

CHAPTER 2

How the backpropagation algorithm works

In the [last chapter](#) we saw how neural networks can learn their weights and biases using the gradient descent algorithm. There was, however, a gap in our explanation: we didn't discuss how to compute the gradient of the cost function. That's quite a gap! In this chapter I'll explain a fast algorithm for computing such gradients, an algorithm known as *backpropagation*.

The backpropagation algorithm was originally introduced in the 1970s, but its importance wasn't fully appreciated until a [famous](#)

1986 paper by [David Rumelhart](#), [Geoffrey Hinton](#), and [Ronald Williams](#). That paper describes several neural networks where backpropagation works far faster than earlier approaches to learning, making it possible to use neural nets to solve problems which had previously been insoluble. Today, the backpropagation algorithm is the workhorse of learning in neural networks.

This chapter is more mathematically involved than the rest of the book. If you're not crazy about mathematics you may be tempted to skip the chapter, and to treat backpropagation as a black box whose details you're willing to ignore. Why take the time to study those details?

The reason, of course, is understanding. At the heart of backpropagation is an expression for the partial derivative $\partial C / \partial w$ of the cost function C with respect to any weight w (or bias b) in the network. The expression tells us how quickly the cost changes when we change the weights and biases. And while the expression is somewhat complex, it also has a beauty to it, with each element having a natural, intuitive interpretation. And so backpropagation isn't just a fast algorithm for learning. It actually gives us detailed insights into how changing the weights and biases changes the overall behaviour of the network. That's well worth studying in detail.

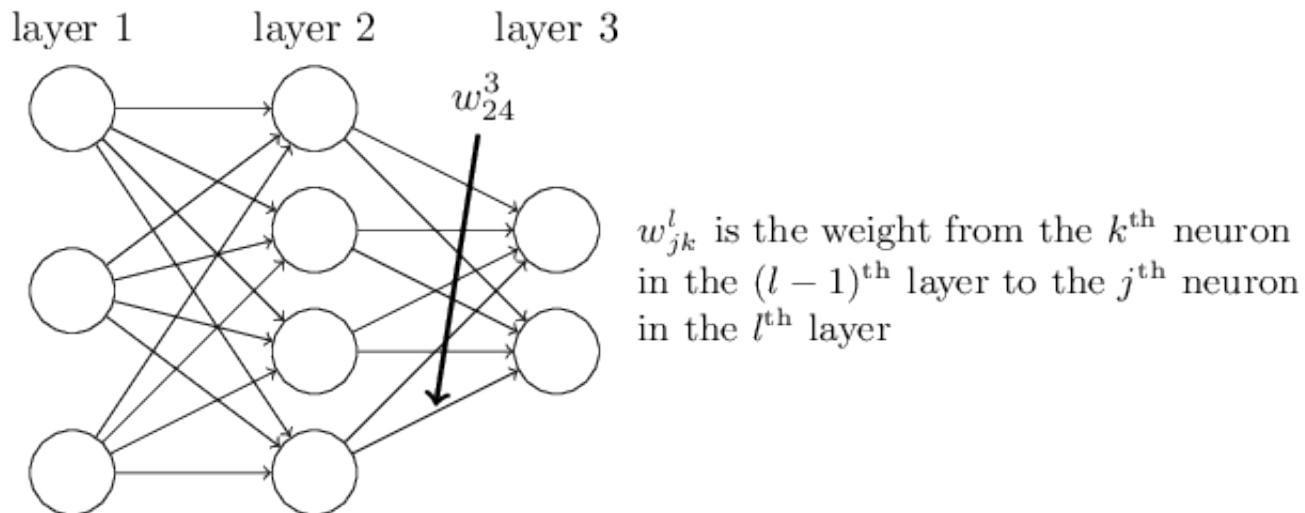
With that said, if you want to skim the chapter, or jump straight to the next chapter, that's fine. I've written the rest of the book to be accessible even if you treat backpropagation as a black box. There are, of course, points later in the book where I refer back to results from this chapter. But at those points you should still be able to understand the main conclusions, even if you don't follow all the reasoning.

