

Jacobians and Gradients for Cartesian Differential Categories

ACT 2021

JS Pacaud Lemay
(he/him)



NSERC
CRSNG

Email: jsplemay@gmail.com

Website: <https://sites.google.com/view/jspl-personal-webpage>



Introduction to Cartesian Differential Categories

- Cartesian differential categories (CDC) come equipped with a differential combinator D that formalizes the directional derivative from multivariable calculus (definition in a few slides...)



Blute, R. F. and Cockett, J. R. B. and Seely, R. A. G., [Cartesian Differential Categories](#)

For every map $A \xrightarrow{f} B$, the differential combinator produces its derivative $A \times A \xrightarrow{D[f]} B$.

- The main example is the category of Euclidean spaces \mathbb{R}^n and smooth functions $\mathbb{R}^n \xrightarrow{F} \mathbb{R}^m$, where the differential combinator is precisely the classical derivative $\mathbb{R}^n \times \mathbb{R}^n \xrightarrow{D[F]} \mathbb{R}^m$.
- An important class of examples are the Cartesian *closed* differential categories, which provide the categorical semantics of the differential λ -calculus
 -  Ehrhard, T. and Regnier, L., [The differential lambda-calculus](#)
 -  Manzonetto, G., [What is a categorical model of the differential and the resource \$\lambda\$ -calculus?](#)
- Other interesting examples of CDC include any category with finite biproducts, polynomials over a semiring, the coKleisli category of differential categories, the differential objects of tangent categories, both free and cofree CDC, the Abelian functor calculus model, etc.

Recent Applications of Cartesian Differential Categories

CDC (and their variants) have found numerous applications in computer science such as in:

- Causal computation:



Sprunger, D. and Katsumata, S., [Differentiable causal computations via delayed trace](#)

- Incremental Computation:



Alvarez-Picallo, M. and Ong, C.-H. L., [Change actions: models of generalised differentiation](#)

- Game Theory:



Laird, J. and Manzonetto, G. and McCusker, G., [Constructing differential categories and deconstructing categories of games](#)

- Differentiable Programming:



Abadi, M. and Plotkin, G., [A simple differentiable programming language](#)



Cruttwell, G. and Gallagher, J. and Pronk, D., [Categorical semantics of a simple differential programming language.](#)

- Machine Learning, with the introduction of Cartesian *reverse* differential categories:



Cockett, R., Cruttwell, G., Gallagher, J., Lemay, J. S. P., MacAdam, B., Plotkin, G., & Pronk, D. [Reverse derivative categories.](#)

which have been shown to be a suitable setting for reverse gradient descent:



Cruttwell, G. and Gavranović, B. and Ghani, N. and Wilson, P. and Zanasi, F. [Categorical Foundations of Gradient-Based Learning.](#)



Wilson, P. and Zanasi, F. [Reverse Derivative Ascent: A Categorical Approach to Learning Boolean Circuits.](#)

Motivation: Jacobians and Gradients

- Important concepts in differential calculus are the Jacobian matrix and the gradients, and are also fundamental tools in automatic differentiation and machine learning algorithms.
- Jacobians and gradients have yet to be formally defined in a CDC...
- It is desirable to provide a (coordinate-free) characterization of Jacobians and gradients in a CDC, specifically if one wishes to formalize said algorithms in a CDC.
- This need for Jacobians and gradients is expressed by Katsumata and Sprunger where they state following in the conclusion of their paper



Sprunger, D. and Katsumata, S., [Differentiable causal computations via delayed trace](#)

“Though we would like to say our abstract treatment of differentiation can be used directly by machine learning practitioners, it appears this is not the case yet. The derivative of a morphism in a Cartesian differential category is not the same as having an explicit Jacobian or gradient. A gradient can be recovered from this morphism by applying it to all the basis vectors, but when there are millions of parameters in a machine learning model, this idea is computationally disastrous.”

Today's Story: Jacobians and gradients in the context of Cartesian differential categories.

