CIS 419/519

Primer - Single Variable, Vector and Matrix Calculus

In this primer, we start off by giving a brief recap on key concepts of single variable calculus, following by an extension of these concepts to vectors and matrices. In many areas of machine learning, the problems and analysis often involve partial derivatives with respect to vectors and matrices. Some of these areas include optimization, probability, regression and classification. Hence, it is important to understand single-variable calculus, as well as vector and matrix calculus in order to appreciate the mathematics behind the applications in machine learning. Here we provide a list of useful concepts, derivatives and differentiation rules often encountered in machine learning, taken from the references stated at the end of the notebook. For a more comprehensive list, refer to [1] and [2]. Note that the *denominator layout* convention is adopted in this document.

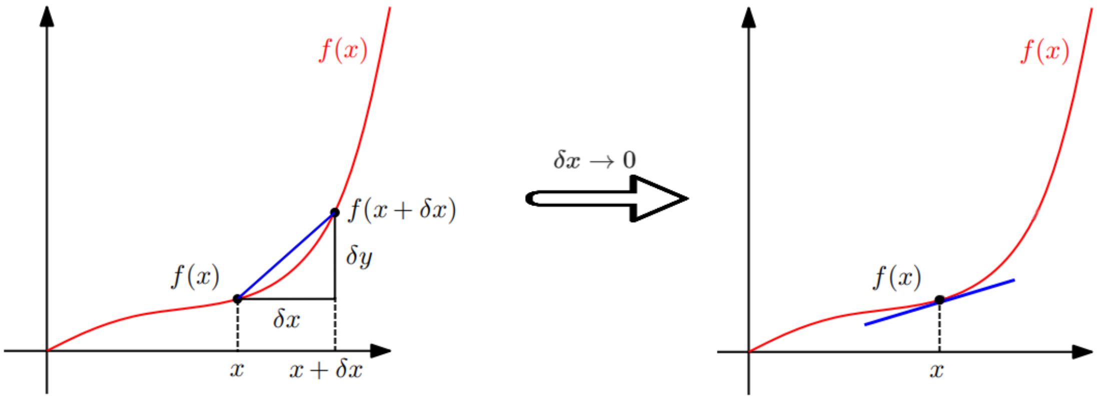
Key concepts in single variable calculus

Derivative:

Given $h > 0$, the derivative of f with respect to x is defined as

$$\frac{df}{dx} := \lim_{\delta x \to 0} \frac{f(x + \delta x) - f(x)}{\delta x}. \tag{1}$$

This can be interpreted as the limit of the difference quotient or slope of the function. In other words, we first evaluate f at two points, x and $x + \delta x$ and compute the value of the difference quotient, i.e. $\frac{f(x+\delta x)-f(x)}{(x+\delta x)-x} = \frac{f(x+\delta x)-f(x)}{\delta x}$, which is the slope of the secant line (blue) shown in the left figure below. By shifting the two points closer to each other, i.e., $\delta x \to 0$, this secant line becomes a tangent line, shown in the figure on the right. The slope of this tangent line (blue line in the right plot) is the derivative of f evaluated at x, $\frac{df}{dx}$.



