

Machine learning for engineering and science applications
Professor Dr. Balaji Srinivas
Department of Computer Mechanical Engineering
Indian Institute of Technology Madras
Derivatives, Gradient, Hessian, Jacobian, Taylor Series

This week we will be dealing with optimisation and as you would know from your experience in school as well as in college, almost all optimisation involves you to find out derivatives. So in this video we will be looking at derivatives so little bit of warning both this video and the next one which will deal with what is called matrix calculus, they will be widely advanced material.

Some of it you will once again be already familiar with in the one-dimensional context or in the context of scalars and we will be looking at the context of vectors also. We have only a few slides to go through both in this video as well as in the next but the materials are little bit dense so please concentrate on this material and if it is not very-very clear, you will still get clarified as the course goes on okay.

(Refer Slide Time: 1:05)

Derivatives

$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$ (Handwritten note)

$\frac{dy}{dx} = f'(x)$ (Handwritten note)

$z = f(x, y)$ (Handwritten note)

$\frac{\partial z}{\partial x} = \frac{\partial f}{\partial x}$ (Handwritten note)

$\vec{x} = (x, y, z) \in \mathbb{R}^3$ (Handwritten note)

$\vec{x} = (x, y) \in \mathbb{R}^2$ (Handwritten note)

$\frac{\partial z}{\partial x} [x, y=1]$ (Handwritten note)

- Derivatives measure how one quantity changes when there is a small change in another
- Geometrically, in one dimension, this can be given as the slope of the tangent

$$f'(a) = \frac{df}{dx}(x=a) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$
- In higher dimensions (functions of many variables/vectors), we have the idea of partial derivatives $f: \mathbb{R}^n \rightarrow \mathbb{R}$

$$\frac{\partial f}{\partial x_i}(a_1, \dots, a_n) = \lim_{h \rightarrow 0} \frac{f(a_1, \dots, a_i + h, \dots, a_n) - f(a_1, \dots, a_i, \dots, a_n)}{h}$$

https://en.wikipedia.org/wiki/Differential_calculus#Tangent_to_a_curve

So let us first look at the idea of derivatives which is essential for any sort of optimisation. So derivatives typically measure how one quantity changes when there is a small change in another. So if you have something like dY/dX it means how much does Y change given that X changes by a certain amount okay. So, as you would know geometrically in a simple scalar case we look at this as the slope of a tangent okay. So if this is the curve $Y = F$ of X

then if this is the point let us say p , if you differentiate Y with respect to X at X equal to p you will get the slope of this tangent, of course you can denote this as $\frac{dY}{dX}$ at X equal to P or you can denote this as F' , some people will simply call it p or people will call it F' X equal to p so there are multiple ways of denoting this, you would be familiar with all of this once again from your prior experience.

So we know that this slope essentially is can be written as limits of a small perturbation of X , so $X + H - F$ of X by H , this of course is the limits of this secants ok as they go towards this point and become a tangent, so the slope at this point X and $X + H$, so you find out the difference in values and as this limit tends to 0, the slope will tend to a finite value and that is what we call the derivative of the slope at that point. Now, when you have higher dimensions, by higher dimensions I simply mean you still have a scalar function but X now is a vector okay. So in that case, X vector could be something like let us say X_1, X_2, X_3 , or the figure that I will show shortly could be X factor is X_1, X_2 which means X belongs to \mathbb{R}^2 , this is the case where X belongs to \mathbb{R}^3 okay.

So in such a case we can have partial derivative, so let us look at such an example let us say Z if F of X and Y okay, now if you want to denote or visualize Z , you simply have the variables X and Y , as they change Z changes and you see here 1 whole surface okay for Z . Now I could want to know what is $\text{Del } Z$, $\text{Del } X$ ok that is at a particular point let us say this point I might want to know if I just change X and I keep Y fix you would have seen such thing in thermodynamics perhaps but if change X and keep Y fix, you might want to know how much does Z changes.

Now the way to see that geometrically through let us say you draw a cross-section something of this sort okay. Let us say Y is fixed at Y equal to in this case 1 and you can try and find out what this derivative is. A generalisation of this idea is with N variables okay, so here F is a function that takes in a vector, in this case the vector is A , which is in \mathbb{R}^N which has N components and it gives back a single scalar.

And if you want to find out Del of $\text{Del } X$ I that is just like in this case I want the derivative with respect to X then all you do is you change only that variable so in this case for example, I change only the i th variable, I perturbed it by a little bit so I do A_i goes to $A_i + H$ and then find out how much does the function changes when I just change this variable and that limit as H tends to 0 is what is called the partial derivative of F with respect to the variable X_i okay.

