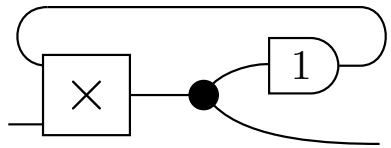
The differential calculus of causal functions

Bart Jacobs and David Sprunger*

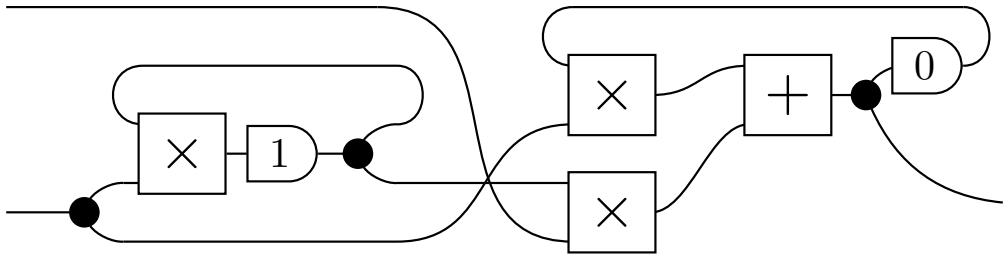
G0 seminar
April 17, 2019

Where we're heading

In February, I laid out a case for why the derivative of this function:



Is this:



Today, I will give a more direct route to this causal derivative, inspired by traditional calculus and using less categorical machinery.

How we get there

1. Causal functions
2. Definition of causal derivatives
3. Rules of causal derivatives
4. Example application: Elman networks

Causal functions on sequences

A^ω is the set of *A-valued infinite sequences*. The entries of $\sigma \in A^\omega$ are $\sigma_k \in A$ for $k \in \mathbb{N}$, so $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_k, \ldots)$.

Causal functions on sequences

A^ω is the set of *A-valued infinite sequences*. The entries of $\sigma \in A^\omega$ are $\sigma_k \in A$ for $k \in \mathbb{N}$, so $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_k, \ldots)$.

Slicing extracts a finite list from an infinite sequence:

$$(\cdot)_{j:k} : \sigma \mapsto (\sigma_j, \sigma_{j+1}, \ldots, \sigma_k)$$

(We also sometimes use slicing on finite lists.)

Causal functions on sequences

A^ω is the set of *A-valued infinite sequences*. The entries of $\sigma \in A^\omega$ are $\sigma_k \in A$ for $k \in \mathbb{N}$, so $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_k, \ldots)$.

Slicing extracts a finite list from an infinite sequence:

$$(\cdot)_{j:k} : \sigma \mapsto (\sigma_j, \sigma_{j+1}, \ldots, \sigma_k)$$

(We also sometimes use slicing on finite lists.)

Definition

A function on sequences $f : A^\omega \to B^\omega$ is *causal* if it satisfies $\sigma_{0:k} = \tau_{0:k} \to f(\sigma)_{0:k} = f(\tau)_{0:k}$ for all input sequences $\sigma, \tau \in A^\omega$ and $k \in \mathbb{N}$.

Intuitively, the first k outputs of f only depend on the first k inputs.

Finite approximants

Lemma

The following are equivalent:

1. *a causal function* $f : A^\omega \to B^\omega$,
2. *a sequence of functions* $u_k : A^{k+1} \to B$, *and*
3. *a sequence of functions* $t_k : A^{k+1} \to B^{k+1}$ *satisfying* $t_k(x_{0:k}) = [t_{k+1}(x)]_{0:k}$ *for all* $x \in A^{k+2}$.

Proof.

$(1 \Rightarrow 2)$ The *pointwise approximation* of f is the sequence $U_k(f)(x) \triangleq f(x : \sigma)_k$ for $x \in A^{k+1}$. $(1 \Rightarrow 3)$ The *stringwise approximation* of f is the sequence $T_k(f)(x) \triangleq f(x : \sigma)_{0:k}$ for $x \in A^{k+1}$.

