

Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 08
Lecture - 56

Linear least squares - Part 1

Today we will look at least squares problems. I think the last time I just introduced it, but we did not get into the details of it. So, we look at least squares problems which as I mentioned are probably one of the most commonly found problems in engineering. So, there is linear least squares and non-linear least squares we look at both. And once you are done with that the next part of the course will be constraint optimization which will start after the quiz ok. So, let us look at least squares optimization.

In terms of motivation, we do not need to re-motivate it. We had given two models if you remember. What was the first model? There was radioactive decay, right. There was one model.

So, let me just write that one example down. So, this was example 1 was something like this. So, over here what was our t was our measurement time and what would we call x in plain English what would you call x parameters of the model. Model and the other model that we had was for example, something simple like this. What is the difference between these two models? In terms of nomenclature the first model would be it is non-linear in the parameters right.

So, this is a non-linear in the parameters and this is linear, loosely speaking we call it linear, but technically what should we call it? Affine I mean there could be a constant also over here which I have not mentioned. So, these are the two broad families of problems once you have got a model the model is not going to come to you from optimization the model is going to come to you from your domain knowledge ok. And as I mentioned in the very first lecture this is where your modeling accuracy comes into play ok. Having done this we had written the problem in the following way what is the problem to be solved right. let us say I take m measurements for each measurement I have a model and I have the models prediction and this I have to try to what bring as close to each other as possible right that is the basic idea.

So, if the measurement is let us say some y_j correspondingly the prediction is going to be $y_j(x)$. This is my you could say my j th measurement and what I am going to do is I am going to square this, j is going to go from 1 to m right and this is what I want to minimize as much as possible right. So, what is the tuning knob in my hand? I mean it is a very basic question, but you will be surprised how often students get tripped up in this. What is the tuning knob in my hand? x . So, I will write this as this $\arg \min_x$.

This just tells me that x is the variable of optimization which is in my hand and after I solve this \hat{x} , \hat{x} here means it is the best x right ok. And so, this is why we call this a least squares optimization. I think we had spoken roughly until this point. Now, to progress any further we obviously need to know some or not know, but do some calculus on this so that we can take derivatives set them to 0 look at second derivatives and so on ok. So, there is a because these

problems occur so frequently there is a standard terminology for solving or mentioning most of this.

So, each of these. So, let me just make a mention over here this is total number of measurements ok. Is there a common word that we use to denote the expression inside this red bracket? In the optimization world this expression within the red bracket what do we also call? Not quadratic, a general word objective function or another very common word is cost function ok. So, cost function in many papers you will in fact see this written as $c \cdot f$ and you will wonder what does this $c \cdot f$ this means cost function or objective function ok. And the common symbol for it let us say let us call it $f(x)$.

NPTEL
Note Title

Least Squares Optimization

eg. 1) $y(x,t) = x_1 + x_2 \exp(-x_3 t)$ N.L.
 ↘ meas. time
 ↘ model params

$y(x,t) = x_1 + x_2 t + x_3 t^2$ L.

Now, when I say $f(x)$ you should realize that x here is a vector or a scalar, it is a vector because there can be many parameters in the model. So, the common way of writing this expression this one term over m is called a residual. So, let us say $r_j(x) = y_j - y(x_j)$. I am going to call this a residual. Why residual? Because that is the residue left when I match my data with my prediction ok and the common notation is this $r_j(x)$. So, if I have all of these residual quantities are there for each measurement I have these residuals, I can as well stack it and make a vector out of it right.

So, the residual vector then is simply has the expression $r(x)$ ok. So, $r(x)$ is going to be a vector. So, $r_1(x), r_2(x), \dots, r_m(x)$ because there are m measurements and transpose. ok and the common expression is written like this ok. So, $j = 1$ to m $r_j(x)$ this half is introduced just to make math a little bit easier otherwise you will have a 2 floating around everywhere.

So, do not worry too much about it ok. Now how many obviously, I am going to call x as a n dimensional vector. So, $x \in \mathbb{R}^n$ right. So, let us look at the numbers over here. If I have $m \geq n$ ok.

