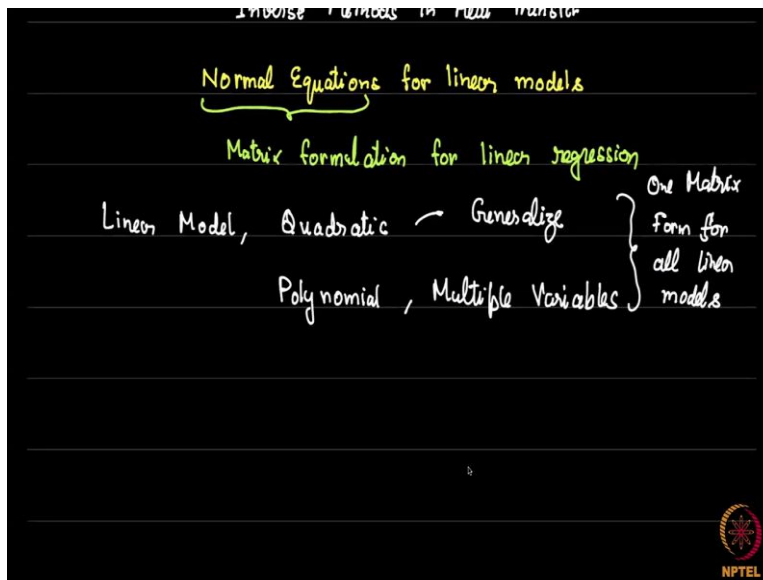


**Inverse Methods in Heat Transfer**  
**Prof: Balaji Srinivasan**  
**Department of Mechanical Engineering**  
**Indian Institute of Technology, Madras**

**Lecture No 12**  
**Introduction to Normal Equations for Linear Models**

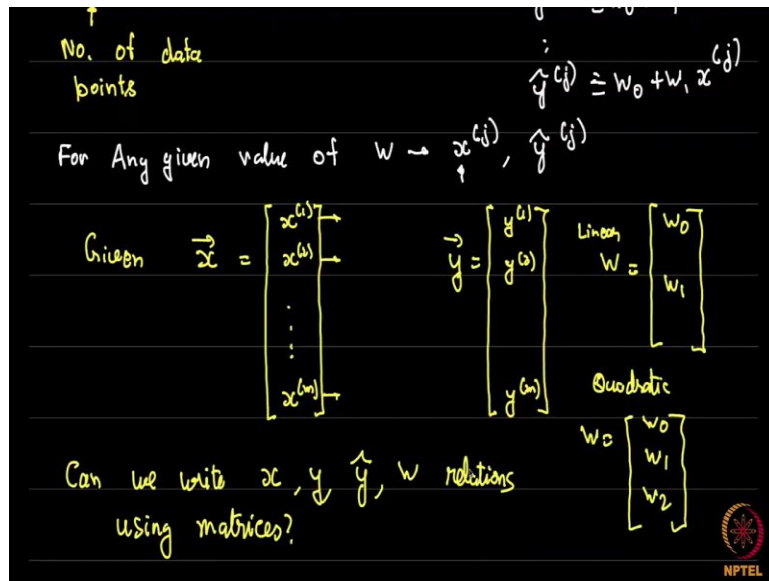
(Refer Slide Time 00:24)



Welcome back. This is week three of this course on inverse methods in heat transfer. This video is about the so-called normal equations for linear models. These normal equations are methods, basically the Matrix formulation for linear regression. This Matrix formulation for linear regression as you can see is a compact form. Now one thing of course is, this makes things a little bit easier to code in the general case. remember that in the last videos we had looked at a linear model, we had also looked at quadratic models.

So, what we will see is that we can generalize this and make the form of linear regression work for any polynomial and also multiple variables. I had briefly alluded to this in the last week, but what we will do in this video is write one single form that encompasses for all linear models. In the future videos for this week, we will also see that the same form can be used for things that kind of look non-linear, but a case can be linearized. So, let us proceed with this Matrix formulation.

(Refer Slide Time 00:24)



So, recall what we had I will write this as a table. Recall what we had as the general linear model. So, what we used to do for a general linear model is just have a lot of data points and we would have  $x$  which is the input variable, in the case of the slab problem this was the physical location for example and we also had  $y$ , which is from the experiment or what I called Ground truth. Apart from this we also have  $\hat{y}$  which is the prediction from our model.