Recap of Jacobians and Gradients

- Recall that for a smooth function $\mathbb{R}^n \xrightarrow{F=\langle f_1, \dots, f_n \rangle} \mathbb{R}^m$, its Jacobian matrix at point $\vec{x} \in \mathbb{R}^n$ is given by the $n \times m$ matrix $\mathbf{J}(F)(\vec{x})$ whose coordinates are the partial derivatives:

$$\mathbf{J}(F)(\vec{x}) := \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\vec{x}) & \frac{\partial f_1}{\partial x_2}(\vec{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\vec{x}) \\ \frac{\partial f_2}{\partial x_1}(\vec{x}) & \frac{\partial f_2}{\partial x_2}(\vec{x}) & \dots & \frac{\partial f_2}{\partial x_n}(\vec{x}) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\vec{x}) & \frac{\partial f_m}{\partial x_2}(\vec{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\vec{x}) \end{bmatrix}$$

- The Jacobian matrix at \vec{x} 's associated \mathbb{R} -linear function is called the total derivative of F , and evaluating this linear function at \vec{y} results in the derivative $D[F](\vec{x}, \vec{y})$
- Thus, the Jacobian of F can be interpreted as a map $\mathbb{R}^n \xrightarrow{\mathbf{J}(F)} \text{LIN}(\mathbb{R}^n, \mathbb{R}^m)$ – which is a special case of both the Fréchet derivative and the Gateaux derivative.
- For a smooth function $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$, its gradient at $\vec{x} \in \mathbb{R}^n$ is the transpose of its Jacobian at \vec{x} :

$$\nabla(f)(\vec{x}) = \mathbf{J}(F)(\vec{x})^T \in \mathbb{R}^n = \left[\frac{\partial f}{\partial x_1}(\vec{x}) \quad \frac{\partial f}{\partial x_2}(\vec{x}) \quad \dots \quad \frac{\partial f}{\partial x_n}(\vec{x}) \right]$$

- Thus, the gradient can be interpreted as a map $\mathbb{R}^n \xrightarrow{\nabla(f)} \text{LIN}(\mathbb{R}, \mathbb{R}^n)$.

So there seems to be currying and closed structure going on...

This closed idea...

- As suggested by Katsumata and Sprunger:


“We think that by adding some structure to Cartesian differential categories, such as a **designated closed subcategory**, we could give a theoretical treatment allowing for more explicit representation of Jacobians.”

one's first attempt might be to consider a Cartesian closed setting, such as models of the differential λ -calculus, and then define the Jacobian as the curry of the derivative.

- While this is a very reasonable and promising idea where one can get quite far in extracting the main properties of the Jacobian, there is a flaw!
- Unfortunately, this approach would exclude numerous examples of CDC, specifically those with a “finite-dimensional flavour” and many machine learning related models... such as the category of real smooth functions.

So another approach is required...

How to fix it: internal linear homs!

- Note that codomain of the Jacobian is $\text{LIN}(\mathbb{R}^n, \mathbb{R}^m)$, which is isomorphic to \mathbb{R}^{nm} ... so an object in the category of smooth functions.
- So while the category of smooth functions is not Cartesian closed, it does have a sensible notion of internal *linear* homs.
- Internal linear homs are a key concept in Vákár's recent work on automatic differentiation:
 Vákár, M., **CHAD: Combinatory Homomorphic Automatic Differentiation**
- There is a canonical notion of linearity which is induced by the differential combinator in a CDC, which coincides with the classical notion of linearity... So...

Summary of main ideas:

- **Main Definition:** **Linearly closed Cartesian differential category**, which is a CDC with internal linear hom $\mathcal{L}(A, B)$ (which can be interpreted as an object which represents the set of linear maps from A to B), a bilinear evaluation map $\mathcal{L}(A, B) \times A \xrightarrow{\epsilon_\ell} B$, and the ability of currying maps which are linear in their second argument.
- The Jacobian of $A \xrightarrow{f} B$ is defined as the curry of $A \times A \xrightarrow{D[f]} B$, $A \xrightarrow{J(f) := \lambda_\ell(D[f])} \mathcal{L}(A, B)$
- For gradients, we require the extra notion of linear transposes $\mathcal{L}(A, B) \xrightarrow{\tau} \mathcal{L}(B, A)$. So the gradient of $A \xrightarrow{f} B$ is equal to the transpose of its Jacobian, $A \xrightarrow{\nabla(f) := \tau \circ J(f)} \mathcal{L}(B, A)$.
- In future work, these formal notions of Jacobians and gradient will be particularly useful when generalizing and applying automatic differentiation and machine learning algorithms, such as back-propagation or (reverse) gradient descent, in the setting of a C(R)DC.