So, here is where we need to first operate from common sense. $m \geq n$ means what? I have more measurement than variables is that does from a common sense point of view is that a good place to be or a bad place to be? It is a good place to be in because you have enough measurements are there I mean if I have like 10 parameters of the model and one measurement and there is that famous saying I do not know who that physicist is that if you give me 3 points I can fit an elephant for you right. Anything any model will fit it if I have less data and too many parameters right. So, this is where you want to be and this is also the more realistic case right. What about this? There may be some situations where data may be extremely expensive.

For example, the example I mean with the Geiger counter right radioactive measurements you cannot send a person to make too many measurements. So, you have less number of measurements right. So, there are situations when this also happens this is more challenging and the uniqueness of the solution all of these properties are questionable ok. So, in principle this is the situation where you can have infinite solutions ok. When you have infinite solutions then some solutions come with some special properties right.

So, let us take a very simple case something that you have all studied from class 11. When you have a linear system of equation that means a fact matrix, fact matrix has infinite solutions. Is there any special solution out of these infinite solutions? I have infinite number of solutions that is that everyone agrees on, but is there any one solution which has a special property? there is a minimum norm solution that is possible. Of all the solutions you can find one solution which has minimum norm ok. Minimum norm and we will discuss this at the end of this class when we look at linear least squares.

NPTEL

Problem:

$$\hat{x} = \underset{x}{\operatorname{argmin}} \sum_{j=1}^m |y_j - y(x, t_j)|^2$$

total no of meas.

Least squares optimization.

$f(x) = \text{Cost function (c.f.) / Objective fn}$

$y_j - y(x, t_j) \rightarrow \text{residual} \rightarrow r_j(x)$

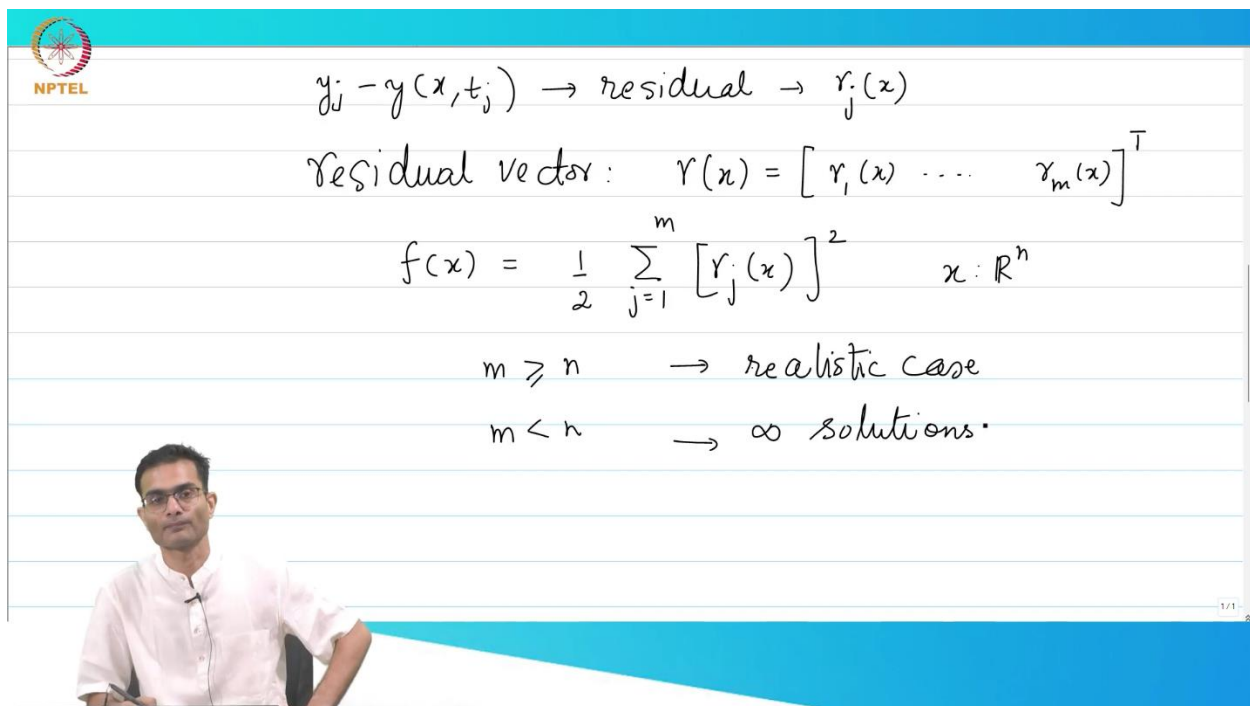
Residual vector: $r(x) = [r_1(x) \dots r_m(x)]^T$

