In Lecture 2 we needed the gradient $\nabla f = (\partial f/\partial x, \partial f/\partial y, \partial f/\partial z)$ for the sum-of-squares loss function

$$f(x,y,z) = (x-3y-3z+5)^2 + (x-2y-2z)^2 + (x+y-5z-3)^2. \tag{1}$$

in order to run gradient descent. You have probably spent a great deal of time in calculus class doing calculations like that one. The main message of differential calculus is that no matter how complicated a function looks, its derivatives can be obtained by following simple rules (sum rule, product rule, chain rule) and applying a handful of substitutions when the function cannot be simplified further ($de^x/dx = e^x$, $d\sin x/dx = \cos x$, etc.)

If we were to program a computer to calculate gradients, how should we go about it? Experience suggests that a "top-down" approach tends to work. To calculate $\partial f/\partial x$ in (1), we would start with the sum rule to break up the task into three simpler calculations:

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(x-3y-3z+5)^2 + \frac{\partial}{\partial x}(x-2y-2z)^2 + \frac{\partial}{\partial x}(x+y-5z-3)^2. \tag{2}$$

The first term can be handled like this:

$$\begin{aligned}
\frac{\partial}{\partial x}(x-3y-3z+5)^2 &= 2(x-3y-3z+5)\cdot\frac{\partial}{\partial x}(x-3y-3z+5) && \text{chain rule}\\
&= 2(x-3y-3z+5)\cdot\Big(\frac{\partial}{\partial x}x + \frac{\partial}{\partial x}(-3y-3z+5)\Big) && \text{sum rule again}\\
&= 2(x-3y-3z+5)\cdot(1+0) && \text{substitutions}\\
&= 2(x-3y-3z+5).
\end{aligned}$$

The substitutions we used is $\partial x/\partial x = 1$, $\partial(\text{expression that does not contain } x)/\partial x = 0$. By a similar chain of reasoning, $\partial(x-2y-2z)^2/\partial x = 2(x-2y-2z)$ and $\partial(x+y-5z-3)^2/\partial x = 2(x+y-5z-3)$. Plugging all back into (2) we find

$$\frac{\partial f}{\partial x} = 2(x-3y-3z+5) + 2(x-2y-2z) + 2(x+y-5z-3). \tag{3}$$

Repeating the same logic we also find

$$\begin{aligned}
\frac{\partial f}{\partial y} &= -6(x-3y-3z+5) + -4(x-2y-2z) + 2(x+y-5z-3),\\
\frac{\partial f}{\partial z} &= -6(x-3y-3z+5) + -4(x-2y-2z) + -10(x+y-5z-3).
\end{aligned} \tag{4}$$

This algorithm is called *forward propagation*. The gradient of a complex function is obtained by "propagating" the gradients of its constitutent parts forward.

Algorithm: FP (Forward Propagation for polynomials)
Input: A polynomial $f(x,y,\ldots,z)$.
1 If f is a constant, output $\nabla f = (0,0,\ldots,0)$.
2 If f is a variable, say x, output $\nabla f = (1,0,\ldots,0)$.
3 If f is of the form $g+h$, output $FP(g) + FP(h)$ (sum rule).
4 If f is of the form $g \times h$, output $g \times FP(h) + h \times FP(g)$ (product rule).

To understand Forward Propagation we need to talk about the representation of functions like (1) and (2). In Lecture 4 we saw two representations of functions: the list representation and the Fourier representation. Neither of them tells us how to compute f or its derivatives.

1 Circuits

In computer science it is often useful to represent expressions like (1) and (3) not as equations but as labeled graphs called *circuits*. The circuit representation of f in (1) is shown in Figure 1.

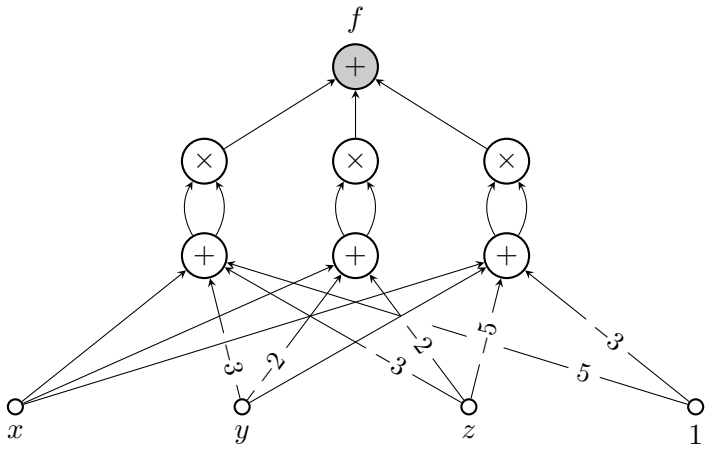
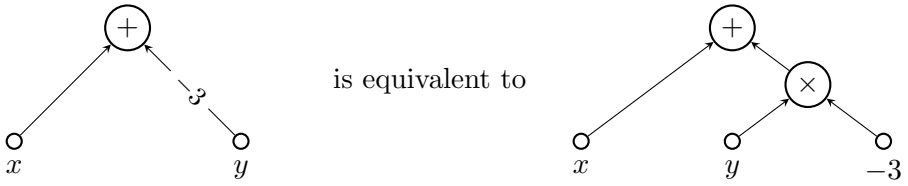


