

Backprop as Functor: A compositional perspective on supervised learning

Brendan Fong David I. Spivak Rémy Tuyéras*

Department of Mathematics,
Massachusetts Institute of Technology

Abstract

A supervised learning algorithm searches over a set of functions $A \rightarrow B$ parametrised by a space P to find the best approximation to some ideal function $f: A \rightarrow B$. It does this by taking examples $(a, f(a)) \in A \times B$, and updating the parameter according to some rule. We define a category where these update rules may be composed, and show that gradient descent—with respect to a fixed step size and an error function satisfying a certain property—defines a monoidal functor from a category of parametrised functions to this category of update rules. A key contribution is the notion of request function. This provides a structural perspective on backpropagation, as well as a broad generalisation of neural networks.

1 Introduction

Machine learning, and in particular the use of neural networks, has rapidly become remarkably effective at real world tasks [9]. A significant contributor to this success has been the backpropagation algorithm. Backpropagation gives a way to compute the derivative of a function via message passing on a network, significantly speeding up learning. Yet, while the power of this approach has been impressive, it is also somewhat mysterious. What structures make backpropagation so effective, and how can we interpret, predict, and generalise it?

In recent years, monoidal categories have been used to formalise the use of networks in computation and reasoning—amongst others, applications include

*We thank Patrick Schultz and Daniel Trewartha for useful discussions. Work supported by AFOSR FA9550-14-1-0031 and FA9550-17-1-0058.

circuit diagrams, Markov processes, quantum computation, and dynamical systems [4, 1, 2, 10]. This paper responds to a need for more structural approaches to machine learning by using categories to provide an algebraic, compositional perspective on learning algorithms and backpropagation.

Consider a supervised learning algorithm. The goal of a supervised learning algorithm is to find a suitable approximation to a function $f: A \rightarrow B$. To do so, the supervisor provides a list of pairs $(a, b) \in A \times B$, each of which is supposed to approximate the values taken by f , i.e. $b \approx f(a)$. The supervisor also defines a space of functions over which the learning algorithm will search. This is formalised by choosing a set P and a function $I: P \times A \rightarrow B$. We denote the function at parameter $p \in P$ as $I(p, -): A \rightarrow B$. Then, given a pair $(a, b) \in A \times B$, the learning algorithm takes a current hypothetical approximation of f , say given by $I(p, -)$, and tries to improve it, returning some new best guess, $I(p', -)$. In other words, a supervised learning algorithm includes an *update* function $U: P \times A \times B \rightarrow P$ for I .

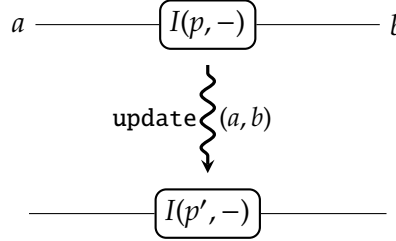


Figure 1. Given a training datum (a, b) , a learning algorithm updates p to p' .

To make this compositional, we ask the following question. Suppose we are given two learning algorithms, as described above, one for approximating functions $A \rightarrow B$ and the other for functions $B \rightarrow C$. Can we piece them together to make a learning algorithm for approximating functions $A \rightarrow C$? We will see that the answer is no, because something is missing.

To construct a learning algorithm, we would need a parameterised function $A \rightarrow C$ as well as an update rule. It is easy to take the given parameterised functions $I: P \times A \rightarrow B$ and $J: Q \times B \rightarrow C$ and produce one from A to C . Indeed, take $P \times Q$ as the parameter space and define the parametrised function $P \times Q \times A \rightarrow C$; $(p, q, a) \mapsto J(q, I(p, a))$. We call the function $J(-, I(-, -)): P \times Q \times A \rightarrow C$ the *composite parametrised function*.

The problem comes in defining the update rule for the composite learner. Our algorithm must take as training data pairs (a, c) in $A \times C$. However, to use the given update functions, written U and V for updating I and J respectively, we must produce training data of the form (a', b') in $A \times B$ and (b'', c'') in $B \times C$. It is straightforward to produce a pair in $B \times C$ —take $(I(p, a), c)$ —but there is no natural pair (a', b') to use as training data for I . The choice of b' should encode something about the information in both c and J , and nothing of the sort has been specified.

Thus to complete the compositional picture, we must add to our formalism a way for the second learning algorithm to pass back elements of B . We will call this a *request* function, because it is as though J is telling I what input b' would have been more helpful. The request function for J will be of the form $s: Q \times B \times C \rightarrow B$: given a hypothesis q and training data (b'', c'') , it returns $b' := s(q, b'', c'')$ for I to use. As such, the request function is a way of ‘backpropagating’ the output back toward the earlier learners in a network.

In this paper we will show that learning becomes *compositional*—i.e. we can define a learning algorithm $A \rightarrow C$ from learning algorithms $A \rightarrow B$ and $B \rightarrow C$ —as long as each learning algorithm consists of these four components:

- a parameter space P ,
- a function $I: P \times A \rightarrow B$,
- an update function $U: P \times A \times B \rightarrow P$, and
- a request function $r: P \times A \times B \rightarrow A$.

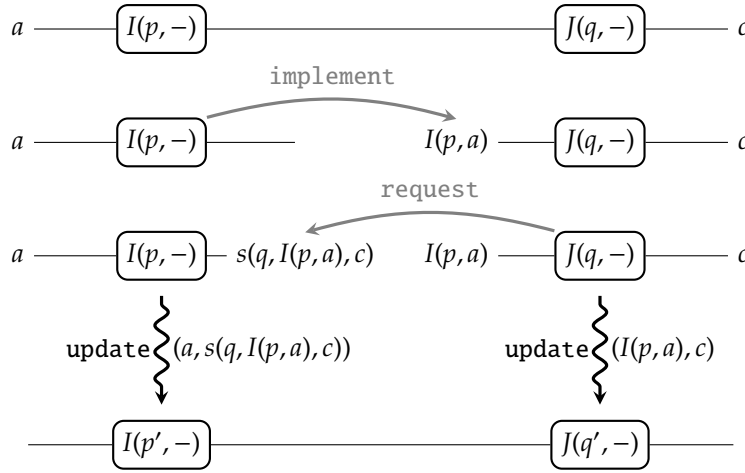


Figure 2. A request function allows an update function to be defined for the composite $J(q, I(p, -))$.

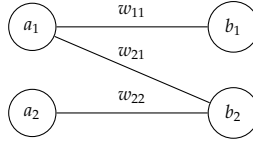
More precisely, we will show that learning algorithms (P, I, U, r) form the morphisms of a category **Learn**. A category is an algebraic structure that models composition. More precisely, a category consists of *types* A, B, C , and so on, *morphisms* $f: A \rightarrow B$ between these types, and a *composition rule* by which morphisms $f: A \rightarrow B$ and $g: B \rightarrow C$ can be combined to create a morphism $A \rightarrow C$. Thus we can say that learning algorithms form a category, as we have informally explained above. In fact, they have more structure because they can be composed not only in series but also in parallel, and this too has a clean algebraic description. Namely, we will show that **Learn** has the structure of a symmetric monoidal category.

Our aim thus far has been to construct an algebraic description of learning algorithms, and we claim that the category **Learn** suffices. In particular, then, our framework should be broad enough to capture known methods for constructing

supervised learning algorithms; such learning algorithms should sit inside **Learn** as a particular kind of morphism. Here we study neural networks.

Let us say that a *neural network layer of type* (n_1, n_2) is a subset $C \subseteq [n_1] \times [n_2]$, where $n_1, n_2 \in \mathbb{N}$ are natural numbers, and $[n] = \{1, \dots, n\}$ for any $n \in \mathbb{N}$. The numbers n_1 and n_2 represent the number of nodes on each side of the layer, C is the set of connections, and the inclusion $C \subseteq [n_1] \times [n_2]$ encodes the connectivity information, i.e. $(i, j) \in C$ means node i is connected to node j .

If we additionally fix a function $\sigma: \mathbb{R} \rightarrow \mathbb{R}$, which we call the *activation function*, then a neural network layer defines a parametrised function $\mathbb{R}^{|C|+n_2} \times \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$. For example, the layer $C = \{(1, 1), (2, 1), (2, 2)\} \subseteq [2] \times [2]$, has $|C| = 3$ connections and we draw it as follows



This layer defines the parametrised function $I: \mathbb{R}^5 \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$, given by

$$I(w_{11}, w_{21}, w_{22}, w_1, w_2, a_1, a_2) := \left(\sigma(w_{11}a_1 + w_1), \sigma(w_{21}a_1 + w_{22}a_2 + w_2) \right).$$

A neural network is a sequence of layers of types $(n_0, n_1), (n_1, n_2), \dots, (n_{k-1}, n_k)$. By composing the parametrised functions defined by each layer as above, a neural network itself defines a parametrised function $P \times \mathbb{R}^{n_0} \rightarrow \mathbb{R}^{n_k}$. Note that this function is always differentiable if σ is.

To go from a differentiable parametrised function to a learning algorithm, one typically specifies a suitable error function e and a step size ε , and then uses an algorithm known as gradient descent.

Our main theorem is that, under general conditions, gradient descent is compositional. This is formalised as a functor $\text{Para} \rightarrow \text{Learn}$, where Para is a category where morphisms are differentiable parametrised functions between finite dimensional Euclidean spaces. In brief, the functoriality means that given two differentiable parametrised functions I and J , we get the same result if we (i) use gradient descent to get learning algorithms for I and J , and then compose those learning algorithms, or (ii) compose I and J as parametrised functions, and then use gradient descent to get a learning algorithm.

Our main theorem is the following.

Theorem. Fix $\varepsilon > 0$ and $e(x, y): \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ such that $\frac{\partial e}{\partial x}(z, -): \mathbb{R} \rightarrow \mathbb{R}$ is invertible for each $z \in \mathbb{R}$. Then there is a faithful, injective-on-objects, symmetric monoidal functor

$$L: \text{Para} \longrightarrow \text{Learn}$$

sending each parametrised function $I: P \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ to the learning algorithm (P, I, U_I, r_I) defined by

$$U_I(p, a, b) := p - \varepsilon \nabla_p E_I(p, a, b)$$

and

$$r_I(p, a, b) := f_a(\nabla_a E_I(p, a, b)),$$

where $E_I(p, a, b) := \sum_i e(I(p, a)_i, b_i)$ and f_a denotes the component-wise application of the inverse to $\frac{\partial e}{\partial x}(a_i, -)$ for each i .

