define.

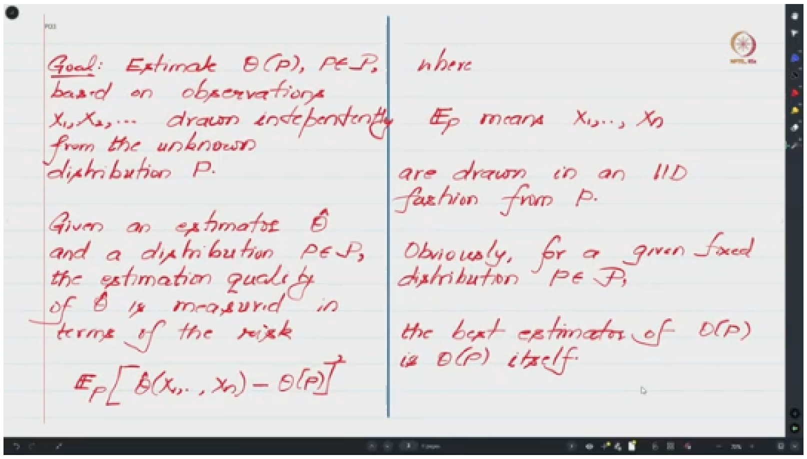
Let $\mathcal{P}$ be a class of distributions and

$$\theta : \mathcal{P} \longmapsto \mathbb{R}^d$$

be a function.

$$\mathcal{P} = \{\underbrace{N(\mu,\sigma^2)}: \underline{\underline{\mu}} \in \mathbb{R}\}$$

$$\mu : N(\mu,\sigma^2) \longmapsto \mu \in \mathbb{R} \quad (d=1)$$



Goal: Estimate $\theta(P)$, $P \in \mathcal{P}$, based on observations $X_1, X_2, \dots$ drawn independently from the unknown distribution P.

Given an estimator $\hat{\theta}$ and a distribution $P \in \mathcal{P}$, the estimation quality of $\hat{\theta}$ is measured in terms of the risk

$$\mathbb{E}_P[\hat{\theta}(X_1, \dots, X_n) - \theta(P)]^2$$

where

$\mathbb{E}_P$ means $X_1, \dots, X_n$ are drawn in an IID fashion from P.

Obviously, for a given fixed distribution $P \in \mathcal{P}$, the best estimator of $\theta(P)$ is $\theta(P)$ itself.

Instead, what you have access to are these random observations that are drawn from this unknown distribution P, and using these observations, the goal is to estimate theta of P. So, one way one can talk about the risk that is involved with regards to an estimator can be defined in the following sense. Let me elaborate.

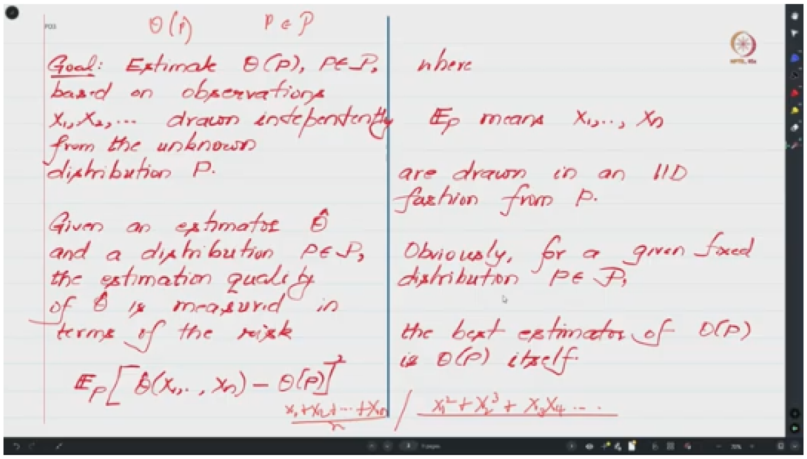
Now, given these observations, you may come up with some rule. For example, in our previous class, we had come up with this simple averaging rule that took all the samples, took their simple average, and proposed that as an estimate of the mean. But you could work with other estimators as well. So, what this risk concept is over here is that suppose you have some estimator theta hat. The estimator theta hat's job is to take in your observations, that is, X1, X2, X3, and so on, and spit out an estimate of this quantity of interest, which is theta of P.

And the risk is basically the squared error, which is the expected value of theta hat of X1 to Xn minus theta of P, the square of that. Is that okay? So the way to interpret this is, you get to see these, uh, you know, observations of this unknown distribution P, right? And then you come up with some rule to estimate theta of P, right? So, for example, this rule could basically be X1 plus X2 plus dot dot dot plus Xn over n. So this could be one rule, but maybe you do not like this rule.

Tomorrow you may want to, you know, do X1 squared plus X2 cubed plus X3 X4 and so on. So you could come up with some fancy rule like this. Right, and you know, propose that as a way to estimate your quantity of interest, right? So whatever your choice of estimator, we will define the risk associated with this estimator by the expected value of theta hat of X1 to Xn minus theta P squared, right? And the P over here indicates that X1 to Xn has been drawn independently from this unknown distribution P. So, these are your random variables, and theta hat is a function of these random variables; hence, this itself is a random variable, and we are looking at the expectation of the difference between this quantity and this quantity. In particular, we are looking at the square of this difference.

So, now one can ask: what is the best estimator for finding theta of P? So, of course, if there is only one fixed distribution, if there is only one fixed distribution, then, of course, the best estimator of theta of P is theta of P itself. So if there is, let us say, only one

Gaussian distribution with mean mu, then the best estimator for the mean of this distribution is mu itself. So if someone can say mu, well, that is the best estimator.



$\theta(P)$ $P \in \mathcal{P}$

Goal: Estimate $\theta(P)$, $P \in \mathcal{P}$, based on observations $X_1, X_2, \ldots$ drawn independently from the unknown distribution P.

Given an estimator $\hat{\theta}$ and a distribution $P \in \mathcal{P}$, the estimation quality of $\hat{\theta}$ is measured in terms of the risk

$$\mathbb{E}_P\left[\hat{\theta}(X_1, \ldots, X_n) - \theta(P)\right]^2$$

$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$

where

$\mathbb{E}_P$ means $X_1, \ldots, X_n$

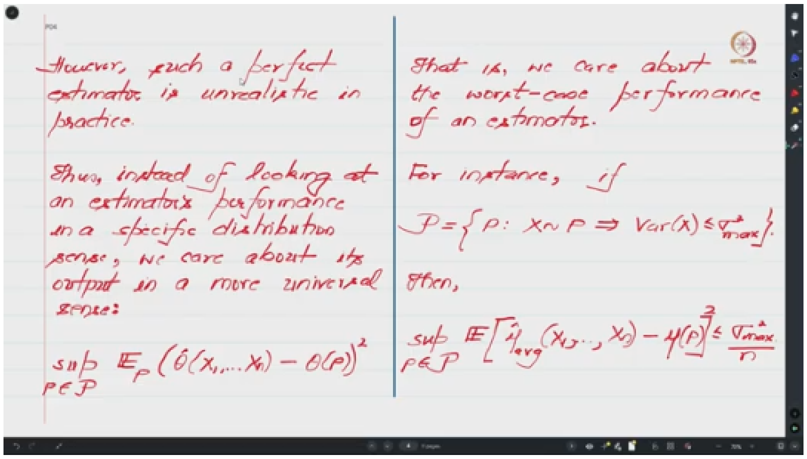
are drawn in an IID fashion from P.

Obviously, for a given fixed distribution $P \in \mathcal{P}$,

the best estimator of $\theta(P)$ is $\theta(P)$ itself

$X_1^2 + X_2^3 + X_3 X_4 \ldots$

However, such estimators are unrealistic in practice. Hence, instead of focusing on one distribution at a time, We evaluate or look at the performance of an estimator in a more universal sense. Instead of looking at it in the context of a specific distribution, we evaluate the power or the potential of an estimator in a universal sense. So, what do I mean by that?



However, such a perfect estimator is unrealistic in practice.

