

**Introduction To Adaptive Signal Processing**  
**Prof. Mrityunjoy Chakraborty**  
**Department of Electronics and Electrical Communication Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture No # 14**

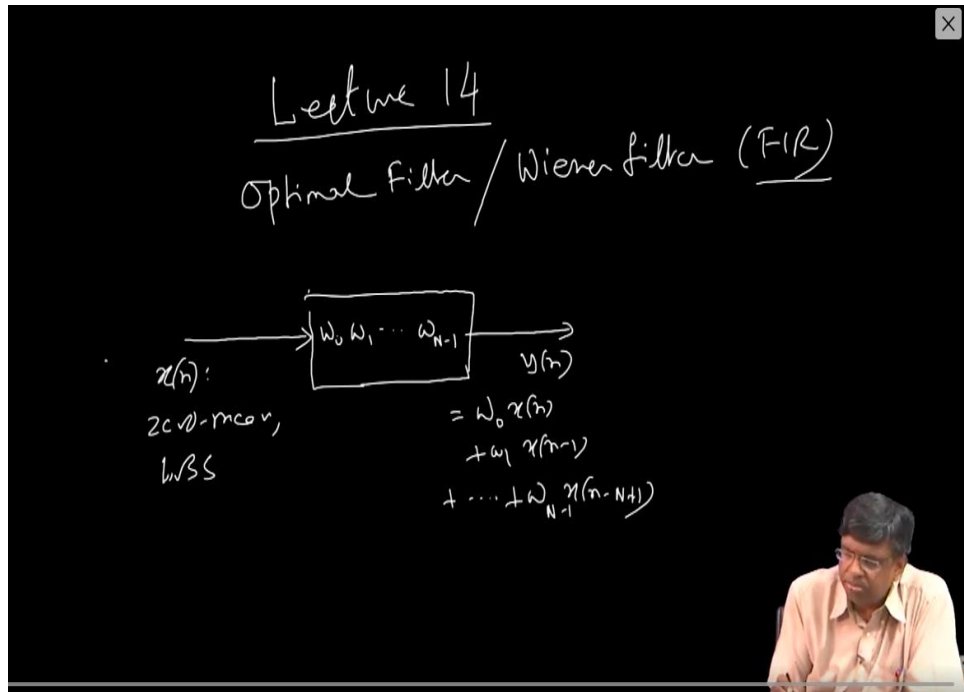
**Optimal FIR Filter**

So, we begin this lecture where we bring a new topic now. And this new topic will be using all the material that we covered earlier. Based on that, but this is the key topic in adaptive signal processing which is optimal filter. Also called Wiener Norbert Wiener the famous man and this is FIR filter. Essentially because IIR filter has this additional requirement of maintaining stability and causality that the poles must be within unit circle, but so far whatever research has gone into it and lots have gone into it over 50 years. I mean there is no guarantee nobody could give a guarantee that the filter that they will derive will have you know no stability issues and all that ok.

And ok even if optimal IIR filters are there going to adaptive IIR filter from there is impossible because when it is adaptive IIR filter there is no guarantee that you know we will be able to ensure its stability and causality more about that later. Let us consider FIR filter first. Here basic structure is this there is a random process which is maybe 0 mean WSS. Probability density, joint density all those things can be of any type that we are not you know doing here.

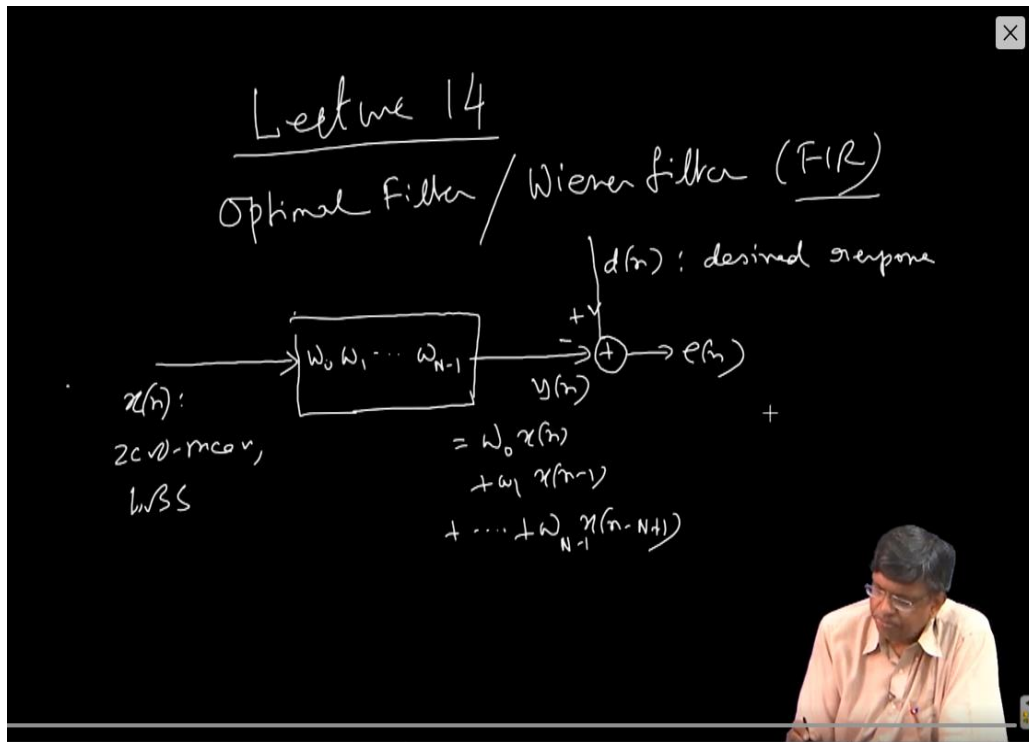
And you want to pass it through a filter FIR filter of coefficients  $W_0, W_1 \dots W_{N-1}$  and filter output is nothing, but a linear combination of  $X_n$  and then  $X_{n-1}$  there is  $W_0 X_n + W_1 X_{n-1} + \dots + W_{N-1} X_{n-N+1}$  this all right. This is a linear combination of current input and  $N-1$  capital  $N-1$  past input they are linear combination multiplied by coefficients and added. These coefficients

