SEBASTIAN RASCHKA

# Introduction to Artificial Neural Networks and Deep Learning

with Applications in Python

# Introduction to Artificial Neural Networks

**with Applications in Python**

**Sebastian Raschka**

DRAFT

Last updated: February 12, 2019

This book will be available at http://leanpub.com/ann-and-deeplearning.

Please visit https://github.com/rasbt/deep-learning-book for more information, supporting material, and code examples.

# Contents

# Website

Please visit the GitHub repository to download the code examples accompanying this book and other supplementary material.

If you like the content, please consider supporting the work by buying a copy of the book on Leanpub. Also, I would appreciate hearing your opinion and feedback about the book, and if you have any questions about the contents, please don't hesitate to get in touch with me via mail@sebastianraschka.com. Happy learning!

*Sebastian Raschka*

# About the Author

Sebastian Raschka received his doctorate from Michigan State University developing novel computational methods in the field of computational biology. In summer 2018, he joined the University of Wisconsin–Madison as Assistant Professor of Statistics. Among others, his research activities include the development of new deep learning architectures to solve problems in the field of biometrics. Among his other works is his book "Python Machine Learning," a bestselling title at Packt and on Amazon.com, which received the ACM Best of Computing award in 2016 and was translated into many different languages, including German, Korean, Italian, traditional Chinese, simplified Chinese, Russian, Polish, and Japanese.

Sebastian is also an avid open-source contributor and likes to contribute to the scientific Python ecosystem in his free-time. If you like to find more about what Sebastian is currently up to or like to get in touch, you can find his personal website at https://sebastianraschka.com.

# Acknowledgements

I would like to give my special thanks to the readers, who provided feedback, caught various typos and errors, and offered suggestions for clarifying my writing.

- Appendix A: Artem Sobolev, Ryan Sun
- Appendix B: Brett Miller, Ryan Sun
- Appendix D: Marcel Blattner, Ignacio Campabadal, Ryan Sun, Denis Parra Santander
- Appendix F: Guillermo Monecchi, Ged Ridgway, Ryan Sun, Patric Hindenberger
- Appendix H: Brett Miller, Ryan Sun, Nicolas Palopoli, Kevin Zakka

# Appendix D

# Calculus and Differentiation Primer

Calculus is a discipline of mathematics that provides us with *tools* to analyze rates of change, or decay, or motion. Both Isaac Newton and Gottfried Leibniz developed the foundations of calculus independently in the 17th century. Although we recognize Gottfried and Leibniz as the founding fathers of calculus, this field, however, has a very long series of contributors, which dates back to the ancient period and includes Archimedes, Galileo, Plato, Pythagoras, just to name a few [Boyer, 1970].

In this appendix we will only concentrate on the subfield of calculus that is of most relevance to machine and deep learning: differential calculus. In simple terms, differential calculus is focused on instantaneous rates of change or computing the slope of a *linear* function. We will review the basic concepts of computing the derivatives of functions that take on one or more parameters. Also, we will refresh the concepts of the chain rule, a rule that we use to compute the derivatives of composite functions, which we so often deal with in machine learning.

## D.1 Intuition

So, what *is* the derivative of a function? In simple terms, the derivative a function is a function's instantaneous *rate of change*. Now, let us start this section with a visual explanation, where we consider the function

$$f(x) = 2x \tag{D.1}$$

shown in the graph in Figure D.1.

