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 2017-10-31

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.

© 2016 - 2017 Sebastian Raschka

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 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

Appendix D - Calculus and Differentiation Primer

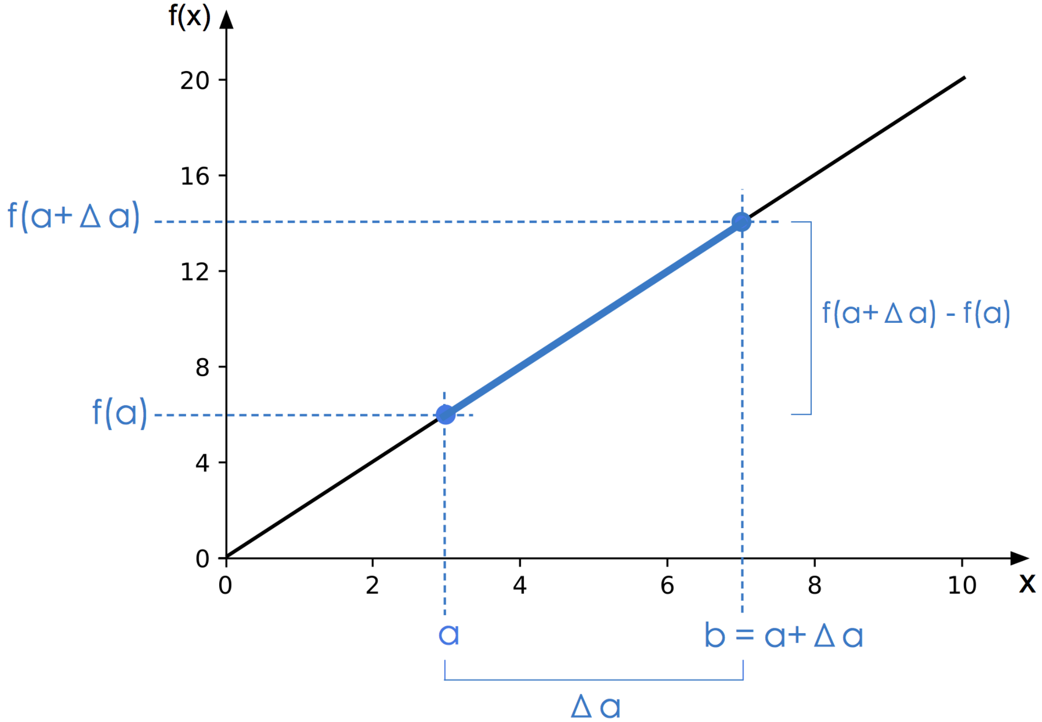
Calculus is a discipline of mathematics that provides us with *tools* to analyze rates of change, or decay, or motion. Both Isaac Newton and Gottfried Leibniz developed the foundations of calculus independently in the 17th century. Although we recognize Gottfried and Leibniz as the founding fathers of 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[4].

In this appendix we will only concentrate on the subfield of calculus that is of most relevance to machine and deep learning: differential calculus: 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.

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 $f(x) = 2x$ shown in the graph below:

[4] Boyer, Carl B. "The history of the calculus." The Two-Year College Mathematics Journal 1.1 (1970): 60-86.



Given this linear function above, 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 Δa 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 $a + \Delta a$:

$$\text{Slope} = \frac{f(a + \Delta a) - f(a)}{a + \Delta a - a}. \tag{1}$$

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

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

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)*:

$$f'(x) = \frac{df}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}, \tag{3}$$

where $\lim_{\Delta x \to 0}$ means "as the change in x becomes infinitely small (for instance, Δx 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[5].

Notation

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 small change in the output of a function f and dx as a small change of its input x – so we can interpret the ratio $\frac{df}{dx}$ as the slope of a point in a function graph.

