

Deep Learning with Parametric Lenses

Geoffrey S. H. Cruttwell, Bruno Gavranović, Neil Ghani, Paul Wilson, and Fabio Zanasi

Mount Allison University, Canada

University of Strathclyde, United Kingdom

University of Strathclyde, United Kingdom

Independent, United Kingdom

University College London, United Kingdom, and University of Bologna, Italy

Abstract

We propose a categorical semantics for machine learning algorithms in terms of lenses, parametric maps, and reverse derivative categories. This foundation provides a powerful explanatory and unifying framework: it encompasses a variety of gradient descent algorithms such as ADAM, AdaGrad, and Nesterov momentum, as well as a variety of loss functions such as MSE and Softmax cross-entropy, and different architectures, shedding new light on their similarities and differences. Furthermore, our approach to learning has examples generalising beyond the familiar continuous domains (modelled in categories of smooth maps) and can be realised in the discrete setting of Boolean and polynomial circuits. We demonstrate the practical significance of our framework with an implementation in Python.

Keywords: Neural network, Deep Learning, String diagram, Symmetric Monoidal Category, Cartesian Differential Category

1. Introduction

The last decade has witnessed a surge of interest in machine learning, fuelled by the numerous successes and applications that these methodologies have found in many fields of science and technology. As machine learning techniques become increasingly pervasive, algorithms and models become more sophisticated, posing a significant challenge both to the software developers and the users that need to interface, execute and maintain these systems. In spite of this rapidly evolving picture, the formal analysis of many learning algorithms mostly takes place at a heuristic level [49], or using definitions that fail to provide a general and scalable framework for describing machine learning. Indeed, it is commonly acknowledged through academia, industry, policy makers and funding agencies that there is a pressing need for a unifying perspective, which can make this growing body of work more systematic, rigorous, transparent and accessible both for users and developers [41, 53].

Consider, for example, one of the most common machine learning scenarios: supervised learning with a neural network. This technique trains the model towards a certain task, e.g. the recognition of patterns in a data set (*cf.* Figure 1). There are several different ways of implementing this scenario. Typically, at their core, there is a *gradient update* algorithm (often called the “optimiser”), depending on a given *loss function*, which updates in steps the parameters of the network, based on some *learning rate* controlling the “scaling” of the update. All of these components can vary independently in a supervised learning algorithm and a number of choices is available for loss maps (quadratic error, Softmax cross entropy, dot product, etc.) and optimisers (Adagrad [23], Momentum [45], and Adam [36], etc.).

This scenario highlights several questions: is there a uniform mathematical language capturing the different components of the learning process? Can we develop a unifying picture of the various optimisation techniques, allowing for their comparative analysis? Moreover, it should be noted

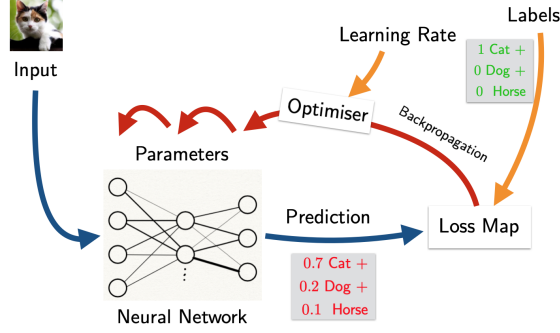


Figure 1: An informal illustration of gradient-based learning. This neural network is trained to distinguish different kinds of animals in the input image. Given an input X , the network predicts an output Y , which is compared by a ‘loss map’ with what would be the correct answer (‘label’). The loss map returns a real value expressing the error of the prediction; this information, together with the *learning rate* (a weight controlling how much the model should be changed in response to error) is used by an *optimiser*, which computes by gradient-descent the update of the parameters of the network, with the aim of improving its accuracy. The neural network, the loss map, the optimiser and the learning rate are all components of a supervised learning system, and can vary independently of one another.

that supervised learning is not limited to neural networks. For example, supervised learning is surprisingly applicable to the discrete setting of boolean circuits [60] where continuous functions are replaced by boolean-valued functions. Can we identify an abstract perspective encompassing both the real-valued and the boolean case? In a nutshell, this paper seeks to answer the question:

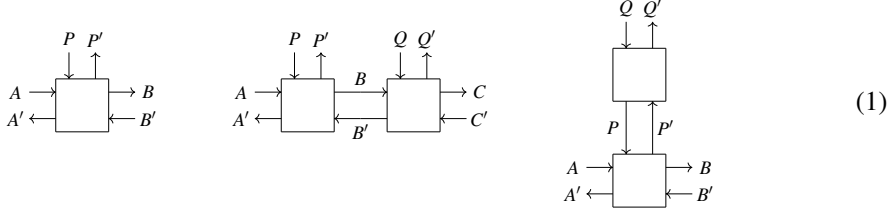
*what are the fundamental mathematical structures
underpinning gradient-based learning?*

Our approach to this question stems from the identification of three fundamental aspects of the gradient-descent learning process:

- (I) computation is **parametric**, e.g. in the simplest case we are given a function $f : P \times X \rightarrow Y$ and learning consists of finding a parameter $p : P$ such that $f(p, -)$ is the best function according to some criteria. Specifically, the weights on the internal nodes of a neural network are a parameter which the learning is seeking to optimize. Parameters also arise elsewhere, e.g. in the loss function (see later).
- (II) information flows **bidirectionally**: in the forward direction, the computation turns inputs via a sequence of *layers* into predicted outputs, and then into a loss value; in the reverse direction, backpropagation is used to propagate the changes *backwards* through the layers, and then turn them into parameter updates.
- (III) the basis of parameter update via gradient descent is **differentiation** e.g. in the simple case we differentiate the function mapping a parameter to its associated loss to reduce that loss.

We model bidirectionality via lenses [13, 6, 33] and based upon the above three insights, we propose the notion of **parametric lens** as the fundamental semantic structure of learning. In a nutshell, a parametric lens is a process with three kinds of interfaces: inputs, outputs, and parameters. On each interface, information flows both ways, i.e. computations are bidirectional. These data are best explained with our graphical representation of parametric lenses, with inputs A, A' , outputs B, B' , parameters P, P' , and arrows indicating information flow (below left). The graphical notation

also makes evident that parametric lenses are *open systems*, which may be composed along their interfaces (below center and right).



This pictorial formalism is not just an intuitive sketch: as we will show, it can be understood as a completely formal (graphical) syntax using the formalism of *string diagrams* [44], in a way similar to how other computational phenomena have been recently analysed e.g. in quantum theory [16], control theory [4, 8], and digital circuit theory [29].

It is intuitively clear how parametric lenses express aspects (I) and (II) above, whereas (III) will be achieved by studying them in a space of ‘differentiable objects’ (in a sense that will be made precise). The main technical contribution of our paper is showing how the various ingredients involved in learning (the model, the optimiser, the error map and the learning rate) can be uniformly understood as being built from parametric lenses.

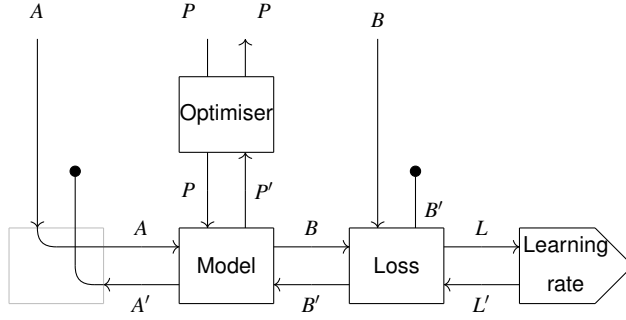


Figure 2: The parametric lens that captures the learning process informally sketched in Figure 1. Note each component is a lens itself, whose composition yields the interactions described in Figure 1. Defining this picture formally will be the subject of Sections 3-4.

We will use *category theory* as the formal language to develop our notion of parametric lenses, and make Figure 2 mathematically precise. The categorical perspective brings several advantages, which are well-known, established principles in programming language semantics [59, 1, 48]. Three of them are particularly important to our contribution, as they constitute distinctive advantages of our semantic foundations:

Abstraction Our approach studies which categorical structures are sufficient to perform gradient-based learning. This analysis abstracts away from the standard case of neural networks in several different ways: as we will see, it encompasses other models (namely Boolean circuits), different kinds of optimisers (including Adagrad, Adam, Nesterov momentum), and error maps (including quadratic and softmax cross entropy loss). These can be all understood as parametric lenses, and different forms of learning result from their interaction.

Uniformity As seen in Figure 1, learning involves ingredients that are seemingly quite different: a model, an optimiser, a loss map, etc. We will show how all these notions may be seen as instances of the categorical definition of a parametric lens, thus yielding a remarkably

uniform description of the learning process, and supporting our claim of parametric lenses being a fundamental semantic structure of learning.

Compositionality The use of categorical structures to describe computation naturally enables *compositional reasoning* whereby complex systems are analysed in terms of smaller, and hence easier to understand, components. Compositionality is a fundamental tenet of programming language semantics; in the last few years, it has found application in the study of diverse kinds of computational models, across different fields— see e.g. [54, 28, 16, 8]. As made evident by Figure 2, our approach models a neural network as a parametric lens, resulting from the *composition* of simpler parametric lenses, capturing the different ingredients involved in the learning process. Moreover, as all the simpler parametric lenses are themselves composable, one may engineer a different learning process by simply plugging a new lens on the left or right of existing ones. This means that one can glue together smaller and relatively simple networks to create larger and more sophisticated neural networks.

We now give a synopsis of our contributions:

- In Section 2, we introduce the tools necessary to define our notion of **parametric lens**. First, in Section 2.1, we introduce a notion of parametric categories, which amounts to a functor $\mathbf{Para}(-)$ turning a category \mathcal{C} into one $\mathbf{Para}(\mathcal{C})$ of ‘parametric \mathcal{C} -maps’. Second, we recall *lenses* (Section 2.2). In a nutshell, a lens is a categorical morphism equipped with operations to view and update values in a certain data structure. Lenses play a prominent role in functional programming [57], as well as in the foundations of database theory [35] and more recently game theory [28]. Considering lenses in \mathcal{C} simply amounts to the application of a functorial construction $\mathbf{Lens}(-)$, yielding $\mathbf{Lens}(\mathcal{C})$. Finally, we recall the notion of a *cartesian reverse differential category* (CRDC): a categorical structure axiomatising the notion of differentiation [15] (Section 2.4). We wrap up in Section 2.3, by combining these ingredients into the notion of parametric lens, formally defined as a morphism in $\mathbf{Para}(\mathbf{Lens}(\mathcal{C}))$ for a CRDC \mathcal{C} . In terms of our desiderata (I)-(III) above, note that $\mathbf{Para}(-)$ accounts for (I), $\mathbf{Lens}(-)$ accounts for (II), and the CRDC structure accounts for (III).
- As seen in Figure 1, in the learning process there are many components at work: the model, the optimiser, the loss map, the learning rate, etc.. In Section 3, we show how the notion of parametric lens provides a uniform characterisation for such components. Moreover, for each of them, we show how different variations appearing in the literature become instances of our abstract characterisation. The plan is as follows:
 - In Section 3.1, we show how the combinatorial **model** subject of the training can be seen as a parametric lens. The conditions we provide are met by the ‘standard’ case of neural networks, but also enables the study of learning for other classes of models. In particular, another instance are Boolean circuits: learning of these structures is relevant to binarisation [18] and it has been explored recently using a categorical approach [60], which turns out to be a particular case of our framework. We continue by describing internals of a model, and translating several examples of models in deep learning to their categorical form. This includes linear layers, biases, activations, convolutional layers, but also general techniques such as weight tying and batching.
 - In Section 3.2, we show how the **loss maps** associated with training are also parametric lenses. Our approach covers the cases of quadratic error, Boolean error, Softmax cross entropy, but also the ‘dot product loss’ associated with the phenomenon of deep dreaming [40, 38, 22, 52].
 - In Section 3.3, we model the **learning rate** as a parametric lens. This analysis also allows us to contrast how learning rate is handled in the ‘real-valued’ case of neural networks with respect to the ‘Boolean-valued’ case of Boolean circuits.

- In Section 3.4, we show how **optimisers** can be modelled as ‘reparameterisations’ of models as parametric lenses. As case studies, in addition to basic gradient ascent and descent, we consider the stateful variants: Momentum [45], Nesterov Momentum [58], Adagrad [23], and Adam (Adaptive Moment Estimation) [36], as well as optimiser composition (Subsection 3.4.1). Also, on Boolean circuits, we show how the reverse derivative ascent of [60] can be also regarded in such way.
- In Section 4, we study how the composition of the lenses defined in Section 3 yields a description of different kinds of learning processes.
 - Section 4.1 is dedicated to modelling supervised **learning of parameters**, in the way described in Figure 1. This amounts essentially to study of the composite of lenses expressed in Figure 2, for different choices of the various components. In particular we look at (i) quadratic loss with basic gradient descent, (ii) softmax cross entropy loss with basic gradient descent, (iii) quadratic loss with Nesterov momentum, and (iv) learning in Boolean circuits with XOR loss and basic gradient ascent.
 - In Section 4.2 we describe how a system traditionally considered as unsupervised can be recast to its supervised form: Generative Adversarial Networks ([30, 3]). We define this model abstractly as a parametric lens, and describe how a particular instantiation thereof — Wasserstein GAN ([3]) — arises as a supervised learning system with the dot product loss and the gradient descent-ascent optimiser.
 - In order to showcase the flexibility of our approach, in Section 4.3 we depart from our ‘core’ case study of parameter learning, and turn attention to supervised **learning of inputs**, also called **deep dreaming** — the idea behind this technique is that, instead of the network parameters, one updates the inputs, in order to elicit a particular interpretation [40, 38, 22, 52]. Deep dreaming can be easily expressed within our approach, with a different rearrangement of the parametric lenses involved in the learning process, see (11) below. The abstract viewpoint of categorical semantics provides a mathematically precise and visually captivating description of the differences between the usual parameter learning process and deep dreaming.
- In Section 5 we describe a proof-of-concept Python **implementation**, available at [19], based on the theory developed in this paper. This code is intended to show more concretely the payoff of our approach. Model architectures, as well as the various components participating in the learning process, are now expressed in a uniform, principled mathematical language, in terms of lenses. As a result, computing network gradients is greatly simplified, as it amounts to lens composition. Moreover, the modularity of this approach allows one to more easily tune the various parameters of training.
We show our library via a number of experiments, and prove correctness by achieving accuracy on par with an equivalent model in Keras, a mainstream deep learning framework [12]. In particular, we create a working non-trivial neural network model for the MNIST image-classification problem [37].
- Finally, in Sections 6 and 7, we discuss related and future work.

Note this paper extends a previous conference version [20]. Section 3.1 has been extended with examples of different architectures and techniques in deep learning, while Section 4.2, which considers unsupervised learning, and Remark 9, followed by a discussion on the axioms of CRDCs needed in our framework, are new. Also, we included missing proofs and complementary background material (see in particular Appendices A-B).

2. Categorical Toolkit

In this section we describe the three categorical components of our framework, each corresponding to an aspect of gradient-based learning: (I) the **Para** construction (Section 2.1), which builds

a category of parametric maps, (II) the **Lens** construction, which builds a category of “bidirectional” maps (Section 2.2), and (III) the combination of these two constructions into the notion of “parametric lenses” (Section 2.3). Finally (IV) we recall Cartesian reverse differential categories — categories equipped with an abstract gradient operator.

Notation We shall use $f; g$ for sequential composition of morphisms $f: A \rightarrow B$ and $g: B \rightarrow C$ in a category, 1_A for the identity morphism on A , and I for the unit object of a symmetric monoidal category.

2.1 Parametric Maps

In supervised learning one is typically interested in approximating a function $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ for some n and m . To do this, one begins by building a neural network, which is a smooth map $f: \mathbb{R}^p \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ where \mathbb{R}^p is the set of possible weights of that neural network. Then one looks for a value of $q \in \mathbb{R}^p$ such that the function $f(q, -): \mathbb{R}^n \rightarrow \mathbb{R}^m$ closely approximates g . We formalise these maps categorically via the **Para** construction [26, 27, 10, 34].