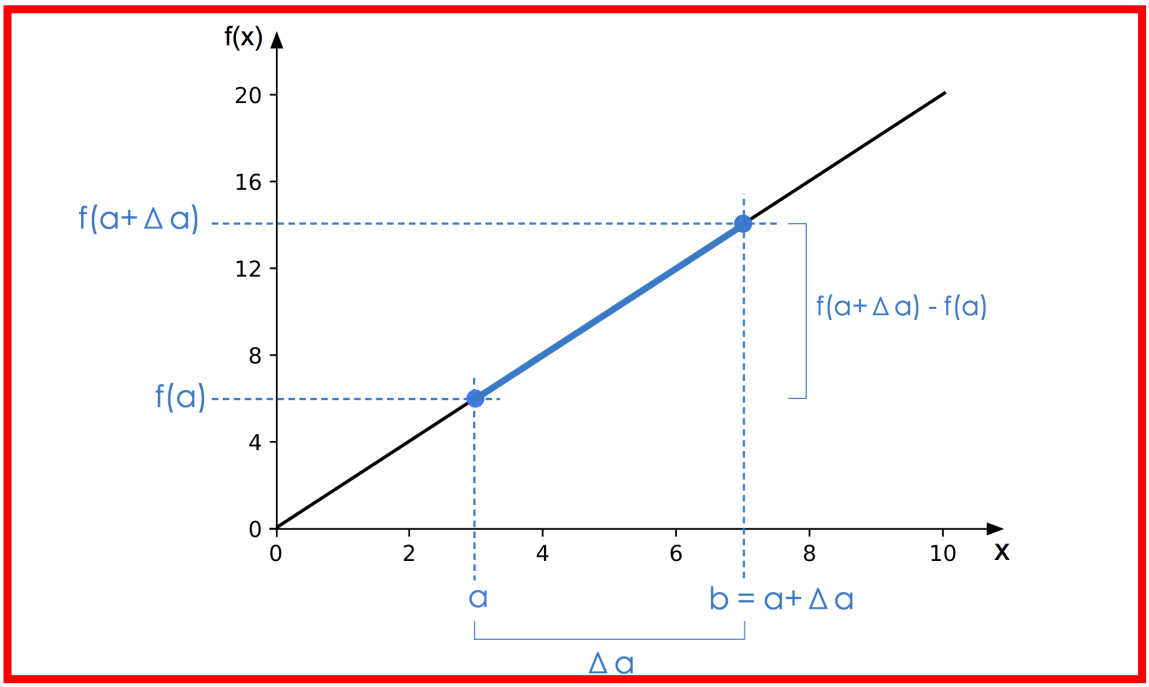


**Figure D.1:** Graph of a linear function, $f(x) = 2x$.

Given the linear function in Equation D.1, we can interpret the "rate of change" as the *slope* of this function. And to compute the slope of a function, we take an arbitrary *x-axis* value, say $a$, and plug it into this function: $f(a)$. Then, we take another value on the *x-axis*, let us call it $b = a + \Delta a$, where $\Delta$ is the change between $a$ and $b$. Now, to compute the *slope* of this linear function, we divide the change in the function's output $f(a + \Delta a)$ by the change in the function's input $a + \Delta a$:

$$\text{Slope} = \frac{f(a + \Delta a) - f(a)}{a + \Delta a - a}. \tag{D.2}$$

In other words, the slope is simply the fraction of the change in $a$ and the function's output:

$$\text{Slope} = \frac{f(a + \Delta a) - f(a)}{a + \Delta a - a} = \frac{f(a + \Delta a) - f(a)}{\Delta a}. \tag{D.3}$$

Now, let's take this intuition, the *slope of a linear function*, and formulate the general definition of the derivative of a continuous function *f(x)*:

$$f'(x) = \frac{df}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}, \tag{D.4}$$

where $\lim_{\Delta x \to 0}$ means "as the change in $x$ becomes infinitely small (for instance, $\Delta x$ approaches zero)." Since this appendix is merely a refresher rather than a comprehensive calculus resource, we have to skip over some important concepts such as *Limit Theory*. So, if this is the first time you encounter calculus, I recommend consulting additional resources such as "Calculus I, II, and III" by Jerrold E. Marsden and Alan Weinstein [1].

**Infobox D.1.1 Derivative Notations**

The two different notations $\frac{df}{dx}$ and $f'(x)$ both refer to the derivative of a function $f(x)$. The former is the "Lagrange notation," and the latter is called "Leibniz notation," respectively. In Leibniz notation, $\frac{df}{dx}$ is sometimes also written as $\frac{d}{dx}f(x)$, and $\frac{d}{dx}$ is an operator that we read as "differentiation with respect to x." Although the Leibniz notation looks a bit verbose at first, it plays nicely into our intuition by regarding $df$ as a small change in the output of a function $f$ and $dx$ as a small change of its input $x$. Hence, we can interpret the ratio $\frac{df}{dx}$ as the slope of a point in a function graph.

Based on the linear function introduced at the beginning of this section (Equation D.1), let us use the concepts introduced in this section to compute the derivative of this function from basic principles. Given the function $f(x) = 2x$, we have

$$f(x + \Delta x) = 2(x + \Delta x) = 2x + 2\Delta x, \tag{D.5}$$

so that

$$\begin{aligned} \frac{df}{dx} &= \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{2x + 2\Delta x - 2x}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{2\Delta x}{\Delta x} \\ &= \lim_{\Delta x \to 0} 2. \end{aligned} \tag{D.6}$$

[1] http:/ /www.cds.caltech.edu/ marsden/volume/Calculus/

We conclude that the derivative of $f(x) = 2x$ is simply a constant, namely $f'(x) = 2$.

Applying these same principles, let us take a look at a slightly more interesting example, a quadratic function,

$$f(x) = x^2, \tag{D.7}$$

as illustrated in Figure D.2.