This theorem has a number of consequences. For now, let us name just three. The first is that we may train a neural network by using the training data on the whole network to create training data for each subunit, and then training each subunit separately. To some extent this is well-known: it is responsible for speedups due to backpropagation, as one never needs to compute the derivatives of the function defined by the entire network. However the fact that this functor is symmetric monoidal structure shows that we can vary the backpropagation algorithm to factor the neural network into richer sub-parts than simply carving it layer by layer.

Second, it gives a sufficient condition—which is both straightforward and general—under which an error function works well under backpropagation.

Finally, it shows that backpropagation can be applied far more generally than just to neural networks: it is compositional for all differentiable parametrised functions. As a consequence, it shows that backpropagation gives a sound method for computing gradient descent even if we introduce far more general elements into neural networks than the traditional linear and activation functions.

Overview

In Section 2, we define the category **Learn** of learning algorithms. We may then immediately get to the main theorem in Section 3: given a choice of error function and step size, gradient descent and backpropagation give a functor from the category of parametrised functions to the category of learning algorithms. In Section 4, we broaden this view to show how it relates to neural networks. Next, in Section 5, we note that the category **Learn** has additional structure beyond just that of a symmetric monoidal category: it has bimonoid structures that allow us to split and merge connections to form networks. We also show this is useful in understanding the construction of individual neurons, and in weight tying and convolutional neural nets. We then explicitly compute an example of functoriality from neural nets to learning algorithms (§6), and discuss implications for this framework (§7). An appendix provides more technical aspects of the proof of the main theorem, and a brief, diagram-driven introduction to category theory.

2 The category of learners

In this section we define a symmetric monoidal category **Learn** that models supervised learning algorithms and their composites. See Appendix C for back-

ground on categories and string diagrams.

Definition 2.1. Let A and B be sets. A supervised learning algorithm, or simply learner, $A \rightarrow B$ is a tuple (P, I, U, r) where P is a set, and I , U , and r are functions of types:

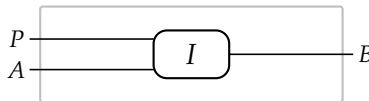
$$\begin{aligned} I: P \times A &\rightarrow B, \\ U: P \times A \times B &\rightarrow P, \\ r: P \times A \times B &\rightarrow A. \end{aligned}$$

We call P the *parameter space*; it is just a set. The map I implements a parameter value $p \in P$ as a function $I(p, -): A \rightarrow B$. We think of a pair $(a, b) \in A \times B$ as a *training datum*; it pairs an input a with an output b . The map $U: P \times A \times B \rightarrow P$ is the *update function*; given a ‘current’ parameter p and a training datum $(a, b) \in A \times B$, it produces an ‘updated’ parameter $U(p, a, b) \in P$. This can be thought of as the learning step. The idea is that the updated function $I(U(p, a, b), -) \in B$ would hopefully send a closer to b than the function $I(p, -)$ did, though this is not a requirement and is certainly not always true in practice. Finally, we have the *request function* $r: P \times A \times B \rightarrow A$. This takes the same datum and produces a ‘requested value’ $r(p, a, b) \in A$. The idea is this value will be sent to upstream learners for their own training.

Remark 2.2. The request function is perhaps a little mysterious at this stage. Indeed, it is superfluous to the definition of a standalone learning algorithm: all we need for learning is a space P of functions $I(p, -)$ to search over, and a rule U for updating our parameter p in light of new information. As we emphasised in the introduction, the request function is crucial in *composing* learning algorithms: there is no composite update rule without the request function.

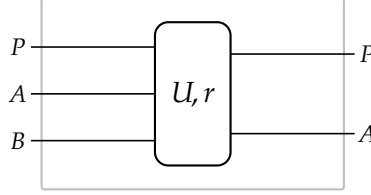
Another way to understand the role of the request function comes from experiments in machine learning. Fixing some parameter p and hence a function $I(p, -)$, the request function allows us to choose a desired output b , and then for any input a return a new input $a' := r(p, a, b)$. In the case of backpropagation, we will see we then have the intuition that $I(p, a')$ is closer to b than $I(p, a)$. For example, if we are classifying images, and b is the value indicating the classification ‘cat’, then a' will be a more ‘cat-like’ version of the image a . This is similar in spirit to what has been termed inversion or ‘dreaming’ in neural nets [8].

Remark 2.3. Using string diagrams¹ in (\mathbf{Set}, \times) , we can draw an implementation function I as follows:



¹String diagrams are an alternative, but nonetheless still formal, syntax for morphisms in a monoidal category. See Appendix C for more details.

One can do the same for U and r , though we find it convenient to combine them into a single update–request function $(U, r): P \times A \times B \rightarrow P \times A$. This function can be drawn as follows:



Let (P, I, U, r) and (P', I', U', r') be learners of the type $A \rightarrow B$. We consider them to be equivalent if there is a bijection $f: P \xrightarrow{\cong} P'$ such that the following hold for each $p \in P$, $a \in A$, and $b \in B$:

$$\begin{aligned} I'(f(p), a) &= I(p, a), \\ U'(f(p), a, b) &= U(p, a, b), \\ r'(f(p), a, b) &= r(p, a, b). \end{aligned}$$

Proposition 2.4. *There exists a symmetric monoidal category **Learn** whose objects are sets and whose morphisms are equivalence classes of learners.*

The proof of Proposition 2.4 is given in Appendix A. For now, we simply specify the composition, identities, monoidal product, and braiding for this symmetric monoidal category.

Composition. Suppose we have a pair of learners

$$A \xrightarrow{(P, I, U, r)} B \xrightarrow{(Q, J, V, s)} C.$$

The composite learner $A \rightarrow C$ is defined to be $(P \times Q, I * J, U * V, r * s)$, where the implementation function is

$$(I * J)(q, p, a) := J(q, I(p, a))$$

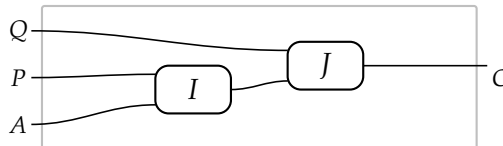
the update function is

$$(U * V)(q, p, a, c) := \left(U(p, a, s(q, I(p, a), c)), V(q, I(p, a), c) \right),$$

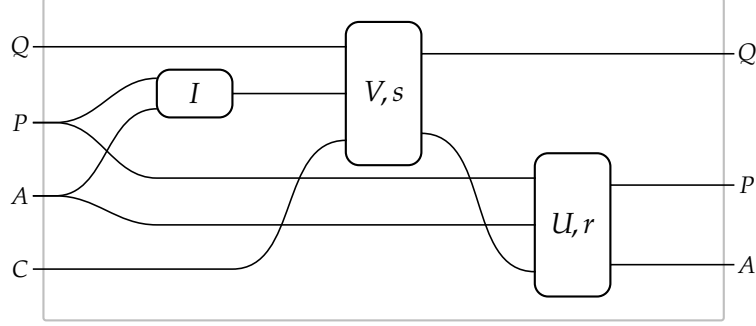
and the request function

$$(r * s)(q, p, a, c) := r(p, a, s(q, I(p, a), c)).$$

Let us also present the composition rule using string diagrams in (\mathbf{Set}, \times) . Given learners (P, I, U, r) and (Q, J, V, s) as above, the composite implementation function can be written as



while the composite update–request function $(U * V, r * s)$ can be written as:



Here the splitting represents the diagonal map $A \rightarrow A \times A$, i.e. $a \mapsto (a, a)$. We hope that the reader might find visually tracing through these diagrams helpful for making sense of the composition rule; see the introduction for further intuition.

Identities. For each object A , we have the identity map

$$(\mathbf{1}, \text{id}, !, \pi_2): A \longrightarrow A,$$

where $\mathbf{1}$ is a one element set, $\text{id}: \mathbf{1} \times A \rightarrow A$ is the second projection (as this is a bijection, we abuse notation to write this projection as id), $!: \mathbf{1} \times A \times A \rightarrow \mathbf{1}$ is the unique function, and $\pi_2: A \times A \rightarrow A$ is the second projection.

Monoidal product. The monoidal product of objects A and B is simply their cartesian product $A \times B$ as sets. The monoidal product of morphisms $(P, I, U, r): A \rightarrow B$ and $(Q, J, V, s): C \rightarrow D$ is defined to be $(P \parallel Q, I \parallel J, U \parallel V, r \parallel s)$, where the implementation function is

$$(I \parallel J)(p, q, a, c) := (I(p, a), J(q, c))$$

the update function is

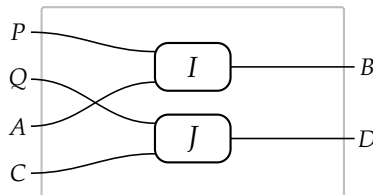
$$(U \parallel V)(p, q, a, c, b, d) := (U(p, a, b), V(q, c, d))$$

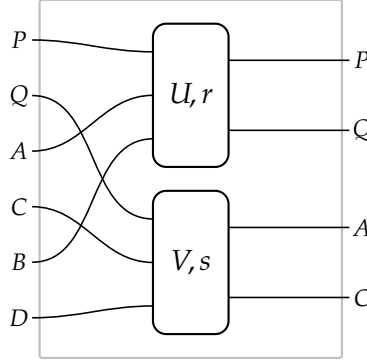
and the request function is

$$(r \parallel s)(p, q, a, c, b, d) := (r(p, a, b), s(q, c, d)).$$

We use the notation \parallel because monoidal product can be thought of as parallel—rather than series—composition.

We also present this in string diagrams:





Braiding. A symmetric braiding $A \times B \rightarrow B \times A$ is given by $(\mathbf{1}, \sigma, !, \sigma \circ \pi_2)$ where $\sigma: A \times B \rightarrow B \times A$ is the usual swap function $(a, b) \mapsto (b, a)$.

A proof that this is a well-defined symmetric monoidal category can be found in Appendix A.

3 Gradient descent and backpropagation

In this section we show that gradient descent and backpropagation define a strict symmetric monoidal functor from the symmetric monoidal category of differentiable parametrised functions between finite dimensional Euclidean spaces to the symmetric monoidal category **Learn** of learning algorithms.

We first define the category of differentiable parametrised functions. A *Euclidean space* is one of the form \mathbb{R}^n for some $n \in \mathbb{N}$. We call n the *dimension* of the space, and write an element $a \in \mathbb{R}^n$ as (a_1, \dots, a_n) , or simply $(a_i)_i$, where each $a_i \in \mathbb{R}$.

