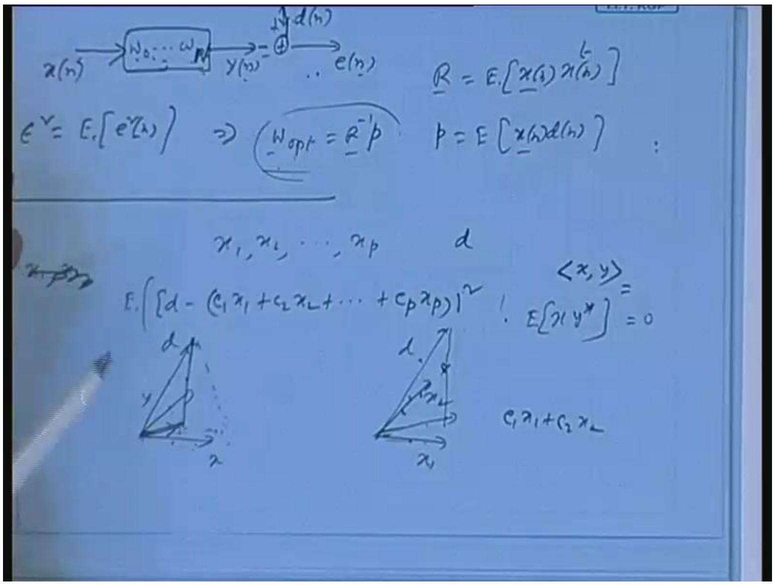
Adaptive Signal Processing
Prof. M. Chakraborty
Department of Electronics & Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture - 6
Steepest Descent Technique

Dealing with this optimal filters or wiener filters with an FIR filter just a quick recap, we had zero mean discrete time random process x n it was pass through a filter w 0 w p purpose is to generate an output y n, which is a very good estimate of a desired sequence d n.

(Refer Slide Time: 01:13)



$\underline{R} = E[\underline{x}(n)\underline{x}^t(n)]$

$\epsilon^2 = E[e^2(n)] \Rightarrow (\underline{w}_{opt} = \underline{R}^{-1}\underline{p})$ $\quad p = E[\underline{x}(n)d(n)]$

$x_1, x_2, \cdots, x_p \qquad d$

$E([d - (c_1x_1 + c_2x_2 + \cdots + c_px_p)]^2 \qquad \langle x, y \rangle = E[xy^*] = 0$

$c_1x_1 + c_2x_2$

So, if you take the error between the two that error which is a random process. So, as I told you that there is no point in minimizing that absolute value of an error for a particular n; because next time you absorb the e n the waveform will change. So, you better take the variance of e n and because of stationarity as you saw yesterday because of stationarity it will become independent of n. So, variance will give the power after all d n is zero means x n is zero means therefore, y n is zero mean, so e n is zero mean. So, absolute I mean expected value of e n square will be the variance of e n and if there is a power if that is minimized because e n will depend on d n and y n and y n depends of the weights.

So, if you can choose the weights such that minimum I mean this mean square value of this error is minimized then those weights will be optimal weights very good weights.
Then, you saw that e n is I mean if you take the variants that is indeed a quadratic function of the filter weights w 0 to w p, which then has either a maxima or minima only one unique. In all case of rest it is minimal, so we found out the minima value that gives the optimal filter for which the variants is minimum. So, there is if you have design epsilon square is equal to you see I have already eliminate n here because of stationarity we also verify it after substituting for e n in the expression yesterday we found that n goes from everywhere.

This was a quadratic function of the weights w 0 to w p this was the quadratic function of the weights w 0 to w p. So, we minimize this will give some formula for matrix differentiation also. So, minimize this and that led to optimal filter w opt as R inverse p this is called the FIR wiener filter. This is the FIR wiener filter and this gives the information about R matrix R is E x n in general x n Hermitian, but in our case we are dealing with real; yes I forgot to mention that we are dealing with the real valued case.
So, that is why it was transpose this is called the correlation matrix p was E x n d n. This was the definition using the R and p you can compute the optimal filter that is the wiener filter that will give raise to minimum value of epsilon square all that we have seen yesterday.

Actually, I tell you something now and again I will come back to that later after the several lectures that actually you know what happens in this estimation thing. Here, you are having x n then x n minus 1 up to x n minus p they are multiplied by the weights w 0 w 1 and w p are now added. In general, suppose I have a situation like this in general you have got a random set of random variables x 1 x 2. It is very difficult things if changes that one here does it go out of the spin x 1 x 2 suppose you have got a set of random variables x 1 x 2 dot dot dot x p zero mean random variable say could be complex valued could be real valued.

We want to estimate another random variable d as a linear combination of these that is called linear estimation. What we do there we are trying to find out the coefficient c 1 x 1 plus c 2 x 2 plus dot dot dot c p x p. This coefficient c 1 c p are to be found out so that; if you take the error d minus this the error I am assuming real otherwise simply take the mod square. So, square expected value of that that is minimum that will be a quadratic function of c 1 to c p and therefore, this are the minima you find it out. What is actually happening here? You know you suppose see the analogy forget about the random variable suppose I give you some vectors in a three dimension world or have been two dimension world one vector say; they may not be an orthogonal one vector say x another y and this another third vector you call d; you want to estimate as a linear combination of y and x.

What you will do? You will project it orthogonally on this plane; so that on this plane this guy is orthogonal do the orthogonal to this x is this axis and this much this projection; what is the meaning of this projection? Projection is such that the norm square of the difference that is minimum, because it is not this thing in geometry if I do not project it orthogonally on the plane if take it in this again this way I will have so much of difference; so much of this projection this we will difference them. Obviously, norm square of that will be longer than this larger than this. Any project any estimate of d what is the best estimate of d, staying in the plane of plane span by x and y.

What is the best estimate of p? You project it orthogonally take this component that is the best estimate because then only I has minimum norm square. This is a simple geometry if you take any other vector connected like this that will have that difference will have larger norm square value than the original one not only that the projection and you will get only one unique such thing. There is no two side this you cannot get two such com vectors in this plane, which we have you know same norm square I mean you will that have same difference. You cannot have another vector which we if you take the difference the difference norm square the difference vector same as the norm square of these difference vector that cannot happen.