So, typically the input to the inverse problem or this  $x$  column and the  $y$  column and  $\hat{y}$  is what comes after we guess for the parameters of the model  $w$ . So, we have a whole bunch of data points. Let us say 1, 2 so on and so forth. And I had said we could have something like  $m$  is the number of data points that we have or let us say in the case of our slab, that is the number of sensors or the thermocouples that we have.

So, let us say  $x$ , I am going to write a superscript for a particular reason  $x_1$  is the first location,  $x_2$  is the second location. so on and so forth till  $x_m$  being the  $m$ -th location. Now corresponding to this, we have ground truth or the experimental values  $y_1, y_2$  so on and so forth into  $y_m$ . Now for a linear model for example, let us say we had the simple linear model which is our usual model which is,

$$\hat{y} = w_0 + w_1 x$$

So, this is the usual linear model. This case is how we predict each one of these data points. So, for example,

$$\hat{y}^{(1)} = w_0 + w_1 x^{(1)}$$

and

$$\hat{y}^{(2)} = w_0 + w_1 x^{(2)}$$

and in general,

$$\hat{y}^{(j)} = w_0 + w_1 x^{(j)}$$

Now, you might be wondering why I am spending so much time on this. This is because frequently as we go ahead these superscripts and subscripts can get quickly confusing.

So, I just want to make sure that, everything is really clear to you as we proceed. So, once we use this. So, suppose we guess or for any given value of  $w$ , given  $x^{(j)}$ , you can find out  $\hat{y}^{(j)}$ , that is what I meant to say. So, you can similarly find out once you know  $x^{(1)}$  you can find out  $\hat{y}^{(1)}$ ,  $\hat{y}^{(2)}$ ,  $\hat{y}^{(m)}$ .

Then you compare these 2 and find out the error etcetera, which we did in the last week all right. Now what I want to do is to write this entire thing as a matrix, which is one way of thinking about it is given  $\vec{x}$ , which is all these data points,

$$\vec{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ x^{(m)} \end{bmatrix}$$

I am writing this as a column Matrix  $x^{(m)}$  where each  $x$  refers to a new data point and given  $\vec{y}$  which is,

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ y^{(m)} \end{bmatrix}$$

and also given  $w$ .

Now  $w$  for example in our case for a linear model is  $w_0, w_1$  if it is a linear model.

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

If it is a quadratic model, then  $w$  is as you might remember it has three coefficients  $w_0, w_1, w_2$ .

$$w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

So, our key question is this can we write these relations, that is the relationship between  $x, y, \hat{y}$  and  $w$  using matrices. Now why do we need matrices as you will see within this video, once you have matrices you can generalize every single linear form of model that we will be covering both in the that we had covered in the last week as well as we will be covering in this week.

**(Refer Slide Time 08:02)**

Linear Case

$$X = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(m)} \end{bmatrix} \Rightarrow$$

$$XW = \begin{bmatrix} 1 & x^{(1)} \\ \vdots & \vdots \\ 1 & x^{(m)} \end{bmatrix} \rightarrow$$

$$W = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad XW = \begin{bmatrix} w_0 + w_1 x^{(1)} \\ w_0 + w_1 x^{(2)} \\ \vdots \\ w_0 + w_1 x^{(m)} \end{bmatrix}$$

Quadratic Case

$$W = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} \quad XW = \begin{bmatrix} w_0 + w_1 x + w_2 x^2 \\ \vdots \\ w_0 + w_1 x + w_2 x^2 \end{bmatrix}$$

$XW = ?$

So, let us come back here and look at this model. So, the data here can Now be abstracted into a simple model. So, the way we are going to do it, I have done the same thing for a quadratic model here. this is for a linear model and this is for a quadratic model. Now notice the differences between the 2 and this will give you a clue on how to sort of do the Matrix operation in general.

Now this of course you can think of a serial number here 1, 2 etcetera on the left-hand side, but the important thing here is that I have one column for the  $x$  and another column for the  $x$  square. Now why do we do this that is because  $\hat{y} = w_0 + w_1x + w_2x^2$  and now, you can imagine an Excel sheet where this whole thing is written and what you would do in that Excel sheet, of course is have one column for just the serial number, one column for all the  $x$  values and then another column for  $x^2$  values.

And then maybe you can have a column for  $\hat{y}$  and calculate this as some  $w_0 + w_1$  the first Column +  $w_2$  the second column. If you add a cubic model, you would again do the same thing except you would have one extra column for the  $x^3$ . Once again this becomes obvious as I will go further and further within this video, on how this helps us generalize to any sort of linear model.

