The Mathematical Engineering of Deep Learning

Book Draft

Benoit Liquet, Sarat Moka and Yoni Nazarathy

June 15, 2022

A. Some Multivariable Calculus - DRAFT

This appendix provides key results and notation from multivariable calculus. It is not an exhaustive summary of multi-variable calculus but rather contains the results needed for content of the book.

A.1. Vectors and Functions in $\mathbb{R}^n$

Denote the set of all the real numbers by $\mathbb{R}$ and the real coordinate space of dimension n by $\mathbb{R}^n$. Each element of $\mathbb{R}^n$ is an n dimensional vector, interpreted as a column of the form

$$u = (u_1, \ldots, u_n) = [u_1, \ldots, u_n]^\top = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}.$$

The *Euclidean norm* of $u \in \mathbb{R}^n$, measuring the geometric length of u and also known as the L_2 norm, is

$$\|u\|_2 = \sqrt{u^\top u} = \left(\sum_{i=1}^{n} u_i^2 \right)^{1/2}.$$

Here the scalar $u^\top v$ is the *inner product* between two vectors $u, v \in \mathbb{R}^n$. The Euclidean norm is a special case of the L_p norm which is defined via,

$$\|u\|_p = \left(\sum_{i=1}^{n} |u_i|^p \right)^{1/p},$$

for $p \geq 1$. When p in $\|\cdot\|_p$ is not specified, we interpret $\|\cdot\|$ as the L_2 norm.

Focusing on the L_2 norm and the inner product $u^\top v$, the *Cauchy-Schwartz* inequality is,

$$|u^\top v| \leq \|u\| \|v\|, \tag{A.1}$$

where the two sides are equal if and only if u and v are linearly dependent (that is, $u = c\,v$ for some $c \in \mathbb{R}$). Also, the *Euclidean distance* (or, simply the *distance*) between u and v is defined as

$$\|u - v\| = \left(\sum_{i=1}^{n} (u_i - v_i)^2 \right)^{1/2}.$$

An important consequence of the Cauchy-Schwartz inequality is that the Euclidean norm satisfies the *triangle inequality*: For any $u, v \in \mathbb{R}^n$,

$$\|u + v\| \leq \|u\| + \|v\|. \tag{A.2}$$

To see this, observe that

$$\begin{aligned}
\|u + v\|^2 &= \|u\|^2 + \|v\|^2 + 2u^\top v \\
&\leq \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| \\
&= (\|u\| + \|v\|)^2 .
\end{aligned}$$

Convergence of a sequence of vectors can be defined via scalar converge of the distance. That is, a sequence of vectors $u^{(1)}, u^{(2)}, \ldots$ in $\mathbb{R}^n$ is said to *converge* to a vector $u \in \mathbb{R}^n$, denoted via $\lim_{k\to\infty} u^{(k)} = u$, if

$$\lim_{k\to\infty} \|u^{(k)} - u\| = 0.$$

That is, if for every $\epsilon > 0$ there exists an N_0 such that for all $k \geq N_0$,

$$\|u^{(k)} - u\| < \epsilon.$$

Let $f : \mathbb{R}^n \to \mathbb{R}$ be an n-dimensional multivariate function that maps each vector $u = (u_1, \ldots, u_n)^\top \in \mathbb{R}^n$ to a real number. Then, the function is said to be *continuous* at $u \in \mathbb{R}^n$ if for any sequence $u^{(1)}, u^{(2)}, \ldots$ such that $\lim_{k\to\infty} u^{(k)} = u$, we have that

$$\lim_{k\to\infty} f(u^{(k)}) = f(u).$$

Alternatively, f is continuous at $u \in \mathbb{R}^n$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$|f(u) - f(v)| < \epsilon,$$

for every $v \in \mathbb{R}^n$ with $\|u - v\| < \delta$. Continuity of f at u implies that the values of f at u and at v can be made arbitrarily close by setting the point v to be arbitrarily close to u.

