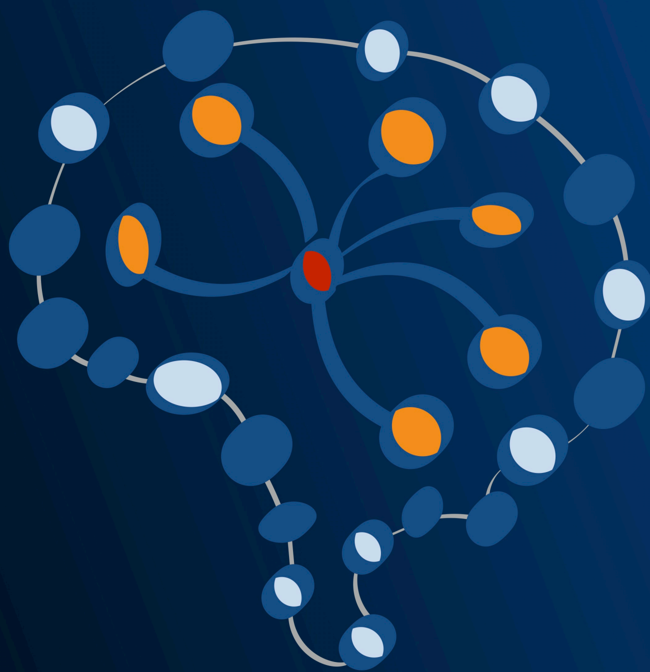
SEBASTIAN RASCHKA



Introduction to Artificial Neural Networks and Deep Learning

A Practical Guide
with Applications in Python

Introduction to Artificial Neural Networks and Deep Learning

A Practical Guide with Applications in Python

Sebastian Raschka

This book is for sale at http://leanpub.com/ann-and-deeplearning

This version was published on 2017-02-24

Leanpub

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

Contents

Website

Please visit the GitHub repository[1] to download code examples used in this book.

If you like the content, please consider supporting the work by buying a copy of the book on Leanpub[2].

I would appreciate hearing your opinion and feedback about the book! Also, if you have any questions about the contents, please don't hesitate to get in touch with me via mail@sebastianraschka.com or join the mailing list[3].

Happy learning!

Sebastian Raschka

[1] https://github.com/rasbt/deep-learning-book

[2] https://leanpub.com/ann-and-deeplearning

[3] https://groups.google.com/forum/#!forum/ann-and-dl-book

Acknowledgments

I would like to give my special thanks to the readers, who 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

Appendix B – Algebra Basics

In this appendix, we will refresh a subset of the fundamental Algebra concepts that are relevant for following the mathematics behind the machine learning algorithms presented in this book. While this appendix aims to bring you up to speed regarding *Algebra Basics*, please bear in mind that it is not a comprehensive algebra resource. Thus, I recommend you to consult a more comprehensive resource, such as *Beginning and Intermediate Algebra* by Tyler Wallace [4], if you wish to study algebra in more depth.

What is Algebra? Algebra is a subfield of mathematics that uses numbers and letters (or symbols) in consent to describe relationships. Algebra is often used to solve problems where at least one quantity is unknown or may vary (we call these *variables*). In simple terms, we can think of algebra as a language that combines letters, symbols, and numbers, to generalize arithmetic operations and to solve equations with variables.

Exponents

Almost every equation or function in computer science and engineering involves exponents. And based on the assumption that the vast majority of readers have worked with exponents before, this section provides a summary of the most important concepts in tabular form, which is intended to serve as a refresher as well as a look-up reference.

[4]Wallace, Tyler. "Beginning and Intermediate Algebra." (2010)., http://www.wallace.ccfaculty.org/book/book.html

Basic Principles

	Rule	Example
1	$x^0 = 1$	$5^0 = 1$
2	$x^1 = x$	$5^1 = 5$
3	$x^n = x \cdot x \cdot x \cdot \ldots \cdot x$	$5^3 = 5 \cdot 5 \cdot 5 = 125$
4	$x^{-n} = \frac{1}{x^n}$	$5^{-2} = \frac{1}{5^2} = \frac{1}{25}$
5	$x^{1/2} = \sqrt{x}$	$25^{1/2} = \sqrt{25} = 5$
6	$x^{1/n} = \sqrt[n]{x}$	$125^{1/3} = \sqrt[3]{125} = 5$
7	$x^{m/n} = \sqrt[n]{x^m}$	$125^{2/3} = \sqrt[3]{125^2} = 25$