are so called filter coefficients. They are filtered because they work on current data and  $N$  minus 1 previous data and continuously for every  $N$  that is why it is a filter.



But this purpose of this filter is one thing that is there may be some target response  $dN$  which is a desired response.

Sometimes called target signal, training signal and all those. Purpose is to see that you design the coefficient such that  $Y_N$  is a very good estimate of  $dN$  all right. So,  $Y_N$  is possibly the best estimate of  $dN$ . So, that means if I take the error between the two plus here minus here call it  $E_N$ .  $E_N$  should have minimum strength all right.



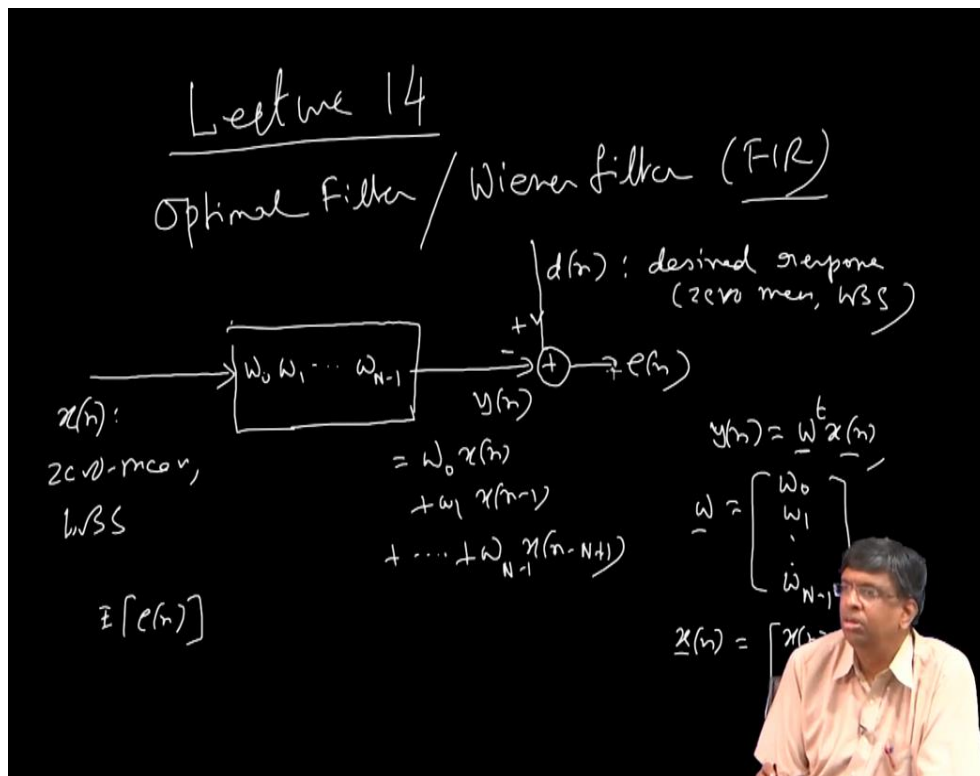
Now certain notations this  $Y_N$  you can write as  $Y_N$  you can write as some  $W$  vector transpose  $X_N$  vector where  $W$  vector is  $W_0, W_1$ . So, these coefficients put in a vector form  $X_N$  see the underline it means a vector. It is a current data vector sometimes called regression vector it is  $X_N$ . You understand  $W$  transpose  $X_N$  will be what this is a row column vector. So, after transpose it will become a row vector and then multiplying.