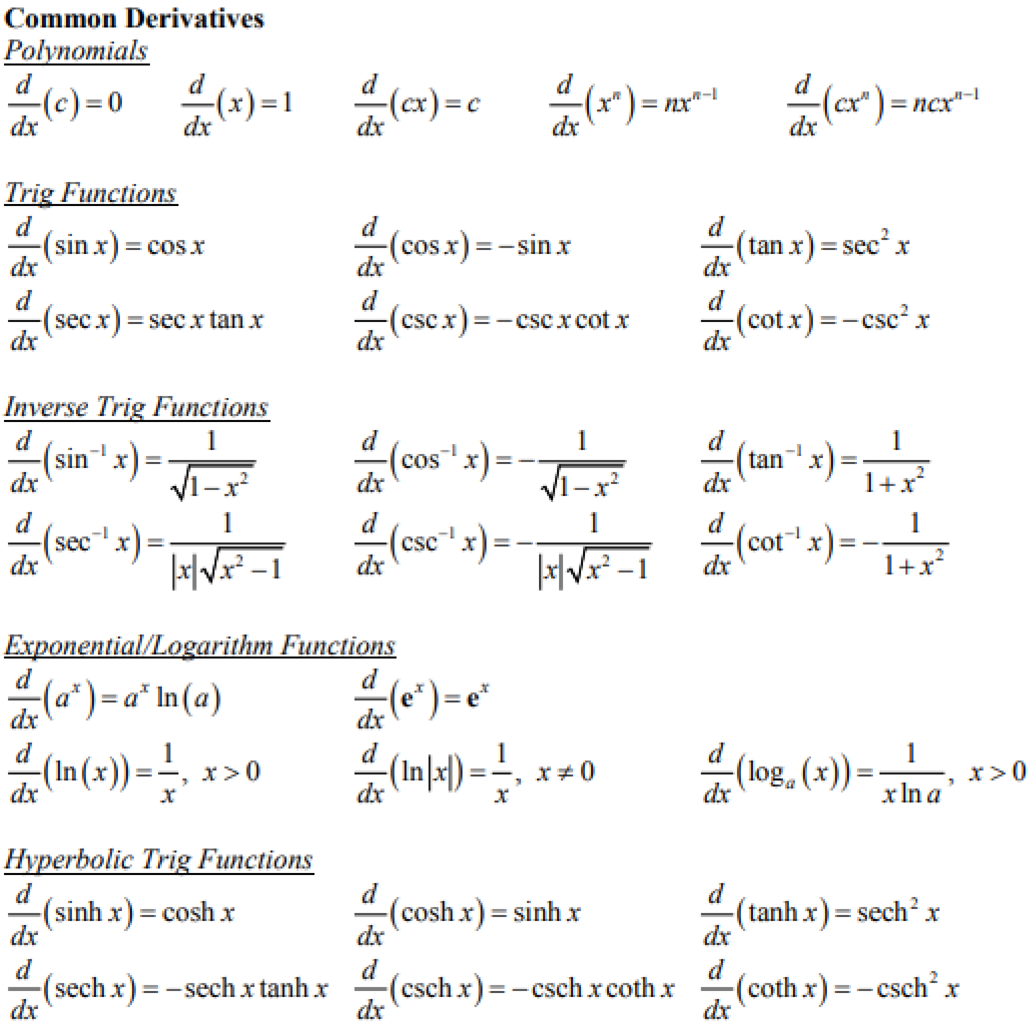
Maxima and Minima:

One of the most important applications of calculus is its ability to sniff out the maximum or the minimum of a function. In optimization problems we are looking for the largest value or the smallest value that a function can take. By finding the minima and maxima of a function we determine where is the function at a high or a low point. In a smoothly changing function a maximum or minimum is always where the function flattens out i.e when the slope of the function or the derivative of the function is 0.

Take $f(\mathbf{x})$ to be a function of x. Then the value of x for which the derivative of $f(\mathbf{x})$ with respect to x is equal to zero corresponds to a maximum or a minimum point of the function $f(\mathbf{x})$. To deterrmine whether the point is a maximum or a minimum, we perform the second derivate test by taking the derivative of the slope $\frac{df}{dx}$ or the second derivative of the original function $f(\mathbf{x})$. When a function's slope is zero at x, and the second derivative at x is less than 0, it is a local maximum and greater than 0, it is a local minimum.

Common Derivatives:

A list of common derivatives are given in this figure [3]:



Common Derivatives

Polynomials

$\frac{d}{dx}(c)=0$ $\quad$ $\frac{d}{dx}(x)=1$ $\quad$ $\frac{d}{dx}(cx)=c$ $\quad$ $\frac{d}{dx}(x^n)=nx^{n-1}$ $\quad$ $\frac{d}{dx}(cx^n)=ncx^{n-1}$

Trig Functions

$\frac{d}{dx}(\sin x)=\cos x$ $\quad$ $\frac{d}{dx}(\cos x)=-\sin x$ $\quad$ $\frac{d}{dx}(\tan x)=\sec^2 x$

$\frac{d}{dx}(\sec x)=\sec x\tan x$ $\quad$ $\frac{d}{dx}(\csc x)=-\csc x\cot x$ $\quad$ $\frac{d}{dx}(\cot x)=-\csc^2 x$

