Many dimensions

In the previous lecture we looked at differenciation, but only in one dimension, $f(x)$ for example. Often we are interested in function of more than dimension, $z(x, y)$, the height of the ground at a point (x, y); the functions of physics, temperature for example, that have a value for every point (x, y, z), the loss function of an deep learning network that depends on million of parameters. All of these and many more. It is natural to ask how differentiation works for functions of more than one variable, the laws of physics depends on it, as does optmization a deep learning network.

In fact, the definitions start out fairly straight forward. The derivative with respect to x tells us the rate of change as x changes, say $f(x, y)$:

$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(x + h, y) - f(x)}{h} \tag{1}$$

Basically the y does nothing while the derivative in x is being calculated. You will notice that the notation has changed slightly, with $\partial f/\partial x$ having the curly d, pronounced "del" or "partial" and we call this the **partial derivative**.

This is an old notation intended to distinguish what we are looking at here, partial derivatives, from the so called total derivative. This happens when one of the variables depends on the other, say, for example you have a path in (x, y) space you might write it as $(x, y(x))$ so changing x will change the function in two ways, one way is because x itself changes and a second because changing x changes y; in this case the derivative is called the **total derivative**; in fact finding the total derivative is not that common and we won't consider it here, but the notational difference between differentiating in one dimension and in many is annoying since it is essentially the same calculation and because there are times when you have a function with variables and parameters and you don't know which notation to use. Anyway, try not to worry about the notation, as is typical in mathematics it is powerful and useful but not as clear cut as you'd expect and at times annoying.

Anyway, obviously we can also define the partial derivative with respect to y:

$$\frac{\partial f}{\partial y} = \lim_{h \to 0} \frac{f(x, y + h) - f(x)}{h} \tag{2}$$

It is sometimes useful to write a shorthand

$$f_x = \frac{\partial f}{\partial x} \tag{3}$$

and

$$f_y = \frac{\partial f}{\partial y} \tag{4}$$

As with the dots and primes in one-dimensional, the proper fraction-like notation is clearly better since it expresses in notation some of the properties of the derivative. However, the other notation is also commonly used because we are lazy.

Here is an example:

$$f(x, y) = \sin x \cos y \tag{5}$$

and $f_x = \cos x \cos y$ whereas $f_y = -\sin x \sin y$. Here is another example

$$f(x, y) = e^{x^2 y} \tag{6}$$

then $f_x = 2xy \exp(x^2 y$ and $f_y = x^2 \exp(x^2 y)$.

Nabla

Obviously, $\partial f/\partial x$ gives the rate of change in the x direction and $\partial f/\partial y$ the rate of change in the y direction. It will often be useful to put these together as a vector, this is called the gradient:

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) \tag{7}$$

The symbol ∇ is often just called "the gradient operator", it is also called "nabla" the name given to it by the mathematician Peter Tait; sometimes it is called "del". The symbol, and the concept of the gradient was introduced by the nineteenth century Irish mathematician William Rowan Hamilton, how also invented the four-dimensional generalization of complex numbers called quoternions. He used the nabla symbol because it is just the Greek letter capital delta, Δ, upside down and was therefore easy for typesetters; the name nabla is from the Greek word for harp. Anyway, whatever its name, here is an example, if $f(x, y) = 2x^3 y^2 + y^2$ then

$$\nabla f = (6x^2 y^2, 4x^3 y + 2y) \tag{8}$$

Often we are interested in how a function $f(x, y)$ changes in some direction that isn't specifically the x or y directions, for this there is the concept of

the **derivative along a vector**:

$$\nabla_{\mathbf{w}} f(x,y) = \lim_{h\to 0} \frac{f(x+hw_1, y+hw_2) - f(x,y)}{h} \tag{9}$$

where $\mathbf{w} = (w_1, w_2)$; often a unit vector is used and then we would refer to the *derivative in the direction of* $\mathbf{w}$. Without proving any theorem, you can hopefully see intuitively that the rate f changes along $\mathbf{w}$ is the rate f is changing in the x direction by the amount of $\mathbf{w}$ in the x direction plus the rate f is changing in the y direction by the amount of $\mathbf{w}$ in the y direction. In short, it can be proved that:

$$\nabla_{\mathbf{w}} f(x,y) = w_1 \frac{\partial f}{\partial x} + w_2 \frac{\partial f}{\partial y} \tag{10}$$

or, written using the dot product

$$\nabla_{\mathbf{w}} f(x,y) = \nabla f \cdot \mathbf{w} \tag{11}$$

This leads to a nice interpretation of the gradient. For two vectors $\mathbf{u}$ and $\mathbf{v}$ the dot product is given by

$$\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos\theta \tag{12}$$

where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$. Thus, for given lengths, the maximum dot product is when the two vectors point in the same direction. Now since

$$\nabla_{\mathbf{w}} f(x,y) = \nabla f \cdot \mathbf{w} \tag{13}$$

this means the direction along which f changes most is the direction of ∇f, in other words, the gradient points in the direction of highest rate of change. If we are thinking of $f(x,y)$ as giving the height of some landscape over coordinates x and y, this means ∇f "points straight up the hill".

Summary

This set of notes revises basic calculus using the old-fashioned notion of infinitessimals. We gave a list of standard derivatives and looked at the chain rule.