So,  $W_0, X_N, W_1, X_N$  minus and dot dot which is what this. Now one thing you can say that why not design the coefficients where  $d_N$  equal to  $Y_N$  that you can do if you create  $d_N$  to this right hand side and there are unknowns  $W_0, W_1, W$  capital  $N$  minus 1. So, capital  $N$  number of unknowns data is known  $X_N, X_N$  minus 1 this data is known. So, this is just one equation right hand side is  $d_N$  which is known and left hand side is this linear combination where  $X_N$  known  $X_N$  minus 1 known and all that  $W_0, W_1$  dot dot dot  $W$  capital  $N$  minus 1 they are unknown. So, one equation so many unknown.

So, you have infinite solution you can pick up any solution put that here. So, then  $Y_N$  will be equal to  $d_N$  and you are very happy  $N$  is 0, but relax that will not So, then  $Y_N$  will be

equal to  $d_N$  and you are very happy  $N$  is 0, but relax that will not work. Because once you design the coefficients that if you put here. Next time at  $N$  plus 1th clock it will be  $W_0 X_N$  plus 1,  $W_1 X_N$  and dot dot dot that linear combination there is no guarantee it will be satisfying  $d_N$  plus 1. If at that time mirror margin could be very huge.

So, that is why this is not a procedure. So, procedure is to design the coefficients so that in an expected sense average sense  $E_n$  has minimum strength again you cannot minimize expected value of  $E_n$ . Remember  $E_n$  is  $d_N$  minus  $Y_N$ ,  $d_N$  is a random sequence this is also WSS 0 mean.  $E_n$  is random because  $d_N$  minus  $Y_N$   $d_N$  is random  $Y_N$  is obtained from the input samples and input samples are random  $X_N$  is a random variable  $X_N$  minus 1 is a random variable so on and so forth. So,  $Y_N$  is a random variable  $d_N$  is a random variable so difference is a random variable.



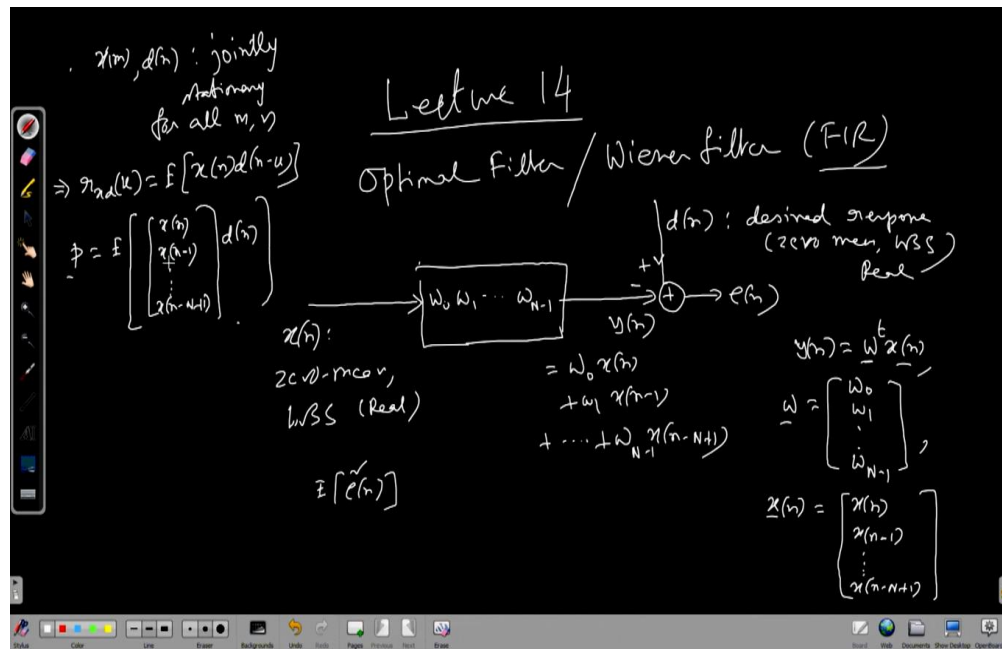
But you cannot just minimize this because that does not mean that error will even if you minimize it does not mean error will be less because  $E_n$  is 0 mean. So, sometimes it will go positive sometimes negative. So average could still be very minimal 0 maybe. So

expected value of this could be minimum possible maybe 0, but that does not mean that this is not varying over a wide range which is basically which does not serve any purpose. I want its power rather I should have this  $E[n^2]$   $E[n]$  is real here we are dealing with real case we have been WSS real this also real.

So, that is why  $E[n]$  is real so we just take square of it there is no complex. So  $E[n^2]$  is the power average and then expected value of that is the average power which is the variance here and that is fine because whether  $E[n]$  goes positive or negative  $E[n^2]$  is always positive and expected value that will give you the power. So,  $E[n]$  is power if it is less that means power of  $E[n]$  is less means its variance is less. That is around mean it will vary over a smaller range the increments around the mean which is 0 will not be having large spread because then power will be you know average power will go up. If it is spread if it is variation around the mean is small there are in an expected sense this also will be less average power.

