INFR10086 Machine Learning (MLG) Semester 1, 2022/23

Tutorial 1: Vector-Matrix Calculus

In this tutorial, we will briefly review functions of vectors and matrices and their derivatives. Because the notations can sometimes be confusing, we will put extra emphasis on types.

1 Type Theory

Types are used to specify the possible values a variable can take. For example, we can say that a variable x is a real number, and this is often written as $x \in \mathbb{R}$ or $x : \mathbb{R}$. For the purpose of this course, types are synonymous to sets. For example, $\mathbb{R}$ is the set of all possible real numbers, and it is also a type.

Other commonly used types are vectors and matrices, often written as $\mathbb{R}^d$ for d-dimensional vectors, and $\mathbb{R}^{m \times n}$ for $m \times n$ matrices.

Types of functions are defined by the input and output types. For example, $\mathcal{X} \to \mathcal{Y}$ is a function type with input type $\mathcal{X}$ and output type $\mathcal{Y}$. If a function f is of type $\mathcal{X} \to \mathcal{Y}$ or simply $f : \mathcal{X} \to \mathcal{Y}$, then we know $f(x)$ is of type $\mathcal{Y}$ if x is of type $\mathcal{X}$.

Sometimes a function can take a function as input and returns another function. In other words, the input and output types can both be functions. For example, if we have $T : (\mathbb{R}^d \to \mathbb{R}) \to (\mathbb{R} \to \mathbb{R})$, for a function $f : \mathbb{R}^d \to \mathbb{R}$, $T(f)$ is a function $\mathbb{R} \to \mathbb{R}$.

2 Derivatives

For a function $f : \mathbb{R} \to \mathbb{R}$, the **derivative** of f at x is defined as

$$Df(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}. \tag{1}$$

Note that Df is a function, and $Df(x)$ is the evaluation of Df at x. The term $Df(x)$ should be parsed as $(Df)(x)$, not $D(f(x))$.

Discussion. What is the type of Df?

For multivariate functions $f : \mathbb{R}^d \to \mathbb{R}$, the concept of derivative can be extended to **directional derivative**. Since we only have a definition of derivative on the type $\mathbb{R} \to \mathbb{R}$, we need to construct a function of type $\mathbb{R} \to \mathbb{R}$ to make use of the definition of derivative. A simple approach is to define $g(t) = f(x + tv)$ for a particular point $x \in \mathbb{R}^d$ and a particular direction $v \in \mathbb{R}^d$. The function g is $\mathbb{R} \to \mathbb{R}$, and Dg is defined. We also do not care much about the derivative for different t's, so we can limit ourselves to the point x, i.e., when $t = 0$. This leads to the definition of the directional

derivative. The derivative at $x \in \mathbb{R}^d$ along a direction $v \in \mathbb{R}^d$ is defined as

$$D_v f(x) = Dg(0) = \lim_{h \to 0} \frac{f(x + (0+h)v) - f(x+0v)}{h} = \lim_{h \to 0} \frac{f(x+hv) - f(x)}{h}.$$

Discussion. What is the type of $D_v f$?

Recall that the standard basis $\{e_1, e_2, \ldots, e_d\}$ for $\mathbb{R}^d$ is

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \ldots. \tag{2}$$

We will write

$$D_{e_1} f = \frac{\partial f}{\partial x} \qquad D_{e_2} f = \frac{\partial f}{\partial y} \qquad D_{e_3} f = \frac{\partial f}{\partial z}$$

Discussion. What is the type of a basis for $\mathbb{R}^d$?

Discussion. What is the type of $\frac{\partial f}{\partial x}$?

Since $D_{e_1} f$ is a function of $\mathbb{R}^d \to \mathbb{R}$, it is perfectly valid to ask what the directional derivatives of $D_{e_1} f$ are and write $D_v D_{e_1} f$. For example, we will write

$$D_{e_2} D_{e_1} f = \frac{\partial^2 f}{\partial y \partial x} \qquad D_{e_1} D_{e_2} f = \frac{\partial^2 f}{\partial x \partial y}.$$

Discussion. We can treat D_v as an operator on functions. What is type of D_v?

