

Backprop as Functor: A compositional perspective on supervised learning

Brendan Fong

David Spivak

Rémy Tuyéras

Department of Mathematics,
Massachusetts Institute of Technology

Computer Science and Artificial Intelligence Lab,
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, giving a broad generalisation of neural networks and linking it with structures from bidirectional programming and open games.

I. INTRODUCTION

Machine learning, and in particular the use of neural networks, has rapidly become remarkably effective at real world tasks [18]. 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 circuit diagrams, Markov processes, quantum computation, and dynamical systems [8], [1], [6], [19]. 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.

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

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 .

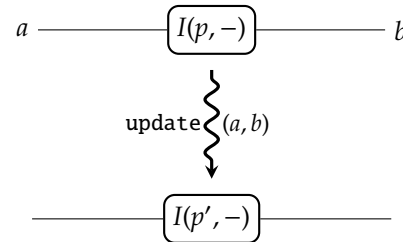


Figure 1. Given a training datum (a, b) , a learning algorithm updates p to p' .

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 for the composite, 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 $b' := s(q, b'', c'')$. Now we have the desired training data (a, b') for I . The request function is thus a way of ‘backpropagating’ the output back toward the earlier learners in a network.

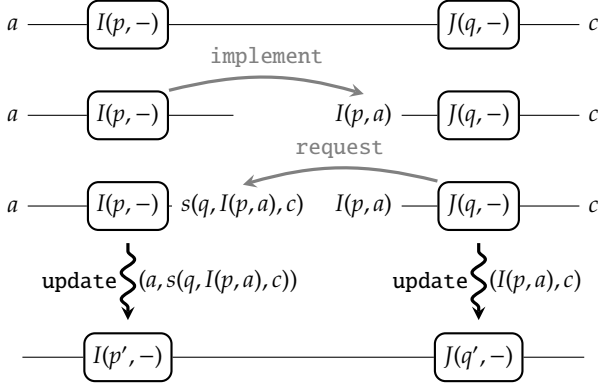


Figure 2. A request function allows an update function to be defined for the composite $J(q, I(p, -))$.

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

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 composi-

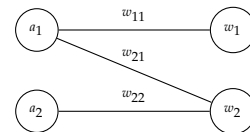
tion. More precisely, a category consists of *types* A, B, C , and so on, *morphisms* $f : A \rightarrow B$ between these types, and a *composition rule* by which morphisms $f : A \rightarrow B$ and $g : B \rightarrow C$ can be combined to create a morphism $A \rightarrow C$. Thus we can say that learning 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.

This novel category **Learn**, synthesised from the above analysis of learning algorithms, nonetheless curiously resembles and is closely related to lenses [3] and open games [11], two well-known structures that also model compositional, bidirectional exchange information between interacting systems. We return to this briefly in Section VII.

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 on the right is connected to node j on the left.

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 $I : \mathbb{R}^{|C|+n_2} \times \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$. The $\mathbb{R}^{|C|}$ factor encodes numbers called *weights* and the \mathbb{R}^{n_2} factor encodes numbers called the *biases*. For example, the layer $C = \{(1, 1), (2, 1), (2, 2)\} \subseteq [2] \times [2]$, has $n_2 = 2$ biases and $|C| = 3$ weights. The biases are represented by the right hand nodes below, while the weights are represented by the edges:



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

\mathbb{R}^2 , given by

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

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 $P \times \mathbb{R}^{n_0} \rightarrow \mathbb{R}^{n_k}$ for some P . 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 $I: P \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ between finite dimensional Euclidean spaces, where the parameter space $P = \mathbb{R}^p$ is also Euclidean.

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. More precisely, we have the following:

Theorem. Fix $\varepsilon > 0$ and $e(x, y): \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ such that $\frac{\partial e}{\partial x}(x_0, -): \mathbb{R} \rightarrow \mathbb{R}$ is invertible for each $x_0 \in \mathbb{R}$. Then there is a faithful, injective-on-objects, symmetric monoidal functor

$$L_{\varepsilon, e}: \text{Para} \longrightarrow \text{Learn}$$

sending each differentiable 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

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 composites of linear functions and activation functions.