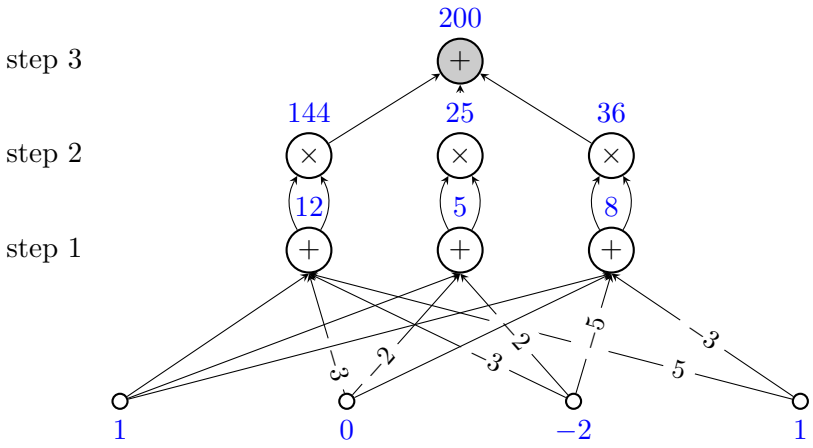
FIGURE 1: The circuit representation of (1).

In general, a circuit is a directed acyclic graph in which every source vertex is labeled by an input variable or a constant. Every other vertex—called a *gate*—is labeled by a basic operation *op*. You may think of these basic operations as instructions that can be executed in hardware at unit cost. The basic operations of an *arithmetic circuit* are plus and times. Richer circuit models may allow others like divide and exponentiate.

In this example some of the edges incident to a + gate are labeled by constants describing a scaling factor. If they bother you, you can think of these factors as "syntactic sugar" for multiplication by a constant. For example:



The primary measure of computational complexity for a circuit is its number of gates. If each gate takes one time unit to evaluate then the size of the circuit is the time it takes to do the computation sequentially. The circuit in Figure 1 has size 7. A secondary complexity measure is the *depth* of the circuit. This is the length of its longest directed path. It is the minimum time it takes to evaluate the circuit in parallel, assuming each gate takes unit time. The circuit in Figure 1 has depth 3. Here is an example evaluation:



2 Analysis of Forward Propagation

Forward propagation is friendly to circuit representations. Given a circuit computing f as input, it produces a circuit for its partial derivatives as output. The portion of the circuit that computes $\partial f/\partial x$ is shown in Figure 2.

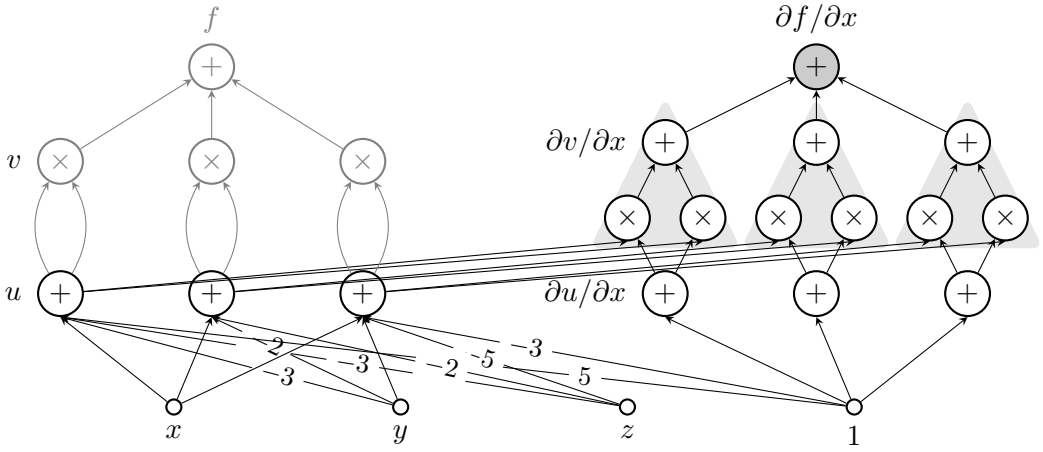


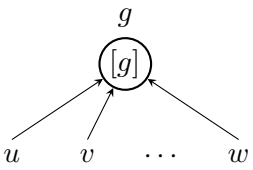
FIGURE 2: A portion of circuit produced by Forward Propagation on input (1).

The sub-circuits computing $\partial f/\partial y$ and $\partial f/\partial z$ have the same structure. The only differences are the factors on the wires coming out of the 1 node.

The overall size of the circuit for ∇f in Figure 2 is $3 + 3 \cdot 13 = 42$. The part of the circuit computing f has 3 gates and there are 13 additional gates participating in the computation of each partial derivative. In general, if $f(x_1, \ldots, x_n)$ has circuit size s, Forward Propagation produces a circuit for ∇f of size $\Theta(ns)$.