Definition 1 (Parametric category). *Let $(\mathcal{C}, \otimes, I)$ be a strict^a symmetric monoidal category. We define a category $\mathbf{Para}(\mathcal{C})$ with*

- *objects those of \mathcal{C} ;*
- *a map from A to B is a pair (P, f) , with P an object of \mathcal{C} and $f: P \otimes A \rightarrow B$;*
- *the identity on A is the pair $(I, 1_A)$ (since \otimes is strict monoidal, $I \otimes A = A$);*
- *the composite of maps $(P, f): A \rightarrow B$ and $(P', f'): B \rightarrow C$ is the pair $(P' \otimes P, (1_{P'} \otimes f); f')$;*

Example 2. *Take the category **Smooth** whose objects are natural numbers and whose morphisms $f: n \rightarrow m$ are smooth maps from \mathbb{R}^n to \mathbb{R}^m . This is a strict symmetric monoidal category with product given by addition. As described above, the category $\mathbf{Para}(\mathbf{Smooth})$ can be thought of as a category of neural networks: a map in this category from n to m consists of a choice of p and a map $f: \mathbb{R}^p \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ with \mathbb{R}^p representing the set of possible weights of the neural network.*

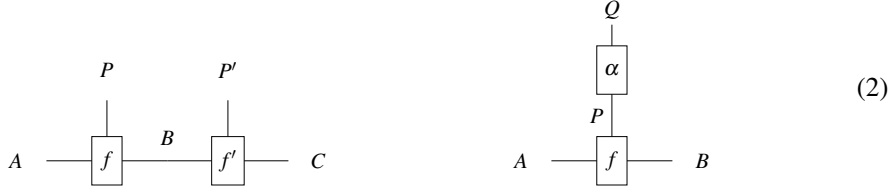
As we will see in the next sections, the interplay of the various components at work in the learning process becomes much clearer once represented the morphisms of $\mathbf{Para}(\mathcal{C})$ using the pictorial formalism of *string diagrams*, which we now recall. In fact, we will mildly massage the traditional notation for string diagrams (below left), by representing a morphism $f: A \rightarrow B$ in $\mathbf{Para}(\mathcal{C})$ as below right.



This is to emphasise the special role played by P , reflecting the fact that in machine learning data and parameters have different semantics. String diagrammatic notations also allows to neatly represent composition of maps $(P, f): A \rightarrow B$ and $(P', f'): B \rightarrow C$ (below left), and

^aOne can also define $\mathbf{Para}(\mathcal{C})$ in the case when \mathcal{C} is non-strict; however, the result would be not a category but a bicategory.

“reparameterisation” of $(P, f) : A \rightarrow B$ by a map $\alpha : Q \rightarrow P$ (below right), yielding a new map $(Q, (\alpha \otimes 1_A); f) : A \rightarrow B$.



Intuitively, reparameterisation changes the parameter space of $(P, f) : A \rightarrow B$ to some other object Q , via some map $\alpha : Q \rightarrow P$. We shall see later that gradient descent and its many variants can naturally be viewed as reparameterisations.

Note coherence rules in combining the two operations in (2) just work as expected, as these diagrams can be ultimately ‘compiled’ down to string diagrams for monoidal categories.

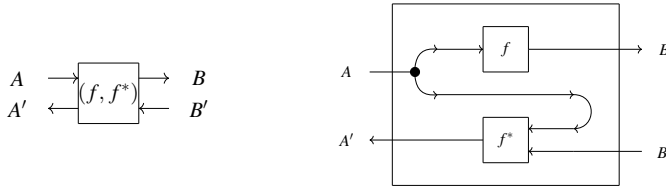
2.2 Lenses

In machine learning (or even learning in general) it is fundamental that information flows both forwards and backwards: the ‘forward’ flow corresponds to a model’s predictions, and the ‘backwards’ flow to *corrections* to the model. The category of lenses is the ideal setting to capture this type of structure, as it is a category consisting of maps with both a “forward” and a “backward” part.

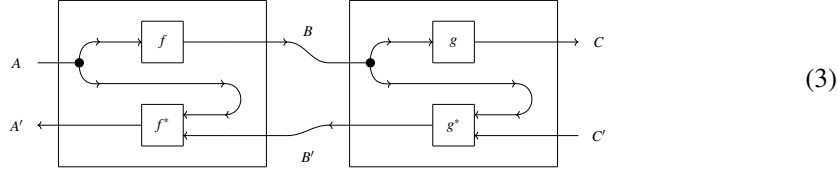
Definition 3. For any Cartesian category \mathcal{C} , the category of (bimorphic) lenses in \mathcal{C} , $\mathbf{Lens}(\mathcal{C})$, is the category with the following data:

- Objects are pairs (A, A') of objects in \mathcal{C} , written as $\begin{pmatrix} A \\ A' \end{pmatrix}$;
- A map from $\begin{pmatrix} A \\ A' \end{pmatrix}$ to $\begin{pmatrix} B \\ B' \end{pmatrix}$ consists of a pair (f, f^*) (also written as $\begin{pmatrix} f \\ f^* \end{pmatrix}$) where $f : A \rightarrow B$ (called the **get** or **forward** part of the lens) and $f^* : A \times B' \rightarrow A'$ (called the **put** or **backwards** part of the lens);
- The identity on $\begin{pmatrix} A \\ A' \end{pmatrix}$ is the pair $\begin{pmatrix} 1_A \\ \pi_1 \end{pmatrix}$;
- The composite of $\begin{pmatrix} f \\ f^* \end{pmatrix} : \begin{pmatrix} A \\ A' \end{pmatrix} \rightarrow \begin{pmatrix} B \\ B' \end{pmatrix}$ and $\begin{pmatrix} g \\ g^* \end{pmatrix} : \begin{pmatrix} B \\ B' \end{pmatrix} \rightarrow \begin{pmatrix} C \\ C' \end{pmatrix}$ is given by get $f; g$ and put $\langle \pi_0, \langle \pi_0; f, \pi_1 \rangle; g^* \rangle; f^*$.

The embedding of $\mathbf{Lens}(\mathcal{C})$ into the category of Tambara modules over \mathcal{C} (see [7, Thm. 23]) provides a rich string diagrammatic language, in which lenses may be represented with forward/backward wires indicating the information flow. In this language, a morphism $\begin{pmatrix} f \\ f^* \end{pmatrix} : \begin{pmatrix} A \\ A' \end{pmatrix} \rightarrow \begin{pmatrix} B \\ B' \end{pmatrix}$ is written as below left, which can be ‘expanded’ as below right.



It is clear in this language how to describe the composite of $\begin{pmatrix} f \\ f^* \end{pmatrix} : \begin{pmatrix} A \\ A' \end{pmatrix} \rightarrow \begin{pmatrix} B \\ B' \end{pmatrix}$ and $\begin{pmatrix} g \\ g^* \end{pmatrix} : \begin{pmatrix} B \\ B' \end{pmatrix} \rightarrow \begin{pmatrix} C \\ C' \end{pmatrix}$:



Remark 4. Note $\mathbf{Lens}(\mathcal{C})$ is a monoidal category, with $\begin{pmatrix} A \\ A' \end{pmatrix} \otimes \begin{pmatrix} B \\ B' \end{pmatrix}$ defined as $\begin{pmatrix} A \times B \\ A' \times B' \end{pmatrix}$. However, in general $\mathbf{Lens}(\mathcal{C})$ is not itself Cartesian. This is easy to see when looking at even a terminal object: if T is a terminal object in \mathcal{C} , then in general $\begin{pmatrix} T \\ T \end{pmatrix}$ will not be a terminal object in $\mathbf{Lens}(\mathcal{C})$ — if it was, there would be a unique lens $\begin{pmatrix} !_A \\ !_{A'} \end{pmatrix} : \begin{pmatrix} A \\ A' \end{pmatrix} \rightarrow \begin{pmatrix} T \\ T \end{pmatrix}$ whose put part would need to be a (unique) map $A \times T \rightarrow A'$, but in general there are many such maps.

2.3 Parametric Lenses

The fundamental category where supervised learning takes place is the composite $\mathbf{Para}(\mathbf{Lens}(\mathcal{C}))$ of the two constructions in the previous sections:

Definition 5. The category $\mathbf{Para}(\mathbf{Lens}(\mathcal{C}))$ of *parametric lenses* on \mathcal{C} is defined as follows.

- Its objects are pairs of objects $\begin{pmatrix} A \\ A' \end{pmatrix}$ of objects from \mathcal{C} ;
- A morphism from $\begin{pmatrix} A \\ A' \end{pmatrix}$ to $\begin{pmatrix} B \\ B' \end{pmatrix}$, called a *parametric lens*^b, is a choice of parameter pair $\begin{pmatrix} P \\ P' \end{pmatrix}$ and a lens $\begin{pmatrix} f \\ f^* \end{pmatrix} : \begin{pmatrix} P \\ P' \end{pmatrix} \times \begin{pmatrix} A \\ A' \end{pmatrix} \rightarrow \begin{pmatrix} B \\ B' \end{pmatrix}$ where $f : P \times A \rightarrow B$ and $f^* : P \times A \times B' \rightarrow P' \times A'$

String diagrams for parametric lenses are built by simply composing the graphical languages of the previous two sections — see (1), where respectively a morphism, a composition of morphisms, and a reparameterisation are depicted.

Given a generic morphism in $\mathbf{Para}(\mathbf{Lens}(\mathcal{C}))$ as depicted in (1) on the left, one can see how it is possible to “learn” new values from f : it takes as input an input A , a parameter P , and a change B' , and outputs a change in A , a value of B , and a change P' . This last element is the key component for supervised learning: intuitively, it says how to change the parameter values to get the neural network closer to the true value of the desired function.

The question, then, is how one is to define such a parametric lens given nothing more than a neural network, ie., a parametric map $(P, f) : A \rightarrow B$. This is precisely what the gradient operation provides, and its generalization to categories is explored in the next subsection.

2.4 Cartesian Reverse Differential Categories

Fundamental to all types of gradient-based learning is, of course, the gradient operation. In most cases this gradient operation is performed in the category of smooth maps between Euclidean spaces. However, recent work [60] has shown that gradient-based learning can also work well

^bIn [26], these are called *learners*. However, in this paper we study them in a much broader light; see Section 6.

in other categories; for example, in a category of boolean circuits. Thus, to encompass these examples in a single framework, we will work in a category with an abstract gradient operation.

Definition 6. A *Cartesian left additive category* [15, Defn. 1] consists of a category \mathcal{C} with chosen finite products (including a terminal object), and an addition operation and zero morphism in each homset, satisfying various axioms. A *Cartesian reverse differential category* (CRDC) [15, Defn. 13] consists of a Cartesian left additive category \mathcal{C} , together with an operation which provides, for each map $f : A \rightarrow B$ in \mathcal{C} , a map $R[f] : A \times B \rightarrow A$ satisfying various axioms (see (Def. 58)).

For $f : A \rightarrow B$, the pair $\begin{pmatrix} f \\ R[f] \end{pmatrix}$ forms a lens from $\begin{pmatrix} A \\ A \end{pmatrix}$ to $\begin{pmatrix} B \\ B \end{pmatrix}$. We will pursue the idea that $R[f]$ acts as backwards map, thus giving a means to “learn” f .

Note that assigning type $A \times B \rightarrow A$ to $R[f]$ hides some relevant information: B -values in the domain and A -values in the codomain of $R[f]$ do not play the same role as values of the same types in $f : A \rightarrow B$: in $R[f]$, they really take in a tangent vector at B and output a tangent vector at A (cf. the definition of $R[f]$ in **Smooth**, Example 7 below). To emphasise this, we will type $R[f]$ as a map $A \times B' \rightarrow A'$ (even though in reality $A = A'$ and $B = B'$), thus meaning that $\begin{pmatrix} f \\ R[f] \end{pmatrix}$ is actually a lens from $\begin{pmatrix} A \\ A' \end{pmatrix}$ to $\begin{pmatrix} B \\ B' \end{pmatrix}$. This typing distinction will be helpful later on, when we want to add additional components to our learning algorithms.

The following two examples of CRDCs will serve as the basis for the learning scenarios of the upcoming sections.

Example 7. The category **Smooth** (Example 2) is Cartesian with product given by addition, and it is also a Cartesian reverse differential category: given a smooth map $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, the map $R[f] : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ sends a pair (x, v) to $J[f]^T(x) \cdot v$: the transpose of the Jacobian of f at x in the direction v . For example, if $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ is defined as $f(x_1, x_2) := (x_1^3 + 2x_1x_2, x_2, \sin(x_1))$, then $R[f] : \mathbb{R}^2 \times \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is given by

$$(x, v) \mapsto \begin{bmatrix} 3x_1^2 + 2x_2 & 0 & \cos(x_1) \\ 2x_1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

Using the reverse derivative (as opposed to the forward derivative) is well-known to be much more computationally efficient for functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ when $m \ll n$ (for example, see [31]), as is the case in most supervised learning situations (where often $m = 1$).

Example 8. Another CRDC is the symmetric monoidal category $\text{POLY}_{\mathbb{Z}_2}$ [15, Example 14] with objects the natural numbers and morphisms $f : A \rightarrow B$ the B -tuples of polynomials $\mathbb{Z}_2[x_1 \dots x_A]$. When presented by generators and relations these morphisms can be viewed as a syntax for boolean circuits, with parametric lenses for such circuits (and their reverse derivative) described in [60].

Remark 9. The definition of a CRDC (see Def. 58) satisfies 7 axioms describing the interaction of the reverse differential operator with the rest of the cartesian left-additive structure. As pointed out in [14, Sec. 2.3], the last two axioms are independent of the others. In this paper we will additionally see that these two axioms are also not needed to compositionally model the update step of supervised learning.

The remark above can be stated abstractly in two steps. Firstly, we note that a CRDC \mathcal{C} for each morphism f defines a lens $\left(\begin{smallmatrix} f \\ R[f] \end{smallmatrix}\right)$ whose backwards map is additive in the second component (see Def. 54 for the definition of additivity in the second component). This defines a subcategory $\mathbf{Lens}_A(\mathcal{C})$ of $\mathbf{Lens}(\mathcal{C})$, and the following functor.

Theorem. *Lenses with backward passes additive in the second component form a functor*

$$\mathbf{Lens}_A : \mathbf{CLACat} \rightarrow \mathbf{CLACat}$$

Proof. See Appendix. Definition 56 contains the definition of the category \mathbf{CLACat} , Definition 59 the definition of $\mathbf{Lens}_A(\mathcal{C})$, Prop. 60 the proof that $\mathbf{Lens}_A(\mathcal{C})$ is a cartesian left-additive category and Prop. 61 the action of \mathbf{Lens}_A on morphisms. \square

The second step is the observation that a coalgebra of this functor gives us a choice of a cartesian left-additive category \mathcal{C} and a cartesian left-additive functor $\mathbf{R}_{\mathcal{C}} : \mathcal{C} \rightarrow \mathbf{Lens}_A(\mathcal{C})$ such that a number of axioms are satisfied: precisely the first five axioms of a CRDC.

Proposition 10 ((compare [15, Prop. 31])). *A coalgebra of the copointed \mathbf{Lens}_A endofunctor gives rise to a cartesian left-additive category \mathcal{C} equipped with a reverse differential combinator R which satisfies the first five axioms of a cartesian reverse derivative category.*

Proof. A coalgebra of \mathbf{Lens}_A consists of a category \mathcal{C} and a cartesian left-additive functor $\mathbf{R}_{\mathcal{C}} : \mathcal{C} \rightarrow \mathbf{Lens}_A(\mathcal{C})$ such that the following diagram commutes.

$$\begin{array}{ccc} \mathbf{Lens}_A(\mathcal{C}) & \xrightarrow{\varepsilon_{\mathcal{C}}} & \mathcal{C} \\ \mathbf{R}_{\mathcal{C}} \uparrow & \nearrow & \\ \mathcal{C} & & \end{array} \quad (4)$$

The data of the functor $\mathbf{R}_{\mathcal{C}}$ is equivalent to the data of a reverse differential combinator R : every map $f : A \rightarrow B$ in \mathcal{C} is mapped to a lens whose forward part is by the Eq. 4 above restricted to be f itself, leaving the only choice involved in this functor the one of the backward part. What remains to prove is that this backward part satisfies the first five axioms of a CRDC. We do this in Appendix B. \square

We will see in the next section how only the data of the functor $\mathbf{R}_{\mathcal{C}} : \mathcal{C} \rightarrow \mathbf{Lens}_A(\mathcal{C})$ is used to model supervised learning, justifying our claim that only the first five axioms of a CRDC are used.