Now reduced to a one-dimensional problem this is what it would look like, this is simply the cross-section of this function at Y equal to 1 and if I want the slope now then all I will do is, let us say I will change X by a little bit. So suppose I want $\text{Del } Z, \text{Del } X$ at X equal to 1, Y equal to 1 then I take a cross-section, where Y is fixed at 1 and evaluate the slope at X equal to 1 by just changing X and that slope will actually gave me the value of this partial derivative okay, so this is the idea of partial derivative again you should be familiar with this from multivariable calculus before.

(Refer Slide Time: 6:20)

The slide contains the following text and equations:

- The **gradient** is the multivariable generalization of the derivative
- It is a vector the components of which denote the partial derivatives in each direction

$$\nabla_x f(x_1, \dots, x_n) = \left[\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right]^T$$
- It can be used to calculate the directional partial derivative of f along the direction \vec{v}

$$D_{\vec{v}} f(x) = \lim_{\alpha \rightarrow 0} \frac{\partial f(x + \alpha \vec{v})}{\partial \alpha} = \nabla_x f(x) \cdot \vec{v} \quad \nabla_x f = \left[\frac{\partial f}{\partial x_1} \dots \frac{\partial f}{\partial x_n} \right]^T \cdot \vec{v} = \frac{\partial f}{\partial x_1}$$

Now we can generalise this idea okay, of derivatives to what is called the gradient which we will be using very-very often okay once again. So let us say you have let us call this X, Y and F of X Y, you can call it Z = F of X Y okay, so you have a curve of this sort okay. There are several noticeable things here, so suppose I want to say how much does the value of the function change at this point that notion by itself does not become intelligent unless you say how much does it changes with X.

So you have $\text{Del } F \text{ Del } X$ and you also have $\text{Del } F \text{ Del } Y$ okay. In fact, instead of just looking at these 2 directions so $\text{Del } F \text{ Del } X$ would be the change in the direction X and $\text{Del } F \text{ Del } Y$ will be the change in direction Y, you could ask a 3rd direction, I could call it $\text{Del } F \text{ Del } V$, where V is some arbitrary direction okay. So if this is X, this is Y, V could be some 3rd direction altogether.

So the gradient is defined as basically a concatenation or putting together of all these partial derivative, so in this case with the two-dimensional case we have to search partial derivatives,

in the n dimensional case you will have n such partial derivatives and you would basically write the gradient of F in my case would be $\text{Del } F \text{ Del } X_1, \text{Del } F \text{ Del } X_2$ okay. So in the n dimensional case it is $\text{Del } F \text{ Del } X_1, \text{Del } F \text{ Del } X_2, \dots$ so on and so forth up till $\text{Del } F \text{ Del } X_n$, and I have put a transpose there to show that this is a vector, some people eliminate the transpose, some people put the transpose either is fine. In this case now notice this is a vector and we look at a more general case of this in the next video which would be matrix calculus video but this gradient is used very-very often okay.

Now what does the gradient physically represents? Okay, so if we see that here are couple of figures to clarify this idea, so let us look at the 1st figure. The 1st figure is just showing the shading, now imagine this curve here, if it is collapsed okay imagine it is a bunch of springs and you just collapse it and you will see these things here, the projection here are called contours, what does the contour mean? If I take this contour and raise it up to the curve, it has all the values at this value of X and Y all of these places Z has the same value okay, so these are what are called level sets or contours which we will look at a little bit later in this video series also.

Now this is shaded according to value for example, here the value of the function is high, here the value of the function is low so the place where the value of the function is high is shaded as dark black and later on it is shaded white okay. Now the gradient notice is a vector and the direction of the gradient tells you in which direction is the change the sharpest okay, so the change is the highest in the direction of the gradient okay. So in this case for example all the change is the sharpest in the horizontal direction okay. This of course is colour-coded now okay red means high, blue means low so this is just simply colour-coded but it is the same idea. Some of you who have worked in fluid mechanics might have seen this or even in other fields. So now you notice this, these are arrows here and the arrows are aligned along the direction of maximum change.

