Combinatory Adjoints and Differentiation

Martin Elsman
DIKU, U. Copenhagen
mael@diku.dk

Fritz Henglein
DIKU, U. Copenhagen
henglein@diku.dk

Robin Kaarsgaard
U. Edinburgh
Robin.Kaarsgaard@ed.ac.uk

Mikkel Kragh Mathiesen
DIKU, U. Copenhagen
mkm@di.ku.dk

Robert Schenck
DIKU, U. Copenhagen
rschenck@di.ku.dk

We develop a compositional approach for automatic and symbolic differentiation based on categorical constructions in functional analysis where derivatives are linear functions on abstract vectors rather than being limited to scalars, vectors, matrices or tensors represented as multi-dimensional arrays.

We show that both symbolic and automatic differentiation can be performed using a differential calculus for generating linear functions representing Fréchet derivatives based on rules for primitive, constant, linear and bilinear functions as well as their sequential and parallel composition. Linear functions are represented in a combinatory domain-specific language.

Finally, we provide a calculus for symbolically computing the adjoint of a derivative without using matrices, which are too inefficient to use on high-dimensional spaces. The resulting symbolic representation of a derivative retains the data-parallel operations from the input program. The combination of combinatory differentiation and computing formal adjoints turns out to be behaviorally equivalent to reverse-mode automatic differentiation. In particular, it provides opportunities for optimizations where matrices are too inefficient to represent linear functions.

1 Introduction

Automatic differentiation (AD) [21] is the discipline of computing derivatives for functions given by programs. It is used in gradient-based optimization, neural networks, probabilistic inference [3, sec. 4] and has numerous applications in computer vision, natural language processing, computational science, bioinformatics, quantitative finance, computational economics, and in many other areas. For example, backpropagation, which is used in machine learning to train neural networks, is an instance of reverse mode AD. Building tools to implement and compute derivatives from programs automatically, efficiently, and precisely, has far-reaching impact potential.

1.1 Contributions

In this paper we develop a general framework for expressing and reasoning about functions, their derivatives and the adjoints of these in combinatory form.

We make the following novel contributions:

- We present a general framework for constructing Hilbert spaces. The constructions freely combine tensor products and direct sums. Direct sums generalize both homogeneous data types, such as order-k tensors (scalars, vectors, matrices, and so on) and inhomogeneous types such as tuple and record types. Abstract tensor products express tensor decomposition of matrices, which are asymptotically more efficient than using matrices for low-rank matrices.

Jeremy Gibbons and Max S. New (Eds.): Ninth Workshop on Mathematically Structured Functional Programming (MSFP 2022)
EPTCS 360, 2022, pp. 1–26, doi:10.4204/EPTCS.360.1

- We identify five general differentiation rules for calculating Fréchet derivatives, which represent derivatives as linear functions, by structural recursion on functions given in combinatory form. The combinatory form of a function thus distills its differential properties. Intuitively, differentiating a function in point-full notation consists mostly of (implicitly) turning it into combinatory form.
- We exhibit the generalized product rule, which is applicable to arbitrary bilinear functions operating on spaces of any dimension as a general rule not previously exploited. Bilinear functions on high-dimensional data are common, including matrix multiplication, outer product, dot product, zip (Hadamard) product and any composition of a linear function with a bilinear function. To differentiate a bilinear function we only need to know that it is bilinear since its derivative is expressed in terms of itself.
- We provide affine interpretation of a function in combinatory form, which computes both the output value of a function at a given input and returns a symbolic (term) representation of its Fréchet derivative. Symbolic rather than functional representations facilitate optimization using (multi)linear and tensor algebra equalities.
- We further demystify reverse-mode automatic differentiation by identifying its essence as symbolically computing the adjoint of the Fréchet derivative in combinatory form. The adjoint of a linear function f is a representation of its transpose, the continuation passing style version of f. In adjoints linear continuations are represented by their duals, ordinary first-order vectors, which facilitates and explains how a linear function can be executed efficiently in reverse.
- We provide an adjoint calculus for symbolically calculating the adjoints of linear functions in combinatory form. We identify *relational reduction* and *tensor contraction* as natural parallel linear operations since they provide their own adjoints.
- We illustrate how combinatory differentiation and combinatory adjoint calculation can be used to derive the backpropagation code for neural networks such that all data parallelism is preserved.

More speculatively, we believe our combinatory setting is useful for a differential and adjoint calculus on functions and linear functions. The Hilbert space setting seems to provide a promising setting in which both database and analytic functions can be specified, differentiated and reversed by taking adjoints.

1.2 Outline

We assume basic familiarity with functional analysis, which, as a framework, generalizes both multivariate and tensor calculus by operating on arbitrary elements of structured vector spaces instead of restricting them to tuples of scalars or multi-dimensional arrays that represent tensors. The relevant notions are introduced in the remainder of this and the next section.

In Section 3 we informally present a list of primitive, constant, linear and bilinear analytic functions that can be combined freely by sequential and parallel composition to complex analytic functions in point-free notation. In Section 4 we then formulate a calculus for symbolically differentiating functions in combinatory form such that the derivatives of parallel functions are rendered in point-free notation, as parallel linear functions. In Section 5 we show how this gives rise to affine interpretation of an analytic function in combinatory form: The interpreter returns not only the value of a function on its input, but also a compact combinatory representation of its derivative whose size is largely independent of the dimensionality of the vector spaces involved. In Section 6 we show how the inner product operator can be used to uniquely represent linear continuations by ordinary first-order vectors. This gives rise to symbolically computed adjoints, which run a linear function efficiently "in reverse" and thus provide

reverse-mode AD. We illustrate combinatory differentiation on neural networks in Section 7 and discuss related work in Section 8.

1.3 Background

Definition 1.1 (Fréchet derivative)**.** For a function $f : V \to W$ on Banach spaces V, W, the *linear function* $A \in V \multimap W$ is the *Fréchet derivative of f at v* if it satisfies

$$f(v+dv) \approx f(v) + A(dv),$$

that is

$$\lim_{||dv||_V \to 0} \frac{||f(v+dv) - (f(v)+A(dv))||_W}{||dv||_V} = 0$$

where $||\ldots||_U$ is the norm that comes with the Banach space U.

The *Fréchet derivative of $f : V \to W$* is the partial function $f' : V \to (V \multimap W)$ that maps a vector $v \in V$ to the Fréchet derivative of f at v.

See Appendix A for other notions of derivatives, including Gateaux derivatives.

2 Sets and spaces

We provide general constructions for defining *inner product spaces* over $\mathbb{R}$ and their implicit completions to real Hilbert spaces. These provide a model of symbolic derivatives as Fréchet derivatives.

An inner product space over $\mathbb{R}$ is a vector space V over $\mathbb{R}$ equipped with an inner product

$$\odot : V \times V \to_2 \mathbb{R}$$

that is symmetric, $v_1 \odot v_2 = v_2 \odot v_1$, and positive definite, $v \odot v > 0$ for all $v \neq 0$. A real Hilbert space is an inner product space over $\mathbb{R}$ that is also a complete metric space with respect to the distance function $d(v,w) = \|v - w\|$ where $\|v\| = \sqrt{v \odot v}$.

A continuous function $f : V \to W$ on real Hilbert spaces V, W is *linear* if $f(u+v) = f(u) + f(v)$ and $f(k \cdot v) = k \cdot f(v)$; we write $f : V \multimap W$ if f is continuous and linear.

A continuous binary function $\diamond : U \times V \to W$ is *bilinear* if $(u\diamond) : V \multimap W$ and $(\diamond v) : U \multimap W$ are linear for all $u \in U, v \in V$ where $(u\diamond)$ and $(\diamond v)$ are defined by $(u\diamond)(v) = u \diamond v = (\diamond v)(u)$. We write $f : U \times V \to_2 W$ if f is continuous and bilinear.

Proviso: Henceforth all functions will implicitly be continuous.

2.1 Sets

We provide a language for defining *index sets*. These are used to construct direct sum spaces.

$$X, Y \quad ::= \quad \mathbf{n} \mid X \times Y \mid X + Y$$

where $n \in \mathbb{N}$, $\mathbf{n}$ is the initial segment $\{1, \ldots, n\}$ of natural numbers; $S \times T$ and $S + T$ the Cartesian product, respectively disjoint union of S and T.

The constructible index sets are finite, which ensure that the constructions are metrically complete. We believe the theory, being essentially algebraic, can be extended to infinite denumerable sets. We stick to finite index sets and thus finite-dimensional Hilbert spaces in this paper, however.

2.2 Spaces

Below we provide constructions for Hilbert spaces generated by the following terms:

$$U,V,W \quad ::= \quad 0 \mid K \mid \oplus_{x\in X} V_x \mid V \otimes W$$

where V_x may depend on $x \in X$.

2.2.1 Atomic spaces

The trivial vector space 0 consists of the single element 0.

K stands for the underlying field of our vector spaces, here $\mathbb{R}$. Its elements are the elements of $\mathbb{R}$ as a field. Its operations as a vector space are the corresponding field operations.

2.2.2 Direct sum space

The Hilbert space $V = \oplus_{x\in X} V_x$ for denumerable X is the *(external) direct sum* of a family of Hilbert spaces V_x indexed by $x \in X$. Its elements are maps m from X such that $m(x) \in V_x$ and $\sum_{x\in X}(m(x) \odot_{V_x} m(x)) < \infty$. We write m_x for the result of applying the map to highlight that x is an element of an index set, not a vector. Its operations are defined by component-wise lifting, where the inner product is

$$(v \odot_V v') \quad = \quad \sum_{x\in X} (v_x \odot_{V_x} v'_x)$$

The summation is defined since $\sum_{x\in X}(v_x \odot_{V_x} v'_x) \leq \sum_{x\in X}(v_x \odot_{V_x} v_x) + \sum_{x\in X}(v'_x \odot_{V_x} v'_x) < \infty$. Note it is trivially well-defined for finite X.

V comes with linear injection and projection functions

$$\begin{array}{rcl} \iota_y^X & : & V_y \multimap \oplus_{x\in X} V_x \\ \pi_y^X & : & \oplus_{x\in X} V_x \multimap V_y \end{array}$$

for $y \in X$, and the *zipped apply* operator

$$\Pi_{x\in X} f_x \quad : \quad \oplus_{x\in X} V_x \to \oplus_{x\in X} W_x$$

for a family of functions $f_x \in V_x \to W_x$ indexed by $x \in X$. They satisfy

$$\begin{array}{rcll} \pi_y^X \circ \Pi_{x\in X} f_x \circ \iota_y^X & = & f_y & \\ \pi_z^X \circ \Pi_{x\in X} f_x \circ \iota_y^X & = & 0_{yz} & \text{if } y \neq z \end{array}$$

where $0_{yz} : V_y \multimap W_z$ maps all $v_x \in V_x$ to $0 \in W_z$. The zipped apply operator preserves linearity, that is

$$\Pi_{x\in X} f_x \quad : \quad \oplus_{x\in X} V_x \multimap \oplus_{x\in X} W_x$$

for $f_x : V_x \multimap W_x$. A special case of this is

$$\Delta \quad : \quad \oplus_{x\in X}(V_x \multimap W_x) \multimap (\oplus_{x\in X} V_x \multimap \oplus_{x\in X} W_x)$$

defined by

$$\Delta(\oplus_{x\in X} f_x)(\oplus_{x\in X} v_x) \quad = \quad \oplus_{x\in X}(f_x(v_x))$$

which will later play the role of gathering derivatives acting on the individual differentials of a collection into a derivative that acts on all differentials in parallel.

