19
Calculus

19.1 Calculus in One Variable

In this section, we briefly review calculus in one variable.

19.1.1 Exponential and Logarithmic Functions

When we multiply the same number a by n times, we denote it as a^n. The *exponential function* is a natural extension of this concept.

Definition 19.1.1 (Exponential Function)**.** *There is a unique function $f : \mathbb{R} \to \mathbb{R}$ such that $f(n) = e^n$ for any $n \in \mathbb{N}$ and $f(x+y) = f(x)f(y)$ for any $x, y \in \mathbb{R}$. This function is called the **exponential function** and is denoted as e^x or $\exp(x)$.*

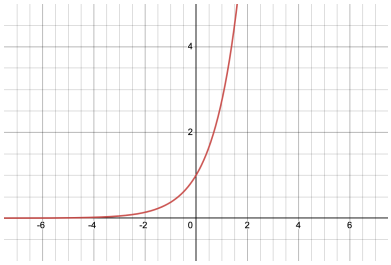

Figure 19.1: The graph of the exponential function.

Proposition 19.1.2. *The following properties hold for the exponential function:*

1. $\exp(x) > 0$ *for any* $x \in \mathbb{R}$
2. $\exp(x)$ *is increasing*
3. $\lim_{x \to -\infty} \exp(x) = 0$
4. $\lim_{x \to \infty} \exp(x) = \infty$
5. $\exp(-x) = \frac{1}{\exp(x)}$

We are also interested in the inverse function of the exponential function.

Definition 19.1.3 (Logarithmic Function)**.** *The **logarithmic function** $\log : (0, \infty) \to \mathbb{R}$ is defined as the inverse function of the exponential function. That is, $\log(x) = y$ where $x = e^y$.*

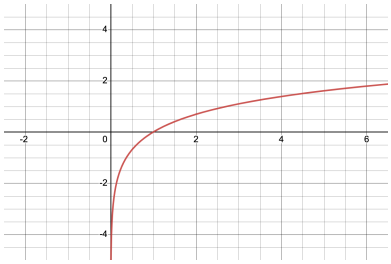

Figure 19.2: The graph of the logarithmic function.

Proposition 19.1.4. *The following properties hold for the logarithmic function:*

1. $\log(x)$ *is increasing*
2. $\lim_{x \to 0+} \log(x) = -\infty$
3. $\lim_{x \to \infty} \log(x) = \infty$
4. $\log(xy) = \log(x) + \log(y)$

19.1.2 Sigmoid Function

In Machine Learning, a slight variant of the exponential function, known as the *sigmoid function* is widely used.

Definition 19.1.5 (Sigmoid Function)**.** *The* ***sigmoid function*** *denoted as* $\sigma : \mathbb{R} \to \mathbb{R}$ *is defined as*

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

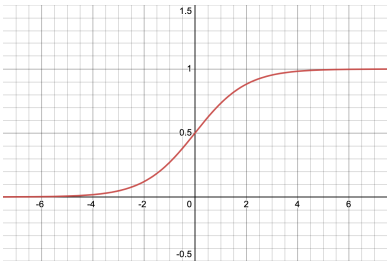

Figure 19.3: The graph of the sigmoid function.

Proposition 19.1.6. *The following properties hold for the sigmoid function:*

1. $0 < \sigma(x) < 1$ *for any* $x \in \mathbb{R}$
2. $\sigma(x)$ *is increasing*
3. $\lim_{x \to -\infty} \sigma(x) = 0$
4. $\lim_{x \to \infty} \sigma(x) = 1$
5. *The graph of* σ *is symmetrical to the point* $\left(0, \frac{1}{2}\right)$*. In particular,*

$$\sigma(x) + \sigma(-x) = 1$$

Because of the last property in Proposition 19.1.6, the sigmoid function is well suited for binary classification (*e.g.*, in logistic regression in Chapter 1). Given some output value x of a classification model, we interpret it as the measure of confidence that the input is of label 1, where we implicitly assume that the measure of confidence that the input is of label 2 is $-x$. Then we apply the sigmoid function to translate this into a probability distribution over the two labels.