Thus, instead of looking at an estimator's performance in a specific distribution sense, we care about its output in a more universal sense:

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P\left(\hat{\theta}(X_1, \ldots X_n) - \theta(P)\right)^2$$

That is, we care about the worst-case performance of an estimator.

For instance, if

$\mathcal{P} = \{P : X \sim P \Rightarrow Var(X) \leq \sigma^2_{max}\}$.

Then,

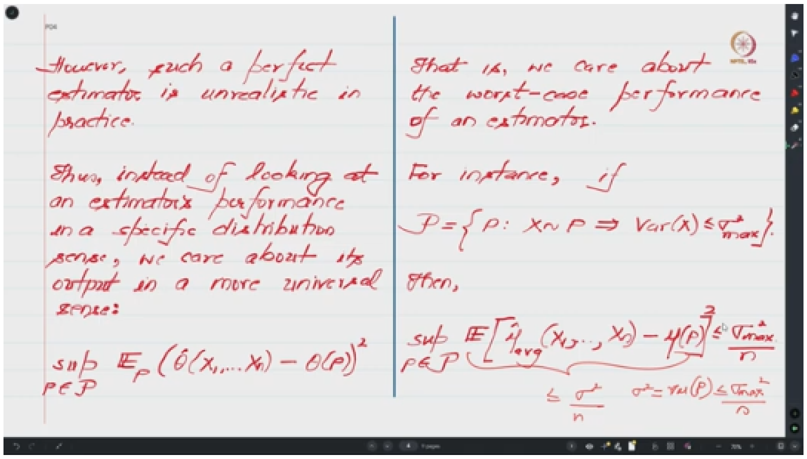
$$\sup_{P \in \mathcal{P}} \mathbb{E}\left[\hat{\mu}_{avg}(X_1, \ldots, X_n) - \mu(P)\right]^2 \leq \frac{\sigma^2_{max}}{n}$$

So, instead of focusing on one distribution at a time, what we ask is, you first give us an estimator, theta hat. So, theta hat is basically a rule which says how to combine these observations X1 to Xn, right? So, theta hat is a rule, and now what we say is, how good is this way of combining? Now, instead of evaluating how good this way of combining is for a specific distribution, we will look at in some sense the worst-case performance or worst-case performance of this estimator. In particular, what we will do is we will look at the supremum of this error or this risk for every P in calligraphic P. Let me again repeat. So, theta hat is one fixed rule, that is, one fixed estimator. For this estimator, we are going to look at the worst-case error, and the worst here is over the choice of this input or unknown distribution P in this class of distributions calligraphic P. So, that is what I have mentioned over here, that is, you know, to test or check the power of an estimator, we care about the worst-case performance of this estimator.

So, for example, let us say our distributions or the class of distributions includes all those distributions whose variance—that is, for any random variable X that has the distribution P—its variance is upper bounded by sigma max squared. Let me repeat: this is the collection of all those distributions. Such that the variance associated with those distributions is all upper bounded by sigma max squared. So, this is some quantity that acts like a universal constant, upper bounding the variance of all these distributions. So, now, based on our analysis from our previous two lectures, one can show that if you work

with the simple averaging algorithm—the one that just takes the input observations and takes their sample average, which I denote as mu hat average—then how good is such an algorithm at computing the mean of this unknown distribution using samples? In our previous class, we have shown that this expression for any P is upper bounded by sigma max squared over n. This is something we have proved in our previous two classes, right? And this quantity over here is irrespective of what this P is. This quantity here only depends on this bound sigma max squared. In particular, for a specific distribution, this error is upper bounded by sigma squared over n, where sigma squared is the variance.

of P, right? And we have been told that this is upper bounded by sigma max squared over n, and hence one can conclude that for a fixed P, this expression here is upper bounded by sigma squared over n, which is independent of P. Hence, if you take the supremum of this expression over all P in calligraphic P, this expression will still be upper bounded by sigma squared over n, which means that the worst-case performance of your simple averaging estimator is sigma squared over n, right? So, one can ask: is there a better estimator in this worst-case sense, right? And this is the spirit of this minimax estimator. In particular, the optimal estimator is the one that gives the minimax risk.



However, such a perfect estimator is unrealistic in practice.

Thus, instead of looking at an estimator's performance in a specific distribution sense, we care about its output in a more universal sense:

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\left(\hat{\theta}(X_1,\dots X_n)-\theta(P)\right)^2$$

That is, we care about the worst-case performance of an estimator.

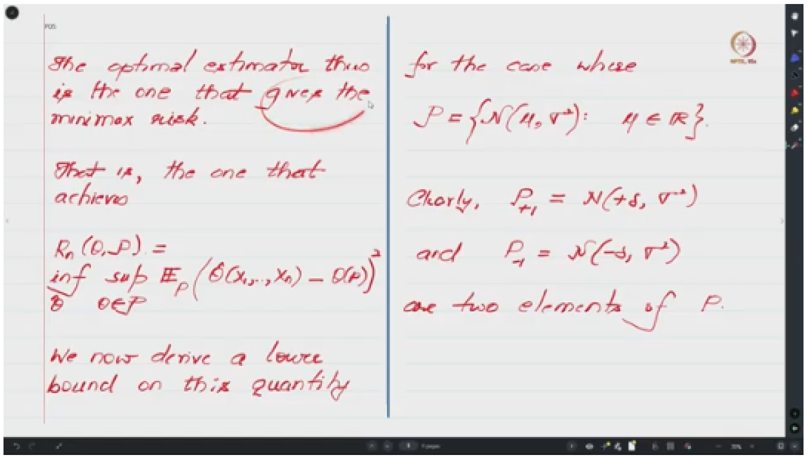
For instance, if

$$\mathcal{P}=\{P: X\sim P \Rightarrow Var(X)\le\sigma_{max}^2\}.$$

Then,

$$\sup_{P\in\mathcal{P}} \mathbb{E}\underbrace{\left[\hat{\mu}_{avg}(X_1,\dots,X_n)-\mu(P)\right]^2}_{\le\frac{\sigma^2}{n}}\le\frac{\sigma_{max}^2}{n}$$

$$\sigma^2=Var(P)\le\frac{\sigma_{max}^2}{n}$$



The optimal estimator thus is the one that gives the minimax risk.

That is, the one that achieves

$$R_n(\theta,\mathcal{P})=\inf_{\hat{\theta}}\sup_{\theta\in\mathcal{P}} \mathbb{E}_P\left(\hat{\theta}(X_1,\dots,X_n)-\theta(P)\right)^2$$

We now derive a lower bound on this quantity

for the case where

$$\mathcal{P}=\{\mathcal{N}(\mu,\sigma^2): \mu\in\mathbb{R}\}.$$

Clearly, $P_{+1}=\mathcal{N}(+\delta,\sigma^2)$

and $P_{-1}=\mathcal{N}(-\delta,\sigma^2)$

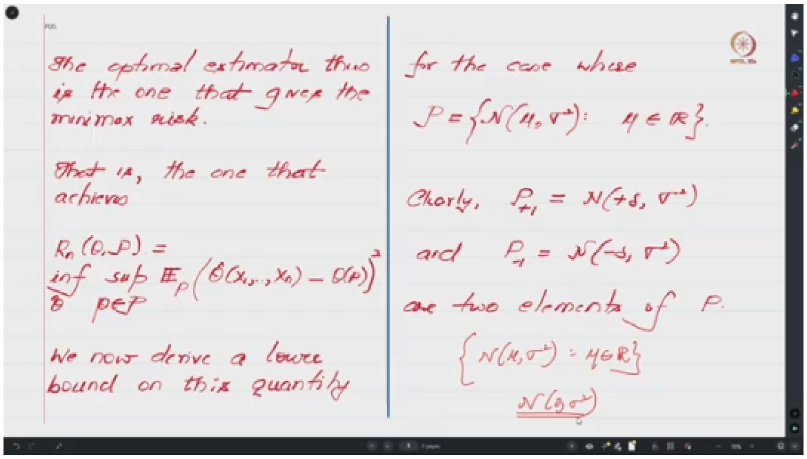
are two elements of $\mathcal{P}$.