Warm up: a fast matrix-based approach to computing the output from a neural network

Before discussing backpropagation, let's warm up with a fast matrix-based algorithm to compute the output from a neural network. We actually already briefly saw this algorithm [near the end of the last chapter](#), but I described it quickly, so it's worth revisiting in detail. In particular, this is a good way of getting comfortable with the notation used in backpropagation, in a familiar context.

Let's begin with a notation which lets us refer to weights in the network in an unambiguous way. We'll use w_{jk}^l to denote the weight for the connection from the k^{th} neuron in the $(l - 1)^{\text{th}}$ layer to the j^{th} neuron in the l^{th} layer. So, for example, the diagram below shows

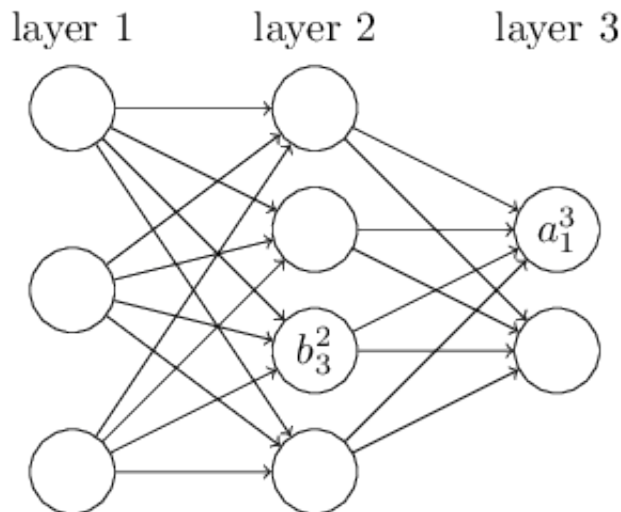
the weight on a connection from the fourth neuron in the second layer to the second neuron in the third layer of a network:



This notation is cumbersome at first, and it does take some work to master. But with a little effort you'll find the notation becomes easy and natural. One quirk of the notation is the ordering of the j and k

indices. You might think that it makes more sense to use j to refer to the input neuron, and k to the output neuron, not vice versa, as is actually done. I'll explain the reason for this quirk below.

We use a similar notation for the network's biases and activations. Explicitly, we use b_j^l for the bias of the j^{th} neuron in the l^{th} layer. And we use a_j^l for the activation of the j^{th} neuron in the l^{th} layer. The following diagram shows examples of these notations in use:



With these notations, the activation a_j^l of the j^{th} neuron in the l^{th} layer is related to the activations in the $(l - 1)^{\text{th}}$ layer by the equation (compare Equation (4) and surrounding discussion in the last chapter)

$$a_j^l = \sigma \left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l \right), \quad (23)$$

where the sum is over all neurons k in the $(l - 1)^{\text{th}}$ layer. To rewrite this expression in a matrix form we define a *weight matrix* w^l for each layer, l . The entries of the weight matrix w^l are just the weights connecting to the l^{th} layer of neurons, that is, the entry in the j^{th} row and k^{th} column is w_{jk}^l . Similarly, for each layer l we define a *bias vector*, b^l . You can probably guess how this works - the components of the bias vector are just the values b_j^l , one component for each neuron in the l^{th} layer. And finally, we define an activation vector a^l whose components are the activations a_j^l .

The last ingredient we need to rewrite (23) in a matrix form is the idea of vectorizing a function such as σ . We met vectorization briefly in the last chapter, but to recap, the idea is that we want to apply a function such as σ to every element in a vector v . We use the obvious notation $\sigma(v)$ to denote this kind of elementwise application of a function. That is, the components of $\sigma(v)$ are just $\sigma(v)_j = \sigma(v_j)$. As an example, if we have the function $f(x) = x^2$ then the vectorized form of f has the effect

$$f\left(\begin{bmatrix} 2 \\ 3 \end{bmatrix}\right) = \begin{bmatrix} f(2) \\ f(3) \end{bmatrix} = \begin{bmatrix} 4 \\ 9 \end{bmatrix}, \quad (24)$$

that is, the vectorized f just squares every element of the vector.