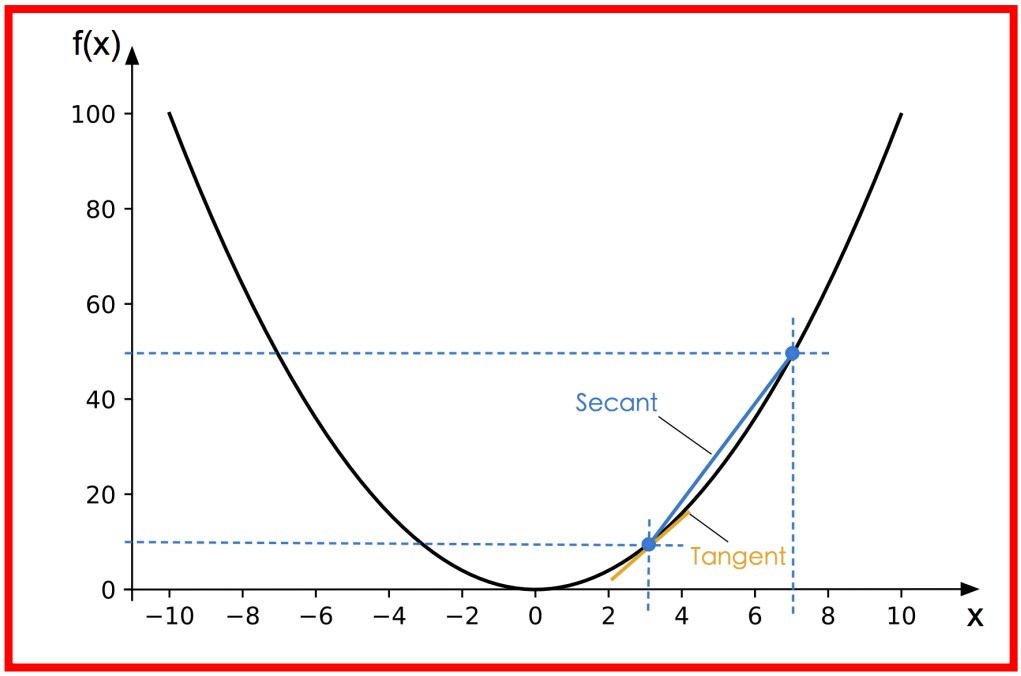


**Figure D.2:** Graph of a quadratic function, $f(x) = x^2$.

As we can see in Figure D.2, this quadratic function (Equation D.7) does not have a constant slope, in contrast to a linear function. Geometrically, we can interpret the derivative of a function as the slope of a tangent to a function graph at any given point. And we can approximate the slope of a tangent at a given point by a secant connecting this point to a second point that is infinitely close, which is where the $\lim_{\Delta x \to 0}$ notation comes from. (In the case of a linear function, the tangent is equal to the secant between two points.)

Now, to compute the derivative of the quadratic function $f(x) = x^2$, we can, again, apply the basic concepts we used earlier, using the fact that

$$f(x + \Delta x) = (x + \Delta x)^2 = x^2 + 2x\Delta x + (\Delta x)^2. \tag{D.8}$$

Now, computing the derivative, we get

$$\begin{aligned}
\frac{df}{dx} &= \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{x^2 + 2x\Delta x + (\Delta x)^2 - x^2}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{2x\Delta x + (\Delta x)^2}{\Delta x} \\
&= \lim_{\Delta x \to 0} 2x + \Delta x.
\end{aligned} \tag{D.9}$$

And since $\Delta x$ approaches zero due to the limit, we arrive at $f'(x) = 2x$, which is the derivative of $f(x) = x^2$.

## D.2 Derivatives of Common Functions

After we gained some intuition in the previous section, this section provides tables and lists of the basic rules for computing function derivatives for our convenience – as an exercise, readers are encouraged to apply the *basic principles* to derive these rules.

The following table, Table D.1, in this subsection lists derivatives of commonly used functions; the intention is that we can use it as quick look-up table. As mentioned earlier, we can obtain these derivates using the basic principles we discussed at the beginning of this appendix. For instance, we just used these basic principles to compute the derivative of a linear function (Table D.1, row 3) and a quadratic function (Table D.1, row 4) earlier on.

| | Function $f(x)$ | Derivative with respect to $x$ |
|---|---|---|
| 1 | $a$ | $0$ |
| 2 | $x$ | $1$ |
| 3 | $ax$ | $a$ |
| 4 | $x^2$ | $2x$ |
| 5 | $x^a$ | $ax^{a-1}$ |
| 6 | $a^x$ | $\log(a)a^x$ |
| 7 | $\log(x)$ | $1/x$ |
| 8 | $\log_a(x)$ | $1/(x\log(a))$ |
| 9 | $\sin(x)$ | $\cos(x)$ |
| 10 | $\cos(x)$ | $-\sin(x)$ |
| 11 | $\tan(x)$ | $\sec^2(x)$ |

**Table D.1:** Derivatives of common functions.

## D.3 Common Differentiation Rules

In addition to the *constant rule* (Table D.1, row 1) and the *power rule* (Table D.1, row 5), the following table lists the most common differentiation rules that we often encounter in practice. Although we will not go over the derivations of these rules, it is highly recommended to memorize and practice them. Most machine learning concepts heavily rely on applications of these rules, and in the following sections, we will pay special attention to the last rule in this list, the chain rule.

| | Function | Derivative |
|---|---|---|
| Sum Rule | $f(x) + g(x)$ | $f'(x) + g'(x)$ |
| Difference Rule | $f(x) - g(x)$ | $f'(x) - g'(x)$ |
| Product Rule | $f(x)g(x)$ | $f'(x)g(x) + f(x)g'(x)$ |
| Quotient Rule | $f(x)/g(x)$ | $[g(x)f'(x) - f(x)g'(x)]/[g(x)]^2$ |
| Reciprocal Rule | $1/f(x)$ | $-[f'(x)]/[f(x)]^2$ |
| Chain Rule | $f(g(x))$ | $f'(g(x))g'(x)$ |

**Table D.2:** Common differentiation rules.

## D.4 The Chain Rule – Computing the Derivative of a Composition of Functions

The chain rule is essential to understanding *backpropagation*; thus, let us discuss it in more detail. In its essence, the chain rule is just a *mental crutch* that we use to differentiate composite functions, functions that are nested within each other. For example,

$$F(x) = f(g(x)). \tag{D.10}$$

To differentiate such a function $F$, we can use this *chain rule*, which we can break down to a three-step procedure. First, we compute the derivative of the outer function $(f')$ with respect to the inner function $(g)$. Second, we compute the derivative of the inner function $(g')$ with respect to its function argument $(x)$. Third, we multiply the outcome of step 1 and step 2:

$$F'(x) = f'(g(x))g'(x). \tag{D.11}$$

Since this notation may look quite daunting, let us use a more visual approach, breaking down the function $F$ into individual steps as illustrated in Figure D.3: We take the argument $x$, feed it to $g$, then, we take the outcome of $g(x)$ and feed it to $f$.

