Lecture 7: SCALAR AND VECTOR DERIVATIVES *

10-606 MATHEMATICAL FOUNDATIONS FOR MACHINE LEARNING

1 Scalar derivatives

We will use the following notation for scalar derivatives:

- For a function $f \in \mathbb{R} \to \mathbb{R}$, we write $f' \in \mathbb{R} \to R$ for its derivative with respect to its argument. If the argument is called x, we can also write $\frac{d}{dx} f$.
- If a function depends on more than one variable, we write $\frac{\partial}{\partial x} f$ or $\frac{\partial}{\partial y} f$ to indicate a *partial* derivative: the derivative with respect to one variable while holding the others constant.
- Second and higher derivatives are f'', $\ddot{f}$, $\frac{d^2}{dx^2} f$, or $\frac{\partial^2}{\partial x \partial y} f$.
- For a function f, we write $f\big|_{\hat{x}}$ or $f(x)\big|_{x=\hat{x}}$ to represent evaluation at $\hat{x}$. This means the same thing as $f(\hat{x})$ but is sometimes clearer: it lets us keep one name (x) for the variable we are differentiating, and another name ($\hat{x}$) for the value we are substituting at the end.

1.1 Scalar derivative identities

Some most common identities for working with scalar derivatives that you should be familiar with are

- Differentiation and partial differentiation are linear operators, e.g., $(af + bg)' = af' + bg'$.
- Chain rule: if we want $\frac{d}{dx} f(g(x))$, then we use

$$\frac{df}{dx} = \frac{df}{dg}\frac{dg}{dx} \quad \text{or equivalently} \quad \frac{d}{dx} f(g(x)) = f'(g(x))\, g'(x).$$

- Product rule: the derivative of the product of two functions can be computed as $(fg)' = f'g + fg'$.

1.2 Common functions

Here are some useful derivatives of scalar functions. In each expression, x is the variable of interest; all other symbols represent constants.

- The derivative of a constant is zero: $\frac{d}{dx} a = 0$.
- The derivative of a monomial x^k is kx^{k-1}. This works even for negative and fractional values of k. One special case is x^0, where by convention we treat $0x^{-1}$ as equal to zero everywhere.
- The derivative of $\sin x$ is $\cos x$; the derivative of $\cos x$ is $-\sin x$.
- The derivative of e^{ax} is ae^{ax}. If we're using some other base b, we rewrite $b^x = e^{x \ln b}$ and then use the identity above.
- The derivative of $\ln x$ is x^{-1}. Again we can easily switch to another base: $\log_b x = \ln x / \ln b$.

*Compiled on Saturday 1st July, 2023 at 16:45

Knowledge check

1. **Math**: What is the derivative of $f(x) = 3x^7 + 5\sin x + \ln x$?
   - **Answer**: $\frac{df}{dx} = 21x^2 + 5\cos x + \frac{1}{x}$
2. **Math**: What is the derivative of $f(x) = \exp(\frac{\cos x}{x^2})$?
   - **Answer**: $\frac{df}{dx} = \exp(\frac{\cos x}{x^2})(\frac{\sin x}{x^2} - \frac{2\cos x}{x^3})$

2 Vector-valued derivatives

It's also useful to think about functions that return vectors or take vectors as arguments. If f is a vector-valued function of a real argument, $f \in \mathbb{R} \to \mathbb{R}^n$, we can write it as a vector whose components are real-valued functions,

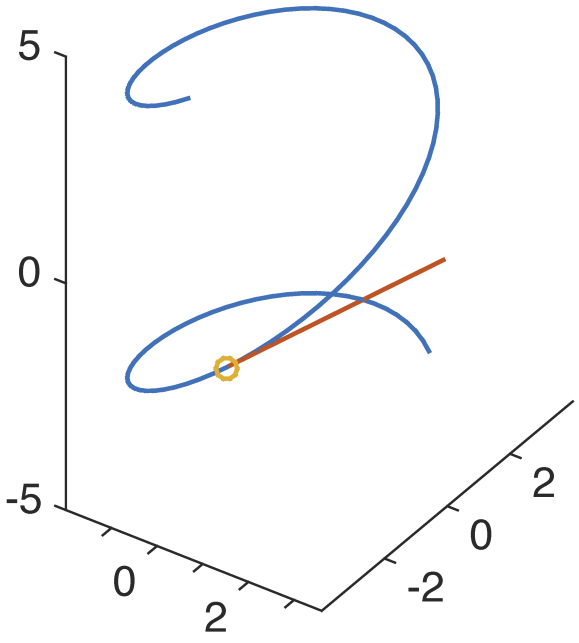
$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{bmatrix}$$