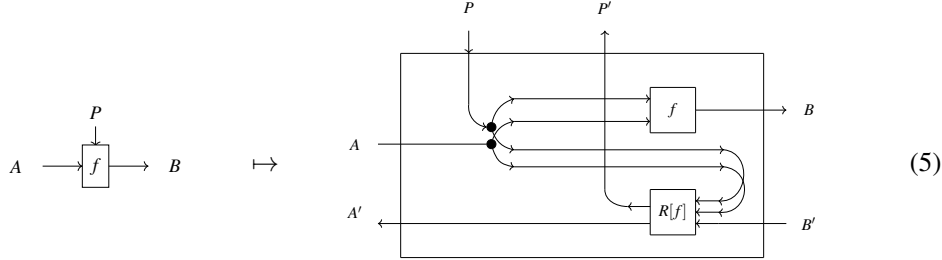
3. Components of learning as Parametric Lenses

As seen in the introduction, in the learning process there are many components at work: a model, an optimiser, a loss map, a learning rate, etc. In this section we show how each such component can be understood as a parametric lens. Moreover, for each component, we show how our framework encompasses several variations of the gradient-descent algorithms, thus offering a unifying perspective on many different approaches that appear in the literature.

3.1 Models as Parametric Lenses

We begin by characterising the models used for training as parametric lenses. In essence, our approach identifies a set of abstract requirements necessary to perform training by gradient descent, which covers the case studies that we will consider in the next sections.

The leading intuition is that a suitable model is a parametric map, equipped with a reverse derivative operator. Using the formal developments of Section 2, this amounts to assuming that a model is a morphism in $\mathbf{Para}(\mathcal{C})$, for a CRDC \mathcal{C} . In order to visualise such morphism as a parametric lens, it then suffices to apply under $\mathbf{Para}(-)$ the canonical morphism $\mathbf{R}_{\mathcal{C}}: \mathcal{C} \rightarrow \mathbf{Lens}(\mathcal{C})$ (which exists for any CRDC \mathcal{C} , see Prop. 10)^c, mapping f to $\left(\begin{smallmatrix} f \\ R[f] \end{smallmatrix}\right)$. This yields a functor $\mathbf{Para}(\mathbf{R}_{\mathcal{C}}): \mathbf{Para}(\mathcal{C}) \rightarrow \mathbf{Para}(\mathbf{Lens}(\mathcal{C}))$, pictorially defined as



Example 11 (Neural networks). As noted previously, to learn a function of type $\mathbb{R}^n \rightarrow \mathbb{R}^m$, one constructs a neural network, which can be seen as a function of type $\mathbb{R}^p \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ where \mathbb{R}^p is the space of parameters of the neural network. As seen in Example 2, this is a map in the category $\mathbf{Para}(\mathbf{Smooth})$ of type $\mathbb{R}^n \rightarrow \mathbb{R}^m$ with parameter space \mathbb{R}^p . Then one can apply the functor in (5) to present a neural network together with its reverse derivative operator as a parametric lens, i.e. a morphism in $\mathbf{Para}(\mathbf{Lens}(\mathbf{Smooth}))$.

Example 12 (Boolean and Polynomial circuits). For learning of Boolean circuits as described in [60], the recipe is the same as in Example 11, except that the base category is $\mathbf{POLY}_{\mathbb{Z}_2}$ (see Example 8). The important observation here is that $\mathbf{POLY}_{\mathbb{Z}_2}$ is a CRDC, see [15, 60], and thus we can apply the functor in (5). Note this setting can be generalised to circuits over any polynomial ring, see [62].

Note a model/parametric lens f takes as inputs an element of A , a parameter P , an element of B' (a change in B) and outputs an element of B , a change in A , and a change in P . This is not yet sufficient to do machine learning! When we perform learning, we want to input a parameter P and a pair $A \times B$ and receive a new parameter P . Instead, f expects a change in B (not an element of B) and outputs a change in P (not an element of P). Deep dreaming, on the other hand, wants to return an element of A (not a change in A). Thus, to do machine learning (or deep dreaming) we need to add additional components to f ; we will consider these additional components in the next sections.

We now proceed to describe the internals of a model, and translate several examples of models in deep learning to their categorical form. But before doing so, we clarify some terminology. While ‘layers’ and ‘models’ are both parametric maps, the former typically refers to *components* of larger models, while the latter refers to the final model to be learned in the manner described in Section 4.1).

Remark 13. An unfortunate ambiguity in deep learning terminology is the second meaning of ‘layer’. For example, the ‘hidden layer’ of a model refers to internal *values* of a neural network, corresponding to the ‘wires’ of a string diagram. We will avoid using the term ‘layer’ in this sense unless explicitly noted.

^cHere we are treating $\mathbf{R}_{\mathcal{C}}$ as postcomposed with the inclusion $\mathbf{Lens}_A(\mathcal{C}) \hookrightarrow \mathbf{Lens}(\mathcal{C})$.

In deep learning, one often speaks of ‘model architectures’. This can mean either a specific model (e.g., ResNet50 [32]) or a family of models employing a particular technique. For example, one says a model has a ‘convolutional architecture’ when it contains a convolutional layer (Example 20). Examples of layers, models, and architectures are given in the following sections.

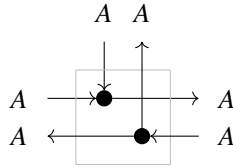
3.1.1 Layers

The *dense* or *fully-connected* layer is a component of many neural network architectures. In the categorical viewpoint, a dense layer is the composition of linear, bias, and activation layers, which we describe first. Unless explicitly noted, we will assume for simplicity that most layers are maps in **Para(Smooth)**. However, many of the maps defined here only require a *multiplication* map as additional structure, and so can be defined in any cartesian distributive category [61] such as POLY_S .

Example 14 (Linear layer). A **linear** or **matrix-multiply** layer is the parametric map $(\mathbb{R}^{nm}, \text{linear} : \mathbb{R}^{nm} \times \mathbb{R}^n \rightarrow \mathbb{R}^m)$, where $\text{linear}(M, x) = M \cdot x$ is the matrix-vector product of M interpreted as matrix coefficients and the vector x .

Note that layers are best thought of as *families* of maps. For example, there is a linear layer morphism for all choices of dimension m and n .

Example 15 (Bias layer). A **bias** layer is a parametric map $(\mathbb{R}^n, +)$, where $+: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the cartesian left-additive addition map. The reverse derivative of $+$ is the copy map, so we may depict the bias layer as a parametric lens as below.



An *activation layer* $(I, \alpha : A^n \rightarrow A^n)$ is a typically nonlinear, trivially parametric map often applied to the output of another layer. Many are simply the n -fold tensor product of a univariate function $A \rightarrow A$.

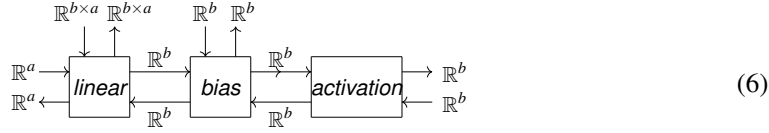
Example 16 (Sigmoid Activation). The **sigmoid** activation layer $(I, \text{sigmoid} : \mathbb{R}^n \rightarrow \mathbb{R}^n)$ is the n -fold tensor product $\sigma \times \dots \times \sigma$ of the **sigmoid function** $\sigma(x) = \frac{\exp(x)}{\exp(x)+1}$.

Note that unlike other layers considered so far, while **sigmoid** is a map in **Smooth**, it is not a map in POLY_S . An example of an activation layer which is *not* in **Smooth** is the ReLU map.

Example 17 (ReLU Activation). The ‘Rectified Linear Unit’ activation function is the map $\text{ReLU}(x) = \delta_{>0}(x) \cdot x$ where $\delta_{>0}$ is the positive indicator function. The ReLU activation layer $\text{ReLU} : A^n \rightarrow A^n$ is again the n -fold tensor product of this function. Although ReLU is not a smooth map, some presentations of RDCs can be extended via Theorem 3.1 of [61] to include the positive indicator function $\delta_{>0} : A \rightarrow A$. The ReLU function can then be expressed in terms of this function, and its reverse derivative can be derived as $R[\text{ReLU}](x, \delta_y) = \delta_{>0}(x) \cdot \delta_y$.

The combination of linear, bias, and choice of activation layer gives a *dense* or *fully-connected* layer.

Example 18 (Dense Layer). A *dense* or *fully-connected* layer is the following composition.



Where some choice of input dimension $m \in \mathbb{N}$, output dimension $m \in \mathbb{N}$, and activation layer $\alpha : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is assumed. Note that the activation layer has no parameters.

The final two examples of layers we cover here are *convolutional* and *max-pooling* layers, which are common in models for image-processing tasks.

Example 19 (Convolutional Layer). A *convolutional* layer is a map $(\mathbb{R}^{k^2}, \text{convolve2D} : \mathbb{R}^{k^2} \times \mathbb{R}^{m^2} \rightarrow \mathbb{R}^{n^2})$ where convolve2D denotes the discrete 2D convolution of a $k \times k$ kernel and an $m \times m$ image. The output of a convolutional layer is an $n \times n$ image with $n = \max(m, k) - \min(m, k) + 1$.

A number of further variations of the convolutional layer exist in the literature, but the basic idea is to use 2D convolution to map a kernel (the parameters) over the input. Convolutional layers are frequently composed with max-pooling layers.

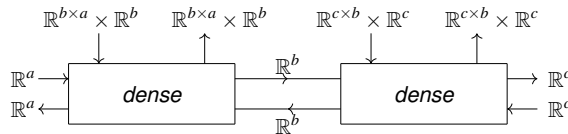
Example 20 (Max-Pooling Layer). A *max-pooling* layer $(I, \text{maxpool} : S^{(kn)^2} \rightarrow S^n)$ computes the maximum of each of the n^2 size- $k \times k$ subregions of the input image.

However, as with the ReLU layer, max-pooling layers cannot be thought of as maps in **Smooth**. Nevertheless, by again appealing to [61, Theorem 3.1], one can extend a presentation of RDCs to include a function $\max : 2 \rightarrow 1$ from which the max-pooling layer and its reverse derivative can be derived.

3.1.2 Architectures

We now consider some examples of neural network architectures defined in terms of the layers above. Since both layers and architectures are just parametric maps, we can consider the layers by themselves as architectures already, and in fact the *linear* and *dense* layers are sufficient to solve some simple machine learning problems. The first non-trivial architecture we consider is the ‘single hidden layer’ network.

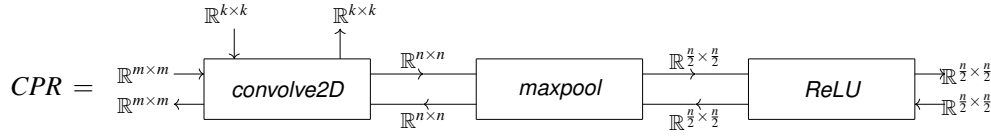
Example 21 (Hidden Layer Network). A neural network with a single ‘hidden layer’ (in the sense of Remark 13) is the composition of two dense maps.



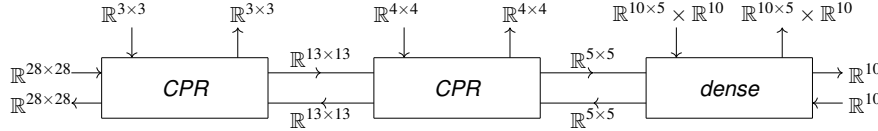
We emphasize that the term ‘hidden layer’ here ambiguously refers to the central wires labeled \mathbb{R}^b rather than the *dense* morphisms.

This architecture is demonstrated in detail in the experiments [19] accompanying this paper. Also included in our experiments is a convolutional model for classifying images of handwritten digits (the MNIST [37] dataset). A simplified version is below.

Example 22 (Convolutional Architecture). *First, define a CPR layer as the composition of a convolution, max-pooling, and ReLU layer:*



where $n = \max(m, k) - \min(m, k) + 1$. One can then define a convolutional architecture for classifying 28×28 -pixel images of the MNIST dataset into digits 0 – 9 as follows.

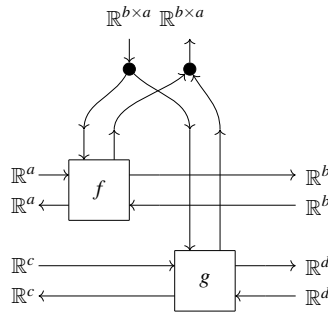


Lastly, we mention the architecture of Generative Adversarial Networks which we define and thoroughly discuss in Section 4.2.

3.1.3 Weight Tying and Batching

A number of general techniques are employed in designing deep learning models. We now describe two examples of these techniques in terms of their categorical interpretations. The first is *weight-tying*, which is required in Subsection 4.2 to define a more complex architecture for unsupervised learning: the GAN.

Example 23 (Weight Tying). *Weight tying is the sharing of parameters between two different components. Categorically speaking, this means using the copy map on parameters as below.*

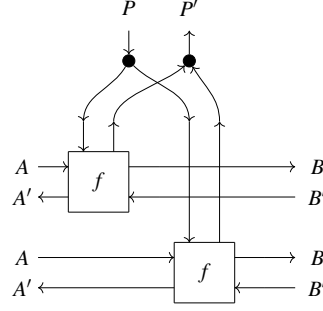


Note that f and g have the same parameters, but might be applied to different parts of the input. In this sense, one can think of convolutional layers (Example 20) as using weight-tying.

A related technique is *batching*. So far, we have considered learning as updating a model with a single data example (x, y) at each timestep. However, it is common for efficiency purposes to update the model using a *batch* of examples: a finite number n of examples (x_i, y_i) for $i \in \{0 \dots n\}$.

Example 24 (Batching). *Suppose we have a model $(P, f : P \times A \rightarrow B)$. The batched model with **batch size** n is a parametric map $(P, f' : P \times A^n \rightarrow B^n)$ where f' consists of n copies of f applied to each input, but with the same parameters. For example, when $n = 2$, the batch update is as*

follows.



The above diagrams highlight the relationship between weight-tying and batching. However, note that these techniques serve different purposes: while weight-tying can be used to reduce the number of weights in a model, batching is used for efficiency reasons to update a model with multiple examples in parallel.

3.2 Loss Maps as Parametric Lenses

Another key component of any learning algorithm is the choice of loss map. This gives a measurement of how far the current output of the model is from the desired output. In standard learning in **Smooth**, this loss map is viewed as a map of type $B \times B \rightarrow \mathbb{R}$. However, in our setup, this is naturally viewed as a parametric map from B to \mathbb{R} with parameter space B .^d We also generalize the codomain to an arbitrary object L .

Definition 25. A loss map on B consists of a parametric map $(B, \text{loss}) : \mathbf{Para}(\mathcal{C})(B, L)$ for some object L .

Note that we can precompose a loss map $(B, \text{loss}) : B \rightarrow L$ with a neural network $(P, f) : A \rightarrow B$ (below left), and apply the functor in (5) (with $\mathcal{C} = \mathbf{Smooth}$) to obtain the parametric lens below right.

$$\begin{array}{c} P \\ \downarrow \\ A \rightarrow [f] \rightarrow B \end{array} \quad \begin{array}{c} B \\ \downarrow \\ [\text{loss}] \rightarrow L \end{array} \quad \mapsto \quad \begin{array}{c} P \quad P' \\ \downarrow \quad \uparrow \\ A \rightarrow [f] \rightarrow B \\ A' \leftarrow [R[f]] \leftarrow B' \end{array} \quad \begin{array}{c} B \quad B' \\ \downarrow \quad \uparrow \\ [\text{loss}] \rightarrow L \\ [R[\text{loss}]] \leftarrow L' \end{array} \quad (7)$$

This is getting closer to the parametric lens we want: it can now receive inputs of type B . However, this is at the cost of now needing an input to L' ; we consider how to handle this in the next section.

Example 26 (Quadratic error). In **Smooth**, the standard loss function on \mathbb{R}^b is quadratic error: it uses $L = \mathbb{R}$ and has parametric map $e : \mathbb{R}^b \times \mathbb{R}^b \rightarrow \mathbb{R}$ given by

$$e(b_t, b_p) = \frac{1}{2} \sum_{i=1}^b ((b_p)_i - (b_t)_i)^2$$

^dHere the loss map has its parameter space equal to its input space. However, putting loss maps on the same footing as models lends itself to further generalizations where the parameter space is different, and where the loss map can itself be learned. See Generative Adversarial Networks, [10, Figure 7.].

where we think of b_t as the “true” value and b_p the predicted value. This has reverse derivative $R[e] : \mathbb{R}^b \times \mathbb{R}^b \times \mathbb{R} \rightarrow \mathbb{R}^b \times \mathbb{R}^b$ given by $R[e](b_t, b_p, \alpha) = \alpha \cdot (b_p - b_t, b_t - b_p)$ — note α suggests the idea of learning rate, which we will explore in Section 3.3.