$(2 \Rightarrow 1)$ $f : \sigma \mapsto \tau$ iff $\tau_k = u_k(\sigma_{0:k})$. $(3 \Rightarrow 2)$ $u_k \triangleq \pi_{k+1} \circ t_k$. □

Defining a causal function pointwise

Suppose $(A, +_A, \cdot_A, 0_A)$ is a vector space (over $\mathbb{R}$). Then A^ω is also an $\mathbb{R}$-vector space using the following:

1. Define $+_{A^\omega} : A^\omega \times A^\omega \to A^\omega$ pointwise by $U_k(+_{A^\omega})((\sigma_0, \tau_0), (\sigma_1, \tau_1), \ldots, (\sigma_k, \tau_k)) = \sigma_k +_A \tau_k$.
2. For each $r \in \mathbb{R}$, define $r \cdot_{A^\omega} (-) : A^\omega \to A^\omega$ pointwise by $U_k(r \cdot_{A^\omega} (-))(\sigma_0, \sigma_1, \ldots, \sigma_k) = r \cdot_A \sigma_k$.
3. The zero sequence is 0_A in each position.

Defining a causal function pointwise

Suppose $(A, +_A, \cdot_A, 0_A)$ is a vector space (over $\mathbb{R}$). Then A^ω is also an $\mathbb{R}$-vector space using the following:

1. Define $+_{A^\omega} : A^\omega \times A^\omega \to A^\omega$ pointwise by $U_k(+_{A^\omega})((\sigma_0, \tau_0), (\sigma_1, \tau_1), \ldots, (\sigma_k, \tau_k)) = \sigma_k +_A \tau_k$.
2. For each $r \in \mathbb{R}$, define $r \cdot_{A^\omega} (-) : A^\omega \to A^\omega$ pointwise by $U_k(r \cdot_{A^\omega} (-))(\sigma_0, \sigma_1, \ldots, \sigma_k) = r \cdot_A \sigma_k$.
3. The zero sequence is 0_A in each position.

Note that $+_{A^\omega}$ defined above is really $+_{A^\omega} : (A \times A)^\omega \to A^\omega$. We will use the isomorphisms like the one between $(A \times A)^\omega$ and $A^\omega \times A^\omega$ without pointing it out in the future.

Stringwise approximations are also useful

Lemma

The composition of two causal functions $f : A^\omega \to B^\omega$ *and* $g : B^\omega \to C^\omega$ *is another causal function* $g \circ f : A^\omega \to C^\omega$*. Their composite is also the unique causal function satisfying* $T_k(g \circ f) = T_k(g) \circ T_k(f)$.

Stringwise approximations are also useful

Lemma

The composition of two causal functions $f : A^\omega \to B^\omega$ and $g : B^\omega \to C^\omega$ is another causal function $g \circ f : A^\omega \to C^\omega$. Their composite is also the unique causal function satisfying $T_k(g \circ f) = T_k(g) \circ T_k(f)$.

Characterizing the composition of causal functions using only pointwise approximants is harder—they aren't composable on the nose.

Outline

Our goal: derivatives of causal functions

As a reminder, $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \mathbb{R}^n$ means there is a linear map $Jf(x) : \mathbb{R}^n \to \mathbb{R}^m$ such that $f(x + \Delta x) \approx f(x) + Jf(x)(\Delta x)$.

Our goal: derivatives of causal functions

As a reminder, $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \mathbb{R}^n$ means there is a linear map $Jf(x) : \mathbb{R}^n \to \mathbb{R}^m$ such that $f(x + \Delta x) \approx f(x) + Jf(x)(\Delta x)$. One way $\approx$ is formalized is

$$\lim_{\Delta x \to 0} \frac{\|f(x + \Delta x) - f(x) - Jf(x)(\Delta x)\|}{\|\Delta x\|} = 0$$

Our goal: derivatives of causal functions

As a reminder, $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \mathbb{R}^n$ means there is a linear map $Jf(x) : \mathbb{R}^n \to \mathbb{R}^m$ such that $f(x + \Delta x) \approx f(x) + Jf(x)(\Delta x)$. One way $\approx$ is formalized is

$$\lim_{\Delta x \to 0} \frac{\|f(x + \Delta x) - f(x) - Jf(x)(\Delta x)\|}{\|\Delta x\|} = 0$$

Linear maps $Jf(x) : \mathbb{R} \to \mathbb{R}$ are 1-1 with real numbers.