Laws of Exponents

The following table lists the fundamental laws of exponents that we can apply to manipulate algebraic equations:

	Law	Example
1	$x^m \cdot x^n = x^{m+n}$	$5^3 \cdot 5^2 = 5^5 = 3125$
2	$x^m / x^n = x^{m-n}$	$\frac{5^3}{5^2} = 5^1 = 5$
3	$\left(x^m\right)^n = x^{m \cdot n}$	$(5^2)^3 = 5^6 = 15,625$
4	$(xy)^n = x^n \cdot y^n$	$(2 \cdot 5)^2 = 2^2 \cdot 5^2 = 100$
5	$\left(\frac{x}{y}\right)^n = \frac{x^n}{y^n}$	$\left(\frac{5}{2}\right)^2 = \left(\frac{5^2}{2^2}\right) = \frac{25}{4}$
6	$\left(\frac{x}{y}\right)^{-n} = \left(\frac{y}{x}\right)^n$	$\left(\frac{5}{2}\right)^{-2} = \left(\frac{2}{5}\right)^2 = \left(\frac{2^2}{5^2}\right) = \frac{4}{25}$

While derivations of the exponent rules are beyond the scope of this algebra refresher, notice that these rules are all closely related to each other, and they are based on the basic principles that were presented in the previous section. For instance, while $x^0 = 1$ may seem a bit unintuitive at first, we can demonstrate why this relation is true by applying the second rule from the table above:

$$1 = \frac{x^n}{x^n} = x^{n-n} = x^0. \tag{1}$$

Roots

This section lists the basic concepts and rules regarding roots. The root of a number *x* is a number that, if multiplied by itself a certain number of times, equals *x*. Written in a different form, the equation $\sqrt[m]{x} = y$ is true *if and only if* $y^m = x$, where m is the root *index* and x is the *radicand*.

	Equation	**Example**
1	$(\sqrt{x})^2 = \sqrt{x^2} = x$ for all positive numbers	$(\sqrt{4})^2 = \sqrt{4^2} = 4$
2	$\sqrt{x \cdot y} = \sqrt{x} \cdot \sqrt{y}$	$\sqrt{36} = \sqrt{9 \cdot 4} = \sqrt{9} \cdot \sqrt{4} = 3 \cdot 2 = 6$
3	$\sqrt{\frac{x}{y}} = \frac{\sqrt{x}}{\sqrt{y}}$	$\sqrt{\frac{9}{16}} = \frac{\sqrt{9}}{\sqrt{16}} = \frac{3}{4}$
4	$\sqrt{x+y} \neq \sqrt{x} + \sqrt{y}$ $\sqrt{x^2+y^2} \neq x + y$	$\sqrt{16+9} \neq 4 + 3$

Properties of Logarithms

Logarithms are closely related to exponents; given the same base, the logarithm is the inverse of the exponential function and vice versa. For instance, if we are given an equation $2^x = 9$, we can obtain the value of the variable x by applying the *base 2* logarithm:

$$\log_2(2^x) = x. \tag{2}$$

In a more general form, we can express this relationship as follows:

$$\log_b(a) = c \iff b^c = a, \tag{3}$$

where a is the *power*, b is the *base*, and c is the *exponent*.

The following table summarizes the most important properties of logarithms that we use in practice:

	Property	**Example**
1	$\log_b 1 = 0$	$\log_2 1 = 0$
2	$\log_b b = 1$	$\log_2 2 = 1$
3	$\log_b b^a = a$	$\log_2 2^3 = 3$
4	$b^{\log_b a} = a$	$2^{log_2 8} = 2^3 = 8$
5	$\log_b(m \cdot n) = \log_b m + \log_b n$	$\log_2(2{\cdot}4) = \log_2 2{+}\log_2 4 = 1{+}2 = 3$
6	$\log_b\left(\frac{m}{n}\right) = \log_b m - \log_b n$	$\log_2\left(\frac{4}{2}\right) = \log_2 4{-}\log_2 2 = 2{-}1 = 1$
7	$\log_b(m^p) = p\log_b m$	$\log_2(2^3) = 3\log_2 2 = 3$
8	$\log_b m = \frac{\log_a m}{\log_a b}$	$\log_2 8 = \frac{\log_{10} 8}{\log_{10} 2} = 3$