What do I mean by that? The optimal estimator is the one that minimizes or achieves the infimum specified over here. So, let us look at this expression a bit more clearly. This is your mean squared error for your choice of estimator. When the input distribution is p, I should actually say this is p over here, not theta, so this should be p. So, when the input distribution is p, this is the mean squared error, and with the supremum, this is the worst-case error for this fixed estimator.

Right, and we would like to ask which estimator minimizes the worst-case error, right? The estimator that achieves the minimax risk is what we will refer to as the optimal estimator. Is that okay? So, this is how we will evaluate which estimator is good. See, again, for a specific problem—if you fix P, this P in calligraphic P—and if you want to estimate theta of P, then, of course, the estimator that always outputs theta of P will always win.

We will not be able to compete against that algorithm unless we also output theta of P, which is typically unknown. So, what we are instead asking is, instead of focusing on one distribution, we will look at the performance of a given estimator on a class of distributions. Is this okay? So, if you have, let us say, two distributions or a family of distributions, let us say, n mu square over r, right? So, you know, in this family, right?

So, of course, one of the distributions is, you know, the Gaussian distribution with mean 0. So, for this distribution, right, the best estimator is 0, right? However, if you always give 0 as the answer, right, then you are not going to do well when your mu is actually not equal to 0. Is this okay? So, in this sense, we want to evaluate the performance of any estimator in a worst-case sense over here.



The optimal estimator thus is the one that gives the minimax risk.

That is, the one that achieves

$R_n(\Theta, \mathcal{P}) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\left(\hat{\theta}(X_1, \ldots, X_n) - \theta(P)\right)^2$

We now derive a lower bound on this quantity

for the case where

$\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$.

Clearly, $P_{+1} = \mathcal{N}(+\delta, \sigma^2)$

and $P_{-1} = \mathcal{N}(-\delta, \sigma^2)$

are two elements of $\mathcal{P}$.

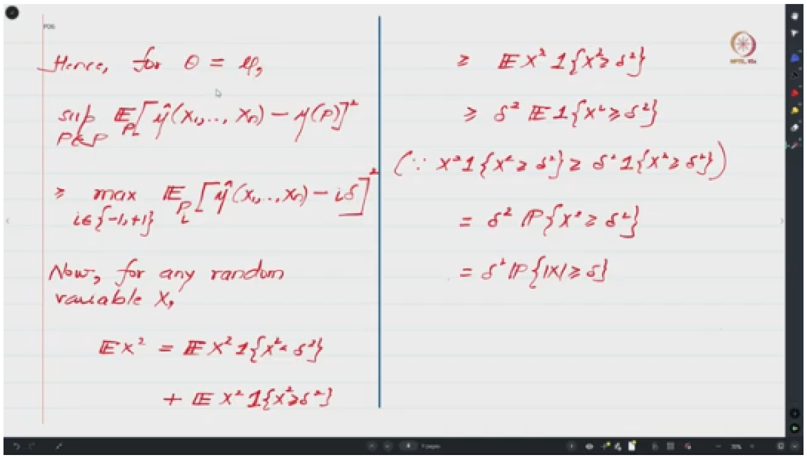
$\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$

$\underline{\mathcal{N}(0, \sigma^2)}$

Right, and in the previous few slides, I mentioned that we would like to compute a lower bound, right? And the lower bound that we are going to derive is a lower bound on this quantity, right? And what we are going to show is that, you know, the worst-case error for the average—you know, the simple averaging estimator—matches in an order sense with respect to a lower bound on this, and we will use this to conclude that indeed the averaging estimator is the you know, optimal estimator in a minimax sense, okay? So, you will soon see what I mean by that, okay? So, I'm going to give an overview of this proof, right? In particular, how to derive a lower bound on this quantity. Unfortunately, I won't have time to go into the full details, but I will try to cover as much as possible. And for those who wish to know more, I request you to look at these lecture notes by John Duchi. He has some lecture notes on his webpage.

I will provide the description of that in one of the comments here so that you can access it and use those notes to actually get all the details about what I am going to discuss now. So, to ground the problem, let us assume that our class of distributions includes all Gaussian distributions with different possible means but with the same fixed variance. So, across all these distributions, the variance is fixed. However, the means can vary. So, this is our class of distributions.

In particular, you know, we will use this symbol P plus 1 to denote the Gaussian distribution with mean plus delta and variance sigma square, and P minus 1 to denote the Gaussian distribution with mean minus delta and variance sigma square. So, you can see that both P plus 1 and P minus 1 belong to this class of distributions. So, these are two specific elements within this distribution, and one can check that the mean between them is, you know, at least 2 delta away, and this fact we will actually make use of in some of our later discussions. So, now the question is: can we discuss the minimax risk associated with this class of distributions with regard to estimating the mean mu? So, for theta equals mu, that is, you know, we are now going to focus on estimating the mean of this underlying distribution.



Hence, for $\theta = \mu$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\hat{\mu}(X_1, \ldots, X_n) - \mu(P)\right]^2$$

$$\geq \max_{i \in \{-1,+1\}} \mathbb{E}_{P_i}\left[\hat{\mu}(X_1, \ldots, X_n) - i\delta\right]^2$$

Now, for any random variable X,

$$\mathbb{E}X^2 = \mathbb{E}X^2 \mathbb{1}\{X^2 < \delta^2\} + \mathbb{E}X^2 \mathbb{1}\{X^2 \geq \delta^2\}$$

$$\geq \mathbb{E}X^2 \mathbb{1}\{X^2 \geq \delta^2\}$$

$$\geq \delta^2 \mathbb{E}\mathbb{1}\{X^2 \geq \delta^2\}$$

$$(\because X^2 \mathbb{1}\{X^2 \geq \delta^2\} \geq \delta^2 \mathbb{1}\{X^2 \geq \delta^2\})$$

$$= \delta^2 \mathbb{P}\{X^2 \geq \delta^2\}$$

$$= \delta^2 \mathbb{P}\{|X| \geq \delta\}$$

The worst-case error for any estimator mu hat. So, I am going to use mu hat in place of theta hat to emphasize that we are now focusing on estimating the mean of this unknown distribution P, and let us say you have some rule mu hat. Based on these observations X1 to Xn, then the worst-case error associated with this estimator is given over here. And since you have, you know, supremum over here, so if I restrict my attention to the two specific distributions that I spoke about and take the max of this expression only over those two distributions. Then it is easy to see that this supremum is actually lower bounded by the, you know, max over two specific distributions. Recall that our goal was to find a lower bound on this minimax risk, in particular this quantity Rn theta p that I had defined on this slide right, so this.

Quantity is lower bounded by this quantity, and what we are going to do is that we are going to find a lower bound on this quantity right and use that as a lower bound on the minimax risk right, and recall our goal is to show that the order of this lower bound matches the performance of your simple averaging estimator. So, now we are going to do a bit of algebra to lower bound this quantity further. So, what we will do is, we will make use of the fact that for any random variable, let us say x, its expected value of x square is equal to expected value of x square indicator x square is strictly less than delta square. And expected value of X square, indicator X square greater than delta square. Since these two indicators add up to 1, that fact can be used to show that this expectation is actually equal to the sum of these expectations.

Now, since we are working with squares, one can easily see that this expectation is non-negative. This expression is lower bounded by the second quantity and that is what we have. Now, x square, sorry I have already written this. So, x square times indicator x square greater than delta square, trivially is bigger than delta square indicator x square greater than delta square. Because on any sample point where this event fails to hold, this indicator will be 0.

right and hence this left hand side will be 0 and the right hand side will also be 0 hence both sides are equal. On the other hand when this event indeed holds that is x square is bigger than delta square on this event since x square is bigger than delta square this random variable is can be shown to be you know lower bounded by this random variable for all sample point and from the monotonicity of the expectation one can show that you know the expectation of this random variable is lower bounded by the expectation of this random variable. However, since delta square is constant, I can pull it outside that expectation and I will be left with delta square expected value of this indicator. Since the expectation of an indicator is just the probability of that event, one can conclude that expected value of X square is lower bounded by delta square.