Our goal: derivatives of causal functions

As a reminder, $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \mathbb{R}^n$ means there is a linear map $Jf(x) : \mathbb{R}^n \to \mathbb{R}^m$ such that $f(x + \Delta x) \approx f(x) + Jf(x)(\Delta x)$. One way $\approx$ is formalized is

$$\lim_{\Delta x \to 0} \frac{\|f(x + \Delta x) - f(x) - Jf(x)(\Delta x)\|}{\|\Delta x\|} = 0$$

Linear maps $Jf(x) : \mathbb{R} \to \mathbb{R}$ are 1-1 with real numbers. Linear maps $Jf(x) : \mathbb{R}^n \to \mathbb{R}^m$ are (Jacobian) matrices:

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \dots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \dots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}$$

Our goal: derivatives of causal functions

As a reminder, $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \mathbb{R}^n$ means there is a linear map $Jf(x) : \mathbb{R}^n \to \mathbb{R}^m$ such that $f(x + \Delta x) \approx f(x) + Jf(x)(\Delta x)$. One way $\approx$ is formalized is

$$\lim_{\Delta x \to 0} \frac{\|f(x + \Delta x) - f(x) - Jf(x)(\Delta x)\|}{\|\Delta x\|} = 0$$

Linear maps $Jf(x) : \mathbb{R} \to \mathbb{R}$ are 1-1 with real numbers. Linear maps $Jf(x) : \mathbb{R}^n \to \mathbb{R}^m$ are (Jacobian) matrices:

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \dots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \dots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}$$

What is $J+(x, y)$ for $+ : \mathbb{R}^2 \to \mathbb{R}$?

Our goal: derivatives of causal functions

As a reminder, $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \mathbb{R}^n$ means there is a linear map $Jf(x) : \mathbb{R}^n \to \mathbb{R}^m$ such that $f(x + \Delta x) \approx f(x) + Jf(x)(\Delta x)$. One way $\approx$ is formalized is

$$\lim_{\Delta x \to 0} \frac{\|f(x + \Delta x) - f(x) - Jf(x)(\Delta x)\|}{\|\Delta x\|} = 0$$

Linear maps $Jf(x) : \mathbb{R} \to \mathbb{R}$ are 1-1 with real numbers. Linear maps $Jf(x) : \mathbb{R}^n \to \mathbb{R}^m$ are (Jacobian) matrices:

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \dots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \dots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}$$

What is $J+(x, y)$ for $+ : \mathbb{R}^2 \to \mathbb{R}$? What about $J(\cdot)(x, y)$?

Linear causal functions

The derivative of a causal function at a sequence will be an appropriate linear map. We have already described a vector space structure on $(\mathbb{R}^n)^\omega$—this gives the appropriate notion of linear.

Definition

A function on sequences $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ is a *linear causal map* if it is (1) causal and (2) linear with respect to the natural vector space structure on $(\mathbb{R}^n)^\omega$.

Linear causal functions

The derivative of a causal function at a sequence will be an appropriate linear map. We have already described a vector space structure on $(\mathbb{R}^n)^\omega$—this gives the appropriate notion of linear.

Definition

A function on sequences $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ is a *linear causal map* if it is (1) causal and (2) linear with respect to the natural vector space structure on $(\mathbb{R}^n)^\omega$.

Examples:

1. $\mathsf{dup}_{(\mathbb{R}^n)^\omega} : (\mathbb{R}^n)^\omega \to (\mathbb{R}^n)^\omega \times (\mathbb{R}^n)^\omega$ given by $U_k(\mathsf{dup}_{(\mathbb{R}^n)^\omega})(\sigma_{0:k}) = \langle \sigma_k, \sigma_k \rangle$.
2. $+_{(\mathbb{R}^n)^\omega} : (\mathbb{R}^n)^\omega \times (\mathbb{R}^n)^\omega \to (\mathbb{R}^n)^\omega$ given by $U_k(+_{(\mathbb{R}^n)^\omega})(\langle \sigma, \tau \rangle_{0:k}) = \sigma_k + \tau_k$.
3. $0 : (\mathbb{R}^n)^\omega \to (\mathbb{R}^n)^\omega$ given by $U_k(0)(\sigma_{0:k}) = 0$.