This projection is unique all is things in a more abstract space; we will prove later using abstract vector space, but you can see this orthogonal projection exists it is unique; exist and unique and this gives a minimum norm square value for the difference that is why you say. So, this component is the projection and that will be called the linear estimate, because that what is this vector it will be something of some constant times x some constant time y; any vector in the plane is something times x plus something times y. You are in familiar with their i vector j vector k vector I am calling x and y vector. So, projection will be a linear combination of x and y and that linear combination for which the difference has the minimum norm square error.

Similarly, in the case of random variables you can further timing we can only view later; we will prove this things you know is a vector space treat well. We will prove this things more rigorously, but here you can view suppose you have got two vectors two random variables x 1 and x 2 you view them as mass kind of vector say. We have got a fellow d another variable. You want to find the best estimate of d as a linear combination of x 1 x 2. So, you can assume you can imagine all kind of plane or space span by x 1 and x 2 where every third element is a linear combination of some linear combination of x 1 and x 2.

In that case, which one will you take as a best linear combination? So, that the error is minimize in terms of non-square so; that means, you have to orthogonally project, but the here it has perfect three dimensional geometry even their motion of angle, you know ninety degree. You know orthogonal perpendicularity they are in the case of random variable there is no angle. So, I will tell you this angle this so called dot product they are very special case of a more general dot product that dot product is called inner product value of we will discuss later in the course of this as we go along in this course. I am just giving a hint before end, because you can always ask me sir why did you suddenly choose a filter.

Then you say I choose a filter and I take a y n and then try to minimize mod e n square I mean epsilon square as a function of this, but why filter why not something else that questions no idea I mean you people did not ask I said on my own impose the filter. I on my own impose the filter then said this epsilon square depend on the filter weights. So, why not minimize it, but question is to estimate d n why at all filter why not some other structure have may be that could be better estimate. I am tried to generate the motivation for this filter why suddenly filter came.

So, you see here you know I mean two I mean angle is a by product of the notion of dot product; when dot product between two vectors is zero then we say the two vectors are orthogonal and the angle between them we given them ninety degree. So, the dot product thing is more general the whether two dot dot product is zero; then you say two vectors are orthogonal here. This is projection means the difference is orthogonal with this entire plane that is if you take the two axis it is orthogonal to both and therefore, orthogonal to any in the plane, but there orthogonality has a clear cut notion in terms of dot product that is that cos theta cos ninety zero.

Here I do not have that, but I will tell you theta is no consequence. Here also we define I more general dot product all these things we will prove method we will I am telling you we will be taken up later more rigorously. Here I am only showing analogy; analogy does not hold in mathematics only is things will be proved will be taken up rigorously later. But I am trying to generate the motivation for this filter that is why I am bringing this motion of projection and all that. Here between any two random variable say x and y if they are complex value in general you take the correlation between them that is called the dot product between x and y, which we call actually inner product.

Dot product is a special case more general term is inner product and you normally so x. So, dot y p dot q instead this is a notation x y that dot notation goes essentially the same, but x comma y this is the dot product in our terminology; we call it inner product between x and y this is a more general term. This is defined it should satisfy some property also those things are not showing. Those properties are really satisfied that is why it is taken of in mathematics; otherwise it cannot be it would not do any good. But physically all you can see one thing; correlation gives you what if the correlation is high both are zero mean. So, this correlation has its covariance, so this is high; that means, the two vectors are highly correlated.

So, they are alike they just like you know two vectors in this three dimensional wall; if they are close to each other then only that dot product is less. So, dot product is high dot product is high similarly if they are correlate I mean if this correlation goes up, then only they are I mean like they are not I mean there is no you know physical notion of near or far here I do not say they are close, but they are quite alike. They have lot of similarities or some correlation. If they are not then they are I mean like here two vectors if they are far away; there is some then there are at orthogonal at ninety degree. There is a for these two vectors can go it cannot till further, then again that other angle start simply, so dot product become zero.

Similarly, here also correlation this correlation is a kind of major of the dot product between x and y so if correlation is high the two vectors are quite alike. So, I can as though they are close to each other. If they are uncorrelated means covariates zero and covariates and correlation here are same because mean is zero that is if this is zero, then we say they are orthogonal because they are not like each other. So, equivalently I mean it is like the situations of two vectors are ninety degree with each other that kind of situation here using that.

Another thing is another property of the dot product there was then if you take the any vector in the 3D world and take the dot product with itself that must be a real quantity; there is a major of the length or power energy whatever you know length square or if you do to give a something energetic the energy whatever it is a major of that. Similar thing would hold here also you see if you if you take the inner product with itself E of mod x square, which is a mean square value that will be the actually is power. It is real quantity it gives major of how powerful it is a an how powerful how less powerful it is and all that.

So, this is actually this was satisfy basic properties of some basic axioms for dot product that will do later. So, with that inner product if you want to know estimate d as a linear combination of x 1 and x 2; you should find out c 1 x 1 and c 2 x 2 something see I am drawing the similar figure, but actually the random variable in case there is no such you know angle or line and all that it is only for analogy purpose please understand that. So, I should find out a vector, which is c 1 x 1 plus c 2 x 2 some linear combination; so that d minus this d minus this that is orthogonal to that is this vector is orthogonal to this and this orthogonal means inner product there is a correlation between this guy and this guy zero. Correlation between this guy and this guy zero

Here also, you see that you will then get only one such orthogonal projection it will exist it is unique and all that, which actually transform that basic vector space theory which I again I will repeat we will discuss later. We will see that I can easily show I will show in fact, here but if you really try to final linear combination and take that error that error will be orthogonal to this; so that orthogonally property will be satisfy this is called the this is the linear estimation, linear also orthogonal projection based estimation. In the case of filter instead of x 1 x 2 x 3 dot dot dot you have got the random variables x n x n minus 1 x n minus 2 up to x n minus say can be known whatever; it will be N x n minus capital N those random variables, where they that the successive once we have from correlation amongst them and use their mutual correlation and their correlation with d n to estimate d n as a linear combination of this.