Example 27 (Boolean error). In $\text{POLY}_{\mathbb{Z}_2}$, the loss function on \mathbb{Z}^b which is implicitly used in [60] is a bit different: it uses $L = \mathbb{Z}^b$ and has parametric map $e : \mathbb{Z}^b \times \mathbb{Z}^b \rightarrow \mathbb{Z}^b$ given by

$$e(b_t, b_p) = b_t + b_p.$$

(Note that this is $+$ in \mathbb{Z}_2 ; equivalently this is given by XOR.) Its reverse derivative is of type $R[e] : \mathbb{Z}^b \times \mathbb{Z}^b \times \mathbb{Z}^b \rightarrow \mathbb{Z}^b \times \mathbb{Z}^b$ given by $R[e](b_t, b_p, \alpha) = (\alpha, \alpha)$.

Example 28 (Softmax cross entropy). The Softmax cross entropy loss is a \mathbb{R}^b -parametric map $\mathbb{R}^b \rightarrow \mathbb{R}$ defined by

$$e(b_t, b_p) = \sum_{i=1}^b (b_t)_i ((b_p)_i - \log(\text{Softmax}(b_p)_i))$$

where $\text{Softmax}(b_p) = \frac{\exp((b_p)_i)}{\sum_{j=1}^b \exp((b_p)_j)}$ is defined componentwise for each class i .

We note that, although b_t needs to be a probability distribution, at the moment there is no need to ponder the question of interaction of probability distributions with the reverse derivative framework: one can simply consider b_t as the image of some logits under the Softmax function.

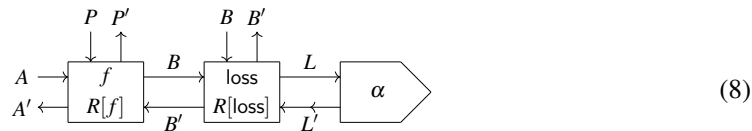
Example 29 (Dot product). In *Deep Dreaming* (Section 4.3) we often want to focus only on a particular element of the network output \mathbb{R}^b . This is done by supplying a one-hot vector b_t as the ground truth to the loss function $e(b_t, b_p) = b_t \cdot b_p$ which computes the dot product of two vectors. If the ground truth vector y is a one-hot vector (active at the i -th element), then the dot product performs masking of all inputs except the i -th one. Note the reverse derivative $R[e] : \mathbb{R}^b \times \mathbb{R}^b \times \mathbb{R} \rightarrow \mathbb{R}^b \times \mathbb{R}^b$ of the dot product is defined as $R[e](b_t, b_p, \alpha) = (\alpha \cdot b_p, \alpha \cdot b_t)$.

3.3 Learning Rates as Parametric Lenses

After models and loss maps, another ingredient of the learning process are *learning rates*, which we formalise as follows.

Definition 30. A *learning rate* α on L consists of a lens from $\begin{pmatrix} L \\ L' \end{pmatrix}$ to $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ where 1 is a terminal object in \mathcal{C} .

Note that the get component of the learning rate lens must be the unique map to 1, while the put component is a map $L \times 1 \rightarrow L'$; that is, simply a map $\alpha^* : L \rightarrow L'$. Thus we can view α as a parametric lens from $\begin{pmatrix} L \\ L' \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ (with trivial parameter space) and compose it in $\text{Para}(\text{Lens}(\mathcal{C}))$ with a model and a loss map (cf. (7)) to get



Example 31. In standard supervised learning in **Smooth**, one fixes some $\varepsilon > 0$ as a learning rate, and this is used to define α : α is simply constantly $-\varepsilon$, ie., $\alpha(l) = -\varepsilon$ for any $l \in L$.

Example 32. In supervised learning in $\text{POLY}_{\mathbb{Z}_2}$, the standard learning rate is quite different: for a given L it is defined as the identity function, $\alpha(l) = l$.

Other learning rate morphisms are possible as well: for example, one could fix some $\varepsilon > 0$ and define a learning rate in **Smooth** by $\alpha(l) = -\varepsilon \cdot l$. Such a choice would take into account how far away the network is from its desired goal and adjust the learning rate accordingly.

3.4 Optimisers as Reparameterisations

In this section we consider how to implement gradient descent, ascent, and other gradient updates into our framework. To this aim, note that the parametric lens $\left(\begin{smallmatrix} f \\ R[f] \end{smallmatrix} \right)$ representing our model (see (5)) outputs a P' , which represents a *change* in the parameter space. Now, we would like to receive not just the requested change in the parameter, but the new parameter itself. This is precisely what gradient update accomplishes, when formalised as a lens. We start by describing gradient ascent and gradient descent.

Definition 33 (Gradient ascent). Let \mathcal{C} be a CRDC. Gradient ascent on $P : \mathcal{C}$ is a lens

$$\left(\begin{smallmatrix} id_P \\ +_P \end{smallmatrix} \right) : \left(\begin{smallmatrix} P \\ P \end{smallmatrix} \right) \rightarrow \left(\begin{smallmatrix} P \\ P' \end{smallmatrix} \right)$$

where $+_P : P \times P' \rightarrow P$ is the monoid structure of P .^c

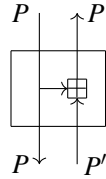


Figure 3: Gradient Ascent

Intuitively, such a lens allows one to receive the requested change in parameter and implement that change by adding that value to the current parameter. By its type, we can now “plug” the gradient descent lens $G : \left(\begin{smallmatrix} P \\ P \end{smallmatrix} \right) \rightarrow \left(\begin{smallmatrix} P \\ P' \end{smallmatrix} \right)$ above the model $\left(\begin{smallmatrix} f \\ R[f] \end{smallmatrix} \right)$ in (5) — formally, this is accomplished as a *reparameterisation* of the parametric morphism $\left(\begin{smallmatrix} f \\ R[f] \end{smallmatrix} \right)$, cf. Section 2.1. This gives us Figure 4 (left).

Example 34 (Gradient ascent in Smooth). In **Smooth**, the gradient ascent reparameterisation will take the output from P' and add it to the current value of P to get a new value of P .

^cNote that as in the discussion in Section 2.4, we are implicitly assuming that $P = P'$; we have merely notated them differently to emphasize the different “roles” they play (the first P can be thought of as “points”, the second as “vectors”).

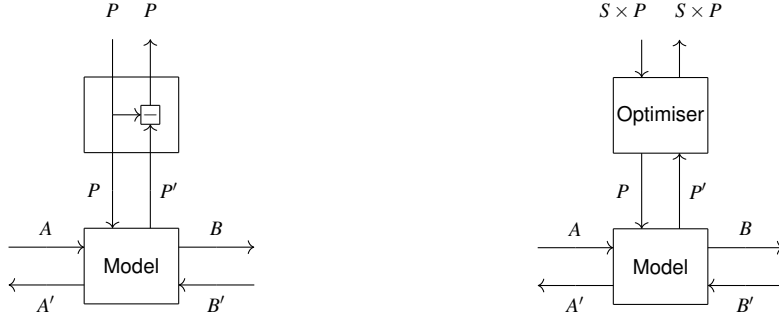


Figure 4: Model reparameterised by basic gradient descent (left) and a generic stateful optimiser (right).

Example 35 (Gradient ascent in Boolean circuits). *In the CRDC $\text{POLY}_{\mathbb{Z}_2}$, the gradient ascent reparameterisation will again take the output from P' and add it to the current value of P to get a new value of P ; however, since $+$ in \mathbb{Z}_2 is the same as XOR, this can also be seen as taking the XOR of the current parameter and the requested change; this is exactly how this algorithm is implemented in [60].*

Definition 36 (Gradient descent). *Let \mathcal{C} be a CRDC where every monoid is additionally a commutative group.^f Gradient descent on P is a lens*

$$\begin{pmatrix} id_P \\ -_P \end{pmatrix} : \begin{pmatrix} P \\ P \end{pmatrix} \rightarrow \begin{pmatrix} P \\ P' \end{pmatrix}$$

where $-_P : (p, p') = p - p'$.

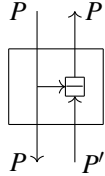


Figure 5: Gradient Descent

In **Smooth** this instantiates to usual gradient descent. Gradient ascent in $\text{POLY}_{\mathbb{Z}_2}$ is equal to gradient descent because XOR is its own inverse. Intuitively in $\text{POLY}_{\mathbb{Z}_2}$ there is always only one direction we can move (other than staying still): it's flipping the bit. Gradient descent and ascent are not usually seen as a lens — but they fit precisely into this picture that we are creating.

Other variants of gradient descent also fit naturally into this framework by allowing for additional input/output data with P . In particular, many of them keep track of the history of previous updates and use that to inform the next one. This is easy to model in our setup: instead of asking for a lens $\begin{pmatrix} P \\ P \end{pmatrix} \rightarrow \begin{pmatrix} P \\ P' \end{pmatrix}$, we ask instead for a lens $\begin{pmatrix} S \times P \\ S \times P \end{pmatrix} \rightarrow \begin{pmatrix} P \\ P' \end{pmatrix}$ where S is some “state” object.

^fSince a homomorphism between groups needs to satisfy *less* equations than a monoid homomorphism, this means that every monoid homomorphism is also a group homomorphism. This in turn means there are no extra conditions we need to impose on such a CRDC equipped with group objects.

Definition 37. A *stateful parameter update* consists of a choice of object S (the *state* object) and a lens $U : \begin{pmatrix} S \times P \\ S \times P \end{pmatrix} \rightarrow \begin{pmatrix} P \\ P' \end{pmatrix}$.

Again, we view this optimiser as a reparameterisation which may be “plugged in” a model as in Figure 4 (right). Let us now consider how several well-known optimisers can be implemented in this way.

Example 38 (Momentum). In the momentum variant of gradient descent, one keeps track of the previous change and uses this to inform how the current parameter should be changed. Thus, in this case, we set $S = P$, fix some $\gamma > 0$, and define the **momentum** lens $\begin{pmatrix} U \\ U^* \end{pmatrix} : \begin{pmatrix} P \times P \\ P \times P \end{pmatrix} \rightarrow \begin{pmatrix} P \\ P' \end{pmatrix}$ by $U(s, p) = p$ and $U^*(s, p, p') = (s', p + s')$, where $s' = -\gamma s + p'$. Note momentum recovers gradient descent when $\gamma = 0$.

In both standard gradient descent and momentum, our lens representation has trivial get part. However, as soon as we move to more complicated variants, this is not anymore the case, as for instance in Nesterov momentum below.

Example 39 (Nesterov momentum). In Nesterov momentum, one uses the momentum from previous updates to tweak the input parameter supplied to the network. We can precisely capture this by using a small variation of the lens in the previous example. Again, we set $S = P$, fix some $\gamma > 0$, and define the **Nesterov momentum** lens $\begin{pmatrix} U \\ U^* \end{pmatrix} : \begin{pmatrix} P \times P \\ P \times P \end{pmatrix} \rightarrow \begin{pmatrix} P \\ P' \end{pmatrix}$ by $U(s, p) = p + \gamma s$ and U^* as in the previous example.

Example 40 (Adagrad). Given any fixed $\varepsilon > 0$ and $\delta \sim 10^{-7}$, Adagrad [23] is given by $S = P$, with the lens whose get part is $(g, p) \mapsto p$. The put is $(g, p, p') \mapsto (g', p + \frac{\varepsilon}{\delta + \sqrt{g'}} \odot p')$ where $g' = g + p' \odot p'$ and \odot is the elementwise (Hadamard) product. Unlike with other optimization algorithms where the learning rate is the same for all parameters, Adagrad divides the learning rate of each individual parameter with the square root of the past accumulated gradients.

Example 41 (Adam). Adaptive Moment Estimation (Adam) [36] is another method that computes adaptive learning rates for each parameter by storing exponentially decaying average of past gradients (m) and past squared gradients (v). For fixed $\beta_1, \beta_2 \in [0, 1)$, $\varepsilon > 0$, and $\delta \sim 10^{-8}$, Adam is given by $S = P \times P$, with the lens whose get part is $(m, v, p) \mapsto p$ and whose put part is $\text{put}(m, v, p, p') = (\hat{m}', \hat{v}', p + \frac{\varepsilon}{\delta + \sqrt{\hat{v}'}} \odot \hat{m}')$ where $m' = \beta_1 m + (1 - \beta_1)p'$, $v' = \beta_2 v + (1 - \beta_2)p'^2$, and $\hat{m}' = \frac{m'}{1 - \beta_1^t}$, $\hat{v}' = \frac{v'}{1 - \beta_2^t}$.

Note that, so far, optimisers/reparameterisations have been added to the P/P' wires. In order to change the model’s parameters (Fig. 4). In Section 4.3 we will study them on the A/A' wires instead, giving *deep dreaming*.

3.4.1 Can we compose optimisers?

Even though not explicitly acknowledged in the literature, optimisers can be composed, and this composition plays an important role in settings where deep learning intersects with multivariable optimisation. In such settings we’re interested in their *parallel* composition, therefore giving a

positive answer to the above question.[‡] Parallel composition of optimisers arises out of the fact that optimisers are lenses, and lenses are a monoidal category (Remark 4). In such settings we might have an optimiser of two variables which descends in one direction, and ascends in the other one, for instance.

Definition 42 (Gradient descent-ascent (GDA)). *Given objects P and Q , gradient descent-ascent on $P \times Q$ is a lens*

$$\left(\begin{smallmatrix} P \times Q \\ \text{gda} \end{smallmatrix} \right) : \left(\begin{smallmatrix} P \times Q \\ P \times Q \end{smallmatrix} \right) \rightarrow \left(\begin{smallmatrix} P \times Q \\ P' \times Q' \end{smallmatrix} \right)$$

where $\text{gda}(p, q, p', q') = (p - p', q + q')$.

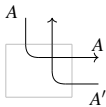
In **Smooth** this gives an optimiser which descends on P and ascends on Q . In $\text{POLY}_{\mathbb{Z}_2}$ this map ends up computing the same update function on both parameter spaces: the one that just flips the underlying bit. This is something that ends up preventing us from modelling GANs in this setting (compare with Ex. 48 where both positive and negative polarity of the optimiser map is needed).

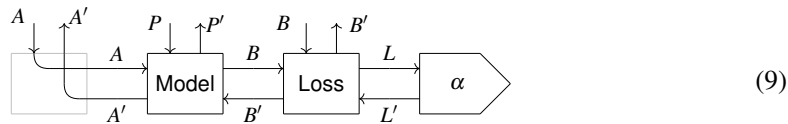
When it comes to optimisers of two parameters, gradient descent-ascent is a particular type of an optimiser that is a product of two optimisers. But not all optimisers can be factored in such a way, much like a general monoidal product doesn't necessarily have to be cartesian. A good example of this is an optimiser on two parameters called *competitive gradient descent* ([47]). We don't explicitly define or use it in this paper, instead inviting the reader to the aforementioned reference for more information.

4. Learning with Parametric Lenses

In the previous section we have seen how all the components of learning can be modeled as parametric lenses. We now study how all these components can be put together to form learning systems. We cover the most common examples of supervised learning, also discussing different kinds of layers, architectures, and techniques such as weight tying and batching. We also consider unsupervised learning, in the form of Generative Adversarial Networks. Finally, in addition to systems that learn *parameters*, we study systems that learn their *inputs*. This is a technique commonly known as deep dreaming, and we present it as a natural counterpart of supervised learning of parameters.

Before we describe these systems, it will be convenient to represent all the inputs and outputs of our parametric lenses as parameters. In (8), we see the P/P' and B/B' inputs and outputs as parameters; however, the A/A' wires are not. To view the A/A' inputs as parameters, we compose that system with the parametric lens η we now define. The parametric lens η has the type $\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \right) \rightarrow \left(\begin{smallmatrix} A \\ A' \end{smallmatrix} \right)$ with parameter space $\left(\begin{smallmatrix} A \\ A' \end{smallmatrix} \right)$ defined by $(\text{get}_\eta = 1_A, \text{put}_\eta = \pi_1)$ and can be depicted

graphically as . Composing η with the rest of the learning system in (8) gives us the closed parametric lens



[‡]One might wonder whether optimisers can be composed in sequence as well. The apparent sequential composability of optimisers is unfortunately an artefact of our limited view without dependent types.