With these notations in mind, Equation (23) can be rewritten in the beautiful and compact vectorized form

$$a^l = \sigma(w^l a^{l-1} + b^l). \quad (25)$$

This expression gives us a much more global way of thinking about how the activations in one layer relate to activations in the previous layer: we just apply the weight matrix to the activations, then add the bias vector, and finally apply the σ function*. That global view is often easier and more succinct (and involves fewer indices!) than the neuron-by-neuron view we've taken to now. Think of it as a way of escaping index hell, while remaining precise about what's going on. The expression is also useful in practice, because most matrix libraries provide fast ways of implementing matrix multiplication,

vector addition, and vectorization. Indeed, the [code](#) in the last chapter made implicit use of this expression to compute the behaviour of the network.

When using Equation [\(25\)](#) to compute a^l , we compute the intermediate quantity $z^l \equiv w^l a^{l-1} + b^l$ along the way. This quantity turns out to be useful enough to be worth naming: we call z^l the *weighted input* to the neurons in layer l . We'll make considerable use of the weighted input z^l later in the chapter. Equation [\(25\)](#) is sometimes written in terms of the weighted input, as $a^l = \sigma(z^l)$. It's also worth noting that z^l has components $z_j^l = \sum_k w_{jk}^l a_k^{l-1} + b_j^l$, that is, z_j^l is just the weighted input to the activation function for neuron j in layer l .

The two assumptions we need about the cost function

The goal of backpropagation is to compute the partial derivatives $\partial C/\partial w$ and $\partial C/\partial b$ of the cost function C with respect to any weight w or bias b in the network. For backpropagation to work we need to make two main assumptions about the form of the cost function. Before stating those assumptions, though, it's useful to have an example cost function in mind. We'll use the quadratic cost function from last chapter (c.f. Equation (6)). In the notation of the last section, the quadratic cost has the form

$$C = \frac{1}{2n} \sum_x \|y(x) - a^L(x)\|^2, \quad (26)$$

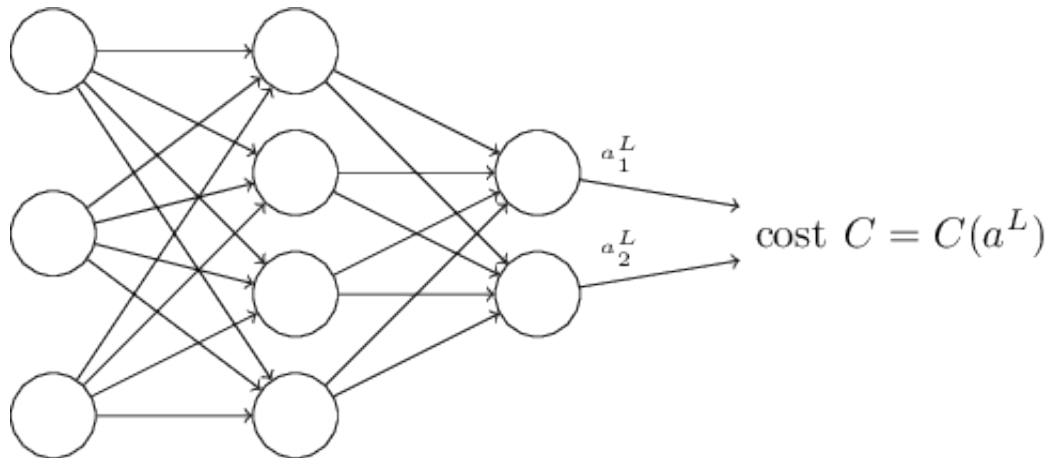
where: n is the total number of training examples; the sum is over individual training examples, x ; $y = y(x)$ is the corresponding desired output; L denotes the number of layers in the network; and $a^L = a^L(x)$ is the vector of activations output from the network when x is input.