19.1.3 Differentiation

Definition 19.1.7 (Derivative)**.** *Given a function* $f : \mathbb{R} \to \mathbb{R}$*, its* ***derivative*** f' *is defined as*

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

We alternatively denote $f'(x)$ *as* $\frac{d}{dx} f(x)$.

Example 19.1.8. *The derivative of the exponential function is itself:*

$$\exp'(x) = \exp(x)$$

and the derivative of the logarithmic function is:

$$\log'(x) = \frac{1}{x}$$

In general, there are more than two variables, that are related to each other through a composite function. The *chain rule* helps us find the derivative of the composite function.

Definition 19.1.9 (Chain Rule)**.** *If there are functions $f, g : \mathbb{R} \to \mathbb{R}$ such that $y = f(x)$ and $z = g(y)$, then*

$$(g \circ f)'(x) = g'(f(x))f'(x) = \frac{d}{dy}g(f(x)) \cdot \frac{d}{dx}f(x)$$

or equivalently

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

19.2 Multivariable Calculus

In this section, we introduce the basics of multivariable calculus, which is widely used in Machine Learning. Since this is a generalization of the calculus in one variable, it will be useful to pay close attention to the similarity with the results from the previous section.

19.2.1 Mappings of Several Variables

So far, we only considered functions of the form $f : \mathbb{R} \to \mathbb{R}$ that map a real value x to a real value y. But now we are interested in mappings $f : \mathbb{R}^n \to \mathbb{R}^m$ that map a vector $\vec{\mathbf{x}} = (x_1, \ldots, x_n)$ with n coordinates to a vector $\vec{\mathbf{y}} = (y_1, \ldots, y_m)$ with m coordinates. In general, a *function* is a special case of a *mapping* where the range is $\mathbb{R}$. If the mappings are of the form $f : \mathbb{R}^n \to \mathbb{R}$ (*i.e.*, $m = 1$), it can still be called a *function* of several variables.

First consider an example where $m = 1$.

Example 19.2.1. *Let $f(x_1, x_2) = x_1^2 + x_2^2$ be a function in two variables. This can be understood as mapping a point $\vec{\mathbf{x}} = (x_1, x_2)$ in the Cartesian coordinate system to its squared distance from the origin. For example, $f(3, 4) = 25$ shows that the point the squared distance between $(3, 4)$ and the origin $(0, 0)$ is 25.*

When $m > 1$, we notice that each coordinate $y_1, \ldots, y_m$ is a function of $x_1, \ldots, x_n$. Therefore, we can decompose f into m functions $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(\vec{\mathbf{x}}) = (f_1(\vec{\mathbf{x}}), \ldots, f_m(\vec{\mathbf{x}}))$$

Example 19.2.2. *Let $f(x_1, x_2) = (x_1^2 x_2, x_1 x_2^2)$ be a mapping from $\mathbb{R}^2$ to $\mathbb{R}^2$. Then we can decompose f into two functions f_1, f_2 in two variables where*

$$f_1(x_1, x_2) = x_1^2 x_2$$
$$f_2(x_1, x_2) = x_1 x_2^2$$

19.2.2 Softmax Function

The *softmax function* is a multivariable function widely used in Machine Learning, especially for multi-class classification (see Chapter 4, Chapter 10). It takes in a vector of k values, each corresponding to a particular class, and outputs a probability distribution over the k classes — that is, a vector of k non-negative values that sum up to 1. The resulting probability is *exponentially proportional* to the input value of that class. We formally write this as:

Definition 19.2.3 (Softmax Function)**.** *Given a vector $\vec{\mathbf{z}} = (z_1, z_2, \ldots, z_k) \in \mathbb{R}^k$, we define the **softmax function** as a probability function $softmax : \mathbb{R}^k \to [0,1]^k$ where the "probability of predicting class i" is:*

$$softmax(\vec{\mathbf{z}})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{19.1}$$

Problem 19.2.4. *Show that for $k = 2$, the definition of the softmax function is equivalent to the sigmoid function (after slight rearrangement/renaming of terms).*

The sigmoid function is used for binary classification, where it takes in a single real value and converts it to a probability of one class (and the probability of the other class can be inferred as its complement). The softmax function is used for multi-class classification, where it takes in k real values and converts them to k probabilities, one for each class.