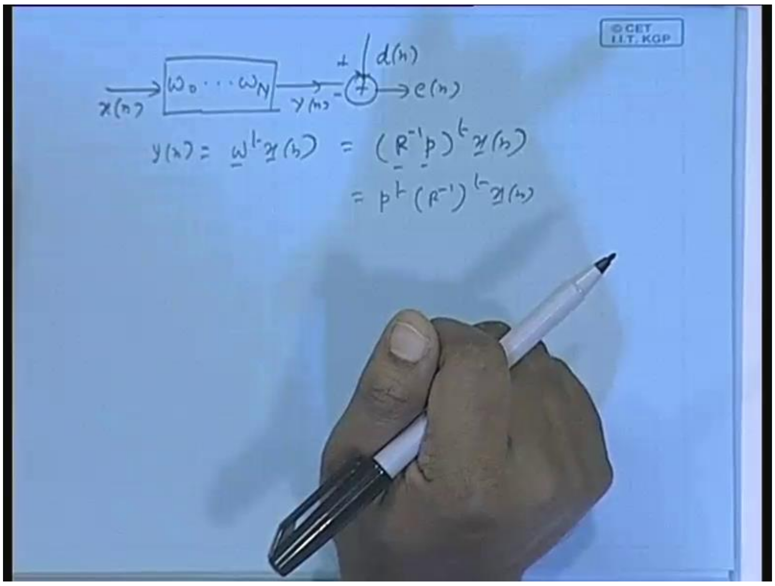
So, basically I am trying to project the random variable d n on a plane or a space span by x n x n minus 1 x n minus 2 dot dot dot x n minus capital N. So, n dimensional n axis as a linear combination of this those combine of coefficients of this. They will define the orthogonal based project this will be the orthogonal projection of d n; it will be a linear combination of those elements, this is the orthogonal projection the error will be orthogonal to each of the x’s x n x n minus 1 x n minus 2 up to x n minus n. That is if you take the really correlation between e n and x n e n and x n minus 1 e n and x n minus 2 dot dot dot it should be zero.

You will see that if you really choose this filter weights those correlations will be zero. No, why we choose linear estimation? Yes, why we choose linear estimation? So, linear estimation we have got all this theory; now in this orthogonal projection like this unique and everything and linear otherwise you can make it non-linear, but then there will be lot of complications in that convergence problem may come or implementation problem may come all those things. Linear we always in to be very honest we always prefer linearity once I mean from complexity point of view and also from our inability to handle the you know the enormity of the I mean maths involved, where you would be able to get solution and all that.

That is a that that is you know fundamental thing that books will be silent allowed nobody will that do, but i know whenever you have linear problem you can solve you can get close form results and they work in practice. But I am trying to give you the physical meaning of this linear estimation it is actually a projection of an unknown vector or target vector or target variable whatever or a plane instead of plane I could a space span by few axis few components x n x n minus 1 up to x n minus up to x n minus capital N. So, as a linear combination of them where the combine or coefficients of this you get y n y n should be the orthogonal projection of d n on that plane span by those fellows the projection error is this. So, must be orthogonal to them.

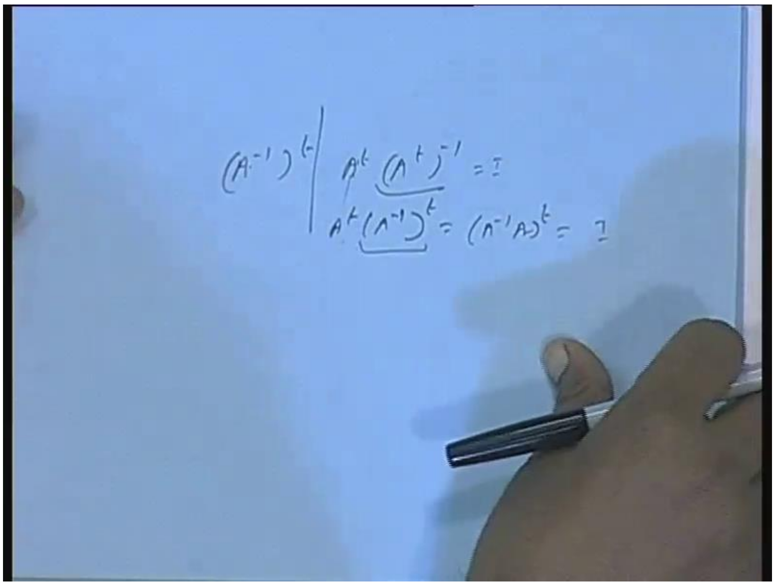
We will see now that they are in the orthogonal, but I am telling you I will again start from this few of the afterwards later I am will just giving a hint; so that you can understand why this filter k. Basically, I am doing nothing else, but the classical linear mean square estimation only whereas given a set of random variables you trying to find linearly combine them. So, that the combination becomes an orthogonal projection of that unknown vector on the space plane on the space span by those given variables. So, this is nothing but that that is why I wanted to say. We can now see on the orthogonal thing.

(Refer Slide Time: 19:21)



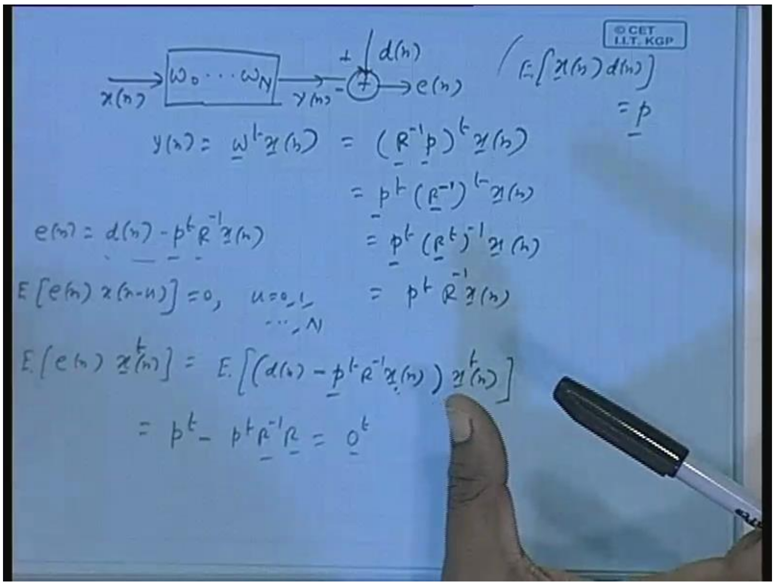
That is, what we have here w 0 to w n we have got x n we have got y n; I am repeating that figure for the sake of continuity d n e n. So, what is y n? y n is your w transpose I am dealing with real case here later I will deal with complex that time I may ask you to prove this is for the real case all transpose no Hermitian transpose; w transpose x n and I what is w transpose here? w is R inverse p optimal filter have already put R inverse p transpose x n and you know transpose of a product of matrices is what p transpose R inverse transpose. This is p transpose R inverse transpose and inverse and then transpose is same as transpose and an inverse you know or not how do you say that?