Okay, so what assumptions do we need to make about our cost function, C , in order that backpropagation can be applied? The first assumption we need is that the cost function can be written as an average $C = \frac{1}{n} \sum_x C_x$ over cost functions C_x for individual training examples, x . This is the case for the quadratic cost function, where the cost for a single training example is $C_x = \frac{1}{2} \|y - a^L\|^2$. This

assumption will also hold true for all the other cost functions we'll meet in this book.

The reason we need this assumption is because what backpropagation actually lets us do is compute the partial derivatives $\partial C_x / \partial w$ and $\partial C_x / \partial b$ for a single training example. We then recover $\partial C / \partial w$ and $\partial C / \partial b$ by averaging over training examples. In fact, with this assumption in mind, we'll suppose the training example x has been fixed, and drop the x subscript, writing the cost C_x as C . We'll eventually put the x back in, but for now it's a notational nuisance that is better left implicit.

The second assumption we make about the cost is that it can be written as a function of the outputs from the neural network:



For example, the quadratic cost function satisfies this requirement, since the quadratic cost for a single training example x may be written as

$$C = \frac{1}{2} \|y - a^L\|^2 = \frac{1}{2} \sum_j (y_j - a_j^L)^2, \quad (27)$$

and thus is a function of the output activations. Of course, this cost function also depends on the desired output y , and you may wonder why we're not regarding the cost also as a function of y . Remember, though, that the input training example x is fixed, and so the output y is also a fixed parameter. In particular, it's not something we can modify by changing the weights and biases in any way, i.e., it's not something which the neural network learns. And so it makes sense to regard C as a function of the output activations a^L alone, with y merely a parameter that helps define that function.

The Hadamard product, $s \odot t$

The backpropagation algorithm is based on common linear algebraic operations - things like vector addition, multiplying a vector by a matrix, and so on. But one of the operations is a little less commonly used. In particular, suppose s and t are two vectors of the same dimension. Then we use $s \odot t$ to denote the *elementwise* product of the two vectors. Thus the components of $s \odot t$ are just $(s \odot t)_j = s_j t_j$. As an example,

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \odot \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 * 3 \\ 2 * 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 8 \end{bmatrix}. \quad (28)$$

This kind of elementwise multiplication is sometimes called the *Hadamard product* or *Schur product*. We'll refer to it as the Hadamard product. Good matrix libraries usually provide fast

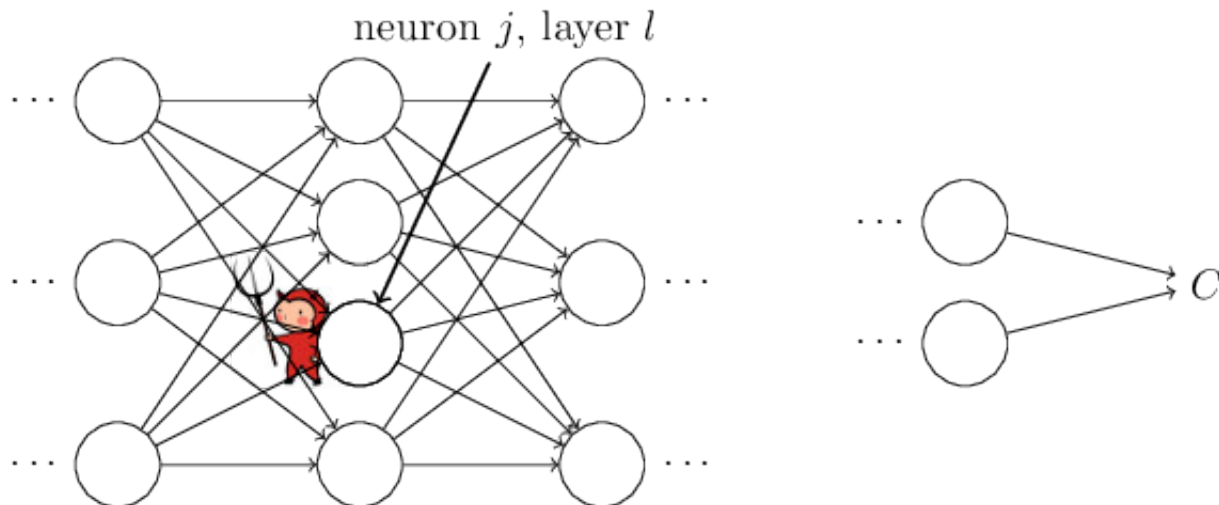
implementations of the Hadamard product, and that comes in handy when implementing backpropagation.