Also here, we can derive the different properties from each other. For example, the last property in this list can be obtained as follows, starting with the basic properties of logarithms:

$$\begin{aligned} \log_b m &= c \\ b^c &= m. \end{aligned} \tag{4}$$

Taking the desired logarithm on both sides, we have

$$\log_a b^c = \log_a m. \tag{5}$$

Next, we move the exponent outside, using the rule from row 7:

$$c\log_a b = \log_a m. \tag{6}$$

Finally, we divide by the logarithm on the right, $\log_a(b)$ and substitute c by $\log_b m$:

$$c = \frac{\log_a m}{\log_a b} \log_b m = \frac{\log_a m}{\log_a b} \tag{7}$$

Natural Logarithms

Throughout this book, we will mostly work with *natural logarithms* – logarithms to the base of the mathematical constant e. When we are working with such natural logarithms, we will follow the common convention and omit writing the base of the logarithm explicitly. For example, instead of writing $\log_e(e^x) = x$, we will use the simpler notation $\log(e^x) = x$.

Order of Operations

This section provides a brief overview of the basic laws of algebra, which are based on the following order of operations:

1. Perform all calculations inside parentheses.
2. Perform multiplication and division operations from left to right.
3. Perform addition and subtraction from left to right.

Commutative, Associative, and Distributive Laws

The *commutative law* says that we can disregard the order of two real numbers if we perform operations such as addition or multiplication:

$$x + y = y + x \tag{8}$$

and

$$x \cdot y = y \cdot x. \tag{9}$$

For example, $5 + 2 = 2 + 5$ and $5 \cdot 2 = 2 \cdot 5$.

Per *associative law*, we can disregard the grouping of numbers in operations such as addition and multiplication:

$$(x + y) + z = x + (y + z) \tag{10}$$

and

$$(x \cdot y) \cdot z = x \cdot (y \cdot z). \tag{11}$$

For example, $(5 + 2) + 3 = 5 + (2 + 3)$ and $(5 \cdot 2) \cdot 3 = 5 \cdot (2 \cdot 3)$.

The *distributive law* is extremely useful for multiplying and factoring algebraic expressions. With respect to multiplication and addition, the distributive law says that

$$x \cdot (y + z) = x \cdot y + x \cdot z, \tag{12}$$

for example,

$$\begin{aligned} 5 \cdot (3 + 2) &= 5 \cdot 3 + 5 \cdot 2 \\ &= 15 + 10 \\ &= 25. \end{aligned} \tag{13}$$

Notice that this is similar to performing the addition operation inside the parentheses first, for instance,

$$\begin{aligned} 5 \cdot (3 + 2) &= 5 \cdot 5 \\ &= 25. \end{aligned} \tag{14}$$

Notation

In practice, we often omit the center "dot" symbol ($\cdot$) as multiplication operator whenever it is obvious from the context. For example, we could write $2x$ instead of $2 \cdot x$ or $5(3 + 2)$ instead of $5 \cdot (3 + 2)$.

Removing Parentheses and Expanding

If an expression, enclosed by parentheses, is preceded by a *plus sign*, we can simply remove the parentheses:

$$\begin{aligned} 3 + (5x - 2y + 4) &= 3 + 5x - 2y + 4 \\ &= 5x - 2y + 7. \end{aligned} \tag{15}$$

However, if there is a *minus sign* in front of the opening parenthesis, we have to be more careful and reverse addition and subtraction inside the parentheses:

$$\begin{aligned} 3 - (5x - 2y + 4) &= 3 - (-1)(5x - 2y + 4) \\ &= 3 - 5x + 2y - 4 \\ &= -5x + 2y - 1. \end{aligned} \tag{16}$$

As we have seen in the previous subsection, we can get rid of parentheses by using the distributive law, which says that we have to multiply a factor onto everything inside the parentheses:

$$\begin{aligned} &5(3x + 7) \\ &= 15x + 35. \end{aligned} \tag{17}$$

Of course, the same is true for division, since division is simply *multiplication by the reciprocal*:

$$\begin{aligned} 5/(5x + 10) &= \frac{1}{5}(5x + 10) \\ &= x + 2 \end{aligned} \tag{18}$$

Now, if we are multiplying two expressions containing variables, we can simply use the distributive law and distribute the terms of the first expression over the second:

$$\begin{aligned} (a + b)(c + d) &= a(c + d) + b(c + d) \\ &= ac + ad + bc + bd \end{aligned} \tag{19}$$

For example,

$$
\begin{aligned}
(2x-3)(3x+4) &= 2x(3x+4) - 3(3x+4) \\
&= 6x^2 + 8x - 9x + 12 \\
&= 6x^2 - x + 12.
\end{aligned} \tag{20}
$$

Polynomials

At the end of the previous section, we multiplied two *binomials* – binomials are *polynomials* with two terms. A *polynomial* is an expression that contains multiple terms (*poly* means "many" in Greek).