3 Taylor Series

Let's focus again on $f : \mathbb{R} \to \mathbb{R}$. Suppose we would like to approximate $f(x)$ locally at $x = 0$ with an n-degree polynomial $a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$. We can write

$$f(x) \approx a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n. \tag{3}$$

By repeatedly differentiating both sides, we have

$$f(0) = a_0 \qquad f'(0) = a_1 \qquad f''(0) = (2!)a_2 \qquad \ldots \qquad f^{(n)}(0) = (n!)a_n. \tag{4}$$

In other words,

$$a_0 = f(0) \qquad a_1 = f'(0) \qquad a_2 = \frac{f''(0)}{2!} \qquad \dots \qquad a_n = \frac{f^{(n)}(0)}{n!}. \tag{5}$$

If we plug the coefficients back, we can say that $f(x)$ at $x = 0$ can be approximated by

$$f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n, \tag{6}$$

a series known as the Taylor series. We can apply this approximation at any point $x = a$, and the approximation would be

$$f(x) \approx \sum_{i=0}^{n} \frac{f^{(i)}(a)}{i!}(x-a)^i, \tag{7}$$

where we conveniently define $f^{(0)}(a) = f(a)$ and $0! = 1$. Since the series is only an approximation, to make it an actual equality, we can have

$$f(x) = \sum_{i=0}^{n} \frac{f^{(i)}(a)}{i!}(x-a)^i + E_n(x), \tag{8}$$

where $E_n(x)$ is a remainder term that depends on n, the degree of the polynomial.

For example, we can write

$$f(x) = f(a) + f'(a)(x-a) + E_1(x). \tag{9}$$

In words, we approximate $f(x)$ at $x = a$ locally with a linear function (literally a line) $f(a) + f'(a)(x-a)$.

4 Gradients and Hessians

We can have approximation for multivariate functions of type $\mathbb{R}^d \to \mathbb{R}$. The extension is again based on a direction v at the point x. In particular, we *hope* to get

$$f(x+v) = f(x) + T_x(v) + \|v\|E_1(x,v), \tag{10}$$

where $T_x(v)$ is a linear transformation and $E_1(x,v) \to 0$ when $\|v\| \to 0$. Note that we only hope that this happens. Let's derive $T_x(v)$ if this were to happen. We first substitute $v = hu$ for another vector $u \in \mathbb{R}^d$ and get

$$f(x+hu) = f(x) + T_x(hu) + \|hu\|E_1(x,hu). \tag{11}$$

We now intentionally rearrange the equation into

$$\frac{f(x+hu) - f(x)}{h} = \frac{T_x(hu)}{h} + \frac{\|hu\|}{h}E_1(x,hu) \tag{12}$$

$$= T_x(u) + \frac{|h|}{h}\|u\|E_1(x,hu) \tag{13}$$

where we use the fact that $T_x(u)$ is linear in u, i.e., $T_x(hu) = hT_x(u)$. If we let $h \to 0$ on both sides, because $E(x, hu) \to 0$ when $h \to 0$, we have

$$D_u f(x) = T_x(u). \tag{14}$$

Now we can express u as $\sum_{i=1}^{d} u_i e_i$, a weighted sum of the standard basis. Because of linearity of T_x, we have

$$T_x\left(\sum_{i=1}^{d} u_i e_i\right) = \sum_{i=1}^{d} u_i T_x(e_i) = \sum_{i=1}^{d} u_i D_{e_i} f(x) \tag{15}$$

$$= \begin{bmatrix} D_{e_1} f(x) & D_{e_2} f(x) & \cdots & D_{e_d} f(x) \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix} = \nabla f(x)^\top u, \tag{16}$$

where

$$\nabla f(x) = \begin{bmatrix} D_{e_1} f(x) \\ D_{e_2} f(x) \\ \vdots \\ D_{e_d} f(x) \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_d}(x) \end{bmatrix} \tag{17}$$

is called the **gradient** of f at x. The linear transformation T_x is also known as the **total derivative** at x. In sum,

$$D_u f(x) = T_x(u) = \nabla f(x)^\top u. \tag{18}$$

Discussion. What is the type of T_x?

Discussion. What is the type of ∇f?

Discussion. What is the type of ∇ if we treat it as an operator?

We can similarly derive

$$f(x + v) = f(x) + \nabla f(x)^\top v + \frac{1}{2!} v^\top H(x) v + \|v\|^2 E_2(x, v), \tag{19}$$

where $E_2(x, v) \to 0$ when $\|v\| \to 0$ and

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(x) & \frac{\partial^2 f}{\partial x_d \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d}(x) \end{bmatrix} \tag{20}$$

is called the **Hessian** of f at x. The Hessian of f is also sometimes written as $\nabla^2 f$, though it should never be interpreted as $\nabla(\nabla f)$. Deriving this is beyond the scope of this tutorial.