Ok now let's give some details!

But not full details because of time...for all the nitty-gritty details see the paper!

Definition

A **Cartesian differential category** is a category \mathbb{X} with finite products such that:

- Each hom-set $\mathbb{X}(A, B)$ is a commutative monoid with zero $0 \in \mathbb{X}(A, B)$ and addition $\mathbb{X}(A, B) \times \mathbb{X}(A, B) \xrightarrow{+} \mathbb{X}(A, B)$, such that pre-composition preserves the additive structure:

$$(f + g) \circ a = f \circ a + g \circ a \qquad 0 \circ a = 0$$

- A **differential combinator** D , which is a family of operators $\mathbb{X}(A, B) \xrightarrow{D} \mathbb{X}(A \times A, B)$,

$$\frac{f : A \rightarrow B}{D[f] : A \times A \rightarrow B}$$

where $D[f]$ is called the derivative of f , and which satisfies seven axioms which capture the basics of the derivative from differential calculus (such as the chain rule, additivity in its second argument, symmetry of the partial derivatives, etc.)

To help us, we will use the following term logic ¹:

$$D[f](a, b) := \frac{df(x)}{dx}(a) \cdot b$$

¹There is a sound & complete term logic for CDC. Anything we can prove using the term logic, holds in any CDC. Super useful!

Example

Define SMOOTH as the category whose objects are the Euclidean real vector spaces \mathbb{R}^n and whose maps are the real smooth functions $\mathbb{R}^n \xrightarrow{F} \mathbb{R}^m$ between them. SMOOTH is a CDC where the differential combinator is defined as the directional derivative of a smooth function.

A smooth function $\mathbb{R}^n \xrightarrow{F} \mathbb{R}^m$ is in fact a tuple $F = \langle f_1, \dots, f_m \rangle$ of smooth functions $\mathbb{R}^n \xrightarrow{f_i} \mathbb{R}$.

Using the convention that $\vec{x} \in \mathbb{R}^n$ are column vectors, the derivative $\mathbb{R}^n \times \mathbb{R}^n \xrightarrow{D[F]} \mathbb{R}^m$ is defined as multiplying the Jacobian matrix of F at the first argument \vec{x} , which is an $m \times n$ matrix $\mathbf{J}(F)(\vec{x})$, with the second argument \vec{y} , seen as an $n \times 1$ matrix:

$$D[F](\vec{x}, \vec{y}) := \mathbf{J}(F)(\vec{x})\vec{y} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\vec{x}) & \frac{\partial f_1}{\partial x_2}(\vec{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\vec{x}) \\ \frac{\partial f_2}{\partial x_1}(\vec{x}) & \frac{\partial f_2}{\partial x_2}(\vec{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\vec{x}) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\vec{x}) & \frac{\partial f_m}{\partial x_2}(\vec{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\vec{x}) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \frac{\partial f_1}{\partial x_i}(\vec{x})y_i \\ \vdots \\ \sum_{i=1}^n \frac{\partial f_m}{\partial x_i}(\vec{x})y_i \end{bmatrix}$$

When $m = 1$, for a smooth function $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$, $D[f](\vec{x}, \vec{y})$ is precisely the directional derivative of f at point \vec{x} and along the vector \vec{y} .

Cartesian Differential Categories - Linear Maps

An important class of maps in a CDC are the linear maps and maps which are linear in certain arguments. Essentially, a map is linear in an argument if when differentiating with respect to that argument (and keeping the other arguments constant), one gets back the starting map.