Polynomials may consist of constants, variables, and exponents, but they never divide a term by a variable. Thus, polynomials also have no negative exponents, for example, $\sqrt{x} = x^{1/2}$ is not a polynomial.

In *standard form*, we write a polynomial as a combination of monomials sorted by their degree. For instance, the following polynomial, written in standard form, has three terms:

$$4x^3 + 2x^2 + y. \tag{21}$$

The degree of a polynomial is defined by the degree of its highest term, and the degree of the highest term is defined by the sum of its exponents. For instance, the polynomial $4x^3+2x^2+y$ has degree 3.

However, note that the following polynomial has degree 7, since $3 + 4 = 7$:

$$4x^3y^4 + 2x^2 + y. \tag{22}$$

In general form, we can express a univariate nth-degree polynomial (a polynomial containing only a single variable) as follows:

$$a_n x^n + a_{n-1}x^{n-1} + ... + a_2x^2 + a_1x + a_0, \tag{23}$$

where $a_n, ...a_0$ are coefficients. In practice, however, we will mostly work with polynomials with a degree between 0 and 3:

Degree	Common Name	Example
0	constant	$2x^0 = 2$
1	linear	$2x + 1$
2	quadratic	$2x^2 + 3x + 1$
3	cubic	$2x^3 + 5x^2 + x - 5$

Summation and Products of Sequences

In the previous section, we expressed a univariate nth-degree polynomial as

$$a_n x^n + a_{n-1} x^{n-1} + \ldots + a_2 x^2 + a_1 x + a_0. \tag{24}$$

In practice, when we are expressing sequences with an indexed variable, we use the summation symbol $\sum$ for our convenience. For instance, we could rewrite the previous expression as

$$\sum_{i=0}^{n} a_i x^i, \tag{25}$$

where i is the *index of summation*, and n is the *upper bound of summation*. For a simpler example, consider the following:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ \ldots \ + x_n. \tag{26}$$

Similarly, we use the *Capital Pi Notation*, $\prod$, to express products of sequences with an index variable more concisely. For example,

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \ \ldots \ \cdot x_n. \tag{27}$$

Algebra of Sets

Intuitively, a *set* represents a collection of different and unique elements. In this section, we will go over basic set theory, discussing the properties and laws of such sets.

Set Notation

There are several different notations for defining sets; the three most common ones are listed below:

1. Verbal notation: *Let S be the set of letters in the word "algebra."*
2. Roster notation: $\{a, l, g, e, b, r\}$
3. Set-builder notation: $\{x \mid x \text{ is a letter in "algebra"}\}$

Although, the different set notations can be used equivalently, it is sometimes more convenient or appropriate to use one notation over the other. For example, we can define the set of integers as