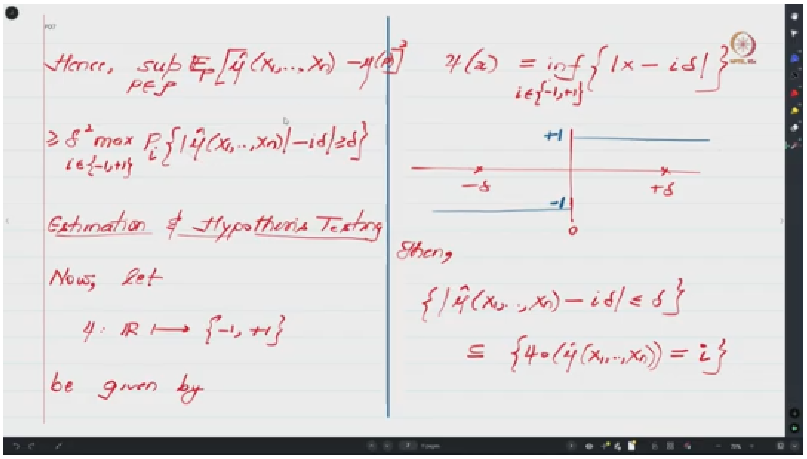
times probability of X square greater than equal to delta square. Is this okay? Now, this event and this event are one and the same because I am using these absolute values. Hence, I can replace this event by absolute value of X because whenever absolute value of X is bigger than delta, X square is bigger than equal to delta square. Similarly, when X square is bigger than equal to delta square, one can conclude that that can only happen when absolute value of X is bigger than equal to delta.

Hence, one can conclude that the expected value of X squared is actually lower bounded by this quantity. So, let us use this fact over here. So here, you can see that you have some random variable over here, and we are going to bound this expression by some quantity like this. And I would again like to highlight what I have written over here. So here, as I told you, we were taking the supremum over all possible P's.

On the right side, we are going to restrict our attention to these two distributions, that is P minus 1 and P plus 1, where we call P minus 1 the Gaussian distribution where the mean is minus delta, and P plus 1 is with mean plus delta, right. So, when we are working with PI, right. When I is between one of these two elements, then the mean is actually I delta. So, for example, when I is plus 1, this is plus delta, and when I is minus 1, this is minus delta, and accordingly, this expression can be written in this fashion. Again, as I told you, this lower bound appears because instead of looking at the supremum over all possible distributions, we are focusing our attention on the two distributions P minus 1 and P plus 1. So, we are going to use this bound over here to obtain a lower bound on this quantity.

And it is easy to see that this supremum, based on whatever we discussed, is lower bounded by delta squared times the max over i belonging to minus 1 plus 1, P of i, the absolute value of this difference being greater than or equal to delta. So, the expected X squared is greater than or equal to delta squared times P of the absolute value of X bigger than or equal to delta. Hence, whatever we have over here, this quantity can be lower bounded by the probability that this difference is actually bigger than or equal to delta.

So, by doing some simple algebra, one can show that the worst-case error is actually lower bounded by the quantity that we have over here. This is just some simple algebra.



Hence, $\sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\hat{\mu}(X_1,\ldots,X_n) - \mu(P)\right]^2$

$\geq \delta^2 \max_{i \in \{-1,+1\}} P_i\left\{|\hat{\mu}(X_1,\ldots,X_n)| - i\delta| \geq \delta\right\}$

Estimation & Hypothesis Testing

Now, let

$\psi : \mathbb{R} \longmapsto \{-1,+1\}$

be given by

$\psi(x) = \inf_{i \in \{-1,+1\}} \{|x - i\delta|\}$

Then,

$\{|\hat{\mu}(X_1,\ldots,X_n) - i\delta| \leq \delta\}$

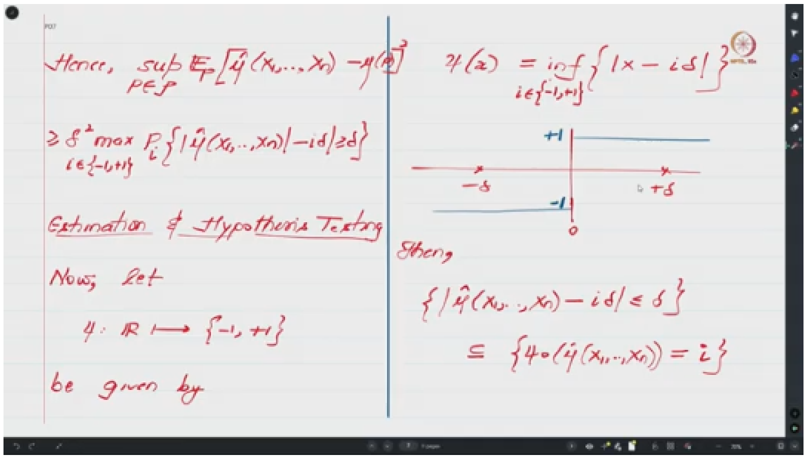
$\subseteq \{\psi \circ (\hat{\mu}(X_1,\ldots,X_n)) = i\}$

So, the key thing that we are going to do is to relate this estimation error—the worst-case estimation error—to this problem of hypothesis testing. Is this okay? So, this is where we will transform one problem of estimation to that of hypothesis testing, and we will use some knowledge about this hypothesis testing problem to obtain some lower bound on a quantity like this. So, let me explain what I mean by hypothesis testing, right? So, here, we have been given two distributions, right? One whose mean is minus delta and the other whose mean is plus delta, right?

And in hypothesis testing, what we have to do is, whenever we observe some samples, right? We have to decide whether those samples come from this distribution or that distribution. So, in some sense, this is like an easier problem: instead of wanting to estimate the mean, we are only wanting to say whether it is coming from this distribution or that distribution. So, this is what we are trying to do, and this is what the hypothesis testing problem is. So, in the estimation problem, we want to estimate the mean, whereas in the hypothesis testing,

having looked at these samples, we want to know whether these samples come from the distribution whose mean is minus delta or from the distribution whose mean is plus delta. So, toward making this formal connection between this estimation problem and this hypothesis testing problem, let us define this function psi in the following way. So, this psi actually goes from the set of real numbers to this set of outputs, minus 1 and plus 1.

So, whenever it says minus 1, we have to conclude that the hypothesis testing problem is suggesting that it is minus delta, and whenever it is plus 1, then the hypothesis test is suggesting that the samples perhaps have come from this distribution with mean plus delta, right?



Hence, $\sup_{P \in \mathcal{P}} E_P[(\hat{\mu}(X_1, \ldots, X_n) - \mu(P))^2]$

$\geq \delta^2 \max_{i \in \{-1,+1\}} P_i\{|\hat{\mu}(X_1, \ldots, X_n) - i\delta| \geq \delta\}$

Estimation & Hypothesis Testing

Now, let

$\psi : \mathbb{R} \longmapsto \{-1, +1\}$

be given by

$\psi(x) = \inf_{i \in \{-1,+1\}} \{|x - i\delta|\}$

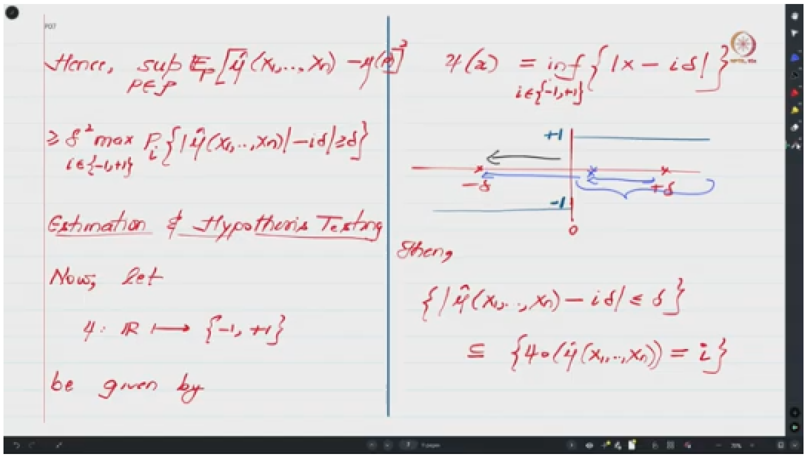
Then,

$\{|\hat{\mu}(X_1, \ldots, X_n) - i\delta| \leq \delta\}$

$\subseteq \{\psi \circ (\hat{\mu}(X_1, \ldots, X_n)) = i\}$

So, keeping that in mind, let us define this psi of x to be the inf of the absolute value of x minus i delta. So, what do I mean by that? So, whenever your x is over here, right on the right side of your origin, I hope you agree that the outcome of this definition will be plus 1 because whenever you are over here, right, this value will be closer to plus Delta as compared to, you know, the distance to minus Delta. So, because this value is closer to this one, you can conclude that you know, the I that achieves the infimum when X is to the right of the origin is plus 1, and hence psi of X is plus 1. On the other hand, whenever X, let us say, lies on this side right here, the, you know, the outcome of your psi of X will be minus 1. So, whenever I give you an X