So, therefore, we will minimize this. Now certain things are given to us here one is  $x_n$   $d_n$  they are jointly stationary  $x_n$  minus sorry  $x_n$  minus  $m$  and  $d_n$  jointly stationary, maybe, I rewrite it,  $x_m$  let me put it  $m$   $d_n$  they are jointly stationary for all  $m$   $n$  meaning and they are real meaning if you take the correlation between  $x$  and  $d$   $r_{xk}$  that is  $e$  you can take any index here  $n$   $d$   $n$  minus  $k$  no star here because all are real gap is  $k$ . So, it should depend only on  $k$  even if one sample is  $x$  another is  $d$  the correlation between them depends only on the gap. So, it is called cross correlation because one is the random process  $x$  another is  $d$  both are not  $x$ . So, not auto correlation, but cross, but even in the cross correlation that depends on the on the gap  $k$  then only you say they are jointly stationary and that is what is given to us number 1 ok.

In fact, what is given to us is this vector  $p$  where  $p$  is  $e$  of this data vector this is the data vector times  $d_n$ .

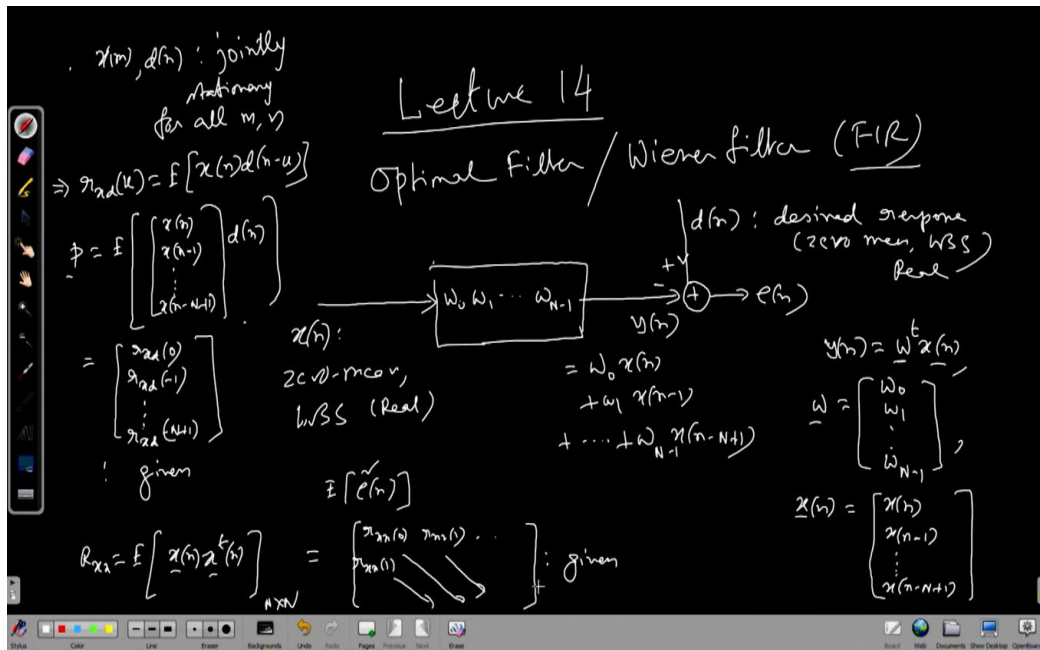


So,  $x_n d_n$  will be  $x_{n-d_0} x_{n-1} d_{n-1} d_n$ . So, gap is  $n-1$  minus  $n$  gap is minus 1. So, it will be  $x_{n-d} x_{n-d-1}$  so on and so forth. So, it is basically  $x_{n-d_0} x_{n-d-1} \dots x_{n-d_{N-1}}$  last one is  $n$  this index minus So, it is basically  $x_{n-d_0} x_{n-d-1} \dots x_{n-d_{N-1}}$  last one is  $n$  this index minus this index so small  $n$  cancels out so minus capital  $N$  plus 1.

So, this is given because we have, we must be given some information. So, exact  $x_n$  exact  $d_n$  we are not given neither are we bother about them because just one observation of this sequence or that sequence does not matter because they will change in the next time, but statistical properties is given that is the cross-correlation values between this  $d_n$  and  $x_n$  at least for this many for 0 lag or minus 1 lag lag or gap minus capital  $N$  plus 1 gap and at least that is given to us ok. That is one information given to us another thing that is given to us is  $r_{xx}$  and we have studied this earlier  $x_n x_n^T$  just transpose not Hermitian transpose because conjugation has no meaning here because we are for simplicity we are considering real case later we consider  $x_n d_n^2$  complex case of course, ok. So, right now all are real  $x_n d_n$  and therefore, the filter coefficients are not real. This  $x_n x_n^T$  we have already seen earlier it takes a top plate structure and it is a Hermitian matrix it was

like this  $r_{xx} 0$  and this continues  $r_{xx} 1$  this continues and this symmetric matrix now not conjugate symmetric.