We write $V_1 \times \ldots \times V_n$ or $V_1 \oplus \ldots \oplus V_n$ for $\oplus_{i\in \mathbf{n}} V_i$. In particular, $V_1 \times V_2 = V_1 \oplus V_2 = \oplus_{i\in \mathbf{2}} V_i$ is the direct sum of V_1 and V_2, whose elements are the pairs (v_1, v_2) such that $v_1 \in V_1$ and $v_2 \in V_2$.

2.2.3 Copower space

For set X and space V, the *copower* V^X is the direct sum, where each space in the family is the same $V_x = V$:

$$V^X = \oplus_{x \in X} V.$$

As special cases we have $\mathbb{R}^{\mathbf{n}}$ as the space of n-ary vectors of scalars. In particular, $V \times V = V^{\mathbf{2}}$. Note that the exponents are sets, not numbers. This is reflected in the notation $\mathbb{R}^{\mathbf{n}}$: an element is a finite map m from $\mathbf{n}$ to $\mathbb{R}$, which can conveniently be written using tuple notation $(m_1, \ldots, m_n)$. For example $(5,8,22) \in \mathbb{R}^{\mathbf{3}}$ is syntactic sugar for $\{1 \mapsto 5, 2 \mapsto 8, 3 \mapsto 22\}$.

For relation $R \subseteq X \times Y$ where X, Y are finite we define *relational reduction*

$$\begin{array}{rcl} \mathrm{red}_R & : & V^X \multimap V^Y \\ (\mathrm{red}_R(v))_y & = & \displaystyle\sum_{(x,y) \in R} v_x \end{array}$$

Many useful functions can be defined in terms of relational reduction. Let $Y \subseteq X$ be finite. The functions

$$\begin{array}{rcll} f^X & : & V^X \to W^X & \text{if } f : V \to W \\ \mathrm{rep}_Y & : & V \multimap V^X & \\ \Sigma_Y & : & V^X \multimap V & \\ + & : & V^{\mathbf{2}} \multimap V & \\ \mathrm{dup} & : & V \multimap V^{\mathbf{2}} & \\ \mathrm{scan}_n & : & V^{\mathbf{n}} \multimap V^{\mathbf{n}} & \\ \langle f_y \rangle_{y \in Y} & : & U \multimap \oplus_{y \in Y} V_y & \text{if } f_y : U \multimap V_y \\ [g_x]_{x \in X} & : & \oplus_{x \in X} V_x \multimap W & \text{if } g_x : V_x \multimap W \end{array}$$

are defined by

$$\begin{array}{rcl} f^X & = & \Pi_{x \in X} f \\ \mathrm{rep}_Y & = & \mathrm{red}_{\mathbf{1} \times Y} \bullet \iota_1^{\mathbf{1}} \\ \displaystyle\sum_Y & = & \pi_1^{\mathbf{1}} \bullet \mathrm{red}_{Y \times \mathbf{1}} \\ + & = & \displaystyle\sum_{\mathbf{2}} \\ \mathrm{dup} & = & \mathrm{rep}_{\mathbf{2}} \\ \mathrm{scan}_n & = & \mathrm{red}_{\{(i,j) \mid 1 \le i \le j \le n\}} \\ \langle f_y \rangle_{y \in Y} & = & \Pi_{y \in Y} f_y \bullet \mathrm{rep}_Y \\ [g_x]_{x \in X} & = & \displaystyle\sum_X \bullet \Pi_{x \in X} g_x \end{array}$$

2.2.4 Tensor product space

$W = U \otimes V$ is the tensor product space of U and V. Its finite elements are the formal terms generated by

$$w \quad ::= \quad 0 \mid k \cdot w \mid w_1 + w_2 \mid u \otimes v$$

where $k \in \mathbb{R}, u \in U, v \in V$ that are identified modulo the vector space axioms and the equalities

$$\begin{aligned}(k \cdot v) \otimes w &= k \cdot (v \otimes w) = v \otimes (k \cdot w)\\ (v_1 + v_2) \otimes w &= (v_1 \otimes w) + (v_2 \otimes w)\\ v \otimes (w_1 + w_2) &= (v \otimes w_1) + (v \otimes w_3).\end{aligned}$$

We write $[w]_\otimes$ for the equivalence class of w under these equalities and define

$$\begin{aligned}0_W &= [0]_\otimes\\ v_1 +_W v_2 &= [v_1 + v_2]_\otimes\\ k \cdot_W v &= [k \cdot v]_\otimes\end{aligned}$$

W is metrically complete for finite-dimensional U, V; otherwise metric completion of the equivalence classes $[w]_\otimes$ is required. The equalities guarantee that the functions are well-defined and $(W, 0_W, +_W, \cdot_W)$ forms a Hilbert space such that

$$\otimes : U \times V \to_2 W$$

is bilinear, that is pointwise linear in each of its arguments. Indeed, the operation $\otimes$ and the space $U \otimes V$ are constructed to be *universal*: For every bilinear function $\diamond : U \times V \to_2 T$ there exists a unique linear function $\bar{\diamond} : U \otimes V \multimap T$ such that $\diamond = \bar{\diamond} \circ \otimes$.

Furthermore, we define the inner product

$$\odot : W \times W \to_2 \mathbb{R}$$

to be the unique bilinear function that satisfies

$$(u_1 \otimes v_1) \odot (u_2 \otimes v_2) \quad = \quad (u_1 \odot u_2) \cdot (v_1 \odot v_2).$$

3 Functions in combinatory form

We provide a domain-specific language for specifying analytic functions on Hilbert spaces in combinatory form, that is in point-free notation. In combinatory form, all subterms are closed functions; in particular, a subterm does not have implicit dependencies on an environment. This facilitates formulation of a compositional differential calculus for calculating Fréchet derivatives.

3.1 Tensor contraction

We provide a language constant for a single bilinear function. It would be sufficient to provide the tensor product $\otimes$ as sole bilinear function since it is universal in the sense that all bilinear functions $f : U \times V \to_2 W$ factor into $f = \bar{f} \bullet (\otimes)$ for a unique $\bar{f} : U \otimes V \multimap W$, the characteristic universal property of $\otimes$. For reasons to become clear later, we provide *tensor contraction*

$$* \quad : \quad (W \otimes V) \times (V \otimes U) \to_2 (W \otimes U)$$

instead. It is defined as the unique bilinear function satisfying

$$(w \otimes v) * (v' \otimes u) \quad = \quad (v \odot v') \cdot (w \otimes u)$$

3.2 Unitary operators

We have a large number of useful *natural unitary operators*; these are natural linear isomorphisms that are *isometric*, i.e. preserve norms. We list a few of them here.

$$\begin{array}{rcccl}
\langle_ & : & V \longleftrightarrow_1 \mathbb{R}\otimes V & : & |_| \\
_\rangle & : & V \longleftrightarrow_1 V\otimes\mathbb{R} & : & |_| \\
_^T & : & (V\otimes W) \longleftrightarrow_1 (W\otimes V) & : & _^T \\
assoc & : & ((U\otimes V)\otimes W) \longleftrightarrow_1 (U\otimes(V\otimes W)) & : & assoc^{-1} \\
distrib & : & (\oplus_{x\in X} V_x)\otimes W \longleftrightarrow_1 \oplus_{x\in X}(V_x\otimes W) & : & distrib^{-1} \\
zip & : & (\oplus_{x\in X} V_x)\oplus(\oplus_{x\in X} W_x) \longleftrightarrow_1 \oplus_{x\in X}(V_x\oplus W_x) & : & unzip
\end{array}$$

They are defined by

$$\begin{aligned}
\langle v &= 1\otimes v \\
|k\otimes v| &= k\cdot v \\
v\rangle &= v\otimes 1 \\
|v\otimes k| &= k\cdot v \\
(v\otimes w)^T &= w\otimes v \\
assoc((u\otimes v)\otimes w) &= u\otimes(v\otimes w) \\
assoc^{-1}(u\otimes(v\otimes w)) &= (u\otimes v)\otimes w \\
distrib((\oplus_{x\in X} v_x)\otimes w) &= \oplus_{x\in X}(v_x\otimes w) \\
(zip(v,w))_x &= (v_x, w_x)
\end{aligned}$$

where $\oplus_{x\in X} v_x$ is notation for the element of $\oplus_{x\in X} V_x$ that maps x to the value $v_x \in V_x$. It turns out that the inverse of a unitary operator is also its adjoint; this will be useful later.

A derived isometric isomorphism is

$$V^X \otimes W^Y \longleftrightarrow_1 (V\otimes W)^{X\times Y}$$

and in particular

$$\mathbb{R}^{\mathbf{m}} \otimes \mathbb{R}^{\mathbf{n}} \longleftrightarrow_1 \mathbb{R}^{\mathbf{m}\times\mathbf{n}}.$$

In other words, all the elements of the tensor product of $\mathbb{R}^{\mathbf{m}}$ and $\mathbb{R}^{\mathbf{n}}$ can be represented by $m\times n$ matrices. Our construction of $\mathbb{R}^{\mathbf{m}}\otimes\mathbb{R}^{\mathbf{n}}$ using symbolic operators 0, $\cdot$, $+$ and $\otimes$ provides more space efficient representations for low-rank matrices, however. For example, every rank-1 $m\times n$ matrix corresponds to $v\otimes w$, its *tensor decomposition*, for some $v\in\mathbb{R}^{\mathbf{m}}, w\in\mathbb{R}^{\mathbf{n}}$. This term representation is of size $O(m+n)$ rather than requiring $m\cdot n$ entries in a matrix. (Note that a rank-1 matrix may have no 0-entries.) Matrix/vector multiplication can be performed with only n multiplications instead of $m\cdot n$ multiplications when using the matrix representation.

The tensor and inner product operators are special cases of tensor contraction via the $\langle\_$ and $\_\rangle$ unitary operators:

$$\begin{aligned}
v\otimes w &= v\rangle * \langle w \\
v_1 \odot v_2 &= |\langle v_1 * v_2\rangle|
\end{aligned}$$

Note that these are parsed as $(v\rangle) * (\langle w)$ and $|(\langle v) * (w\rangle)|$, respectively.