So, let us look at both these cases the linear case and the quadratic case and now start thinking about, how to abstract this into a matrix form. So, the way I am going to do it is to define something called a design matrix. A design Matrix is a matrix where all rows correspond to input data and columns correspond to features. For now, let us forget about this second part which is corresponding to features and just concentrate on input data.

So, if I look at this data set, the input data is purely just  $x$ , there is nothing else there. but I am going to make a small change here. So, for the linear case, I am going to define my design matrix as this capital  $X$  which consists of,

$$X = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x^{(m)} \end{bmatrix}$$

and I am going to augment it by just one a constant one everywhere here. you will shortly see why we add this. So, this really is here is the actual data and this we call the bias or let us just call it the constant term.

Now why do we do this? This will become a little bit clearer, when I actually do the quadratic data. So, for quadratic, my  $X$  the design variable for this, it is useful for you to refer to this. Now look at what the data we gave was the data input data was  $x$  as well as  $x^2$ . Now you might say  $x^2$  can be inferred from  $x$  but that is not the point, as far as the model is concerned  $x$  is has to be given separately and  $x^2$  has to be set even separately because it is multiplying different coefficients.

So, let us come back here and with an analogy to the linear case, this would be one all these biases, then I am going to call this  $x_1$  because we have to put a square up till  $x_m$  and  $x_m^2$ . So,

once again this, here is the data and this here is the constant. So, far we have done nothing I have just written a new matrix X we can also write the size of this matrix X you can see that there are 1, 2, 3, 4 up till m rows.

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_m & x_m^2 \end{bmatrix}$$

And in this case, you can see 2 columns. in this case, x has m rows and it has 3 columns all right. Now let us look at the Matrix w. w in the linear case is,

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

and in the quadratic case W is,

$$w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

Now w here is a  $2 \times 1$  matrix and w here is a  $3 \times 1$  matrix. So, I am going to ask a simple question. you can see that at the very least just by comparing the sizes, you can always do an operation like x times w because this is  $m \times 2$ , this is  $2 \times 1$ , this has to be a matrix some matrix which will have size  $m \times 1$ .

Similarly, here xw will be  $m \times 3$  and  $3 \times 1$  this will have some matrix here which is going to have size again  $m \times 1$ . So, let us now look at what this matrix XW is? Let us go back here let us look at XW. The very first row of xw is going to be first row into first column, which is going to be 1 times  $w_0$  which is  $w_0 + w_1x^{(1)}$ . Now what does this remind you of let me just do it once more, let us look at the second row  $w_0 + w_1x^{(2)}$  and the last one is  $w_0 + w_1x^{(m)}$ .

$$xw = \begin{bmatrix} w_0 + w_1x^{(1)} \\ w_0 + w_1x^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ w_0 + w_1x^{(m)} \end{bmatrix}$$

**(Refer Slide Time 16:10)**


$XW = Y$

For a polynomial of order  $n$

$$\hat{y} = w_0 + w_1x + w_2x^2 \dots + w_nx^n$$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \quad (n+1) \times 1$$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^n \end{bmatrix} \quad m \times (n+1)$$

$$\hat{Y} = \begin{bmatrix} \hat{y} \\ \vdots \\ \hat{y} \end{bmatrix} \quad m \times 1$$


Now if you look at this, this is exactly what you would expect, if you were writing the elements of  $\hat{y}$ . why because,

$$\hat{y}^{(j)} = w_0 + w_1x^{(j)}$$

So, this is nothing but the prediction that you will get if you put  $x^{(1)}$  into the model. So, you are just going to get your model prediction  $\hat{y}$ . So, this is  $\hat{y}^{(1)}$ , the next one is  $\hat{y}^{(2)}$  and the last one is  $\hat{y}^{(m)}$ . So, at least in the linear case we can immediately see, that  $xw$  equal to  $\hat{y}$ . Now what about the quadratic case we can again check  $xw$  is like this.

$$xw = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix}$$

Now let us notice what  $x$  is, 1 multiplying by  $w_0$  is  $w_0 + w_1x_1 + x_1^2w_2$ , which is exactly once again. I will write the sound  $w_0 + w_1x_1 + w_2x_1^2$ . Similarly, the next term will be  $w_0 + w_1x_2 + w_2x_2^2$ . Notice that I am moving around the subscript and superscript a little bit, just because we do not want to get confused with the squares and the last term will be  $w_0 + w_1x_m + w_2x_m^2$ .