For Euclidean spaces $A = \mathbb{R}^n$ and $B = \mathbb{R}^m$, define a *differentiable parametrised function* to be a pair (P, I) , where P is a Euclidean space and $I: P \times A \rightarrow B$ is a differentiable function. We consider two parametrised functions (P, I) and (P', I') to be *equivalent* if there is a bijection $f: P \xrightarrow{\cong} P'$ such that for all $p \in P$ and $a \in A$, we have $I'(f(p), a) = I(p, a)$, and such that f and f^{-1} are differentiable. We shall abuse notation to simply write the equivalence class of (P, I) , where $I: P \times \mathbb{R}^n \rightarrow \mathbb{R}^m$, as $I: \mathbb{R}^n \rightarrow \mathbb{R}^m$.

Differentiable parametrised functions between Euclidean spaces form a symmetric monoidal category.

Definition 3.1. We write **Para** for the strict symmetric monoidal category whose objects are Euclidean spaces and whose morphisms $\mathbb{R}^n \rightarrow \mathbb{R}^m$ are equivalence classes of differentiable parametrised functions $\mathbb{R}^n \rightarrow \mathbb{R}^m$.

Composition of $I: \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $J: \mathbb{R}^m \rightarrow \mathbb{R}^\ell$ is given by

$$(I * J)(q, p, a) = J(q, I(p, a)).$$

The monoidal product of objects \mathbb{R}^n and \mathbb{R}^m is the object \mathbb{R}^{n+m} , while the monoidal product of morphisms $I: \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $J: \mathbb{R}^\ell \rightarrow \mathbb{R}^k$ is given by

$$(I \parallel J)(p, q, a, c) = (I(p, a), J(q, c)).$$

The braiding $\mathbb{R}^n \parallel \mathbb{R}^m \rightarrow \mathbb{R}^m \parallel \mathbb{R}^n$ is given by $(a, b) \mapsto (b, a)$.

It is straightforward to check this is a well defined symmetric monoidal category. We are now in a position to state the main theorem.

Theorem 3.2. Fix a real number $\varepsilon > 0$ and $e(x, y): \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ differentiable such that $\frac{\partial e}{\partial x}(z, -): \mathbb{R} \rightarrow \mathbb{R}$ is invertible for each $z \in \mathbb{R}$. Then we can define a faithful, injective-on-objects, symmetric monoidal functor

$$L: \text{Para} \longrightarrow \text{Learn}$$

that sends each parametrised function $I: P \times A \rightarrow B$ to the learner (P, I, U_I, r_I) defined by

$$U_I(p, a, b) := p - \varepsilon \nabla_p E_I(p, a, b)$$

and

$$r_I(p, a, b) := f_a \left(\nabla_a E_I(p, a, b) \right),$$

where $E_I(p, a, b) := \sum_j e(I(p, a)_j, b_j)$, and f_a is component-wise application of the inverse to $\frac{\partial e}{\partial x}(a_i, -)$ for each i .

Proof (sketch). The proof of this theorem amounts to observing that the chain rule is functorial given the above setting. The key points are the use of the chain rule to show the functoriality of the P -part of the update function and the request function. A full proof is given in Appendix B. \square

We call ε the *step size*, e the *error function*, and E_I the *total error* (with respect to I). We also call the functors L , so named because they turn parametrised functions into a *learning* algorithms, the *gradient descent/backpropagation functors*.

Remark 3.3. The update function U_I encodes what is known as *gradient descent*: the parameter p is updated by moving it an ε -step in the direction that most reduces the total error E_I .

The request function r_I encodes the *backpropagation* value, passing back, up to the invertible function f_a , the gradient of the total error with respect to the input a . In particular, the functoriality of L says that the following two update functions are equal:

- The update function $U_{((I \parallel J) * K) * M}$, which represents gradient descent on the composite of parametrised functions $((I \parallel J) * K) * M$.
- The update function $((U_I \parallel U_J) * U_K) * U_M$, which represents the composite, according to the structure in **Learn**, of the update functions U_I, U_J, U_K , and U_M together with the request functions r_I, r_J, r_K , and r_M .

This shows that we may compute gradient descent by local computation of the gradient together with local message passing. This is the backpropagation algorithm.

Example 3.4 (Quadratic error). Quadratic error is given by the error function $e(x, y) := \frac{1}{2}(x - y)^2$, so that the total error is given by

$$E_I(p, a, b) = \frac{1}{2} \sum_j (I_j(p, a) - b_j)^2 = \frac{1}{2} \|I(p, a) - b\|^2.$$

In this case $\frac{\partial e}{\partial x}(x, -)$ is the function $y \mapsto x - y$. This function is its own inverse, so we have $f_x(y) := x - y$.

Fixing some step size $\varepsilon > 0$, we have

$$\begin{aligned} U_I(p, a, b) &:= p - \varepsilon \nabla_p E_I(p, a, b) \\ &= \left(p_k - \varepsilon \sum_j (I_j(p, a) - b_j) \frac{\partial I_j}{\partial p_k} \right)_k \end{aligned}$$

and similarly

$$\begin{aligned} r_I(p, a, b) &:= a - \nabla_a E_I(p, a, b) \\ &= \left(a_i - \sum_j (I_j(p, a) - b_j) \frac{\partial I_j}{\partial a_i} \right)_i. \end{aligned}$$

Thus given this choice of error function, the functor L of Theorem 3.2 just implements, as update function, the usual gradient descent with step size ε with respect to the quadratic error.

Remark 3.5. Note that the requests in Example 3.4 have an interesting form, appearing as ‘gradient descent with step size 1’. One might wonder, for example, why the step size ε is not important.

Theorem 3.2 shows, however, that this is somewhat of a coincidence. What is important about requests, and hence the messages passed backward in backpropagation, is the fact that they are constructed inverting certain partial derivatives and applying the result to the gradient of the total error with respect to the input. We interpret this as a corrected input value, that if used would reduce the total error with respect to the given output and current parameter value. In particular, the resemblance of the request values to gradient descent in Example 3.4 is just an artifact of the choice of quadratic error.

4 From networks to parametrised functions

In the previous section we showed that gradient descent and backpropagation are a method, dependent on a choice of error function and step size, for defining a functor from differentiably parametrised functions to supervised learning algorithms. But backpropagation is often considered an algorithm executed on

a neural net. How do neural nets come into the picture? As we shall see, neural nets are a method for defining parametrised functions.

This method, like backpropagation itself, is also compositional—it respects the gluing together of neural networks. To formalise this, we thus first define a category **NNet** of neural networks. Implementation of each neural net will then define a parametrised function, and in fact we get a functor from **NNet** to **Para**. Note that just as defining a gradient descent/backpropagation functor depends on a choice, so too does defining an implementation functor. Namely, we must choose an activation function.

Recall from the introduction that a neural network layer of type (n, m) is a subset of $[n] \times [m]$, where $n, m \in \mathbb{N}$ and $[n] = \{1, \dots, n\}$. A k -layer neural network of type (n, m) is a sequence of neural network layers of types $(n_0, n_1), (n_1, n_2), \dots, (n_{k-1}, n_k)$, where $n_0 = n$ and $n_k = m$. A neural network of type (n, m) is a k -layer neural network for some k .

Given a neural network of type (n, m) and a neural network of type (m, p) we may concatenate them to get a neural network of type (n, p) . Note that when $n = m$, we consider the 0-layer neural network to be a morphism. Concatenating any neural network on either side with the 0-layer neural network does not change it.

Definition 4.1. *The category **NNet** of neural networks has as objects natural numbers and as morphisms $n \rightarrow m$ neural networks of type (n, m) . Composition is given by concatenation of neural networks. The identity morphism on n is the 0-layer neural network.*

Observe that since composition is just concatenation, it is immediately associative, and we have indeed defined a category.

Proposition 4.2. *Given a differentiable function $\sigma: \mathbb{R} \rightarrow \mathbb{R}$, we have a functor*

$$I: \mathbf{NNet} \longrightarrow \mathbf{Para}.$$

On objects, I maps each natural number n to the n -dimensional Euclidean space \mathbb{R}^n .

On morphisms, each 1-layer neural network $C: n \rightarrow m$ is mapped to the parametrised function

$$I_C: \mathbb{R}^{|C|+m} \times \mathbb{R}^n \longrightarrow \mathbb{R}^m;$$

$$((w_{ji}, w_j), x_i) \longmapsto \left(\sigma \left(\sum_i w_{ji} x_i + w_j \right) \right)_{1 \leq j \leq m}.$$

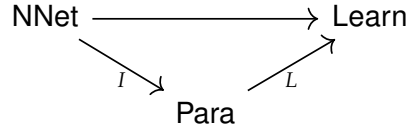
Given a neural net $N = C_1, \dots, C_n$, the image under I is the composite of the image of each layer:

$$I_N = I_{C_1} * \dots * I_{C_n}$$

We call σ the *activation function*. We also call the w_{ji} *weights*, where $(i, j) \in C$, and call the w_j *biases*, where $j \in n_2$.

Proof. The proof of this proposition is straightforward. Note in particular that the image I_C of each layer C is differentiable, and so their composites are too. Also note that the image of a 0-layer neural net is the empty composite, so identities map to identities. Composition is preserved by definition. \square

Composing an implementation functor I with a gradient descent/backpropagation functor L , we get a functor



This states that, given choices of activation function, error function, and step size, a neural net defines a supervised learning algorithm, and does so in a compositional way.

A symmetric monoidal structure can be given by generalising the category **NNet** to the category where morphisms are directed acyclic graphs with interfaces; details on such a category can be found in [3]. We will say a few more words about this discussion in Section 7.

In Section 6, we will compute an extended example of using neural nets to compositionally define supervised learning algorithms. Before this, however, we discuss additional compositional structure available to us in **Learn**, **Para**, and, although we shall not see it here, the aforementioned monoidal generalisation of **NNet**.

5 Networking in Learn

Our formulation of supervised learning algorithms as morphisms in a monoidal category means learning algorithms can be formed by combining other learning algorithms in sequence and in parallel. In fact, as hinted at by neural networks themselves, more structure is available to us: we are able to form new learning algorithms by combining others in networks of learners where wires can split and merge. Formally, this means each object in the category of learners is equipped with the structure of a bimonoid.