OPTIMIZATION THEORY AND ALGORITHMS

Is there any from an engineering point of view is there any special appeal for a minimum norm solution? Minimum norm can often mean minimum energy. So, that has a desirable aspect to it. So, even though you may have infinite solutions you may be still be able to arrive at a minimum norm solution and minimum norm can be for example, minimum energy for a physics

application for some sensor network application it may be minimum power of the nodes whatever right. So, those solutions are still possible. There is another very interesting area of research now it is about I would say 10 to 15 years old which is called compressive sensing which is a very beautiful subfield of signal processing where you can show there are theoretical guarantees for this that you in spite of m being less than n you can arrive at a unique solution ok.

It is so this is a very nice possible topic for your course project ok. Many students in the past iterations have enjoyed this topic. So, just telling you it is very counterintuitive, but it is good fun. But for now this is the case that we are going to look at where common sense dictates there is more measurements than the number of variables over here. So, with my with my residual vector R defined this way is there yet another very compact way of writing this cost function $f(x)$ instead of this laborious summation that I have written can I write it in some other way? Half norm r square L_2 norm.



NPTEL

$$y_j - y(x, t_j) \rightarrow \text{residual} \rightarrow r_j(x)$$

$$\text{Residual vector: } r(x) = [r_1(x) \dots r_m(x)]^T$$

$$f(x) = \frac{1}{2} \sum_{j=1}^m [r_j(x)]^2 \quad x: \mathbb{R}^n$$

$$m \geq n \rightarrow \text{realistic case}$$

$$m < n \rightarrow \infty \text{ solutions.}$$

So let us look at these two objects over here. So I have at the top over here. Now $r(x)$. Is $r(x)$ like a function also? Does it take an input and give an output? So what is the input? Dimension of the input? n . n variable x enters into this guy.

What comes out of it? How many measurements do we have? m right. So, it is a vector right. Similarly, $f(x)$ that is our objective function or cost function. What does it take as input? \mathbb{R}^n comes in and out comes simply \mathbb{R} right. So, which one of these guys is going to have a Jacobian a non-trivial Jacobian? It is going to be r .

Now finally, when I am when I am looking at $f(x)$ which I have written over here I want to if I want to calculate a stationary point of f what am I going to be looking at? Given f you want a stationary point what is the quantity that I need to look at? ∇f right.

Now, ∇f because f is written in terms of this r it will be it will make our life a little simpler if you look at anticipate that what is going to happen when I start differentiating f with respect to x_1 x_2 up to x_n . I am going to see a norm r square that is going to appear everywhere. So, before we jump into ∇f let us just look at the Jacobian of r it will help us simplify the math because I have to take a derivative of r anyway and a derivative of r is also known as Jacobian. So, let us look at. So, remember we had a if I give you a function f and I ask you ∇f , it was a column vector like this. Everyone remembers this, but when I gave you a function like r which is $\mathbb{R}^n \rightarrow \mathbb{R}^m$, how did we write the Jacobian of r ? sorry small r there was a transpose operation if you remember right.

NPTEL

$m < n \rightarrow \infty \text{ solutions}$
 $[\text{Compressive Sensing}]$

$r(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$
 $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$

$f \rightarrow \nabla f \rightarrow \begin{bmatrix} \partial f / \partial x_1 \\ \vdots \\ \partial f / \partial x_n \end{bmatrix}$

Jacobian of $r(x)$?
 $J_{ji} \equiv \begin{pmatrix} \partial r_j \\ \partial x_i \end{pmatrix} \rightarrow \begin{bmatrix} (\nabla r_1)^T \leftrightarrow f \\ (\nabla r_2)^T \leftrightarrow \\ \vdots \end{bmatrix} \begin{bmatrix} \partial r_1 / \partial x_1 & \partial r_1 / \partial x_2 & \dots & \partial r_1 / \partial x_n \end{bmatrix}$

OPTIMIZATION THEORY AND ALGORITHMS

So, every row of the Jacobian was what? ∇ of each of those residual terms right. So, let us write down. So, is this right right? So, I have $J_{i,j}$. So, J is now the row, i is the column right. So, if I am keeping the row number fixed and I am changing the columns right, I am getting ∇r_1 is one row, ∇r_2 is one row right. So, if you visualize it, it is like this.