Since we introduced the linear function $f(x) = 2x$ at the beginning of this section, let us use these concepts to compute the derivative of this function from basic principles:

Given the function $f(x) = 2x$, we have

$$f(x + \Delta x) = 2(x + \Delta x) = 2x + 2\Delta x, \tag{4}$$

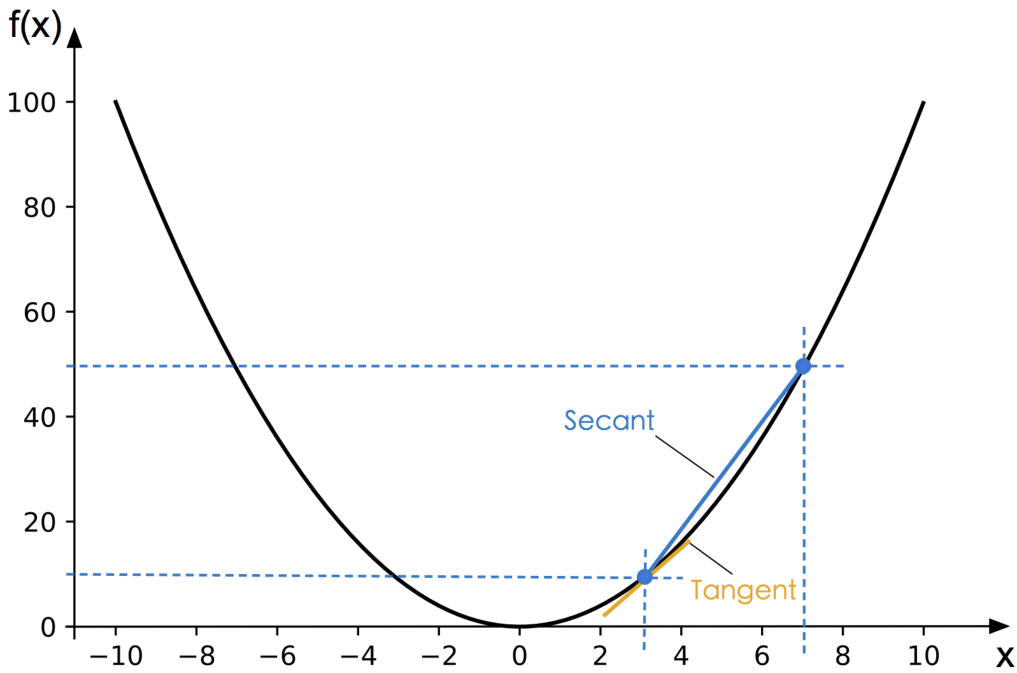
so that

$$\begin{aligned} \frac{df}{dx} &= \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{2x + 2\Delta x - 2x}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{2\Delta x}{\Delta x} \\ &= \lim_{\Delta x \to 0} 2. \end{aligned} \tag{5}$$

[5] 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 $f(x) = x^2$:



As we can see in the figure above, this function 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

$$f(x + \Delta x) = (x + \Delta x)^2 = x^2 + 2x\Delta x + (\Delta x)^2. \tag{6}$$

Now, computing the derivative, we get

$$
\begin{aligned}
\frac{df}{dx} &= \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{x^2 + 2x\Delta x + (\Delta x)^2 - x^2}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{2x\Delta x + (\Delta x)^2}{\Delta x} \\
&= \lim_{\Delta x \to 0} 2x + \Delta x.
\end{aligned} \tag{7}
$$

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

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 – you are encouraged to apply the *basic principles* to derive these rules, however.

The following table 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 (3) and a quadratic function (4) earlier on.

Table D1. Derivatives of common functions.

	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	$\tan(x)$	$\sec^2(x)$

Common Differentiation Rules

In addition to the *constant rule* (Table D1 row 1) and the *power rule* (Table D1 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.

Table D2. Common differentiation rules.

	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	$f(g(x))$	$f'(g(x))g'(x)$

The Chain Rule – Computing the Derivative of a Composition of Functions

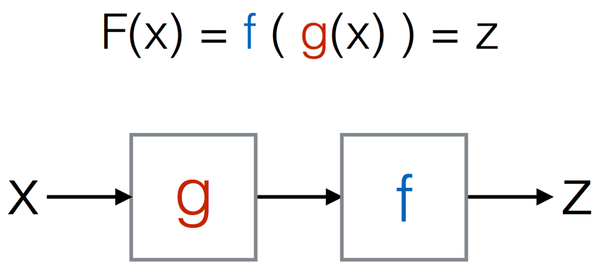
The chain rule is essential to understanding *backpropagation*; thus, let us discuss it in more detail. In its essence, the chain rule is just a *mental crutch* that we use to differentiate composite functions, functions that are nested within each other. For example,