So, if it is to the left of the origin, this psi of x will give an output as minus 1, and whenever it is to the right of the origin, it will output as plus 1. So, that is what I have, you know, drawn over here. So, you can see that this psi function takes the value of plus 1 on the right side and minus 1 on the left side, and at the origin, we can come up with some rule for breaking ties. Is this okay? From this definition of psi, I hope you agree that, you know, let us say mu hat, whatever your estimation rule is, right?



Hence, $\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\hat{\mu}(X_1,\dots,X_n)-\mu(P)\right]^2$

$\geq \delta^2 \max_{i\in\{-1,+1\}} P_i\left\{|\hat{\mu}(X_1,\dots,X_n)|-i\delta| \geq \delta\right\}$

Estimation & Hypothesis Testing

Now, let

$\psi : \mathbb{R} \longmapsto \{-1,+1\}$

be given by

$\psi(x) = \inf_{i\in\{-1,+1\}} \left\{|x - i\delta|\right\}$

Then,

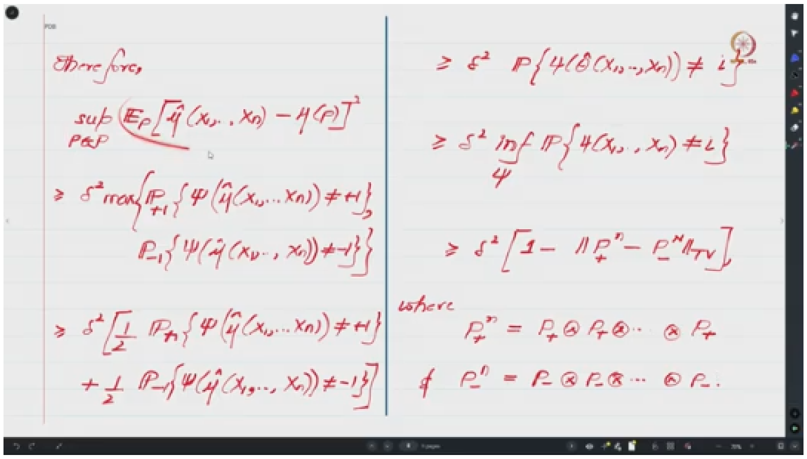
$\left\{|\hat{\mu}(X_1,\dots,X_n) - i\delta| \leq \delta\right\}$

$\subseteq \left\{\psi\circ(\hat{\mu}(X_1,\dots,X_n)) = i\right\}$

So, let us say it comes up with some value of the, you know, underlying distribution—sorry, the mean of the underlying distribution, right? And let us say this difference is less than or equal to delta, right? So, let us say this difference for some reason is less than or equal to delta, or whenever this is less than delta, one can immediately see that, you know, if you apply psi on this value. So, this mu hat will take as input these samples and spit out a number because this is trying to estimate the mean, and the mean right now is, you know, some number on the real line—specifically, it is minus delta or plus delta. So, this algorithm that we have will spit out a number, and this number will lie somewhere on the real line.

So, whenever this distance to I delta is less than or equal to delta, one can conclude that if you apply psi on this expression. That will be exactly i. Like, for example, if your mu hat sits somewhere over here, right? Then, if you apply psi on this, then indeed the answer will be i, right? Because when you are sitting here, right? We have already shown that the value of psi of x is plus 1, and when you are sitting over here, the value is minus 1.

So, one can see that whenever you have such an event, when you apply psi on that, it will give us i. Is this okay? So, from this observation, one can see that if you take the supremum of this expectation over all possible P's, right? We have already shown that this is lower bounded by, you know, the max between, you know, this quantity over here, right, and this event. is, you know, a subset of this. So, in particular, notice that—so let me restate it—this event is a subset of this, and what we have over here is a complement of that event. So, since this is a subset of this, So, the complement of this will be a superset of this, and one can conclude that this probability is lower bounded by the complement of this probability. So, that is what I have written over here.



Therefore,

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\hat{\psi}(x_1,\dots,x_n)-\psi(P)\right]^2$$

$$\geq \delta^2 \max\left\{\mathbb{P}_{+1}\{\Psi(\hat{\psi}(x_1\dots x_n))\neq +1\},\ \mathbb{P}_{-1}\{\Psi(\hat{\psi}(x_1,\dots,x_n))\neq -1\}\right\}$$

$$\geq \delta^2\left[\frac{1}{2}\mathbb{P}_{+1}\{\Psi(\hat{\psi}(x_1\dots x_n))\neq +1\} + \frac{1}{2}\mathbb{P}_{-1}\{\Psi(\hat{\psi}(x_1,\dots,x_n))\neq -1\}\right]$$

$$\geq \delta^2\ \mathbb{P}\{\Psi(\hat{\theta}(x_1,\dots,x_n))\neq i\}$$

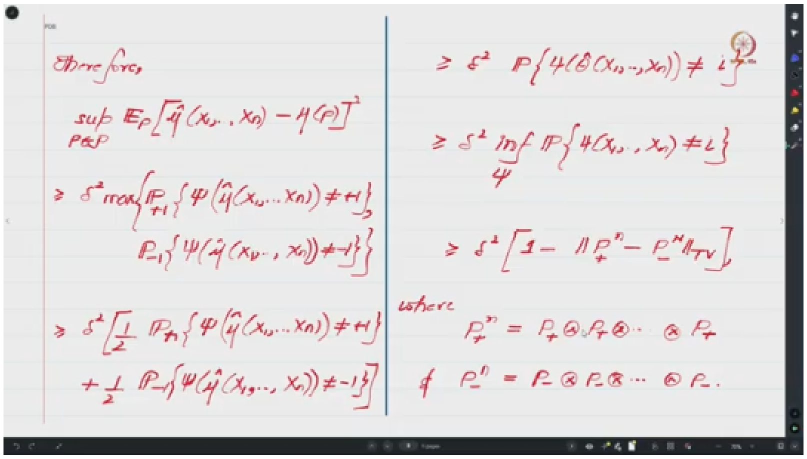
$$\geq \delta^2 \inf_{\Psi} \mathbb{P}\{\Psi(x_1,\dots,x_n)\neq i\}$$

$$\geq \delta^2\left[1-\|P_+^n-P_-^n\|_{TV}\right]$$

where

$$P_+^n = P_+\otimes P_+\otimes\cdots\otimes P_+$$

& $$P_-^n = P_-\otimes P_-\otimes\cdots\otimes P_-.$$



Therefore,

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\hat{\psi}(x_1,\dots,x_n)-\psi(P)\right]^2$$

$$\geq \delta^2 \max\left\{\mathbb{P}_{+1}\{\Psi(\hat{\psi}(x_1\dots x_n))\neq +1\},\ \mathbb{P}_{-1}\{\Psi(\hat{\psi}(x_1,\dots,x_n))\neq -1\}\right\}$$

$$\geq \delta^2\left[\frac{1}{2}\mathbb{P}_{+1}\{\Psi(\hat{\psi}(x_1\dots x_n))\neq +1\} + \frac{1}{2}\mathbb{P}_{-1}\{\Psi(\hat{\psi}(x_1,\dots,x_n))\neq -1\}\right]$$