Definition

In a CDC \mathbb{X} :

- i A map $A \xrightarrow{f} B$ is **linear** if $\frac{df(x)}{dx}(a) \cdot b = f(b)$
- ii A map $A \times B \xrightarrow{f} C$ is **linear in its second argument** if f is linear with respect to the partial derivative in its second argument, that is, $\frac{df(x,y)}{d(x,y)}(a, b) \cdot (0, c) = f(a, c)$;
- iii A map $A \times B \xrightarrow{f} C$ is **bilinear** if $\frac{df(x,y)}{d(x,y)}(a, b) \cdot (c, d) = f(a, d) + f(c, b)$

Define $\text{Lin}[\mathbb{X}]$ to be the subcategory of linear maps of \mathbb{X} .

One of the axioms of the differential combinator tells us that $D[f]$ is linear in its second argument.

Example

In SMOOTH, being linear in the CDC sense is the same as being \mathbb{R} -linear in the classical sense. For example, $\mathbb{R}^n \xrightarrow{F} \mathbb{R}^m$ is linear in the CDC sense if and only if it is \mathbb{R} -linear in the classical sense:

$$F(s\vec{x} + t\vec{y}) = sF(\vec{x}) + tF(\vec{y})$$

So $\text{Lin}[\text{SMOOTH}]$ is the category of \mathbb{R} -linear maps between the \mathbb{R} -vector spaces \mathbb{R}^n .

Definition

A **linearly closed Cartesian differential category** is a CDC \mathbb{X} such that:

- For each pair of objects A and B , there is an object $\mathcal{L}(A, B)$, called the **internal linear hom**;
- A *bilinear map* $\mathcal{L}(A, B) \times A \xrightarrow{\epsilon_\ell} B$ called the **evaluation map**;
- For every map $A \times B \xrightarrow{f} C$ which is *linear in its second argument*, there exists a unique map $A \xrightarrow{\lambda_\ell(f)} \mathcal{L}(B, C)$, $\lambda_\ell(f)(a) = \lambda_\ell y. f(a, y)$, called the **linear curry** of f , such that:

$$\begin{array}{ccc}
 A \times B & \xrightarrow{\lambda_\ell(f) \times 1_B} & \mathcal{L}(B, C) \times B \\
 & \searrow f & \downarrow \epsilon_\ell \\
 & & C
 \end{array}$$

$\mathcal{L}(A, B)$ should be interpreted as the internal version of $\text{Lin}[\mathbb{X}](A, B)$, so the linear maps from A to B . Being linearly closed does not imply that $\text{Lin}[\mathbb{X}]$ is Cartesian closed. However, if a CDC \mathbb{X} has a linear \otimes representation:



Blute, R. F. and Cockett, J. R. B. and Seely, R. A. G., [Cartesian Differential Storage Categories](#)

then we conjecture that being linearly closed is equivalent to $\text{Lin}[\mathbb{X}]$ being monoidal closed – this approach in the monoidal setting has been studied independently by Gallagher and MacAdam.

Definition

In a linearly closed CDC, the **Jacobian** of a map $A \xrightarrow{f} B$ is the map $A \xrightarrow{\mathbf{J}(f)} \mathcal{L}(A, B)$ defined as the linear curry of $A \times A \xrightarrow{D[f]} B$, that is, $\mathbf{J}(f) := \lambda_{\ell}(D[f])$ and so $\mathbf{J}(f)(a) = \lambda_{\ell}y. \frac{df(x)}{dx}(a) \cdot y$.

Example

SMOOTH is a linearly closed CDC where the internal linear hom is $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) = \mathbb{R}^{nm}$ and the evaluation map $\mathbb{R}^{nm} \times \mathbb{R}^n \xrightarrow{\epsilon_{\ell}} \mathbb{R}^m$ is defined as:

$$\epsilon_{\ell}(\vec{x}, \vec{y}) := \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ x_{n+1} & x_{n+2} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{(m-1)n+1} & x_{(m-1)n+2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \vdots \\ \sum_{i=1}^n x_{(m-1)n+i} y_i \end{bmatrix}$$

When $m = 1$, the evaluation map is given by the dot product: $\epsilon_{\ell}(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y} = \sum_{i=1}^n x_i y_i$.