Its derivative is then also a vector-valued function, of the same shape as f. Its components are the derivatives of the component functions:

$$\frac{d}{dx} f = \begin{bmatrix} \frac{df_1}{dx} \\ \frac{df_2}{dx} \\ \vdots \\ \frac{df_n}{dx} \end{bmatrix}$$

We can think of f as representing a curve in $\mathbb{R}^n$. The derivative $\frac{df}{dx}$ represents a tangent vector to this curve: the instantaneous velocity of a point moving along the curve as the argument x changes at a unit rate. The length of the tangent vector tells us the speed of the point, and the components tell us its direction.

Here's an example of a function in $\mathbb{R} \to \mathbb{R}^3$ and its derivative at a particular point:



Note that this plot doesn't show the argument x explicitly: instead it is implicit in the position of the point along the curve. If we wanted to show x explicitly, we could color the curve or add grid marks to show what values of x correspond to what values of $f(x)$.

3 Gradients

Now suppose that $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value. Then the *gradient* of f with respect to A is the matrix of partial derivatives, defined as:

$$\nabla_A f(A) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

i.e., an $m \times n$ matrix with

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$$

Note that the size of $\nabla_A f(A)$ is always the same as the size of A. So if A is just a vector $x \in \mathbb{R}^n$, then

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

It is very important to remember that the gradient of a function is defined only if the function is real-valued, that is, if it returns a scalar value. We can not, for example, take the gradient of $Ax, A \in \mathbb{R}^{n \times n}$ with respect to x, since this quantity is vector-valued. Instead, we would need to apply the notation of the previous section and compute vector-valued derivatives.

It follows directly from the linearity of partial derivatives that $\nabla_x(af(x) + bg(x)) = a\nabla_x f(x) + b\nabla_x g(x)$.

3.1 Chain rule for vectors

With the above notation, the chain rule for vector functions looks similar to how it did for scalar functions. Suppose $f \in \mathbb{R}^n \to \mathbb{R}$ takes multiple arguments and $g \in \mathbb{R} \to \mathbb{R}^n$ returns multiple values, so that $f(g(x))$ makes sense. Then we have

$$\frac{df}{dx} = \frac{df}{dg}^T \frac{dg}{dx}$$

This looks nearly identical to the scalar chain rule, with the addition of a transpose in order to make the matrix multiplication make sense (note, this is purely a byproduct of how we have chosen to represent these vector derivatives and our notation choice that vectors are by default column vectors). If we write out the dot product, we get

$$\frac{df}{dx} = \sum_{i=1}^{n} \frac{\partial f}{\partial g_i} \frac{dg_i}{dx}$$

which may be familiar as the rule for calculating the *total derivative* of f with respect to x. In words, to calculate the change in f, we sum up the effects of all of the changes in all of the inputs to f.

Knowledge check

1. **Math**: What is the gradient of $f(x) = x_1 x_2^2 x_3^3$?
   - **Answer**:
   $$\nabla_x f(x) = \begin{bmatrix} x_2^2 x_3^3 \\ 2x_1 x_2 x_3^3 \\ 3x_1 x_2^2 x_3^2 \end{bmatrix}.$$
2. **Math**: Let $g(x) = \begin{bmatrix} g_1(x) = \sin x \\ g_2(x) = \ln x \end{bmatrix}$ and let $f(g) = g_1 + g_1 g_2 + g_2$. What is $\frac{df}{dx}$
   - **Answer**:
   $$\frac{df}{dg} = \begin{bmatrix} 1 + g_2 \\ g_1 + 1 \end{bmatrix} = \begin{bmatrix} 1 + \ln x \\ 1 + \sin x \end{bmatrix} \text{ and } \nabla_x f(x) = \begin{bmatrix} \cos x \\ \frac{1}{x} \end{bmatrix}$$
   $$\text{Thus } \frac{df}{dx} = \begin{bmatrix} 1 + \ln x \\ 1 + \sin x \end{bmatrix}^T \begin{bmatrix} \cos x \\ \frac{1}{x} \end{bmatrix} = \cos x + \cos x \ln x + \frac{1 + \sin x}{x}$$