3.3 Linear functions

In addition to the unitary operators, the following are linear functions:

$$\begin{array}{rcll}
(v*) & : & V \otimes U \multimap W \otimes U & \text{if } v \in W \otimes V \\
(*w) & : & W \otimes V \multimap W \otimes U & \text{if } w \in V \otimes U \\
0_{V,W} & : & V \multimap W & \\
\iota_y^X & : & V_y \multimap \oplus_{x\in X} V_x & \text{if } y \in X \\
\pi_y^X & : & \oplus_{x\in X} V_x \multimap V_y & \text{if } y \in X \\
\Pi_{x\in X} f_x & : & \oplus_{x\in X} V_x \multimap \oplus_{x\in X} W_x & \text{if } f_x : V_x \multimap W_x \\
\Delta f & : & \oplus_{x\in X} V_x \multimap \oplus_{x\in X} W_x & \text{if } f : \oplus_{x\in X}(V_x \multimap W_x) \\
\langle f_x \rangle_{x\in X} & : & V \multimap \oplus_{x\in X} W_x & \text{if } f_x : V \multimap W_x \\
\mathrm{red}_R & : & V^X \multimap V^Y & \text{if } R \subseteq X \times Y \text{ is compact} \\
f^X & : & U^X \multimap V^X & \text{if } f : U \multimap V \\
\mathrm{id}_V & : & V \multimap V & \\
g \bullet f & : & U \multimap W & \text{if } f : U \multimap V, g : V \multimap W
\end{array}$$

3.4 Constant functions

We have the *constant functions*

$$K_w \quad : \quad V \to W \quad \text{if } w \in W$$

defined by $K_w(v) = w$.

3.5 Primitive functions

We furthermore assume we have named primitive functions $p_1, \ldots, p_n$ denoting analytic functions with associated derivative functions that are expressible as combinator expressions. For example, for each $k \in \mathbb{Z}/\{0\}$ we have the function $\_^k : \mathbb{R} \to \mathbb{R}$ with associated derivative $(\_^k)'(x) = ((k \cdot x^{k-1})\cdot)$. Note the $\cdot$ at the end; it is there since the Fréchet derivative at x is not a value from $\mathbb{R}$, but an element of $\mathbb{R} \multimap \mathbb{R}$, which is isomorphic with, but not the same as, $\mathbb{R}$. Similarly, we have $\ln : \mathbb{R} \to \mathbb{R}$ with associated $\ln'(x) = (x^{-1}\cdot)$; $\sin : \mathbb{R} \to \mathbb{R}$ with $\sin'(x) = ((\cos x)\cdot)$; $\cos : \mathbb{R} \to \mathbb{R}$ with $\cos'(x) = ((-\sin x)\cdot)$ and so on. Note that ln is only defined on $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x > 0\}$ and is thus, in particular, not analytic on all of $\mathbb{R}$.

We defer the subtleties of handling partially defined and not-everywhere differentiable functions in this paper to future work and assume henceforth for simplicity that our primitive functions are analytic on their entire domain.

In practice almost all primitive functions are functions on scalars and returning scalars. Primitive operators and functions on high-dimensional spaces are typically linear or bilinear.

3.6 Function composition

Every constant, linear, bilinear and primitive function constructed so far is an analytical function.

Finally we have sequential and parallel composition of analytical functions:

$$\begin{array}{rcll}
g \circ f & : & U \to W & \text{if } f : U \to V, g : V \to W \\
\Pi_{x\in X} f_x & : & \oplus_{x\in X} V_x \to \oplus_{x\in X} W_x & \text{if } f_x : V_x \to W_x \text{ for all } x \in X, X \text{ finite}
\end{array}$$

4 Fréchet differential calculus

Recall that $f' : V \to (V \multimap W)$ denotes the Fréchet derivative of $f : V \to W$. The linear function $f'(v)$ is the tangent of f at v. We provide differentiation rules for functions in combinatory form.

Theorem 4.1. *The following differentiation rules are valid for analytic functions on Hilbert spaces:*

$$
\begin{array}{rcll}
(g \circ f)'(v) & = & g'(f(v)) \bullet f'(v) & (1)\\
K_w{}'(v) & = & 0 & (2)\\
h'(v) & = & h \qquad \textit{if } h : V \multimap W & (3)\\
\diamond'(u,v) & = & (u\diamond) \bullet \pi_2 + (\diamond v) \bullet \pi_1 \qquad \textit{if } \diamond : U \times V \to_2 W & (4)\\
(\Pi_{x\in X} f_x)'(v) & = & \Delta((\Pi_{x\in X} f_x{}')(v)) \qquad \textit{if } f_x : V_x \to W_x & (5)
\end{array}
$$

Rule 1 is the *chain rule* for sequential composition. It expresses that the derivatives of g at $f(v)$ and of f at v are combined by composition $\bullet$ of linear functions.

Rules 2, 3 and 4 are for constant, linear and bilinear functions, respectively. Note in particular Rule 4, the *generalized product rule*. It is applicable to any bilinear function. The derivative of any bilinear function can be written in terms of the function itself; we do not need access to its definition, only its name. The same is true for linear functions; they are their own derivatives. All we need to know is that a function is linear to differentiate it. We will see that adjoint differentiation, which underlies reverse-mnode AD, requires processing its definition, however.

Finally, Rule 5 is for differentiating *parallel composition*. It is worth looking at special cases of it. Let $X = \mathbf{2}$, that is $\Pi_{x\in\mathbf{2}} f_x = f_1 \times f_2 : V_1 \times V_2 \to W_1 \times W_2$. We can calculate

$$
\begin{array}{rcl}
(f_1 \times f_2)'(v_1, v_2) & = & \Delta((f_1{}' \times f_2{}')(v_1, v_2))\\
& = & \Delta(f_1{}'(v_1), f_2{}'(v_2))\\
& = & f_1{}'(v_1) \times f_2{}'(v_2)
\end{array}
$$

Let us consider $f^X = \Pi_{x\in X} f$ where $f \in V \to W$.

$$
\begin{array}{rcl}
(f^X)'(v) & = & (\Pi_{x\in X} f)'(v)\\
& = & \Delta((\Pi_{x\in X} f')(v))\\
& = & \Delta(f'^X(v))
\end{array}
$$

In words, to differentiate f^X at value $v \in V^X$, we need to compute the derivative of f at each element v_x of V^X. This yields an element of $(V \multimap W)^X$; finally, Δ gathers these component-wise derivatives into a single derivative.

5 Affine interpretation

A function $h : V \to W$ is *affine* if it is the sum of a constant and a linear function, that is

$$h(v) = w + g(v)$$

for some $w \in W$ and $g \in V \multimap W$. Note that w and g are uniquely determined by h. We call them the *constant* and *linear* component of h, respectively, and write $h \in V \to_{\leq 1} W$ if h is affine.

$$
\begin{array}{rcll}
(g \circ f)^{[1]}(x) & = & \textbf{let } (fx, f'x) = f^{[1]}(x) \textbf{ in} & \\
 & & \textbf{let } (gfx, g'fx) = g^{[1]}(fx) \textbf{ in} & \\
 & & \quad (gfx, g'fx \bullet f'x) & \\
{K_w}^{[1]}(x) & = & (w, 0) & \\
h^{[1]}(x) & = & (h(x), h) & \text{if } h : V \multimap W \\
\diamond^{[1]}(x) & = & \underline{\textbf{let}}\ (u, v) = x \textbf{ in} & \\
 & & \quad (u \diamond v, (u \diamond) \bullet \pi_2 + (\diamond v) \bullet \pi_1) & \text{if } \diamond : U \times V \to_2 W \\
(\Pi_{y \in Y} f_y)^{[1]}(x) & = & \textbf{let } (w, d) = \mathit{unzip}((\Pi_{y \in Y} (\lambda x. {f_y}^{[1]}(x)))(x)) \textbf{ in} & \\
 & & \quad (w, \Delta(d)) \qquad \text{if } f_y : V_y \to W_y &
\end{array}
$$

Figure 1: Affine interpretation of functions in combinatory form. See Section 5.2 for an explanation of the underlined **let** .

We say that $g : V \to_{\leq 1} W$ is the *affine approximation of* $f : V \to W$ *at* $v \in V$ and write $f(v) \approx g$ if

$$\lim_{||dv||_V \to 0} \frac{||f(v+dv) - g(dv)||_W}{||dv||_V} = 0$$

Proposition 5.1. *A function has at most one affine approximation at v, written* $f^{[1]}(v)$*, where* $f^{[1]}(v)(dv) = f(v) + f'(v)(dv)$.

Thinking about differentiation in terms of computing affine approximations is useful since computing derivatives compositionally requires computing a function's value paired with its derivative [13]. The components of the affine approximation of a function in combinatory form can be computed by structural recursion. See Figure 1.

The parallel composition rule specializes to tuples and copowers as follows:

$$
\begin{array}{rcl}
(f_1 \times f_2)^{[1]}(x_1, x_2) & = & \textbf{let } (fx_1, f'x_1) = {f_1}^{[1]}(x_1), (fx_2, f'x_2) = {f_2}^{[1]}(x_2) \textbf{ in } ((fx_1, fx_2), (f'x_1, f'x_2)) \\
{f^X}^{[1]}(v) & = & \textbf{let } (w, d) = \mathit{unzip}({f^{[1]}}^X(v)) \textbf{ in } (w, \Delta(d))
\end{array}
$$

Note that unzipping the outputs of each component is the price we pay for separating the collective output into a value and a derivative component.

Theorem 5.2. *Assume* $p^{[1]}(v) = (p(v), p'(v))$ *for all primitive functions. Then* $f^{[1]}(v) = (f(v), f'(v))$ *for all functions in combinatory form.*

5.1 Automatic differentation

The affine approximation rules give rise to an interpreter

$$\text{eval}^{[1]}[\![\_]\!] : \text{Term}(V \to W) \to V \to \text{Term}(W \times (V \multimap W))$$

where $\text{Term}(V \to W)$ is a language for representing functions in combinatory form, including $\text{Term}(V \multimap W)$ as a (sub)language for representing linear functions in combinatory form: Just replace $t^{[1]}$ in Figure 1 by $\text{eval}^{[1]}[\![t]\!]$. When applied to a combinatory term t denoting f and a concrete value v, it returns a term containing the value $w = f(v)$ and a combinatory representation t' denoting the derivative $f'(v)$. This term t' can be optimized using the rules of linear and tensor algebra prior to applying an interpreter

$\mathrm{eval}^{(0)}[\![\_]\!] : \mathrm{Term}(V \multimap W) \to V \to W$ to t' and an input differential value dv, which yields the output differential $f'(v)(dv)$.

Behaviorally this corresponds to forward-mode automatic differentiation (AD), but is essentially different. In *elemental* and *tensor-based* forward-mode AD [18, 1] we have an interpreter $\mathrm{eval}^{(\mathrm{f})}[\![\_]\!]$ for a term $t : \mathrm{Term}(\oplus_{i\in\mathbf{n}}(V_i^{\mathbf{2}}) \to \oplus_{j\in\mathbf{m}}(V_j^{\mathbf{2}}))$ representing f such that

$$\mathrm{eval}^{(\mathrm{f})}[\![t]\!] \quad : \quad \oplus_{i\in\mathbf{n}}(V_i^{\mathbf{2}}) \to \oplus_{j\in\mathbf{m}}(V_j^{\mathbf{2}}).$$

requires both values and associated differentials as inputs at the same time. It computes

$$\mathrm{eval}^{(\mathrm{f})}[\![t]\!]((v_1,dv_1),\ldots,(v_n,dv_n)) \quad = \quad ((y_1,dy_1),\ldots,(y_m,dy_m)).$$

where $(y_1,\ldots,y_m) = f(v_1,\ldots,v_n)$ and $(dy_1,\ldots,dy_m) = f'(v_1,\ldots,v_n)(dv_1,\ldots,dv_n)$.