Overview

In Section II, we define the category Learn of learning algorithms. We present the main theorem in Section III: 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 IV, we broaden this view to show how it relates to neural networks. Next, in Section V, 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 (§VI), and discuss implications for this framework (§VII). The extended version [10] of this article provides appendices with more technical aspects of the proof of the main theorem, and a brief, diagram-driven introduction to relevant topics in category theory.

II. THE CATEGORY OF LEARNERS

In this section we define a symmetric monoidal category Learn that models supervised learning algorithms and their composites. See extended version [10] for background on categories and string diagrams.

Definition II.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:

$$I: P \times A \rightarrow B, \\ U: P \times A \times B \rightarrow P, \\ r: P \times A \times B \rightarrow A.$$

We call P the *parameter space*; it is just a set. The map I implements a parameter value $p \in P$ as a function

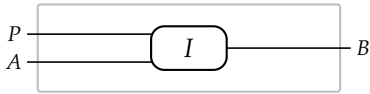
$I(p, -): A \rightarrow B$. We think of a pair $(a, b) \in A \times B$ as a *training datum*; it pairs an input a with an output b . The map $U: P \times A \times B \rightarrow P$ is the *update function*; given a ‘current’ parameter p and a training datum $(a, b) \in A \times B$, it produces an ‘updated’ parameter $U(p, a, b) \in P$. This can be thought of as the learning step. The idea is that the updated function $I(U(p, a, b), -): A \rightarrow B$ 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 that this value will be sent to upstream learners for their own training.

Remark II.2. The request function is perhaps a little mysterious at this stage. Indeed, it is superfluous to the definition of a stand-alone 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 $I(p, a)$ is. 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 [17].

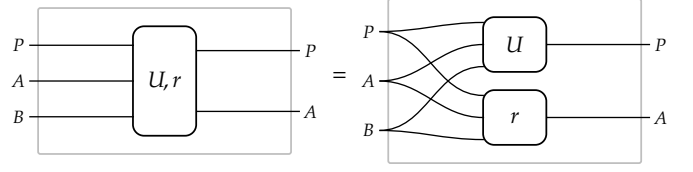
A third way to understand the request function is by analogy with other compositional structures: the request function plays an analogous role to the put function in an asymmetric lens [3] and the coplay function in an open game [11].

Remark II.3. Using string diagrams¹ in (\mathbf{Set}, \times) , we can draw an implementation function I as follows:



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 $f: P \rightarrow P'$ such that the following hold for each $p \in P$, $a \in A$, and $b \in B$:

$$I'(f(p), a) = I(p, a),$$

$$U'(f(p), a, b) = f(U(p, a, b)),$$

$$r'(f(p), a, b) = r(p, a, b).$$

In fact, a stronger notion of equivalence—the equivalence relation generated by the existence of a surjection f with these properties—also makes semantic sense, but we use this definition as it gives rise to faithfulness in the main theorem (Theorem III.2).

Proposition II.4. *There exists a symmetric monoidal category \mathbf{Learn} whose objects are sets and whose morphisms are equivalence classes of learners.*

The proof of Proposition II.4 is given in Appendix A of the extended version [10]. For now, we simply specify the composition, identities, monoidal product, and braiding for this symmetric monoidal category. Note that although we write in terms of representatives, each of these is well defined, respecting the equivalence relation on learners.

a) *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

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

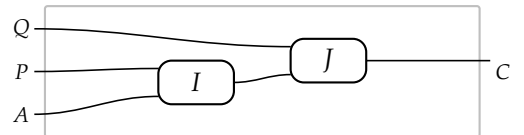
the update function is

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

and the request function

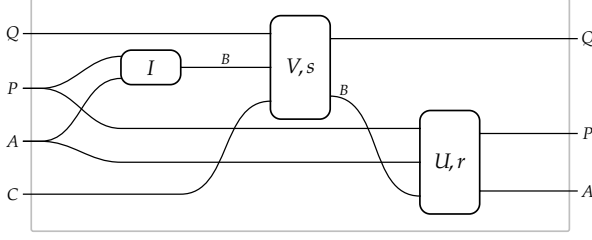
$$(r * s)(p, q, a, c) := r(p, a, s(q, I(p, a), c)).$$

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



¹String diagrams are an alternative, but nonetheless still formal, syntax for morphisms in a monoidal category. See extended version [10] for more details.