So,  $r_{xx} 1$  again will come here continues and dot dot dot it is like this. So, this is also given the auto correlation matrix of order, you know if the length is capital N, So, it is basically N cross N this is given ok.

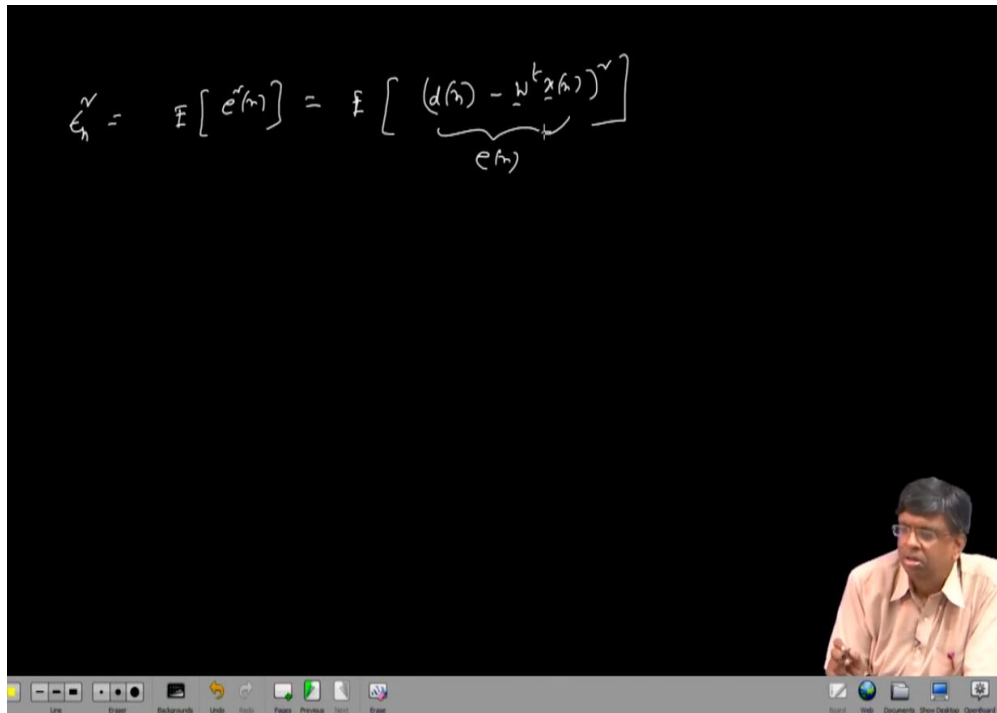


So, this information auto correlation matrix is given for input process  $x_n$  at least up to the size capital N cross capital N and cross correlation values at least this many these values ok that is given we have to now design the filters. So, that the variance of the error is minimized That is what we have to do is E of E square n and let me give it in N epsilon square I could put a you know index N because after all the it is a function of N here, but very soon we will see that because of stationarity this entire function will be free of the index N and that is why I am not writing any N here in advance because I know what it will be. So, known N here, but to start with you could put a N or if you want I can put a N here later I will erase it and this is the variance this is basically what is.

So, square of what is  $e_n$ ?  $e_n$  means desired response minus the filter output and filter

output is  $y_n$   $y_n$  was  $W^T x_n$ . This is your  $e_n$  and you have to take square of it and then the expected value. So, it depends on this filter coefficients present in this vector is a filter coefficient vector. So,  $W_0$  is a function of filter weight  $W_0 W_1 \dots W_{N-1}$ . So, capital  $N$  number of variables is a function of, So, I have to minimize it.

So, I have to take partial derivative with respect to  $W_0$  equal to 0 with respect to  $W_1$  equal to 0 so on and so forth and by that process I have to find out  $W_0 W_1 \dots$ .



$$\epsilon_n^2 = E[e_n^2] = E\left[\underbrace{(d(n) - \sum_k W_k x_k(n))}_{e(n)}^2\right]$$

That is called the optimal filter for which this will be the variance will be minimum. Let us see one thing if I expand it now certain tricks, I will apply this is a scalar any scalar and its transpose they are same because scalar is nothing, but a 1 cross 1 matrix. So, I can if I take a transpose of it I get by the same thing because 1 cross 1 matrix transposition is the same. So, I can write it as  $e_n$  first that is  $d_n$  minus into again  $e_n$ , but that  $e_n$  I will write as transpose of  $e_n$  transpose of  $e_n$   $e_n$   $e_n$  transpose  $e_n$  transpose is  $e_n$  itself.