(Refer Slide Time: 20:41)



Suppose give a matrix is a square matrix A inverse transpose and I say A transpose inverse they are same how simple theory. Consider this, suppose I consider this case A transpose A transpose inverse then this should be I , but A transpose this also you can do I you can see, but A transpose has only unique inverse. If will take A transpose and this guy now what transpose, so A inverse goes here A goes here transpose; A transpose first A inverse here and A inverse is I transpose I. So that means, this fellow and this fellow both are the inverse of this, but this has the unique inverse it is square matrix if it is inverse is unique inverse. So, these two are same.

(Refer Slide Time: 21:46)



$E[\underline{x}(n)d(n)] = \underline{p}$

$y(n) = \underline{w}^t\underline{x}(n) = (\underline{R}^{-1}\underline{p})^t\underline{x}(n)$

$= \underline{p}^t(\underline{R}^{-1})^t\underline{x}(n)$

$= \underline{p}^t(\underline{R}^t)^{-1}\underline{x}(n)$

$= p^t R^{-1}\underline{x}(n)$

$e(n) = d(n) - \underline{p}^t \underline{R}^{-1}\underline{x}(n)$

$E[e(n)x(n-k)] = 0, \quad k = 0, 1, \dots, N$

$E[e(n)\underline{x}^t(n)] = E[(d(n) - \underline{p}^t R^{-1}\underline{x}(n))\underline{x}^t(n)]$

$= p^t - p^t\underline{R}^{-1}\underline{R} = \underline{0}^t$

So, this is p transpose R transpose inverse x n; sometimes some skipping the underscore underlined you know, but you should put them I might skip them that is my mistake. And R transpose is R where R is Hermitian real Hermitian here. So, p transpose R x n p transpose R x n R inverse x n R inverse x n. So, what is e n? e n is d n minus this guy. So, I have to show that en is orthogonal to x n orthogonal to x n minus 1 orthogonal to x n minus 2 dot dot dot dot. So, you should try to prove it an orthogonal means correlation zero. All are zero means correlation covariates means same from now onwards throughout the course no and all everything will have been zero mean.

So that means, you have to prove that E of expected value of these two. So, x n minus k equal to zero k could be zero one dot dot dot up to capital N this, what you have to prove? So, you will replace e n by this expression here and x n minus k multiply there may prove this equal to zero. Are you following me? If you try if you try to prove in that way I mean just do it will be difficult just look at this d n into x n minus k fine d n into x n minus k, but p transpose R inverse x n again into x n minus k x n is a vector.

So, this one component multiplied by the x n minus k; expected value expect value will not occur p or R because they are already constant not random; it will work on this you will bet a kind of clumsy real after that do what, but you see how clever your mathematics can be instead of doing it like this I say fine you want to do it for k equal to zero k equal to one up to k equal to N. So, why not do together? E of then we have we have to prove this equal to we have to prove that this is equal to zero zero zero zero zero zero right and the above into do that you get the result. Now, you substitute E of d n minus p transpose R inverse x n; some book have said foolishly they have done some elaborate maths and proved it for I am proving it in one and two lines.

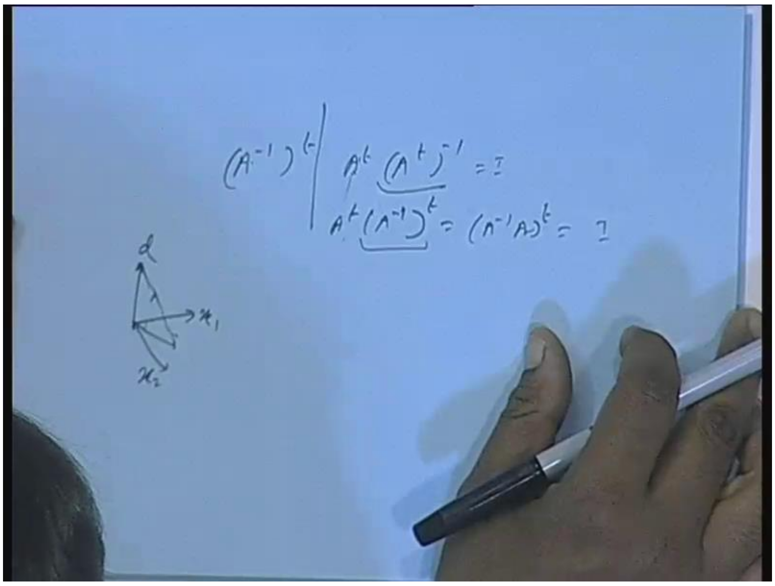
I am now putting e n I am replacing e n by this guy this is my e n, but I have brought back x n here. Instead of x n suppose I make a change here it is not an x n I make it x transpose n. I make it x transpose n that is e n into a row vector first element to x n that should be zero second elements x n minus that should be zero. So, result should be a row vector or zeros that is what I have to prove. For that combination I put x transpose here that is another thing and, so x transpose here. Now, consider this d n into x transpose n d n is a scalar and I given the definition E of x n vector into d n scalar is a vector p that I give. So, instead of x n if you make it row.