Linear causal functions, continued

Lemma

Let $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *be a causal function. TFAE:*

1. f *is linear,*
2. $U_k(f) : (\mathbb{R}^n)^{k+1} \to \mathbb{R}^m$ *is linear for all* $k \in \mathbb{N}$*, and*
3. $T_k(f) : (\mathbb{R}^n)^{k+1} \to (\mathbb{R}^m)^{k+1}$ *is linear for all* $k \in \mathbb{N}$*.*

Linear causal functions, continued

Lemma

Let $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ be a causal function. TFAE:

1. *f is linear,*
2. *$U_k(f) : (\mathbb{R}^n)^{k+1} \to \mathbb{R}^m$ is linear for all $k \in \mathbb{N}$, and*
3. *$T_k(f) : (\mathbb{R}^n)^{k+1} \to (\mathbb{R}^m)^{k+1}$ is linear for all $k \in \mathbb{N}$.*

This lets us define linear causal functions by giving linear approximants. That's the trick we need to define derivatives.

Derivatives of causal functions

Definition

A causal function $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ is *differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ if and only if $U_k(f) : (\mathbb{R}^n)^{k+1} \to \mathbb{R}^m$ is differentiable at $\sigma_{0:k}$ for all $k \in \mathbb{N}$.

If f is differentiable at σ, the *derivative of* f *at* σ is the linear causal function $\mathcal{D}^* f(\sigma) : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ satisfying $U_k(\mathcal{D}^* f(\sigma)) = J(U_k(f))(\sigma_{0:k})$.

Derivatives of causal functions

Definition

A causal function $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ is *differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ if and only if $U_k(f) : (\mathbb{R}^n)^{k+1} \to \mathbb{R}^m$ is differentiable at $\sigma_{0:k}$ for all $k \in \mathbb{N}$.

If f is differentiable at σ, the *derivative of* f *at* σ is the linear causal function $\mathcal{D}^* f(\sigma) : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ satisfying $U_k(\mathcal{D}^* f(\sigma)) = J(U_k(f))(\sigma_{0:k})$.

You could equally well use stringwise approximants in the above definition. In that case, $T_k(\mathcal{D}^* f(\sigma)) = J(T_k(f))(\sigma_{0:k})$.

Example I: sequence sum

The causal function $+ : (\mathbb{R}^2)^\omega \to \mathbb{R}^\omega$ is its own derivative at every point, meaning $\mathcal{D}^*+(\sigma, \tau) = +$. By definition, $U_k(+)(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \sigma_k + \tau_k$.

Example I: sequence sum

The causal function $+ : (\mathbb{R}^2)^\omega \to \mathbb{R}^\omega$ is its own derivative at every point, meaning $\mathcal{D}^*+(\sigma, \tau) = +$. By definition, $U_k(+)(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \sigma_k + \tau_k$. Then $J(U_k(+))(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \begin{bmatrix} 0 & \ldots & 0 & 1 & 1 \end{bmatrix}$,

Example I: sequence sum

The causal function $+ : (\mathbb{R}^2)^\omega \to \mathbb{R}^\omega$ is its own derivative at every point, meaning $\mathcal{D}^*+(\sigma, \tau) = +$. By definition, $U_k(+)(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \sigma_k + \tau_k$. Then $J(U_k(+))(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \begin{bmatrix} 0 & \ldots & 0 & 1 & 1 \end{bmatrix}$, so $J(U_k(+))(\langle \sigma, \tau \rangle_{0:k})(\langle \Delta\sigma, \Delta\tau \rangle_{0:k}) = \Delta\sigma_k + \Delta\tau_k$.

Example I: sequence sum

The causal function $+ : (\mathbb{R}^2)^\omega \to \mathbb{R}^\omega$ is its own derivative at every point, meaning $\mathcal{D}^*+(\sigma, \tau) = +$. By definition, $U_k(+)(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \sigma_k + \tau_k$. Then $J(U_k(+))(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \begin{bmatrix} 0 & \ldots & 0 & 1 & 1 \end{bmatrix}$, so $J(U_k(+))(\langle \sigma, \tau \rangle_{0:k})(\langle \Delta\sigma, \Delta\tau \rangle_{0:k}) = \Delta\sigma_k + \Delta\tau_k$. This means $\mathcal{D}^*+(\sigma, \tau)(\Delta\sigma, \Delta\tau) = \Delta\sigma + \Delta\tau$.