$$F(x) = f(g(x)). \tag{8}$$

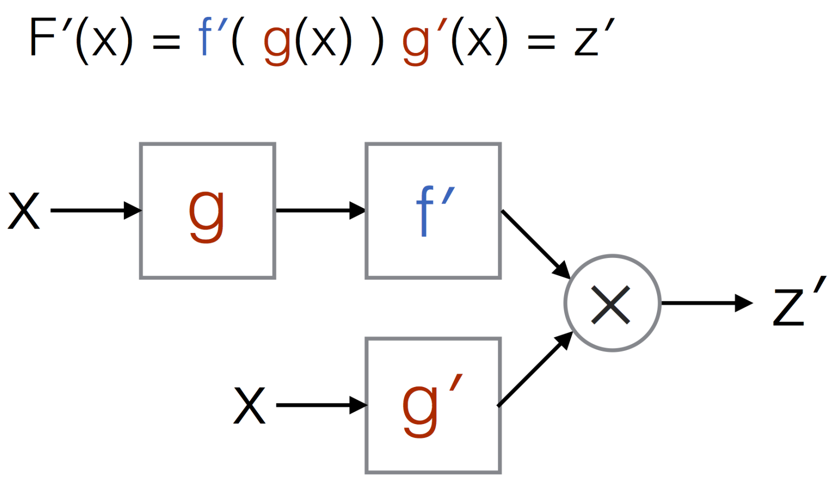
To differentiate such a function F, we can use this *chain rule*, which we can break down to a three-step procedure. First, we compute the derivative of the outer function (f') with respect to the inner function (g). Second, we compute the derivative of the inner function (g') with respect to its function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

$$F'(x) = f'(g(x))g'(x). \tag{9}$$

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



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



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

$$\frac{d}{dx}\left[f(g(x))\right] = \frac{df}{dg} \cdot \frac{dg}{dx}. \tag{10}$$

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

A Chain Rule Example

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

$$f(x) = \log(\sqrt{x}). \tag{11}$$

Step 0: Organization

First, we identify the innermost function:

$$g(x) = \sqrt{x}. \tag{12}$$

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

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

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:

$$\frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}, \tag{14}$$

which lets us arrive at

$$\frac{df}{dx} = \frac{d}{dg}\log(g) \cdot \frac{d}{dx}\sqrt{x}. \tag{15}$$

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:

$$\frac{d}{dg}\log(g) = \frac{1}{g} = \frac{1}{\sqrt{x}}. \tag{16}$$

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

$$g(x) = \sqrt{x} = x^{1/2}. \tag{17}$$

Then, we can use the *power rule* (Table D1 row 5) to arrive at

$$\frac{d}{dx} x^{1/2} = \frac{1}{2} x^{-1/2} = \frac{1}{2\sqrt{x}}. \tag{18}$$

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 $f(x) = \log(\sqrt{x})$:

$$\frac{df}{dx} = \frac{1}{\sqrt{x}} \cdot \frac{1}{2\sqrt{x}} = \frac{1}{2x}. \tag{19}$$

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:

$$F(x) = f(g(h(u(v(x))))). \tag{20}$$

Then, we compute the derivative as

$$\begin{aligned} \frac{dF}{dx} = \frac{d}{dx} F(x) &= \frac{d}{dx} f(g(h(u(v(x))))) \\ &= \frac{df}{dg} \cdot \frac{dg}{dh} \cdot \frac{dh}{du} \cdot \frac{du}{dv} \cdot \frac{dv}{dx}. \end{aligned} \tag{21}$$

As we can see, 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.