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. To repeat the intuition from the introduction, suppose given current parameters $p \in P$ and $q \in Q$ and training data $a \in A$ and $c \in C$. I takes p and a and produces some $b \in B$ for training the second component. Along with q and c , b is used to compute an updated parameter q' together with a value b' for training the first component. Along with p and a , b' is used to compute an updated parameter p' together with a value a' .

b) *Identities*: For each object A , we have the identity map

$$(\mathbb{R}^0, \text{id}, !, \pi_2): A \longrightarrow A,$$

where $\text{id}: \mathbb{R}^0 \times A \rightarrow A$ is the second projection (as this is a bijection, we abuse notation to write this projection as id), $!: \mathbb{R}^0 \times A \times A \rightarrow \mathbb{R}^0$ is the unique function, and $\pi_2: \mathbb{R}^0 \times A \times A \rightarrow A$ is the projection onto the final factor (again, ignoring the \mathbb{R}^0).

c) *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 $(P \times Q, I \parallel J, U \parallel V, r \parallel s)$, where the implementation function is

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

the update function is

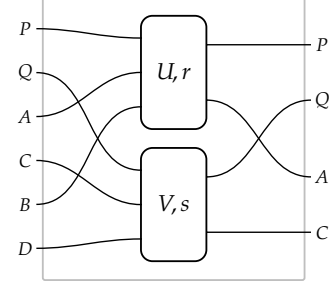
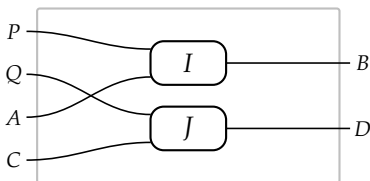
$$(U \parallel V)(p, q, a, c, b, d) := (U(p, a, b), V(q, c, d))$$

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:



d) *Braiding*: A symmetric braiding $A \times B \rightarrow B \times A$ is given by $(\mathbb{R}^0, \sigma, !, \sigma \circ \pi)$ where $\sigma: A \times B \rightarrow B \times A$ is the usual swap function $(a, b) \mapsto (b, a)$ and $\pi: \mathbb{R}^0 \times (A \times B) \times (B \times A) \rightarrow B \times A$ is again the projection onto the final factor.

A proof that this is a well-defined symmetric monoidal category can be found in the extended version [10].

III. GRADIENT DESCENT AND BACKPROPAGATION

In this section we show that gradient descent and backpropagation define a strong symmetric monoidal functor from a symmetric monoidal category Para , 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* $A \rightarrow B$ to be a pair (P, I) , where P is a Euclidean space and $I: P \times A \rightarrow B$ is a differentiable function. We call two such pairs (P, I) , (P', I') equivalent if there exists a differentiable bijection $f: P \rightarrow P'$ such that for all $p \in P$ and $a \in A$ we have $I'(f(p), a) = I(p, a)$. Differentiable parametrised functions between Euclidean spaces form a symmetric monoidal category.

Definition III.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 $(P, I): \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $(Q, J): \mathbb{R}^m \rightarrow \mathbb{R}^\ell$ is given by $(P \times Q, I * J)$ where

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

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 $(P, I): \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $(Q, J): \mathbb{R}^m \rightarrow \mathbb{R}^\ell$ is given by $(P \times Q, I \parallel J)$ where

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

The braiding $\mathbb{R}^n \parallel \mathbb{R}^m \rightarrow \mathbb{R}^m \parallel \mathbb{R}^n$ is given by (\mathbb{R}^0, σ) where $\sigma(a, b) = (b, a)$.

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 III.2. Fix a real number $\varepsilon > 0$ and $e(x, y): \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ differentiable such that $\frac{\partial e}{\partial x}(x_0, -): \mathbb{R} \rightarrow \mathbb{R}$ is invertible for each $x_0 \in \mathbb{R}$. Then we can define a faithful, injective-on-objects, strong symmetric monoidal functor

$$L_{\varepsilon, e}: \text{Para} \longrightarrow \text{Learn}$$

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 $E_I(p, a, b) := \sum_j e(I_j(p, a), b_j)$, 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 the extended version [10]. \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 $L_{\varepsilon, e}$, so named because they turn parametrised functions into a learning algorithms, the *gradient descent/backpropagation functors*.