For a smooth function $\mathbb{R}^n \xrightarrow{F=\langle f_1, \dots, f_m \rangle} \mathbb{R}^m$, taking $\mathbf{J}(F)(\vec{x}) = \lambda_{\ell}(D[F])(\vec{x})$ results precisely in the Jacobian matrix of F at \vec{x} interpreted as column vector of size nm :

$$\mathbf{J}(F)(\vec{x}) = \left[\frac{\partial f_1}{\partial x_1}(\vec{x}), \dots, \frac{\partial f_1}{\partial x_n}(\vec{x}), \frac{\partial f_2}{\partial x_1}(\vec{x}), \dots, \frac{\partial f_m}{\partial x_n}(\vec{x}) \right]^{\top}$$

which when post-composed in the evaluation map results in laying it out back into a $m \times n$ matrix.

Gradients and Transposes

Definition

A linearly closed CDC has a **linear transpose** if for every pair of objects A and B , there is a linear map $\mathcal{L}(A, B) \xrightarrow{\tau} \mathcal{L}(B, A)$ such that τ satisfies three axioms capturing the classical notion of transpose (idempotent, swaps order of composition, preserves identities).

The **gradient** of a map $A \xrightarrow{f} B$ is the map $A \xrightarrow{\nabla(f)} \mathcal{L}(B, A)$ defined as the transpose of its Jacobian $\nabla(f) = \tau \circ \mathbf{J}(f)$.

Proposition

A linearly closed CDC equipped with a linear transpose is precisely a linearly closed Cartesian reverse differential category. The gradient is then linear curry of the reverse derivative:

$$\frac{f : A \rightarrow B}{\mathbf{R}[f] : A \times B \rightarrow A} \qquad \nabla(f) := \lambda_\ell(\mathbf{R}[f])$$

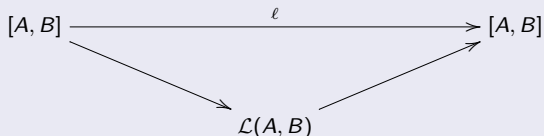
Example

SMOOTH is a linearly closed CDC with a linear transpose (where τ is simply the matrix transpose operation), or equivalently, SMOOTH is a linearly closed CRDC.

Cartesian Closed vs Linearly Closed

Proposition

A Cartesian closed differential category is linearly closed if and only if a canonical linear idempotent splits.



This linear idempotent ℓ exists for any Cartesian closed differential category via linearization:



Cockett, R.; Lemay, J-S. P., [Linearizing Combinators](#)

Proposition

If \mathbb{X} is a Cartesian closed differential category, then its linear idempotent completion $LS[\mathbb{X}]$



is a linearly closed Cartesian closed differential category.

Cockett, R.; Gallagher, J., [Categorical models of the differential \$\lambda\$ -calculus](#)

Some Final Thoughts

- We conjecture that an equivalent alternative axiomatization of a linearly closed Cartesian differential category can be done simply in terms of $\mathcal{L}(-, -)$, ε_ℓ , and \mathbf{J} , where one would define the differential combinator as $D[-] = \varepsilon_\ell \circ (\mathbf{J}(-) \times 1)$
- We conjecture that an equivalent alternative axiomatization of a linearly closed Cartesian reverse differential category can be done simply in terms of $\mathcal{L}(-, -)$, ε_ℓ , ∇ , and τ , where one would define the reverse differential combinator as $R[-] = \varepsilon_\ell \circ (\nabla(-) \times 1)$.
- It should also be possible to generalize other important notions from classical differential calculus such as the divergence, the curl, the Laplacian, and the Hessian. For example, the Hessian matrix is defined as the Jacobian matrix of the gradient, so:

$$A \xrightarrow{\mathbf{H}(f) := \mathbf{J}(\nabla(f))} \mathcal{L}(A, \mathcal{L}(B, A))$$

That's all folks!

HOPE YOU ENJOYED MY TALK!

THANKS FOR LISTENING!

MERCI!

Email: jsplemay@gmail.com

Website: <https://sites.google.com/view/jspl-personal-webpage>