Inverse Trig Functions

$\frac{d}{dx}(\sin^{-1} x)=\frac{1}{\sqrt{1-x^2}}$ $\quad$ $\frac{d}{dx}(\cos^{-1} x)=-\frac{1}{\sqrt{1-x^2}}$ $\quad$ $\frac{d}{dx}(\tan^{-1} x)=\frac{1}{1+x^2}$

$\frac{d}{dx}(\sec^{-1} x)=\frac{1}{|x|\sqrt{x^2-1}}$ $\quad$ $\frac{d}{dx}(\csc^{-1} x)=-\frac{1}{|x|\sqrt{x^2-1}}$ $\quad$ $\frac{d}{dx}(\cot^{-1} x)=-\frac{1}{1+x^2}$

Exponential/Logarithm Functions

$\frac{d}{dx}(a^x)=a^x\ln(a)$ $\quad$ $\frac{d}{dx}(\mathbf{e}^x)=\mathbf{e}^x$

$\frac{d}{dx}(\ln(x))=\frac{1}{x},\ x>0$ $\quad$ $\frac{d}{dx}(\ln|x|)=\frac{1}{x},\ x\neq 0$ $\quad$ $\frac{d}{dx}(\log_a(x))=\frac{1}{x\ln a},\ x>0$

Hyperbolic Trig Functions

$\frac{d}{dx}(\sinh x)=\cosh x$ $\quad$ $\frac{d}{dx}(\cosh x)=\sinh x$ $\quad$ $\frac{d}{dx}(\tanh x)=\operatorname{sech}^2 x$

$\frac{d}{dx}(\operatorname{sech} x)=-\operatorname{sech} x\tanh x$ $\quad$ $\frac{d}{dx}(\operatorname{csch} x)=-\operatorname{csch} x\coth x$ $\quad$ $\frac{d}{dx}(\coth x)=-\operatorname{csch}^2 x$

Taylor Series (Optional on first reading):

For a smooth function f that is infinitely differentiable, we define the Taylor series of f at a point x_0 as

$$T_\infty(x) := \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k. \tag{2}$$

Taylor series have two primary applications. In theoretical applications, when we want to understand a function that is too complex, an application of the Taylor series allows us to approximate it into a polynomial. In numerical applications, functions like e^x or $\cos(x)$ are difficult for machines to compute. They can store tables of values at a fixed precision, but it still leaves open questions like "What is the 1000^{th} digit of $\cos(1)$?" Taylor series are often helpful to answer such questions.

Differentiation rules:

Denoting the derivative of $f(x)$ and $g(x)$ as $f(x)'$ and $g(x)'$ and $g \circ f := g(f(x))$, c is any real number, we have

$$(cf)' = cf' \tag{3}$$

$$(f \pm g)' = f' \pm g' \tag{4}$$

$$(fg)' = f'g + fg' \qquad \text{(Product Rule)} \tag{5}$$

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2} \qquad \text{(Quotient Rule)} \tag{6}$$

$$(g \circ f)' = g'(f)f' \qquad \text{(Chain Rule)} \tag{7}$$

Higher order derivatives (Optional on first reading):

The second derivative of f is defined as the derivative of the first derivative,

$$\frac{d^2 f}{dx^2} = f''(x) = f^{(2)}(x) := \frac{d}{dx}\big(f'(x)\big) \tag{8}$$

and in general, the n^{th} derivative of f is defined as the derivative of the $(n-1)^{th}$ derivative,

$$\frac{d^n f}{dx^n} = f^{(n)}(x) := \frac{d}{dx}\left(f^{(n-1)}(x)\right) \tag{9}$$

Multivariable Calculus

In this section, common derivatives and differentiation rules with respect to vectors and matrices are listed. Observe that most of the basic rules such as product, quotient, chain and sum rules still apply.