So,  $e_n$  into  $e_n$   $n$  square. Now you break it  $E$  of  $d_n$  and  $d_n$  transpose  $d_n$  transpose is  $d_n$ . So,  $E$  of  $d$  square  $N$  which will be the  $d_n$  is 0 means, So, this is the variance of  $d_n$  and  $d_n$



is Wss, So, variance does not depend on N. So, variance of  $dN$  minus  $E$  of  $W$  transpose  $xN$   $dN$  ok.

$$\begin{aligned}
 e_n^v &= E[e_n^v] = E[(d(n) - \underline{W}^T \underline{x}(n))^v] \\
 &= E[(d(n) - \underline{W}^T \underline{x}(n)) (d(n) - \underline{W}^T \underline{x}(n))^v] \\
 &= E[d(n)] - E[\underline{W}^T \underline{x}(n) d(n)] + \underbrace{\quad}_{\text{Var of } d(n)}
 \end{aligned}$$

Then other two terms minus if it is  $dN$  if it is  $W$  transpose  $xN$  then  $dN$  transpose next is  $dN$  transpose  $W$  transpose  $xN$  next if it is  $W$  transpose  $xN$   $dN$  transpose next is  $dN$   $W$  transpose  $xN$  transpose ok. It is like you know separately if you call it  $A$  in general if you have something like you know a vector  $A$  it is a scalar here, but in general a vector  $A$  plus a vector  $B$  suppose again  $W$  transpose  $xN$  is a scalar, but I am doing a more general thing suppose  $A$  plus  $B$  or maybe this I will do later forgive it.

You can see one thing  $W$  transpose  $xN$   $dN$  alright I will write the other term later. Now here what does it mean  $xN$  into  $dN$   $xN$  consists of all the random variables and  $dN$  also random variables. So, product is a random variable and  $W$  transpose times that. That means, if you quality a vector may be random vector  $hN$ . So, you have got the elements of  $hN$  because because of course,  $xN$  is a column vector  $dN$  is a scalar.

So,  $dN$  times every element of  $xN$  and that you get  $hN$  and  $hN$  is random because  $dN$  is random and all the elements of  $xN$  are random. So,  $hN$  now  $W$  transpose row vector times

$\mathbf{h}_N$  column vector what we will have  $W_0$  times the first guy here  $W_1$  time second guy here and dot dot dot added then expect it. But the expectation that can be pushed inside that summation we have seen the expectation is linear. So, expectation of a summation of variable random variable means it is a summation of expected value of the value of each of the random variable and this random variable will be some  $W$  term from here and some term from here, but this is constant not random. So,  $E$  will come directly on the term from here.

So, it will be  $W_0 E$  of first component of  $\mathbf{h}_N$  plus  $W_1 E$  of second component of  $\mathbf{h}_N$  and dot dot dot which is which means it will be  $\mathbf{W}^T$  then  $E$  of  $\mathbf{h}_N$ . Then you have got the first component there is  $W_0$  times the first component of  $E$  of  $\mathbf{h}_N$   $W_1$  and like that and  $\mathbf{h}_N$  is  $\mathbf{x}_N \mathbf{d}_N$  which is given to be cross correlation vector  $\mathbf{p}$ . In the previous page you have seen. Similar manner the other term is  $\mathbf{d}_N \mathbf{W}^T \mathbf{N} \mathbf{x}_N^T$ . Now these two terms are same because this is row  $\mathbf{x}_N \mathbf{d}_N$  if you call it  $\mathbf{h}_N$ .

So,  $\mathbf{W}^T \mathbf{h}_N$  and here it is  $\mathbf{d}_N$  this is nothing but  $\mathbf{d}_N \mathbf{x}_N^T \mathbf{N} \mathbf{W}$  alright. So, this is if  $\mathbf{h}_N$  is  $\mathbf{x}_N \mathbf{d}_N$  what is  $\mathbf{h}_N^T \mathbf{N}$ ? This is  $\mathbf{h}_N^T \mathbf{N}$  alright because  $\mathbf{h}_N^T \mathbf{N}$  means  $\mathbf{d}_N$  is a scalar it is like a 1 by 1 matrix. So,  $\mathbf{d}_N$  will come in the front  $\mathbf{d}_N \mathbf{x}_N^T \mathbf{N}$ , but  $\mathbf{d}_N$  transpose is  $\mathbf{d}_N$ , because it is a scalar after all and  $\mathbf{x}_N^T$  transpose. So, this is  $\mathbf{h}_N^T$ .

$$\begin{aligned}
 e_h^v &= E[e_h^v] = E[(d(n) - \underline{W}^T \underline{x}(n))^v] \\
 &= E[(d(n) - \underline{W}^T \underline{x}(n)) (d(n) - \underline{W}^T \underline{x}(n))^v] \\
 &= E[d(n)^v] - E[\underline{W}^T \underline{x}(n) d(n)] - E[d(n) (\underline{W}^T \underline{x}(n))^v] \\
 &\quad \underbrace{\qquad\qquad\qquad}_{\sigma_d^2: \text{variance of } d(n)} \quad \underbrace{\qquad\qquad\qquad}_{\underline{W}^T E[\underline{x}(n) d(n)]} \quad \underbrace{\qquad\qquad\qquad}_{\underline{h}^v(n)}
 \end{aligned}$$

So,  $\underline{W}^T \underline{h}(n)$  row vector and column vector or  $\underline{h}(n)^T \underline{W}$  you know they are same.