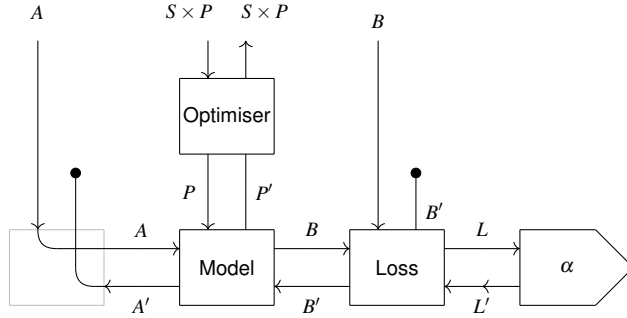
This composite is now a map in $\mathbf{Para}(\mathbf{Lens}(\mathcal{C}))$ from $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ to $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$; all its inputs and outputs are now vertical wires, ie., parameters. Unpacking it further, this is a lens of type $\begin{pmatrix} A \times P \times B \\ A' \times P' \times B' \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ whose get map is the terminal map, and whose put map is of the type $A \times P \times B \rightarrow A' \times P' \times B'$. It can be unpacked as the composite

$$\begin{aligned} \text{put}(a, p, b_t) &= (a', p', b'_t) & \text{where} & & b_p &= f(p, a) \\ & & & & (b'_t, b'_p) &= R[\text{loss}](b_t, b_p, \alpha(\text{loss}(b_t, b_p))) \\ & & & & (p', a') &= R[f](p, a, b'_p) \end{aligned}$$

In the next two sections we consider further additions to the image above which correspond to different types of supervised learning.

4.1 Supervised Learning of Parameters

The most common type of learning performed on (9) is supervised learning of *parameters*. This is done by reparameterising (cf. Section 2.1) the image in the following manner. The parameter ports are reparameterised by one of the (possibly stateful) optimisers described in the previous section, while the backward wires A' of inputs and B' of outputs are discarded. This finally yields the complete picture of a system which learns the parameters in a supervised manner:



Fixing a particular optimiser $\begin{pmatrix} U \\ U^* \end{pmatrix} : \begin{pmatrix} S \times P \\ S \times P \end{pmatrix} \rightarrow \begin{pmatrix} P \\ P' \end{pmatrix}$ we again unpack the entire construction. This is a map in $\mathbf{Para}(\mathbf{Lens}(\mathcal{C}))$ from $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ to $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ whose parameter space is $\begin{pmatrix} A \times S \times P \times B \\ S \times P \end{pmatrix}$. In other words, this is a lens of type $\begin{pmatrix} A \times S \times P \times B \\ S \times P \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ whose get component is the terminal map. Its put map has the type $A \times S \times P \times B \rightarrow S \times P$ and unpacks to $\text{put}(a, s, p, b_t) = U^*(s, p, p')$, where

$$\begin{aligned} \text{put}(a, s, p, b_t) &= U^*(s, p, p') & \text{where} & & \bar{p} &= U(s, p) \\ & & & & b_p &= f(\bar{p}, a) \\ & & & & (b'_t, b'_p) &= R[\text{loss}](b_t, b_p, \alpha(\text{loss}(b_t, b_p))) \\ & & & & (p', a') &= R[f](\bar{p}, a, b'_p) \end{aligned}$$

While this formulation might seem daunting, we note that it just explicitly specifies the computation performed by a supervised learning system. The variable \bar{p} represents the parameter supplied to the network by the stateful gradient update rule (in many cases this is equal to p); b_p represents the prediction of the network (contrast this with b_t which represents the ground truth from the dataset). Variables with a tick ' represent changes: b'_p and b'_t are the changes on predictions and true values respectively, while p' and a' are changes on the parameters and inputs.

Furthermore, this arises automatically out of the rule for lens composition (3); what we needed to specify is just the lenses themselves.

We justify and illustrate our approach on a series of case studies drawn from the literature. This presentation has the advantage of treating all these instances uniformly in terms of basic constructs, highlighting their similarities and differences. First, we fix some parametric map $(\mathbb{R}^p, f) : \mathbf{Para}(\mathbf{Smooth})(\mathbb{R}^a, \mathbb{R}^b)$ in **Smooth** and the constant *negative* learning rate $\alpha : \mathbb{R}$ (Example 31). We then vary the loss function and the gradient update, seeing how the put map above reduces to many of the known cases in the literature.

Example 43 (Quadratic error, basic gradient descent). *Fix the quadratic error (Example 26) as the loss map and basic gradient update (Example 34). Then the aforementioned put map simplifies. Since there is no state, its type reduces to $A \times P \times B \rightarrow P$, and we have $\text{put}(a, p, b_t) = p + p'$, where $(p', a') = R[f](p, a, \alpha \cdot (f(p, a) - b_t))$.*

Note that α here is simply a constant, and due to the linearity of the reverse derivative (Def 6), we can slide the α from the costate into the basic gradient update lens. Rewriting this update, and performing this sliding we obtain a closed form update step

$$\text{put}(a, p, b_t) = p + \alpha \cdot (R[f](p, a, f(p, a) - b_t); \pi_0)$$

where the negative descent component of gradient descent is here contained in the choice of the negative constant α .

This example gives us a variety of *regression* algorithms solved iteratively by gradient descent: it embeds some parametric map $(\mathbb{R}^p, f) : \mathbb{R}^a \rightarrow \mathbb{R}^b$ into the system which performs regression on input data - where a denotes the input to the model and b_t denotes the ground truth. If the corresponding f is linear and $b = 1$, we recover simple linear regression with gradient descent. If the codomain is multi-dimensional, i.e. we are predicting multiple scalars, then we recover multivariate linear regression. Likewise, we can model a multi-layer perceptron or even more complex neural network architectures performing supervised learning of parameters simply by changing the underlying parametric map.

Example 44 (Softmax cross entropy, basic gradient descent). *Fix Softmax cross entropy (Example 28) as the loss map and basic gradient update (Example 34). Again the put map simplifies. The type reduces to $A \times P \times B \rightarrow P$ and we have*

$$\text{put}(a, p, b_t) = p + p'$$

where $(p', a') = R[f](\bar{p}, a, \alpha \cdot (\text{Softmax}(f(p, a)) - b_t))$. The same rewriting performed on the previous example can be done here.

This example recovers *logistic regression*, e.g. classification.

Example 45 (Mean squared error, Nesterov Momentum). *Fix the quadratic error (Example 26) as the loss map and Nesterov momentum (Example 39) as the gradient update. This time the put map $A \times S \times P \times B \rightarrow S \times P$ does not have a simplified type. The implementation of put reduces to*

$$\begin{aligned} \text{put}(a, s, p, b_t) &= (s', p + s') & \text{where} & & \bar{p} &= p + \gamma s \\ & & & & (p', a') &= R[f](\bar{p}, a, \alpha \cdot (f(\bar{p}, a) - b_t)) \\ & & & & s' &= -\gamma s + p' \end{aligned}$$

This example with Nesterov momentum differs in two key points from all the other ones: i) the optimiser is stateful, and ii) its get map is not trivial. While many other optimisers are stateful,

the non-triviality of the get map here showcases the importance of lenses. They allow us to make precise the notion of computing a “lookahead” value for Nesterov momentum, something that is in practice usually handled in ad-hoc ways. Here, the algebra of lens composition handles this case naturally by using the get map, a seemingly trivial, unused piece of data for previous optimisers.

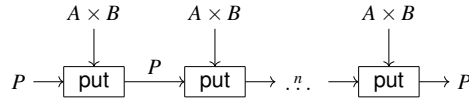
Our last example, using a different base category $\text{POLY}_{\mathbb{Z}_2}$, shows that our framework captures learning in not just continuous, but discrete settings too. Again, we fix a parametric map $(\mathbb{Z}^p, f) : \text{POLY}_{\mathbb{Z}_2}(\mathbb{Z}^a, \mathbb{Z}^b)$ but this time we fix the identity learning rate (Example 32), instead of a constant one.

Example 46 (Basic learning in Boolean circuits). *Fix XOR as the loss map (Example 27) and the basic gradient update (Example 35). The put map again simplifies. The type reduces to $A \times P \times B \rightarrow P$ and the implementation to $\text{put}(a, p, b_t) = p + p'$ where $(p', a') = R[f](p, a, f(p, a) + b_t)$.*

A sketch of learning iteration. Having described a number of examples in supervised learning, we outline how to model learning iteration in our framework. Recall the aforementioned put map whose type is $A \times P \times B \rightarrow P$ (for simplicity here modelled without state S). This map takes an input-output pair (a_0, b_0) , the current parameter p_i and produces an updated parameter p_{i+1} . At the next time step, it takes a potentially different input-output pair (a_1, b_1) , the updated parameter p_{i+1} and produces p_{i+2} . This process is then repeated. We can model this iteration as a composition of the put map with itself, as a composite $(A \times \text{put} \times B); \text{put}$ whose type is $A \times A \times P \times B \times B \rightarrow P$. This map takes two input-output pairs $A \times B$, a parameter and produces a new parameter by processing these datapoints in sequence. One can see how this process can be iterated any number of times, and even represented as a string diagram.

But we note that with a slight reformulation of the put map, it is possible to obtain a conceptually much simpler definition. The key insight lies in seeing that the map $\text{put} : A \times P \times B \rightarrow P$ is essentially an endo-map $P \rightarrow P$ with some extra inputs $A \times B$; it’s a parametric map!

In other words, we can recast the put map as a parametric map $(A \times B, \text{put}) : \mathbf{Para}(\mathcal{C})(P, P)$. Being an endo-map, it can be composed with itself. The resulting composite is an endo-map taking two “parameters”: input-output pair at the time step 0 and time step 1. This process can then be repeated, with **Para** composition automatically taking care of the algebra of iteration.



This reformulation captures the essence of parameter iteration: one can think of it as a trajectory $p_i, p_{i+1}, p_{i+2}, \dots$ through the parameter space; but it is a *trajectory parameterised by the dataset*. With different datasets the algorithm will take a different path through this space and learn different things.

4.2 Unsupervised Learning

Many kinds of systems that are traditionally considered unsupervised can be recast to their supervised form. One example is a Generative Adversarial Network ([30, 3]). This is a neural network architecture that lies in the centre of the intersection of deep learning and game theory. It is a system of two neural networks trained with “competing” optimisers. One neural network is called *the generator* whose optimiser is, as usual, tasked with moving in the direction of the negative gradient of the loss. However, the other network — called *the discriminator* — has an optimiser which is tasked with moving in the *positive*, i.e. ascending direction of the gradient of the total loss — maximising the loss. The actual networks are wired in such a way (Fig. 6) where the

discriminator effectively serves as a loss function to the generator, i.e. being the generator's only source of information on how to update. Dually, taking the vantage point of the discriminator, the generator serves as an ever changing source of training data.

Definition 47 (GAN). Fix three objects Z, X and L in \mathcal{C} (respectively called “the latent space”, “the data space” and “the payoff space”). Then given two parametric morphisms

$$(P, g) : \mathbf{Para}(\mathcal{C})(Z, X) \quad \text{and} \quad (Q, d) : \mathbf{Para}(\mathcal{C})(X, L)$$

a **generative adversarial network** is a morphism $(P \times Q, \text{GAN}_{g,d}) : \mathbf{Para}(\mathcal{C})(Z \times X, L \times L)$ where $\text{GAN}_{g,d}$ is defined as the composite

$$Z \times X \times P \times Q \cong Z \times P \times X \times Q \xrightarrow{g \times X \times \Delta_Q} X \times X \times Q \times Q \cong X \times Q \times X \times Q \xrightarrow{d \times d} L \times L$$

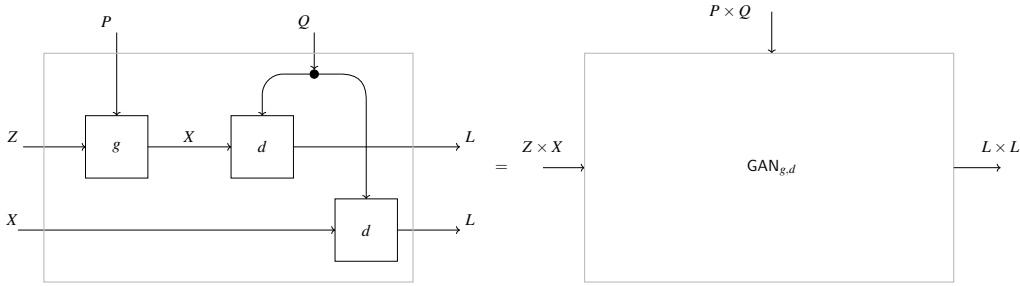


Figure 6: A generative adversarial network as a parametric morphism.

Its string diagram representation is shown in Fig. 6 where we see that a GAN consists of two parallel tracks. We will see in Ex. 48 how the first one will be used to process latent vectors, and the second one to process samples from a chosen dataset. Despite the fact that there are two boxes labeled d , they are weight tied, making them behave like a singular unit.

We can easily state what the reverse derivative of $\text{GAN}_{g,d}$ is in terms of its components:

$$\begin{aligned} R[\text{GAN}_{g,d}](z, x_r, p, q, \alpha_g, \alpha_r) &= (z', x'_r, p', q'_g + q'_r) \quad \text{where} \quad \begin{aligned} (x'_g, q'_g) &= R[d](g(z, p), q, \alpha_g) \\ (x'_r, q'_r) &= R[d](x_r, q, \alpha_r) \\ (z', p') &= R[g](z, p, x'_g) \end{aligned} \end{aligned} \tag{10}$$

The pair $(\text{GAN}_{g,d}, R[\text{GAN}_{g,d}])$ yields a parametric lens of type $(Z \times X)(Z' \times X') \rightarrow (L \times L)(L' \times L')$ (Fig. 7), which we interpret as follows.

It consumes two pieces of data, “a latent vector” $z : Z$, a “real” sample from the dataset $x_r : X$, in addition to the parameter $p : P$ for the generator and a parameter $q : Q$ for the discriminator. What happens then are two independent evaluations done by the discriminator. The first one uses the generator’s attempt of producing a sample from the dataset (the latent vector which was fed into it, producing $g(z, p) : X$) as input to the discriminator, producing a payoff $d((g, z, p), q) : L$ for this particular sample. The second one uses the actual sample from the dataset x_r , producing the payoff $d(x_r, q) : L$.

By choosing $\text{GAN}_{g,d}$ as the parametric map representing our supervised learning model, we can differentiate it as in (Fig. 7), and, with the appropriate choice of a loss function, produce the learning system in the literature called *Wasserstein GAN* ([3]).

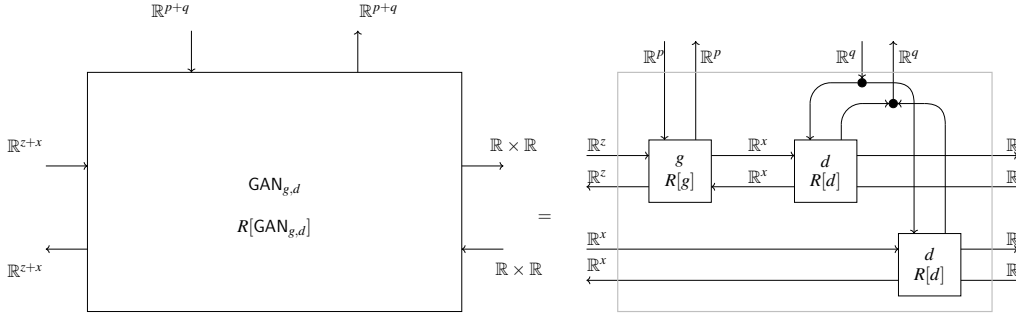


Figure 7: A generative adversarial network under the image of $\text{Para}(\mathbf{R}_{\mathcal{C}})$.

