MATH5836: Data and Machine Learning

Week 0: Basics of Calculus

Sarat Moka *UNSW, Sydney*

Key Topics

Reference:

- Appendix A of *Mathematical Engineering of Deep Learning* by Liquet, Moka, and Nazarathy: `https://deeplearningmath.org/`

0.1.1 Vectors and Functions in $\mathbb{R}^n$

Notation

- $\mathbb{R}$: The set of all real numbers.
- $\mathbb{R}^n$: The n-dimensional real coordinate space, where each element is a vector represented as a column:

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

Euclidean Norm and Inner Product

- The **Euclidean norm** (or L_2 norm) of $u \in \mathbb{R}^n$ is $\|u\|_2 = \sqrt{u^\top u} = \left(\sum_{i=1}^n u_i^2 \right)^{1/2}$.
- The **inner product** of two vectors $u, v \in \mathbb{R}^n$ is the scalar $u^\top v = \sum_{i=1}^n u_i v_i$.
- The **cosine of the angle** between u and v is:

$$\cos\theta = \frac{u^\top v}{\|u\|_2 \|v\|_2}. \tag{0.1}$$

L_p Norm

- For $p \geq 1$, the L_p norm of $u \in \mathbb{R}^n$ is

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

- The default norm $\|u\|$ (without subscript) refers to the L_2 norm.

Key Inequalities

- **Cauchy-Schwarz inequality**: For any $u, v \in \mathbb{R}^n$,

$$|u^\top v| \leq \|u\| \|v\|, \tag{0.2}$$

  with equality **if and only if** u and v are linearly dependent (i.e., $u = cv$ for some $c \in \mathbb{R}$).
- **Triangle inequality**: For any $u, v \in \mathbb{R}^n$, we have $\|u + v\| \leq \|u\| + \|v\|$.

  Exercise 0.1.1.1

  By expanding $\|u+v\|^2$ and using the Cauchy-Schwarz inequality, establish the triangle inequality.

Euclidean Distance

- The distance between $u, v \in \mathbb{R}^n$ is $\|u - v\| = \left(\sum_{i=1}^{n}(u_i - v_i)^2\right)^{1/2}$.

Convergence in $\mathbb{R}^n$

- A sequence $\{u^{(k)}\}$ in $\mathbb{R}^n$ **converges** to a vector $u \in \mathbb{R}^n$ (denoted $u^{(k)} \to u$) if $\lim_{k\to\infty} \|u^{(k)} - u\| = 0$.
- That is, for every $\varepsilon > 0$, there exists a positive integer N_0 such that $\|u^{(k)} - u\| < \varepsilon$ for all $k \geq N_0$.

Exercise 0.1.1.2

Consider the sequence of vectors $\{u^{(k)}\}$ in $\mathbb{R}^2$ defined by

$$u^{(k)} = \left[\frac{1}{k},\ 2 + \frac{3}{k}\right]^\top \quad \text{for } k = 1, 2, 3, \ldots.$$

Then,

- Identify the proposed limit vector $u \in \mathbb{R}^2$ for this sequence. Justify your answer using component-wise limits.
- Compute the Euclidean distance $\|u^{(k)} - u\|$ and show that

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

- Generalize your reasoning: In $\mathbb{R}^n$, if a sequence $\{u^{(k)}\}$ satisfies $\lim_{k\to\infty} u_i^{(k)} = u_i$ for every component $i = 1, \ldots, n$, prove that $u^{(k)} \to u$ in the Euclidean norm.

Continuity of Functions

- **Scalar functions ($f : \mathbb{R}^n \to \mathbb{R}$):**
  - f is **continuous at** u if for **every** sequence $u^{(k)} \to u$,

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

  - Or, equivalently, for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|f(u) - f(v)| < \varepsilon \quad \text{whenever} \quad \|u - v\| < \delta.$$

- **Vector-valued functions ($f : \mathbb{R}^n \to \mathbb{R}^m$):**

– Expressed as

$$f(u) = [f_1(u) \ \cdots \ f_m(u)]^\top, \tag{0.3}$$

where each component $f_i : \mathbb{R}^n \to \mathbb{R}$ is a scalar function.

– f is **continuous at** u if **each** f_i is continuous at u.

– f is **continuous on a set** $\mathcal{U} \subseteq \mathbb{R}^n$ if it is continuous at every $u \in \mathcal{U}$.