To understand Forward Propagation better I find it helpful to work with a general variant that works for circuits with arbitrary gates, not only plus and times. To explain it we'll need a bit of notation. In a circuit C, each gate computes a function whose inputs are the functions computed at its children. We will denote by g both the node in the circuit graph and the function computed by it. Let $[g]$ stand for the type of gate at node g. For example, in the circuit in Figure 1, $[f]$ is the sum of its three inputs.

Given a gate operation $[g]$ with d inputs and one output, its gradient $[\nabla g]$ can be viewed as some other gate operation with (at most) d inputs and d outputs. For instance, if $[op](u, v, w) = u + v + w$ then $[\nabla g](u, v, w) = (1, 1, 1)$, i.e., three copies of the constant 1. If $[g](u, v, w) = uvw$, then $[\nabla g](u, v, w) = (vw, wu, uv)$. If $[g](u) = \exp u$ then $[\nabla g](u) = \exp u$ and if $[g](u) = 1/u$ then $[\nabla g](u) = -1/u^2$.



computes the function $g = [g](u, v, \ldots, w)$.

The partial derivatives of g with respect to the input variables $x, y, \ldots, z$ can be computed from the partial derivatives of u, v, w using the *chain rule*. For example, if $[g]$ is a multiplication gate with three inputs then

$$\frac{\partial g}{\partial x} = \frac{\partial u}{\partial x} \cdot (vw) + \frac{\partial v}{\partial x} \cdot (wu) + \frac{\partial w}{\partial x} \cdot (uv).$$

This is the sum of products between the partial derivatives of the functions represented by the children of g and the gradient of g evaluated at those children:

$$\begin{aligned} \frac{\partial g}{\partial x} &= \left(\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}, \ldots, \frac{\partial w}{\partial x}\right) \cdot [\nabla g](u, v, \ldots, w) \\ &= \frac{\partial u}{\partial x} \cdot \partial_u[g] + \frac{\partial v}{\partial x} \cdot \partial_v[g] + \cdots + \frac{\partial w}{\partial x} \cdot \partial_w[g] \end{aligned} \tag{5}$$

Here, $\partial_v[g]$ stands for the v-th component of the gradient gate $[\nabla g]$. This is the partial derivative of the gate with respect to argument v:

$$[\nabla g](u, v, \ldots, w) = (\partial_u[g], \partial_v[g], \ldots, \partial_w[g]).$$

The chain rule (5) is valid for any gate. The sum and product rules are special cases of it. Forward Propagation applies it systematically to transform a circuit C into its gradient ∇C.

Algorithm: *FP* (General Forward Propagation)
Input: A circuit C with n inputs.
1 For every gate g in C with children $u, v, \ldots, w$,
2 Create a new gate ∇g.
3 For every input variable x,
4 Create a new node $\partial g/\partial x$.
5 Connect $\partial u/\partial x, \partial v/\partial x, \ldots, \partial w/\partial x$ and ∇g to $\partial g/\partial x$ using the chain rule (5).
6 Output the resulting circuit ∇C.

We can apply the chain rule inductively to conclude that Forward Propagation correctly calculates all the partial derivatives.

Theorem 1. *Each gate $\partial g/\partial x$ in ∇C computes the partial derivative of g with respect to input variable x.*

How large is the resulting circuit? In step 4 each gate g of C is augmented with $\Theta(n)$ new gates, one for every partial derivative. In step 5 each wire (edge) of C contributes an extra times gate to each of these partial derivatives. If C has s gates and w wires then ∇C will have $\Theta(n \cdot (s + w))$ gates in asymptotic notation.

Can we do better? As an example, the output of Forward Propagation for C in Figure 1 is quite a bit larger than C: ∇C has size 42, while C has only size 7. In this example a much smaller circuit for the gradient of the same function exists. It is shown in Figure 3.

I constructed this circuit "by hand" after inspecting (3) and (4). The partial derivatives share many common terms that can be reused in the circuit for ∇f. Can this construction procedure be automated for general f?

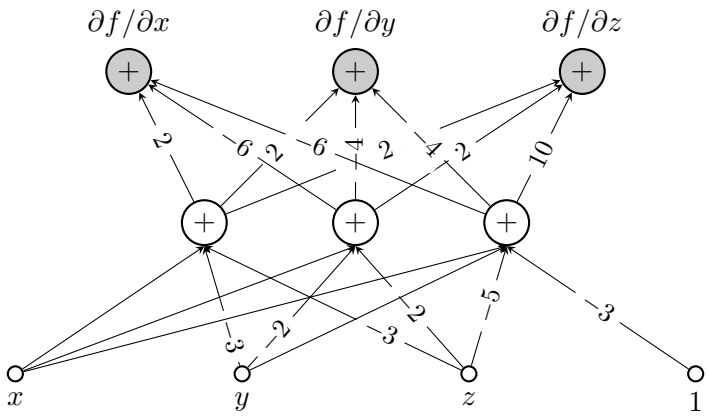


FIGURE 3: A small circuit for the gradient of (1).