Example 48 (GANs, Dot product, GDA). Fix $\text{GAN}_{g,d}$ as the parametric map (Def. 47), gradient descent-ascent (Ex. 42) as the optimiser and dot product (Ex. 29) as the loss function. Then put becomes a map of type $Z \times X \times P \times Q \times L \rightarrow P \times Q$ and its implementation reduces to

$$\text{put}(z, x, p, q, b_t) = (p - p', q + q') \quad \text{where} \quad (z', x', p', q') = R[\text{GAN}_{g,d}](z, x, p, q, \alpha \cdot b_t)$$

We can further unpack the label

$$\begin{aligned} \text{put}(z, x_r, p, q, b_{tg}, b_{tr}) &= (p - p', q + q'_g + q'_r) \quad \text{where} \quad \begin{aligned} (x'_g, q'_g) &= R[d](g(z, p), q, \alpha \cdot b_{tg}) \\ (z', p') &= R[g](z, p, x'_g) \\ (x'_r, q'_r) &= R[d](x_r, q, \alpha \cdot b_{tr}) \end{aligned} \end{aligned}$$

This brings us to the last step, where by linearity of the backward pass we can extract α and components of b_t out:

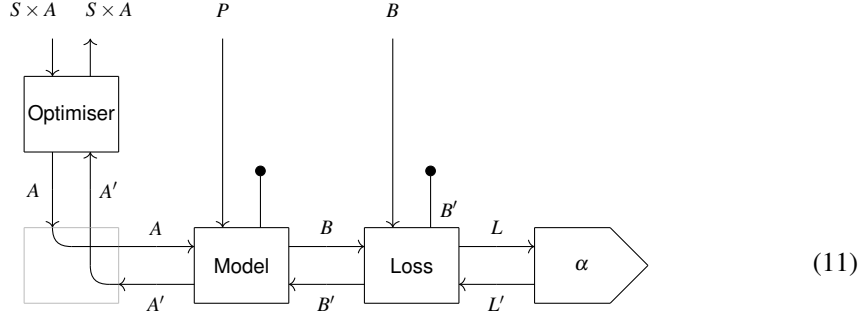
$$\begin{aligned} \text{put}(z, x_r, p, q, b_{tg}, b_{tr}) &= (p - \alpha b_{tg} p', q + \alpha(b_{tg} q'_g + b_{tr} q'_r)) \quad \text{where} \quad \begin{aligned} (x'_g, q'_g) &= R[d](g(z, p), q, 1) \\ (z', p') &= R[g](z, p, x'_g) \\ (x'_r, q'_r) &= R[d](x_r, q, 1) \end{aligned} \end{aligned}$$

The ultimate representation is a form in which it makes it possible to see how the update recovers that of Wasserstein GANs. The last missing piece is to note that the supervision labels y_t here are effectively “masks”. Just like in standard supervised learning an input-output pair (x_i, y_i) consisted of an input value and a corresponding label which guided the direction in which the output $f(x_i, p)$ should’ve been improved, here the situation is the same. Given any latent vector z its corresponding “label” is the learning signal $b_{tg} = 1$ which does not change anything in the update, effectively signaling to the generator’s and discriminator’s optimisers that they should descend (minimizing the assigned loss, making the image more realistic next time), and respectively ascend (maximizing the loss, becoming better at detecting when its input is a sample generated by the generator). On the other hand, given any real sample x_r its corresponding “label” is the learning signal $b_{tr} = -1$ which signals to the discriminator’s optimiser that it should do the opposite of what it usually does; it should descend, causing it to assign a lower loss value actual samples from the dataset. In other words, the input-output pairs here are always of the form $((z, x)_i, (1, -1)_i)$, making this GAN in many ways a *constantly* supervised model. Nonetheless, these different “forces” that pull the discriminator in different directions depending on the source of the input, coupled with the ever-changing generated inputs make GANs have intrinsically complex dynamics that are still being studied.

The fact that we were able to encode Wasserstein GAN in this form in our framework is a consequence of its simple formulation of its loss function, which is effectively given by subtraction [3, Theorem 3].

4.3 Deep Dreaming: Supervised Learning of Inputs

We have seen that reparameterising the parameter port with gradient descent allows us to capture supervised parameter learning. In this section we describe how reparameterising the *input port* provides us with a way to enhance an input image to elicit a particular interpretation. This is the idea behind the technique called Deep Dreaming, appearing in the literature in many forms [40, 38, 22, 52].



Deep dreaming is a technique which uses the parameters p of some trained classifier network to iteratively dream up, or amplify some features of a class b on a chosen input a . For example, if we start with an image of a landscape a_0 , a label b of a “cat” and a parameter p of a sufficiently well-trained classifier, we can start performing “learning” as usual: computing the predicted class for the landscape a_0 for the network with parameters p , and then computing the distance between the prediction and our label of a cat b . When performing backpropagation, the respective changes computed for each layer tell us how the activations of that layer should have been changed to be more “cat” like. This includes the first (input) layer of the landscape a_0 . Usually, we discard this changes and apply gradient update to the parameters. In deep dreaming we *discard the parameters* and *apply gradient update to the input* (see (11)). Gradient update here takes these changes and computes a new image a_1 which is the same image of the landscape, but changed slightly so to look more like whatever the network thinks a cat looks like. This is the essence of deep dreaming, where iteration of this process allows networks to dream up features and shapes on a particular chosen image [39].

Just like in the previous subsection, we can write this deep dreaming system as a map in $\mathbf{Para}(\mathbf{Lens}(\mathcal{C}))$ from $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ to $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ whose parameter space is $\begin{pmatrix} S \times A \times P \times B \\ S \times A \end{pmatrix}$. In other words, this is a lens of type $\begin{pmatrix} S \times A \times P \times B \\ S \times A \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ whose get map is trivial. Its put map has the type $S \times A \times P \times B \rightarrow S \times A$ and unpacks to

$$\begin{aligned} \text{put}(s, a, p, b_t) &= U^*(s, a, a') & \text{where} & & \bar{a} &= U(s, a) \\ & & & & b_p &= f(p, \bar{a}) \\ & & & & (b'_t, b'_p) &= R[\text{loss}](b_t, b_p, \alpha(\text{loss}(b_t, b_p))) \\ & & & & (p', a') &= R[f](p, \bar{a}, b'_p) \end{aligned}$$

We note that deep dreaming is usually presented without any loss function as a maximisation of a particular activation in the last layer of the network output [52, Section 2.]. This maximisation is done with gradient ascent, as opposed to gradient descent. However, this is just a special case of

our framework where the loss function is the dot product (Example 29). The choice of the particular activation is encoded as a one-hot vector, and the loss function in that case essentially masks the network output, leaving active only the particular chosen activation. The final component is the gradient *ascent*: this is simply recovered by choosing a positive, instead of a negative learning rate [52]. We explicitly unpack this in the following example.

Example 49 (Deep dreaming, dot product loss, basic gradient update). *Fix **Smooth** as base category, a parametric map $(\mathbb{R}^p, f) : \mathbf{Para}(\mathbf{Smooth})(\mathbb{R}^a, \mathbb{R}^b)$, the dot product loss (Example 29), basic gradient update (Example 34), and a positive learning rate $\alpha : \mathbb{R}$. Then the above put map simplifies. Since there is no state, its type reduces to $A \times P \times B \rightarrow A$ and its implementation to*

$$\text{put}(a, p, b_t) = a + a' \quad \text{where } (p', a') = R[f](p, a, \alpha \cdot b_t).$$

Like in Example 43, this update can be rewritten as

$$\text{put}(a, p, b_t) = a + \alpha \cdot (R[f](p, a, b_t); \pi_1)$$

making a few things apparent. This update does not depend on the prediction $f(p, a)$: no matter what the network has predicted, the goal is always to maximise particular activations. Which activations? The ones chosen by b_t . When b_t is a one-hot vector, this picks out the activation of just one class to maximise, which is often done in practice.

While we present only the most basic image, there is plenty of room left for exploration. The work of [52, Section 2.] adds an extra regularisation term to the image. In general, the neural network f is sometimes changed to copy a number of internal activations which are then exposed on the output layer. Maximising all these activations often produces more visually appealing results. In the literature we did not find an example which uses the Softmax-cross entropy (Example 28) as a loss function in deep dreaming, which seems like the more natural choice in this setting. Furthermore, while deep dreaming commonly uses basic gradient descent, there is nothing preventing the use of any of the optimiser lenses discussed in the previous section, or even doing deep dreaming in the context of Boolean circuits. Lastly, learning iteration which was described in at the end of previous subsection can be modelled here in an analogous way.

5. Implementation

We provide a proof-of-concept implementation as a Python library — full usage examples, source code, and experiments can be found at [19]. We demonstrate the correctness of our library empirically using a number of experiments implemented both in our library and in Keras [12], a popular framework for deep learning. For example, one experiment is a model for the MNIST image classification problem [37]: we implement the same model in both frameworks and achieve comparable accuracy. Note that despite similarities between the user interfaces of our library and of Keras, a model in our framework is constructed as a composition of parametric lenses. This is fundamentally different to the approach taken by Keras and other existing libraries, and highlights how our proposed algebraic structures naturally guide programming practice

In summary, our implementation demonstrates the advantages of our approach. Firstly, computing the gradients of the network is greatly simplified through the use of lens composition. Secondly, model architectures can be expressed in a principled, mathematical language; as morphisms of a monoidal category. Finally, the modularity of our approach makes it easy to see how various aspects of training can be modified: for example, one can define a new optimization algorithm simply by defining an appropriate lens. We now give a brief sketch of our implementation.

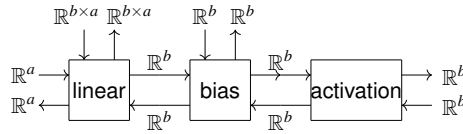
5.1 Constructing a Model with Lens and Para

We model a lens $\begin{pmatrix} f \\ f^* \end{pmatrix}$ in our library with the `Lens` class, which consists of a pair of maps `fwd` and `rev` corresponding to f and f^* , respectively. For example, we write the identity lens $\begin{pmatrix} 1_A \\ \pi_2 \end{pmatrix}$ as follows:

```
identity = Lens(lambda x: x, lambda x_dy: x_dy[1])
```

The composition (in diagrammatic order) of `Lens` values `f` and `g` is written `f >> g`, and monoidal composition as `f @ g`. Similarly, the type of **Para** maps is modeled by the `Para` class, with composition and monoidal product written the same way. Our library provides several primitive `Lens` and `Para` values.

Let us now see how to construct a single layer neural network from the composition of such primitives. Diagrammatically, we will construct a model consisting of a single dense layer, as in Example 18 and below.



Recall that the parameters of `linear` are the coefficients of a $b \times a$ matrix, and the underlying lens has as its forward map the function $(M, x) \rightarrow M \cdot x$, where M is the $b \times a$ matrix whose coefficients are the $\mathbb{R}^{b \times a}$ parameters, and $x \in \mathbb{R}^a$ is the input vector. The `bias` map is even simpler: the forward map of the underlying lens is simply pointwise addition of inputs and parameters: $(b, x) \rightarrow b + x$. Finally, the `activation` map simply applies a nonlinear function (e.g., `sigmoid`) to the input, and thus has the trivial (unit) parameter space. The representation of this composition in code is straightforward: we can simply compose the three primitive `Para` maps as in (6):

```
def dense(a, b, activation):
    return linear(a, b) >> bias(b) >> activation
```

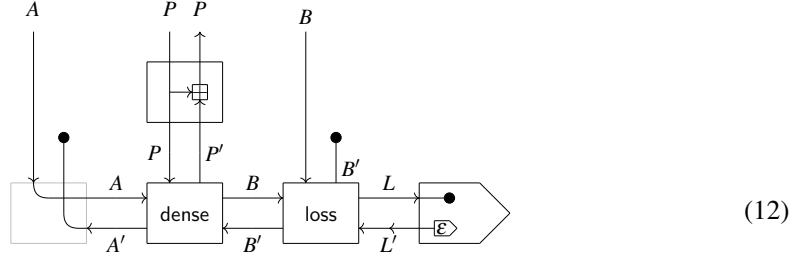
Note that by constructing model architectures in this way, the computation of reverse derivatives is greatly simplified: we obtain the reverse derivative ‘for free’ as the `put` map of the model. Furthermore, adding new primitives is also simplified: the user need simply provide a function and its reverse derivative in the form of a `Para` map. Finally, notice also that our approach is truly compositional: we can define a hidden layer neural network with n hidden units simply by composing two dense layers, as follows:

```
dense(a, n, activation) >> dense(n, b, activation)
```

5.2 Learning

Now that we have constructed a model, we also need to use it to *learn* from data. Concretely, we will construct a full parametric lens as in Figure 2 then extract its `put` map to iterate over the dataset.

By way of example, let us see how to construct the following parametric lens, representing basic gradient descent over a single layer neural network with a fixed learning rate:



This morphism is constructed essentially as below, where $\text{apply_update}(\alpha, f)$ represents the ‘vertical stacking’ of α atop f :

```
apply_update(basic_update, dense) >> loss >> learning_rate( $\epsilon$ )
```

Now, given the parametric lens of (12), one can construct a morphism $\text{step} : B \times P \times A \rightarrow P$ which is simply the put map of the lens. Training the model then consists of iterating the **step** function over dataset examples $(x, y) \in A \times B$ to optimise some initial choice of parameters $\theta_0 \in P$, by letting $\theta_{i+1} = \text{step}(y_i, \theta_i, x_i)$.

Note that our library also provides a utility function to construct **step** from its various pieces:

```
step = supervised_step(model, update, loss, learning_rate)
```

For an end-to-end example of model training and iteration, we refer the interested reader to the experiments accompanying the code [19].

6. Related Work

The work [26] is closely related to ours, in that it provides an abstract categorical model of back-propagation. However, it differs in a number of key aspects. We give a complete lens-theoretic explanation of *what* is back-propagated via (i) the use of CRDCs to model gradients; and (ii) the **Para** construction to model parametric functions and parameter update. We thus can go well beyond [26] in terms of examples - their example of smooth functions and basic gradient descent is covered in our subsection 4.1.

We also explain some of the constructions of [26] in a more structured way. For example, rather than considering the category **Learn** of [26] as primitive, here we construct it as a composite of two more basic constructions (the **Para** and **Lens** constructions). The flexibility could be used, for example, to compositionally replace **Para** with a variant allowing parameters to come from a different category, or lenses with the category of optics [46] enabling us to model things such as control flow using prisms.

One more relevant aspect is functoriality. We use a functor to augment a parametric map with its backward pass, just like [26]. However, they additionally augmented this map with a loss map and gradient descent using a functor as well. This added extra conditions on the partial derivatives of the loss function: it needed to be invertible in the 2nd variable. This constraint was not justified in [26], nor is it a constraint that appears in machine learning practice. This led us to reexamine their constructions, coming up with our reformulation that does not require it. While loss maps and optimisers are mentioned in [26] as parts of the aforementioned functor, here they are extracted out and play a key role: loss maps are parametric lenses and optimisers are reparameterisations. Thus, in this paper we instead use **Para**-composition to add the loss map to the model, and **Para** 2-cells to add optimisers. The mentioned inverse of the partial derivative of the loss map in the

2nd variable was also hypothesised to be relevant to deep dreaming. We have investigated this possibility thoroughly in our paper, showing it is gradient update which is used to dream up pictures. We also correct a small issue in Theorem III.2 of [26]. There, the morphisms of **Learn** were defined up to an equivalence (pg. 4 of [26]) but, unfortunately, the functor defined in Theorem III.2 does not respect this equivalence relation. Our approach instead uses 2-cells which comes from the universal property of **Para** — a 2-cell from $(P, f) : A \rightarrow B$ to $(Q, g) : A \rightarrow B$ is a lens, and hence has two components: a map $\alpha : Q \rightarrow P$ and $\alpha^* : Q \times P \rightarrow Q$. By comparison, we can see the equivalence relation of [26] as being induced by map $\alpha : Q \rightarrow P$, and not a lens. Our approach highlights the importance of the 2-categorical structure of learners. In addition, it does not treat the functor $\mathbf{Para}(\mathcal{C}) \rightarrow \mathbf{Learn}$ as a primitive. In our case, this functor has the type $\mathbf{Para}(\mathcal{C}) \rightarrow \mathbf{Para}(\mathbf{Lens}(\mathcal{C}))$ and arises from applying **Para** to a canonical functor $\mathcal{C} \rightarrow \mathbf{Lens}(\mathcal{C})$ existing for any reverse derivative category, not just **Smooth**. Lastly, in our paper we took advantage of the graphical calculus for **Para**, redrawing many diagrams appearing in [26] in a structured way.