$$xw = \begin{bmatrix} w_0 + w_1x_1 + x_1^2w_2 \\ w_0 + w_1x_2 + w_2x_2^2 \\ \vdots \\ w_0 + w_1x_m + w_2x_m^2 \end{bmatrix}$$

So, put together these are of course exactly the predictions of the quadratic model. So, once again here too,

$$xw = \hat{y}$$

In general, you can see this fairly quickly, the suppose I had a cubic model and I added one extra column here and made that  $x^3$  and my W's went to  $w_0, w_1, w_2, w_3$ , then once again  $xw = \hat{y}$ . So, we can write in general for any polynomial model we can write  $xw = \hat{y}$ .

This happens to be also the general formula for any linear model and I will show that to you shortly. Now let us compare sizes for a polynomial of order n. So, which means a polynomial of order n would look like,

$$\hat{y} = w_0 + w_1x + w_2x^2 + \dots + w_nx^n$$

w itself will be of size  $w_0, w_1$ , up until  $w_n$ .

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{bmatrix}$$

So, this is a  $(n + 1) \times 1$  matrix because of the  $w_0$  you have n of these terms then you have this added term  $w_0$ . So, this is  $(n + 1) \times 1$  matrix.

Similarly, X is going to be

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdot & \cdot & \cdot & x_1^n \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_m & x_m^2 & \cdot & \cdot & \cdot & x_m^n \end{bmatrix}$$

So, if you look at this x is a matrix of m rows and n + 1 columns and  $\hat{y}$  and y are both  $m \times 1$  matrices all right. So, this is what happens, when you have General polynomial model.


**(Refer Slide Time 20:59)**



in features (independent variables)  
Thermocouples

$$X = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & \dots & x_5^{(i)} \\ \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \end{bmatrix} \quad W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_5 \end{bmatrix}$$

$\hat{y} = X W$  General Expression for any linear case



But for General linear models, we may have more input features. So, for example we could have a case, where we have instead of one dimension, 2 dimensional models. So, for example we had temperature as a function of  $x$  in one dimension, but it could be temperature of  $x$  comma  $y$  and since we do not want to confuse the  $y$  which is also an output variable, we are going to call it  $x_1$  and  $x_2$ .

$$T(x_1, x_2)$$

So, instead of a slab, let us say you have a plate and you want thermocouples at a lot of places. you need a model which depends on both  $x$  as well as  $y$  location and of course three dimensional, let us say you have some semiconductor stuff like that and you have a chip and you want a temperature model within that of course. Now it is going to be a function of three variables. So, 3D you will have things like  $T(x_1, x_2, x_3)$ .

Now let us say you have a case like this and you have a linear model. Now these things here are called features or attributes. It is more accurate to call them attributes but I am going to call them features. we of course call them in the usual language independent variables. Now let us say you are solving an inverse problem with not one independent variable which is the only case that we did till now. But we are looking at multiple independent variables what does that look like.

So, in this case, if you have multiple features or multiple independent variables, let us say you have a temperature model then you could have the temperature, which I am going to call  $\hat{y}$  again, it is a hypothesis would depend on,

$$\hat{y} = w_0 + w_1x_1 + w_2x_2$$

This would be the linear model in 2 dimensions. Notice here I have used a subscript to mean the dimension and not the data point, ideally speaking if you have multiple such sensors.

Let us say you have seven such sensors here or eight such sensors here, then what you would do is, suppose I want the fourth sensor, let us say or the fifth sensor then I would say something like  $\hat{y}^{(5)}$  would be the x location of that five multiplied by  $w_1$  + the y location of 5 multiplied by  $w_2$  and + some extra term. So, this would be a linear model. Now what happens to a quadratic model?

$$\hat{y}^{(5)} = w_0 + w_1x_1^{(5)} + w_2x_2^{(5)}$$

So, that would be something like so, suppose we know that the temperature also has a heat source such that the entire thing is going to be quadratic, then you would have something like,

$$\hat{y} = w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2$$

$\hat{y}$  equal to  $w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2$  + a quadratic model would also have cross terms let us say  $w_5x_1x_2$ . Now this looks like a very complicated model but luckily this is where our matrix formulation is extremely helpful. So, I want to point out again this is a general model that handles linear quadratic any type of polynomial.