Remark III.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 the gradient of the total error with respect to the input a , as modified by the invertible function f_a . To pick an example, the functoriality of $L_{\varepsilon, e}$ 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 III.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 $\frac{\partial e}{\partial x}(x_0, -)$ is the function $y \mapsto x_0 - y$. This function is its own inverse, so we have $f_{x_0}(y) := x_0 - y$.

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

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

and similarly

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

Thus given this choice of error function, the functor $L_{\varepsilon, e}$ of Theorem III.2 just implements, as update function, the usual gradient descent with step size ε with respect to the quadratic error.

Remark III.5. Comparing the requests r_I to the updates U_I in Example III.4, one may notice that they are similar, except that the former seem to be missing the ε . One might wonder why the two are different or where the ε factor has gone.

Theorem III.2 shows, however, that in fact the similarity between r_I and U_I is something of a coincidence. What is important about requests, and hence the messages passed backward in backpropagation, is the fact that they are constructed by inverting certain partial derivatives which are then applied to the gradient of the total error with respect to the input. We interpret the result as a ‘corrected’ input value that, if used instead, 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 III.4 is just an artifact of the choice of quadratic error function $e(x, y) := \frac{1}{2}(x - y)^2$.

IV. FROM NETWORKS TO PARAMETRISED FUNCTIONS

In the previous section we showed that gradient descent and backpropagation—for a given choice of error function and step size—define a functor from differentiable 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 from network architectures.

This method, like backpropagation itself, is also compositional—it respects the gluing together of neural networks. To formalise this, we first define a category \mathbf{NNet} of neural networks. Implementation of each neural net will then define a parametrised function, and in fact we get a functor $I: \mathbf{NNet} \rightarrow \mathbf{Para}$. Note that just as defining a gradient descent/backpropagation functor depends on a choice (namely, of error function and step size), so too does defining I . Namely, we must choose an activation function.

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

Given a neural network of type (m, n) and a neural network of type (n, p) we may concatenate them to get a neural network of type (m, p) . Note that when $m = n$, 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 IV.1. *The category \mathbf{NNet} of neural networks has as objects natural numbers and as morphisms $m \rightarrow n$ neural networks of type (m, n) . Composition is given by concatenation of neural networks. The identity morphism on n is the 0-layer neural network.*

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

Proposition IV.2. *Given a differentiable function $\sigma: \mathbb{R} \rightarrow \mathbb{R}$, we have a functor*

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

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

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

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

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

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

$$I_N^\sigma = I_{C_1}^\sigma * \dots * I_{C_k}^\sigma$$

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

Proof. The proof of this proposition is straightforward. Note in particular that the image I_C^σ 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^σ with a gradient descent/backpropagation functor $L_{\varepsilon, \rho}$, we get a functor

$$\begin{array}{ccc} \mathbf{NNet} & \xrightarrow{\quad} & \mathbf{Learn} \\ & \searrow I^\sigma \quad \nearrow L_{\varepsilon, \rho} & \\ & \mathbf{Para} & \end{array}$$

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

A symmetric monoidal structure, both on \mathbf{NNet} and on the above functors, can be given by generalising the category \mathbf{NNet} to the category where morphisms are directed acyclic graphs with interfaces; details on such a category can be found in [7], see also Remark V.5.

In Section VI, we will compute an extended example of the use of neural nets to compositionally define supervised learning algorithms. Before this, however, we discuss additional compositional structure available to us in \mathbf{Learn} , \mathbf{Para} , and the aforementioned monoidal generalisation of \mathbf{NNet} .

V. NETWORKING IN \mathbf{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 \mathbf{FVect} of linear maps between Euclidean spaces sits inside the category \mathbf{Para} of parametrised functions; we simply consider each linear map as parametrised by the trivial parameter space \mathbb{R}^0 . Given a choice of step size and error function, and hence a functor $L_{\varepsilon, \rho}: \mathbf{Para} \rightarrow \mathbf{Learn}$ as in Theorem III.2, we thus have an inclusion

$$\mathbf{FVect} \hookrightarrow \mathbf{Para} \xrightarrow{L_{\varepsilon, \rho}} \mathbf{Learn}.$$

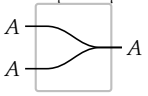
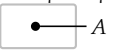
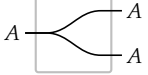
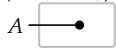
This allows us to construct a learning algorithm—that is, a morphism in \mathbf{Learn} —as the image of any morphism in \mathbf{FVect} , and the output algorithms obey the same equations as the input linear maps. In particular, from graphical linear algebra [2], [4] we know that each object in \mathbf{FVect} is equipped with a bimonoid structure, so we

can use our functor $L_{\varepsilon, e}$ to equip each object of the form \mathbb{R}^n in **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 III.4).