$$\geq \delta^2\ \mathbb{P}\{\Psi(\hat{\theta}(x_1,\dots,x_n))\neq i\}$$

$$\geq \delta^2 \inf_{\Psi} \mathbb{P}\{\Psi(x_1,\dots,x_n)\neq i\}$$

$$\geq \delta^2\left[1-\|P_+^n-P_-^n\|_{TV}\right]$$

where

$$P_+^n = P_+\otimes P_+\otimes\cdots\otimes P_+$$

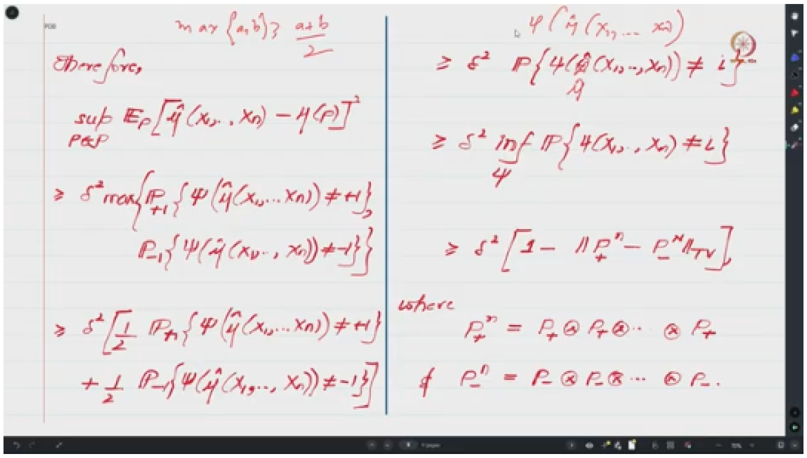
& $$P_-^n = P_-\otimes P_-\otimes\cdots\otimes P_-.$$

So, this expression is lower bounded by what we had on the previous slide, and those quantities are actually lower bounded by these two expressions over here. And since we had a max there, I am taking a max over here as well. So, this is the max between this. So, now notice that, you know, this was like a, you know, we started out with an estimation error, right? And now we have translated that error into a hypothesis testing, right? So, what we are saying is whether the estimate will be, you know, the hypothesis will be that the samples have been generated from this distribution with mean plus delta, or the samples have been generated with some, I mean, from the distribution whose mean is minus 1. So, in this way, we have transferred, you know, the problem on risk estimation to error related to hypothesis testing, right? And we will do a bunch of algebra to, you know, sort of obtain a lower bound on this hypothesis testing problem—like, for example, you know, the max of two quantities is always lower bounded by, you know, the average, right.

So, all I am saying is that the max of any two numbers. a and b, right? This is lower bounded by (a plus b) over 2, right? So, by using this simple fact, one can say that the max of these two quantities is lower bounded by half times their sum. And now you can interpret this half as the probability of choosing the underlying distribution. So, the underlying distribution is either the one with mean plus delta or the one with mean minus delta. Right, and one can imagine this half to be the probability with which the underlying distribution is chosen. Hence, one can subsume this half and show that this quantity is actually equal to this quantity, where not only are X1 to Xn random variables, but I is also a random variable. So, you can imagine this experiment to be: first, you pick I,

which is either -1 or +1 with probability half, right? Then, conditional on this choice of I, we sample X1 to Xn and reveal these observations. Right? And this estimator—sorry, I have used theta hat here; I should use mu hat. This estimator mu hat will then take in these samples, and this psi function will make a prediction of the underlying distribution. And what we have shown is that this risk error we discussed is lower bounded by this quantity over here. Is this okay? And now what we can do is this psi of mu hat.

You know, mu hat, right? So, this psi of mu hat—let me do that again—psi of mu hat of X1 to Xn, this is one way to figure out what the hypothesis is. So, since we are working with one psi—this specific psi—we can say that this expression, this psi of mu hat, which looks at the samples and guesses what your hypothesis is. So, this is just one of your psi functions.



$\max\{a,b\} \geq \frac{a+b}{2}$

Therefore,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\hat{\mu}(x_1,..,x_n) - \mu(P)\right]^2$$

$$\geq \delta^2 \max\left\{ \mathbb{P}_{+1}\{\psi(\hat{\mu}(x_1,...x_n)) \neq +1\}, \; \mathbb{P}_{-1}\{\psi(\hat{\mu}(x_1,..,x_n)) \neq -1\} \right\}$$

$$\geq \delta^2\left[\frac{1}{2}\,\mathbb{P}_{+1}\{\psi(\hat{\mu}(x_1,...x_n)) \neq +1\} + \frac{1}{2}\,\mathbb{P}_{-1}\{\psi(\hat{\mu}(x_1,..,x_n)) \neq -1\}\right]$$

$\psi(\hat{\mu}(x_1,...\,x_n))$

$$\geq \delta^2 \; \mathbb{P}\{\psi(\hat{\mu}(x_1,..,x_n)) \neq i\}$$

$$\geq \delta^2 \inf_{\psi} \mathbb{P}\{\psi(x_1,..,x_n) \neq i\}$$

$$\geq \delta^2\left[1 - \|P_+^n - P_-^n\|_{TV}\right]$$

where

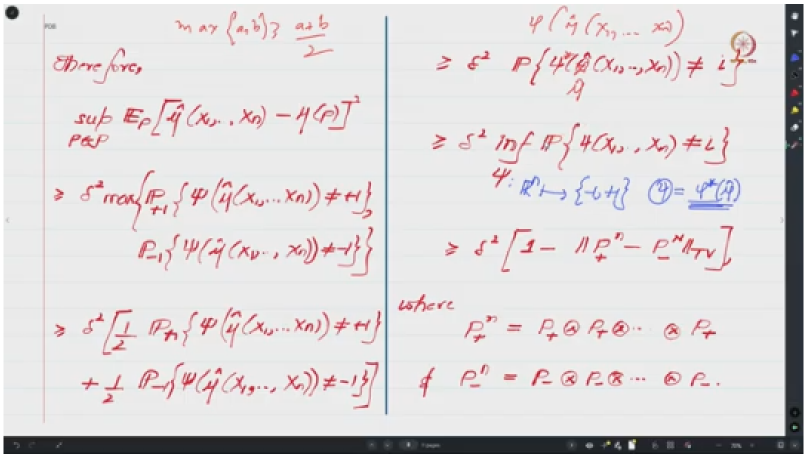
$$P_+^n = P_+ \otimes P_+ \otimes \cdots \otimes P_+$$

& $P_-^n = P_- \otimes P_- \otimes \cdots \otimes P_-$.

So, one can, you know, say that this specific—I mean, the probability for this specific psi function—is actually lower-bounded by the inf over all possible psi functions. Is this okay? Where psi actually goes from you know, goes from Rn to minus 1 plus 1, right? So, you know, your psi of—so, maybe I will call this with some special notation, let us say psi star over here.

Then for this choice of psi star—so again, I would like to emphasize—let us say this is your psi star function. So, this is some specific way of doing hypothesis testing. So, in this case, your psi function is basically psi star of, you know, mu hat. So, this function over here takes in as input x1 to xn, first applies your mu hat, and then applies psi star. So, in this way, you can interpret this psi to be of this form over here, and hence, you know, the probability over a specific choice of psi is lower-bounded by the inf.

over all possible psi functions. Now, why are we doing this? Well, from information theory, you know, the details can be found in these John Bucci lecture notes. One can show that this infimum is actually lower-bounded by the total variation distance between the n-fold product of your p plus 1 distribution and p minus 1 distribution. So, this n and n should be the same.