The type of $\mathrm{eval}^{(\mathrm{f})}[\![t]\!]$ camouflages that the output values $y_1,\ldots,y_n$ do not depend on the second components $dv_1,\ldots,dv_n$, and that for fixed $v_1,\ldots,v_n$ the $dy_1,\ldots,dy_m$ are linear functions of $dv_1,\ldots,dv_n$. Note, in particular, that both values and differentials must be provided before execution can start.

In our formulation $\mathrm{eval}^{[1]}[\![t]\!]$ requires no input differentials to run the code, only the input values $v_1,\ldots,v_n$, which manifests that the output values do not depend on any differentials. Furthermore, the derivative is returned as a term in a language that guarantees that it denotes a linear function.

5.2 Symbolic differentiation

The affine approximation rules are carefully written to facilitate *symbolic differentiation* by applying $\mathrm{eval}^{[1]}[\![t]\!]$ to a symbolic variable x. This amounts to *specializing* the code of $\mathrm{eval}^{[1]}[\![\_]\!]$ to the concrete t by partial evaluation.

The **let** -expressions without underlining can be eliminated by substitution, that is rewriting **let** $(u,v) = (f,g)$ **in** g to $g[f/u,g/v]$ during partial evaluation. Since the let-bound variables have single occurrences the size of the expression does not grow. The **let** -expression for bilinear functions should not be eliminated, however, since its let-bound variables are used twice. Substituting them would cause *expression swell*. Conversely, not substituting them avoids expression swell: The size of the symbolically differentiated expression is linear in the size of the input expression. Retaining **let** in the output is the reason for having

$$\mathrm{eval}^{[1]}[\![\_]\!] : \mathrm{Term}(V \to W) \to V \to \mathrm{Term}(W \times (V \multimap W))$$

rather than

$$\mathrm{eval}^{[1]}[\![\_]\!] : \mathrm{Term}(V \to W) \to V \to (W \times \mathrm{Term}(V \multimap W)).$$

This supports and generalizes to non-elemental symbolic differentiation that expression swell is a myth [33]: Retaining sharing when applying the (generalized) product rule is both necessary and sufficient to avoid it.

6 Adjoints

The *dual vector space* V^* of vector space V is the vector space of *linear functionals*, also called *covectors*, $V \multimap \mathbb{R}$ where $\mathbb{R}$ is the underlying field of V. By the Riesz representation theorem, the inner product induces an isomorphism *dual* defined by

$$\begin{aligned} \mathit{dual} &: V \multimap V^* \\ \mathit{dual}(v) &= (\odot v). \end{aligned}$$

In particular, $dual^{-1}(\odot v) = v$, and v and $(\odot v)$ are called *duals* of each other.[1]

Some applications require computing the dual of a covector. For example, given a *scalar* function $f : V \to \mathbb{R}$, the *gradient* $\nabla f : V \to V$ is defined by

$$\nabla f(v) = dual^{-1}(f'(v)).$$

If we implement covectors as functions that can only be applied, the only way of implementing $dual^{-1}$ is by applying it to each of the base vectors of V, which is problematic if V is of high dimension, say a million or a billion.

Similarly, sometimes we may want to implement the *transpose*

$$\begin{array}{rcl} f^\dagger & : & W^* \multimap V^* \\ f^\dagger & = & (\bullet f) \end{array}$$

of $f : V \multimap W$. The transpose is the *continuation-passing style* version of f where a linear continuation is passed as the first argument. To wit, we have $f^\dagger(\kappa)(v) = \kappa(f(v))$ where κ is the continuation.

A general idea permeating mathematical and computer science applications of linear algebra is representing a covector by its dual vector v with an indication that it represents $(\odot v)$ ("I am contravariant"), not v itself.

We would thus like to find a linear function $f^* : W \multimap V$ that implements the transpose $f^\dagger$ by using ordinary vectors rather than covectors to represent linear continuations; that is, it should be the case that $f^*(w) = v$ whenever $f^\dagger(\odot w) = (\odot v)$.

Definition 6.1. $f^* : W \multimap V$ is the *adjoint* of $f : V \multimap W$ if

$$f^\dagger(\odot w) = (\odot v) \Leftrightarrow f^*(w) = v$$

By the Riesz representation theorem we immediately have that

Proposition 6.2. *f^* exists and is unique for Hilbert spaces.*

The defining property of an adjoint can be restated as the familiar property where f is pushed from one argument to the other argument of the inner product.

Proposition 6.3. *$f^* : W \multimap V$ is the adjoint of $f : V \multimap W$ if and only if $f(v) \odot w = v \odot f^*(w)$ for all $v \in V, w \in W$.*

We can implement $dual^{-1}$ using the adjoint:

Proposition 6.4. *Let $f : V \multimap \mathbb{R}$, that is $f \in V^*$. Then $dual^{-1}(f) = f^*(1)$ and thus $\nabla f(v) = (f'(v))^*(1)$.*

Linear functions are built from other linear functions. We provide general rules for calculating adjoints symbolically of linear functions in combinatory form.

6.1 Adjoint calculation

Adjoints can be calculated symbolically for linear functions in combinatory form.

[1] For finite index sets the constructible Hilbert spaces are finite-dimensional. Note that the Riesz representation theorem also holds for infinite-dimensional Hilbert spaces.

Theorem 6.5. *Let X, Y be finite sets, $R \subseteq X \times Y$, and $R^T = \{(y,x) \mid (x,y) \in R\}$. Then:*

$$\begin{aligned}
\mathrm{id}^* &= \mathrm{id} \\
(g \bullet f)^* &= f^* \bullet g^* \\
0^* &= 0 \\
(v*)^* &= (v^T *) \\
(*w)^* &= (*w^T) \\
(\iota_x^X)^* &= \pi_x^X \\
(\Pi_{x \in X} f_x)^* &= \Pi_{x \in X} {f_x}^* \\
{\mathrm{red}_R}^* &= \mathrm{red}_{R^T}
\end{aligned}$$

Furthermore, the inverses of unitary operators are also their adjoints.

These rules are not accidentally symmetric; indeed the above language of linear functions has been *designed* to yield these symmetries, where each construct has an adjoint construct. This is the reason for using tensor contraction $*$ instead of $\otimes$ as the primitive universal bilinear function.

The adjoint of a partially applied tensor contraction $(v*)$ is the *transpose*[2] v^T of v. Recall that $\_^T : V \otimes W \multimap W \otimes V$ swaps components of formal tensor product sums.

Note also that the adjoint of red_R is red_{R^T}; in particular, if R is a function, R^T is generally not a function. Allowing for R to be a relation rather than restricting it to be a function makes expressing its adjoint in terms of red possible.

The adjoints for other operations can be derived from their definitions in terms of these primitive functions and constructs. For example, we can derive

$$\begin{aligned}
(k\cdot)^* &= (k\cdot) \\
(\cdot v)^* &= (\odot v)
\end{aligned}$$

6.2 Adjoint differentiation

The *adjoint derivative* of $f : V \to W$ at $v \in V$ is $(f'(v))^*$. We can compute it by employing our affine interpreter $\mathrm{eval}^{[1]}[\![f]\!]$; applying it to v; extracting the term representing $f'(v)$; applying the adjoint calculation rules of Theorem 6.5 to calculate a term representating $(f'(v))^*$; and finally applying the derived adjoint to output differentials to compute input differentials.[3] Alternatively, we can construct the adjoint derivative during affine interpretation; see Figure 2.

This provides us with a method behaviorally equivalent to reverse-mode automatic differentiation. In the first phase, the value of f at v and the term of $(f'(v))^*$, which includes all—and only— the relevant intermediate results of $f(v)$ are computed by $f^{[1r]}$ from v alone. Only in the second phase, the output term representing $(f(v))^*$ is interpreted as a function by applying it to output differentials. After the first phase, the adjoint derivative can be optimized using algebraic simplications, and it can be compiled for efficient data parallel execution on a GPU.

Going one step further, the first phase can be done symbolically, which amounts to a specialization of the adjoint affine interpreter to the particular source code for f. The result of doing so can be compiled for efficient data parallel execution before the values of v and output differential dy are available.

[2]Not to be confused with the transpose of a linear function, which it is related, but different.

[3]When used in the adjoint direction, from output to input, the variables containing differentials are often called adjoint variables.

$$
\begin{array}{rcll}
(g \circ f)^{[1r]}(x) & = & \textbf{let } (fx, f'xa) = f^{[1r]}(x) \textbf{ in} & \\
 & & \textbf{let } (gfx, g'fxa) = g^{[1r]}(fx) \textbf{ in} & \\
 & & \quad (gfx, f'xa \bullet g'fxa) & \\
K_w{}^{[1r]}(x) & = & (w, 0) & \\
h^{[1r]}(x) & = & (h(x), h^*) & \text{if } h : V \multimap W \\
\diamond^{[1r]}(x) & = & \underline{\textbf{let}}\ (u, v) = x \textbf{ in} & \\
 & & \quad (u \diamond v, \iota_2^{\mathbf{2}} \bullet (u\diamond)^* + \iota_1^{\mathbf{2}} \bullet (\diamond v)^*) & \text{if } \diamond : U \times V \to_2 W \\
(\Pi_{y \in Y} f_y)^{[1r]}(x) & = & \textbf{let } (w, d) = \mathit{unzip}((\Pi_{y \in Y} (\lambda x. f_y{}^{[1r]}(x)))(x)) \textbf{ in} & \\
 & & \quad (w, \Delta(d)) & \text{if } f_y : V_y \to W_y
\end{array}
$$

Figure 2: Adjoint affine interpretation of functions in combinatory form

7 Applications

We illustrate combinatory differentiation by applying it to neural networks. Additional examples showing the application of equational reasoning to derive derivatives (sic!) can be found in Appendix B.