And it also handles multiple Dimensions or multiple features. Now how does it handle that by a very simple trick, let us take this expression. I am now going to call it,

$$\hat{y} = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5$$

where it is our understanding that  $x_3$  corresponds to  $x_1^2$ ,  $x_4$  corresponds to  $x_2^2$  and  $x_5$  corresponds to  $x_1x_2$ . Now you might say that but this is not linear. the point is it need not be linear in x. why because these are given data points the problem, we are solving for is the problem of solving for w.

So, all that matters is for linear models are linear in parameters and not in features or independent variables. for example, when we had a simple quadratic model such as this, it was not linear in x but it was linear in W. So, it is only  $w_0$ ,  $w_1$  and  $w_2$  you do not have terms like  $w_0^2$ ,  $w_1^2$ ,  $w_2^2$ , etcetera. All you have are terms that are linear in w. When you look at this formula  $xw = \hat{y}$  the main thing here is this is linear in w, x is basically a constant matrix.

So, it does not matter whether it is linear non-linear whatever. So, for example I could make up a model with an additional term, if I wish like  $w_6$  times  $x_6$  where  $x_6$  could be sine of  $x_1$  anything of that sort anything of that sort is still linear as long as you have only  $w_6$  here and not  $w_6$  square or  $w_1, w_2, w_3, w_6$  etcetera as long as that is the case this model is always linear.

Now the second thing is when we look at an Excel sheet or that kind of format, all you need to do is you need to substitute, you had x, x-square here and y here in serial number instead of that you will write x1 which is the x dimension then you will write x2 which is the y dimension then you will write x1-square as another column x2-square is another column. so on and so forth and you can actually make up an entire design Matrix that way.

For example, in this case the design Matrix capital X is going to look like a whole bunch of ones and then after that,

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdot & \cdot & \cdot & x_5^{(1)} \end{bmatrix}$$

Remember these are just Columns of our Excel sheet except for this first column. this first column is just adding attack done just to make the matrix multiplication work well. So, after that this will be x1 at the first point, this will be x2 at the first point remember each one of these corresponds to the thermocouples.

The superscript corresponds to the location of the thermocouple and the subscript here corresponds to our Excel sheet entry. the first 2 are simply the x location and the y location. The third one would be as we already know x3 was x1-square, it is the square of the location. x4 is the square of the y-location and x5 is x-location into y-location and then w is

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \cdot \\ \cdot \\ \cdot \\ w_5 \end{bmatrix}$$

And once again if you see the model, it looks exactly the same  $\hat{y}$  equal to x times w. why for example at the first point  $\hat{y} = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5$ . So, you can basically use this Matrix in order to write the relationship. Now for any linear model not only the linear model that we saw which was very simple  $w_0 + w_1x$  but this is through the in general.

So, this is the general expression as I wrote pick for any linear case but how does this help us solve the optimization problem.

(Refer Slide Time 30:55)

$$J = \frac{1}{2} (Y^T Y - W^T X^T Y - Y^T X W + W^T X^T X W)$$

Minimize  $J$  wrt all possible values of  $W$

$J$  is a scalar,  $W$  is a vector

So, let us look at that now. so, solving a general regression problem. Now if you recall last week, we had 2 independent Solutions. even though there was some pattern, that we could notice there were 2 independent solutions for a linear case and for a quadratic case and you can imagine redriving it for a cubic case or for a case where you are in 2 dimensions and you have multiple you have 2 dimensions and quadratic or three dimensions and cubic etcetera that becomes more and more messy.

Can we write one single expression that works for every case and we are going to do that. this is what is called the normal equations approach. So, our model is as follows. I am going to use a particular couple of notations. Notations are this, if I say some vector  $v$  this thing is called  $\|v\|^2$ . This is how we pronounce it, this part is called the norm with the double absolute sign that part is called Norm of  $V$ . This is nothing but all the components of  $V$ ,

$$\|v\|^2 = v_1^2 + v_2^2 + \dots v_n^2$$

So, if  $V$  is a vector which goes from,

$$V = \begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{bmatrix}$$

$\|v\|^2 = v_1^2 + v_2^2 + \dots + v_n^2$ . Now another way of writing this is  $V^T$ , if  $V$  is this Matrix,  $V^T$  all of you would be aware is basically just the transposed case.

$$V^T = [V_1 \quad V_2 \quad \dots \quad V_n]$$

It is the row Matrix and  $V$  is the column matrix. So,  $V^T V$  you can now see is,

$$\begin{aligned} V^T V &= v_1^2 + v_2^2 + \dots + v_n^2 \\ &= \|v\|^2 \end{aligned}$$

under the usual is the same as Norm of  $V$  Square  $\|v\|^2$ .