Partial Derivatives and Gradients

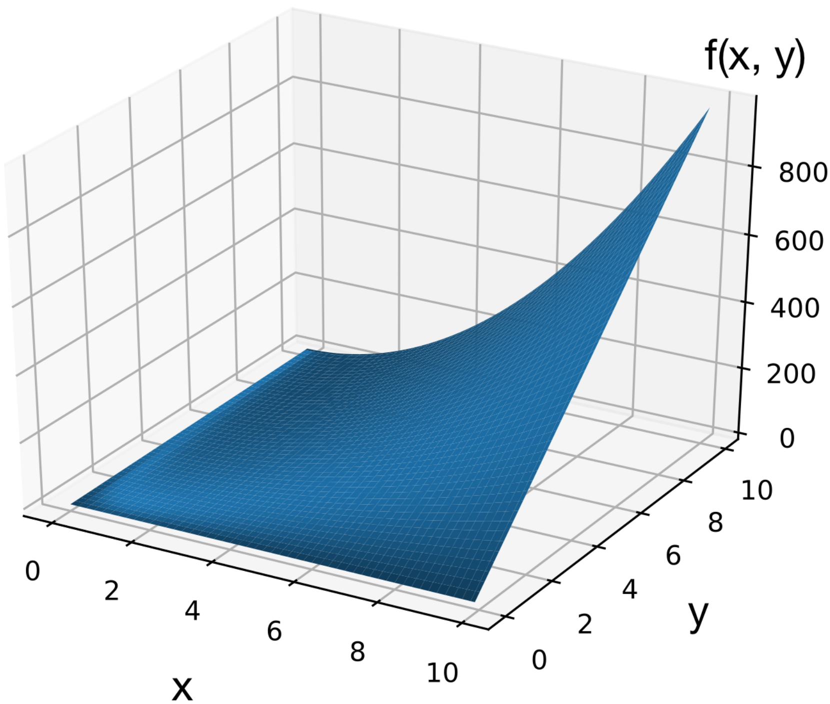
Throughout the previous sections, we only looked at univariate functions, functions that only take one input variable, for example, $f(x)$. In this section, we will compute the derivatives of multivariable functions $f(x, y, z, ...)$. Note that we still consider scalar-valued functions, which return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

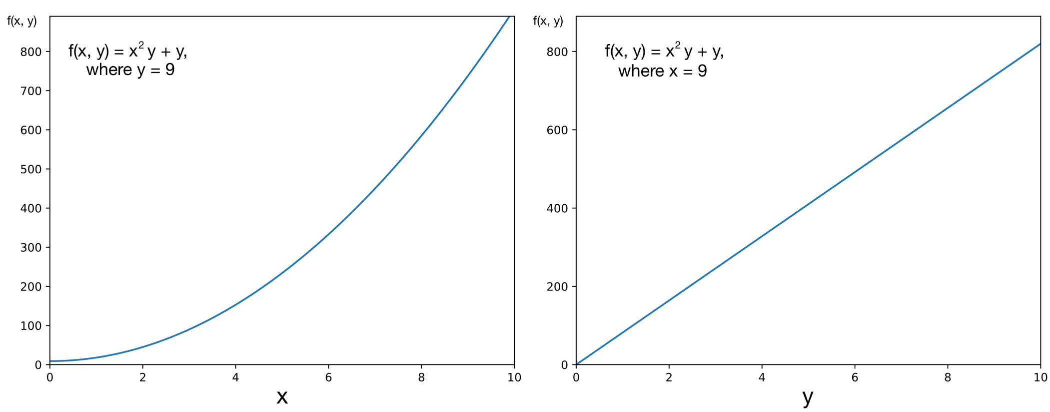
$$\nabla f = \begin{bmatrix} \partial f / \partial x \\ \partial f / \partial y \\ \partial f / \partial z \\ \vdots \end{bmatrix}. \tag{22}$$

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 ∂ 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 this concept, let us walk through a concrete example, where we will compute the gradient of the function $f(x, y) = x^2 y + y$. The following plot shows a graph of this function for different values of x and y:



The following subfigures illustrate how the function looks like if we treat either x or y as a constant:



Intuitively, we can think of the two graphs above as slices of the multivariable function graph. 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:

$$\nabla f(x, y) = \begin{bmatrix} \partial f / \partial x \\ \partial f / \partial y \end{bmatrix}, \tag{23}$$

where

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x} x^2 y + y = 2xy \tag{24}$$

(via the power rule and constant rule), and

$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y} x^2 y + y = x^2 + 1. \tag{25}$$

So, the gradient of the function f is defined as

$$\nabla f(x, y) = \begin{bmatrix} 2xy \\ x^2 + 1 \end{bmatrix}. \tag{26}$$

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 example, we write the second derivative of a function f as

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2}. \tag{27}$$

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

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x} x^2 y + y\right) = \frac{\partial}{\partial x} 2xy = \frac{\partial}{\partial x} = 2y. \tag{28}$$

Given a multivariable function with two arguments, we can in fact compute four distinct second order partial derivatives:

$$\frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial y^2}, \frac{\partial^2 f}{\partial x \partial y}, \text{ and } \frac{\partial^2 f}{\partial y \partial x}, \tag{29}$$

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

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right). \tag{30}$$

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:

$$f\big(g, h\big) = g^2 h + h, \tag{31}$$

where $g(x) = 3x$, and $h(x) = x^2$.