0.1.2 Derivatives

Partial Derivatives and Gradient

- For $f : \mathbb{R}^n \to \mathbb{R}$, the **partial derivative** with respect to u_i is

$$\frac{\partial f(u)}{\partial u_i} = \lim_{h \to 0} \frac{f(u_1, \ldots, u_i + h, \ldots, u_n) - f(u)}{h}. \tag{0.4}$$

- The **gradient** of f at u aggregates all partial derivatives as

$$\nabla f(u) = \left[\frac{\partial f(u)}{\partial u_1}, \ldots, \frac{\partial f(u)}{\partial u_n} \right]^\top. \tag{0.5}$$

- For matrix inputs $U \in \mathbb{R}^{n \times m}$, the gradient is structured as

$$\frac{\partial f(U)}{\partial U} = \begin{bmatrix} \frac{\partial f(U)}{\partial u_{1,1}} & \cdots & \frac{\partial f(U)}{\partial u_{1,m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(U)}{\partial u_{n,1}} & \cdots & \frac{\partial f(U)}{\partial u_{n,m}} \end{bmatrix}. \tag{0.6}$$

Directional Derivatives

- The **directional derivative** of f at u in direction $v \in \mathbb{R}^n$ is

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

- Key relationships:
  – Partial derivatives are directional derivatives along coordinate axes: $\nabla_{e_i} f(u) = \frac{\partial f(u)}{\partial u_i}$.
  – For differentiable f, we have $\nabla_v f(u) = v^\top \nabla f(u)$.
  – Maximum directional derivative occurs when $v \propto \nabla f(u)$, by the Cauchy-Schwarz inequality.

Exercise 0.1.2.1

Prove the three key relationships stated above.

Jacobians

- For $f : \mathbb{R}^n \to \mathbb{R}^m$ with $f(u) = [f_1(u), \ldots, f_m(u)]^\top$, the **Jacobian** is

$$J_f(u) = \begin{bmatrix} \frac{\partial f_1(u)}{\partial u_1} & \cdots & \frac{\partial f_1(u)}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(u)}{\partial u_1} & \cdots & \frac{\partial f_m(u)}{\partial u_n} \end{bmatrix}. \tag{0.7}$$

- Transposed Jacobian notation: $\frac{\partial f(u)}{\partial u} = J_f(u)^\top$.

Hessians

- For $f : \mathbb{R}^n \to \mathbb{R}$, the **Hessian** captures second-order derivatives:

$$\nabla^2 f(u) = \begin{bmatrix} \frac{\partial^2 f}{\partial u_1^2} & \cdots & \frac{\partial^2 f}{\partial u_1 \partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial u_n \partial u_1} & \cdots & \frac{\partial^2 f}{\partial u_n^2} \end{bmatrix}. \tag{0.8}$$

- Symmetry: If second derivatives are continuous, $\frac{\partial^2 f}{\partial u_i \partial u_j} = \frac{\partial^2 f}{\partial u_j \partial u_i}$ (Schwarz's theorem).
- Hessian as Jacobian: $\nabla^2 f(u) = J_{\nabla f}(u)$.

Differentiability

- A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is **differentiable** at u if

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

  where $A = J_f(u)$ (the Jacobian).
- For scalar f, the derivative is $\nabla f(u)^\top$.
- **Continuously differentiable** functions have continuous partial derivatives.

0.1.3 The Multivariable Chain Rule

Chain Rule for Compositions

- For $f = g \circ h$, where $h : \mathbb{R}^n \to \mathbb{R}^k$ and $g : \mathbb{R}^k \to \mathbb{R}$, we have

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

- For vector-valued $g : \mathbb{R}^k \to \mathbb{R}^m$,

$$J_f(u) = J_g(h(u)) J_h(u). \tag{0.9}$$

Matrix Derivative of an Affine Transformation

- For $z = Wu + b$ and $y = g(z)$, the derivative with respect to W is

$$\frac{\partial y}{\partial W} = \frac{\partial y}{\partial z} u^\top. \tag{0.10}$$

Jacobian Vector Products (JVP) and Vector Jacobian Products (VJP)

- For composite $f = h_L \circ \cdots \circ h_1$, let

$$g_\ell(u) = h_\ell\left(h_{\ell-1}\left(\cdots(h_1(u))\cdots\right)\right).$$

  for each $\ell = 1, \ldots, L$. Then, by recursive application of (0.9), we obtain

$$J_f(u) = J_{h_L}\left(g_{L-1}(u)\right) J_{h_{L-1}}\left(g_{L-2}(u)\right) \cdots J_{h_1}(u). \tag{0.11}$$

  Note that from the definition of the Jacobian, the j-th column of $J_f(u)$ is the m dimensional vector

$$\frac{\partial f(u)}{\partial u_j} = \left(\frac{\partial f_1(u)}{\partial u_j}, \ldots, \frac{\partial f_m(u)}{\partial u_j}\right) = J_f(u) e_j,$$

  where e_j is the j-th unit vector of appropriate dimension. Therefore, the **JVP** $\partial f(u)/\partial u_j$ can be computed recursively via

$$v_\ell = J_{h_\ell}(g_{\ell-1}(u)) v_{\ell-1},$$

  starting with $v_0 = e_j$ and $g_0(u) = u$.