Example 7.1. A k-layer neural network N_k consists of a composition of k layers with the i-th layer given by

$$g_i(x_i, W_i, b_i) = h_i{}^{\mathbf{m}_i}(W_i \star x_i + b_i),$$

where $W_i \in \mathbb{R}^{\mathbf{m}_i \times \mathbf{m}_{i-1}}$, $b_i \in \mathbb{R}^{\mathbf{m}_i}$, and $x_i \in \mathbb{R}^{\mathbf{m}_{i-1}}$ along with a loss function $l(v, y) : \mathbb{R}^{\mathbf{m}_k} \to \mathbb{R}$:

$$
\begin{aligned}
&N_k : \mathbb{R}^{\mathbf{m}_0} \times \mathbb{R}^{\mathbf{m}_1 \times \mathbf{m}_0} \times \mathbb{R}^{\mathbf{m}_1} \times \mathbb{R}^{\mathbf{m}_2 \times \mathbf{m}_1} \times \mathbb{R}^{\mathbf{m}_2} \times \cdots \times \mathbb{R}^{\mathbf{m}_k \times \mathbf{m}_{k-1}} \times \mathbb{R}^{\mathbf{m}_k} \times \mathbb{R}^{\mathbf{m}_k} \to \mathbb{R} \\
&N_k(x, W_1, b_1, W_2, b_2, \ldots, W_k, b_k, y) = l(g_k(W_k, b_k(\cdots(g_2(W_2, b_2(g_1(W_1, b_1, x))))))), y).
\end{aligned}
$$

For simplicity, we use $l(v, y) = (v - y) \odot (v - y)$ for the loss function. In point-free form, the i-th layer of the network must propagate the inputs for all subsequent layers; g_i and l in point-free form are

$$
\begin{aligned}
g_i &= \langle h_i{}^{\mathbf{m}_i} \circ ((\star) \circ \langle \pi_2^{\mathbf{n}_i}, \pi_1^{\mathbf{n}_i} \rangle + \pi_3^{\mathbf{n}_i}), \pi_4^{\mathbf{n}_i}, \ldots, \pi_{n_i}^{\mathbf{n}_i} \rangle, \\
l &= (\odot) \circ \mathrm{dup} \circ (\pi_1^{\mathbf{2}} - \pi_2^{\mathbf{2}}),
\end{aligned}
$$

where $\mathbf{n}_i = 2(k + 2 - i)$. Hence, the entire network is constructed as

$$N_k = l \circ g_k \circ \cdots g_2 \circ g_1.$$

Applying the differentiation rules of Theorem 4.1, we differentiate g_i and l

$$
\begin{aligned}
&g_i'(x_i, W_i, b_i, \ldots, W_k, b_k, y) \\
&= \{\text{definition of } g_i\} \\
&\quad \langle {h_i}^{\mathbf{m}_i} \circ ((\star) \circ \langle \pi_2^{\mathbf{n}_i}, \pi_1^{\mathbf{n}_i} \rangle + \pi_3^{\mathbf{n}_i}), \pi_4^{\mathbf{n}_i}, \ldots, \pi_{n_i}^{\mathbf{n}_i} \rangle'(x_i, W_i, b_i, \ldots, W_k, b_k, y) \\
&= \{\text{by Rules 1, 3, and 5}\} \\
&\quad \langle ({h_i}^{\mathbf{m}_i})'(W_i \star x_i + b_i) \bullet ((\star)'(\langle W_i, x_i \rangle) \bullet \langle \pi_2^{\mathbf{n}_i}, \pi_1^{\mathbf{n}_i} \rangle + \pi_3^{\mathbf{n}_i}), \pi_4^{\mathbf{n}_i}, \ldots, \pi_{n_i}^{\mathbf{n}_i} \rangle \\
&= \{\text{by Rules 5 and 4 }\} \\
&\quad \langle \Delta({h_i'}^{\mathbf{m}_i}(W_i \star x_i + b_i)) \bullet (((W_i \star) \bullet \pi_2^{\mathbf{2}} + (\star x_i) \bullet \pi_1^{\mathbf{2}}) \bullet \langle \pi_2^{\mathbf{n}_i}, \pi_1^{\mathbf{n}_i} \rangle + \pi_3^{\mathbf{n}_i}), \pi_4^{\mathbf{n}_i}, \ldots, \pi_{n_i}^{\mathbf{n}_i} \rangle,
\end{aligned}
$$

$$
\begin{aligned}
&l'(v, y) \\
&= \{\text{definition of } l\} \\
&\quad ((\odot) \circ \mathrm{dup} \circ (\pi_1^{\mathbf{2}} - \pi_2^{\mathbf{2}}))'(v, y) \\
&= \{\text{by Rule 1}\} \\
&\quad (\odot)'(v - y, v - y) \bullet (\mathrm{dup} \circ (\pi_1^{\mathbf{2}} - \pi_2^{\mathbf{2}}))'(v, y) \\
&= \{\text{by Rules 1, 4, and 3}\} \\
&\quad (((v - y) \odot) \bullet \pi_2^{\mathbf{2}} + (\odot(v - y)) \bullet \pi_1^{\mathbf{1}}) \bullet (\mathrm{dup} \bullet (\pi_1^{\mathbf{2}} - \pi_2^{\mathbf{2}})).
\end{aligned}
$$

Repeated application of Rule 1 now yields the entire differentiated network

$$
\begin{aligned}
N_k'(x, W_1, b_1, W_2, b_2, \ldots, W_k, b_k, y) = {} & l'((g_k \circ \cdots \circ g_1)(x, W_1, b_1, W_2, b_2, \ldots, W_k, b_k, y)) \\
& \bullet\ g_k'((g_{k-1} \circ \cdots \circ g_1)(x, W_1, b_1, W_2, b_2, \ldots, W_k, b_k, y)) \\
& \bullet \cdots \bullet\ g_1'(x, W_1, b_1, W_2, b_2, \ldots, W_k, b_k, y).
\end{aligned}
$$

Straight-forward application of Theorem 6.5 may subsequently be used to obtain the adjoint of N_k', $(N_k'(x, W_1, b_1, W_2, b_2, \ldots, W_k, b_k, y))^* : \mathbb{R} \multimap \mathbb{R}^{\mathbf{m}_0} \times \mathbb{R}^{\mathbf{m}_1 \times \mathbf{m}_0} \times \cdots \times \mathbb{R}^{\mathbf{m}_k}$.

8 Discussion

We have provided a functional-analysis based compositional framework for differentiation and adjoint differentiation that encompasses both symbolic and automatic differentiation. It highlights that, very generally, adjoint differentiation is the combination of symbolic Fréchet differentiation and symbolic calculation of adjoints over Hilbert spaces, where both derivatives and adjoints retain the data parallelism in their input functions.

Why Hilbert spaces? A Hilbert space is a vector space that is equipped with an inner product $\odot$ and is metrically complete. The inner product is crucial: it establishes an isomorphism with the dual space such that ordinary first-order vectors can be used to represent linear functionals rather than having to code these as procedures in a programming language. This representation trick is the essence of adjoints, which run linear functions in reverse, from output differential to input differential and, in particular, compute gradients as the input differentials resulting from a single evaluation of the adjoint derivative to the output differential 1.

The metric completness is, in some sense, irrelevant: it only pops up in the definition of Fréchet derivative and checking that it constitutes a valid model of the differentiation rules.

Why symbolic tensor products? We could define the tensor product of $\mathbb{R}^{\mathbf{m}}$ and $\mathbb{R}^{\mathbf{n}}$ to be the matrix space $\mathbb{R}^{\mathbf{m}\times\mathbf{n}}$, but matrices as data structures for derivatives are too inefficient for large values of m,n. Symbolic tensor decompositions can provide more efficient representations [16].

For example, Griewank [19] gives $f : \mathbb{R}^{\mathbf{n}} \to \mathbb{R}^{\mathbf{m}}$ defined by

$$f(x) = b\sin(a^T x)$$

with $a \in \mathbb{R}^{\mathbf{n}}$ and $b \in \mathbb{R}^{\mathbf{n}}$ as an example where computing the Jacobian derivative of f at x_0 requires $m \cdot n$ multiplications, whereas the original function requires only $n+m$ multiplications. But this is only due to insisting on representing the derivative as a Jacobian matrix. Translated into combinatory form and employing our differentiation rules we arrive at a corresponding representation of the matrix as

$$c \cdot (b \otimes a)$$

where $c = \cos(a^T x_0)$. Note that this is the output: itt uses symbolic scalar product and tensor product operators. This representation can be computed using only n scalar multiplications. Furthermore, applying it to a vector $dx \in \mathbb{R}^{\mathbf{n}}$ produces the term $d \cdot b$ where $d = c \cdot (a \odot dx)$, which requires only $n+1$ multiplications. Using the matrix equivalent of $c \cdot (b \otimes a)$ takes $m \cdot n$ multiplications.

8.1 Related work

The origin of symbolic differentiation using electronic computers for functions on scalar variables dates back to the 1950s [30]. Forward-mode AD for scalar variables was discovered independently by a number of researchers in the 1950s and 1960s [21]. The history of reverse-mode AD dates back to the early 1970s and is surveyed by Griewank [19]. Linnainmaa [34, 35] observed early on that reverse-mode AD on scalar variables consists of building a computation graph [2] and then reversing its dependency arrows, which is tantamount to transposing sparse matrices in a sequential composition of matrix multiplications. This observation has since been made repeatedly in both elemental and tensor settings.

Derivatives of functions on scalar variables are conventionally represented by Jacobian matrices. Computing the matrix with a minimum number of steps is NP-hard, however [40]. As we have shown, matrices are often not even a good data structure for derivatives, however. As we have shown, they can be represented more compactly and efficiently using a combinatory language for linear functions, including symbolic operators for scalar multiplication, addition and tensor product [37, 25].

Functional languages have served well for exploring AD techniques both as a host language for capturing AD techniques [31, 32], and as the language under investigation, featuring, for instance, multivariate functions, higher-dimensional data, higher-order functions, higher-degree differentiation (e.g., through a lazy infinite tower of derivatives) [15, 22], and even differentiation of formal languages [14]. Whereas many of the above-mentioned features are well-suited for forward-mode AD (no memoization of primal values is needed), capturing the essence of reverse-mode AD has proven difficult. We believe this is due to using λ-calculus formulation [45] rather than a combinatory formulation, representing the "tape" as (the code of) a function or procedure [13, 51] and/or employing matrices to represent linear functions instead of asymptotically more compact and efficient data structures made possible by symbolic tensor products and useful constants (identity, projections and injections).

Reverse-mode AD for higher-order languages has been studied by Mazza [38] and on capturing reverse-mode AD by building a library for functional representation general differentiation based on the specifications of the functionality . Our work follows Elliott's [13] lead:[4] It is also based on adjoint affine

[4]While inspired by Elliott's elegant presentation of Fréchet derivatives in a Haskell framework [15], our work on functional-analysis based AD started in 2015 and developed independently of Elliott's work, but has so far remained unpublished except for a presentation at a Workshop in honor of Tom Reps's 60th birthday in 2016 [24].

interpretation, but additionally employs specialized and efficient representations for linear functions, supports sums, tensor products and copowers, avoids expression swell, identifies and exploits general differentiation rules for bilinear operators, and supports relational reduction and parallel composition.

Other work has focused on exploring how AD techniques can be applied in a functional parallel setting while preserving the parallel properties of functions, also in the differentiated code [43, 44]. Recent work [46] on reverse- and forward-mode AD operators embedded in Futhark [26] has shown that even nested parallelism can be handled in reverse mode effectively with excellent GPU-utilization and performance. Our combinatory language features powerful parallel operations and general differentiation and adjoint rules that retain semantic data parallelism, but does not devise a general implementation method for compact representation of relations and efficient parallel implementation of relational reduction. This is future work.

An approach to combinatory differentiation based on category theory rather than linear algebra is *differential categories* [6, 10] and its many variations (e.g., [7, 9, 11]). In brief, a differential category is an additive monoidal category with a differential combinator and a modality allowing differentiable morphisms to be identified by their signature. The variation most closely resembling the one presented here is that of *reverse derivative categories* [11], as they can be thought of as categories of smooth maps equipped with the ability to take adjoints (a "dagger"). This approach is ultimately closer to symbolic differentiation rather than AD, though it could be interesting to integrate our approach into a notion of differential category with a distinction between semantic and syntactic data (see also [12]).