Proposition V.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 in the image of this functor can be equipped with the structure of a bimonoid.*

Explicitly, the bimonoid maps are given as follows. Note they all have trivial parameter space \mathbb{R}^0 , so we denote the unique update function $! : \mathbb{R}^0 \times A \times B \rightarrow \mathbb{R}^0$.

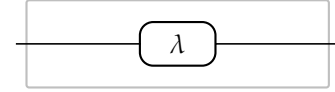
	Implementation	Request
Multiplication μ $(1, I_\mu, !, r_\mu)$ 	$I_\mu(a_1, a_2) = a_1 + a_2$	$r_\mu((a_1, a_2), a_3) = (a_3 - a_2, a_3 - a_1)$
Unit η $(1, I_\eta, !, r_\eta)$ 	$I_\eta(0) = 0$	$r_\eta(a) = 0$
Comultiplication δ $(1, I_\delta, !, r_\delta)$ 	$I_\delta(a) = (a, a)$	$r_\delta(a_1, (a_2, a_3)) = a_2 + a_3 - a_1$
Counit ϵ $(1, I_\epsilon, !, r_\epsilon)$ 	$I_\epsilon(a) = 0$	$r_\epsilon(a) = 0$

Remark V.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 is instead given by $r'_\mu(a_1, a_2, a_3) = (a_3, a_3)$ and the request function on the comultiplication is instead given by $r'_\delta(a_1, a_2, a_3) = a_2 + a_3$. While this is a rather strange error function—minimising error entails sending outputs to 0—the existence of such structures is interesting.

A choice of bimonoid structures, such as that given by Proposition V.1, allows us to interpret network diagrams in the monoidal category $(\mathbf{Learn}, \parallel)$ from Proposition II.4. Indeed, they give canonical interpretations of splitting, joining, initializing, and discarding wires.

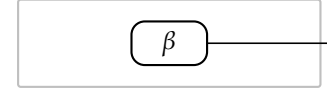
Example V.3 (Building neurons). As learning algorithms implemented with respect to quadratic error (see Example III.4) and some step size ε , neural networks have a rather simple structure: they are generated by three basic learning algorithms—scalar multiplication λ , bias β , and an activation function σ —together with the bimonoid multiplication μ and comultiplication δ given by Proposition V.1.

The scalar multiplication learning algorithm $\lambda : \mathbb{R} \rightarrow \mathbb{R}$, which we shall represent graphically by the string diagram in **Learn**²



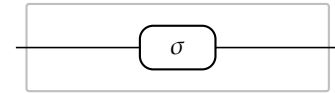
is given by the parameter space \mathbb{R} , implementation function $\lambda(w, x) = wx$, update function $U_\lambda(w, x, y) = w - \varepsilon x(wx - y)$, and request function $r_\lambda(w, x, y) = x - w(wx - y)$.

The bias learning algorithm $\beta : \mathbb{R}^0 \rightarrow \mathbb{R}$, 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

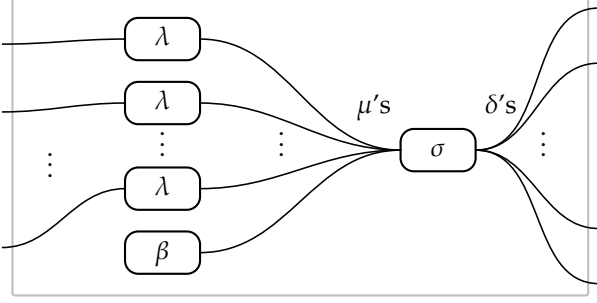


has trivial parameter space, and is specified by some choice of activation function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, together with the trivial update function and the request function $r_\sigma(x, y) = x - (x - y) \frac{\partial \sigma}{\partial x}(x)$.