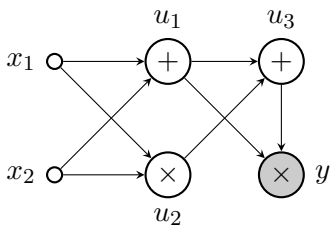
3 Backpropagation

Backpropagation is an ingenious alternative algorithm for calculating the gradient of a circuit C. If C has n inputs, s nodes, and w wires, Backpropagation produces a circuit for ∇C of size $O(n+s+w)$. This is a big improvement over Forward Propagation for circuits with many inputs.

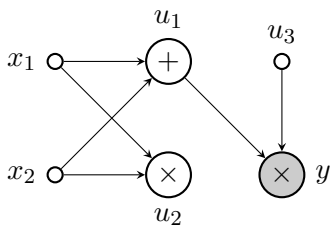
The backpropagation algorithm calculates partial derivatives by visiting the vertices in reverse topological order, starting from the output y and ending with the sources $x_1, \ldots, x_n$. Upon visiting node z, backpropagation constructs a new gate that calculates the partial derivative $\partial y/\partial z$. To explain what this means we need to define the partial derivative of one circuit gate y with respect to another gate z.

Definition 2. Assume C is a circuit, y is a gate, and z is a vertex (input or gate). Let $(C - z)$ be the circuit obtained by removing all edges that point to z from C and turning z into an input (if it is not already one). Then y computes some function $f(z, \text{other inputs})$ in $(C - z)$. The partial derivative $\partial y/\partial z$ is the derivative of f with respect to z.

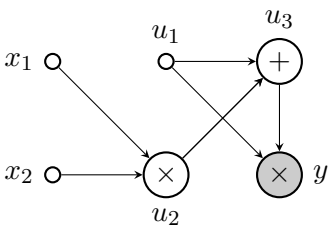
Let's work out an example. Consider the following circuit that computes the function $y(x_1, x_2) = (x_1 + x_2 + x_1x_2)(x_1 + x_2)$. The ordering $x_1, x_2, u_1, u_2, u_3, y$ is a topological sort of its vertices. To its right are the circuits $(C - u_3)$ and $(C - u_1)$.



circuit C



circuit $(C - u_3)$



circuit $(C - u_1)$

The circuit $(C - u_3)$ has inputs x_1, x_2, u_3. The function computed by y in $(C - u_3)$ is $u_1 \cdot u_3 = (x_1 + x_2) \cdot u_3$. The partial derivative $\partial y/\partial u_3$ is then $\partial(u_1 \cdot u_3)/\partial u_3 = u_1 = x_1 + x_2$. The last simplification was possible because u_1 does not depend on u_3 (as it precedes it in the topological sort).

The circuit $(C - u_1)$ has inputs x_1, x_2, u_1. The function computed by y in $(C - u_1)$ is $u_1 \cdot u_3^* = u_1 \cdot (u_1 + u_2) = u_1 \cdot (u_1 + x_1 + x_2)$. Here u_3^* stands for the function computed by u_3 in $(C - u_1)$: This is not the same as the function computed by u_3 in C! The input u_1 appears twice in this expression, once explicitly as an argument of the product gate y and once implicitly via u_3^* which itself depends on u_1. We can calculate the partial derivative of this function using the product rule for derivatives:

$$\frac{\partial y}{\partial u_1} = \frac{\partial(u_1 \cdot u_3^*)}{\partial u_1} = \frac{\partial u_1}{\partial u_1} \cdot u_3^* + u_1 \cdot \frac{\partial u_3^*}{\partial u_1} = 1 \cdot (u_1 + u_2) + u_1 \cdot \frac{\partial(u_1 + u_2)}{\partial u_1} = (u_1 + u_2) + u_1 = 2u_1 + u_2.$$

This expresses the desired partial derivative as a small circuit in terms of the gates u_1 and u_2, which are already present in C. We can compute $\partial y/\partial u_3$ from the existing gates for u_1 and u_2 and a bit of extra work (one addition and one multiplication). Is there a general method for this?

Let us first rework the expression for $\partial y/\partial u_1$ in a more systematic way. Recall that for a gate g taking inputs a, b, $\partial_a[g]$ is the partial derivative of the gate operation $[g]$ with respect to argument a. For example, for a product gate $[g](a, b) = a \cdot b$ we have $\partial_a[g] = b$. In general, $\partial_a[g]$ is not the same as $\partial g/\partial a$. In the circuit $(C - u_1)$, $\partial_{u_1}[y] = \partial(u_1 u_3^*)/\partial u_1 = u_3^*$, while $\partial y/\partial u_1 = u_3^* + u_1$. The reason is that u_1 affects y partially through the edge (u_1, y) and partially through the path (u_1, u_3, y). In this example we can represent y as $y = [y]([u_3](u_1, u_2), u_1)$. By the chain rule,

$$\frac{\partial y}{\partial u_1} = \frac{\partial y}{\partial u_3} \cdot \partial_{u_1}[u_3] + \partial_{u_1}[y].$$

As we already calculated, $\partial y/\partial u_3 = u_1$. As for the other terms, u_3 is a sum gate so $\partial_{u_1}[u_3] = 1$ and u_1 is a product gate so $\partial_{u_1}[y] = u_3 = u_1 + u_2$. We obtain again $\partial y/\partial u_1 = 2u_1 + u_2$.