We can extend the above continuity definitions to multivariate vector valued functions of the form $f : \mathbb{R}^n \to \mathbb{R}^m$ that map every n dimensional real-valued vector to an m dimensional real-valued vector. Such functions can be written as

$$f(u) = [f_1(u), \ldots, f_m(u)]^\top, \tag{A.3}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ for each $i = 1, \ldots, m$. Then, the function f is called continuous at u if each f_i is continuous at u.

We say that the function f is *continuous on* a set $\mathcal{U} \subseteq \mathbb{R}^n$ if f is continuous at *each point* in $\mathcal{U}$.

A.2. Derivatives

Consider an n-dimensional multivariate function $f : \mathbb{R}^n \to \mathbb{R}$. The *partial derivative* $\frac{\partial f(u)}{\partial u_i}$ of f with respect u_i is the derivative taken with respect to the variable u_i while keeping all other variables constant. That is,

$$\frac{\partial f(u)}{\partial u_i} = \lim_{h \to 0} \frac{f(u_1, \ldots, u_{i-1}, u_i + h, u_{i+1}, \ldots, u_n) - f(u)}{h}. \tag{A.4}$$

Suppose that the partial derivative (A.4) exists for all $i = 1, \ldots, n$. Then the *gradient* of f at u, denoted by $\nabla f(u)$ or $\frac{\partial f(u)}{\partial u}$, is a concatenation of the partial derivatives of f with respect to all its variables, and it is expressed as a vector:

$$\nabla f(u) = \frac{\partial f(u)}{\partial u} = \left[\frac{\partial f(u)}{\partial u_1}, \ldots, \frac{\partial f(u)}{\partial u_n}\right]^\top. \tag{A.5}$$

The gradient $\nabla f(u)$ is a vector capturing the direction of the steepest ascent at u. Further, $h||\nabla f(u)||$ is the increase in f when moving in that direction for infinitesimal distance h.

In some situations, instead of a vector form, variables of the function are represented as a matrix. In that scenario, multivariate functions are of form $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, that is, f maps matrices $U = (u_{i,j})$ of dimension $n \times m$ to real values $f(U)$. If the partial derivative $\frac{\partial f(U)}{\partial u_{i,j}}$ exists for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$, it is convenient to use the notation $\frac{\partial f(U)}{\partial U}$ to denote the collection of the partial derivatives of f with respect to all its variables as a matrix of the same dimension $n \times m$:

$$\frac{\partial f(U)}{\partial U} = \begin{bmatrix} \frac{\partial f(U)}{\partial u_{1,1}} & \cdots & \frac{\partial f(U)}{\partial u_{1,m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(U)}{\partial u_{n,1}} & \cdots & \frac{\partial f(U)}{\partial u_{n,m}} \end{bmatrix}. \tag{A.6}$$

Directional Derivatives

The *directional derivative* of $f : \mathbb{R}^n \to \mathbb{R}$ at u in the direction $v \in \mathbb{R}^n$ is the scalar defined by

$$\nabla_v f(u) = \lim_{h \to 0} \frac{f(u + hv) - f(u)}{h}.$$

The directional derivative generalizes the notion of the partial derivative. In fact, the partial derivative $\frac{\partial f(u)}{\partial u_i}$ is the directional derivative at u in the direction of the vector e_i which consists of 1 at the i-th coordinate and zeros everywhere else. This simply follows from the observation that

$$\nabla_{e_i} f(u) = \lim_{h \to 0} \frac{f(u_1, \ldots, u_{i-1}, u_i + h, u_{i+1}, \ldots, u_n) - f(u)}{h} = \frac{\partial f(u)}{\partial u_i}.$$

As consequence, if the gradient of f exists at u, the directional derivative exists in every direction v and we have

$$\nabla_v f(u) = v^\top \nabla f(u). \tag{A.7}$$

One way to see (A.7) in the case of continuity of the partial derivatives is via a Taylor's theorem based first-order approximation (see Theorem 2):

$$f(u + hv) = f(u) + (h\, v)^\top \nabla f(u) + O(h^2),$$

where $O(h^k)$ denotes a function such that $O(h^k)/h^k$ goes to a constant as $h \to 0$. Thus,

$$\frac{f(u + hv) - f(u)}{h} = v^\top \nabla f(u) + O(h).$$

Now take the limit $h \to 0$ on both the sides to get (A.7).