The categorical semantics of both forward and reverse mode AD with higher types was recently given a unified treatment in [50], with models based on so-called *biadditive* categories: indexed categories with biproducts at each index, preserved by reindexing. Interestingly, when applying the Grothendieck construction $\int(-)$ to a biadditive category $\mathbf{C}$, the resulting fibred category $\int \mathbf{C}$ describes forward mode AD, while its dual $\int \mathbf{C}^{\mathrm{op}}$ describes reverse mode AD. This highlights the formal connection between duality (via adjoints) and reverse mode AD, as also argued in [11] and in Section 6.2.

On the practical side, a variety of systems provide tooling for automatically differentiating source code. These tools include (but are far from limited to) Python tools such as Autograd [36], JAX [8, 47], C/C++ tools such as Adept [27], ADOL-C [20] and Tapenade [23], DiffSharp for F# [3], tools for MATLAB [5, 41], Julia [28], FutharkAD [46] and even tools for the LLVM IR [39]. Most of these tools feature both forward-mode and backward-mode AD and are therefore applicable for a variety of domains and applications, including physics simulation [47], finance [22, 4, 17] and economics [49]. AD has also received renewed attention due to its application to deep learning, where backpropagation is reverse-mode AD for scalar functions, as shown in Example 7.1. AD techniques have therefore been incorporated, either directly or indirectly (through library APIs), into most of the major general machine learning frameworks, including Caffe [29], TensorFlow [1], and PyTorch [42]. For a general overview, consult [3]. Work has also been done at benchmarking many of the commonly used AD tools [48].

Acknowledgements. This work was made possible by Independent Research Fund Denmark grants *FUTHARK: Functional Technology for High-performance Architectures*, *Deep Probabilistic Programming for Protein Structure Prediction (DPP)*, and DFF–International Postdoc 0131-00025B. We would like to thank Gabriele Keller, Ken Friis Larsen and Dimitrios Vytionitis for collaborative discussions over the last six years that have greatly helped in developing the foundations of combinatory differentiation and our colleagues on FUTHARK and DPP, in particular Cosmin Oancea, Troels Henriksen, Thomas Hamelryck and Ola Rønning. Furthermore, the second author would like to thank Conal Elliott for stimulating exchanges on AD in the period he was working on his ICFP 2018 paper [13]. We greatly appreciate and thank the three anonymous referees for their recommendations.

References

[1] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin et al. (2016): *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*. arXiv preprint arXiv:1603.04467, doi:10.48550/ARXIV.1603.04467.

[2] F. L. Bauer (1974): *Computational Graphs and Rounding Error*. SIAM Journal on Numerical Analysis 11(1), pp. 87–96, doi:10.1007/BF01386233.

[3] Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul & Jeffrey Mark Siskind (2018): *Automatic Differentiation in Machine Learning: A Survey*. arXiv:1502.05767 [cs, stat], doi:10.48550/ARXIV.1502.05767. arXiv:1502.05767.

[4] C. H. Bischof, H. M. Bücker & B. Lang (2002): *Automatic Differentiation for Computational Finance*. In E. J. Kontoghiorghes, B. Rustem & S. Siokos, editors: Computational Methods in Decision-Making, Economics and Finance, chapter 15, Applied Optimization 74, Kluwer Academic Publishers, Dordrecht, pp. 297–310, doi:10.1007/978-1-4757-3613-7_15.

[5] Christian H. Bischof, H. Martin Bücker, Bruno Lang, Arno Rasch & Andre Vehreschild (2002): *Combining Source Transformation and Operator Overloading Techniques to Compute Derivatives for MATLAB Programs*. In: Proceedings of the Second IEEE International Workshop on Source Code Analysis and Manipulation (SCAM 2002), IEEE Computer Society, Los Alamitos, CA, USA, pp. 65–72, doi:10.1109/SCAM.2002.1134106.

[6] R. F. Blute, J. R. B. Cockett & R. A. G. Seely (2006): *Differential categories*. Mathematical Structures in Computer Science 16(6), pp. 1049–1083, doi:10.1017/S0960129506005676.

[7] R. F. Blute, J. R. B. Cockett & R. A. G. Seely (2009): *Cartesian differential categories*. Theory and Applications of Categories 22(23), pp. 622–672.

[8] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne & Qiao Zhang (2018): *JAX: Composable Transformations of Python+NumPy Programs*. Available at `http://github.com/google/jax`.

[9] J. R. B. Cockett, G. S. H. Cruttwell & J. D. Gallagher (2011): *Differential restriction categories*. Theory and Applications of Categories 25(21), pp. 537–613, doi:10.48550/ARXIV.1208.4068.

[10] J. R. B. Cockett & J.-S. Lemay (2017): *There Is Only One Notion of Differentiation*. In Dale Miller, editor: 2nd International Conference on Formal Structures for Computation and Deduction (FSCD 2017), Leibniz International Proceedings in Informatics (LIPIcs) 84, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 13:1–13:21, doi:10.4230/LIPIcs.FSCD.2017.13.

[11] R. Cockett, G. Cruttwell, J. Gallagher, J.-S. Pacaud Lemay, B. MacAdam, G. Plotkin & D. Pronk (2020): *Reverse Derivative Categories*. In M. Fernández & A. Muscholl, editors: 28th EACSL Annual Conference on Computer Science Logic (CSL 2020), Leibniz International Proceedings in Informatics (LIPIcs) 152, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 18:1–18:16, doi:10.4230/LIPIcs.CSL.2020.18.

[12] O. Danvy (1996): *Type-Directed Partial Evaluation*. In: Proceedings of the 23rd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '96), ACM, pp. 242–257, doi:10.1145/237721.237784.

[13] Conal Elliott (2018): *The simple essence of automatic differentiation*. Proceedings of the ACM on Programming Languages 2(ICFP), p. 70, doi:10.1145/355586.364791.

[14] Conal Elliott (2021): *Symbolic and automatic differentiation of languages*. Proceedings of the ACM on Programming Languages 5(ICFP), pp. 1–18, doi:10.1016/S0019-9958(61)80020-X.

[15] Conal M. Elliott (2009): *Beautiful Differentiation*. In: Proceedings of the 14th ACM SIGPLAN International Conference on Functional Programming, ICFP '09, Association for Computing Machinery, New York, NY, USA, pp. 191–202, doi:10.1145/1596550.1596579.

[16] Patrick Gelß (2017): *The Tensor-Train Format and Its Applications: Modeling and Analysis of Chemical Reaction Networks, Catalytic Processes, Fluid Flows, and Brownian Dynamics*. Ph.D. thesis, Freie Universität Berlin.

[17] Michael B. Giles & Paul Glasserman (2006): *Smoking Adjoints: fast evaluation of Greeks in Monte Carlo calculations*. In Chris Kenyon & Andrew Green, editors: *Landmarks in XVA: From Counterparty Risk to Funding Costs and Capital*, chapter 25, Risk books, Infopro digital, Houndsditch, London.

[18] Andreas Griewank (1989): *On Automatic Differentiation*. In: *Mathematical Programming: Recent Developments and Applications*, pp. 83–108. ISBN 978-0792304906.

[19] Andreas Griewank (2012): *Who invented the reverse mode of differentiation*. *Documenta Mathematica, Extra Volume ISMP*, pp. 389–400.

[20] Andreas Griewank, David Juedes & Jean Utke (1996): *Algorithm 755: ADOL-C: A Package for the Automatic Differentiation of Algorithms Written in C/C++*. *ACM Transactions on Mathematical Software* 22(2), pp. 131–167. Available at `http://doi.acm.org/10.1145/229473.229474`.

[21] Andreas Griewank & Andrea Walther (2008): *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Siam, doi:10.1137/1.9780898717761.

[22] Esben Bistrup Halvorsen (2012): *Calculating Key Ratios for Financial Products using Automatic Differentiation and Monte Carlo Simulation*. Student Project, Department of Computer Science, University of Copenhagen (DIKU).

[23] Laurent Hascoet & Valérie Pascual (2013): *The Tapenade Automatic Differentiation Tool: Principles, Model, and Specification*. *ACM Trans. Math. Softw.* 39(3), doi:10.1145/2450153.2450158.

[24] Fritz Henglein (2016): *Automatic Differentiation: From Functional Analysis to Functional Programming*. Presentation, Reps at Sixty Workshop at Static Analysis Symposium.

[25] Fritz Henglein, Robin Kaarsgaard & Mikkel Kragh Mathiesen (2022): *The Programming of Algebra*. In: *Proc. 9th Workshop on Mathematically Structured Functional Programming (MSFP)*, Electronic Proceedings in Theoretical Computer Science (EPTCS), Munich, Germany.

[26] Troels Henriksen, Niels G. W. Serup, Martin Elsman, Fritz Henglein & Cosmin E. Oancea (2017): *Futhark: Purely Functional GPU-programming with Nested Parallelism and In-place Array Updates*. In: *Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2017, ACM, New York, NY, USA, pp. 556–571, doi:10.1145/3062341.3062354.

[27] Robin J. Hogan (2014): *Fast Reverse-Mode Automatic Differentiation Using Expression Templates in C++*. *ACM Transactions on Mathematical Software* 40(4), pp. 26:1–26:24, doi:10.1145/2560359.

[28] Michael Innes (2019): *Don't Unroll Adjoint: Differentiating SSA-Form Programs*. *arXiv:1810.07951 [cs]*, doi:10.48550/ARXIV.1810.07951. arXiv:1810.07951.

[29] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama & Trevor Darrell (2014): *Caffe: Convolutional Architecture for Fast Feature Embedding*. In: *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, Association for Computing Machinery, New York, NY, USA, pp. 675–678, doi:10.1145/2647868.2654889.

[30] Harry G Kahrimanian (1953): *Analytical differentiation by a digital computer*. *MA Thesis, Temple University*.

[31] Jerzy Karczmarczuk (1998): *Functional Differentiation of Computer Programs*. In: *Proceedings of the Third ACM SIGPLAN International Conference on Functional Programming*, ICFP '98, Association for Computing Machinery, New York, NY, USA, pp. 195–203, doi:10.1145/289423.289442.

[32] Jerzy Karczmarczuk (1999): *Functional Coding of Differential Forms*. In: *Scottish Workshop on Functional Programming*. ISBN 978-1-84150-024-9.

[33] Sören Laue (2019): *On the equivalence of forward mode automatic differentiation and symbolic differentiation*. *arXiv preprint arXiv:1904.02990*, doi:10.48550/arXiv.1904.02990.

[34] Seppo Linnainmaa (1970): *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors*. Master's Thesis (in Finnish), Univ. Helsinki, pp. 6–7.

[35] Seppo Linnainmaa (1976): *Taylor Expansion of the Accumulated Rounding Error*. BIT 16(2), pp. 146–160, doi:10.1007/BF01931367.

[36] Dougal Maclaurin (2016): *Modeling, Inference and Optimization with Composable Differentiable Procedures*. Ph.D. thesis, Harvard University.

[37] Mikkel Kragh Mathiesen (2016): *Infinite-Dimensional Linear Algebra for Efficient Query Processing*. Master's thesis, Department of Computer Science, Unversity of Copenhagen (DIKU).

[38] Damiano Mazza & Michele Pagani (2021): *Automatic Differentiation in PCF*. Proceedings of the ACM on Programming Languages 5(POPL), pp. 1–27, doi:10.1145/3434309. arXiv:2011.03335.