Discussion. What is the type of H or $\nabla^2 f$?

Discussion. Is it type-correct to write $\nabla(\nabla f)$? If so, what's the type of $\nabla(\nabla f)$?

5 Matrix Calculus

Writing gradients and Hessians the way they are written quickly becomes cumbersome when there are many vectors and matrices involved. To simplify notation, when $f : \mathbb{R}^d \to \mathbb{R}$, we define

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}, \tag{21}$$

a derivative of f with respect to a vector.

Discussion. When f is of type $\mathbb{R}^d \to \mathbb{R}$, what is the type of $\frac{\partial f}{\partial x}$?

For example, we know that

$$\frac{\partial x^\top a}{\partial x_n} = \frac{\partial}{\partial x_n}\left(\sum_{i=1}^{d} a_i x_i\right) = a_n. \tag{22}$$

Using the shorthand, we can now write

$$\frac{\partial x^\top a}{\partial x} = \begin{bmatrix} \frac{\partial x^\top a}{\partial x_1} \\ \frac{\partial x^\top a}{\partial x_2} \\ \vdots \\ \frac{\partial x^\top a}{\partial x_d} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} = a \tag{23}$$

Discussion. We just wrote $\frac{\partial x^\top a}{\partial x} = a$, but a is a vector. Shouldn't $\frac{\partial x^\top a}{\partial x}$ be a function of $\mathbb{R}^d \to \mathbb{R}^d$?

As another example, we know that

$$\frac{\partial x^\top Ax}{\partial x_n} = \frac{\partial}{\partial x_n}\left(\sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j A_{ij}\right) = 2A_{nn}x_n + \sum_{j\neq n} x_j A_{nj} + \sum_{i\neq n} x_i A_{in} + 0 \tag{24}$$

$$= \sum_{j=1}^{d} x_j A_{nj} + \sum_{i=1}^{d} x_i A_{in} = A_{n\cdot}x + A_{\cdot n}^\top x, \tag{25}$$

where $A_{\cdot n}$ is the n-th column of A and $A_{n\cdot}$ is the n-th row of A. Using the shorthand, we can now write

$$\frac{\partial x^\top Ax}{\partial x} = \begin{bmatrix} \frac{\partial x^\top Ax}{\partial x_1} \\ \frac{\partial x^\top Ax}{\partial x_2} \\ \vdots \\ \frac{\partial x^\top Ax}{\partial x_d} \end{bmatrix} = \begin{bmatrix} A_{1\cdot}x + A_{\cdot 1}^\top x \\ A_{2\cdot}x + A_{\cdot 2}^\top x \\ \vdots \\ A_{d\cdot}x + A_{\cdot d}^\top x \end{bmatrix} = (A + A^\top)x \tag{26}$$

We can also extend the notation to derivatives with respect to a matrix. Specifically, we define

$$\frac{\partial f}{\partial X} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1d}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2d}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f}{\partial x_{d1}} & \frac{\partial f}{\partial x_{d2}} & \cdots & \frac{\partial f}{\partial x_{dd}} \end{bmatrix}. \tag{27}$$

Since

$$\frac{\partial a^\top Xb}{\partial x_{mn}} = \frac{\partial}{\partial x_{mn}}\left(\sum_{i=1}^{d}\sum_{j=1}^{d} a_i b_j X_{ij}\right) = a_m b_n, \tag{28}$$

we can now write

$$\frac{\partial a^\top Xb}{\partial X} = \begin{bmatrix} \frac{\partial a^\top Xb}{\partial x_{11}} & \frac{\partial a^\top Xb}{\partial x_{12}} & \cdots & \frac{\partial a^\top Xb}{\partial x_{1d}} \\ \frac{\partial a^\top Xb}{\partial x_{21}} & \frac{\partial a^\top Xb}{\partial x_{22}} & \cdots & \frac{\partial a^\top Xb}{\partial x_{2d}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial a^\top Xb}{\partial x_{d1}} & \frac{\partial a^\top Xb}{\partial x_{d2}} & \cdots & \frac{\partial a^\top Xb}{\partial x_{dd}} \end{bmatrix} = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_d \\ a_2b_1 & a_2b_2 & \cdots & a_2b_d \\ \vdots & \vdots & & \vdots \\ a_db_1 & a_db_2 & \cdots & a_db_d \end{bmatrix} = ab^\top. \tag{29}$$