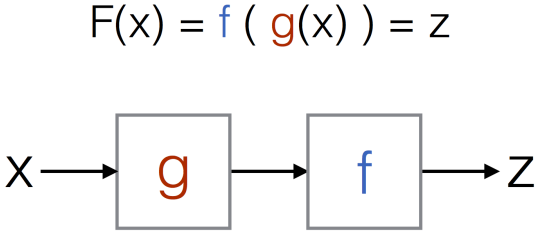


**Figure D.3:** Visual decomposition of a function

Using the chain rule, Figure D.4 illustrates how we can derive $F(x)$ via two parallel steps: We compute the derivative of the inner function $g$ (i.e., $g'(x)$) and multiply it by the outer derivative $f'(g(x))$.

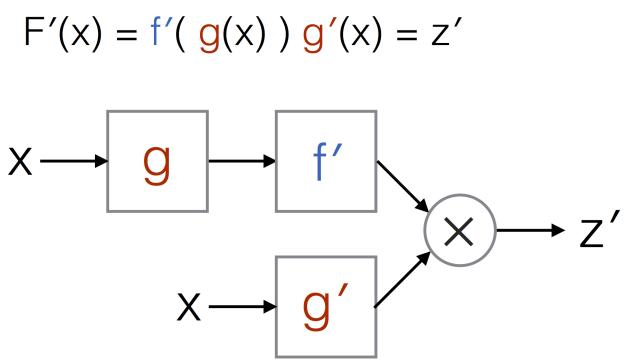


**Figure D.4:** Concept of the chain rule

Now, for the rest of the section, let us use the Leibniz notation, which makes these concepts easier to follow:

$$\frac{d}{dx}[f(g(x))] = \frac{df}{dg} \cdot \frac{dg}{dx}. \tag{D.12}$$

(Remember that the equation above is equivalent to writing $F'(x) = f'(g(x))g'(x)$.)

### D.4.1 A Chain Rule Example

Let us now walk through an application of the chain rule, working through the differentiation of the following function:

$$f(x) = \log(\sqrt{x}). \tag{D.13}$$

**Step 0: Organization**

First, we identify the innermost function:

$$g(x) = \sqrt{x}. \tag{D.14}$$

Using the definition of the inner function, we can now express the outer function in terms of $g(x)$:

$$f(x) = \log(g(x)). \tag{D.15}$$

But before we start executing the chain rule, let us substitute in our definitions into the familiar framework, differentiating function $f$ with respect

to the inner function $g$, multiplied by the derivative of $g$ with respect to the function argument:

$$\frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}, \tag{D.16}$$

which lets us arrive at

$$\frac{df}{dx} = \frac{d}{dg}\log(g) \cdot \frac{d}{dx}\sqrt{x}. \tag{D.17}$$

**Step 1: Derivative of the outer function**

Now that we have set up everything nicely to apply the chain rule, let us compute the derivative of the outer function with respect to the inner function:

$$\frac{d}{dg}\log(g) = \frac{1}{g} = \frac{1}{\sqrt{x}}. \tag{D.18}$$

**Step 2: Derivative of the inner function**

To find the derivative of the inner function with respect to $x$, let us rewrite $g(x)$ as

$$g(x) = \sqrt{x} = x^{1/2}. \tag{D.19}$$

Then, we can use the *power rule* (Table D.1 row 5) to arrive at

$$\frac{d}{dx}x^{1/2} = \frac{1}{2}x^{-1/2} = \frac{1}{2\sqrt{x}}. \tag{D.20}$$

**Step 3: Multiplying inner and outer derivatives**

Finally, we multiply the derivatives of the outer (step 1) and inner function (step 2), to get the derivative of the function $f(x) = \log(\sqrt{x})$:

$$\frac{df}{dx} = \frac{1}{\sqrt{x}} \cdot \frac{1}{2\sqrt{x}} = \frac{1}{2x}. \tag{D.21}$$

## D.5 Arbitrarily Long Function Compositions

In the previous sections, we introduced the chain rule in context of two nested functions. However, the chain rule can also be used for an arbitrarily long function composition. For example, suppose we have five different functions, $f(x), g(x), h(x), u(x)$, and $v(x)$, and let $F$ be the function composition:

$$F(x) = f(g(h(u(v(x))))). \tag{D.22}$$

Then, we compute the derivative as

$$\begin{aligned} \frac{dF}{dx} = \frac{d}{dx}F(x) &= \frac{d}{dx}f(g(h(u(v(x))))) \\ &= \frac{df}{dg} \cdot \frac{dg}{dh} \cdot \frac{dh}{du} \cdot \frac{du}{dv} \cdot \frac{dv}{dx}. \end{aligned} \tag{D.23}$$

As we can see in Equation D.23, composing multiple function is similar to the previous two-function example; here, we create a chain of derivatives of functions with respect to their inner function until we arrive at the innermost function, which we then differentiate with respect to the function parameter $x$.