Now if you have a more complex curve something of this sort, once again you can draw the gradient field, why is it the gradient field at any point? I have an F , I have $\text{Del } F \text{ Del } X_1$ and I have $\text{Del } F \text{ Del } X_2$, these 2 put together define a vector and that vector is what is drawn here, longer arrows means higher gradients and shorter arrows means lower gradient. Now one useful way of utilising the gradient vector is as I told you before, you might not only want $\text{Del } F \text{ Del } X$ and $\text{Del } F \text{ Del } Y$, you might also want $\text{Del } F \text{ Del } V$, where V is some other direction. So suppose X and Y are orthogonal and V is a 3rd direction, suppose you want Del

$\nabla f \cdot \mathbf{V}$, what does that mean? Physically it means if I move in the direction \mathbf{V} or $\hat{\mathbf{V}}$, how much will the function change?

And this is fairly easy, all you do is take the gradient which we have defined before, this is $\nabla f = \nabla_x f, \nabla_y f, \dots$ so on and so forth up to $\nabla_{x_n} f$, this vector dotted with the direction \mathbf{V} ok. You can simply see special cases if \mathbf{V} was $\hat{\mathbf{i}}$ or the x_1 direction then gradient in the direction \mathbf{V} should be $\nabla_x f$ which is correct okay, so this retains the meaning of partial derivatives. Similarly, if you take the direction 2 you will get $\nabla_y f$ so on and so forth, so for the coordinate axis this kind of reduces trivially but in the general case you simply take a dot product along that direction.



(Refer Slide Time: 12:41)

Hessian

- The Hessian is the gradient of the gradient
 - It is the equivalent of the second derivative in scalar calculus and has similar uses
- For $f: \mathbb{R}^n \rightarrow \mathbb{R}$, we have $H_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$ is the Hessian which is a $n \times n$ matrix

$$H_{i,j} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

- Note that the Hessian is a symmetric matrix → For real $f \Rightarrow$ Hessian has real eigen values / eigenvectors

So next we come to the idea of hessian, this is basically the gradient of the gradient. Now remember, we will see this once again in the next slide, the gradient is a vector now you are trying to find out how much does this vector change okay as you move in space. Now why would he would use some such complicated quantity because it is equivalent of the 2nd derivative in scalar calculus.

So all the uses that we had for 2nd derivative like finding out whether something is a maximum or minimum all those uses also pass on to the hessian as we will see in some of the videos in this week okay. So suppose F is a function, remember what this means F is taking in, it is a box that takes in a vector as input so \mathbf{X} is a vector and what it gives out is a scalar.

In such a case the hessian is defined as $\nabla^2 F = \nabla_{x_i} \nabla_{x_j} F$ okay, so hessian is a matrix, every entry of the matrix is basically a partial derivative, F is a scalar so first entry is

for example, Del square F Del X 1 square so the N comma 2 entries; Del square F, Del X and Del X 2 so on and so forth, this is N cross N matrix. You can also notice that this is a symmetric matrix, notice that these 2 derivatives are just the same, Del square X Del X 1 Del X 2 is the same as Del X 2 Del X 1. So hessian is a symmetric matrix and from our linear algebra we would know that from real F this means hessian has real eigenvalues okay eigenvectors, so we will use this property little bit later.

(Refer Slide Time: 14:43)

Jacobian

$\nabla \rightarrow \text{Scalar} \rightarrow \text{vector}$
 $\nabla^2 \rightarrow \text{Hessian} \rightarrow \nabla^2 f$


- The Jacobian is the equivalent of the gradient for vector valued functions
 - The Hessian can be seen as the gradient (Jacobian) of a gradient (which is a vector)
- For $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$, we have $J_{i,j} = \frac{\partial^2 f(x)_i}{\partial x_j}$ is the Jacobian which is a $\mathbb{R}^{m \times n}$

$$J = \nabla_x f = \begin{bmatrix} \frac{\partial^2 f_1}{\partial x_1} & \frac{\partial^2 f_1}{\partial x_2} & \dots & \frac{\partial^2 f_1}{\partial x_n} \\ \frac{\partial^2 f_2}{\partial x_1} & \frac{\partial^2 f_2}{\partial x_2} & \dots & \frac{\partial^2 f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f_m}{\partial x_1} & \frac{\partial^2 f_m}{\partial x_2} & \dots & \frac{\partial^2 f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

So one other quantity that we will like to define is that of a Jacobian, the Jacobian is equivalent of a gradient for vector valued functions, so the gradient remember then we had defined simple gradient was from scalar to vector but this assume that the value of the function itself is a scalar. Now you can define a more general case where you have a vector input and a vector output, a Hessian that we just looked at is very similar, you can see this as the Jacobian of a gradient so the hessian took in the gradient of F and gave out Del square F, where F is a scalar but grad of F is now a vector. In general, we define the Jacobian as Del square F of X I by Del X J, remember that since F is now a vector, it is going to have a Ith component.