Example I: sequence sum

The causal function $+ : (\mathbb{R}^2)^\omega \to \mathbb{R}^\omega$ is its own derivative at every point, meaning $\mathcal{D}^*+(\sigma, \tau) = +$. By definition, $U_k(+)(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \sigma_k + \tau_k$. Then $J(U_k(+))(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \begin{bmatrix} 0 & \ldots & 0 & 1 & 1 \end{bmatrix}$, so $J(U_k(+))(\langle \sigma, \tau \rangle_{0:k})(\langle \Delta\sigma, \Delta\tau \rangle_{0:k}) = \Delta\sigma_k + \Delta\tau_k$. This means $\mathcal{D}^*+(\sigma, \tau)(\Delta\sigma, \Delta\tau) = \Delta\sigma + \Delta\tau$.

An intuition which can also be useful is to think about $f(x + \Delta x) - f(x)$. In this case, f is sequence sum:

$$[(\sigma + \Delta\sigma) + (\tau + \Delta\tau)] - [\sigma + \tau] = \Delta\sigma + \Delta\tau$$

Example II: Cauchy product

The Cauchy product of sequences $\times : (\mathbb{R}^2)^\omega \to \mathbb{R}^\omega$ is defined by $U_k(\times)(\sigma_{0:k}, \tau_{0:k}) = \sum_{i=0}^{k} \sigma_i \cdot \tau_{k-i}$. Writing out the first few terms,

$$\begin{aligned} \sigma \times \tau = (&\sigma_0\tau_0, \\ &\sigma_0\tau_1 + \sigma_1\tau_0, \\ &\sigma_0\tau_2 + \sigma_1\tau_1 + \sigma_2\tau_0, \ldots) \end{aligned}$$

Example II: Cauchy product

The Cauchy product of sequences $\times : (\mathbb{R}^2)^\omega \to \mathbb{R}^\omega$ is defined by $U_k(\times)(\sigma_{0:k}, \tau_{0:k}) = \sum_{i=0}^{k} \sigma_i \cdot \tau_{k-i}$. Writing out the first few terms,

$$\begin{aligned} \sigma \times \tau = (&\sigma_0\tau_0, \\ &\sigma_0\tau_1 + \sigma_1\tau_0, \\ &\sigma_0\tau_2 + \sigma_1\tau_1 + \sigma_2\tau_0, \ldots) \end{aligned}$$

Now we find its causal derivative at $(\sigma, \tau) \in (\mathbb{R}^2)^\omega$.

$$J(U_k(\times))(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \begin{bmatrix} \tau_k & \sigma_k & \tau_{k-1} & \sigma_{k-1} & \cdots & \tau_0 & \sigma_0 \end{bmatrix}$$

Example II: Cauchy product

The Cauchy product of sequences $\times : (\mathbb{R}^2)^\omega \to \mathbb{R}^\omega$ is defined by $U_k(\times)(\sigma_{0:k}, \tau_{0:k}) = \sum_{i=0}^{k} \sigma_i \cdot \tau_{k-i}$. Writing out the first few terms,

$$\begin{aligned} \sigma \times \tau = (&\sigma_0\tau_0, \\ &\sigma_0\tau_1 + \sigma_1\tau_0, \\ &\sigma_0\tau_2 + \sigma_1\tau_1 + \sigma_2\tau_0, \ldots) \end{aligned}$$

Now we find its causal derivative at $(\sigma, \tau) \in (\mathbb{R}^2)^\omega$.

$$J(U_k(\times))(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \begin{bmatrix} \tau_k & \sigma_k & \tau_{k-1} & \sigma_{k-1} & \cdots & \tau_0 & \sigma_0 \end{bmatrix}$$

$$J(U_k(\times))(\langle \sigma, \tau \rangle_{0:k})(\langle \Delta\sigma, \Delta\tau \rangle_{0:k}) = \sum_{i=0}^{k} \Delta\sigma_i \cdot \tau_{k-i} + \sum_{i=0}^{k} \sigma_i \cdot \Delta\tau_{k-i}$$

Example II: Cauchy product

The Cauchy product of sequences $\times : (\mathbb{R}^2)^\omega \to \mathbb{R}^\omega$ is defined by $U_k(\times)(\sigma_{0:k}, \tau_{0:k}) = \sum_{i=0}^{k} \sigma_i \cdot \tau_{k-i}$. Writing out the first few terms,

$$\begin{aligned}\sigma \times \tau = (&\sigma_0\tau_0,\\ &\sigma_0\tau_1 + \sigma_1\tau_0,\\ &\sigma_0\tau_2 + \sigma_1\tau_1 + \sigma_2\tau_0, \ldots)\end{aligned}$$