- On the other hand, since the i-th row of $J_f(u)$ is the gradient $\nabla f_i(u)$, we have

$$\begin{aligned} \nabla f_i(u) &= e_i^\top J_f(u) \\ &= \left[\cdots\left[\left[e_i^\top J_{h_L}\left(g_{L-1}(u)\right)\right] J_{h_{L-1}}\left(g_{L-2}(u)\right)\right]\cdots\right] J_{h_1}(u). \end{aligned} \tag{0.12}$$

  That is, for each $i = 1, \ldots, m$, the **VJP** $\nabla f_i(u)$ can be obtained by recursively via

$$v_\ell^\top = v_{\ell-1}^\top J_{h_{L-\ell+1}}(g_{L-\ell}(u)),$$

  starting with $v_0 = e_i$ and $g_0(u) = u$.

0.1.4 Taylor's Theorem

Univariate Case

- A function $f : \mathbb{R} \to \mathbb{R}$ is **k-times continuously differentiable** on an open interval $\mathcal{U}$ if all k-th order derivatives exist and are continuous on $\mathcal{U}$.

- **Taylor's Theorem in $\mathbb{R}$**: For k-times continuously differentiable univariate real-valued function f and $u, v \in \mathcal{U}$, we have

$$f(u) = \sum_{i=0}^{k} \frac{(u-v)^i}{i!} \frac{\mathrm{d}^i f(v)}{\mathrm{d}u^i} + O\left(|u-v|^{k+1}\right)$$
$$= P_k(u) + O\left(|u-v|^{k+1}\right). \quad (0.13)$$

where the **Taylor Polynomial**

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

and the Big-O notation $O(r^k)$ represents a function that, as $r \to 0$, satisfies $|O(r^k)| \leq C|r^k|$ for some constant $C > 0$, indicating that the remainder term, $f(u) - P_k(u)$, vanishes at least as fast as r^k.

- **Linear approximation ($k = 1$)**: $f(u) \approx P_1(u) = f(v) + (u - v)f'(v)$.
- **Quadratic approximation ($k = 2$)**: $f(u) \approx P_2(u) = f(v) + (u - v)f'(v) + \frac{(u-v)^2}{2} f''(v)$.

Multivariate Case

- **Multi-index Notation**: For $\alpha = (\alpha_1, \ldots, \alpha_n)$:

$$|\alpha| = \sum_{i=1}^{n} \alpha_i, \quad \alpha! = \prod_{i=1}^{n} \alpha_i!, \quad u^\alpha = \prod_{i=1}^{n} u_i^{\alpha_i}.$$

- Higher-order partial derivative:

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

- **Taylor's Theorem in $\mathbb{R}^n$**: For k-times continuously differentiable multivariate real-valued function f and $u, v \in \mathcal{U}$:

$$f(u) = \sum_{\alpha : |\alpha| \leq k} D^\alpha f(v) \frac{(u-v)^\alpha}{\alpha!} + O\left(\|u-v\|^{k+1}\right). \quad (0.14)$$

- **Key Approximations**:
  - Linear (first-order) approximation around v:

$$f(u) \approx P_1(u) = f(v) + (u-v)^\top \nabla f(v).$$

  - Quadratic (second-order) approximation around v:

$$f(u) \approx P_2(u) = f(v) + (u-v)^\top \nabla f(v) + \frac{1}{2}(u-v)^\top \nabla^2 f(v)(u-v).$$

Linear Approximation with Jacobians and Hessians

- For differentiable $f : \mathbb{R}^n \to \mathbb{R}^m$ with Jacobian J_f, an approximation of $f(u)$ around v is

$$\tilde{f}(u) = f(v) + J_f(v)(u - v). \tag{0.15}$$

- For twice differentiable $g : \mathbb{R}^n \to \mathbb{R}$ with Hessian $\nabla^2 g$, an approximation of the gradient $\nabla g(u)$ around v is

$$\widetilde{\nabla} g(u) = \nabla g(v) + \nabla^2 g(v)(u - v). \tag{0.16}$$

Exercise 0.1.4.1

1. Find the Taylor series expansion of the function $f(x) = e^{2x}$ centered at $x = 0$ (Maclaurin series) up to the quadratic term (x^2).
2. Use your result to approximate the value of $f(0.2)$.