For this, note first that the symmetric monoidal category **FVect** of linear maps between Euclidean spaces sits inside the category **Para** of parametrised functions; we simply consider each linear map as parametrised by the trivial parameter space **1**. Given a choice of step size and error function, and hence a map $L: \text{Para} \rightarrow \text{Learn}$, we thus have an inclusion

$$\mathbf{FVect} \hookrightarrow \mathbf{Para} \xrightarrow{L} \mathbf{Learn}.$$

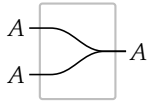
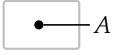
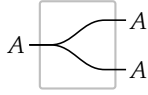

This allows us to construct learning algorithms—that is, morphisms in **Learn**—that are the images of morphisms in **FVect**, and hence obey known equations. In

particular, each object in \mathbf{FVect} is equipped with the structure of a bimonoid, and hence, given a choice of L , we can equip each object of \mathbf{Learn} with a bimonoid structure. This bimonoid structure is what makes the neural network notation feasible: we can interpret the splitting and combining in a way coherent with composition.

In fact, the bimonoids constructed depend only on the choice of error function; we need not specify the step size. As an example, we detail the construction using backpropagation with respect to the quadratic error (Example 3.4).

Proposition 5.1. *Gradient descent with respect to the quadratic error and step size ε defines a symmetric monoidal functor $\mathbf{FVect} \rightarrow \mathbf{Learn}$. This implies each object of \mathbf{Learn} can be equipped with the structure of a bimonoid.*

Explicitly, the bimonoid maps are given as follows. Note they all have trivial parameter space, which means one need not consider an update function.

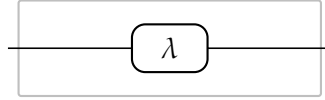
	Implementation	Request
<p>Multiplication, μ</p> <p>$(1, I_\mu, !, r_\mu)$</p> 	$I_\mu: A \times A \longrightarrow A;$ $(a_1, a_2) \mapsto a_1 + a_2$	$r_\mu: (A \times A) \times A \longrightarrow A \times A;$ $(a_1, a_2, a_3) \mapsto (a_3 - a_2, a_3 - a_2)$
<p>Unit, η</p> <p>$(1, I_\eta, !, r_\eta)$</p> 	$I_\eta: \mathbb{R}^0 \longrightarrow A;$ $0 \mapsto 0$	$r_\eta: A \longrightarrow \mathbb{R}^0;$ $a \mapsto 0$
<p>Comultiplication, δ</p> <p>$(1, I_\delta, !, r_\delta)$</p> 	$I_\delta: A \longrightarrow A \times A;$ $a \mapsto (a, a)$	$r_\delta: A \times (A \times A) \longrightarrow A;$ $(a_1, a_2, a_3) \mapsto a_2 + a_3 - a_1$
<p>Counit, ϵ</p> <p>$(1, I_\epsilon, !, r_\epsilon)$</p> 	$I_\epsilon: A \longrightarrow \mathbb{R}^0;$ $a \mapsto 0$	$r_\epsilon: A \longrightarrow A;$ $a \mapsto 0$

Remark 5.2. We actually have many different bimonoid structures in **Learn**: each choice of error function defines one, and these are often distinct. For example, if we choose $e(x, y) = xy$ then the request function on the multiplication instead given by $r'_\mu(a_1, a_2, a_3) = (a_3, a_3)$ and the request function on the comultiplication instead given by $r'_\delta(a_1, a_2, a_3) = a_2 + a_3$. This is a rather strange error function: minimising error entails sending outputs to 0. But we do not rule out that there may be useful bimonoid structures other than the one described above.

A choice of bimonoid structures, such as that given by Proposition 5.1, allows us to interpret network diagrams in **Learn**, as they give canonical interpretations of splitting, joining, initializing, and discarding wires.

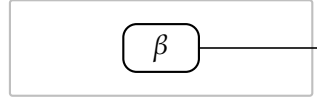
Example 5.3 (Building neurons). As learning algorithms with respect to quadratic error and some step size ε , neural networks have a rather simple structure: they are generated by three basic learning algorithms—scalar multiplication λ , bias β , and an activation function σ —together with the bimonoid multiplication and comultiplication given by Proposition 5.1.

The scalar multiplication learning algorithm $\lambda: \mathbb{R} \rightarrow \mathbb{R}$, which we shall represent graphically by the string diagram in (**Learn**, ||)



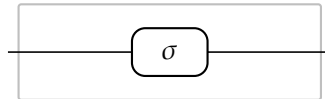
is given by the parameter space \mathbb{R} , implementation function $\lambda(w, x) = wx$, update function $U_\lambda(w, x, y) = w - \varepsilon x(wx - y)$, and request function $r_\lambda(w, x, y) = x - w(wx - y)$.

The bias learning algorithm $\beta: \mathbf{1} \rightarrow \mathbb{R}$, which we represent



is given by the parameter space \mathbb{R} , implementation $\beta(w) = w$, update function $U_\beta(w, y) = (1 - \varepsilon)w + \varepsilon y$, and trivial request function, since it has trivial input space.

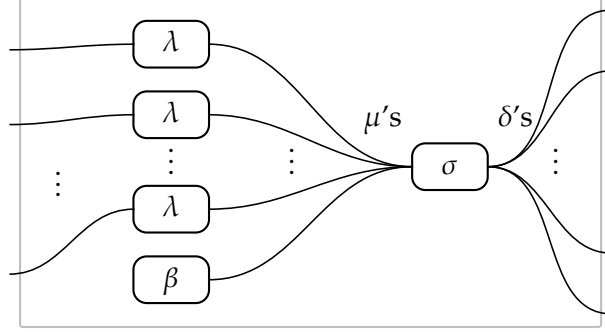
The activation function learning algorithm $\sigma: \mathbb{R} \rightarrow \mathbb{R}$, represented



has trivial parameter space, and is specified by some choice of activation function $\sigma: \mathbb{R} \rightarrow \mathbb{R}$, together with the trivial update function and the request function $r_\sigma(x, y) = x - (x - y)\frac{\partial \sigma}{\partial x}(x)$.

Then, every neuron in a neural network can be understood as a composite of these generators as follows: first, a monoidal product of the required number

of scalar multiplication algorithms and a bias algorithm, then a composite of μ s, an activation function, and finally a composite of δ s.

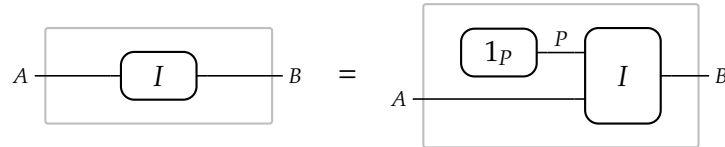


Composing these units using the composition rule in **Learn** further constructs any learning algorithm that can be obtained by gradient descent and backpropagation on a neural network with respect to the quadratic error.

Example 5.4 (Weight tying). Weight tying (or weight sharing) in neural networks is a method by which parameters in different parts of the network are constrained to be equal. It is used in convolutional neural networks, for example, to force the network to learn the same sorts of basic shapes appearing in different parts of an image. This is easily represented in our framework. Before explaining how this works, we first explain a way of factoring morphisms in **Para** into basic parts.

Morphisms in **Para** are roughly generated by morphisms of two different types: trivially parametrised functions and parametrised constants. Given a differentiable function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$, we consider it a *trivially parametrised function* $\mathbb{R}^0 \times \mathbb{R}^n \rightarrow \mathbb{R}^m$, whose parameter space $P = \mathbb{R}^0$ is a point. By a *parametrised constant*, we mean an identity morphism $1_P: P \rightarrow P$, considered as a parametrised function $P \times \mathbb{R}^0 \rightarrow P$.

In particular, every parametrised function can be written as a composite, using the bimonoid structure, of a trivially parametrised function and a parametrised constant. To see this, we use string diagrams in $(\mathbf{Para}, \parallel)$, where as usual we denote a parametrised function $I: P \times A \rightarrow B$ as a box labeled I with input A and output B . It is easy to check that any parametrised function $I: P \times A \rightarrow B$ is the composite of a trivially parametrised function and a parametrised constant as follows

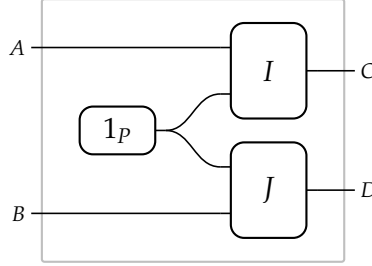


Note that the I on the left hand side and the I on the right hand side represent different parametrised functions: one has domain A and parameter space P , the other has domain $P \times A$ and parameter space \mathbb{R}^0 .

Since these morphisms are the same in **Para**, they correspond to the same learning algorithm, by Theorem 3.2. Looking at the right hand picture, suppose

given a training datum (a, b) . The (\mathbb{R}^0, I) block has trivial parameter space, so updates on it do nothing; however, it is capable of sending a request to the input A and the $(P, 1_P)$ block. The $(P, 1_P)$ block then performs the desired update. Again, the result of doing so must be the same, by the main theorem.

This suggests how one should think of weight tying. The schematic idea, represented in string diagrams, is as follows:



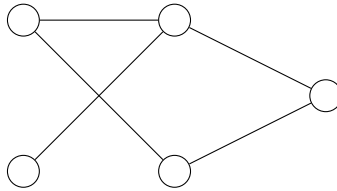
Here I and J are both trivially parametrised functions, while 1_P is a parametrised constant.

The comonoid structure from Proposition 5.1 tells us how the above network will behave as a learning algorithm with respect to quadratic error. The splitting wire will send the same parameter to both implementations I and J , and it will update itself based on the sum of the requests received from I and J .

6 Example: deep learning

In this section we explicitly compute an example of the functoriality of implementing a neural network as a supervised learning algorithm. For this we fix an activation function σ , as well as the quadratic error function, and a step size $\varepsilon > 0$. This respectively defines functors $I: \mathbf{NNet} \rightarrow \mathbf{Para}$ and $L: \mathbf{Para} \rightarrow \mathbf{Learn}$. In particular, we shall see that L implements the usual backpropagation algorithm with quadratic error and step size ε on a neural network with activation function σ .

Consider the single hidden layer network:



Call this network A ; it is a morphism $A: 2 \rightarrow 1$ in the category \mathbf{NNet} of neural networks. Our choice σ of activation function defines a functor $I: \mathbf{NNet} \rightarrow \mathbf{Para}$. The image of A under this functor is the parametrised function:

$$I_A: (\mathbb{R}^5 \times \mathbb{R}^3) \times \mathbb{R}^2 \longrightarrow \mathbb{R};$$

$$(p, q, a) \longmapsto \sigma(q_1 \sigma(p_{11}a_1 + p_{12}a_2 + p_{1b}) + q_2 \sigma(p_{21}a_1 + p_{2b}) + q_b).$$

Here the parameter space is $\mathbb{R}^5 \times \mathbb{R}^3$, since there is a weight for each of the three edges in the first layer, a bias for each of the two nodes in the intermediate column, a weight for each of the two edges in the second, and a bias for the output node. The input space is \mathbb{R}^2 , since there are two neurons on the leftmost side of the network, and the output space is \mathbb{R} , since there is a single neuron on the rightmost side.