## D.6 When a Function is Not Differentiable

A function is only differentiable if the derivative exists for each value in the function's domain (for instance, at each point). Non-differentiable functions may be a bit cumbersome to deal with mathematically; however, they can still be useful in practical contexts such as deep learning. A popular example of a non-differentiable function that is widely used in deep learning is the Rectified Linear Unit (ReLU) function. The ReLU function $f(x)$ is not differentiable because its derivative does not exist at $x = 0$, but more about that later in this section.

One criterion for the derivative to exist at a given point is continuity at that point. However, continuity is not sufficient for the derivative to exist. For the derivative to exist, we require the left-hand and the right-hand limit to exist and to be equal.

Remember that conceptually, the derivative at a given point is defined as the slope of a tangent to the function graph at that point. Or in other words, we approximate the function graph at a given point with a straight

line as shown in Figure D.5. (Intuitively, we can say that a curve, when closely observed, resembles a straight line.)

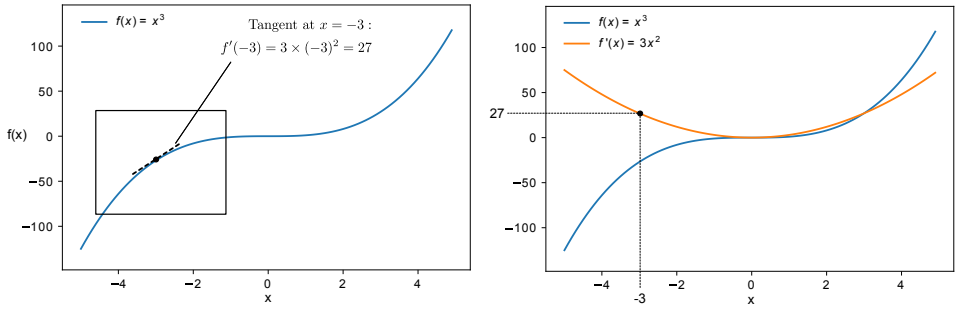


**Figure D.5:** Graph of the function $f(x) = x^3$ with a tangent line to approximate the derivative at point $x = -3$ (left) and the derivative at each point on the function graph (right).

Now, if there are breaks or gaps at a given point, we cannot draw a straight line or tangent approximating the function at that point, because – in intuitive terms – we would not know how to place the tangent. Other common scenarios where derivatives do not exist are sharp turns or corners in a function graph since it is not clear how to place the tangent if we compute the limit from the left or the right side. Finally, any point on a function graph that results in a vertical tangent (parallel to the vertical axis) is not differentiable – note that a vertical line is not a function due to the *one-to-many* mapping condition.

The reason why the derivative of sharp turns or corners (for instance, points on a function graph that are not "smooth") does not exist is that the limit from the left and the right side are different and do not agree. To illustrate this, let us take a look at a simple example, the absolute value function shown in Figure D.6.

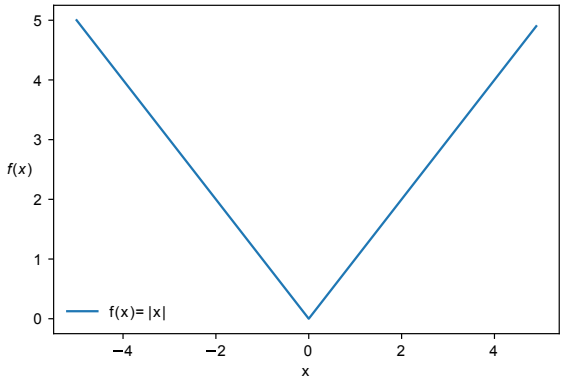


**Figure D.6:** Graph of the "sharp turn"-containing function $f(x) = |x|$

We will now show that the derivative for $f(x) = |x|$ does not exist at the sharp turn at $x = 0$. Recall the definition of the derivative of a continuous function $f(x)$ that was introduced in Section D.1:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}. \tag{D.24}$$

If we substitute $f(x)$ by the absolute value function, $|x|$, we obtain

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

Next, let us set $x = 0$, the point we want to evaluate the equation

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

If the derivative $f'(0)$ exists, it should not matter whether we approach the limit from the left or the right side[2]. So, let us compute the left-side limit first (here, $\Delta x$ represents an infinitely small, negative number):

$$f'(0) = \lim_{\Delta x \to 0^-} \frac{|0 + \Delta x| - |0|}{\Delta x} = \lim_{\Delta x \to 0^-} \frac{|\Delta x|}{\Delta x} = -1.$$

As shown above, the left-hand limit evaluates to $-1$ because dividing a positive number be a negative number yields a negative number. We can now do the same calculation by approaching the limit from the right, where $\Delta x$ is an infinitely small, non-negative number:

[2]Here, "left" and "right" refer to the position of a number on the number line with respect to 0.

$$f'(0) = \lim_{\Delta x \to 0^+} \frac{|0 + \Delta x| - |0|}{\Delta x} = \lim_{\Delta x \to 0^+} \frac{|\Delta x|}{\Delta x} = 1.$$

We can see that the limits are not equal ($1 \neq -1$), and because they do not agree, we have no formal notion of how to draw the tangent line to the function graph at the point $x = 0$. Hence, we say that the derivative of the function $f(x) = |x|$ does not exist (DNE) at point $x = 0$:

$$f'(0) = \text{DNE}.$$

A widely-used function in deep learning applications that is not differentiable at a point[3] is the ReLU function, which was introduced at the beginning of this section. To provide another example of a non-differentiable function, we now apply the concepts of left- and right-hand limits to the piece-wise defined ReLU function (Figure D.7).