The four fundamental equations behind backpropagation

Backpropagation is about understanding how changing the weights and biases in a network changes the cost function. Ultimately, this means computing the partial derivatives $\partial C / \partial w_{jk}^l$ and $\partial C / \partial b_j^l$. But to compute those, we first introduce an intermediate quantity, δ_j^l , which we call the *error* in the j^{th} neuron in the l^{th} layer.

Backpropagation will give us a procedure to compute the error δ_j^l , and then will relate δ_j^l to $\partial C / \partial w_{jk}^l$ and $\partial C / \partial b_j^l$.

To understand how the error is defined, imagine there is a demon in our neural network:



The demon sits at the j^{th} neuron in layer l . As the input to the neuron comes in, the demon messes with the neuron's operation. It adds a little change Δz_j^l to the neuron's weighted input, so that

instead of outputting $\sigma(z_j^l)$, the neuron instead outputs $\sigma(z_j^l + \Delta z_j^l)$. This change propagates through later layers in the network, finally causing the overall cost to change by an amount $\frac{\partial C}{\partial z_j^l} \Delta z_j^l$.

Now, this demon is a good demon, and is trying to help you improve the cost, i.e., they're trying to find a Δz_j^l which makes the cost smaller. Suppose $\frac{\partial C}{\partial z_j^l}$ has a large value (either positive or negative). Then the demon can lower the cost quite a bit by choosing Δz_j^l to have the opposite sign to $\frac{\partial C}{\partial z_j^l}$. By contrast, if $\frac{\partial C}{\partial z_j^l}$ is close to zero, then the demon can't improve the cost much at all by perturbing the weighted input z_j^l . So far as the demon can tell, the

neuron is already pretty near optimal*. And so there's a heuristic sense in which $\frac{\partial C}{\partial z_j^l}$ is a measure of the error in the neuron.

Motivated by this story, we define the error δ_j^l of neuron j in layer l by

$$\delta_j^l \equiv \frac{\partial C}{\partial z_j^l}. \quad (29)$$

As per our usual conventions, we use δ^l to denote the vector of errors associated with layer l . Backpropagation will give us a way of computing δ^l for every layer, and then relating those errors to the quantities of real interest, $\partial C / \partial w_{jk}^l$ and $\partial C / \partial b_j^l$.

You might wonder why the demon is changing the weighted input z_j^l . Surely it'd be more natural to imagine the demon changing the output activation a_j^l , with the result that we'd be using $\frac{\partial C}{\partial a_j^l}$ as our measure of error. In fact, if you do this things work out quite similarly to the discussion below. But it turns out to make the presentation of backpropagation a little more algebraically complicated. So we'll stick with $\delta_j^l = \frac{\partial C}{\partial z_j^l}$ as our measure of error*.

Plan of attack: Backpropagation is based around four fundamental equations. Together, those equations give us a way of computing both the error δ^l and the gradient of the cost function. I state the four equations below. Be warned, though: you shouldn't expect to instantaneously assimilate the equations. Such an

expectation will lead to disappointment. In fact, the backpropagation equations are so rich that understanding them well requires considerable time and patience as you gradually delve deeper into the equations. The good news is that such patience is repaid many times over. And so the discussion in this section is merely a beginning, helping you on the way to a thorough understanding of the equations.