We write the entries of the parameter space $\mathbb{R}^5 \times \mathbb{R}^3$ as $p_{11}, p_{12}, p_{21}, p_{1b}, p_{2b}, q_1, q_2$, and q_b , where p_{ji} represents the weight on the edge from the i th node of the first column to the j th node of the second column, p_{jb} represents the bias at the j th node of the second column, q_j represents the weight on the edge from the j th node of the second column to the unique node of the final column, and q_b represents the bias at the output node.

Suppose we wish to train this network. A training method is given by the functor L , which turns this parametrised function I_A into a supervised learning algorithm. In particular, given a training datum pair (a, c) in $\mathbb{R}^2 \times \mathbb{R}$, we wish to obtain a map $\mathbb{R}^5 \times \mathbb{R}^3 \rightarrow \mathbb{R}^5 \times \mathbb{R}^3$ that updates the value of (p, q) . As we have chosen to define L by using gradient descent with respect to the quadratic error function and an ε step size, this map is precisely the update map given by the L -image of I_A in Learn . That is, this parametrised function maps to the learning algorithm $(\mathbb{R}^5 \times \mathbb{R}^3, I_A, U_A, r_A)$, where

$$\begin{aligned} U_A: (\mathbb{R}^5 \times \mathbb{R}^3) \times \mathbb{R}^2 \times \mathbb{R} &\longrightarrow \mathbb{R}^5 \times \mathbb{R}^3; \\ (p, q, a, c) &\longmapsto \begin{pmatrix} p \\ q \end{pmatrix} - \varepsilon \nabla_{p,q} \frac{1}{2} \|I_A(p, q, a) - c\|^2 \\ &= \begin{pmatrix} p_{11} - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_1\dot{\sigma}(\beta_1)a_1 \\ p_{12} - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_1\dot{\sigma}(\beta_1)a_2 \\ p_{21} - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_2\dot{\sigma}(\beta_2)a_1 \\ p_{1b} - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_1\dot{\sigma}(\beta_1) \\ p_{2b} - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_1\dot{\sigma}(\beta_2) \\ q_1 - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)\sigma(\beta_1) \\ q_2 - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)\sigma(\beta_2) \\ q_b - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma) \end{pmatrix}, \end{aligned}$$

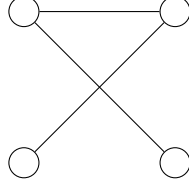
and

$$\begin{aligned} r_A: (\mathbb{R}^5 \times \mathbb{R}^3) \times \mathbb{R}^2 \times \mathbb{R} &\longrightarrow \mathbb{R}^2; \\ (p, q, a, c) &\longmapsto a - \nabla_a \frac{1}{2} \|I_A(p, q, a) - c\|^2 \\ &= \begin{pmatrix} a_1 - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)(q_1\dot{\sigma}(\beta_1)p_{11} + q_2\dot{\sigma}(\beta_2)p_{21}) \\ a_2 - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_1\dot{\sigma}(\beta_1)p_{12} \end{pmatrix} \end{aligned}$$

where γ is such that $I_A(p, q, a) = \sigma(\gamma)$, where $\beta_1 = p_{11}a_1 + p_{12}a_2 + p_{1b}$, where $\beta_2 = p_{21}a_1 + p_{2b}$, and where $\dot{\sigma}$ is the derivative of the activation function σ . (Explicitly, $\gamma = q_1\sigma(p_{11}a_1 + p_{12}a_2 + p_{1b}) + q_2\sigma(p_{21}a_1 + p_{2b}) + q_b$.) Note that U_A executes gradient descent as claimed.

The above expression for U_A is complex. It, however, reuses computations like γ, β_1 , and β_2 repeatedly. To simplify computation, we might try to factor it.

A factorisation can be obtained from the neural net itself. Note that the above net may be written as the composite of two layers. The first layer $B: 2 \rightarrow 2$



maps to the parametrised function

$$I_B: \mathbb{R}^5 \times \mathbb{R}^2 \longrightarrow \mathbb{R}^2;$$

$$(p, a) \longmapsto \begin{pmatrix} \sigma(p_{11}a_1 + p_{12}a_2 + p_{1b}) \\ \sigma(p_{21}a_1 + p_{2b}) \end{pmatrix}$$

which in turn has update and request functions

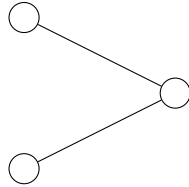
$$U_B: \mathbb{R}^5 \times \mathbb{R}^2 \times \mathbb{R}^2 \longrightarrow \mathbb{R}^5;$$

$$(p, a, b) \longmapsto \begin{pmatrix} p_{11} - \varepsilon(I_B(p, a)_1 - b_1)\dot{\sigma}(\beta_1)a_1 \\ p_{12} - \varepsilon(I_B(p, a)_1 - b_1)\dot{\sigma}(\beta_1)a_2 \\ p_{21} - \varepsilon(I_B(p, a)_2 - b_2)\dot{\sigma}(\beta_2)a_1 \\ p_{1b} - \varepsilon(I_B(p, a)_2 - b_2)\dot{\sigma}(\beta_1) \\ p_{2b} - \varepsilon(I_B(p, a)_2 - b_2)\dot{\sigma}(\beta_2) \end{pmatrix}$$

$$r_B: \mathbb{R}^5 \times \mathbb{R}^2 \times \mathbb{R}^2 \longrightarrow \mathbb{R}^2;$$

$$(p, a, b) \longmapsto \begin{pmatrix} a_1 - (I_B(p, a)_1 - b_1)\dot{\sigma}(\beta_1)p_{11} + (I_B(p, a)_2 - b_2)\dot{\sigma}(\beta_2)p_{21} \\ a_2 - (\sigma(I_B(p, a)_1 - b_1)\dot{\sigma}(\beta_1)p_{12} \end{pmatrix}$$

The second layer $C: 2 \rightarrow 1$



represents the parametrised function

$$I_C: \mathbb{R}^3 \times \mathbb{R}^2 \longrightarrow \mathbb{R};$$

$$(q, b) \longmapsto \sigma(q_1b_1 + q_2b_2 + q_b).$$

which in turn has update and request functions

$$U_C: \mathbb{R}^3 \times \mathbb{R}^2 \times \mathbb{R} \longrightarrow \mathbb{R}^2;$$

$$(q, b, c) \longmapsto \begin{pmatrix} q_1 - \varepsilon(I_C(q, b) - c)\dot{\sigma}(q_1b_1 + q_2b_2 + q_b)b_1 \\ q_2 - \varepsilon(I_C(q, b) - c)\dot{\sigma}(q_1b_1 + q_2b_2 + q_b)b_2 \\ q_b - \varepsilon(I_C(q, b) - c)\dot{\sigma}(q_1b_1 + q_2b_2 + q_b) \end{pmatrix}$$

$$r_C: \mathbb{R}^3 \times \mathbb{R}^2 \times \mathbb{R} \longrightarrow \mathbb{R}^2;$$

$$(q, b, c) \longmapsto \begin{pmatrix} b_1 - (I_C(q, b) - c)\dot{\sigma}(q_1 b_1 + q_2 b_2 + q_b)q_1 \\ b_2 - (I_C(q, b) - c)\dot{\sigma}(q_1 b_1 + q_2 b_2 + q_b)q_2 \end{pmatrix}$$

Thus the layers map respectively to the learners $(\mathbb{R}^5, I_B, U_B, r_B)$ and $(\mathbb{R}^3, I_C, U_C, r_C)$.

Functoriality says that we may recover U_A and r_A as composites $U_A = U_B * U_C$ and $r_A = r_B * r_C$. For example, we can check this is true for the first coordinate p_{11} :

$$\begin{aligned} U_B * U_C(p, q, a, c)_{11} &= p_{11} - \varepsilon(I(p, a)_1 - s(q, I(p, a), c)_1)\dot{\sigma}(\beta_1)a_1 \\ &= p_{11} - \varepsilon(J(q, I(p, a)) - c)\dot{\sigma}(q_1 I_1(p, a) + q_2 I_2(p, a) + q_b)q_1 \dot{\sigma}(\beta_1)a_1 \\ &= U_A(p, q, a, c)_{11} \end{aligned}$$

In particular, the functoriality describes how to factor the expressions for the entries of U_A and r_A in a way that allows us to parallelise the computation and to efficiently reuse expressions.

7 Discussion

To summarise, in this paper we have developed an algebraic framework to describe composition of supervised learning algorithms. In order to do this, we have identified the notion of a request function as the key distinguishing feature of compositional learning. This request function allows us to construct training data for all sub-parts of a composite learning algorithm from training data for just the input and output of the composite algorithm.

This perspective allows us to carefully articulate the structure of the back-propagation algorithm. In particular, we see that:

- An activation function σ defines a functor from neural network architectures to parametrised functions.
- A step size ε and an error function e define a functor from parametrised functions to supervised learning algorithms.
- The update function for the learning algorithm defined by this functor is specified by gradient descent.
- The request function for the learning algorithm defined by this functor is specified by backpropagation.
- Bimonoid structure in the category of learning algorithms allows us to understand neural nets, including variants such as convolutional ones, as generated from three basic algorithms.

We close this paper by making some observations and asking some questions about neural nets from this perspective; we make no claim to originality, and indeed expect that many of these observations are known to the experts, perhaps in other (non-categorical) terms.

7.1 Generalised networked learning algorithms

The category **Learn** contains many more morphisms than those in the images of **Para** under the gradient descent/backpropagation functors L . Indeed, **Learn** does not require us define our update and request functions using derivatives at all. This shows that we can introduce much more general elements than the usual neural nets into machine learning algorithms, and still use a modular, backpropagation-like method to learn.

What might more general learning algorithms look like? As the input and output spaces need not be Euclidean, we could choose parts of our algorithm to learn functions that are constrained to obey certain symmetries, such as periodicity, or equivalently being defined on a torus. We might also learn nonlinear functions like rotations, or find a way to parametrise over network architectures.

There are of course, obvious advantages of using gradient descent: it gives a heuristic argument that the learning algorithm updates the parameter towards minimising some function, which we might interpret as the error. This helps guide the construction of a neural net. Note, however, that the category **Learn** sees none of this structure; it is all in the functors L . Thus we can use define coherent learning algorithms that vary the choice of error function across the network.