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}, \tag{28}$$

while we write the domain of all positive real numbers as

$$\mathbb{R}_{>0} = \{x \in \mathbb{R} \mid x > 0\}. \tag{29}$$

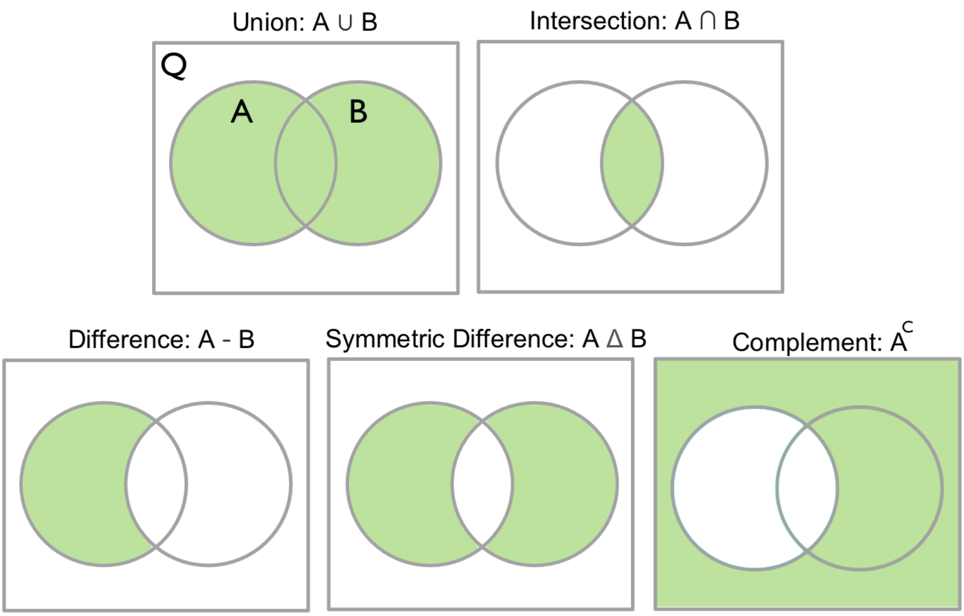
(The symbol $\in$ means "element of." For example, $x \in A$ translates to "x is an element of set A.")

Set Theory

Now, using the set-builder notation, we can summarize the basic set operations as follows:

- Union: $A \cup B = \{x \mid x \in A \text{ or } x \in B, \text{ or both}\}$
- Intersection: $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$
- Difference: $A - B = \{x \mid x \in A \text{ and } x \notin B\}$
- Symmetric difference: $A \Delta B = \{x \mid x \notin A \cap B\}$
- Complement: $A^C = \{x \mid x \notin A\}$

The figure below provides a visual summary of these operations, given the two sets A and B, and the universal set Q – a set which contains all possible elements, including itself.



Interval Notation

Another concept that is related to set theory is the interval notation, which is especially handy when we are describing real-valued functions. Or in other words, the interval notation is essentially just a less complicated form of the set-builder notation when we are working with real numbers and intervals:

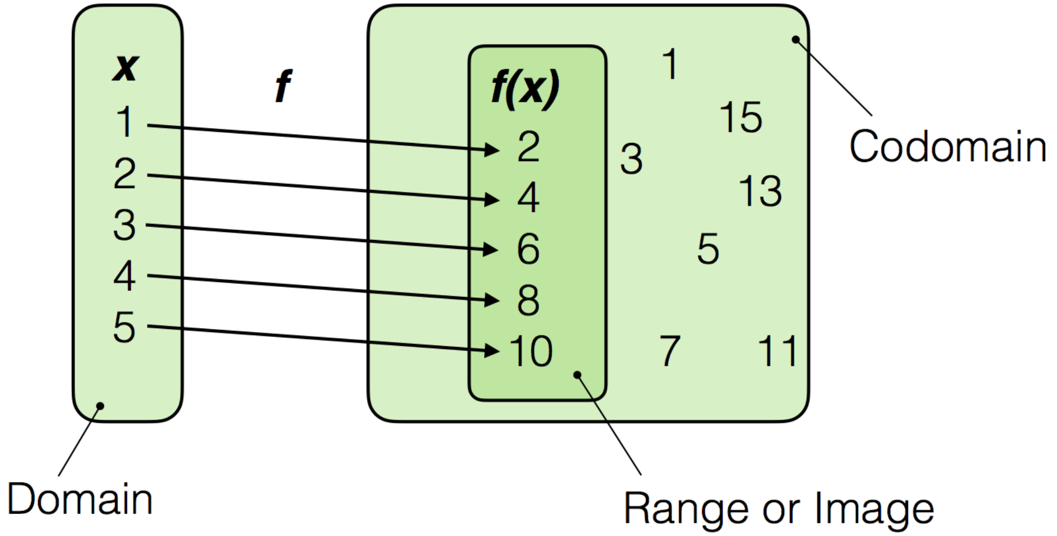
Set of Real Numbers	Interval Notation
$\{x \in \mathbb{R} \| a < x < b\}$	(a, b)
$\{x \in \mathbb{R} \| a \leq x < b\}$	$[a, b)$
$\{x \in \mathbb{R} \| \leq a\}$	$(-\infty, a]$
$\{x \in \mathbb{R} \| \geq a\}$	$[a, \infty)$

Domain and Range of a Function

Now that we have introduced all the essential notation, we can finally discuss *functions*. Intuitively, a function is a relationship between a set of inputs and outputs. We can think of a function f as a box that takes in one input and returns one output, $f : x \rightarrow y$.