If I give you two vectors  $\underline{a}$  and  $\underline{b}$  you can verify they are column vectors. So, if you make it a transpose as a row vector times  $\underline{b}$  column vector. So,  $a_1 b_1 + a_2 b_2 + a_3 b_3$  like that dot dot you will get the same thing if you have  $\underline{b}^T \underline{a}$   $b_1 a_1 + b_2 a_2 + b_3 a_3$  same. So, these two terms will be same because  $\underline{W}^T \underline{h}(n)$  or  $\underline{h}(n)^T \underline{W}$  this is scalar.

So, both are same ok. So, again it will be same as what I have here  $\underline{W}^T \underline{P}$ . So, this entire thing will be same  $\underline{W}^T \underline{P} \underline{W}^T \underline{P}$  and last one  $\underline{W}^T \underline{x}(n) E$  if you take the transpose on it. So,  $\underline{x}(n)^T \underline{W}^T \underline{P} \underline{W}^T \underline{x}(n)$  there is  $\underline{W}$  ok. Again  $\underline{x}(n)$  if you take  $\underline{x}(n)^T$ ,  $\underline{x}(n)$  is a column vector this is a row vector. So, you got a matrix a matrix every element is a product of one data of  $\underline{x}(n)$  another data of  $\underline{x}(n)$ .

So, it is random. So, random matrix if you call it  $\underline{a}(n)$ ,  $\underline{a}(n)$  times  $\underline{W}$  will be again a scalar. So, the entire thing could be maybe  $\underline{h}(n)$ . So,  $\underline{a}(n)$  times  $\underline{W}$  is  $\underline{h}(n)$  which is column vector. So,  $\underline{W}^T \underline{W}$  is not random, So,  $\underline{W}^T$  row vector times  $\underline{h}(n)$  expected value

as you have seen earlier will be  $W^T$  transpose will go out. So, it will be this will be  $W^T$  transpose  $E$  of  $h_1 N$  and now  $h_1 N$  is  $a N$   $W$  and now  $a N$   $W$  see this is square matrix.

So, every term multiplying the terms here and getting added that is how you do  $a N$   $W$  on that if you apply  $E$ ,  $E$  will not work on the elements of  $W$  because they are not random  $E$  will work on the elements here. So, we will get the same thing if you apply  $E$  on the elements of this  $a N$  matrix first and then multiply that by  $W$  because  $E$  does not apply on  $W$ ,  $W$  is not random.

The image shows a handwritten derivation on a blackboard. The derivation starts with the definition of the error signal  $e(n)$  and its expected value  $E[e^2(n)]$ . The error signal is defined as  $e(n) = d(n) - W^T x(n)$ . The expected value is then calculated as follows:

$$E[e^2(n)] = E[(d(n) - W^T x(n))^2]$$

$$= E[d^2(n) - 2d(n)W^T x(n) + (W^T x(n))^2]$$

$$= E[d^2(n)] - 2E[d(n)W^T x(n)] + E[(W^T x(n))^2]$$

The first term  $E[d^2(n)]$  is identified as the variance of  $d(n)$ . The second term  $E[d(n)W^T x(n)]$  is simplified to  $W^T E[x(n)d(n)]$ , which is labeled as  $\phi$ . The third term  $E[(W^T x(n))^2]$  is simplified to  $W^T E[x(n)x^T(n)]W$ , where  $E[x(n)x^T(n)]$  is labeled as  $A(n)$  and  $h_1(n)$ .

At the top right, there is a small diagram showing the dimensions of the vectors and matrices:

$$\begin{matrix} a & 1 \\ a & 1 \\ b & 1 \\ b & 1 \end{matrix}$$

So, it will be  $W^T$  transpose  $E$  of  $a N$  and  $a N$  is this your  $a N$  by mistake I have erased that page, but does not matter I can go to another page. So, it will be  $W^T$  transpose  $e$   $a N$   $W$  and  $a N$  is  $x N$   $x$  transpose  $N$  and  $E$  of that is  $R$  auto correlation matrix  $W^T$  transpose  $R$   $x x$  I am dropping that  $x x$  from here just for simplicity  $W$ . So, expected value of  $e$  square  $N$  is  $\sigma_d^2$  minus that next two terms were same  $W^T$  transpose  $P$  and this  $W^T$  transpose  $R$   $W$ .