7.2 More general error functions

To apply our main theorem, and hence understand backpropagation as a functor, we require the certain derivatives of our chosen error function be invertible. Unfortunately, many commonly used error functions do not quite obey these conditions. For example, *cross entropy* is an error function that is similar to quadratic error, but often leads to faster convergence. Cross entropy is the error function

$$e(x, y) = y \ln x + (1 - y) \ln(1 - x).$$

This does not supply an example of the main theorem, as the derivative is not defined when $x = 0, 1$.

It is, however, quite close to an example. There are two ways in which the practical method differs from our theory. First, instead of using simply summing the error to arrive at our total error E_I , the usual method of using cross entropy takes the average, giving the function

$$E_I(p, a, b) = \frac{1}{n} \sum_{j=1}^n e(I_j(p, a), b_j)$$

where n is the dimension of the codomain vector space B . This is quite straightforward to model, and we show how to do this by incorporating an extra variable α in our generalisation of the main theorem in [Appendix B](#).

The second is more subtle. When $x \neq 0, 1$, cross entropy has the derivative

$$\frac{\partial e}{\partial x}(z, y) = \frac{y - z}{z(1 - z)}.$$

This is invertible for all $z \neq 0, 1$. In practice, we consider (i) training data (a, b) such that $0 \leq a_i, b_j \leq 1$ for all i, j , as well as (ii) $I(p, a)$ such that this implies $0 < I_k(p, a) < 1$ for all k , assuming we start with a suitable initial parameter p and small enough step size ε . In this case $\frac{\partial e}{\partial x}(z, -)$ is invertible at all relevant points, and so we can define request functions.

Indeed, in this case the request function is

$$r_I(p, a, b)_i = a_i - \frac{|A|}{|B|} a_i (1 - a_i) \sum_j \frac{I_j(p, a) - b_j}{I_j(p, a)(1 - I_j(p, a))} \frac{\partial I_j}{\partial a_i}(p, a),$$

while the update function is the standard update rule for gradient descent with respect to the cross entropy.

$$U_I(p, a, b)_k = p_k - \varepsilon \sum_j \frac{I_j(p, a) - b_j}{I_j(p, a)(1 - I_j(p, a))} \frac{\partial I_j}{\partial p_k}(p, a).$$

There is work to be done in generalising the conditions main theorem to accommodate error functions such as cross entropy that fail to have derivatives at some isolated points.

Regardless, note that while in this case it is not straightforward to state backpropagation as a functor from **Para**, this analysis thus nevertheless still sheds light on the compositional nature of the learning algorithm.

7.3 Richer compositional structure on neural nets

It has suited our purposes for this paper to simply consider the category **NNet** of neural networks. On the other hand, we spent some time discussing the monoidal and bimonoid structure on the categories **Para** of parametrised functions and **Learn** of learning algorithms. Moreover, the neural networks intuitively do have both monoidal and bimonoid structure: we can place networks side by side to represent two networks run in parallel, and we can add multiple inputs and duplicate outputs to each node in a neural network as we like.

In fact, the category **NNet** can be generalised to a symmetric monoidal category with bimonoids on each object. This generalisation is the strict symmetric monoidal category **IDAG** of *idags*—interfaced directed acyclic graphs—which has been previously studied as an important structure in concurrency, as well as for its elegant categorical properties [3].

It is also desirable that each functor $I: \mathbf{NNet} \rightarrow \mathbf{Para}$ implementing neural networks as parametrised functions factors as $\mathbf{NNet} \rightarrow \mathbf{IDAG} \rightarrow \mathbf{Para}$, and indeed this can be done. Moreover, the factor $\mathbf{IDAG} \rightarrow \mathbf{Para}$ is a symmetric monoidal functor that preserves bimonoid structures.

7.4 A bicategory of learners

At present, approaches to tuning hyperparameters of a neural network are rather ad hoc. One such hyperparameter is the architecture of the network itself. How many layers does the optimal neural net for a given problem have, and how many nodes should be in each layer?

A bicategory is a generalisation of a category in which there also exist two-dimensional morphisms connecting the ordinary morphisms. Learning algorithms naturally form a bicategory. Indeed, our definition of equivalence class for learning algorithms implicitly uses a notion of morphism between learning algorithms; an equivalence class is just an isomorphism class for the following notion of 2-morphism.

Definition 7.1. *A 2-morphism between learning algorithms is a map between their parameter spaces such that the relevant diagrams commute. That is, suppose we have learners (P, I, U, r) and $(Q, J, V, s): A \rightarrow B$. A 2-morphism $f: (P, I, U, r) \rightarrow (Q, J, V, s)$ is a function $f: P \rightarrow Q$ such that*

$$\begin{aligned} J(f(p), a) &= I(p, a), \\ V(f(p), a, b) &= f(U(p, a, b)), \\ s(f(p), a, b) &= r(p, a, b). \end{aligned}$$

Composition of 2-morphisms between learners is simply given by composition of functions.

Similarly, Para and IDAG are also naturally bicategories. Working in this bicategorical setting gives language for relating different parametrised functions and neural network architectures. Such higher morphisms can encode ideas such as structured expansion of networks, by adding additional neurons or layers. The gradient descent functor restricts to what is called a bifunctor on a certain sub-bicategory of Para, and it would be worthwhile to check whether the functor $\text{IDAG} \rightarrow \text{Para}$ lands in this sub-bicategory.

7.5 Further directions

Traces and recurrent neural networks In monoidal categories, a structure known as a trace often describes the structure of processes that involve feed-back. Does such a structure exist on Learn, and do recurrent neural nets make use of it?

Geometry We defined the category Para of parametrised functions to have Euclidean spaces as objects. It is straightforward to generalise this to a category where the objects are more general manifolds equipped with some differentiable structure. Moreover, deep learning on such structures is an active field of study

[5]. Can we incorporate such work into this categorical setting, viewing a generalised version of gradient descent and backpropagation as defining a functor from this more general category to `Learn`?

Success guarantees While we have defined a structure that models compositional supervised learning algorithms, we have placed no requirements that a learning algorithm converge to any close approximation of a function f when given enough training pairs $(a, f(a))$. If we select training data from a distribution and integrate the error over that distribution, does learning improve the result? Is anything of this sort compositional?

Bibliography

- [1] J. C. Baez, B. Fong, and B. Pollard. A compositional framework for Markov processes. *Journal of Mathematical Physics*, 57(3):033301, 2016. [doi:10.1063/1.4941578](https://doi.org/10.1063/1.4941578)
- [2] B. Coecke and A. Kissinger. *Picturing Quantum Processes A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge University Press, 2017.
- [3] M. Fiore and M. Devesas Campos. The Algebra of Directed Acyclic Graphs. In: B. Coecke, L. Ong, P. Panangaden (eds) *Computation, Logic, Games, and Quantum Foundations. The Many Facets of Samson Abramsky*. Lecture Notes in Computer Science, vol 7860. Springer, Berlin, Heidelberg. Available as [arXiv:1303.0376](https://arxiv.org/abs/1303.0376).
- [4] B. Fong. *The Algebra of Open and Interconnected Systems*. DPhil thesis, University of Oxford, 2016. [arXiv:1609.05382](https://arxiv.org/abs/1609.05382).
- [5] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. [doi:10.1109/MSP.2017.2693418](https://doi.org/10.1109/MSP.2017.2693418).
- [6] A. Joyal and R. Street. The geometry of tensor calculus I. *Advances in Mathematics* 88(1):55–112, 1991. [doi:10.1016/0001-8708\(91\)90003-P](https://doi.org/10.1016/0001-8708(91)90003-P).
- [7] S. Mac Lane. *Categories for the Working Mathematician*, Springer, Berlin, 1998.
- [8] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.5188–5196, 2015.
- [9] M. A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. <http://neuralnetworksanddeeplearning.com>

- [10] D. Vagner, D. Spivak. E. Lerman. Algebras of open dynamical systems on the operad of wiring diagrams. [arXiv:1408.1598](https://arxiv.org/abs/1408.1598).

A Proof of Proposition 2.4

Proof of Proposition 2.4. This follows from routine checking of the axioms; we say a few words about each case.

Identities. The identity axioms are easily checked. For example, to check identity on the left we see that $(P, I, U, r) * (1, \text{id}, !, \pi_2)$ is given by $P \times 1 \cong P$, $I(p, \text{id}(a)) = I(p, a)$, $U * !(p, a, b) = U(p, a, \pi_2(I(p, a), b)) = U(p, a, b)$, and $r * \pi_2(p, a, b) = r(p, a, \pi_2(I(p, a), b)) = r(p, a, b)$.

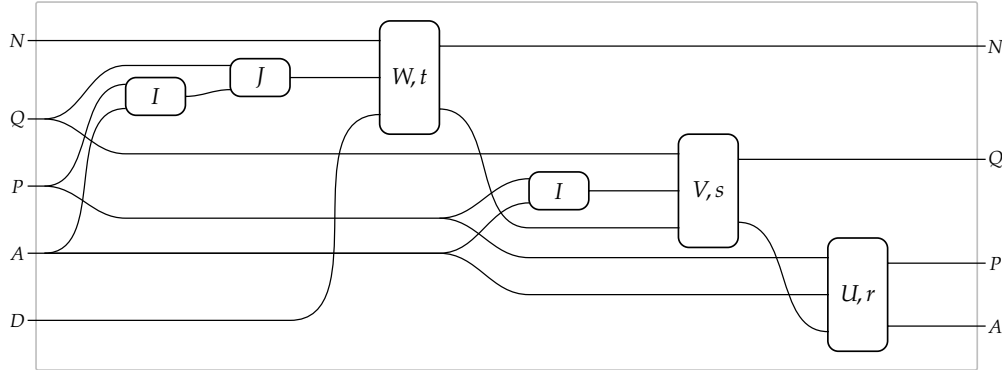
Associativity. Note that the associativity axiom is what requires that our morphisms in **Learn** be *equivalence classes* of learners, and not simply learners themselves: composition of learners is not associative on the nose. Indeed, this is because products of sets are not associative on the nose: we only have isomorphisms $(P \times Q) \times N \cong P \times (Q \times N)$ of sets, not equality. Acknowledging this, associativity is straightforward to prove.