[39] William S. Moses, Valentin Churavy, Ludger Paehler, Jan Hückelheim, Sri Hari Krishna Narayanan, Michel Schanen & Johannes Doerfert (2021): *Reverse-Mode Automatic Differentiation and Optimization of GPU Kernels via Enzyme*. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '21, Association for Computing Machinery, New York, NY, USA, doi:10.1145/3458817.3476165.

[40] Uwe Naumann (2007): *Optimal Jacobian Accumulation Is NP-Complete*. Mathematical Programming 112(2), pp. 427–441, doi:10.1007/s10107-006-0042-z.

[41] Richard D. Neidinger (2010): *Introduction to Automatic Differentiation and MATLAB Object-Oriented Programming*. SIAM Review 52(3), pp. 545–563, doi:10.1137/080743627.

[42] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga & Adam Lerer (2017): *Automatic Differentiation in PyTorch*. In: Proc. 31st Conference on Neural Information Processing Systems (NIPS).

[43] Adam Paszke, Daniel Johnson, David Duvenaud, Dimitrios Vytiniotis, Alexey Radul, Matthew Johnson, Jonathan Ragan-Kelley & Dougal Maclaurin (2021): *Getting to the Point. Index Sets and Parallelism-Preserving Autodiff for Pointful Array Programming*. arXiv:2104.05372 [cs], doi:10.48550/ARXIV.2104.05372. arXiv:2104.05372.

[44] Adam Paszke, Matthew J. Johnson, Roy Frostig & Dougal Maclaurin (2021): *Parallelism-Preserving Automatic Differentiation for Second-Order Array Languages*. In: Proceedings of the 9th ACM SIGPLAN International Workshop on Functional High-Performance and Numerical Computing, FHPNC 2021, Association for Computing Machinery, New York, NY, USA, pp. 13–23, doi:10.1145/3471873.3472975.

[45] Barak A. Pearlmutter & Jeffrey Mark Siskind (2008): *Reverse-Mode AD in a Functional Framework: Lambda the Ultimate Backpropagator*. ACM Transactions on Programming Languages and Systems 30(2), pp. 1–36, doi:10.1145/1330017.1330018.

[46] Robert Schenck, Ola Rønning, Troels Henriksen & Cosmin E. Oancea (2022): *AD for an Array Language with Nested Parallelism*, doi:10.48550/arXiv.2202.10297. arXiv:2202.10297.

[47] Samuel Schoenholz & Ekin Dogus Cubuk (2020): *JAX MD: A Framework for Differentiable Physics*. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin, editors: Advances in Neural Information Processing Systems, 33, Curran Associates, Inc., pp. 11428–11441. Available at `https://proceedings.neurips.cc/paper/2020/file/83d3d4b6c9579515e1679aca8cbc8033-Paper.pdf`.

[48] Filip Šrajer, Zuzana Kukelova & Andrew Fitzgibbon (2018): *A Benchmark of Selected Algorithmic Differentiation Tools on Some Problems in Computer Vision and Machine Learning*. arXiv:1807.10129 [cs], doi:10.48550/arXiv.1807.10129. arXiv:1807.10129.

[49] E. M. Tadjouddine (2009): *Algorithmic Differentiation Applied to Economics*. In S. I. Ao, O. Castillo, C. Douglas, D. D. Feng & J.-A. Lee, editors: Proceedings of the of the International MultiConference of Engineers and Computer Scientists 2009 (IMECS 2009), Hong Kong, March 18–20, 2009, 2, International Association of Engineers, Newswood Limited, pp. 2199–2204.

[50] Matthijs Vákár (2021): *Reverse AD at Higher Types: Pure, Principled and Denotationally Correct*. In Nobuko Yoshida, editor: *ESOP 2021: Programming Languages and Systems*, Springer, pp. 607–634, doi:10.1007/978-3-030-72019-3_22.

[51] Fei Wang, Daniel Zheng, James Decker, Xilun Wu, Grégory M. Essertel & Tiark Rompf (2019): *Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator*. *Proceedings of the ACM on Programming Languages* 3(ICFP), pp. 1–31, doi:10.1145/3341700.

[52] R. E. Wengert (1964): *A Simple Automatic Derivative Evaluation Program*. *Communications of the ACM* 7(8), pp. 463–464, doi:10.1145/355586.364791.

A Derivatives

Informally, the *derivative* of a function f at a particular input value x is a mathematical object that describes how infinitesimal changes dx to x incur changes dy to the result $y = f(x)$ of f at x. There are multiple notions of increasing generality and abstraction in mathematics that make "describing", "infinitesimal" and "changes" precise.

A.1 Leibniz derivative

For a scalar function of one (scalar) variable $f : \mathbb{R} \to \mathbb{R}$, the *Leibniz derivative of f at x* is the number $a \in \mathbb{R}$ that satisfies

$$f(x+dx) \approx f(x) + a \cdot dx$$

where $\cdot$ is multiplication on $\mathbb{R}$ and $\approx$ expresses that the error on the right-hand side vanishes as dx becomes infinitesimally small. Specifically, a is the derivative of f at x if

$$\lim_{|dx|\to 0} \frac{|f(x+dx) - (f(x) + a \cdot dx)|}{|dx|} = 0.$$

For example, for $f(x) = x^2$ we have that 8 is the derivative of f at 4, and 14 is the derivative of f at 7.

The *Leibniz derivative of f* is the function f' that maps x to the derivative of f at x. For example, for $f(x) = x^2$ we have $f'(x) = 2 \cdot x$.

A.2 Jacobi derivative

For a vector-valued function $f : \mathbb{R}^{\mathbf{n}} \to \mathbb{R}^{\mathbf{m}}$, the *Jacobi derivative of f at v* is the $m \times n$-matrix M that satisfies

$$f(v+dv) \approx f(v) + M \star dv.$$

Here $\star$ is matrix/vector multiplication and $\approx$ generalizes the case of scalar functions:

$$\lim_{||dv||\to 0} \frac{||f(v+dv) - (f(v) + M \star dv)||}{||dv||} = 0$$

where $||\ldots||$ is the Euclidean norm. We call M_{ij}, the (i,j)-th entry of M, the *partial derivative of the j-th output of f with respect to its i-th input at v*.

For example, for

$$f \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ x_1 \cdot x_3 \end{bmatrix} : \mathbb{R}^3 \to \mathbb{R}^2$$

the matrix $\begin{bmatrix} 1 & 1 & 0 \\ -2 & 0 & 4 \end{bmatrix}$ is the derivative of f at $\begin{bmatrix} 4 \\ 0 \\ -2 \end{bmatrix}$.

The *Jacobi derivative of f* is the function f' that maps a vector v to the Jacobi derivative of f at v. For example, for f as above we have

$$f'(x_1,x_2,x_3) = \begin{bmatrix} 1 & 1 & 0 \\ x_3 & 0 & x_1 \end{bmatrix}.$$

The *partial derivative of the j-th output of f with respect to its i-th input* is the function $\partial f_{ij}(v) = f'(v)_{ij}$. This is usually written $\frac{\partial f_j}{\partial x_i}$ or even $\frac{\partial y_j}{\partial x_i}$.[5]

The Leibniz derivative is the special case of a Jacobi derivative for $m = n = 1$.

A.3 Fréchet derivative

Jacobi derivatives are restricted to functions of the form $f : \mathbb{R}^{\mathbf{n}} \to \mathbb{R}^{\mathbf{m}}$, that is, finite-dimensional Euclidean spaces over the real numbers. Sometimes it is convenient or even necessary to write functions where inputs are not tuples of scalars, but elements of possibly high-dimensional (or even infinite-dimensional) vector spaces.

Example A.1. A layer of a neural network is parameterized by a weight matrix $W \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$ and bias vector $b \in \mathbb{R}^{\mathbf{m}}$ and takes a data vector $v \in \mathbb{R}^{\mathbf{n}}$ as input, where $|\mathbf{m}|, |\mathbf{n}| \gg 0$ may be in the millions or billions. It can be defined in a single line by

$$g(W,b,v) = \mathrm{map}\, h\, (W \star v + b)$$

where $\mathrm{map}\, h$ applies h to each element of a vector, $h : \mathbb{R} \to \mathbb{R}$ is an activation function such as tanh, $\star$ is matrix/vector multiplication and $+$ is vector addition. It can be straightforwardly evaluated using data-parallel implementations of these operations. Trying to write such a function using only scalar variables would be a bad idea for multiple reasons.

Definition A.2 (Fréchet derivative)**.** For a function $f : V \to W$ on Banach spaces V, W, the *linear function* $A \in V \multimap W$ is the *Fréchet derivative of f at v* if it satisfies

$$f(v+dv) \approx f(v) + A(dv),$$

that is

$$\lim_{||dv||_V \to 0} \frac{||f(v+dv) - (f(v) + A(dv))||_W}{||dv||_V} = 0$$

where $||\ldots||_U$ is the norm that comes with the Banach space U.

The *Fréchet derivative of $f : V \to W$* is the partial function $f' : V \to (V \multimap W)$ that maps a vector $v \in V$ to the Fréchet derivative of f at v.

Example A.3. Consider the function

$$g_{W,b}(v) = \mathrm{map}\, h\, (W \star v + b).$$

[5] We avoid this notation since the choi.ce of variable names for inputs and outputs of a function has nothing to do with the notion of derivative.

Its Fréchet derivative is

$$g_{W,b}{}'(v) = \Delta(\mathrm{map}\, h'\,(W \star v + b)) \bullet (W\star)$$

where $\Delta(f_1, \ldots f_m)(x_1, \ldots, x_m) = (f_1(x_1), \ldots, f_m(x_m))$ is zip-apply and $\bullet$ is linear function composition. For $h = \tanh$ we have

$$\tanh'(x) = ((1 - \tanh^2(x))\cdot).$$

Note that $\tanh'(x) : \mathbb{R} \multimap \mathbb{R}$, which explains the use of the section notation $((1-\tanh^2(x))\cdot)$. Since $\mathbb{R} \multimap \mathbb{R}$ is isomorphic to $\mathbb{R}$, the Leibniz derivative is $\tanh'(x) = 1 - \tanh^2(x)$ via implicit application of the isomorphism to its Fréchet derivative.[6] In particular we can rewrite $g_{W,b}{}'$ as

$$\begin{aligned} g_{W,b}{}'(v) &= \Delta(\mathrm{map}\, h'\,(W \star v + b)) \bullet (W\star) \\ &= \Delta(\mathrm{map}\,(\lambda x.(1-\tanh^2(x))\cdot)(W \star v + b)) \bullet (W\star) \\ &= \mathit{zipWith}(\cdot)(\mathrm{map}(\lambda x.(1-\tanh^2(x)))(W \star v + b)) \bullet (W\star) \end{aligned}$$

Note that the right-hand side is built by composing linear functions into a linear function: $\mathit{zipWith}(\cdot)$ is bilinear and thus $\mathit{zipWith}(\cdot)(\mathrm{map}(\lambda x.(1-\tanh^2(x))))$ is linear; likewise, $\star$ is bilinear and thus the section $(W\star)$ defined by $(W\star)(v) = W \star v$ is linear; and functional composition of linear functions by $\bullet$ preserves linearity.