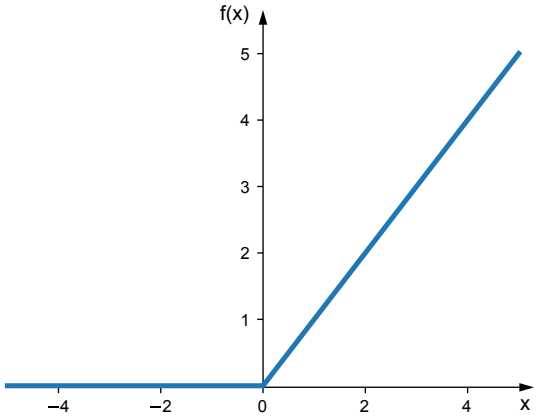


**Figure D.7:** Graph of the ReLU function.

The ReLU function is commonly defined as

$$f(x) = \max(0, x)$$

or

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

[3] Coincidentally, the point where the ReLU function is not defined is also $x = 0$.

(These two function definitions are equivalent.) If we substitute the ReLU equation into Equation D.24, we then obtain

$$f'(x) = \lim_{x \to 0} \frac{\max(0, x + \Delta x) - \max(0, x)}{\Delta x}.$$

Next, let us compute the left- and right-side limits. Starting from the left side, where $\Delta x$ is an infinitely small, negative number, we get

$$f'(0) = \lim_{x \to 0^-} \frac{0 - 0}{\Delta x} = 0.$$

And for the right-hand limit, where $\Delta x$ is an infinitely small, positive number, we get

$$f'(0) = \lim_{x \to 0^+} \frac{0 + \Delta x - 0}{\Delta x} = 1.$$

Again, the left- and right-hand limits are not equal at $x = 0$; hence, the derivative of the ReLU function at $x = 0$ is not defined.

For completeness' sake, the derivative of the ReLU function for $x > 0$ is

$$f'(x) = \lim_{x \to 0} \frac{x + \Delta x - x}{\Delta x} = \frac{\Delta x}{\Delta x} = 1.$$

And for $x < 0$, the ReLU derivative is

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

To summarize, the derivative of the ReLU function is defined as follows:

$$f'(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \\ \text{DNE} & \text{if } x = 0 \end{cases}.$$

**Infobox D.6.1 ReLU Derivative in Deep Learning**

In practical deep learning applications, the ReLU derivative for $x = 0$ is typically set to 0, 1, or 0.5. However, it is extremely rare that $x$ is exactly zero, which is why the decision whether we set the ReLU derivative to 0, 1, or 0.5 has little impact on the parameterization of a neural network with ReLU activation functions.

## D.7 Partial Derivatives and Gradients

Throughout the previous sections, we only looked at univariate functions, functions that only take one input variable, for example, $f(x)$. In this section, we will compute the derivatives of multivariable functions $f(x, y, z, ...)$. Note that we still consider scalar-valued functions, which return a scalar or single value.

While the derivative of a univariate function is a scalar, the derivative of a multivariable function is a vector, the so-called *gradient*. We denote the derivative of a multivariable function $f$ using the gradient symbol $\nabla$ (pronounced "nabla" or "del"):

$$\nabla f = \begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \\ \vdots \end{bmatrix}. \tag{D.25}$$

As we can see, the gradient is simply a vector listing the derivatives of a function with respect to each argument of the function. In Leibniz notation, we use the symbol $\partial$ instead of $d$ to distinguish *partial* from *ordinary* derivatives. The adjective "partial" is based on the idea that a partial derivative with respect to a function argument does not tell the *whole* story about a function $f$. For instance, given a function $f$, the partial derivative $\frac{\partial}{\partial x} f(x, y)$ only considers the change in $f$ if $x$ changes while treating $y$ as a constant.

To illustrate the concept of partial derivatives, let us walk through a concrete example, where we will compute the gradient of the function

$$f(x, y) = x^2 y + y. \tag{D.26}$$

The plot in Figure D.8 shows a graph of this function for different values of $x$ and $y$.