Let $(P, I, U, r): A \rightarrow B$, $(Q, J, V, s): B \rightarrow C$, and $(N, K, W, t): C \rightarrow D$ be learners. The most involved item to check is the associativity of the paired update–request function. Computation shows

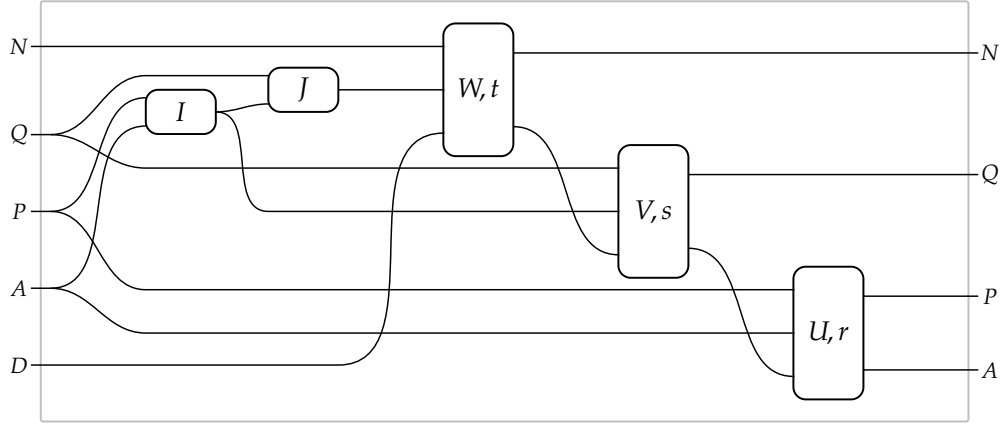
$$\begin{aligned} (U * V) * W &= \left(U(p, a, s(q, I(p, a), \gamma)), V(q, I(p, a), \gamma), W(n, J(q, I(p, a)), d) \right) \\ &= U * (V * W) \end{aligned}$$

where $\gamma = t(n, J(q, I(p, a)), d)$.

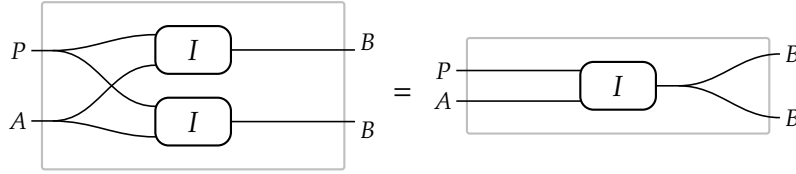
This equality is easier to parse using string diagrams. The composite $(U * V) * W$ is given by the diagram



while the composite $U * (V * W)$ is given by



To prove these two diagrams represent the same function, observe that the function $(I(p, a), I(p, a)): P \times A \rightarrow B \times B$ can be drawn in the following two ways:



This equality, and the associativity of the diagonal map, implies the equality of the previous two diagrams, and hence the associativity of the update and request composites.

Monoidality. It is straightforward to check the above is a monoidal product, with unit given by \emptyset .

Indeed, note that we have now shown that **Learn** is a category. There exists a functor from the category **Set** of sets and functions to **Learn**. This functor maps each set to itself, and each function $f: A \rightarrow B$ to the trivially parametrised function $\bar{f}: 1 \times A \rightarrow B$. Note that (\mathbf{Set}, \times) is a monoidal category, and let α , ρ , and λ respectively denote the associator, right unitor, and left unitor for (\mathbf{Set}, \times) . The images of these maps under this trivial parametrisation functor $(\bar{\cdot})$, written $\bar{\alpha}$, $\bar{\rho}$, and $\bar{\lambda}$, are the corresponding structure maps for $(\mathbf{Learn}, ||)$ as a symmetric monoidal category.

The naturality of these maps, as well as the axioms of a symmetric monoidal category, then follow in a straightforward way from the corresponding facts in **Set**.

Thus we have defined a symmetric monoidal category. □

B Proof of Theorem 3.2

Theorem B.1 (Generalisation of Theorem 3.2). *Fix $\varepsilon > 0$, $\alpha: \mathbb{N} \rightarrow \mathbb{R}_{>0}$, and $e(x, y): \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ differentiable such that $\frac{\partial e}{\partial x}(z, -): \mathbb{R} \rightarrow \mathbb{R}$ is invertible for each $z \in \mathbb{R}$.*

Then we can define a faithful, injective-on-objects, symmetric monoidal functor

$$L: \text{Para} \longrightarrow \text{Learn}$$

that sends each parametrised function $I: P \times A \rightarrow B$ to the learner (P, I, U_I, r_I) defined by

$$U_I(p, a, b) := p - \varepsilon \nabla_p E_I(p, a, b)$$

and

$$r_I(p, a, b) := f_a \left(\frac{1}{\alpha_B} \nabla_a E_I(p, a, b) \right),$$

where f_a is component-wise application of the inverse to $\frac{\partial e}{\partial x}(a_i, -)$ for each i , and

$$E_I(p, a, b) := \alpha_B \sum_j e(I_j(p, a), b_j).$$

Proof. The map is by definition injective-on-objects. Since I maps to (P, I, U_I, r_I) , the map is injective on morphisms, and hence will give a faithful functor.

Let $I: P \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $J: Q \times \mathbb{R}^m \rightarrow \mathbb{R}^\ell$ be parametrised functions. We show that the composite of their images is equal to the image of their composite.

Update functions. By definition the composite of the update functions of I and J is given by

$$\begin{aligned} (U_I * U_J)(p, q, a, c) &= (U_I(p, a, r_J(q, I(p, a), c)), U_J(q, I(p, a), c)) \\ &= (p - \varepsilon \nabla_p E_I(p, a, r_J(q, I(p, a), c)), q - \varepsilon \nabla_q E_J(q, I(p, a), c)), \end{aligned}$$

while the update function of the composite $I * J$ is given by

$$U_{I*J}(p, q, a, c) = (p - \varepsilon \nabla_p E_{I*J}(p, q, a, c), q - \varepsilon \nabla_q E_{I*J}(p, q, a, c)).$$

To show that these are equal, we thus must show that the following equations hold

$$\nabla_p E_I(p, a, r_J(q, I(p, a), c)) = \nabla_p E_{I*J}(p, q, a, c) \quad (1)$$

$$\nabla_q E_J(q, I(p, a), c) = \nabla_q E_{I*J}(p, q, a, c). \quad (2)$$

We first consider Equation 1:

$$\begin{aligned}
& \nabla_p E_I(p, a, r_J(q, I(p, a), c)) \\
&= \nabla_p \alpha_B \sum_i e(I_i(p, a), r_J(q, I(p, a), c)_i) \quad (\text{def } E_I) \\
&= \left(\alpha_B \sum_i \frac{\partial e}{\partial x}(I_i(p, a), r_J(q, I(p, a), c)_i) \frac{\partial I_i}{\partial p_\ell}(p, a) \right)_\ell \quad (\text{def } \nabla_p) \\
&= \left(\alpha_B \sum_i \frac{\partial e}{\partial x} \left(I_i(p, a), f_{I(p, a)} \left(\frac{1}{\alpha_B} \nabla_b E_J(q, I(p, a), c)_i \right) \right) \frac{\partial I_i}{\partial p_\ell}(p, a) \right)_\ell \quad (\text{def } r_J) \\
&= \left(\alpha_B \sum_i \frac{1}{\alpha_B} \nabla_b E_J(q, I(p, a), c)_i \frac{\partial I_i}{\partial p_\ell}(p, a) \right)_\ell \quad (\text{def } f) \\
&= \left(\sum_i \alpha_C \sum_j \frac{\partial e}{\partial x}(J_j(q, I(p, a)), c_j) \frac{\partial J_j}{\partial b_i}(q, I(p, a)) \frac{\partial I_i}{\partial p_\ell}(p, a) \right)_\ell \quad (\text{def } E_J) \\
&= \left(\alpha_C \sum_j \frac{\partial e}{\partial x}(J_j(q, I(p, a)), c_j) \frac{\partial J_j}{\partial p_\ell}(q, I(p, a)) \right)_\ell \quad (\text{chain rule}) \\
&= \nabla_p \alpha_C \sum_j e(J_j(q, I(p, a)), c_j) \quad (\text{def } \nabla_p) \\
&= \nabla_p E_{I*J}(p, q, a, c) \quad (\text{def } E_{J\bullet})
\end{aligned}$$

note the shift to coordinate-wise reasoning, that f is defined as the inverse to ∂e , and the use of the chain rule.

Equation 2 simply follows from the definition of the error function; we need not even take derivatives:

$$E_J(q, I(p, a), c) = \alpha \sum_i e(J(q, I(p, a))_i, c_i) = E_{I*J}(p, q, a, c)_m$$

Thus we have shown $(U_I * U_J)(p, q, a, c) = U_{I*J}$, as desired.

Request functions. We must prove that the following equation holds:

$$r_I(p, a, r_J(q, I(p, a), c)) = r_{I*J}(p, q, a, c)$$

This follows due to the chain rule, in the exact same manner as for updating p , but swapping the roles of a and p in the proof of Equation 1:

$$\begin{aligned}
r_I(p, a, r_J(q, I(p, a), c)) &= f_a \left(\frac{1}{\alpha_B} \nabla_a E_I(p, a, r_J(q, I(p, a), c)) \right) \\
&= f_a \left(\frac{1}{\alpha_B} \nabla_a E_{I*J}(p, q, a, c) \right) \\
&= r_{I*J}(p, q, a, c)
\end{aligned}$$

Identities. The identity on the object A in the category of parametrised functions is the projection $\text{id}_A: \mathbf{1} \times A \rightarrow A$, where the parameter space $\mathbf{1}$ comprises just

a single point. The image of id_A has trivial update function, since the parameter space is trivial. The request function is given by

$$r_{\text{id}_A}(\mathbf{1}, a, b) = f_a\left(\frac{1}{\alpha_B} \nabla_a (\alpha_B \sum_i e(a_i, b_i))\right) = \left(f_a\left(\frac{\partial e}{\partial a_i}(a_i, b_i)\right)\right)_i = b.$$

This is exactly the identity map $(\mathbf{1}, \text{id}_A, !, \pi_2)$ in **Learn**.