Other than [26], there are a few more relevant papers. The work of [21] contains a sketch of some of the ideas this paper evolved from. They are based on the interplay of optics with parameterisation, albeit framed in the setting of diffeological spaces, and requiring cartesian and local cartesian closed structure on the base category. Lenses and Learners are studied in the eponymous work of [25] which observes that learners are parametric lenses. They do not explore any of the relevant **Para** or CRDC structure, but make the distinction between *symmetric* and *asymmetric lenses*, studying how they are related to learners defined in [26]. A lens-like implementation of automatic differentiation is the focus of [24], but learning algorithms aren't studied. A relationship between category-theoretic perspective on probabilistic modeling and gradient-based optimisation is studied in [50] which also studies a variant of the **Para** construction. Usage of Cartesian differential categories to study learning is found in [56]. They extend the differential operator to work on stateful maps, but do not study lenses, parameterisation nor update maps. The work of [27] studies deep learning in the context of Cycle-consistent Generative Adversarial Networks [63] and formalises it via free and quotient categories, making parallels to the categorical formulations of database theory [54]. They do use the **Para** construction, but do not relate it to lenses nor reverse derivative categories. A general survey of category theoretic approaches to machine learning, covering many of the above papers, can be found in [51]. Lastly, the concept of parametric lenses has started appearing in recent formulations of categorical game theory and cybernetics [10, 11]. The work of [10] generalises the study of parametric lenses into parametric optics and connects it to game theoretic concepts such as Nash equilibria.

7. Conclusions and Future Directions

We have given a categorical foundation of gradient-based learning algorithms which achieves a number of important goals. The foundation is principled and mathematically clean, based on the fundamental idea of a *parametric lens*. The foundation covers a wide variety of examples: different optimisers and loss maps in gradient-based learning, different architectures and layer structures, different settings where gradient-based learning happens (smooth functions vs. boolean circuits), adversarial unsupervised learning, and both learning of parameters and learning of inputs (deep dreaming). Finally, the foundation is more than a mere abstraction: we have also shown how it can be used to give a practical implementation of learning, as discussed in Section 5.

There are a number of important directions which are possible to explore because of this work. One of the most exciting ones is a more comprehensive study of neural network architectures through the category-theoretic perspective. Neural network architectures have begun to be studied using category theory adjacent machinery in the context of *Geometric Deep Learning* ([9]) and *Topological Deep Learning* ([43]). Recurrent neural networks, in particular, have been studied in [56], in the context of differential categories and the concept of *delayed trace* introduced in

the same paper. Despite this, a comprehensive categorical study of architectures is still missing in the literature. As first noticed in [42], many architectures such as recurrent and recursive neural network have close parallels to concepts in functional programming such as folds, unfolds and accumulating maps, for instance. As these functional concepts have clear categorical semantics, it is natural to ask whether these categorical semantics can be used to study neural network architectures. We believe the categorical framework presented in this paper can serve as a natural starting point for such a study. Future work includes modelling some classical systems as well, such as the Support Vector Machines [17], which should be possible with the usage of loss maps such as Hinge loss.

In all our settings we have fixed an optimiser beforehand. The work of [2] describes a *meta-learning* approach which sees the optimiser as a neural network whose parameters and gradient update rule can be learned. This is an exciting prospect since one can model optimisers as parametric lenses; and our framework covers learning with parametric lenses.

Future work also includes using the full power of CRDC axioms. In particular, axioms RD.6 or RD.7, which deal with the behaviour of higher-order derivatives, were not exploited in our work, but they should play a role in modelling some supervised learning algorithms using higher-order derivatives (for example, the Hessian) for additional optimisations. Taking this idea in a different direction, one can see that much of our work can be applied to any functor of the form $F : \mathcal{C} \rightarrow \mathbf{Lens}(\mathcal{C})$ - F does not necessarily have to be of the form $f \mapsto \left(\begin{smallmatrix} f \\ R[f] \end{smallmatrix} \right)$ for a CRDC R . Moreover, by working with more generalised forms of the lens category (such as dependent lenses), we may be able to capture ideas related to supervised learning on manifolds. And, of course, we can vary the parameter space to endow it with different structure from the functions we wish to learn. In this vein, we wish to use fibrations/dependent types to model the use of tangent bundles: this would foster the extension of the *correct by construction* paradigm to machine learning, and thereby addressing the widely acknowledged problem of trusted machine learning. The possibilities are made much easier by the compositional nature of our framework. Another key topic for future work is to link gradient-based learning with game theory. At a high level, the former takes little incremental steps to achieve an equilibrium while the later aims to do so in one fell swoop. Formalising this intuition is possible with our lens-based framework and the lens-based framework for game theory [28]. Finally, because our framework is quite general, in future work we plan to consider further modifications and additions to encompass probabilistic, non-gradient based, and other forms of non-supervised learning. This includes genetic algorithms and reinforcement learning.

Acknowledgements Fabio Zanasi acknowledges support from EPSRC EP/V002376/1. Geoff Cruttwell acknowledges support from NSERC.

References

- [1] S. Abramsky and B. Coecke. A categorical semantics of quantum protocols. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science, 2004.*, pages 415–425, 2004.
- [2] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *30th Conference on Neural Information Processings Systems (NIPS)*, 2016.
- [3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, December 2017. arXiv:1701.07875 [cs, stat].
- [4] John C. Baez and Jason Erbele. Categories in Control. *Theory and Applications of Categories*, 30(24):836–881, 2015.
- [5] R. Blute, R. Cockett, and R. Seely. Cartesian Differential Categories. *Theory and Applications of Categories*, 22(23):622–672, 2009.
- [6] Aaron Bohannon, J. Nathan Foster, Benjamin C. Pierce, Alexandre Pilkiewicz, and Alan Schmitt. Boomerang: Resourceful lenses for string data. *SIGPLAN Not.*, 43(1):407–419, January 2008.
- [7] Guillaume Boisseau. String Diagrams for Optics. *arXiv:2002.11480*, 2020.

- [8] Filippo Bonchi, Paweł Sobocinski, and Fabio Zanasi. The calculus of signal flow diagrams I: linear relations on streams. *Inf. Comput.*, 252:2–29, 2017.
- [9] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges, May 2021. arXiv:2104.13478 [cs, stat].
- [10] Matteo Capucci, Bruno Gavranović, Jules Hedges, and E. F. Rischel. Towards foundations of categorical cybernetics. *arXiv:2105.06332*, 2021.
- [11] Matteo Capucci, Neil Ghani, Jérémy Ledent, and Fredrik Nordvall Forsberg. Translating Extensive Form Games to Open Games with Agency. *arXiv:2105.06763*, 2021.
- [12] François Chollet et al. Keras. <https://keras.io>, 2015.
- [13] Bryce Clarke, Derek Elkins, Jeremy Gibbons, Fosco Loregian, Bartosz Milewski, Emily Pillmore, and Mario Román. Profunctor optics, a categorical update. *arXiv:2001.07488*, 2020.
- [14] J. R. B. Cockett and G. S. H. Cruttwell. Differential Structure, Tangent Structure, and SDG. *Applied Categorical Structures*, 22(2):331–417, April 2014.
- [15] J. Robin B. Cockett, Geoff S. H. Cruttwell, Jonathan Gallagher, Jean-Simon Pacaud Lemay, Benjamin MacAdam, Gordon D. Plotkin, and Dorette Pronk. Reverse derivative categories. In *Proceedings of the 28th Computer Science Logic (CSL) conference*, 2020.
- [16] Bob Coecke and Aleks Kissinger. *Picturing Quantum Processes: A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge University Press, 2017.
- [17] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [18] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. BinaryConnect: Training Deep Neural Networks with binary weights during propagations, April 2016. arXiv:1511.00363 [cs].
- [19] Anonymous CRCoauthors. Numeric Optics: A python library for constructing and training neural networks based on lenses and reverse derivatives. <https://github.com/anonymous-c0de/esop-2022>.
- [20] Geoffrey S. H. Cruttwell, Bruno Gavranovic, Neil Ghani, Paul W. Wilson, and Fabio Zanasi. Categorical foundations of gradient-based learning. In *ESOP*, volume 13240 of *Lecture Notes in Computer Science*, pages 1–28. Springer, 2022.
- [21] David Dalrymple. Dioptics: a common generalization of open games and gradient-based learners. *SYCO7*, 2019.
- [22] Alexey Dosovitskiy and Thomas Brox. Inverting convolutional networks with convolutional networks. *arXiv:1506.02753*, 2015.
- [23] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
- [24] Conal Elliott. The simple essence of automatic differentiation (differentiable functional programming made easy). *arXiv:1804.00746*, 2018.
- [25] Brendan Fong and Michael Johnson. Lenses and learners. In *Proceedings of the 8th International Workshop on Bidirectional transformations (Bx@PLW)*, 2019.
- [26] Brendan Fong, David I. Spivak, and Rémy Tuyéras. Backprop as functor: A compositional perspective on supervised learning. In *Proceedings of the Thirty fourth Annual IEEE Symposium on Logic in Computer Science (LICS 2019)*, pages 1–13. IEEE Computer Society Press, June 2019.
- [27] Bruno Gavranovic. Compositional deep learning. *arXiv:1907.08292*, 2019.
- [28] Neil Ghani, Jules Hedges, Viktor Winschel, and Philipp Zahn. Compositional game theory. In *Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS '18*, page 472–481, 2018.
- [29] Dan R. Ghica, Achim Jung, and Aliaume Lopez. Diagrammatic Semantics for Digital Circuits. *arXiv:1703.10247*, 2017.
- [30] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, June 2014. arXiv:1406.2661 [cs, stat].
- [31] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Society for Industrial and Applied Mathematics, 2008.
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [33] Jules Hedges. Limits of bimorphic lenses. *arXiv:1808.05545*, 2018.
- [34] C. Hermida and R. D. Tennent. Monoidal indeterminates and categories of possible worlds. *Theor. Comput. Sci.*, 430:3–22, April 2012.
- [35] Michael Johnson, Robert Rosebrugh, and R.J. Wood. Lenses, fibrations and universal translations. *Mathematical structures in computer science*, 22:25–42, 2012.
- [36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [37] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.
- [38] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. *arXiv:1412.0035*, 2014.
- [39] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going Deeper into Neural Networks, June 2015.

- [40] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv:1412.1897*, 2014.
- [41] Christopher Olah. Neural Networks, Types, and Functional Programming – colah’s blog, September 2015.
- [42] Christopher Olah. Neural Networks, Types, and Functional Programming – colah’s blog, September 2015.
- [43] Mathilde Papillon, Sophia Sanborn, Mustafa Hajij, and Nina Miolane. Architectures of Topological Deep Learning: A Survey on Topological Neural Networks, April 2023. *arXiv:2304.10031* [cs].
- [44] Robin Piedeleu and Fabio Zanasi. An introduction to string diagrams for computer scientists. *CoRR*, abs/2305.08768, 2023.
- [45] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1 – 17, 1964.
- [46] Mitchell Riley. Categories of optics. *arXiv:1809.00738*, 2018.
- [47] Florian Schäfer and Anima Anandkumar. Competitive Gradient Descent, June 2020. *arXiv:1905.12103* [cs, math].
- [48] Peter Selinger. Control categories and duality: on the categorical semantics of the lambda-mu calculus. *Mathematical Structures in Computer Science*, 11(02):207–260, 4 2001.
- [49] Sanjit A. Seshia and Dorsa Sadigh. Towards verified artificial intelligence. *CoRR*, abs/1606.08514, 2016.
- [50] Dan Shiebler. Categorical Stochastic Processes and Likelihood. *Compositionality*, 3(1), 2021.
- [51] Dan Shiebler, Bruno Gavranović, and Paul Wilson. Category Theory in Machine Learning. *arXiv:2106.07032*, 2021.
- [52] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2014.
- [53] The Royal Society. Explainable AI: the basics - policy briefing, 2019.
- [54] David I. Spivak. Functorial data migration. *arXiv:1009.1166*, 2010.
- [55] David I. Spivak. Generalized Lens Categories via functors $\mathcal{C}^{\mathrm{op}} \rightarrow \mathbf{Cat}$, March 2022. *arXiv:1908.02202* [cs, math].
- [56] David Sprunger and Shin-ya Katsumata. Differentiable causal computations via delayed trace. In *Proceedings of the 34th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS ’19*. IEEE Press, 2019.
- [57] Albert Steckermeier. Lenses in functional programming. *Preprint, available at <https://sinusoid.es/misc/lager/lenses.pdf>*, 2015.
- [58] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.
- [59] D. Turi and G. Plotkin. Towards a mathematical operational semantics. In *Proceedings of Twelfth Annual IEEE Symposium on Logic in Computer Science*, pages 280–291, 1997.
- [60] Paul Wilson and Fabio Zanasi. Reverse derivative ascent: A categorical approach to learning boolean circuits. In *Proceedings of Applied Category Theory (ACT)*, 2020.
- [61] Paul Wilson and Fabio Zanasi. An axiomatic approach to differentiation of polynomial circuits. *Journal of Logical and Algebraic Methods in Programming*, 135:100892, 2023.
- [62] Paul W. Wilson and Fabio Zanasi. Categories of differentiable polynomial circuits for machine learning. In *ICGT*, volume 13349 of *Lecture Notes in Computer Science*, pages 77–93. Springer, 2022.
- [63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv:1703.10593*, 2017.

Appendix A. More details on Parametric Categories

As mentioned in the main text, coherence rules in combining the two operations in (2) just work as expected, in the sense that these diagrams can be ultimately ‘compiled’ down to string diagrams for monoidal categories. For example, given maps $(P, f) : A \rightarrow B$, $(Q, g) : B \rightarrow C$ with reparametrisations $\alpha : P' \rightarrow P$, $\beta : Q' \rightarrow Q$, one could either first reparametrise f and g separately and then

compose the results (below left), or compose first then reparametrise jointly (below right):

$$\begin{array}{c}
 \begin{array}{ccc}
 P' & & Q' \\
 \downarrow & & \downarrow \\
 \boxed{\alpha} & & \boxed{\beta} \\
 \downarrow & & \downarrow \\
 P & & Q \\
 \downarrow & & \downarrow \\
 A \longrightarrow \boxed{f} \longrightarrow B & ; & B \longrightarrow \boxed{g} \longrightarrow C
 \end{array} \\
 \\
 \begin{array}{ccc}
 P' & & Q' \\
 \downarrow & & \downarrow \\
 \boxed{\alpha} & & \boxed{\beta} \\
 \downarrow & & \downarrow \\
 P & & Q \\
 \downarrow & & \downarrow \\
 A \longrightarrow \boxed{f} \longrightarrow \boxed{g} \longrightarrow C
 \end{array}
 \end{array} \quad (13)$$

As expected, translating these two operations into string diagrams for monoidal categories yield equivalent representations of the same morphism.

$$\begin{array}{ccc}
 \begin{array}{c}
 Q' \\
 \downarrow \\
 \boxed{\beta} \\
 \downarrow \\
 Q \\
 \downarrow \\
 \boxed{g} \longrightarrow C \\
 \uparrow \\
 B
 \end{array}
 &
 \begin{array}{c}
 P' \\
 \downarrow \\
 \boxed{\alpha} \\
 \downarrow \\
 P \\
 \downarrow \\
 \boxed{f} \longrightarrow B
 \end{array}
 &
 \begin{array}{c}
 Q' \\
 \downarrow \\
 \boxed{\beta} \\
 \downarrow \\
 Q \\
 \downarrow \\
 \boxed{g} \longrightarrow C \\
 \uparrow \\
 B
 \end{array}
 \end{array}
 =
 \begin{array}{ccc}
 \begin{array}{c}
 Q' \\
 \downarrow \\
 \boxed{\beta} \\
 \downarrow \\
 Q \\
 \downarrow \\
 \boxed{g} \longrightarrow C \\
 \uparrow \\
 B
 \end{array}
 &
 \begin{array}{c}
 P' \\
 \downarrow \\
 \boxed{\alpha} \\
 \downarrow \\
 P \\
 \downarrow \\
 \boxed{f} \longrightarrow B
 \end{array}
 &
 \begin{array}{c}
 Q' \\
 \downarrow \\
 \boxed{\beta} \\
 \downarrow \\
 Q \\
 \downarrow \\
 \boxed{g} \longrightarrow C \\
 \uparrow \\
 B
 \end{array}
 \end{array}
 \quad (14)$$

Remark 50. There is a 2-categorical perspective on $\mathbf{Para}(\mathcal{C})$, which we glossed over in this paper for the sake of simplicity. In particular, the reparametrisations described above can also be seen as equipping $\mathbf{Para}(\mathcal{C})$ with 2-cells, giving a 2-categorical structure on $\mathbf{Para}(\mathcal{C})$. This is also coherent with respect to base change: if \mathcal{C} and \mathcal{D} are strict symmetric monoidal categories, and $F: \mathcal{C} \rightarrow \mathcal{D}$ a lax symmetric monoidal functor, then there is an induced 2-functor $\mathbf{Para}(F): \mathbf{Para}(\mathcal{C}) \rightarrow \mathbf{Para}(\mathcal{D})$ which agrees with F on objects. This 2-functor is straightforward: for a 1-cell $(P, f): A \rightarrow B$, it applies F to P and f and uses the (lax) comparison to get a map of the correct type. We will see how this base change becomes important when performing backpropagation on parametric maps (Eq. 5)

Lastly, we mention that $\mathbf{Para}(\mathcal{C})$ inherits the symmetric monoidal structure from \mathcal{C} and that the induced 2-functor $\mathbf{Para}(F)$ respects that structure. This will allow us to compose neural networks not only in series, but also in parallel. For more detail on alternative viewpoints on the \mathbf{Para} construction, including how it can be viewed as the Grothendieck construction of a certain indexed category, see [10].