Common derivatives

Differentiation of a scalar f

Partial Derivatives:

When a function has multiple variables, it helps to try to isolate how each variable changes when others are held constant. This is called a partial derivative, and can be expressed as:

$$\begin{aligned}\frac{\partial f}{\partial x_1} &= \lim_{\delta x \to 0} \frac{f(x_1 + \delta x, x_2, \ldots, x_n) - f(\mathbf{x})}{\delta x}, \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{\delta x \to 0} \frac{f(x_1, x_2, \ldots, x_n + \delta x) - f(\mathbf{x})}{\delta x}\end{aligned}$$

For example, if we have a scalar function $f(x, y) = 2xy^2$, then:

$$\frac{\partial f}{\partial \mathbf{x}} = 2y^2 \tag{10}$$

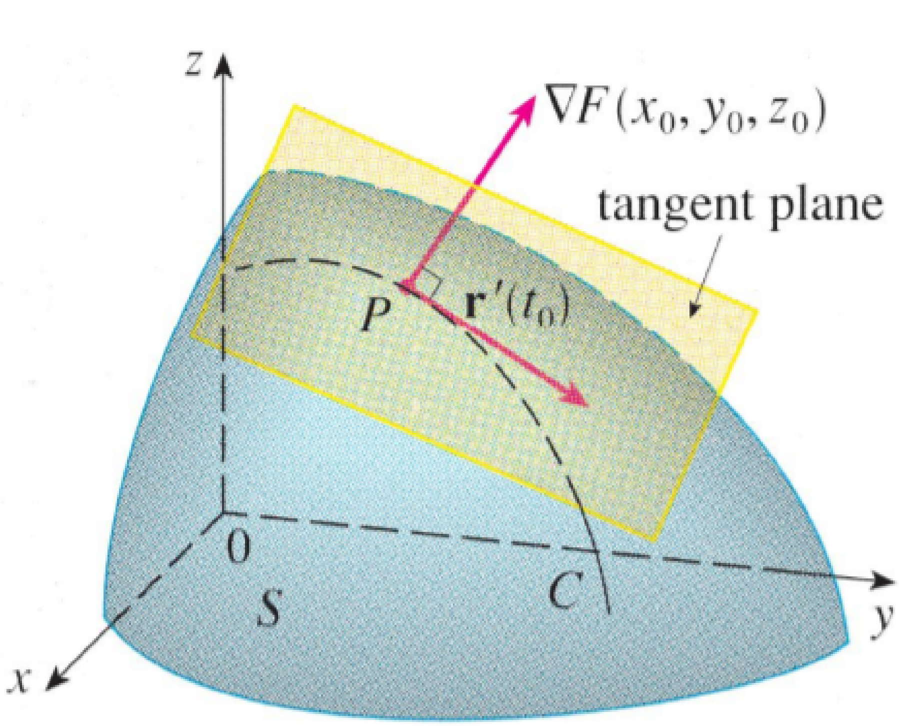
$$\frac{\partial f}{\partial \mathbf{y}} = 4xy \tag{11}$$

Gradient: For a scalar function $f(\mathbf{x})$, where $\mathbf{x}$ is a vector with n variables, $\mathbf{x} := [x_1, \ldots, x_n]^T$, we can define the *gradient* of f with respect to $\mathbf{x}$ as

$$\frac{\partial f}{\partial \mathbf{x}} = \nabla_{\mathbf{x}} f := \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \tag{12}$$

In other words, the gradient is a vector consisting of the partial derivatives of f with respect to its arguments, $x_1, \ldots, x_n$.

Geometrically, the gradient evaluated at any given point $\mathbf{x}$ is a vector that is normal to the tangent plane at $\mathbf{x}$, as shown in the figure below, where an example in 3 dimensions with $\mathbf{x} = [x_0, y_0, z_0]^T$ is plotted. In this figure, $\nabla F(x_0, y_0, z_0)$ denotes the gradient.



Source: [4]

The gradient of a function points in the direction of that functions greatest ascent, or the direction it increases most (of course, this means the opposite of that gradient points in the direction of steepest descent). Thus, by evaluating gradients at different values of our function, we can more efficiently move toward a local maximum or minimum of that function. This can be incredibly useful in machine learning, where we are often trying to find the minimum of a function as part of optimization.