Technically speaking you are supposed to call this the 2 Norm of this and I am supposed to put a subscript 2 etcetera all that I am going to skip at least for now. So, this is  $V^T V$ . Now let us come to our loss function. So, the objective function or the loss function that we are minimizing  $J$  was,

$$J = \frac{1}{2} \sum_{i=1}^m (y^{(i)} - \widehat{y}^{(i)})^2$$

Let us call this  $e_i$ , where  $e_i$  is the error in the  $i$ th data point.

Now of course  $e$  itself is a vector,

$$e = \begin{bmatrix} e^{(1)} \\ e^{(2)} \\ \vdots \\ e^{(m)} \end{bmatrix}$$

So, it is an  $m \times 1$  vector. So, this  $e$  is now a  $m \times 1$  matrix. Now what we want to do of course is express  $J$  also as a matrix. So, remember we had expressed  $x$ ,  $w$ ,  $\hat{y}$  all as matrices. So,  $J$  also we want to represent in a matrix form. Remember  $J$  of course is a scalar. So, this is half of now you can see  $\|e\|^2$ ,

$$J = \frac{1}{2} \|e\|^2$$

why  $e$  is a vector and  $e$  vector are basically,

$$\vec{e} = \vec{y} - \hat{\vec{y}}$$

and each element of  $\|e\|^2$  and added just like before, this  $v$  example is what is happening here.

So, this is  $\|e\|^2$  and which we just saw  $\|e\|^2$  is nothing but half of  $e^T e$ .

$$J = \frac{1}{2} e^T e$$

So, another way of writing J is to say  $J = \frac{1}{2} e^T e$ . Now opening up e, this is half of  $(y - \hat{y})^T (y - \hat{y})$  and we can open this up further. This is half of let us take each term  $y^T y$  minus the second term  $\hat{y}^T y$ , minus the third term  $y^T \hat{y}$  and the final term which is plus of  $\hat{y}^T \hat{y}$ .

$$J = \frac{1}{2} (y - \hat{y})^T (y - \hat{y})$$

$$J = \frac{1}{2} (y^T y - \hat{y}^T y - y^T \hat{y} + \hat{y}^T \hat{y})$$

So, these are the four terms this is of course equal to J. how is this useful I am going to take one further step. Remember that we have a model for y-hat we had the model this of course is an entire vector

$$xw = \hat{y}$$

So,  $\hat{y}$  is equal to x times w. we can also write,

$$\hat{y}^T = (xw)^T = w^T x^T$$

$\hat{y}^T$ , which occurs in a couple of terms  $\hat{y}^T = (xw)^T = w^T x^T$ .

So, we are going to substitute these expressions here and write them out now. So, J equal to half of I am going to use capitals here to indicate that they are vectors; vector signs would get a little bit messy.

$$J = \frac{1}{2} (Y^T Y - W^T X^T Y - Y^T X W + W^T X^T X W)$$

$Y^T Y$ , minus  $\hat{Y}^T$  which is minus  $W^T X^T Y$ , that is this term here, then the next term minus  $Y^T$  times  $\hat{y}$  is just  $xw$ . And the final term is  $\hat{y}^T \hat{y}$  which becomes  $+ \hat{y}^T$  is  $w^T x^T$  and  $\hat{y}$  is  $xw$ .

So, remember this is  $\hat{y}$ . So, this is what we wish to minimize. So, what we wish to minimize, the minimization problem is minimize J with respect to all possible values of w. Now remember J is a scalar. So, it is only one function which we are minimizing but w is a vector, it is a vector of values which goes from  $w_0, w_1$  up until  $w_n$ . So, n we will treat us in general the number of features and m in general as the number of data points. This is the notation that we will be following.

Now this is the giant function, it does not look like we have simplified it looks like we have complicated things but as you will see this actually simplifies things. Now although in the beginning of this video, I started out saying that we will do our normal equations. This video is

already a little bit long this is the objective function we want to minimize. And in the next video I will start with the subjective function and minimize this with respect to  $w$ . So, see you in the next video, Thank you.