$\max\{a,b\} \geq \frac{a+b}{2}$

Therefore,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\hat{\mu}(x_1,\ldots,x_n) - \mu(P)\right]^2$$

$$\geq \delta^2 \max\left\{ \mathbb{P}_{+1}\{\Psi(\hat{\mu}(x_1,\ldots,x_n)) \neq +1\},\ \mathbb{P}_{-1}\{\Psi(\hat{\mu}(x_1,\ldots,x_n)) \neq -1\} \right\}$$

$$\geq \delta^2\left[\frac{1}{2}\mathbb{P}_{+1}\{\Psi(\hat{\mu}(x_1,\ldots,x_n)) \neq +1\} + \frac{1}{2}\mathbb{P}_{-1}\{\Psi(\hat{\mu}(x_1,\ldots,x_n)) \neq -1\}\right]$$

$\Psi(\hat{\mu}(x_1,\ldots,x_n))$

$$\geq \delta^2\, \mathbb{P}\{\Psi^*(\hat{\mu}(x_1,\ldots,x_n)) \neq i\}$$

$$\geq \delta^2 \inf_{\Psi} \mathbb{P}\{\Psi(x_1,\ldots,x_n) \neq i\}$$

$\Psi : \mathbb{R}^n \mapsto \{-1,+1\}$ $\quad \Psi = \Psi^*(\hat{\mu})$

$$\geq \delta^2\left[1 - \|P_+^n - P_-^n\|_{TV}\right]$$

where

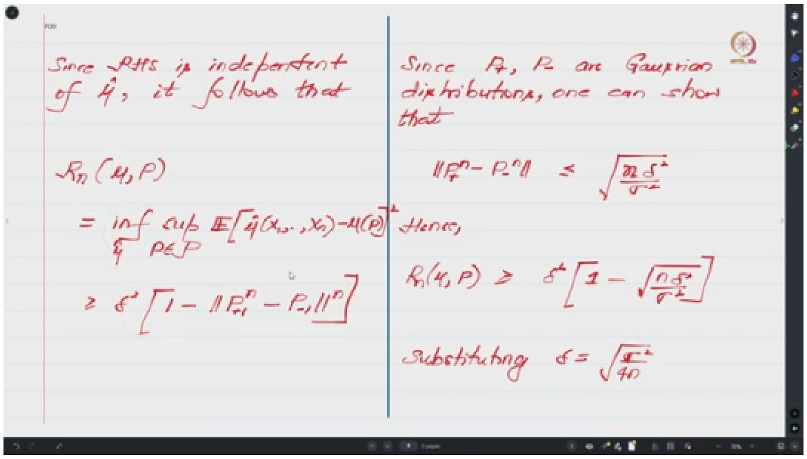
$$P_+^n = P_+ \otimes P_+ \otimes \cdots \otimes P_+$$

$$\& \quad P_-^n = P_- \otimes P_- \otimes \cdots \otimes P_-.$$

what do I mean by the n-fold product distribution you can think of the n-fold product distribution as the joint distribution of n independent samples generated from the same distribution okay so this is the notation for that so basically this is the product distribution so what do I mean by that so P plus 1 let's say of 2 of you know let's say So, here P plus 1 takes you know some value let us say you know. So, this is the product distribution. So, what do I mean by that? I take the Cartesian product of A cross B right where both A and B are subsets of let us say R right.

So, this expression then would be P plus 1 of A times P plus 1 of B okay. So, this is what the n fold product distribution means in particular I have here the twofold distribution in the same way this can be expanded to the n fold distribution. So, if you look at the total variation distance between the n fold product distribution of the p plus 1 distribution and p minus 1, then this quantity actually lower bounds this quantity here. So, how we actually make this jump for this I request you to go through these lecture notes by John Ducey right and since this right hand side is now independent of your mu hat.

So, notice that we had a specific estimator here and then you know once we made this jump from here to here we sort of get rid of the estimator. And we say that whatever be your estimator we this you know hypothesis testing error is actually lower bounded by this quantity and from information theory one can show that this quantity is lower bounded by delta square times the n fold total variation distance between the n fold p plus 1 and n fold p minus 1 distributions. And since this quantity is independent of mu hat one can show that you know by taking you know inf over mu hat here. So, notice that all these calculations were done presuming mu hat is some specific estimator but now if you take inf over mu hats here then that quantity which we denote it as Rn mu p that quantity will also be lower bounded by this expression that we have over here. And since this P plus 1 and P minus 1 are actually Gaussian distributions, they are n fold products total variation distance one can compute and one can show that the total variation distance is actually upper bounded by some quantity like this.



Since RHS is independent of $\hat{\mu}$, it follows that

$R_n(\mu, P)$

$= \inf_{\hat{\mu}} \sup_{P \in \mathcal{P}} \mathbb{E}[\hat{\mu}(X_1, \dots, X_n) - \mu(P)]^2$

$\geq \delta^2 [1 - \|P_{+1}^n - P_{-1}\|^n]$

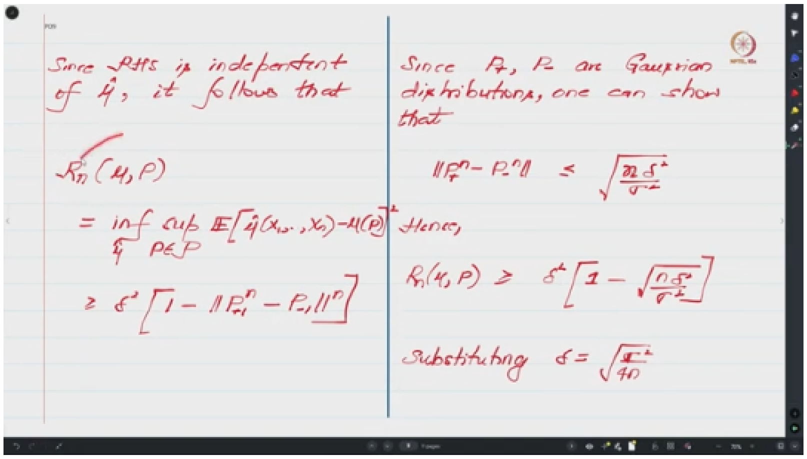
Since P_+, P_- are Gaussian distributions, one can show that

$\|P_+^n - P_-^n\| \leq \sqrt{\frac{n\delta^2}{\sigma^2}}$

Hence,

$R_n(\mu, P) \geq \delta^2 [1 - \sqrt{\frac{n\delta^2}{\sigma^2}}]$

Substituting $\delta = \sqrt{\frac{\sigma^2}{4n}}$



Since RHS is independent of $\hat{\mu}$, it follows that

$R_n(\mu, P)$

$= \inf_{\hat{\mu}} \sup_{P \in \mathcal{P}} \mathbb{E}[\hat{\mu}(X_1, \dots, X_n) - \mu(P)]^2$

$\geq \delta^2 [1 - \|P_{+1}^n - P_{-1}\|^n]$

Since P_+, P_- are Gaussian distributions, one can show that

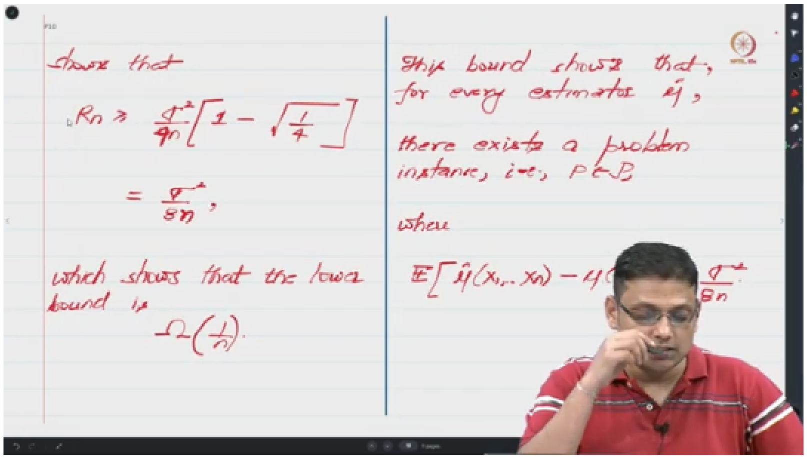
$\|P_+^n - P_-^n\| \leq \sqrt{\frac{n\delta^2}{\sigma^2}}$

Hence,

$R_n(\mu, P) \geq \delta^2 [1 - \sqrt{\frac{n\delta^2}{\sigma^2}}]$

Substituting $\delta = \sqrt{\frac{\sigma^2}{4n}}$

So, recall that p plus 1 and p minus 1 are basically the distributions with means plus delta and minus delta. So, using that fact and using the fact that they are Gaussian, one can show that the total variation distance is actually upper bounded by the square root of n delta squared over sigma squared, right? And hence, one can see that the minimax risk. That is associated with the mean estimation problem is lower bounded by this quantity over here. Now, if you choose this specific value of delta, that is the square root of sigma squared over 4n, and substitute this over here, one can see that the minimax risk estimator is actually lower bounded by sigma squared over 8n. So, you know, I want to give you the gist of what this result is trying to tell us.