For more geometric interpretations, here are some video resources that explains the gradient in different ways:

- https://www.youtube.com/watch?v=QQPz3eXXgQI
- https://www.youtube.com/watch?v=AXH9Xm6Rbfc

Hessian (Optional on first reading): And differentiating this scalar function f again (twice) with respect to a vector $\mathbf{x}$ with n variables, we get the *Hessian*,

$$\frac{\partial f}{\partial \mathbf{x}} := \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{13}$$

We can also differentiate f with respect to a matrix $\mathbf{X}$ of dimensions $n \times m$ to get

$$\frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{n1}} & \cdots & \frac{\partial f}{\partial x_{nm}} \end{bmatrix} \tag{14}$$

Differentiation of a vector $\mathbf{f}$

Consider a vector $\mathbf{f}$ with n variables, $\mathbf{f} := [f_1, \ldots, f_n]^T$. Differentiating this with respect to a scalar x gives

$$\frac{\partial \mathbf{f}}{\partial x} = \left[\frac{\partial f_1}{\partial x}, \ldots, \frac{\partial f_n}{\partial x}\right] \tag{15}$$

Jacobian: If $\mathbf{f}$ is a vector with n variables, $\mathbf{f} := [f_1, \ldots, f_n]^T$, differentiating it with respect to another vector $\mathbf{x}$ gives the *Jacobian*,

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} := \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} \tag{16}$$

The vectors of the Jacobian are the gradients of the respective components of a function.

To get the derivative of a vector $\mathbf{f} := [f_1, \ldots, f_n]$ with respect to a matrix $\mathbf{X}$ of dimensions $n \times m$, it is essentially a concatenation of the derivatives of the scalar elements of $\mathbf{f}$ with respect to $\mathbf{X}$,

$$\frac{\partial \mathbf{f}}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{X}} \\ \vdots \\ \frac{\partial f_n}{\partial \mathbf{X}} \end{bmatrix}, \tag{17}$$

where

$$\frac{\partial f_i}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f_i}{\partial x_{11}} & \cdots & \frac{\partial f_i}{\partial x_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_i}{\partial x_{n1}} & \cdots & \frac{\partial f_i}{\partial x_{nm}} \end{bmatrix}, \quad i = 1, \ldots, n. \tag{18}$$

Useful Identities

In this section, we provide some useful identities that may be encountered in mathematics for machine learning.

Derivatives with respect to vectors

Given a vector $\mathbf{f}$ that is not a function of $\mathbf{x}$ and a matrix $\mathbf{A}$ which is also not a function of $\mathbf{x}$, we have

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \mathbf{0} \quad \text{and} \quad \frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I}, \tag{19}$$

and

$$\frac{\partial \mathbf{f}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{f} \quad \text{and} \quad \frac{\partial \mathbf{f}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T \mathbf{f}, \tag{20}$$

and

$$\frac{\partial \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T \quad \text{and} \quad \frac{\partial \mathbf{x}^T \mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}, \tag{21}$$

and

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}, \qquad \frac{\partial \mathbf{x}^T \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}. \tag{22}$$

Given scalars $f(\mathbf{x})$ and $g(\mathbf{x})$ which are functions of $\mathbf{x}$ and a scalar a that is independent of $\mathbf{x}$,

$$\frac{\partial a f}{\partial \mathbf{x}} = a \frac{\partial f}{\partial \mathbf{x}}, \qquad \frac{\partial (f+g)}{\partial \mathbf{x}} = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial g}{\partial \mathbf{x}}. \tag{23}$$

Given vectors $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ which are functions of $\mathbf{x}$, and a matrix $\mathbf{A}$ that is not a function of $\mathbf{x}$,

$$\frac{\partial \mathbf{f}^T \mathbf{g}}{\partial \mathbf{x}} = \frac{\partial f}{\partial \mathbf{x}} \mathbf{g} + \frac{\partial g}{\partial \mathbf{x}} \mathbf{f}. \tag{24}$$

$$\frac{\partial \mathbf{f}^T \mathbf{A} \mathbf{g}}{\partial \mathbf{x}} = \frac{\partial f}{\partial \mathbf{x}} \mathbf{A} \mathbf{g} + \frac{\partial g}{\partial \mathbf{x}} \mathbf{A}^T \mathbf{f}. \tag{25}$$