It is useful to note that the directional derivative $\nabla_v f(u)$ is maximum in the direction of the gradient in the sense that for all unit length vectors v, the choice $v = \nabla f(u)/\|\nabla f(u)\|$ maximizes $\|\nabla_v f(u)\|$. This is a consequence of the Cauchy-Schwartz inequality (A.1):

$$|\nabla_v f(u)| = |v^\top \nabla f(u)| \leq \|v\|\|\nabla f(u)\| = \|\nabla f(u)\|.$$

Setting $v = \nabla f(u)/\|\nabla f(u)\|$ achieves the equality.

Jacobians

The Jacobian is useful for functions of the form $f : \mathbb{R}^n \to \mathbb{R}^m$ as in (A.3) where each f_i is a real-valued function of u. The *Jacobian* of f at u, denoted by J_f, is the $m \times n$ matrix defined via

$$J_f(u) = \begin{bmatrix} \frac{\partial f_1(u)}{\partial u_1} & \cdots & \frac{\partial f_1(u)}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(u)}{\partial u_1} & \cdots & \frac{\partial f_m(u)}{\partial u_n} \end{bmatrix}.$$

In other words, the i-th row of the Jacobian is the gradient $\nabla f_i(u)$. In some situations, it is convenient to use the notation $\frac{\partial f(u)}{\partial u}$ to denote the transpose of the Jacobian of f at u. That is,

$$\frac{\partial f(u)}{\partial u} = \left(J_f(u)\right)^\top.$$

Hessians

Returning to functions of the form $f : \mathbb{R}^n \to \mathbb{R}$, to describe the curvature of the function f at a given $u \in \mathbb{R}^n$, it is important to consider the second-order partial derivatives at u. These partial derivatives are arranged as an $n \times n$

matrix, called the *Hessian* and defined by

$$\nabla^2 f(u) = \frac{\partial \nabla f(u)}{\partial u} = \begin{bmatrix} \frac{\partial^2 f}{\partial u_1^2} & \frac{\partial^2 f}{\partial u_1 \partial u_2} & \cdots & \frac{\partial^2 f}{\partial u_1 \partial u_n} \\ \frac{\partial^2 f}{\partial u_2 \partial u_1} & \frac{\partial^2 f}{\partial u_2^2} & \cdots & \frac{\partial^2 f}{\partial u_2 \partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial u_n \partial u_1} & \frac{\partial^2 f}{\partial u_n \partial u_2} & \cdots & \frac{\partial^2 f}{\partial u_n^2} \end{bmatrix}, \tag{A.8}$$

where $\frac{\partial^2 f}{\partial u_i \partial u_j} = \frac{\partial f}{\partial u_i}\left(\frac{\partial f}{\partial u_j}\right)$. Note that if all the second order partial derivatives are continuous at u, then the Hessian $\nabla^2 f(u)$ is a symmetric matrix. That is, for all $i, j \in \{1, \dots, n\}$,

$$\frac{\partial^2 f}{\partial u_i \partial u_j} = \frac{\partial^2 f}{\partial u_j \partial u_i}.$$

This result is known as *Schwarz's theorem* or *Clairaut's theorem*. Observe that using the Jacobian, we can treat the Hessian as the Jacobian of the gradient vector. That is,

$$\nabla^2 f(u) = J_{\nabla f}(u).$$

Certain attributes of optimization methods in Chapter 4 are often defined via *positive (semi) definiteness* of the Hessian $\nabla^2 f(\theta)$ at θ. In particular, a symmetric matrix A is said to be *positive semidefinite* if for all $\phi \in \mathbb{R}^d$,

$$\phi^\top A\, \phi \geq 0. \tag{A.9}$$

Furthermore, A is said to be *positive definite* if the inequality in (A.9) is strict for all $\phi \in \mathbb{R}^d \setminus \{0\}$. Note that the matrix A is called is *negative semidefinite* (respectively, *negative definite*) when $-A$ is positive semidefinite (respectively, positive definite).