Shows that

$$R_n \geq \frac{\sigma^2}{4n}\left[1 - \sqrt{\frac{1}{4}}\right]$$

$$= \frac{\sigma^2}{8n},$$

which shows that the lower bound is

$$\Omega\left(\frac{1}{n}\right).$$

This bound shows that, for every estimator $\hat{\mu}$,

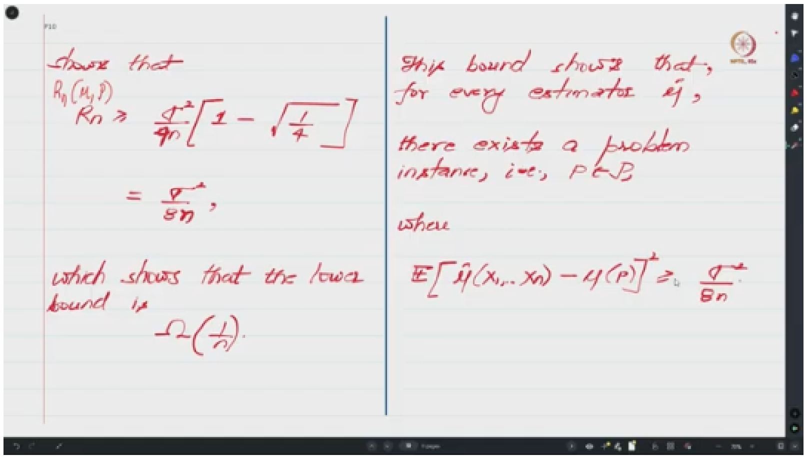
there exists a problem instance, i.e., $P \in \mathcal{B}$

where

$$\mathbb{E}\left[\hat{\mu}(X_1,..X_n) - \mu(\quad \frac{\sigma^2}{8n}$$

This result is saying that the lower bound to this minimax risk that we had defined, which is, you know, Rn of mu, p. This quantity over here is actually lower bounded by this. And notice that this expression is also of the order 1 over n. So, what this bound tells us is that for every estimator mu hat, okay? So, recall that this Rn mu p involved an inf over mu hat and a sup over the problem instances. So, what this bound tells us is that For every estimator mu hat, there exists a problem instance where the squared error will at least be sigma squared over a10, which means that whatever rule you come up with,

for using the samples to estimate the mean, I can find—I mean, at least there may not find, but one can show the existence of a problem instance under which the squared mean squared error is at least sigma squared over a10, right? And we have already shown that the average algorithm, the averaging estimator, right, that achieves this bound. Irrespective of your problem instance, that is, you give me any problem instance, this average estimator achieves this error of sigma squared over n, right? So, one can show that for any other estimator, there is at least one problem instance. Where it suffers or has an error of order 1 over n, and then you have this averaging estimator which, on all problem instances, achieves an error of order 1 over n. So, if you look at it from this order perspective, one can see that the sample average is indeed the optimal estimator from this minimax risk perspective. So, let me now summarize what we have done so far. So, we wanted to study how good is your sample average estimator.



shows that

$R_n(\mu, \mathcal{P})$

$$R_n \geq \frac{\sigma^2}{4n}\left[1 - \sqrt{\frac{1}{4}}\right]$$

$$= \frac{\sigma^2}{8n},$$

which shows that the lower bound is

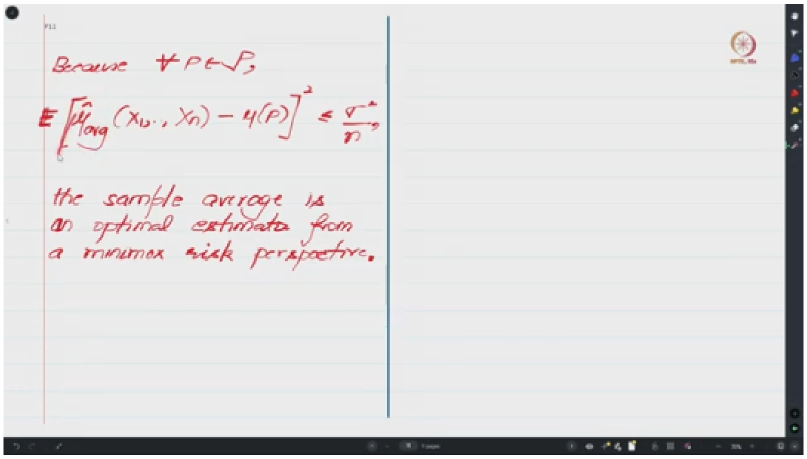
$$\Omega\left(\frac{1}{n}\right).$$

This bound shows that, for every estimator $\hat{\mu}$,

there exists a problem instance, i.e., $P \in \mathcal{P}$

where

$$\mathbb{E}\left[\hat{\mu}(X_1,..X_n) - \mu(P)\right]^2 \geq \frac{\sigma^2}{8n}.$$



Because $\forall P \in \mathcal{P}$,

$$\mathbb{E}\left[\hat{\mu}_{avg}(X_1,..,X_n) - \mu(P)\right]^2 \leq \frac{\sigma^2}{n},$$

the sample average is an optimal estimator from a minimax risk perspective.

On one hand, we asked whether by working with a different step size choice we would get a better convergence rate; the answer was no. Today, we asked whether, other than the sample average, we could have used any other function. Would we have come up with a better convergence rate? The answer again turns out to be no, from an order sense, in particular from this minimax risk perspective. What this result says is that, whatever your estimator is, there is a problem instance

where it will suffer an error of order 1 over n, right? And hence, since the average estimator has this sigma square over n error irrespective of the problem instance, right? It is the optimal, it is an optimal estimator from this minimax risk sense. So, I would like to now stop this discussion on convergence rates of stochastic approximation algorithms by talking about how these ideas extend to more general linear stochastic approximation algorithms and also non-linear algorithms. In linear stochastic algorithms, again, one can show that the step size of the form 1 over n indeed leads to the optimal convergence rate.

However, often one requires that the step size be of the form C over n plus 1, meaning the numerator needs to be multiplied by a constant. And unfortunately, in general linear stochastic approximation, this constant often needs to be chosen based on unknown problem-dependent parameters. So there has been a lot of ongoing research to fix these issues, and one of the popular ideas to address them is what is known as the Polyak-Ruppert averaging, which shows that one can achieve an optimal convergence rate for linear stochastic approximation algorithms even without knowledge of these unknown parameters. Is this okay? And the same idea, at some level, extends also to non-linear stochastic approximation algorithms.

Because at some level, one can show that, you know, if your non-linear stochastic approximation algorithm converges to some limit point, then in the neighborhood of that limit point, this non-linear stochastic approximation can be approximated, you know, in the form of a linear stochastic approximation. Hence, whatever we have studied here often applies in the non-linear context as well. So, to summarize from a convergence rate perspective, often one may need to work with this step size of the form 1 over n plus 1. For this mean estimation problem, this 1 over n plus 1 step size works perfectly well. But for general linear stochastic approximation, one may often need to work with a constant. To achieve the optimal convergence rate, this constant may depend on unknown parameters, and some of the techniques to overcome this are what is known as the Poyack-Ruppert averaging technique. And, you know, the same ideas also extend to non-linear stochastic approximation.

So, with that, I would like to stop. Next week, we will talk about the stability of stochastic approximation algorithms. Until then, thank you, goodbye, and namaste.