In general, suppose we have a gate g whose out-edges point to gates $u, v, \ldots, w$. In the circuit $(C - g)$, the output y depends on g via the gates $u, v, \ldots, w$ (one of which could be y itself), each of which depends on g. The chain rule now expresses $\partial y/\partial g$ as a small circuit that depends on partial derivatives of y with respect to gates that succeed y in the topological sort and some derivatives of the gates:

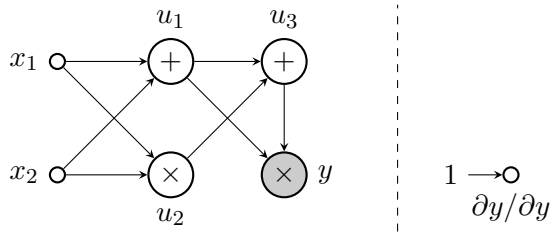
$$\frac{\partial y}{\partial g} = \frac{\partial y}{\partial u} \cdot \partial_g[u] + \frac{\partial y}{\partial v} \cdot \partial_g[v] + \cdots + \frac{\partial y}{\partial w} \cdot \partial_g[w]. \tag{6}$$

suggesting the following iterative algorithm for partial derivatives.

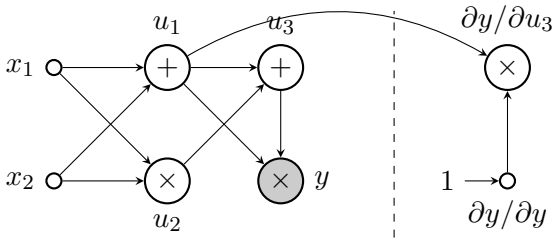
Algorithm *BP* (Backpropagation)
Input: A circuit C.
1 Compute a reverse topological sort s of C's vertices starting with y.
2 For every g in s, construct a subcircuit for $\partial y/\partial g$ using the chain rule (6).
3 Output the resulting circuit ∇C.

We demonstrate an execution of backpropagation on our example circuit C. It is useful to visualize the circuit ∇C by placing the gate $\partial y/\partial z$ in a position mirroring the gate z.

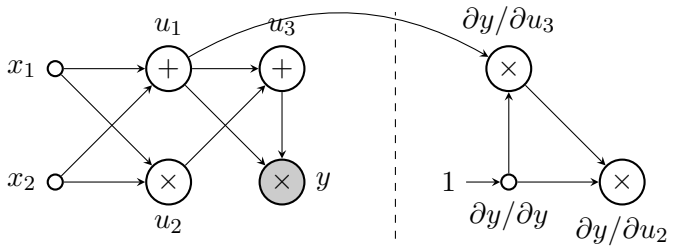
Our reverse topological sort s is the ordering $y, u_3, u_2, u_1, x_2, x_1$. The first vertex y is isolated in the circuit $(C - y)$, so $\partial y/\partial y = 1$.



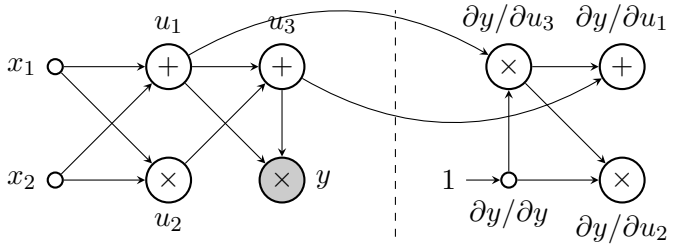
Next in s is u_3. As u_3 has a single out-edge going into y, (6) gives $\partial y/\partial u_3 = \partial y/\partial y \cdot \partial_{u_3}[y] = \partial y/\partial y \cdot u_1$:



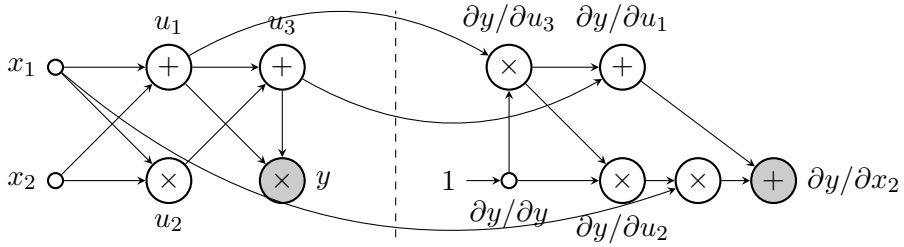
Next is u_2, which has a single edge pointing to u_3. From (6) we get $\partial y/\partial u_2 = \partial y/\partial u_3 \cdot \partial_{u_2}[u_3] = \partial y/\partial u_3 \cdot 1$.