Differentiability

A multivariate vector valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *differentiable* at $u \in \mathbb{R}^n$ if there is an $m \times n$ dimensional matrix A such that

$$\lim_{v \to u} \left(\frac{\|f(u) - f(v) - A(u - v)\|}{\|u - v\|} \right) = 0.$$

Here the limit notation $v \to u$ implies that the limit exists for every sequence $\{v^{(k)} : k \geq 1\}$ such that $\lim_{k\to\infty} v^{(k)} = u$. The matrix A is called *derivative*. If the function f is differentiable at θ, then the derivative at θ is equal to the Jacobian $J_f(\theta)$. In particular, if f is real-valued function (that is, $m = 1$) and differentiable at θ, then the derivative at θ is $\nabla f(\theta)^\top$. If the derivative is continuous on a set $\mathcal{U} \subseteq \mathbb{R}^n$, we say that f is *continuously differentiable* on $\mathcal{U}$, and in that case all the partial derivatives $\frac{\partial f(u)}{\partial u_i}$ are continuous on $\mathcal{U}$.

A.3. The Multivariable Chain Rule

Consider a multivariate vector valued function $h : \mathbb{R}^n \to \mathbb{R}^k$ and a multivariate real-valued function $g : \mathbb{R}^k \to \mathbb{R}$. Suppose that h is differentiable at $u \in \mathbb{R}^n$ and g is differentiable at $h(u) = [h_1(u), \ldots, h_k(u)]^\top$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be the composition $f = g \circ h$ or $f(u) = g(h(u))$. For each $i = 1, \ldots, n$, the *multivariate chain rule* is,

$$\frac{\partial f(u)}{\partial u_i} = \frac{\partial g(h(u))}{\partial v_1} \frac{\partial h_1(u)}{\partial u_i} + \cdots + \frac{\partial g(h(u))}{\partial v_k} \frac{\partial h_k(u)}{\partial u_i},$$

where $\frac{\partial g}{\partial v_i}$ denotes the partial derivative of g with respect to the i-th coordinate. Thus,

$$\frac{\partial f(u)}{\partial u_i} = \begin{bmatrix} \frac{\partial h_1(u)}{\partial u_i} & \cdots & \frac{\partial h_k(u)}{\partial u_i} \end{bmatrix} \nabla g\big(h(u)\big),$$

and combining for all $i = 1, \ldots, n$,

$$\nabla f(u) = J_h(u)^\top \nabla g\big(h(u)\big).$$

Now consider the case where g is also a multivariate vector valued function. That is, suppose $h : \mathbb{R}^n \to \mathbb{R}^k$ is differentiable at u and $g : \mathbb{R}^k \to \mathbb{R}^m$ is differentiable at $h(u)$. Then the composition $f = g \circ h : \mathbb{R}^n \to \mathbb{R}^m$ is a vector valued function with Jacobian,

$$J_f(u) = J_g\big(h(u)\big) J_h(u).$$

A.4. Taylor's Theorem

Once again consider a multivariate real-valued function $f : \mathbb{R}^n \to \mathbb{R}$. If all the k-order derivatives of f are continuous at a point $u \in \mathbb{R}^n$, then *Taylor's theorem* offers an approximation for f within a neighborhood of u in terms of these derivatives. We are particularly interested in cases where $k = 1$ and $k = 2$ as they are crucial in implementation of, respectively, the first-order and the second order optimization methods. It is easy to understand the theorem when the function f is univariate. Hence we start with the univariate case and then move to the general multivariate case. We omit the proof of Taylor's theorem as it is a well known result that can be found in any standard multivariate calculus textbook.