19.2.3 Differentiation

Just like with functions in one variable, we can define differentiation for mappings in several variables. The key point is that now we will define a *partial derivative* for each pair (x_i, y_j) of coordinate x_i of the domain and coordinate y_j of the range.

Definition 19.2.5 (Partial Derivative)**.** *Given a function $f : \mathbb{R}^n \to \mathbb{R}^m$, the **partial derivative of y_j with respect to x_i at the point $\vec{\mathbf{x}}$** is defined as*

$$\left.\frac{\partial y_j}{\partial x_i}\right|_{\vec{\mathbf{x}}} = \lim_{h \to 0} \frac{f_j(x_1, \ldots, x_j + h, \ldots, x_n) - f_j(x_1, \ldots, x_j, \ldots, x_n)}{h}$$

Definition 19.2.6 (Gradient). *If $f : \mathbb{R}^n \to \mathbb{R}$ is a function of several variables, the* gradient of f *is defined as a mapping $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ that maps each input vector to the vector of partial derivatives at that point:*

$$\nabla f(\vec{\mathbf{x}}) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right) \Bigg|_{\vec{\mathbf{x}}}$$

Similarly to the chain rule in one variable, we can define a chain rule for multivariable settings. The key point is that there are multiple ways that a coordinate x_j can affect the value of z_i. Definition 19.2.7 can be thought as applying the chain rule for one variable in each of the paths, and adding up the results.

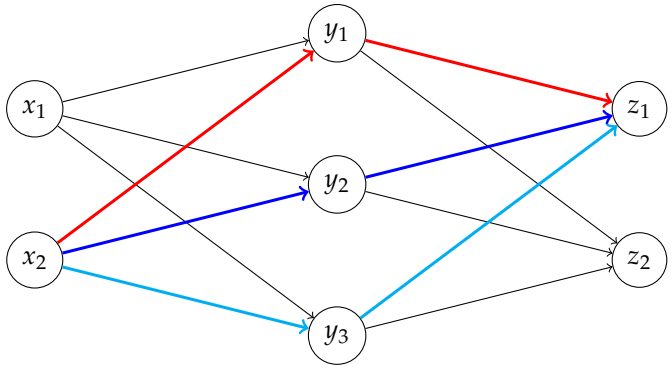

Figure 19.4: A visualization of the chain rule in multivariable settings. Notice that x_2 can affect the value of z_1 in three different paths. The amount of effect from each path will respectively be calculated as $(\partial z_1/\partial y_1)(\partial y_1/\partial x_2)$ (red), $(\partial z_1/\partial y_2)(\partial y_2/\partial x_2)$ (blue), and $(\partial z_1/\partial y_3)(\partial y_3/\partial x_2)$ (cyan).

Definition 19.2.7 (Chain Rule). *If $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^\ell$ are mappings of several variables, where $\vec{\mathbf{y}} = f(\vec{\mathbf{x}})$ and $\vec{\mathbf{z}} = g(\vec{\mathbf{y}})$, the following* ***chain rule*** *holds for each $1 \le i \le \ell$ and $1 \le j \le n$:*

$$\frac{\partial z_i}{\partial x_j} = \sum_{k=1}^{m} \frac{\partial z_i}{\partial y_k} \cdot \frac{\partial y_k}{\partial x_j}$$

Example 19.2.8. *Suppose we define the functions $h = s + t^2$, $s = 3x$, and $t = x - 2$. Then, we can find the partial derivative $\frac{\partial h}{\partial x}$ using the chain rule:*

$$\begin{aligned} \frac{\partial h}{\partial x} &= \frac{\partial s}{\partial x} + \frac{\partial (t^2)}{\partial x} \\ &= \frac{\partial s}{\partial x} + \frac{\partial (t^2)}{\partial t} \cdot \frac{\partial t}{\partial x} \\ &= 3 + 2t \cdot 1 \\ &= 2x - 1 \end{aligned}$$

Problem 19.2.9. *Suppose we define the functions $h = s + t^2$, $s = xy$, and $t = x - 2y$. Compute the partial derivative $\partial h / \partial x$.*