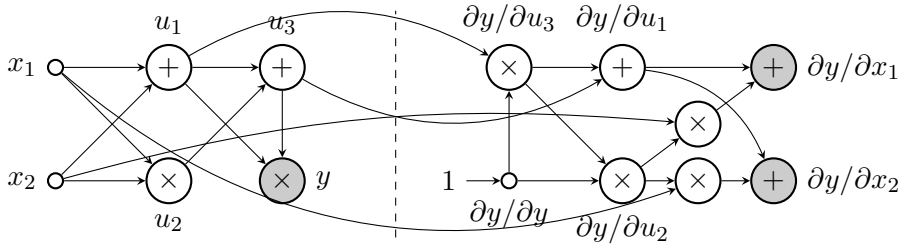
Next is u_1, which has outgoing edges to u_3 and y. Using (6) we get $\partial y/\partial u_1 = \partial y/\partial u_3 \cdot \partial_{u_1}[u_3] + \partial y/\partial y \cdot \partial_{u_1}[y] = \partial y/\partial u_3 \cdot 1 + \partial y/\partial y \cdot u_3$. To keep the picture simple we'll use $\partial y/\partial y = 1$ and omit the multiplicaton gates.



The second to last vertex is x_2, with edges pointing to u_1 and u_2. We have $\partial y/\partial x_2 = \partial y/\partial u_1 \cdot \partial_{x_2}[u_1] + \partial y/\partial u_2 \cdot \partial_{x_2}[u_2] = \partial y/\partial u_1 + \partial y/\partial u_2 \cdot x_1$.



Finally we construct the gate $\partial y/\partial x_1$ using (6) one last time.



At this point it is a good idea to check that this circuit produces correct outputs:

$$\begin{aligned}
\frac{\partial y}{\partial x_1} &= \frac{\partial y}{\partial u_2} \cdot x_2 + \frac{\partial y}{\partial u_1} \\
&= \frac{\partial y}{\partial u_3} \cdot x_2 + \left(u_3 + \frac{\partial y}{\partial u_3} \right) \\
&= u_1 x_2 + (u_3 + u_1) \\
&= (x_1 + x_2) x_2 + (2u_1 + u_2) \\
&= (x_1 + x_2) x_2 + (2x_1 + 2x_2 + x_1 x_2) \\
&= 2x_1 + 2x_2 + 2x_1 x_2 + x_2^2.
\end{aligned}$$

You can verify that this is indeed the partial derivative of $(x_1 + x_2 + x_1x_2)(x_1 + x_2)$ with respect to x_1.

We can now state the correctness of the backpropagation algorithm. Given a set of basic operations B, let ∇B be the set of all partial derivatives of all operations in B.

Theorem 3. *For any circuit C with designated output y and operations coming from B, the output ∇C of Backpropagation is a circuit with operation set $B \cup \nabla B \cup \{+, \times\}$ that contains gates computing $\partial y/\partial u$ for every node u of C. The number of gates in ∇C is at most three times the number of wires plus the number of nodes in C.*

The proof is by strong induction with respect to the ordering s. The gate count comes from formula (6): Each edge (g, u) in G contributes one gate $\partial_u[g]$ in ∇C, one multipilication $\partial y/\partial u \cdot \partial_u[g]$, and at most one addition in the chain rule.

Running backpropagation on the sum-of-squares loss $C(\mathbf{x}) = \|A\mathbf{x} - b\|^2$ produces the circuit $\nabla C(\mathbf{x}) = 2A^T(A\mathbf{x} - b)$, provided the product gates in C are replaced with square gates ($[g](x) = x^2, [\nabla g](x) = 2x$). As an exercise I recommend trying it on the circuit in Figure 1. The output of backpropagation will be the circuit in Figure 3.

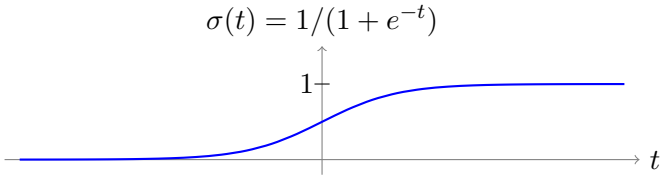
4 Neural networks

When computers took off in the 1950s and 60s scientists began speculating that they can be used to shed light on the workings of human intelligence. Their premise was that the mind can be described as a circuit whose inputs are our sensory perceptions (vision, sound, smell, touch). What are the gates of these "circuits of the mind"? Neuroscience tells us that the basic blocks of the nervous system, including the brain, are neural cells, or neurons. These cells take in electrical signals from the sensory system or from other neurons and amplify or inhibit them depending on their function.

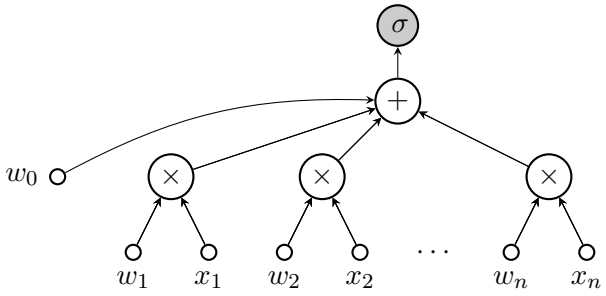
A perceptron is a model of a single neuron. It consists of two types of inputs: n *signals* x_1 up to x_n and an $n + 1$ *weights* $w_0, w_1, \ldots, w_n$. It outputs the value

$$y = \sigma(w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n),$$

where σ is a fixed real-valued function called the *activation function*. One popular choice of activation function is the *logistic function* $\sigma(t) = 1/(1 + e^{-t})$ which increases from 0 to 1:



The perceptron can be implemented by a circuit with multiplications, an addition, and a σ gate:



For example, a perceptron can be used to decide whether the room is dim or bright based on the readings of five "noise sensors" x_1, x_2, x_3, x_4, x_5. Suppose a given sensor outputs 1 if it detects light and -1 if it doesn't. The perceptron $\sigma(x_1 + x_2 + x_3 + x_4 + x_5)$ would then output an overall brightness estimate depending on the number k of sensors that output 1:

k	0	1	2	3	4	5
σ	0.01	0.05	0.27	0.73	0.95	0.99

Perceptrons are not only useful for modelling natural intelligence but also for endowing machines with the ability to make decisions from data. Suppose that you want to estimate your chances y of getting an A in this course. This might depend on a variety of factors such as your homework average x_1, your project grade x_2, and the number of hours x_3 that you studied each week. Your chances y can then reasonably be modeled as the output of some perceptron $\sigma(w_0 + w_1x_1 + w_2x_2 + w_3x_3)$. But how should you choose the weights w_0 to w_3?

The central dogma of Machine Learning is that you can estimate unknown parameters like w_0, w_1, w_2, w_3 from data using a *training algorithm*. From talking to your friends who took the course last year you gathered the following data:

name	x_1	x_2	x_3	A?
Alice	39	31	12	yes
Bob	45	11	3	no
Charlie	43	25	6	yes
	⋮	⋮	⋮	⋮
Zack	50	40	0	yes

For any given student s in the table, you would expect that the "indicator value" $\hat{y}(s)$ for the event that the student received an A should be close to the perceptron's estimate of this value:

$$\hat{y}(s) \approx \sigma(w_0 + w_1x_1(s) + w_2x_2(s) + w_3x_3(s)).$$

This approximation will never be exact; the left-hand side is a binary value (0 or 1) while the right hand side is some real number between 0 and 1. A natural error measure is the square loss

$$\ell(s) = \big(\hat{y}(s) - \sigma(w_0 + w_1x_1(s) + w_2x_2(s) + w_3x_3(s))\big)^2.$$

It is sensible to try to pick the weights w_0, w_1, w_2, w_3 so as to miminize the sum f of the individual losses:

$$\begin{aligned} f &= \ell(\text{Alice}) + \ell(\text{Bob}) + \cdots + \ell(\text{Zack}) \\ &= \big(1 - \sigma(w_0 + 39w_1 + 31w_2 + 12w_3)\big)^2 \\ &\quad + \big(0 - \sigma(w_0 + 45w_1 + 11w_2 + 3w_3)\big)^2 \\ &\quad \vdots \\ &\quad + \big(1 - \sigma(w_0 + 50w_1 + 40w_2 + 0w_3)\big)^2. \end{aligned}$$

Minimizing such expressions can be quite difficult. Gradient Descent is a natural algorithm to try. To run it we need to calculate the gradient at various inputs. That is where a circuit for the gradient comes in handy. In this example f is a function of the four model parameters w_0, w_1, w_2, w_3 but fairly large size. Forward Propagation and Backpropagation will produce circuits for ∇f of comparable size. In models with dozens or more parameters, however, Backpropagation yields orders of magnitude savings in efficiency.

Deep networks

While perceptrons are convenient to work with they are inadequate for more complex data-driven decision tasks. Suppose you want to know whether an image has a cat in it. The inputs $x_1, x_2, \ldots, x_n$ are the pixels and the output y is supposed to equal 1 if $x_1, \ldots, x_n$ form a cat and 0 if they don't. A perceptron would try to base its decision on the value $\sigma(w_0 + w_1x_1 + \cdots + w_nx_n)$ for some weights that represent the importance of different pixels. A weighted sum of pixels cannot take into account high-level features of vision such as edges between objects, foreground and background layers, and so on.

A more expressive model can be obtained by taking multiple perceptrons and organizing them in layers. By analogy with biological neurons we may expect that neurons closer to the inputs should be adequate for lower-level perception tasks such as edge-detection, while neurons closer to the output level may perform more cognitively demanding roles such as object detection and classification (e.g. is it an animal?)

There is a steep price to pay for this complexity: The model now consists of not one but many perceptrons so the number of unknown weights that needs to be estimated from the training data becomes very large. Moreover, the circuits representing these models become more complex. For the training to complete in a reasonable amount of time it is essential to have a systematic and efficient way to evaluate the gradient of the loss. Backpropagation is indispensable for training such deep networks effectively.

5 Circuit complexity

In calculus class we learn that the derivative of a univariate function f is the limiting value of $(f(x+\epsilon) - f(x))/\epsilon$ as ϵ approaches zero. The definition suggests a third "algorithm" for approximating the gradient: Estimate each partial derivative $\partial f/\partial x_i$ by $(f(x_1, \ldots, x_i + \epsilon, \ldots, x_n) - f(x_1, \ldots, x_n))/\epsilon$ for some small ϵ. This algorithm is tricky to implement. First, is unclear how to choose a suitable ϵ. Even if we did, calculating a ratio of numbers close to zero is notoriously sensitive to precision errors. Small changes in numerator and denominator cause the output to swing widely.