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

²Note that these are string diagrams in $(\mathbf{Learn}, \parallel)$, while the string diagrams of Section II were string diagrams in (\mathbf{Set}, \times) . As always, string diagrams represent morphisms in a category, with the domain at the left of the diagram and codomain on the right. For more details see the extended version [10].

composite of δ s.

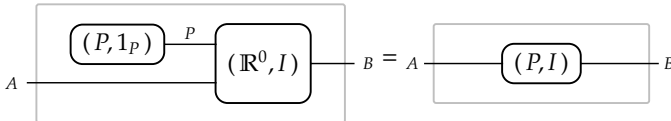


Composing these units using the composition rule in **Learn** further constructs any learning algorithm that can be obtained by gradient descent and backpropagation on a neural network with respect to the quadratic error.

Example V.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 [13]. 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 $P \times \mathbb{R}^0 \rightarrow P$.

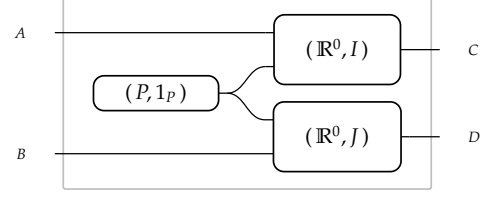
In particular, every parametrised function can be written as a composite, using the bimonoid structure, of a trivially parametrised function and a parametrised constant. To see this, we use string diagrams in $(\mathbf{Para}, \parallel)$, where here we denote a parametrised function $I: P \times A \rightarrow B$ as a box labeled (P, 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



Since these morphisms are the same in **Para**, they correspond to the same learning algorithm, by the functoriality of $L_{\varepsilon, \sigma}$, Theorem III.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:



The comonoid structure from Proposition V.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 .

Remark V.5. It has suited our purposes to simply consider the category **NNet** of neural networks. That said, 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.

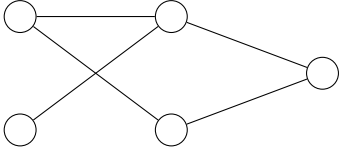
In fact, the category **NNet** can be generalised to a symmetric monoidal category with bimonoids on each object. This generalisation is the strict symmetric monoidal category **IDAG** of *idags*—interfaced directed acyclic graphs—which has been previously studied as an important structure in concurrency, as well as for its elegant categorical properties [7].

It is also desirable that each functor $I^\sigma: \mathbf{NNet} \rightarrow \mathbf{Para}$ 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 symmetric monoidal and preserves bimonoid structures.

VI. 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 $I^\sigma: \mathbf{NNet} \rightarrow \mathbf{Para}$ and $L_{\varepsilon, \sigma}: \mathbf{Para} \rightarrow \mathbf{Learn}$ by Theorem III.2 and Proposition IV.2. In particular, we shall see that $L_{\varepsilon, \sigma}$ implements the usual backpropagation algorithm with quadratic error and step size ε on a neural network with activation function σ . To simplify notation, we'll write I for I^σ .

Consider the following network, which has a single hidden layer:



Call this network A ; it is a morphism $A: 2 \rightarrow 1$ in the category \mathbf{NNet} of neural networks. The image of A under the functor $I: \mathbf{NNet} \rightarrow \mathbf{Para}$ is the parametrised function $I_A: (\mathbb{R}^5 \times \mathbb{R}^3) \times \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by

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

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 $L_{\varepsilon, e}$, 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 $L_{\varepsilon, e}$ 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 $L_{\varepsilon, e}$ -image of I_A in \mathbf{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

$$U_A: (\mathbb{R}^5 \times \mathbb{R}^3) \times \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}^5 \times \mathbb{R}^3$$

is defined by

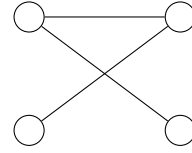
$$U_A(p, q, a, c) = \begin{pmatrix} p \\ q \end{pmatrix} - \varepsilon \nabla_{p, q} \frac{1}{2} \|I_A(p, q, a) - c\|^2 = \begin{pmatrix} p_{11} - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_1\dot{\sigma}(\beta_1)a_1 \\ p_{12} - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_1\dot{\sigma}(\beta_1)a_2 \\ p_{21} - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_2\dot{\sigma}(\beta_2)a_1 \\ p_{1b} - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_1\dot{\sigma}(\beta_1) \\ p_{2b} - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_2\dot{\sigma}(\beta_2) \\ q_1 - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)\sigma(\beta_1) \\ q_2 - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)\sigma(\beta_2) \\ q_b - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma) \end{pmatrix},$$