Monoidal structure. The map is a monoidal functor. That is, the learner given by the monoidal product of parametrised functions is equal to the monoidal product of the learners given by those same functions, up to the standard isomorphisms $\mathbb{R}^n \times \mathbb{R}^m \cong \mathbb{R}^{n+m}$. To see that this is true, suppose we have parametrised functions $I: P \times A \rightarrow B$ and $J: Q \times C \rightarrow D$. Their tensor is $I \otimes J: (P \times Q) \times (A \times C) \rightarrow B \times D$. Note that $E_{I \otimes J}(p, q, a, c, b, d) = E_I(p, a, b) + E_J(q, c, d)$. Thus the update function of their tensor is given by

$$\begin{aligned} & U_{I \otimes J}(p, q, a, c, b, d) \\ &= (p - \varepsilon \nabla_p E_{I \otimes J}(p, q, a, c, b, d), q - \varepsilon \nabla_q E_{I \otimes J}(p, q, a, c, b, d)) \\ &= (p - \varepsilon \nabla_p E_I(p, a, b), q - \varepsilon \nabla_q E_J(q, c, d)) \\ &= (U_I(p, a, b), U_J(q, c, d)) \end{aligned}$$

and similarly the request function is

$$\begin{aligned} & r_{I \otimes J}(p, q, a, c, b, d) \\ &= f_{(a,c)}\left(\frac{1}{\alpha_{B \times D}} \nabla_{(a,c)} E_{I \otimes J}(p, q, a, c, b, d)\right) \\ &= f_{(a,c)}\left(\frac{1}{\alpha_{B \times D}} \nabla_{(a,c)} \alpha_{B \times D} \left(\sum_i e(I(p, a)_i, b_i) + \sum_j e(J(q, c)_j, d_j)\right)\right) \\ &= \left(f_a\left(\frac{1}{\alpha_B} \nabla_a E_I(p, a, b)\right), f_c\left(\frac{1}{\alpha_D} \nabla_c E_J(q, c, d)\right)\right) \\ &= (r_I(p, a, b), r_J(q, c, d)) \end{aligned}$$

Thus image of the tensor is the tensor of the image. □

C Background on category theory

C.1 Symmetric monoidal categories

A symmetric monoidal category is a setting for composition for network-style diagrammatic languages like neural networks. A prop is a particularly simple sort of strict symmetric monoidal category.

First, let us define a category. We specify a **category** **C** using the data:

- a collection X whose elements are called *objects*.

- for every pair (A, B) of objects, a set $[A, B]$ whose elements are called *morphisms*.
- for every triple (A, B, C) of objects, a function $[A, B] \times [B, C] \rightarrow [A, C]$ call the *a composition rule*, and where we write $(f, g) \mapsto f; g$.

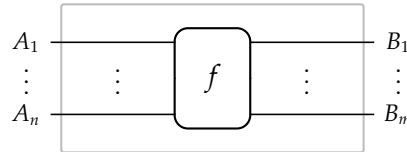
This data is subject to the axioms

- identity: for all objects A there exists $\text{id}_A \in [A, A]$ such that for all $f \in [A, B]$ and $g \in [B, A]$ we have $\text{id}_A; f = f$ and $g; \text{id}_A = g$.
- associativity: for all $f \in [A, B]$, $g \in [B, C]$ and $h \in [C, D]$ we have $(f; g); h = f; (g; h)$.

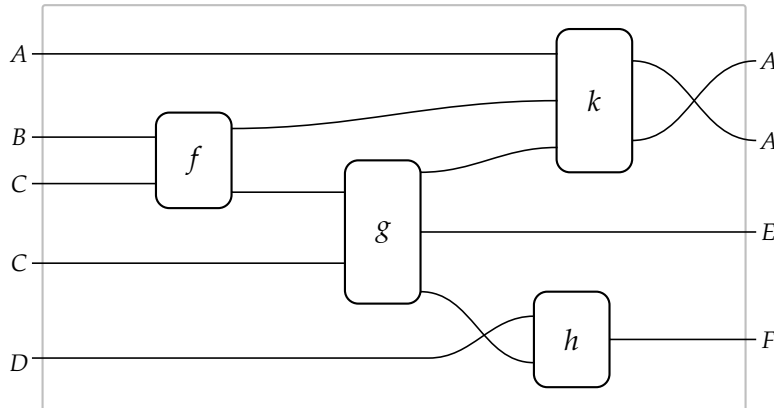
The main object of our interest, however, is a particular type of category, known as a symmetric monoidal category. For a **symmetric monoidal category** \mathcal{C} , we further require the data:

- for every pair (A, B) of objects, another object $A \otimes B$ in X .
- for every quadruple (A, B, C, D) of objects a function $[A, B] \times [C, D] \rightarrow [A \otimes C, B \otimes D]$ called the *monoidal product*.

Using this data, we may draw networks. We think of the objects as being various types of wire, and a morphism f in $[A_1 \otimes \cdots \otimes A_n, B_1 \otimes \cdots \otimes B_m]$ as a box with wires of types A_i on the left and wire of types B_i on the right. Here are some pictures.



By connecting wires of the same type, we can draw more complicated pictures. For example:



The key point of a network is that any such picture must have an unambiguous interpretation as a morphism. The use of string diagrams to represent morphisms in a monoidal category is formalised in [6].

To form what is known as a **strict** symmetric monoidal category, the above data must obey additional axioms that ensure it captures the above intuition of behaving like a network. These axioms are

- interchange: for all $f \in [A, B]$, $g \in [B, C]$, $h \in [D, E]$, $k \in [E, F]$ we have $(f; g) \otimes (h; k) = (f \otimes h); (g \otimes k)$.
- monoidal identity: there exists an object I such that $I \otimes A = A = A \otimes I$.
- monoidal associativity: for all objects A, B, C we have $(A \otimes B) \otimes C = A \otimes (B \otimes C)$.
- symmetry: for all pairs of objects A, B we have morphisms $\sigma_{A,B} \in [A \otimes B, B \otimes A]$ such that $\sigma_{A,B}; \sigma_{B,A} = \text{id}_{A \otimes B}$, and that for all $f \in [A, C]$, $g \in [B, D]$ we have $(f \otimes g); \sigma_{C \otimes D} = \sigma_{A \otimes B}; (f \otimes g)$.

More generally, symmetric monoidal categories require these axioms only to be true up to natural isomorphism. More detail can be found in [7].

Example C.1. An example of a symmetric monoidal category is (Set, \times) , where our objects are a set of each cardinality, and morphisms are functions between them. The monoidal product is given by the cartesian product of sets.

Example C.2. Another example of a symmetric monoidal category is (FVect, \oplus) , where our objects are finite-dimensional vector spaces, morphisms are linear maps, and the monoidal product is given by the direct sum of vector spaces.

C.2 Functors

A functor is a way of reinterpreting one category in another, preserving the algebraic structure. In other words, a functor is the notion of structure preserving map for categories, in analogy with linear transformations as the structure preserving maps for vector spaces, and group homomorphisms as the structure preserving maps for groups.

Formally, given categories \mathbf{C}, \mathbf{D} , a **functor** $F: \mathbf{C} \rightarrow \mathbf{D}$ sends every object A of \mathbf{C} to an object FA of \mathbf{D} , every morphism $f: A \rightarrow B$ in \mathbf{C} to a morphism $Ff: FA \rightarrow FB$ in \mathbf{D} , such that $F1 = 1$ and $Ff; Fg = F(f; g)$.

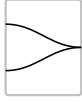
A functor between symmetric monoidal categories is a **symmetric monoidal functor** if $FI_{\mathbf{C}} = I_{\mathbf{D}}$, where I is the monoidal unit for the relevant category, and if there exist isomorphisms $F(A \otimes B) \cong FA \otimes FB$ natural in objects A, B of \mathbf{C} . We say that the functor is a **strict** symmetric monoidal functor if these isomorphisms are in fact equalities.

We also say that a functor is **faithful** if $Ff = Fg$ only when $f = g$, and **injective-on-objects** if the map from objects of \mathbf{C} to objects of \mathbf{D} is injective.

C.3 Bimonoids

A **bimonoid** in a symmetric monoidal category is an object A together with morphisms that obey certain axioms. These morphisms have names and types:

multiplication



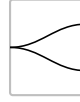
$$\mu: A \otimes A \rightarrow A$$

unit



$$\epsilon: I \rightarrow A$$

comultiplication



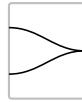
$$\delta: A \rightarrow A \otimes A$$

counit

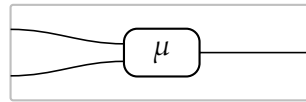


$$\nu: A \rightarrow I$$

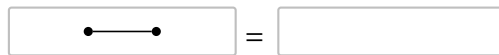
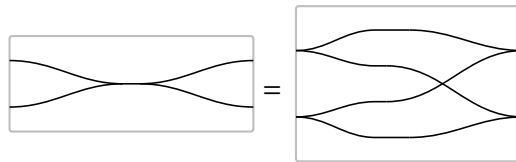
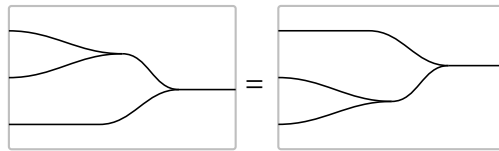
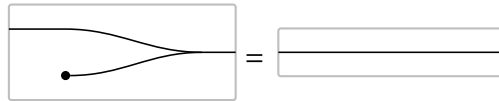
Note that these diagrams are informal, but useful, special depictions of these morphisms. More formally, for example, the diagram



for the multiplication μ is a shorthand for the string diagram

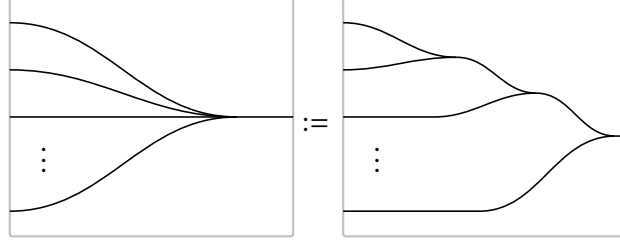


These morphisms must obey the axioms:



and their mirror images.

Note that the second axiom above, called associativity, implies that all maps with codomain 1 constructed using only products of the multiplication and the identity map are equal. It is thus convenient, and does not cause confusion, to define the following notation:



We define the mirror image notation for iterated comultiplications.

These morphisms, and the axioms they obey, allow network diagrams to be drawn. First, the morphisms μ , ϵ , δ , and ν respectively give interpretations to pairwise merging, initializing, splitting in pairs, and deleting edges. The associativity and coassociativity axioms, for example, then give unique interpretation to n -ary merging and n -ary splitting, as described above.

Example C.3. Each object in \mathbf{FVect} can be equipped with the structure of a bimonoid. Indeed, given a vector spaces V , the multiplication $\mu: V \oplus V \rightarrow V$ takes a pair (u, v) to $u + v$, the unit $\epsilon: 0 \rightarrow V$ maps the unique element 0 of the 0-dimensional vector space to the zero vector in V , the comultiplication $\delta: V \rightarrow V \oplus V$ maps v to (v, v) , and the counit $\nu: V \rightarrow 0$ maps every vector $v \in V$ to zero. It is standard linear algebra to check that these maps obey the bimonoid axioms.