So in general we are going to have Jacobian which is going to be M cross N, if it takes in a vector, the size of the vector is N cross 1 and it gives out a vector which is M cross 1, so you can write the whole of the Jacobian as a simple matrix okay. So we will be using Jacobian only very rarely, but some general expressions we will show in the next video.

(Refer Slide Time: 16:23)



Taylor Series

- The Taylor series is a local approximation of a function's value in terms of polynomials
 - It is an extremely useful and widely used idea in multiple fields
 - There are mathematical subtleties which we will be ignoring here
- For $f: \mathbb{R} \rightarrow \mathbb{R}$, recall that the Taylor series is written as

$$f(x) \approx f(x^0) + (x - x^0) \frac{df}{dx} + \frac{1}{2} (x - x^0)^2 \frac{d^2 f}{dx^2} + \dots + \frac{1}{6} (x - x^0)^3 \frac{d^3 f}{dx^3} + \dots$$
- For $f: \mathbb{R}^n \rightarrow \mathbb{R}$, the Taylor series can be written as

$$f(x) \approx f(x^0) + (x - x^0)^T g + \frac{1}{2} (x - x^0)^T H (x - x^0) + \dots$$
- Here, $g = \nabla_x f(x^0)$ and H is the Hessian calculated at x^0 also

So the final idea for this video is that of a Taylor's series, the Taylor's series is extremely useful whenever you try to approximate functions. So this is very-very widely used in science, one of the most commonly used ideas in science, practically anything that people use mathematics and calculus for somewhere or the other Taylor's series will pop-up okay, so this is true even of course for machine learning and optimisation okay. So there are of course many subtle things about this Taylor's series we are not going to look at that, we are just going to look at the single slide for Taylor's series and then we will be using it a little bit later both in optimisation as well as in other parts of machine learning ok.

So remember that when you have scalar function one-dimensional function that the kind that we use in school for example, F of $X = e$ to the power X $\sin X$ or something of that sort okay, you can write the Taylor's series as F of X is F at some other point X^0 okay so you want to approximate the value at some value X given that you know the value at X^0 , you also know derivatives at X^0 , etc so this is the basic idea of Taylor's series okay. So if you have F of X , it = F of $X^0 + X - X^0$ times dF/dX , this df/dx is calculated at X equal to $X^0 +$ half of $X - X^0$ square D square F/dX square, this is also calculated at X^0 .

An example of this which you might or might not have realised is our idea of $S = U T +$ half $A T$ square okay, so that is very-very similar to this you know dX is like U and D square F/dX square is like A , this is the time that has elapsed, this is T -square so the half is very similar to that, S is the total distance travelled okay so that is the special case of the Taylor's series and you can easily explained it in the Taylor's series if you have more than the acceleration. So if

only U and A exist then S expression would be what I told you but if you have U , A and what is called the jerk which is the 3rd derivative of this distance with respect to time then you would have that $+ \frac{1}{6} \frac{d^3 F}{dx^3} (x - x_0)^3 + \dots$ so on and so forth.

So the Taylor's series should be familiar to you but most probably you would have not seen it in the case of vectors okay. So in case X instead of being a scalar it is now a vector, you can now write the Taylor's series, notice the similarities between these 2 expressions, F of X now remember X is a vector is F of X_0 which is the same thing $+ \dots$ now notice this is a vector okay. $(X - X_0)^T G$, G is the gradient okay just for compactness I have written this as G , so instead of dF/dX now you have a full gradient, this is a full vector, this effectively is the dot product between one vector and the other okay.

This remember is something that we had discussed earlier, this is called a quadratic form, so we have $(X - X_0)^T H (X - X_0)$, it is still in the scalar case will be equivalent to $(X - X_0)^2$ square okay, but in the vector case you cannot write it as $(X - X_0)^2$ square, it is $(X - X_0)^T H (X - X_0) + \dots$ higher-order terms.

Luckily practically nowhere especially in the vector case do people use this okay, so this is usually maximum that we will go. So we will go to the 1st order term which is the gradient and the second-order term which is the hessian and this is sufficient for most practical purposes okay. So, as I had said earlier we had defined G as the gradient and H as the hessian also calculated at X_0 .

So this is just some preliminaries for a multivariable calculus, we will be using all these ideas only sparingly but you do need it in terms of rebuilding your intuition, so in case if it is not clear please revisit this video a few times. In the next video we will be looking at a few simple mathematical relations in matrix calculus okay like this one that is slightly advanced material.