and $r_A: (\mathbb{R}^5 \times \mathbb{R}^3) \times \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}^2$ is defined by

$$r_A(p, q, a, c) = a - \nabla_a \frac{1}{2} \|I_A(p, q, a) - c\|^2 = \begin{pmatrix} a_1 - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)(q_1\dot{\sigma}(\beta_1)p_{11} + q_2\dot{\sigma}(\beta_2)p_{21}) \\ a_2 - \varepsilon(I_A(p, q, a) - c)\dot{\sigma}(\gamma)q_1\dot{\sigma}(\beta_1)p_{12} \end{pmatrix}$$

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

$$I_B: \mathbb{R}^5 \times \mathbb{R}^2 \rightarrow \mathbb{R}^2; \\ (p, a) \mapsto \begin{pmatrix} \sigma(p_{11}a_1 + p_{12}a_2 + p_{1b}) \\ \sigma(p_{21}a_1 + p_{2b}) \end{pmatrix}$$

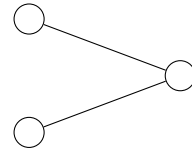
which in turn has update and request functions

$$U_B: \mathbb{R}^5 \times \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}^5; \\ (p, a, b) \mapsto \begin{pmatrix} p_{11} - \varepsilon(I_B(p, a)_1 - b_1)\dot{\sigma}(\beta_1)a_1 \\ p_{12} - \varepsilon(I_B(p, a)_1 - b_1)\dot{\sigma}(\beta_1)a_2 \\ p_{21} - \varepsilon(I_B(p, a)_2 - b_2)\dot{\sigma}(\beta_2)a_1 \\ p_{1b} - \varepsilon(I_B(p, a)_2 - b_2)\dot{\sigma}(\beta_1) \\ p_{2b} - \varepsilon(I_B(p, a)_2 - b_2)\dot{\sigma}(\beta_2) \end{pmatrix}$$

and $r_B: \mathbb{R}^5 \times \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$; where

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

The second layer $C: 2 \rightarrow 1$



represents the parametrised function

$$I_C: \mathbb{R}^3 \times \mathbb{R}^2 \rightarrow \mathbb{R}; \\ (q, b) \mapsto \sigma(q_1b_1 + q_2b_2 + q_b).$$

which in turn has update and request functions

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

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

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

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

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} :

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

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.

VII. 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 backpropagation 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.

- Neural networks provide a simple, compositional language for specifying learning algorithms.
- Composition of learners, along with the fact that gradients are quicker to compute for lower dimensional spaces, expresses the speed up in learning provided by backpropagation.

We close with some remarks on further directions.

A. More general error functions

To apply our main theorem, and hence understand backpropagation as a functor, we require certain derivatives of our chosen error function to be invertible. Some commonly used error functions, however, 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 given by

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

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 A of the extended version [10].

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 main theorem to accommodate error functions such as cross entropy that fail to have derivatives at isolated points. Regardless, note that while in this case it is not straightforward to state backpropagation as a functor from *Para*, our analysis nevertheless still sheds light on the compositional nature of the learning algorithm.

B. Generalised networked learning algorithms

The category *Learn* contains many more morphisms than those in the images of *Para* under the gradient descent/backpropagation functors $L_{\varepsilon, \rho}$. Indeed, *Learn* does not require us to 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/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. Indeed, learning over manifolds equipped with some differentiable structure is an active field of study [5]. We might also learn nonlinear functions like rotations, or find a way to parametrise over network architectures.

There is a clear advantage of using gradient descent: it gives a heuristic argument that the learning algorithm updates towards reducing 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 lies in the functors $L_{\varepsilon, \rho}$. Thus *Learn* lets us construct learning algorithms that vary the notion of error across the network.

Finally, neural networks are useful because they provide a simple, combinatorial language for specifying supervised learning algorithms. In Section V, we saw that this fact can be cast in categorical terms as follows: neural networks are useful as they are the language generated, using the grammar of symmetric monoidal categories, from just a few learners (scalar multiplication, bias, activation functor, monoid multiplication, and comultiplication). Choosing other primitives could provide a new, similarly simple language for specifying learning algorithms tailored to a chosen application.