Nevertheless, suppose you had a magic wand that would make these issues go away. This "dream" algorithm would still be *less* efficient than Backpropagation: It would require evaluating f at $n+1$ different points, leading to a running time of $\Theta(ns)$. It would have the same asymptotic complexity as Forward Propagation.

What is remarkable about Backpropagation is that it allows evaluation of all n partial derivatives of f in much less time than it takes to calculate f at n different inputs. We usually expect that the amount of effort it takes to solve n problems should be n times larger than that for a single problem. Backpropagation bucks this intuition by virtue of its *white-box* nature: It ingeniously exploits the circuit representation of its input f.

This equivalence between the complexity of a function and its gradient has a surprising consequence regarding the circuit complexity of matrix computations. The goal of circuit complexity is to find the smallest possible circuit for computing a given function. For example, the second symmetric polynomial

$$\sigma_2(x_1, \ldots, x_n) = \sum_{i<j} x_i x_j = x_1x_2 + x_1x_3 + \cdots + x_1x_n + x_2x_3 + \cdots + x_{n-1}x_n$$

is expressed here as a sum of $\binom{n}{2}$ products. The circuit implementing it has size $\binom{n}{2} + 1 = \Theta(n^2)$. Is there a smaller circuit for it (when n is large)? Indeed there is: We can rewrite σ_2 as

$$\sigma_2(x_1, \ldots, x_n) = \tfrac{1}{2}(x_1 + \cdots + x_n)^2 - \tfrac{1}{2}(x_1^2 + \cdots + x_n^2),$$

which computes the same σ_2 in size $n + O(1)$.

Figuring out the smallest possible circuit for a given function is very difficult. An important example is the Fourier Transform function FT. FT takes as its input a list representation f of size N and outputs the Fourier representation $\hat{f}$ also of size N. The Fast Fourier Transform can be implemented as a circuit of size $N \log N$ known as the butterfly. Is this the smallest possible circuit for computing FT? The answer is not known.

Another important example in linear algebra is the matrix INVerse. A circuit for INV takes the n^2 entries of a square matrix X as inputs and produces the n^2 entries of the matrix X^{-1} as outputs. (To allow inversion the circuit has the gate $[g](x) = 1/x$ at its disposal.) Calculating the matrix inverse is closely related to solving linear equations. With a bit of work Gaussian Elimination can be turned into a circuit for INV of similar complexity, namely $\Theta(n^3)$. Can we do better?

Computer scientist Volker Strassen answered this question affirmatively in 1969. He came up with an beautiful recursive algorithm for inverting n by n matrices using $\Theta(n^{\log 7}) \approx \Theta(n^{2.81})$ operations. Strassen's

algorithm was the firing shot for what soon became a competitive sport. The current world record for matrix inversion is $\Theta(n^{2.371552})$. These algorithms are complicated and not particularly practical. Yet there is no reason to doubt that further improvements and simplifications are in the cards.

What does this have to do with Backpropagation? To explain its relevance we need to talk about yet another problem: computing the DETerminant of an $n \times n$ matrix X. The determinant of a matrix X is some complicated polynomial in its n^2 entries. How hard is it to calculate?

Gaussian Elimination can also be used to calculate determinants using $\Theta(n^3)$ operations. This is not a coincidence: In his 1969 paper Strassen showed that calculating the determinant of a matrix can be "reduced" to inverting some related matrices. Thus the complexity of calculating the determinant is asymptotically bounded by the complexity of calculating the inverse. If anyone figures out how to invert $n \times n$ matrices in quadratic time—the holy grail of matrix algorithms—they will have also figured out how to calculate the determinant in quadratic time.

This does not sound terribly surprising. After all, the determinant of a matrix is just one number, while its inverse is a whole matrix with n^2 numbers. Surely calculating the determinant should be *easier* than calculating the inverse. It is not! If there is a size-s circuit for calculating the $n \times n$ matrix determinant out there, there will also be a size-$O(s)$ circuit for the inverse.

The reason for this is Backpropagation. By Cramer's rule, the entries of the inverse matrix can be calculated by dividing the determinants of the minors by the determinant of the matrix:

$$(i,j)\text{-th entry of } INV(X) = \frac{DET(X \text{ with } i\text{-th row and } j\text{-th column removed})}{DET(X)}.$$

The determinant of X with the i-th row and j-th column removed is nothing but the partial derivative of $DET(X)$ with respect to variable X_{ij}:

$$(i,j)\text{-th entry of } INV(X) = \frac{\partial DET(X)/\partial X_{ij}}{DET(X)}.$$

Collecting all these entries into the matrix $INV(X)$ gives the beautiful formula

$$INV(X) = \frac{\nabla DET(X)}{DET(X)}$$

so its circuit complexity is at most that of the determinant plus that of the gradient (plus n^2 divisions). Theorem 3 tells us that if DET has a size-s circuit then ∇DET has a size-$O(s)$ circuit, and so does INV.[1]

[1] This reasoning assumes all circuit gates have *bounded fan-in*, namely they take a fixed number of arguments, guaranteeing that the number of wires is proportional to the number of nodes.