Univariate Case

Suppose that $n = 1$, that is, f is a univariate real-valued function. We say that f is *k-times continuously differentiable* on an open interval $\mathcal{U} \subseteq \mathbb{R}$ if f is k times differentiable at every point on $\mathcal{U}$ (i.e., the k-th order derivative $\frac{\mathsf{d}^k f(u)}{\mathsf{d}u^k}$ exists for all $u \in \mathcal{U}$) and $\frac{\mathsf{d}^k f(u)}{\mathsf{d}u^k}$ is continuous on $\mathcal{U}$. If $k = 0$, we interpret $\frac{\mathrm{d}^k f(u)}{\mathrm{d}u^k}$ simply as $f(u)$.

Theorem 2 (**Taylor's Theorem in $\mathbb{R}$**). *Let $f : \mathbb{R} \to \mathbb{R}$ be a k-times continuously differentiable on an open interval $\mathcal{U} \subseteq \mathbb{R}$. Then, for any $u, v \in \mathcal{U}$,*

$$f(u) = \sum_{i=0}^{k} \frac{\mathrm{d}^i f(v)}{\mathrm{d}u^i} \frac{(u-v)^i}{i!} + O\left((u-v)^{k+1}\right). \tag{A.10}$$

The polynomial

$$P_k(u) = \sum_{i=0}^{k} \frac{(u-v)^i}{i!} \frac{\mathrm{d}^i f(v)}{\mathrm{d}u^i}$$

appeared in (A.11) is called k-th order *Taylor polynomial.* Since the remainder

$$R_k(u) = f(u) - P_k(u) \longrightarrow 0, \quad \text{as}\, x \to a,$$

$f(u)$ is approximately equal to $P_k(u)$ for u within a small neighborhood of a. Particularly, for a point u near v, $P_1(u)$ is *linear approximation* of $f(u)$ and $P_2(u)$ is *quadratic approximation* of $f(u)$.

Multivariate Case

Now consider the multivariate case, that is, f is a multivariate real-valued function. In order to state Taylor's theorem for this case, we need some new notion that is relevant only here.

An n-tuple $\alpha = (\alpha_1, \ldots, \alpha_n)$ is called *multi-index* if each α_i is non-negative integer. For a multi-index α, let

$$|\alpha| = \sum_{i=1}^{n} \alpha_i, \quad \alpha! = \alpha_1! \cdots \alpha_n!, \quad \text{and} \quad u^\alpha = u_1^{\alpha_1} \cdots u_n^{\alpha_n},$$

for any $u \in \mathbb{R}^n$. Then, the higher order partial derivatives are expressed as

$$D^\alpha f(u) = \frac{\partial^{|\alpha|} f(u)}{\partial u_1^{\alpha_1} \cdots \partial u_n^{\alpha_n}}.$$

We say that f is *k-times continuously differentiable* on an open set $\mathcal{U} \subseteq \mathbb{R}^n$ if all the higher order partial derivatives $D^\alpha f(u)$ exists and are continuous on $\mathcal{U}$ for all multi-index α such that $|\alpha| \le k$.

Theorem 3 (**Taylor's Theorem in $\mathbb{R}^n$**). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a k-times continuously differentiable on an open set $\mathcal{U} \subseteq \mathbb{R}^n$. Then, for any $u, v \in \mathcal{U}$,*

$$f(u) = \sum_{\alpha : |\alpha| \le k} D^\alpha f(v) \frac{(u-v)^\alpha}{\alpha!} + O\left(\|u-v\|^{k+1}\right). \tag{A.11}$$

The polynomial

$$P_k(u) = \sum_{\alpha : |\alpha| \le k} D^\alpha f(v) \frac{(u-v)^\alpha}{\alpha!}$$

is called k-th order Taylor's polynomial. In particular, for u near v,

$$P_1(u) = \sum_{\alpha : |\alpha| \le 1} D^\alpha f(v) \frac{(u-v)^\alpha}{\alpha!} = f(v) + (u-v)^\top \nabla f(a)$$