This is already scalar that is if you now take I mean mathematical aspect you need transpose of both side transpose you push inside d transpose n, which is d n because it is a scalar x transpose n. So, that will be p transpose. So, first guy is p transpose and you see the beauty is E will work on x n into x transpose n, which is the that is the advantage I get by bringing the total vector here. Instead of having just x n minus k a scalar here if you make it vector immediately I get R and p transpose R inverse there is R. So, p transpose minus p transpose will be zero zero row vector. In fact, you should write zero transpose all my vector are column vectors row vector is always transpose.

So, you can see by one stroke I am very showing some books you know where foolishly they have really carried out this thing vehemently for various case and you know that finally, we prove that is zero where this can be defining this. Just instead of taking what I am trying making it a vector you get transpose there that is all immediately it will come. From this, you get an idea that orthogonality remember either vector space or no vector space, but one thing is there e n is uncorrelated with x n x n minus 1 up to x n minus capital N.

I give the physical interpretation of projection and all that which is the case actually that is the basis of all estimation every filter all these estimation things. But about that rigorous treatment will be take up later, without that I cannot proceed further in the topic up to some point I can I am just giving a be I idea before end. The projection is always it means I mean this filtering always means projection.

(Refer Slide Time: 27:49)



$(A^{-1})^t$

$A^t \underbrace{(A^t)^{-1}}= I$

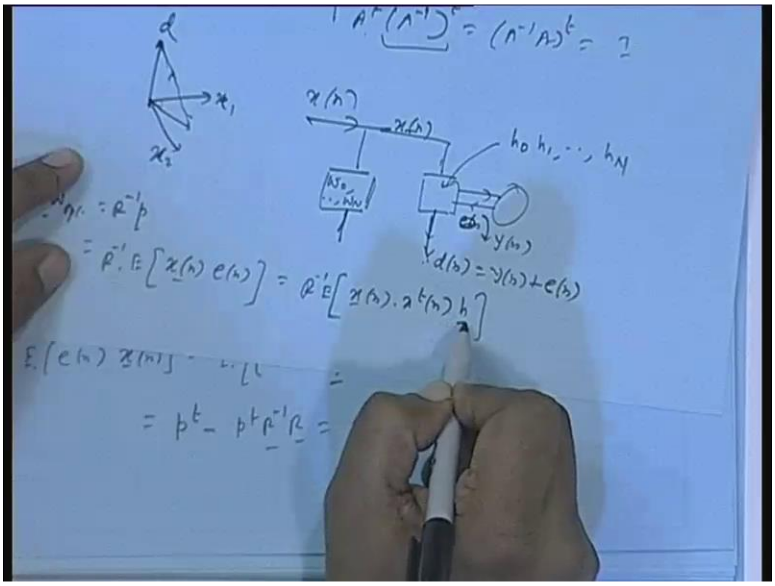
$A^t \underbrace{(A^{-1})^t} = (A^{-1}A)^t = I$

So, from this you can then get get an idea that suppose the vector that you are projecting it is d it is already orthogonal to e 1 e 2; e 1 e 2 may not be orthogonal, but suppose d is already uncorrelated with e 1 e 2 correlation that is p vector is zero vector; p vector there is zero vector because of correlation between d n d n and x n d n and x n minus 1 and dot dot dot zero. But I am dealing with more general case. Say, x 1 instead of x x 1 2 it is given that correlation between d and x 1 or d n x 2 is 0. So, d has to be in this direction. In this case from your physical idea what is will be the linear combination.

So, that the error has minimum norm square zeroes only; otherwise the moment you take anything so anything. This would be the difference and that norm square will be greater than norm square of this. So, projection of this vector on this plane is zero only. If this is perpendicular if you take z axis in a x y plane and project orthogonally z axis onto the x y plane you pit at the origin by the projection and the entire this vector this is the error. So that means, even this filtering frame work it is giving that d n is something, which is totally uncorrelated to each x n and all the samples of x n; that is p vector is a zero vector.

Then the base combination should be what all zeroes and that best estimate will be zero that also comes from that formula. Because formula is R inverse p p is a zero vector. So, based weights are zeroes, but physically this is what it means? Example I have proven this result. No, this situation will arise often this situation will arise often.

(Refer Slide Time: 29:57)



You consider that an eco-consider program remember eco-consider program there was a signal x n; so you are asking this is a hybrid, we modeled as a FIR filter linear FIR filter. So, output of the high actually the signal was to go this way and this fellow signal there is a speaker. This should go this way this should go this way this should come up. But what is happening in this sequence is x n; what is coming out is not just difference this guy speed there is y n. But y n plus part of x n; part of x n e is en. So, what is coming out if you call this d n d n is y n plus e n or e n and x n are correlated because e n comes from x n; e n is correlate with that.

But y n has no correlation with y n x n; two different speaker x n is coming from you, y n is coming from you know correlation uncorrelated suppose zero mean we subtract that mean. So, if I put an optimal filter here to estimate this what will be your filter? Contribution in the filter weight from y n part should be zero the p vector p vector is what if I take d d n as this d n as a the desired stress response. The expected value of d n times an x n vector. So, d n means y n plus e n. So, y n times x n vector expected value of that will be zero because of uncorrelated thing.

Only this part will be there e n into x n vector. So, what will be the filter weights here? Suppose here you have got this coefficients h 0 h 1up to say h N and here we want to find out w zero dot dot dot w N. What will be the optimal once optimal one w opt here will be same as earlier R inverse p, but what is R? We know E of x n into x n transpose.