For a valid function, we require that one input is mapped to one output, a one-to-many mapping is not allowed. In other words, a function must always return the same value for a specific input. This is also called one-to-one mapping. However, note that a function is also valid under the many-to-one condition, for example, consider the function $f(x) = x^2$. Here, both input values $x = -2$ and $x = 2$ produce the same output: $f(-2) = f(2) = 4$.

The *domain* of a function is defined as the set of all possible input values for which the function is defined, and the *codomain* is the set of values that it returns. For practical reasons, since we often do not have full knowledge of the codomain, we use the term *range* (or *image*), which defines the actual set of values produced by the function. So, while the codomain is part of the function definition and defines what possibly comes out, the range is a subset of the codomain. The following figure provides a visual summary of these terms given the function $f(x) = 2x$:



To better illustrate the notion of a domain, let us take a look at an example and consider the following rational function (a rational function is a function that consists of the fractional ratio of two polynomials):

$$f(x) = \frac{\sqrt{3-x}}{x-2}. \tag{30}$$

Using interval notation, we define the domain of this function as $(-\infty, 2) \cup (2, 3]$, since we cannot perform a division by zero, and we cannot take the root of a negative number. (The equivalent set-builder notation of the domain is: $\{x \in \mathbb{R} \mid x < 2 \text{ or } 2 < x <= 3\}$.)

Inverse Functions

Before we take a look at inverse function, let us briefly introduce another important concept that we use frequently in machine learning: *function composition*, the nesting of two or more functions:

$$(f \circ g)(x) = f(g(x)). \tag{31}$$

Now, let g be the inverse function of another function f, so that

- $(f \circ g)(x) = x$ for all x in the domain of f
- $(g \circ f)(x) = f$ for all x in the domain of g

Intuitively, the composition of a function and its inverse is equivalent to the identity function I, where $I(x) = x$. For example, if we define $f(x) = 2x$, the inverse function g must be defined so that it satisfies the equation $(f \circ g)(2) = 2$. Thus, $g(x) = \frac{x}{2}$.

In practice, we use the superscript notation to denote a function inverse, for example, f^{-1}. This notation may look confusing at first, and the origin of this notation may be due to the fact that

$$f^{-1} \circ f = f \circ f^{-1} = I \tag{32}$$

looks similar to the following equation:

$$x^{-1} \cdot x = x \cdot x^{-1} = x \cdot \frac{1}{x} = 1. \tag{33}$$

However, notice that superscript (-1) does not mean that we take the function f to the power of (-1):

$$f^{-1}(x) \neq \frac{1}{f(x)}. \tag{34}$$

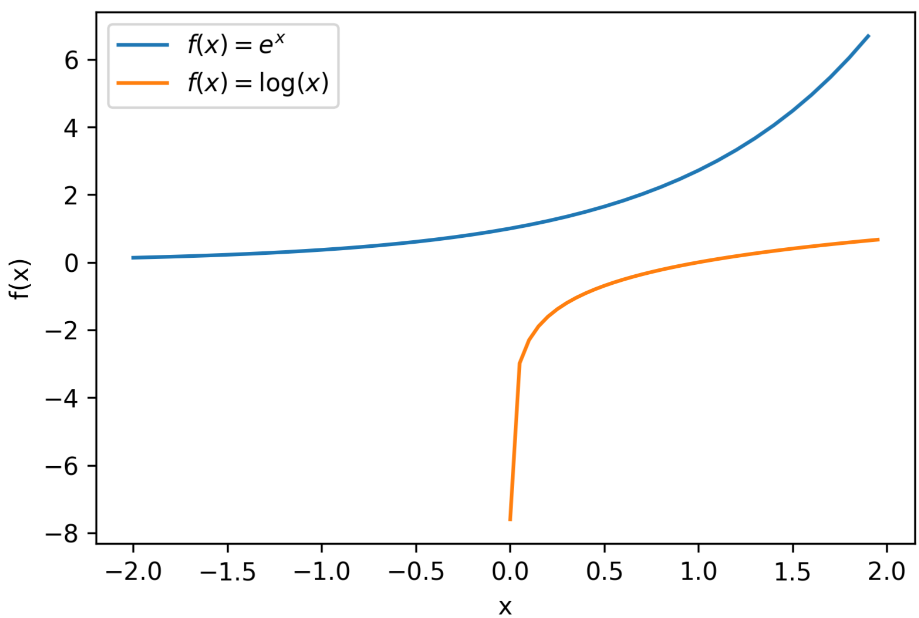
A good example of inverse functions are exponential and logarithmic functions: Logarithms are the inverse of exponential functions and vice versa. For instance, consider the following exponential function

