The Austalian National University Semester 2, 2021
School of Computing Tutorial 5

COMP3670/6670: Introduction to Machine Learning

These exercises will concentrate on vector calculus, and how to compute derivatives of functions that live in higher dimensions.

Preliminaries

The formal definition of the derivative of a function $f : \mathbb{R} \to \mathbb{R}$ is given by

$$\frac{df}{dx} := \lim_{h \to 0} \frac{f(x+h) - f(h)}{h}$$

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of a vector $\mathbf{x}$. The derivative of $f(\mathbf{x})$ with respect to $\mathbf{x}$ is defined as

$$\nabla_{\mathbf{x}} \mathbf{f} = \text{grad} f = \frac{df}{d\mathbf{x}} := \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in (\mathbb{R}^n \to \mathbb{R})^{1 \times n}$$

Note that $\frac{df}{d\mathbf{x}}$ is a row vector, where each element is a function of the form $\mathbb{R}^n \to \mathbb{R}$. We write $\nabla_{\mathbf{x}} f \in (\mathbb{R}^n \to \mathbb{R})^{1 \times n}$. Some authors write $\nabla_{\mathbf{x}} f \in \mathbb{R}^{1 \times n}$ as an abuse of notation for the sake of brevity, and ease of matching dimensions. Keep in mind that each element of the row vector isn't a real number, but itself a function.

Let $\mathbf{g} : \mathbb{R} \to \mathbb{R}^n$ be a function of a scalar t. The derivative of $\mathbf{g}(t)$ with respect to t is defined as

$$\frac{d\mathbf{g}}{dt} := \begin{bmatrix} \frac{dg_1(t)}{dt} \\ \frac{dg_2(t)}{dt} \\ \vdots \\ \frac{dg_n(t)}{dt} \end{bmatrix} \in (\mathbb{R} \to \mathbb{R})^{n \times 1}$$

Note that $\frac{d\mathbf{g}}{dt}$ is a column vector, where each element is itself a function of the form $\mathbb{R} \to \mathbb{R}$. As before, we notate this using an abuse of notation as $\frac{d\mathbf{g}}{dt} \in \mathbb{R}^{n \times 1}$,

The reason why the derivatives are defined this way, is so that the dimensions match when we define the chain rule.

Given $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{g} : \mathbb{R} \to \mathbb{R}^n$, we can define two new functions

$$h : \mathbb{R} \to \mathbb{R}, \quad h(t) = f(\mathbf{g}(t))$$

$$\mathbf{k} : \mathbb{R}^n \to \mathbb{R}^n \quad \mathbf{k}(\mathbf{x}) = \mathbf{g}(f(\mathbf{x}))$$

and we can define their derivatives as

$$\frac{dh}{dt} = \frac{df}{d\mathbf{g}} \frac{d\mathbf{g}}{dt} = \begin{bmatrix} \frac{\partial f(\mathbf{g})}{\partial g_1} & \dots & \frac{\partial f(\mathbf{g})}{\partial g_n} \end{bmatrix} \begin{bmatrix} \frac{\partial g_1}{\partial t} \\ \vdots \\ \frac{\partial g_n}{\partial t} \end{bmatrix} = \sum_{i=1}^{n} \frac{\partial f(\mathbf{g})}{\partial g_i} \frac{\partial g_i}{\partial t}$$

and

$$\frac{d\mathbf{k}}{d\mathbf{x}} = \frac{d\mathbf{g}}{df} \frac{df}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial g_1}{\partial f} \\ \vdots \\ \frac{\partial g_n}{\partial f} \end{bmatrix} \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial f} \frac{\partial f(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial g_1}{\partial f} \frac{\partial f(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial f} \frac{\partial f(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial g_n}{\partial f} \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \mathbf{A}$$

where $\mathbf{A}_{ij} = \frac{\partial g_i}{\partial f} \frac{\partial f(\mathbf{x})}{\partial x_j}$.

(Here, the term $\frac{\partial f(\mathbf{g})}{\partial g_1}$ means to substitute each output component of $\mathbf{g}$ into the inputs for f, and take the partial derivative with respect to the g_i, the i^{th} component of $\mathbf{g}$.)

For a vector valued function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we define the matrix of all first order derivatives as the *Jacobian*, which is given by

$$\mathbf{J} = \nabla_{\mathbf{x}}\mathbf{f} = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}, \quad \mathbf{J}_{ij} = \frac{\partial f_i(\mathbf{x})}{\partial x_j}$$

.

You may also need the definition of matrix multiplication.

If $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{B} \in \mathbb{R}^{m \times p}$, the product $\mathbf{C} = \mathbf{AB}$ is a matrix in $\mathbb{R}^{n \times p}$ satisfying

$$C_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}$$

If $\mathbf{A} \in \mathbb{R}^{n \times m}$, $\mathbf{b} \in \mathbb{R}^{m \times 1}$ and $\mathbf{c} \in \mathbb{R}^{n \times 1}$ then the matrix vector products $\mathbf{Ab}$ and $\mathbf{c}^T\mathbf{A}$ satisfy the properties

$$(\mathbf{Ab})_k = \sum_{j=1}^{m} A_{kj} b_j$$

and

$$(\mathbf{c}^T\mathbf{A})_k = \sum_{i=1}^{n} A_{ik} c_i$$

For $\mathbf{x} \in \mathbb{R}^n$, the Euclidean norm $|| \cdot ||_2$ is given by

$$\|\mathbf{x}\|_2 := \sqrt{\mathbf{x}^T\mathbf{x}}$$

For all problems below, state the dimension of the answer where appropriate.