We can η-expand the combinatory expression on the right-hand side into a more familiar looking term representation:

$$g_{W,b}{}'(v)(dv) = \mathit{zipWith}(\cdot)\ (\mathrm{map}(\lambda x.(1-\tanh^2(x)))(W \star v + b))\ (W \star dv)$$

This holds for any dimensions of W and can be computed entirely symbolically, based on general differentiation rules for composition (chain rule), constant, linear and bilinear function, general second-order operators such as map and a dictionary of derivatives for primitive functions such as tanh.

The Fréchet derivative generalizes the Jacobi derivative. If M is the Jacobi derivative of f at x then $A = (M\star)$ is its Fréchet derivative.

A.4 Gateaux derivative

Definition A.4 (Gateaux differential, Gateaux derivative)**.** For a function $f : V \to W$ on Banach spaces V, W over field K (either $\mathbb{R}$ or $\mathbb{C}$), v an interior point of V, $dv \in V$ an *input differential*, the *output differential* $dy \in W$ is the *Gateaux differential of f at v in the direction dv* if

$$dy = \lim_{t \to 0} \frac{||f(v + t \cdot dv) - f(v)||_W}{||t||_K}$$

where $||\ldots||_W$ is the norm that comes with the Banach space W and the underlying field K, respectively.

The *Gateaux derivative of $f : V \to W$* is the partial function $f' : V \times V \to W$ that maps $v, dv \in V$ to the Gateaux differential $f'(v, dv)$.

Function f is *Gateaux differentiable at v* if its Gateaux differential exists at v for all directions dv.

[6] The pleasant compositional nature and applicability of Fréchet derivatives arrives from *not* performing this isomorphism such that the chain rule is always functional composition of linear functions, no matter which vector space the arguments and results of functions belong to.

Gateaux derivatives generalize directional derivatives in multivariate analysis analogous to Fréchet derivatives generalizing total derivatives. They are more general than Fréchet derivatives in the sense that a function may be Gateaux differentiable at v without also being Fréchet differentiable, but if the Fréchet derivative exists then it determines the Gateaux derivative:

$$f'_{Gateaux}(v,dv) = f'_{Fréchet}(v)(dv).$$

There are more general notions of derivatives. Some distinguish the spaces of vectors and differentials, some apply to continuous functions that that are not conventionally differentiable everywhere such as the absolute-value function $f(x) = |x|$. For the purposes of this paper, Fréchet and Gateaux derivatives are sufficient.

A.5 Gateaux versus Fréchet derivatives for automatic differentiation

Gateaux derivatives are conceptually the basis of *forward-mode automatic differentiation*, since they give rise to interpreting a term (or program) representation of a function as operating on *dual numbers/dual tensors* (v,dv):

$$f^{[fad]}_{Gateaux}(v,dv) = (f(v), f'(v,dv)),$$

which preserves functional composition

$$(g \circ f)^{[fad]}_{Gateaux} = g^{[fad]}_{Gateaux} \circ f^{[fad]}_{Gateaux}$$

and is thus easy to implement by replacing the ordinary implementation of numbers and tensors by dual numbers and tensors, respectively. This camouflages, however, that the first component, called the *primal value*, of the output only depends on the primal value of the input and that the second component, called the *tangent value*, always depends linearly on the input tangent value. These universal properties can be partially recovered by partially evaluating $f^{[fad]}$ with static v and dynamic dv, which, by definition of $f^{[fad]}$, always succeeds with statically computing the primal value and leaving a partially evaluated program representing the derivative behind, but rendered in an expressive programming language that does not inherently capture that this is always a *linear* function. For example, in elemental and tensor-based automatic differentiation, the partially evaluated output will result in a data structure corresponding to a *computation graph* [2, 35, 34], also called tape[7] or trace. Reversing its edges amounts to forming the adjoint of the linear function represented by the computation graph.

First employing Gateaux derivatives underlying the dual number/tensor interpretation of a program just to recover Fréchet derivatives by partial evaluation seems like an unnecessary detour. Following Henglein [24] and Elliott [13], we argue that Fréchet derivatives are better suited for both symbolic and automatic differentiation since they capture and reify that the output value of a function only depends on its input value and the output differential depends on both input value and input differential, but always linearly on the input differential:

$$f^{[fad]}_{Fréchet}(v) = (f(v), f'_{Fréchet}(v)) \in W \times (V \multimap W).$$

Since the derivative is always a linear function, it can be represented in a combinatory *domain-specific language (DSL)* that is closed under linear functions and thus syntactically guarantees linearity of all

[7] It is sometimes referred to as a Wengert tape, which we find surprising since Wengert describes only forward-mode AD in his 2-page article [52].

constructed functions. This facilitates not only universal applicability of properties of linear functions, such as $f(x+y) = f(x) + f(y)$ for any f, but also symbolically computing adjoints, which are only defined for linear functions. Linear functions generated during differentiation or adjoint differentiation (generating the adjoint of the derivative during differentiation) can be represented as ordinary functions (λ-abstractions) [13], of course, but this eliminates the possibility of subsequent optimization using linear and tensor algebra.

Executing adjoint derivatives, possibly after algebraic optimization, as ordinary functions (programs) provides computation of "cheap" gradients [19] for scalar functions. We believe that the functional-analysis based approach including tensor products in this paper provides an implementation and data structure framework for not only cheap parallel computation of gradients for scalar functions, but also for cheap adjoint derivatives for non-scalar functions, where tensor decomposition (formal tensor products), tensor contraction and relational reduction have important roles to play.

B Applications

We consider a series of examples of increasing complexity to illustrate how the functions are represented in point-free notation and how their derivative is computed symbolically.

Example B.1. Let $h(x) = \ln(\sin x)$, that is $h = \ln \circ \sin$ in point-free notation. Thus

$$h'(x) = \ln' y \bullet \sin'(x)$$

where $y = \sin x$ by Rule 1, the chain rule. Since $\ln' y = (\cdot \frac{1}{y})$ and $\sin' x = (\cdot \cos x)$ we get

$$\begin{aligned} h'(x) &= (\ln \circ \sin)'(x) \\ &= \ln'(\sin x) \bullet \sin'(x) \\ &= (\frac{1}{\sin x} \cdot) \bullet (\cos x \cdot) \\ &= (\frac{1}{\sin x} \cdot \cos x \cdot) \\ &= (\frac{\cos x}{\sin x} \cdot) \end{aligned}$$

Example B.2. Consider $y = f(x_1, x_2) = \ln(x_1) + x_1 \cdot x_2 - sin(x_2)$ [3, p.9]. It can be written in point-free form as

$$f = \ln \circ \pi_1 + \pi_1 \hat{} \pi_2 - \sin \circ \pi_2$$

where $\hat{}$ is $\cdot$ lifted to functions: $(f\hat{}g)(x) = f(x)\cdot g(x)$. Employing our rules of differentation we obtain

$$\begin{aligned}
f'(x_1,x_2) &= (\ln\circ\pi_1 + \pi_1\hat{}\pi_2 - \sin\circ\pi_2)'(x_1,x_2)\\
&= (\ln\circ\pi_1)'(x_1,x_2) + (\pi_1\hat{}\pi_2)'(x_1,x_2) - (\sin\circ\pi_2)'(x_1,x_2)\\
&= \ln'(\pi_1(x_1,x_2))\bullet\pi_1{}'(x_1,x_2) +\\
&\quad (\pi_1(x_1,x_2))\cdot)\bullet\pi_2{}'(x_1,x_2)) + (\cdot\,\pi_2(x_1,x_2))\bullet\pi_1{}'(x_1,x_2) -\\
&\quad \sin'(\pi_2(x_1,x_2))\bullet\pi_2{}'(x_1,x_2)\\
&= \ln'(x_1)\bullet\pi_1 +\\
&\quad (x_1\,\cdot)\bullet\pi_2 + (\cdot\,x_2)\bullet\pi_1 -\\
&\quad \sin'(x_2)\bullet\pi_2\\
&= (\frac{1}{x_1}\cdot)\bullet\pi_1 +\\
&\quad (x_1\,\cdot)\bullet\pi_2 + (\cdot\,x_2)\bullet\pi_1 -\\
&\quad (\cos x_2\,\cdot)\bullet\pi_2
\end{aligned}$$

The last line is easily transformed into a familiar looking expressing by recognizing that $\partial x_1 = (\bullet\pi_1)$ and $\partial x_2 = (\bullet\pi_2)$:

$$\begin{aligned}
f'(x_1,x_2)(\partial x_1,\partial x_2) &= (\frac{1}{x_1}\cdot)(\partial x_1) + (x_1\,\cdot)(\partial x_2) + (\cdot\,x_2)(\partial x_1) - (\cos x_2\,\cdot)(\partial x_2)\\
&= \frac{1}{x_1}\cdot\partial x_1 + x_1\cdot\partial x_2 + \partial x_1\cdot x_2 - \cos x_2\cdot\partial x_2
\end{aligned}$$

In Leibniz notation, we can get the partial derivative $\frac{f'(x_1,x_2)}{\partial x_1}$ by setting $\partial x_2 = 0$ and formally dividing the righthand side by ∂x_1. Aanalogously we can compute $\frac{f'(x_1,x_2)}{\partial x_2}$. Finally, writing ∂f instead of f' we get the familiar looking:

$$\begin{aligned}
\frac{\partial f(x_1,x_2)}{\partial x_1} &= \frac{1}{x_1} + x_2\\
\frac{\partial f(x_1,x_2)}{\partial x_2} &= x_1 - \cos x_2
\end{aligned}$$

The derivative for the unreadable version of f is as follows:

$$\begin{aligned}
f'(x_1,x_2) &= ((+)\circ((+)\times((-)\circ\sin))\circ\langle\langle\ln\circ\pi_1,(\cdot)\rangle,\pi_2\rangle)'(x_1,x_2)\\
&= (\llcorner+'\lrcorner)\circ((\llcorner+'\lrcorner)\times(-'\lrcorner\circ\sin' x_2))\circ\langle\langle\ln' x_1\circ\pi_1{}'\lrcorner, x_1\cdot' x_2\rangle,\pi_2{}'\lrcorner\rangle\\
&= (+)\circ((+)\times((-)\circ(\cos x_2\cdot)))\circ\langle\langle(\frac{1}{x_1}\cdot)\circ\pi_1,(+)\circ((x_2\cdot)\times(x_1\cdot))\rangle,\pi_2\rangle
\end{aligned}$$

Since the right-hand side in this example is constructed from linear functions and combinators that preserve linearity we can see that it is linear, as it should be. (The derivative at any point is linear by definition.) More specifically, the code for the derivative is *parametric* in x_1,x_2: it is the same for each point $(x_1,x_2)\in\mathbb{R}\times\mathbb{R}$. It is not a definitional requirement that the derivative at one point have the same expression as the derivative at another point, but when it does it is useful in practice: If we need to compute the derivative at many different points and the derivative is described by the same program at each point, it makes sense to optimize that program prior to executing it. What makes this extra intriguing is that derivatives, always denoting linear functions, are essentially first-order data with strong algebraic properties admitting powerful optimizations.