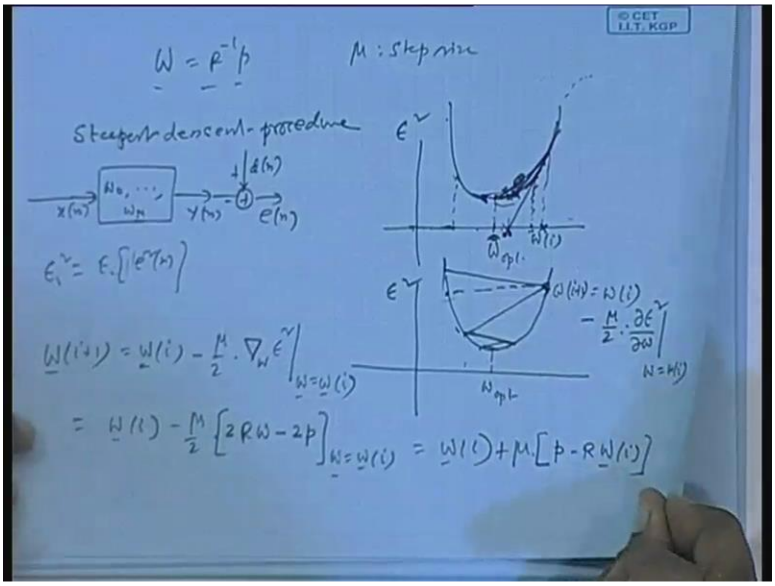
What is p here? p is E of x n vector into e n and what is e n? e n is the output due to x n mind you n is output due to x n. So, basically if you have h vector as h 0 h 1 dot dot dot h what N. Then, h transpose x n that is n or x n transpose h. This is E of x n remains as it is and e n e n is h transpose x n vector or alternatively x transpose n h vector. And h is constant it can go outside expectation this is R R inverse are cancels.

So, you get back your h. So, then optimal weights will be h if you train it by some active algorithm and that converges or that tries to aim at the optimal weights by that adaptive methodology what will hit upon is this coefficient. So, ideally the same e n will be generate same e n will be generated that much will be subtracted you cannot find out y n, y n best estimate will be zero because y n is orthogonal to all the x n. So, projection of y n on the planes spans by a space span by x n x n minus 1 up to x n minus capital N is zero. It is like projecting z axis on x y plane.

Let me plane this as examples you know. So, I got some idea about the filter actually that why this filter because we are essentially doing nothing else, but linear mean square estimation. This is the general linear mean square estimation frame work that is computation of orthogonal projection. We will take this up later more rigorously by without diluting anything. So, far the discussion was based only on analogy. I took the analogy of three dimensional vectors and L dot product orthogonality and said that similar thing exist there also that analogy I used.

But analogy will disappear I will do this things rigorously without bringing any analogy or any with on the physical planes and all that later; guys actual vector space development of base development of filtering an estimation. In fact, this all linear estimation is computation of orthogonal projection and the actual framework for doing that will be vector space framework. Why you have got an idea about how the vector space things comes, but we need a more general notion of vector space here. Analogy will not work the analogy I am just trying to make you we will comfortable with that.

(Refer Slide Time: 35:08)



$\underline{W} = \underline{R}^{-1}\underline{p}$ μ : step size

Steepest descent procedure

$\epsilon_1^2 = E[|e^2(n)|]$

$\underline{W}(i+1) = \underline{W}(i) - \frac{\mu}{2} \cdot \nabla_W \epsilon^2 \Big|_{\underline{W}=\underline{W}(i)}$

$= \underline{W}(i) - \frac{\mu}{2}\left[2RW - 2p\right]_{\underline{W}=\underline{W}(i)} = \underline{W}(i) + \mu\left[p - R\underline{W}(i)\right]$

$W(i+1) = W(i) - \frac{\mu}{2} \cdot \frac{\partial \epsilon^2}{\partial w}\Big|_{W=W(i)}$

Anyway, so we know your W is R inverse p suppose R is giving to you p is giving to you, I say look you know I do not know how to compute inverse of a matrix. I have no idea I know how to compute, I mean I do not know how to compute how to go about it. Suppose, when I can proposed that instead of a close form solution like this may be I can propose an iterative procedure, which at the end of iteration may be we will converge on this. What could be the iterative procedure; so that procedure is called steepest descent, steepest descent procedure. We have seen again I redraw the figure this is w 0 dot dot dot w m y n this, what we are very much familiar with an I think I have drawn it ten times or fifteen times I do not know how many times I drawn.

So, far but anyway this is my reference. So, I have to draw it. We have been seen that epsilon square, which is expected value of this this is a quadratic function of the filter weights. Any now you consider and this real function we know these are complex valued I mean I would have then put a mod here. So, variance is real. So, it is a real function of filter weights, but what kind of function? Quadratic function. Now, if you see a quadratic function of one variable or many variables it has only one unique minima or maxima suppose we are dealing with the case of minima.

So, if it is only unique minima then you start from the minima point and go further away give you the extend you can only see the function going up and up and up it you can never hit a situation where the function after rising for a while start again falling, because then there will be point of inflection minima or max local maxima will result which is not possible. So, usually this will be like a bowl shaped curve. For example suppose I have got only one tap suppose I have got only one. Tap is one what which is some sometimes often used for filter weight or filter coefficient tap.

Suppose, I have got only one tap or one coefficient or one weight say w not w 0 not w 1 not w N, but only one weights w. So, epsilon square will be a quadratic function of w there. So, suppose it should be on it can only have the minima here say, w cap the w opt scalar and this can only from here I mean may be at this point it has this much value so if you go to the right or to the left this can only go up and up and up something like this. It cannot have something like this you know it cannot fall the result will local maxima will come up you understand it can only go on increasing.

You cannot have a situation where it again starts falling then it has to be cross where a point of intersection here, which means derivative will be zero there will be a local maxima. But that is not possibly with quadratic function the moment you derive it and equate to zero you get a first order equation you get only one solution for minima maxima problem that we all know. So, that cannot happen if that means, so then my proposal is that suppose I am running an iterative procedure at i th step of iteration I have got some value of the weight will change in a iterative manner and will progressively take you to the optimal one that is my game.

So, at i th step of interaction i can be zero or one or two or anything. Suppose you are here w i, then the strategy that I should follow is this at this point I look at this function. Find out its gradient? If their gradient is positive some of you are familiar with this thing in your in numerical analysis similar thing is done in new term epsilon and all that. If your gradient is positive no point in going for the right that is because it will be further away from the minima; so you should going the opposite direction. So, the gradient is positive next iterate will be what to the left that is from the given one I should subtract some number.