Here's a preview of the ways we'll delve more deeply into the equations later in the chapter: I'll [give a short proof of the equations](#), which helps explain why they are true; we'll [restate the equations](#) in algorithmic form as pseudocode, and [see how](#) the pseudocode can be implemented as real, running Python code; and,

in [the final section of the chapter](#), we'll develop an intuitive picture of what the backpropagation equations mean, and how someone might discover them from scratch. Along the way we'll return repeatedly to the four fundamental equations, and as you deepen your understanding those equations will come to seem comfortable and, perhaps, even beautiful and natural.

An equation for the error in the output layer, δ^L : The components of δ^L are given by

$$\delta_j^L = \frac{\partial C}{\partial a_j^L} \sigma'(z_j^L). \quad (\text{BP1})$$

This is a very natural expression. The first term on the right, $\partial C / \partial a_j^L$, just measures how fast the cost is changing as a function of the j^{th}

output activation. If, for example, C doesn't depend much on a particular output neuron, j , then δ_j^L will be small, which is what we'd expect. The second term on the right, $\sigma'(z_j^L)$, measures how fast the activation function σ is changing at z_j^L .

Notice that everything in (BP1) is easily computed. In particular, we compute z_j^L while computing the behaviour of the network, and it's only a small additional overhead to compute $\sigma'(z_j^L)$. The exact form of $\partial C / \partial a_j^L$ will, of course, depend on the form of the cost function. However, provided the cost function is known there should be little trouble computing $\partial C / \partial a_j^L$. For example, if we're using the quadratic cost function then $C = \frac{1}{2} \sum_j (y_j - a_j^L)^2$, and so $\partial C / \partial a_j^L = (a_j^L - y_j)$, which obviously is easily computable.

Equation (BP1) is a componentwise expression for δ^L . It's a perfectly good expression, but not the matrix-based form we want for backpropagation. However, it's easy to rewrite the equation in a matrix-based form, as

$$\delta^L = \nabla_a C \odot \sigma'(z^L). \quad (\text{BP1a})$$

Here, $\nabla_a C$ is defined to be a vector whose components are the partial derivatives $\partial C / \partial a_j^L$. You can think of $\nabla_a C$ as expressing the rate of change of C with respect to the output activations. It's easy to see that Equations (BP1a) and (BP1) are equivalent, and for that reason from now on we'll use (BP1) interchangeably to refer to both equations. As an example, in the case of the quadratic cost we have

$\nabla_a C = (a^L - y)$, and so the fully matrix-based form of (BP1) becomes

$$\delta^L = (a^L - y) \odot \sigma'(z^L). \quad (30)$$

As you can see, everything in this expression has a nice vector form, and is easily computed using a library such as Numpy.

An equation for the error δ^l in terms of the error in the next layer, δ^{l+1} : In particular

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l), \quad (\text{BP2})$$

where $(w^{l+1})^T$ is the transpose of the weight matrix w^{l+1} for the $(l + 1)^{\text{th}}$ layer. This equation appears complicated, but each element

has a nice interpretation. Suppose we know the error δ^{l+1} at the $l + 1^{\text{th}}$ layer. When we apply the transpose weight matrix, $(w^{l+1})^T$, we can think intuitively of this as moving the error *backward* through the network, giving us some sort of measure of the error at the output of the l^{th} layer. We then take the Hadamard product $\odot \sigma'(z^l)$. This moves the error backward through the activation function in layer l , giving us the error δ^l in the weighted input to layer l .

By combining (BP2) with (BP1) we can compute the error δ^l for any layer in the network. We start by using (BP1) to compute δ^L , then apply Equation (BP2) to compute δ^{L-1} , then Equation (BP2) again to compute δ^{L-2} , and so on, all the way back through the network.

An equation for the rate of change of the cost with respect to any bias in the network: In particular:

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l. \quad (\text{BP3})$$

That is, the error δ_j^l is *exactly equal* to the rate of change $\partial C / \partial b_j^l$.