∇r_1^T . So, it is going like this ∇r_2^T is going like this. So, that is how my matrix looks like and if you wanted it further explicit $r_1 / \partial x_1$, $r_1 / \partial x_2$. That is how it looks ok. So, you already know $f(x) = \frac{1}{2} \sum_{j=1}^m r_j(x)^2$. Now, I want to calculate ∇f right for calculating ∇f what was the one thing that I need to do calculate for example, $\frac{df}{dx_i}$ right.

So, I can take this expression over here and start differentiating $j = 1$ to m . So, one common mistake again lot of time students make is that mixing up the variable of summation and the variable of differentiation. So, here the dummy index for the summation is j right. So, I should take the derivative for I mean this the subscript for x should not be j because I will land up into that is why I take an i over here very common mistake, but it messes you up.

So, right. So, if I take this derivative what happens? I am taking the derivative of this expression over here. What happens to my 2? 2 goes away right. So, this is will the summation still be there? Has to be there because each r depends on all the x 's. So, it is going to be there.

So, this summation remains there ok. I have a $r_j(x)$ and I have a $\frac{dr_j}{dx_i}(x)$ and I can just for simplicity rearrange them this way. Now, this is $\frac{df}{dx_i}$, I like this I want from $\frac{df}{dx_1} x_2$ all the way up to x_n . What is this guy reminding you of? It is looking like exactly like the term of the Jacobian here, it is the it is looking like the j th column. Right and I have a r_j over there. So, is there a nice way by which we could simplify this expression? It is looking like the column of a I mean you look at j is the guy that is varying in this expression over here, j is the variable of summation i is remaining constant, ok.

$r(x) : \mathbb{R} \rightarrow \mathbb{R}$

$f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$

$f(x) = \frac{1}{2} \sum_{j=1}^m (r_j(x))^2$

Jacobian of $r(x)$?

$J_{ji} \equiv \begin{pmatrix} \frac{\partial r_j}{\partial x_i} \end{pmatrix} \rightarrow \begin{bmatrix} (\nabla r_1)^T \leftrightarrow \\ (\nabla r_2)^T \leftrightarrow \\ \vdots \end{bmatrix} \begin{matrix} \left[\begin{matrix} \frac{\partial r_1}{\partial x_1} & \frac{\partial r_1}{\partial x_2} & \dots & \frac{\partial r_1}{\partial x_n} \end{matrix} \right] \\ \vdots \\ \left[\begin{matrix} \frac{\partial r_m}{\partial x_1} & \frac{\partial r_m}{\partial x_2} & \dots & \frac{\partial r_m}{\partial x_n} \end{matrix} \right] \end{matrix}$

$\nabla f \rightarrow \frac{\partial f}{\partial x_i} = \sum_{j=1}^m r_j(x) \frac{\partial r_j(x)}{\partial x_i} = \sum_{j=1}^m \left(\frac{\partial r_j}{\partial x_i} \right) r_j = J^T(x) r(x)$

OPTIMIZATION THEORY AND ALGORITHMS

So, over here in if i is remaining constant which is x that is corresponding to for example, let us get blue over here. One column like this because over here what is changing $r_1 r_2 r_3$, but x_2 is remaining the same. So, it is a column of my Jacobian matrix. So, I am multiplying the column of a Jacobian matrix, but when I multiply a matrix with a vector what do I multiply a row of the matrix with a column vector. So, this is not exactly like that with j , but it is with j^T .

So, I take J^T this blue circle thing will become a row vector and that is getting multiplied by which vector? r , its components are there $r_j r_1 r_2$ right. So, this is so, this will simply be J^T multiplied by r . Right if you want it to be little bit more explicit you would call this because they are all functions of x ok. So, this is why we spent like 15 seconds trying to compute this Jacobian because it shows up very neatly over here. If I did not have this I will have to write this partial summation I mean the summation with partial derivatives carried forward everywhere everywhere at some point I am going to interchange the j and i and my answer is going to go wrong right.

So, this is kept it this is kept it. So, that I have got my expression for ∇f any questions on this anyone not clear on.