So, I should look at this on the other hand if were here gradient is negative; then I should not go to the left for to this I will add something; that means, from i th iterate I should always subtract the quantity proportional to the gradient. Subtract a quantity proportional to the gradient or other sign of the gradient, but normally I take the proportional with the gradient if the suppose it is w i, I take the gradient is positive. So, positive number and it is subtracted. So, definitely I will go to this side, but if I am here again I say from w i; I will subtract a quantity proportional to the gradient, but gradient is negative. So, basically I will be adding. So, I will go to the right is that all right?

I am saying from the current iterate I will subtract a quantity proportional to the gradient; if the gradient is positive I will indeed subtract positive. So, subtraction, but the gradient is negative as here we subtracting a negative continuous addition. So, I will add this I will go further that will be better and again I am not only taking the sign, I am taking the gradient. So, I am taking the magnitudes of the gradient. So, quantity there is proportional to the gradients means if the gradient is stiff. So, the quantity which I will subtract from the iterate that will have higher magnitude.

I will jump by a huge margin and although there is a gradient is small positive, but small then I will jump back I will go back by a smaller margin. So, then what I am aiming at is something like this suppose I am here suppose I am here. Let me use a different color there may be i two go here. I was here I found the gradient is negative I am quite stiff. So, from w i I subtract a quantity proportional to the proportionality constant is constant it proportional to the gradient. That means gradient was positive. So, I hit somewhere here then again, I find out the gradient sill gradient is positive, but magnitude is much less.

So, again I subtract, but this time I go back by a little margin. Suppose I go back this way this much. Here, I find gradient is negative but again magnitude is not very high, so I go for what by a small margin. So, this way back and forth back and forth that I converge on the optimal one; it is something like this you knows I am drawing the figure again; so it becomes a bit clumsy. Suppose, you are here may be you go here at this point you found the gradient is positive you subtract; so much that you cross this and went to the left. This diagram will be easier to you know will be use more useful to explain.

Then, here you find the gradient to be having some magnitude and negative. So, basically you will add. So, you will go to this side by some amount may be here. Then, again here you find the gradient positive. So, you subtract so we will go somewhere here then here like that. So, finally, you should converge on this; that means from the current weight w i I will go the next weight i is the i th stiff of iteration. So, the next iterate will be what current one minus a quantity proportional to the gradient. This is epsilon square a quantity that proportional there is a proportionality constant that constant is mu, but I take it mu by two because I want this two two cancel some two that will come up.

So, mu by two is a proportionally constant that times is gradient well gradient is evaluate at this that point w equal to w i this is by algorithm, this is called steepest descent algorithm. This is a gradient evaluated at that point w equal to w i, I am subtracting a quantity proportional to the gradient. So, this is the proportionality constant; mu is called the step size mu is the most important parameter, we will be with you throughout this course step size why this. So, important you know if you do not use carefully what can happen you see this, what can happen is this if you if you do not choose carefully you can hit here and then from here you can hit here hit here hit here like this never going down.

What it was if you still mu higher you go like this go up go up go up go up. So, actual for obviously, you can see there should be some upper limit of mu you have to work within that limit then only it will fall and will hit that those limits and we will work out later. So, this is for one variable case that is as though the filter had only one tap w, but now I generalized it vector w i plus 1 is remember. Now I have got all the N capital filter weights; mu is a constant that is why it is a proportionality constant. Constant in every step of iteration and even you have multiple weights constant for all the weights same for all the weights then there are there are other versions, where you know for each separate where you give mu separate mu those are more complicated cases.

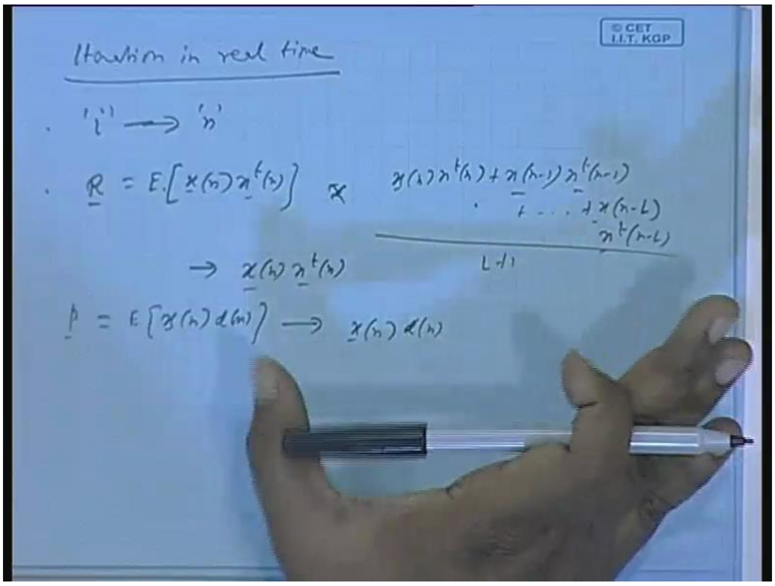
So, w i minus mu i two then i have to do so much of composition I have to find the solution and all that why this point why. How do I choose mu no that there is a theory are you prove convergence then it is very easy. It will come in terms of some correlation figure of the input; that will take you somewhere else you are finding a linear equation you have to find the equation of the line when it cuts either you can do it. Now in N dimension, which will two complete this will not be lying on your plane and all that I do not know what you are aiming at.

So, mu by two into no single gradient, del you remember del is a vector of all partial derivatives that will do the job and this we evaluated yesterday; del w epsilon square and w equal to w i you all know what we came out with last yesterday that figure if you put what was that figure. There was a two there you see there was a term w transpose p and then give rise to p on differentiation by w and there was a term w transpose R w that gives rise to two R w something like that. So, you had twice R w minus twice p and w equal to w. Please check what we found out twice R w minus twice p that is what we found out yesterday I have got w you have to put w equal to w i and you see that two is coming out that is why i put a two here, so that two two go.

So, two will cancel and what you get then is w i plus mu into p minus R w i this is the algorithm this is called steepest descent algorithm. We have to be choose just physically understood that there should be an upper limit of mu there is a procedure to occur. But that is not my concern here this is an algorithm and no matrix inverse is required you of course, know the need the knowledge of p and R, but no matrix inversion is required, but still it is an of line procedure. We have been given the matrix R and p and you are running this sitting at home ideal not using real time data one after another p and R giving to you are just running an iterative algorithm off line.