This is great news, since (BP1) and (BP2) have already told us how to compute δ_j^l . We can rewrite (BP3) in shorthand as

$$\frac{\partial C}{\partial b} = \delta, \quad (31)$$

where it is understood that δ is being evaluated at the same neuron as the bias b .

An equation for the rate of change of the cost with respect to any weight in the network: In particular:

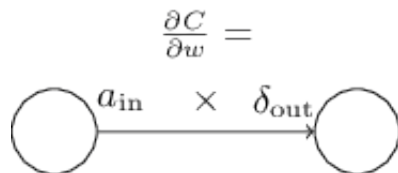
$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l. \quad (\text{BP4})$$

This tells us how to compute the partial derivatives $\partial C / \partial w_{jk}^l$ in terms of the quantities δ^l and a^{l-1} , which we already know how to compute. The equation can be rewritten in a less index-heavy notation as

$$\frac{\partial C}{\partial w} = a_{\text{in}} \delta_{\text{out}}, \quad (32)$$

where it's understood that a_{in} is the activation of the neuron input to the weight w , and δ_{out} is the error of the neuron output from the

weight w . Zooming in to look at just the weight w , and the two neurons connected by that weight, we can depict this as:



A nice consequence of Equation (32) is that when the activation a_{in} is small, $a_{\text{in}} \approx 0$, the gradient term $\partial C / \partial w$ will also tend to be small. In this case, we'll say the weight *learns slowly*, meaning that it's not changing much during gradient descent. In other words, one consequence of (BP4) is that weights output from low-activation neurons learn slowly.

There are other insights along these lines which can be obtained from (BP1)-(BP4). Let's start by looking at the output layer. Consider the term $\sigma'(z_j^L)$ in (BP1). Recall from the [graph of the sigmoid function in the last chapter](#) that the σ function becomes very flat when $\sigma(z_j^L)$ is approximately 0 or 1. When this occurs we will have $\sigma'(z_j^L) \approx 0$. And so the lesson is that a weight in the final layer will learn slowly if the output neuron is either low activation (≈ 0) or high activation (≈ 1). In this case it's common to say the output neuron has *saturated* and, as a result, the weight has stopped learning (or is learning slowly). Similar remarks hold also for the biases of output neuron.

We can obtain similar insights for earlier layers. In particular, note the $\sigma'(z^l)$ term in (BP2). This means that δ_j^l is likely to get small if the neuron is near saturation. And this, in turn, means that any weights input to a saturated neuron will learn slowly*.

Summing up, we've learnt that a weight will learn slowly if either the input neuron is low-activation, or if the output neuron has saturated, i.e., is either high- or low-activation.

None of these observations is too greatly surprising. Still, they help improve our mental model of what's going on as a neural network learns. Furthermore, we can turn this type of reasoning around. The four fundamental equations turn out to hold for any activation function, not just the standard sigmoid function (that's because, as

we'll see in a moment, the proofs don't use any special properties of σ). And so we can use these equations to *design* activation functions which have particular desired learning properties. As an example to give you the idea, suppose we were to choose a (non-sigmoid) activation function σ so that σ' is always positive, and never gets close to zero. That would prevent the slow-down of learning that occurs when ordinary sigmoid neurons saturate. Later in the book we'll see examples where this kind of modification is made to the activation function. Keeping the four equations (BP1)-(BP4) in mind can help explain why such modifications are tried, and what impact they can have.