Derivatives with respect to matrices

Given vectors $\mathbf{f}$ and $\mathbf{g}$ which are not functions of $\mathbf{X}$, and a matrix $\mathbf{A}$ that is not a function of $\mathbf{X}$,

$$\frac{\partial \mathbf{f}^T \mathbf{X} \mathbf{g}}{\partial \mathbf{X}} = \mathbf{f} \mathbf{g}^T, \qquad \frac{\partial \mathbf{f}^T \mathbf{X}^T \mathbf{g}}{\partial \mathbf{X}} = \mathbf{g} \mathbf{f}^T, \tag{26}$$

and

$$\frac{\partial (\mathbf{X}\mathbf{f} + \mathbf{g})^T \mathbf{A} (\mathbf{X}\mathbf{f} + \mathbf{g})}{\partial \mathbf{X}} = (\mathbf{A} + \mathbf{A}^T)(\mathbf{X}\mathbf{f} + \mathbf{g})\mathbf{f}^T, \qquad \frac{\partial (\mathbf{X}\mathbf{f})^T \mathbf{A} (\mathbf{X}\mathbf{g})}{\partial \mathbf{X}} = \mathbf{A}\mathbf{X}\mathbf{g}\mathbf{f}^T +$$

Denoting the trace of a matrix $\mathbf{X}$ as $\operatorname{tr}(\mathbf{X})$, we also have the following useful expressions,

$$\frac{\partial \operatorname{tr}(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{I}, \qquad \frac{\operatorname{tr}(\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}^T, \qquad \frac{\operatorname{tr}(\mathbf{A}\mathbf{X}^T)}{\partial \mathbf{X}} = \mathbf{A}, \qquad \frac{\operatorname{tr}(\mathbf{X}^T \mathbf{A} \mathbf{X})}{\partial \mathbf{X}} = \Big(\mathbf{A} + \mathbf{A}^T$$

While some of the above rules are lengthy, they can be simplified if the matrices are symmetric, i.e., $\mathbf{A} = \mathbf{A}^T$.

An Example of the Chain rule [2]

In many machine learning applications, we find good model parameters by performing gradient descent, which relies on the fact that we can compute the gradient of a loss function with respect to the parameters of the model. For a given loss function, we can obtain the gradient with respect to the model parameters using calculus and applying the chain rule. As an example, consider the function

$$f(x) = \sqrt{x^2 + \exp(x^2)} + \cos(x^2 + \exp(x^2)). \tag{29}$$

Applying the chain rule along with other differentiation rules, the gradient can be computed as

$$\begin{aligned} \frac{df}{dx} &= \frac{2x + 2x\exp(x^2)}{2\sqrt{x^2 + \exp(x^2)}} - \sin(x^2 + \exp(x^2))(2x + 2x\exp(x^2)) \\ &= 2x\left(\frac{1}{2\sqrt{x^2 + \exp(x^2)}} - \sin(x^2 + \exp(x^2))\right)(1 + \exp(x^2)). \end{aligned}$$

Writing out the gradient in this explicit way is often impractical since it often results in a very lengthy expression for a derivative. In the context of deep neural network models, we will study an approach to handle this, called backpropagation, which makes it much more efficient to compute the gradient of a loss function with respect to the model parameters.

More resources:

For those who would like more online resources on calculus, below are some of them,

- Paul Dawkins's online notes on Calculus I (subsequent Calculus modules are also linked from this page): https://tutorial.math.lamar.edu/Classes/CalcI/CalcI.aspx
- Brandon Leonard's video lectures on Calculus I (subsequent Calculus modules are in his youtube page): https://youtube.com/playlist?list=PLF797E961509B4EB5
- Links to several pages for calculus resources: http://calculus.org/

References

[1] K. Petersen and M. Pedersen, “The matrix cookbook, version 20121115”, Technical Univ. Denmark, Kongens Lyngby, Denmark, Tech. Rep, vol. 3274, 2012.

[2] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, Mathematics for machine learning. Cambridge University Press, 2020.

[3] P. Dawkins, Calculus I, Chapter 3: Derivatives, https://tutorial.math.lamar.edu/pdf/calculus_cheat_sheet_derivatives.pdf, 2005.

[4] https://bvmtc.math.tamu.edu/~glahodny/Math251/Section%2012.6.pdf