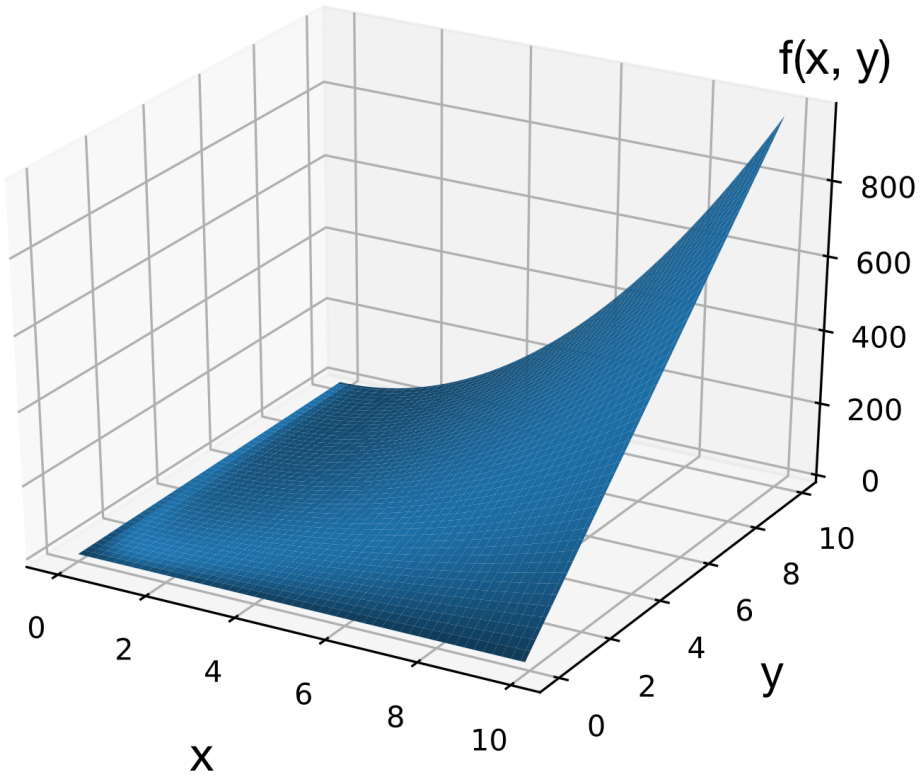


**Figure D.8:** Graph of the function $f(x, y) = x^2y + y$.

The subfigures shown in Figure D.9 illustrate how the function looks like if we treat either $x$ or $y$ as a constant.

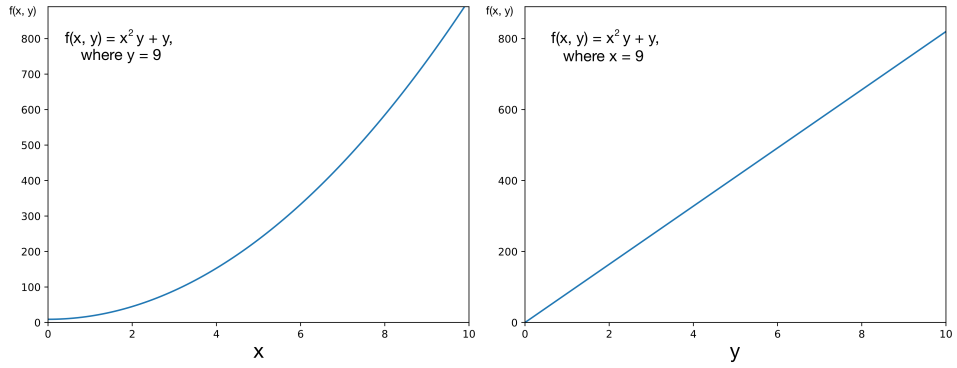


**Figure D.9:** Graph of function $f(x, y) = x^2y + y$ when treating $y$ (left) or $x$ (right) as a constant.

Intuitively, we can think of the two graphs in Figure D.9 as slices of the multivariable function graph shown in Figure D.8. And computing the partial derivative of a multivariable function – with respect to a function's argument – means that we compute the slope of the slice of the multivariable function graph.

Now, to compute the gradient of $f$, we compute the two partial derivatives of that function as follows:

$$\nabla f(x, y) = \begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \end{bmatrix}, \tag{D.27}$$

where

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x} x^2y + y = 2xy \tag{D.28}$$

(via the power rule and constant rule), and

$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y} x^2y + y = x^2 + 1. \tag{D.29}$$

So, the gradient of the function $f$ is defined as

$$\nabla f(x, y) = \begin{bmatrix} 2xy \\ x^2 + 1 \end{bmatrix}. \tag{D.30}$$

## D.8 Second Order Partial Derivatives

Let us briefly go over the notation of second order partial derivatives, since the notation may look a bit strange at first. In a nutshell, the second order partial derivative of a function is the partial derivative of the partial derivative. For instance, we write the second derivative of a function $f$ with respect to $x$ as

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2}. \tag{D.31}$$

For example, we compute the second partial derivative of a function $f(x, y) = x^2 y + y$ as follows:

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x} x^2 y + y\right) = \frac{\partial}{\partial x} 2xy = \frac{\partial}{\partial x} = 2y. \tag{D.32}$$

Note that in the initial definition (Equation D.31) and the example (Equation D.32) both the first and second order partial derivatives were computed with respect to the same input argument, $x$. However, depending on what measurement we are interested in, the second order partial derivative can involve a different input argument. For instance, given a multivariable function with two input arguments, we can in fact compute four distinct second order partial derivatives:

$$\frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial y^2}, \frac{\partial^2 f}{\partial x \partial y}, \text{ and } \frac{\partial^2 f}{\partial y \partial x}, \tag{D.33}$$

where, for example, $\frac{\partial^2 f}{\partial y \partial x}$ is defined as

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right). \tag{D.34}$$

## D.9 The Multivariable Chain Rule

In this section, we will take a look at how to apply the chain rule to functions that take multiple arguments. For instance, let us consider the following function:

$$f(g, h) = g^2 h + h, \tag{D.35}$$

where $g(x) = 3x$, and $h(x) = x^2$. So, as it turns out, our function is a composition of two functions:

$$f(g(x), h(x)) \tag{D.36}$$

Previously, in Section D.4, we defined the chain rule for the univariate case as follows:

$$\frac{d}{dx}[f(g(x))] = \frac{df}{dg} \cdot \frac{dg}{dx}. \tag{D.37}$$

To extend apply this concept to multivariable functions, we simply extend the notation above using the product rule. Hence, we can define the multivariable chain rule as follows:

$$\frac{d}{dx}[f(g(x), h(x))] = \frac{\partial f}{\partial g} \cdot \frac{dg}{dx} + \frac{\partial f}{\partial h} \cdot \frac{dh}{dx}. \tag{D.38}$$