So, as it turns out, our function is a composition of two functions:

$$f\big(g(x), h(x)\big) \tag{32}$$

Remember, we previously defined the chain rule for the univariate case as follows:

$$\frac{d}{dx}\big[f(g(x))\big] = \frac{df}{dg} \cdot \frac{dg}{dx}. \tag{33}$$

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

$$\frac{d}{dx}\big[f(g(x), h(x))\big] = \frac{\partial f}{\partial g} \cdot \frac{dg}{dx} + \frac{\partial f}{\partial h} \cdot \frac{dh}{dx}. \tag{34}$$

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

$$\frac{\partial f}{\partial g} = 2gh \tag{35}$$

and

$$\frac{\partial f}{\partial h} = g^2 + 1. \tag{36}$$

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

$$\frac{dg}{dx} = \frac{d}{dx}3x = 3 \tag{37}$$

$$\frac{dh}{dx} = \frac{d}{dx}x^2 = 2x. \tag{38}$$

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

$$\frac{d}{dx}\big[f(g(x))\big] = [2gh \cdot 3] + [g^2 + 1 + 2x] = g(g + 6h) + 2x + 1. \tag{39}$$

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.

Dot Products

As we remember from the linear algebra appendix, we compute the *dot product* between two vectors as follows: $\begin{bmatrix} a \\ b \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = ax + by$

In vector form, we write the multivariable chain rule

$$\frac{d}{dx}\big[f(g(x), h(x))\big] = \frac{\partial f}{\partial g} \cdot \frac{dg}{dx} + \frac{\partial f}{\partial h} \cdot \frac{dh}{dx}. \tag{40}$$

as follows:

$$\frac{d}{dx}\left[f(g(x), h(x))\right] = \nabla f \cdot \mathbf{v}'(x). \tag{41}$$

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

$$\mathbf{v}(x) = \begin{bmatrix} g(x) \\ h(x) \end{bmatrix}. \tag{42}$$

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

$$\mathbf{v}'(x) = \frac{d}{dx}\begin{bmatrix} g(x) \\ h(x) \end{bmatrix} = \begin{bmatrix} dg/dx \\ dh/dx \end{bmatrix}. \tag{43}$$

So, putting everything together, we have

$$\nabla f \cdot \mathbf{v}'(x) = \begin{bmatrix} \partial f/\partial g \\ \partial f/\partial h \end{bmatrix} \cdot \begin{bmatrix} dg/dx \\ dh/dx \end{bmatrix} = \frac{\partial f}{\partial g} \cdot \frac{dg}{dx} + \frac{\partial f}{\partial h} \cdot \frac{dh}{dx}. \tag{44}$$

The Hessian Matrix

As we mentioned earlier in the section on Second Order Partial Derivatives, we can compute four distinct partial derivatives for a two-variable function:

$$f(x, y). \tag{45}$$

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

$$Hf = \begin{bmatrix} \partial^2 f/\partial x^2 & \partial^2 f/\partial x \partial y \\ \partial^2 f/\partial y \partial x & \partial^2 f/\partial y^2 \end{bmatrix}. \tag{46}$$

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

$$f(x_1, x_2, ..., x_n), \tag{47}$$

we write the Hessian as

$$Hf = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}. \tag{48}$$

The Laplacian Operator

At its core, the Laplacian operator (Δ) 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":

$$\Delta f(g(x), h(x)) = \nabla \cdot \nabla f. \tag{49}$$

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

$$\nabla f(g, h) = \begin{bmatrix} \partial f / \partial g \\ \partial f / \partial h \end{bmatrix}. \tag{50}$$

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

$$\Delta f(g(x), h(x)) = \begin{bmatrix} \partial f / \partial g \\ \partial f / \partial h \end{bmatrix} \cdot \begin{bmatrix} \partial f / \partial g \\ \partial f / \partial h \end{bmatrix} f = \frac{\partial^2 f}{\partial g^2} + \frac{\partial^2 f}{\partial h^2}. \tag{51}$$

And in more general terms, we can define the Laplacian of a function $f(x_1, x_2, ..., x_n)$ as

$$\Delta f = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}. \tag{52}$$