Now you see first right hand side is independent of  $N$  that is why I can write it simply as  $\epsilon$  square no need to put a subscript  $N$  and this  $N$  has disappeared from here because of stationarity like  $E$  of  $x N$   $x$  transpose  $N$  is  $R$  that is independent of  $N$  because of

stationarity or E of  $x_N$  into  $d_N$  that is  $P$  independent of  $N$  because of joint stationarity between  $x_N$  and  $d_N$  alright that is one thing. Another thing is you see this is a quadratic why quadratic first consider this, this will give you  $W^T P W$ . So, it is like  $W_0 P_0 W_1 P_1 \dots$  ok this, but this is first order not  $W_0$  square not  $W_1$  square this is first order, but here  $R W R W$  means you have got  $R$  you have got a rho of  $R W$ . So, this times  $W_0$  this times  $W_1$  all that. So, already  $R W$  means every term consist of elements of  $W_0$  to  $W_N$  and then you are multiplying one the first term here by  $W_0$  second term here by  $W_1$  and all that ok.

Handwritten mathematical derivation on a blackboard:

$$W^T E \left[ \frac{A(n)}{2(n)} \right] W = W^T R W$$

$$\tilde{J} = E[\tilde{e}^2(n)] = b_d^2 - 2 \underbrace{W^T P}_{(w_0 p_0 + w_1 p_1 + \dots + w_{N-1} p_{N-1})} W + \underbrace{W^T R W}_{\text{Quadratic function of the weights } w_0, w_1, \dots, w_{N-1}}$$

$$= \begin{bmatrix} w_0 & w_1 & \dots & w_{N-1} \end{bmatrix} \begin{bmatrix} p_0 & p_1 & \dots & p_{N-1} \end{bmatrix} + \begin{bmatrix} w_0 & w_1 & \dots & w_{N-1} \end{bmatrix} \begin{bmatrix} r_{00} & r_{01} & \dots & r_{0,N-1} \\ r_{10} & r_{11} & \dots & r_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N-1,0} & r_{N-1,1} & \dots & r_{N-1,N-1} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{N-1} \end{bmatrix}$$

This first row times this will be the first guy this is a vector after all column vector first guy will be first row times this that will consist of contain all these elements of  $W$  that will be multiplied by the first element here  $W_0$ . So, it will be  $W_0$  square,  $W_0 W_1$ ,  $W_0 W_2$ ,  $W_0 W_3$ , dot dot dot. Then second row times this will give the second guy of this vector again that will have  $W_0$  to  $W_N$  minus 1 that will be multiplied by  $W_1$  from here. So, we will have  $W_1 W_0$   $W_1$  square  $W_1 W_2$  dot dot dot. So, that way this is the second order component and this is a first order component.

So, model is a second order function and therefore, we should minimize this by taking taking this partial derivatives  $i$  equal to 0 1 dot dot dot dot  $N$  minus 1.

$$W^T E[A(n)] W = W^T R W$$

$x(n)x^T(n)$

$$J = \frac{1}{2} (y_d^T - 2 W^T p + W^T R W)$$

(Quadratic function of the weights  $w_0, w_1, \dots, w_{N-1}$ )

$p = [p_0, p_1, \dots, p_{N-1}]^T$

$$\frac{\partial J}{\partial w_i} = 0, \quad i = 0, 1, \dots, N-1$$

So, every partial derivative by do epsilon square together is a notation it is not epsilon and then square here by epsilon square whole together is a notation. So, deriving that which is for  $w_i$ . So, you will get capital N equations one from each derivative at second order. So, if you do derivation with respect to  $w_0$  or  $w_1$  or  $w_2$  you will get first order equation second order when derived will be first order.

So, we will have a set of first order equations capital N numbers capital N unknowns you solve them and we will get the optimal one which will minimize this ok. But instead of doing like this by individual terms you know we define a derivative operator ok on any function  $f$  a function of the weights. It is nothing, but you took that take the partial derivatives  $f$  like  $f$  is  $\epsilon$  epsilon square here. In general  $\frac{\partial f}{\partial w_0} \frac{\partial f}{\partial w_1} \dots \frac{\partial f}{\partial w_{N-1}}$  I am stacking only  $\frac{\partial f}{\partial w_{N-1}}$  these are definition. Then I do this apply this del operator derivative operator on epsilon square equate to a vector of zeros that is.

So, that will take care of these equation for all the  $i$  because this will be vector  $\frac{\partial \epsilon^2}{\partial w_0} \frac{\partial \epsilon^2}{\partial w_1} \dots \frac{\partial \epsilon^2}{\partial w_{N-1}}$  equal to 0 then  $\frac{\partial \epsilon^2}{\partial w_0}$  equal to 0 and so on and so forth. So, this is the equation we will be solving and that I will do in the next class. Thank you very much.