$$f(x) = b^x, \tag{35}$$

where its inverse is the "base b" logarithm defined as

$$f^{-1}(x) = log_b(x). \tag{36}$$

So, we can say that the domain of logarithmic functions is the range of exponential functions $(0, \infty)$, and the range of logarithmic functions is the domain of exponential functions $(-\infty, \infty)$:



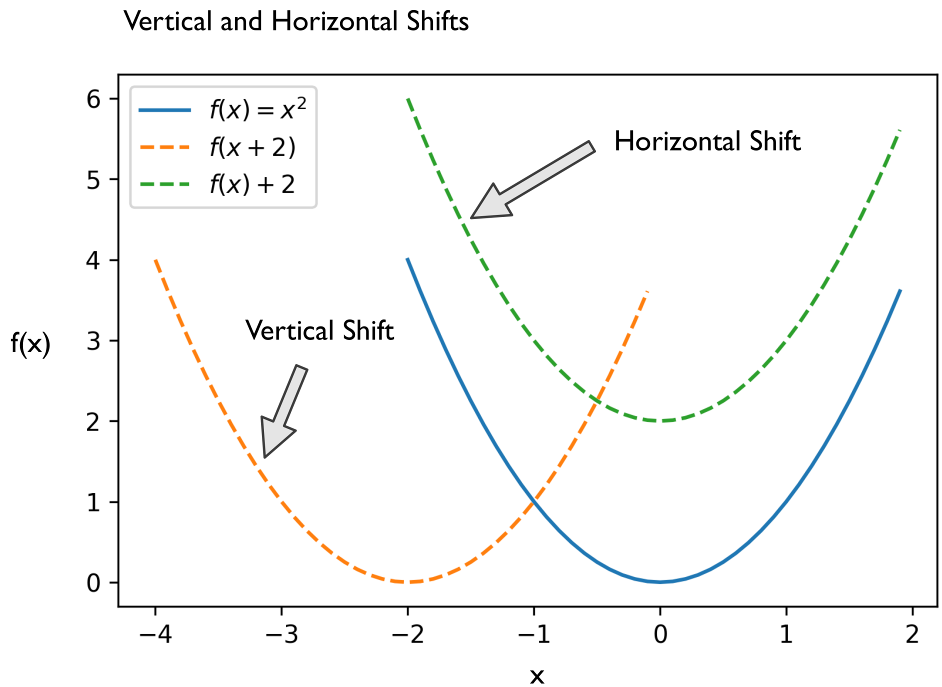
Function Transformations

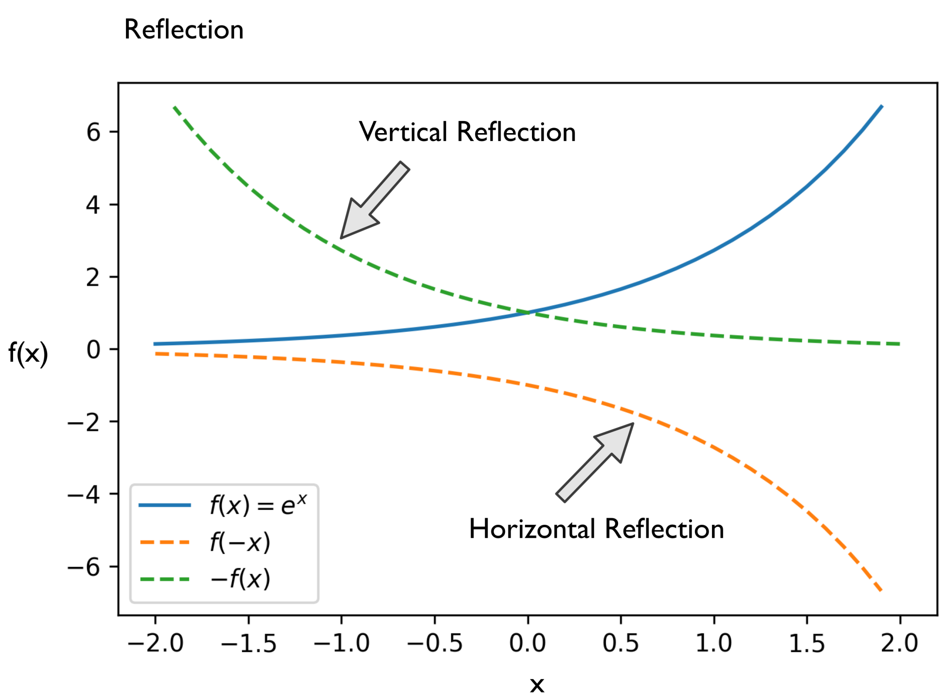
Lastly, this section introduces the basics of *function transformation*, which is a concept that will give us some intuition of how we can modify a function to our needs. Assume we are given a basis function $f(x)$, and it is decorated by the $a, b, c,$ and d in the following way:

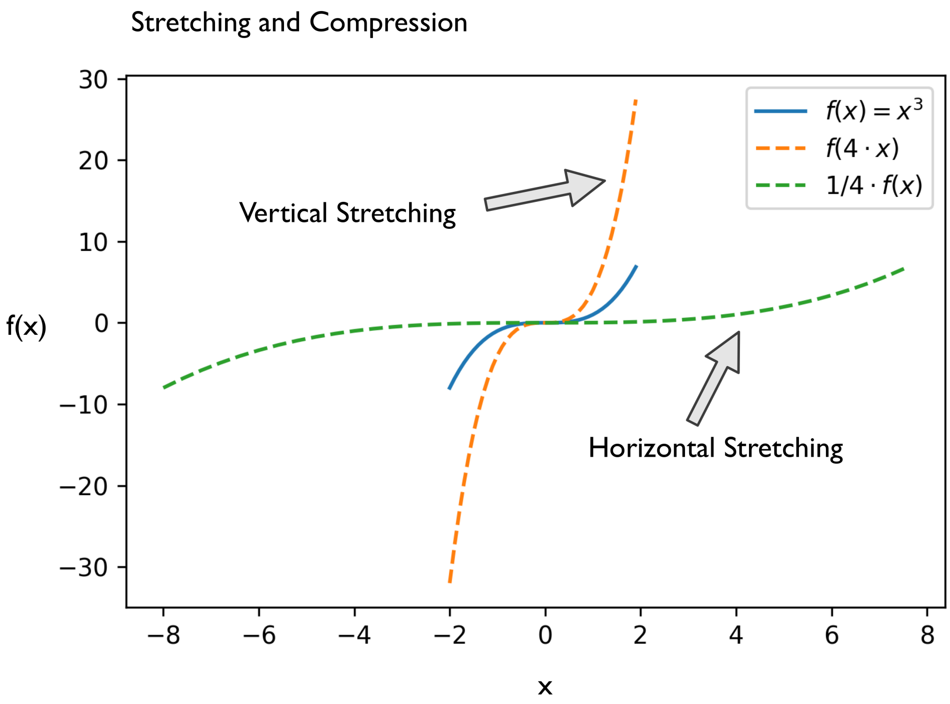
$$af(b(x+c))+d. \tag{37}$$

- a: vertical stretch by a factor of a if $a > 1$; compression if $0 < a < 1$; flip the graph if a is negative.
- b: horizontal shrinking by a factor of b if $b > 1$; stretching if $0 < b < 1$; flip the graph if b is negative.
- c: horizontal shift upwards if c is positive; otherwise, shift the function graph downwards.
- d: vertical shift to the left if d is positive; otherwise, shift the function graph to the right.

The following above-mentioned concepts are illustrated in the figures below:


Vertical and Horizontal Shifts
$f(x) = x^2$
$f(x+2)$
$f(x)+2$
Horizontal Shift
Vertical Shift
f(x)
6
5
4
3
2
1
0
−4
−3
−2
−1
0
1
2
x


Reflection
Vertical Reflection
Horizontal Reflection
$f(x) = e^x$
$f(-x)$
$-f(x)$
f(x)
6
4
2
0
−2
−4
−6
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
x


Stretching and Compression
30
20
10
0
−10
−20
−30
f(x)
−8
−6
−4
−2
0
2
4
6
8
x
$f(x) = x^3$
$f(4 \cdot x)$
$1/4 \cdot f(x)$
Vertical Stretching
Horizontal Stretching