Suppose R to be the same thing is iteration I want to do in real time. So, that my i th iteration index is my n th clock cycle I am look I have a I am switching on a clock zero th clock cycle at that time I will carry out zero th iteration or first clock as for my clock; I will carry out first iteration same iteration.

(Refer Slide Time: 49:48)



Iteration in real time

- 'i' → 'n'
- $\underline{R} = E[\underline{x}(n)\underline{x}^t(n)] \approx \frac{\underline{x}(n)\underline{x}^t(n) + \underline{x}(n-1)\underline{x}^t(n-1) + \ldots + \underline{x}(n-L)\underline{x}^t(n-L)}{L+1}$

$\rightarrow \underline{x}(n)\underline{x}^t(n)$

$\underline{p} = E[\underline{x}(n)d(n)] \rightarrow \underline{x}(n)d(n)$

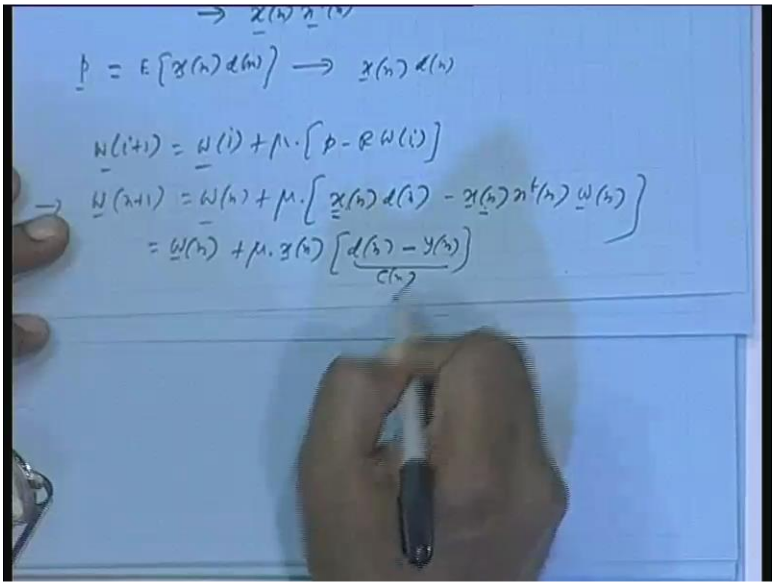
How to that is iteration in real time same iteration. so I am carrying out the iteration over i, but that time i was zero i was one i was two dot dot dot zero th first second third fourth term iterations and all that. I am doing the same iteration, but looking at my clock or some clock. So, iteration is taking place in time so that means, I replace i th term i by n same iteration, but instead of calling i th iteration I have say n th iteration no problem, but n is a real clock cycle samples sample cycle sample index you can also say that is why I am doing the iteration.

Secondly, you had a matrix R you had a matrix R what is R? R is E of x n in the our case transpose x suppose, you really want to have a good idea about R what you should do you take one vector x n multiplied by x transpose n take x n minus one vector multiplied by x transpose n minus one vector and go on adding for say may be hundred such cases divide it hundred you get a good estimate. So, we have statistical averaging you take like you know I mean x n x transpose n plus x n minus one x transpose n minus one dot dot dot say x n minus L x transpose n minus L divided by how many here L plus one. This will be a good estimate. But, suppose I want to replace it by a good estimate I am like a mad man or whatever crazy man.

I say that I do not want to estimate to you good I will take only one element then we add estimate. If you if I suppose, so I am measuring a quantity random variable I should take several cases and then add up I take the average, but if you take only one; then will be wiener estimate because that variance of that will go up because every time a sample is changing is changing you get this estimate or this value or that value, so but suppose I say that fine still; however, crude I will live with it. So that means, I will replace it by simply x n x transpose n.

Similarly, p is x n d n. So, again you should have x n d n plus x n minus one vector into dn minus one plus dot dot dot up to say L times add an average, but suppose again like a mad man I replace it just by one. I am saying that however, could the estimate is suppose we work with this estimate. Finally, in the end if i we can so that things will work with this, then I am things indeed work with this because you know the filter in the iterative procedure x n past in the past case; it will be x n minus one past an internal averaging will take place in a iterative procedure that we will see.

(Refer Slide Time: 53:04)



$$\rightarrow \underline{x}(n)\,\underline{x}^t(n)$$

$$\underline{p} = E[\underline{x}(n)\,d(n)] \rightarrow \underline{x}(n)\,d(n)$$

$$\underline{w}(i+1) = \underline{w}(i) + \mu\cdot[\underline{p} - R\,\underline{w}(i)]$$

$$\rightarrow \underline{w}(n+1) = \underline{w}(n) + \mu\cdot[\underline{x}(n)\,d(n) - \underline{x}(n)\,x^t(n)\,\underline{w}(n)]$$

$$= \underline{w}(n) + \mu\cdot\underline{x}(n)\,[\underbrace{d(n) - y(n)}_{e(n)}]$$

So, if I replace this in that steepest descent algorithm what will that be steepest descent was just one more minute I will finish and i will start from here in the next class. It was this p minus R w i. So, now, this will become w n plus one I replace by n I am doing real time I am doing it as a real time plus mu this is x n d n and this x n x transpose n w n; R is x x transpose w i means w n this is x n d n. Now, take this x n vector out common just one more minute this x n vector common. So, what you get d n here please see this I am going fast d n minus this x n has come out. So, only x transpose w, but x transpose w is what same as w transpose x that is the filter output at n th index; you have got a filter vector that transpose x or x transpose w both are same that is a filter output at what n th index.

So, d n minus y n and this is that error e n this is equal to e n. So, you get this quantity. This is the most famous adaptive filter algorithm this becomes now adaptive because I am working real using real time data x n and d n and all that only I am doing the iterative procedure; so iteration in in real time using the data. No, R matrix no p value vector; so very it converges gives you satisfactory result does it separate, but which we will see, this is called least mean square LMS algorithm. I will talk more about this I will just give a basic derivation. So, I will start from here in the next class.

Thank you very much.