C. 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 usual morphisms. Learners naturally form a bicategory. Indeed, our definition of equivalence of learners implicitly uses this structure; equivalence is just isomorphism for the following notion of 2-morphism.

Definition VII.1. A 2-morphism $f: (P, I, U, r) \rightarrow (Q, J, V, s)$ of learners is a function $f: P \rightarrow Q$ such that $J(f(p), a) = I(p, a)$, $V(f(p), a, b) = f(U(p, a, b))$, and $s(f(p), a, b) = r(p, a, b)$.

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.

D. Learners, lenses, and open games

We defined the category of learners to model the exchange of information between individual learning units, and how this creates a larger, composite learner. Similarly, category theory has been used to abstractly model bidirectional programming languages and databases, using various notions of *lens*, and interacting microeconomic games, resulting in the notion of an *open game*.

These categorical analyses reveal striking structural similarities between these three subjects, unified through the idea that at core, they study how agents exchange and respond to information. Indeed, asymmetric lenses are simply learners with trivial state spaces, and learners themselves are open games obeying a certain singleton best response condition. Writing *Lens* and *Game* for the respective categories (defined in [14] and [11]), this gives embeddings

$$\text{Lens} \hookrightarrow \text{Learn} \hookrightarrow \text{Game}.$$

Via these functors, the implementation function corresponds to the get function of a lens and the play function of an open game, while the request function corresponds to the put and coplay functions. The update function of the learner corresponds to the strategy update function, known as the best response function, for the open game [9], [12]. Moreover, the category *Learn* also embeds into a certain category of symmetric lenses [9].

Lenses come with various notions of ‘well behavedness’, which place compatibility conditions between put and get functions. So far, in the case of learners, we have placed no requirements that an algorithm converge towards a function f when given enough training pairs $(a, f(a))$. Examining the lens-learner relationship may shed insight onto how not only to define structures that learn, but that learn well.

REFERENCES

- [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](#).
- [2] J. C. Baez and J. Erbele. Categories in Control. *Theory and Application of Categories*, 30(24):836–881, 2015.
- [3] A. Bohannon, B. C. Pierce, and J. A. Vaughan. Relational lenses: a language for updatable views. *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '06)*, pp.338–347, 2006. [doi:10.1145/1142351.1142399](#).
- [4] F. Bonchi, P. Sobociński, and F. Zanasi. Interacting Hopf Algebras. *Journal of Pure and Applied Algebra*, 221(1):144–184, 2017. [doi:10.1016/j.jpaa.2016.06.002](#).
- [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](#).
- [6] B. Coecke and A. Kissinger. *Picturing Quantum Processes A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge University Press, 2017.
- [7] 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. [arXiv:1303.0376](#).
- [8] B. Fong. *The Algebra of Open and Interconnected Systems*. DPhil thesis, University of Oxford, 2016. [arXiv:1609.05382](#).
- [9] B. Fong and M. Johnson. Lenses and Learners. To appear in *Proceedings of the Eighth International Workshop on Bidirectional Transformations (Bx2019)*. [arXiv:1903.03671](#)
- [10] B. Fong, D. I. Spivak, R. Tuyéras. Backprop as Functor: A compositional perspective on supervised learning. (Extended version.) [arXiv:1711.10455](#)
- [11] N. Ghani, J. Hedges, V. Winschel, P. Zahn. Compositional Game Theory. *Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS '18)*, pp.472–481, 2018. [doi:10.1145/3209108.3209165](#).
- [12] J. Hedges. From Open Learners to Open Games. *Preprint*. [arXiv:1902.08666](#)
- [13] Y. Le Cun and T. Bengio. Convolutional networks for images, speech, and time series. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks*, pp.255–258, MIT Press, 1995.
- [14] M. Johnson and R. D. Rosebrugh. Spans of lenses. *EDBT/ICDT Workshops*, pp.112–118, 2014.
- [15] 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](#).
- [16] S. Mac Lane. *Categories for the Working Mathematician*, Springer, Berlin, 1998.
- [17] 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.
- [18] M. A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. <http://neuralnetworksanddeeplearning.com>
- [19] D. Vagner, D. Spivak, E. Lerman. Algebras of open dynamical systems on the operad of wiring diagrams. [arXiv:1408.1598](#).