Appendix B. Background on Cartesian Reverse Differential Categories

Here we briefly review the definitions of Cartesian left additive category (CLAC), Cartesian reverse differential category (CRDC) and additive and linear maps in these categories. Note that in this appendix we follow the convention of [15] and write composition in diagrammatic order by juxtaposition of terms (rather than a semicolon) to shorten the form of many of the expressions.

Definition 51. A category \mathcal{C} is said to be **Cartesian** when there are chosen binary products \times , with projection maps π_i and pairing operation $\langle -, - \rangle$, and a chosen terminal object T , with unique maps $!$ to the terminal object.

Definition 52. A **left additive category** [5, Definition 1.1.1] (CLAC) is a category \mathcal{C} such that each hom-set has commutative monoid structure, with addition operation $+$ and zero maps 0 , such that composition on the left preserves the additive structure: for any appropriate f, g, h , $f(g + h) = fg + fh$ and $f0 = 0$.

Definition 53. A map $h : X \rightarrow Y$ in a CLAC is **additive** if it has the property that it preserves additive structure by composition on the right: for any maps $x, y : Z \rightarrow X$, $(x + y); h = x; h + y; h$, and $0; h = 0$.

Definition 54 (Additive in second component, (compare [5, Lemma 1.2.3])). A morphism $f : X \times A \rightarrow B$ is additive in the variable A if it is an additive morphism of type $A \rightarrow B$ in the cartesian left-additive category $\mathbf{CoKl}(X \times -)$, where $\mathbf{CoKl}(X \times -)$ is the coKleisli category of the coreader comonad^h.

Definition 55. A **Cartesian left additive category** [5, Definition 1.2.1] is a left additive category \mathcal{C} which is Cartesian and such that all projection maps π_i are additive.

Definition 56. We call **CLACat** the category whose objects are cartesian left-additive categories and whose morphisms are cartesian left-additive functors (functors which preserve products and commutative monoid structure on objects ([5, Definition 1.3.1])).

Lemma 57. Let \mathcal{C} and \mathcal{D} be cartesian left-additive categories, and $F : \mathcal{C} \rightarrow \mathcal{D}$ a cartesian left-additive functor. Let $f : A \rightarrow B$ be an additive morphism in \mathcal{C} . Then $F(f) : F(A) \rightarrow F(B)$ is also additive.

The central definition of [15] is the following:

Definition 58. A **Cartesian reverse differential category** (CRDC) is a Cartesian left additive category \mathcal{C} which has, for each map $f : A \rightarrow B$ in \mathcal{C} , a map

$$R[f] : A \times B \rightarrow A$$

satisfying seven axioms:

[RD.1] $R[f + g] = R[f] + R[g]$ and $R[0] = 0$.

[RD.2] $\langle a, b + c \rangle R[f] = \langle a, b \rangle R[f] + \langle a, c \rangle R[f]$ and $\langle a, 0 \rangle R[f] = 0$.

[RD.3] $R[1] = \pi_1$, $R[\pi_0] = \pi_1 \iota_0$, and $R[\pi_1] = \pi_1 \iota_1$. [RD.4] For a tupling of maps f and g , the following equality holds:

$$R[\langle f, g \rangle] = (1 \times \pi_0); R[f] + (1 \times \pi_1); R[g]$$

And if $!_A : A \rightarrow T$ is the unique map to the terminal object, $R[!_A] = 0$.

[RD.5] For composable maps f and g ,

$$R[f; g] = \langle \pi_0, \pi_0 f, \pi_1 \rangle (1 \times R[g]) R[f]$$

[RD.6]

$$\langle 1 \times \pi_0, 0 \times \pi_1 \rangle (\iota_0 \times 1) R[R[f]] \pi_1 = (1 \times \pi_1) R[f].$$

[RD.7]

$$(\iota_0 \times 1); R[R[(\iota_0 \times 1) R[f]] \pi_1]; \pi_1 = \text{ex}; (\iota_0 \times 1) R[R[(\iota_0 \times 1) R[f]] \pi_1] \pi_1$$

(where ex is the map that exchanges the middle two variables).

As discussed in [15], these axioms correspond to familiar properties of the reverse derivative:

^hThere are a few other terms for this. One of them is “the writer comonad”, though this is often confused with the writer monad which additionally necessitates a monoid structure on X . It’s also called reader comonad, because of duality to reader monad, and also “product comonad” or “environment comonad”.

- **[RD.1]** says that differentiation preserves addition of maps, while **[RD.2]** says that differentiation is additive in its vector variable.
- **[RD.3]** and **[RD.4]** handle the derivatives of identities, projections, and tuples.
- **[RD.5]** is the (reverse) chain rule.
- **[RD.6]** says that the reverse derivative is linear in its vector variable.
- **[RD.7]** expresses the independence of order of mixed partial derivatives.

We proceed to prove the following theorem in three steps.

Theorem. *Lenses with backward passes additive in the second component form a functor*

$$\mathbf{Lens}_A : \mathbf{CLACat} \rightarrow \mathbf{CLACat}$$

The first step is formally defining the category $\mathbf{Lens}_A(\mathcal{C})$.

Definition 59. *Let \mathcal{C} be a cartesian left-additive category. Then $\mathbf{Lens}_A(\mathcal{C})$ is a wide subcategory of $\mathbf{Lens}(\mathcal{C})$ where*

$$\mathbf{Lens}_A(\mathcal{C}) \left(\begin{smallmatrix} A & B \\ A' & B' \end{smallmatrix} \right) := \mathcal{C}(A, B) \times \mathbf{CoKl}(A \times -)_A(B', A')$$

Compare this with the definition of $\mathbf{Lens}(\mathcal{C})$ via the Grothendieck construction ([55, Prop. 3.10]) where

$$\mathbf{Lens}(\mathcal{C}) \left(\begin{smallmatrix} A & B \\ A' & B' \end{smallmatrix} \right) := \mathcal{C}(A, B) \times \mathbf{CoKl}(A \times -)(B', A')$$

The second step is showing this category is cartesian left-additive.

Proposition 60. *The category $\mathbf{Lens}_A(\mathcal{C})$ is cartesian left-additive.*

Proof. We need to equip $\mathbf{Lens}_A(\mathcal{C})$ with a commutative monoid on every object in a way that's compatible with the cartesian structure.ⁱ That is, for every object $\begin{pmatrix} A \\ A' \end{pmatrix}$ we need to provide two morphisms:

- **Unit** $0_{\begin{pmatrix} A \\ A' \end{pmatrix}} : \begin{pmatrix} 1 \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} A \\ A' \end{pmatrix}$. This is a lens whose forward map we set as the zero map 0_A and the backward map as the delete $!_{1 \times A'}$.
- **Multiplication** $+_{\begin{pmatrix} A \\ A' \end{pmatrix}} : \begin{pmatrix} A \times A \\ A' \times A' \end{pmatrix} \rightarrow \begin{pmatrix} A \\ A' \end{pmatrix}$. This is a lens whose forward map we set as sum $+_A$ and the backward map as copy, i.e. $(A \times A) \times A' \xrightarrow{\pi_2} A' \xrightarrow{\Delta_{A'}} A' \times A'$.

Additionally, these morphisms need to obey the monoid laws. This can be verified by routine. \square

This defines the action of \mathbf{Lens}_A on objects of \mathbf{CLACat} . Action on morphisms is defined below.

ⁱWe don't need to show that this monoid is unique, just that it exists and can be canonically defined.

Proposition 61. *Let $F : \mathcal{C} \rightarrow \mathcal{D}$ be a cartesian left-additive functor. This induces a cartesian left-additive functor $\mathbf{Lens}_A(F)$ between the corresponding categories of lenses:*

$$\begin{array}{ccc} \mathbf{Lens}_A(\mathcal{C}) & \xrightarrow{\mathbf{Lens}_A(F)} & \mathbf{Lens}_A(\mathcal{D}) \\ \begin{array}{c} \left(\begin{array}{c} A \\ A' \end{array} \right) \\ \downarrow \left(\begin{array}{c} f \\ f^* \end{array} \right) \\ \left(\begin{array}{c} B \\ B' \end{array} \right) \end{array} & \xrightarrow{\quad} & \begin{array}{c} \left(\begin{array}{c} F(A) \\ F(A') \end{array} \right) \\ \downarrow \left(\begin{array}{c} F(f) \\ \overline{f^*} \end{array} \right) \\ \left(\begin{array}{c} F(B) \\ F(B') \end{array} \right) \end{array} \end{array} \quad (15)$$

where $\overline{f^*} := F(A') \times F(B') \cong F(A' \times B') \xrightarrow{F(f')} F(A')$.

Proof. We need to prove that $\mathbf{Lens}_A(F)$ is a cartesian left-additive functor. To prove it is a functor, we need to:

- Define its action on objects and morphisms. We have done this in Prop. 61 itself;
- Prove additivity of $\overline{f^*}$. This follows from Lemma. 57;
- Prove identities are preserved. The identity $\left(\begin{array}{c} A \\ \pi_2 \end{array} \right) : \left(\begin{array}{c} A \\ A' \end{array} \right) \rightarrow \left(\begin{array}{c} A \\ A' \end{array} \right)$ in the domain gets mapped to $\left(\begin{array}{c} F(\text{id}_A) \\ F(\pi_2) \end{array} \right)$. By preservation of identities and products of F this is equal to the identity map on $\left(\begin{array}{c} F(A) \\ F(A') \end{array} \right)$.
- Prove composition is preserved. This can be by routine, albeit tedious calculation.

To prove that it is additionally cartesian, we need to show that the image of every comonoid $\left(\left(\begin{array}{c} A \\ A' \end{array} \right), !_{\left(\begin{array}{c} A \\ A' \end{array} \right)}, \Delta_{\left(\begin{array}{c} A \\ A' \end{array} \right)} \right)$ is also a comonoid, and that all maps preserve comonoids. We can understand the first part in terms of actions on the counit and comultiplication of the comonoid.

- **Counit.** The action on the counit unpacks to the pair $\left(\begin{array}{c} F(!_{\left(\begin{array}{c} A \\ A' \end{array} \right)}) \\ F(!_{A \times 1; 0_A}) \end{array} \right)$. By preservation of terminal and additive maps of F this morphism is equal to the counit of $\left(\begin{array}{c} F(A) \\ F(A') \end{array} \right)$.
- **Comultiplication.** The action on the comultiplication unpacks to $\left(\begin{array}{c} F(\Delta_A) \\ F(\pi_{2,3}; +_A) \end{array} \right)$. By F 's preservation of products and additive morphisms this morphism is equal to the comultiplication of $\left(\begin{array}{c} F(A) \\ F(A') \end{array} \right)$.

It is routine to show it obey the corresponding laws and form a comonoid.

For the second part we need to show that the image of every lens $\left(\begin{array}{c} f \\ f^* \end{array} \right) : \left(\begin{array}{c} A \\ A' \end{array} \right) \rightarrow \left(\begin{array}{c} B \\ B' \end{array} \right)$ preserves these comonoids. For the forward part this is true because F preserves products. For the backwards part this is true because F is left-additive.

Lastly, we need to prove that this functor is additionally left-additive. This means that it preserves the monoid $\left(\left(\begin{array}{c} A \\ A' \end{array} \right), 0_{\left(\begin{array}{c} A \\ A' \end{array} \right)}, +_{\left(\begin{array}{c} A \\ A' \end{array} \right)} \right)$ of every object. We unpack the action of $\mathbf{Lens}_A(F)$ on the unit $0_{\left(\begin{array}{c} A \\ A' \end{array} \right)}$ and multiplication $+_{\left(\begin{array}{c} A \\ A' \end{array} \right)}$ below.

- **Unit.** The action on the unit unpacks to the pair $\left(\begin{smallmatrix} F(0_A) \\ F(!_{1 \times A}) \end{smallmatrix} \right)$. By preservation of additive and terminal maps of F this morphism is equal to the unit of $\left(\begin{smallmatrix} F(A) \\ F(A) \end{smallmatrix} \right)$;
- **Multiplication.** The action on the multiplication unpacks to the pair $\left(\begin{smallmatrix} F(+_A) \\ F(\pi_3; \Delta_A) \end{smallmatrix} \right)$. By preservation of coadditive maps and products of F this morphism is equal to the multiplication of $\left(\begin{smallmatrix} F(A) \\ F(A') \end{smallmatrix} \right)$.

Seeing as these monoids in the codomain are of the same form as those in the domain, it is routine to show that they obey the monoid laws. This concludes the proof that $\mathbf{Lens}_A(F)$ is a cartesian left-additive functor. \square

What remains to show is that \mathbf{Lens}_A preserves identities and composition, which follows routinely, concluding the proof of (Thm. 2.4).

This functor has additional structure — it is copointed.^j

Proposition 62 (Copointed structure of \mathbf{Lens}_A). *There is a natural transformation $\varepsilon : \mathbf{Lens}_A \Rightarrow \text{id}_{\mathbf{CLACat}}$ which on components assigns to cartesian left-additive category \mathcal{C} a cartesian left-additive functor $\varepsilon_{\mathcal{C}} : \mathbf{Lens}_A(\mathcal{C}) \rightarrow \mathcal{C}$ which forgets the backward part of the lens.*

Proposition 63 ((compare [15, Prop. 31])). *A coalgebra of the copointed \mathbf{Lens}_A endofunctor gives rise to a cartesian left-additive category \mathcal{C} equipped with a reverse differential combinator R which satisfies the first five axioms of a cartesian reverse derivative category.*

Proof. We have shown how a putative reverse derivative combinator arises out of the functor $\mathbf{R}_{\mathcal{C}} : \mathcal{C} \rightarrow \mathbf{Lens}_A(\mathcal{C})$. What remains to prove is that this combinator satisfies the first five axioms of a CRDC.

- (1) **Additivity of reverse differentiation.** This is recovered by $\mathbf{R}_{\mathcal{C}}$ preserving left-additive structure.
- (2) **Additivity of reverse derivative in the second variable.** This is recovered by definition of \mathbf{Lens}_A — the backward maps are additive in the 2nd component.
- (3) **Coherence with identities and projections.** Coherence with identities is recovered by preservation of identities of the functor $\mathbf{R}_{\mathcal{C}}$, where for every $X : \mathcal{C}$, $\mathbf{R}_{\mathcal{C}}(\text{id}_X) = \text{id}_{\mathbf{R}_{\mathcal{C}}(X)} = (\text{id}_X, \pi_2 : X \times X \rightarrow X)$. Coherence with projections is recovered by $\mathbf{R}_{\mathcal{C}}$ preserving cartesian structure.
- (4) **Coherence with pairings.** Recovered by $\mathbf{R}_{\mathcal{C}}$ preserving cartesian structure.
- (5) **Reverse chain rule.** Recovered by functoriality of $\mathbf{R}_{\mathcal{C}}$.

\square

^jDespite this the functor \mathbf{Lens}_A does not have the comonad structure, for similar reasons that tangent categories do not.