Applying the multivariable chain rule to our multivariable function example $f(g, h) = g^2h + h$, let us start with the partial derivatives:

$$\frac{\partial f}{\partial g} = 2gh \tag{D.39}$$

and

$$\frac{\partial f}{\partial h} = g^2 + 1. \tag{D.40}$$

Next, we take the *ordinary* derivatives of the two functions $g$ and $h$:

$$\frac{dg}{dx} = \frac{d}{dx}3x = 3 \tag{D.41}$$

$$\frac{dh}{dx} = \frac{d}{dx}x^2 = 2x. \tag{D.42}$$

And finally, plugging everything into our *multivariable chain rule* definition, we arrive at

$$\frac{d}{dx}[f(g(x), h(x))] = [2gh \cdot 3] + [g^2 + 1 \cdot 2x] = g^2 + 6gh + 2x. \tag{D.43}$$

## D.10 The Multivariable Chain Rule in Vector Form

After we introduced the general concept of the multivariable chain rule, we often prefer a more compact notation in practice: the multivariable chain rule in vector form.

**Infobox D.10.1 Dot Products**

As we remember from the linear algebra appendix, we compute the *dot product* between two vectors, $\mathbf{a}$ and $\mathbf{b}$ , as follows:

$$\mathbf{a} \cdot \mathbf{b} = \begin{bmatrix} a \\ b \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = ax + by$$

In vector form, we write the multivariable chain rule

$$\frac{d}{dx}[f(g(x), h(x))] = \frac{\partial f}{\partial g} \cdot \frac{dg}{dx} + \frac{\partial f}{\partial h} \cdot \frac{dh}{dx} \tag{D.44}$$

as follows:

$$\frac{d}{dx}[f(g(x), h(x))] = \nabla f \cdot \mathbf{v}'(x). \tag{D.45}$$

Here, $\mathbf{v}$ is a vector listing the function arguments:

$$\mathbf{v}(x) = \begin{bmatrix} g(x) \\ h(x) \end{bmatrix}. \tag{D.46}$$

And the derivative ("v-prime" in Lagrange notation) is defined as follows:

$$\mathbf{v}'(x) = \frac{d}{dx} \begin{bmatrix} g(x) \\ h(x) \end{bmatrix} = \begin{bmatrix} dg/dx \\ dh/dx \end{bmatrix}. \tag{D.47}$$

So, putting everything together, we have

$$\nabla f \cdot \mathbf{v}'(x) = \begin{bmatrix} \partial f/\partial g \\ \partial f/\partial h \end{bmatrix} \cdot \begin{bmatrix} dg/dx \\ dh/dx \end{bmatrix} = \frac{\partial f}{\partial g} \cdot \frac{dg}{dx} + \frac{\partial f}{\partial h} \cdot \frac{dh}{dx}. \tag{D.48}$$

## D.11 The Hessian Matrix

As mentioned earlier in Section D.8 *Second Order Partial Derivatives*, we can compute four distinct partial derivatives for a two-variable function:

$$f(x, y). \tag{D.49}$$

The Hessian matrix is simply a matrix that packages them up:

$$Hf = \begin{bmatrix} \partial^2 f/\partial x^2 & \partial^2 f/\partial x \partial y \\ \partial^2 f/\partial y \partial x & \partial^2 f/\partial y^2 \end{bmatrix}. \tag{D.50}$$

To formulate the Hessian for a multivariable function that takes $n$ arguments,

$$f(x_1, x_2, ..., x_n), \tag{D.51}$$

we write the Hessian as

$$Hf = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}. \tag{D.52}$$

## D.12 The Laplacian Operator

At its core, the Laplacian operator ($\Delta$) is an operator that takes in a function and returns another function. In particular, it is the divergence of the gradient of a function $f$ – a kind of second order partial derivative, or "the dircection that increases the direction most rapidly:"

$$\Delta f(g(x), h(x)) = \nabla \cdot \nabla f. \tag{D.53}$$

Remember, we compute the gradient of a function $f(g, h)$ as follows:

$$\nabla f(g, h) = \begin{bmatrix} \partial f/\partial g \\ \partial f/\partial h \end{bmatrix}. \tag{D.54}$$

Plugging it into the definition of the Laplacian, we arrive at

$$\Delta f(g(x), h(x)) = \begin{bmatrix} \partial f/\partial g \\ \partial f/\partial h \end{bmatrix} \cdot \begin{bmatrix} \partial f/\partial g \\ \partial f/\partial h \end{bmatrix} f = \frac{\partial^2 f}{\partial g^2} + \frac{\partial^2 f}{\partial h^2}. \tag{D.55}$$

And in more general terms, we can define the Laplacian of a function

$$f(x_1, x_2, ..., x_n) \tag{D.56}$$

as

$$\Delta f = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}. \tag{D.57}$$

# Bibliography

[Boyer, 1970] Boyer, C. B. (1970). The history of the calculus. *The Two-Year College Mathematics Journal*, 1(1):60–86.

# Abbreviations and Terms

**AMI** [Amazon Machine Image]
**API** [Application Programming Interface]
**CNN** [Convolutional Neural Network]
**DNE** [Does Not Exist]

# Index