Question 1 **Formal definition of derivative**

Compute the derivative of $f : \mathbb{R} \to \mathbb{R}, f(x) = x^2$ from the formal limit definition of the derivative.

Question 2 **Vector Derivative of Scalar Function**

Given $f : \mathbb{R}^2 \to \mathbb{R}, f(\mathbf{x}) = 2x_1x_2 + x_1 + 3x_2 + 5$, compute $\frac{df}{d\mathbf{x}}$.

Question 3 **Scalar Derivative of Vector Function**

Given $\mathbf{g}(t) : \mathbb{R} \to \mathbb{R}^2, \mathbf{g}(t) = \begin{bmatrix} t^2 \\ e^t \end{bmatrix}$ compute $\frac{d\mathbf{g}}{dt}$.

Question 4 **Derivative of the L2 Norm**

Let $\mathbf{x} \in \mathbb{R}^n$, and define $k : \mathbb{R}^n \to \mathbb{R},\ k(\mathbf{x}) = \|\mathbf{x}\|_2^2 := \mathbf{x}^T\mathbf{x}$. Compute $\frac{dk}{d\mathbf{x}}$.

Question 5 **Chain Rule, Scalar Derivative**

Let $h : \mathbb{R} \to \mathbb{R}, h(t) = f(\mathbf{g}(t))$, where f and $\mathbf{g}$ are defined in Question 2 and Question 3 respectively.

1. Compute $\frac{dh}{dt}$ by using the chain rule.
2. Compute $\frac{dh}{dt}$ by evaluating $f(\mathbf{g}(t))$ first, and then differentiating the entire expression by t. Compare your answer to the above and check that they match.

Question 6 **Chain Rule, Vector Derivative**

Let $\mathbf{k} : \mathbb{R}^n \to \mathbb{R}^n, \mathbf{k}(\mathbf{x}) = \mathbf{g}(f(\mathbf{x}))$, where f and $\mathbf{g}$ are defined in Question 2 and Question 3 respectively.

1. Compute $\frac{d\mathbf{k}}{d\mathbf{x}}$ using the chain rule.
2. Compute $\frac{d\mathbf{k}}{d\mathbf{x}}$ directly by using the Jacobian to differentiate $\mathbf{g}(f(\mathbf{x}))$. Check your answer matches the above using chain rule.

Question 7 **More Derivatives**

1. Let $f : \mathbb{R}^n \to \mathbb{R}, f(\mathbf{x}) = (\mathbf{x}^T\mathbf{x} + 1)^2$.
   Compute $\frac{d}{d\mathbf{x}}f(\mathbf{x})$ using the chain rule. (You can use the previous questions to help you.)
2. Directly compute $\frac{d}{d\mathbf{x}}f(\mathbf{x})$ by expanding out $(\mathbf{x}^T\mathbf{x}+1)^2$ first. Your result should match the above.

Question 8 **Derivative of a Matrix-Vector product**

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{x} \in \mathbb{R}^{n\times 1}$. Show that $\frac{d}{d\mathbf{x}}(\mathbf{A}\mathbf{x}) = \mathbf{A}$.

Question 9 **Linear Regression**

Let $\mathbf{\Phi} \in \mathbb{R}^{n\times m}, \mathbf{w} \in \mathbb{R}^{n\times 1}, \mathbf{t} \in \mathbb{R}^{m\times 1}$.
Let $f : \mathbb{R}^n \to \mathbb{R}, f(\mathbf{w}) = \frac{1}{2}\left\|((\mathbf{w}^T\mathbf{\Phi})^T - \mathbf{t})\right\|_2^2$

1. Verify that f is well defined (the dimensions of all the components match up).
2. Compute $\frac{d}{d\mathbf{w}}f(\mathbf{w})$.

Question 10 **Matrix Gradient**

Given $\mathbf{X} \in \mathbb{R}^{n\times m}$ and some vectors $\mathbf{a} \in \mathbb{R}^{?\times ?}, \mathbf{b} \in \mathbb{R}^{?\times ?}$.

1. What are the dimensions of $\mathbf{a}$ and $\mathbf{b}$ such that $\mathbf{a}^T\mathbf{X}\mathbf{b}$ is well defined?[1] What is the dimension of the result?
2. Compute the matrix gradient $\frac{d}{d\mathbf{X}}\mathbf{a}^T\mathbf{X}\mathbf{b}$.

[1] Note that if $\mathbf{X}$ is square, symmetric and positive definite, then defining $\langle \mathbf{a}, \mathbf{b}\rangle := \mathbf{a}^T\mathbf{X}\mathbf{b}$ gives an inner product.