Summary: the equations of backpropagation

$$\delta^L = \nabla_a C \odot \sigma'(z^L) \quad (\text{BP1})$$

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l) \quad (\text{BP2})$$

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l \quad (\text{BP3})$$

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l \quad (\text{BP4})$$

Problem

- **Alternate presentation of the equations of backpropagation:** I've stated the equations of

backpropagation (notably (BP1) and (BP2)) using the Hadamard product. This presentation may be disconcerting if you're unused to the Hadamard product. There's an alternative approach, based on conventional matrix multiplication, which some readers may find enlightening. (1) Show that (BP1) may be rewritten as

$$\delta^L = \Sigma'(z^L) \nabla_a C, \quad (33)$$

where $\Sigma'(z^L)$ is a square matrix whose diagonal entries are the values $\sigma'(z_j^L)$, and whose off-diagonal entries are zero. Note that this matrix acts on $\nabla_a C$ by conventional matrix multiplication. (2) Show that (BP2) may be rewritten as

$$\delta^l = \Sigma'(z^l) (w^{l+1})^T \delta^{l+1}. \quad (34)$$

(3) By combining observations (1) and (2) show that

$$\delta^l = \Sigma'(z^l)(w^{l+1})^T \dots \Sigma'(z^{L-1})(w^L)^T \Sigma'(z^L) \nabla_a C \quad (35)$$

For readers comfortable with matrix multiplication this equation may be easier to understand than (BP1) and (BP2). The reason I've focused on (BP1) and (BP2) is because that approach turns out to be faster to implement numerically.

Proof of the four fundamental equations (optional)

We'll now prove the four fundamental equations (BP1)-(BP4). All four are consequences of the chain rule from multivariable calculus.

If you're comfortable with the chain rule, then I strongly encourage you to attempt the derivation yourself before reading on.

Let's begin with Equation (BP1), which gives an expression for the output error, δ^L . To prove this equation, recall that by definition

$$\delta_j^L = \frac{\partial C}{\partial z_j^L}. \quad (36)$$

Applying the chain rule, we can re-express the partial derivative above in terms of partial derivatives with respect to the output activations,

$$\delta_j^L = \sum_k \frac{\partial C}{\partial a_k^L} \frac{\partial a_k^L}{\partial z_j^L}, \quad (37)$$

where the sum is over all neurons k in the output layer. Of course, the output activation a_k^L of the k^{th} neuron depends only on the weighted input z_j^L for the j^{th} neuron when $k = j$. And so $\partial a_k^L / \partial z_j^L$ vanishes when $k \neq j$. As a result we can simplify the previous equation to

$$\delta_j^L = \frac{\partial C}{\partial a_j^L} \frac{\partial a_j^L}{\partial z_j^L}. \quad (38)$$

Recalling that $a_j^L = \sigma(z_j^L)$ the second term on the right can be written as $\sigma'(z_j^L)$, and the equation becomes

$$\delta_j^L = \frac{\partial C}{\partial a_j^L} \sigma'(z_j^L), \quad (39)$$

which is just (BP1), in component form.

Next, we'll prove (BP2), which gives an equation for the error δ^l in terms of the error in the next layer, δ^{l+1} . To do this, we want to rewrite $\delta_j^l = \partial C / \partial z_j^l$ in terms of $\delta_k^{l+1} = \partial C / \partial z_k^{l+1}$. We can do this using the chain rule,

$$\delta_j^l = \frac{\partial C}{\partial z_j^l} \tag{40}$$

$$= \sum_k \frac{\partial C}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial z_j^l} \tag{41}$$

$$= \sum_k \frac{\partial z_k^{l+1}}{\partial z_j^l} \delta_k^{l+1}, \tag{42}$$

where in the last line we have interchanged the two terms on the right-hand side, and substituted the definition of δ_k^{l+1} . To evaluate the first term on the last line, note that

$$z_k^{l+1} = \sum_j w_{kj}^{l+1} a_j^l + b_k^{l+1} = \sum_j w_{kj}^{l+1} \sigma(z_j^l) + b_k^{l+1}. \quad (43)$$

Differentiating, we obtain

$$\frac{\partial z_k^{l+1}}{\partial z_j^l} = w_{kj}^{l+1} \sigma'(z_j^l). \quad (44)$$

Substituting back into (42) we obtain

$$\delta_j^l = \sum_k w_{kj}^{l+1} \delta_k^{l+1} \sigma'(z_j^l). \quad (45)$$