Now we find its causal derivative at $(\sigma, \tau) \in (\mathbb{R}^2)^\omega$.

$$J(U_k(\times))(\sigma_0, \tau_0, \ldots, \sigma_k, \tau_k) = \begin{bmatrix}\tau_k & \sigma_k & \tau_{k-1} & \sigma_{k-1} & \cdots & \tau_0 & \sigma_0\end{bmatrix}$$

$$J(U_k(\times))(\langle\sigma, \tau\rangle_{0:k})(\langle\Delta\sigma, \Delta\tau\rangle_{0:k}) = \sum_{i=0}^{k} \Delta\sigma_i \cdot \tau_{k-i} + \sum_{i=0}^{k} \sigma_i \cdot \Delta\tau_{k-i}$$

$$\mathcal{D}^*\times(\sigma, \tau)(\Delta\sigma, \Delta\tau) = \Delta\sigma \times \tau + \sigma \times \Delta\tau.$$

Remark on definition

The uniqueness of the usual derivative is ensured by the norm condition:

$$\lim_{\Delta x \to 0} \frac{\|f(x + \Delta x) - f(x) - Jf(x)(\Delta x)\|}{\|\Delta x\|} = 0$$

But our causal derivative doesn't use a norm on sequence spaces. It does still have a canonical property though: it gives the unique linear causal function whose approximants are the derivatives of the approximants of the original function.

Remark on definition

The uniqueness of the usual derivative is ensured by the norm condition:

$$\lim_{\Delta x \to 0} \frac{\|f(x + \Delta x) - f(x) - Jf(x)(\Delta x)\|}{\|\Delta x\|} = 0$$

But our causal derivative doesn't use a norm on sequence spaces. It does still have a canonical property though: it gives the unique linear causal function whose approximants are the derivatives of the approximants of the original function.

It's also worth noting that our derivative probably cannot be realized by a Fréchet derivative (the above definition, possibly in infinite dimensions), since it does not need a norm.

Outline

Overview of rules

There are three core rules to our differential calculus:

- the causal chain rule,
- the causal parallel rule, and
- the causal linear rule.

Overview of rules

There are three core rules to our differential calculus:

- the causal chain rule,
- the causal parallel rule, and
- the causal linear rule.

With these three rules, we can derive many other standard-looking rules as consequences, including:

- the causal sum rule,
- the causal product rule, and
- the causal reciprocal rule.

Overview of rules

There are three core rules to our differential calculus:

- the causal chain rule,
- the causal parallel rule, and
- the causal linear rule.

With these three rules, we can derive many other standard-looking rules as consequences, including:

- the causal sum rule,
- the causal product rule, and
- the causal reciprocal rule.

Later, we will cover a special rule with no analogue in ordinary calculus: the recurrence rule.

Causal chain rule

Theorem (causal chain rule)

Suppose $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *and* $g : (\mathbb{R}^m)^\omega \to (\mathbb{R}^\ell)^\omega$ *are causal differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ *and* $f(\sigma)$*, respectively. Then* $h = g \circ f$ *is causal differentiable at* σ *and* $\mathcal{D}^* g(f(\sigma)) \circ \mathcal{D}^* f(\sigma)$*.*

Proof.

Let $f_k = T_k(f)$, $g_k = T_k(g)$, and $h_k = T_k(h)$.

$$T_k(\mathcal{D}^*(g \circ f)(\sigma)) = Jh_k(\sigma_{0:k}) = J(g_k \circ f_k)(\sigma_{0:k})$$

Causal chain rule

Theorem (causal chain rule)

Suppose $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *and* $g : (\mathbb{R}^m)^\omega \to (\mathbb{R}^\ell)^\omega$ *are causal differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ *and* $f(\sigma)$*, respectively. Then* $h = g \circ f$ *is causal differentiable at* σ *and* $\mathcal{D}^* g(f(\sigma)) \circ \mathcal{D}^* f(\sigma)$*.*

Proof.

Let $f_k = T_k(f)$, $g_k = T_k(g)$, and $h_k = T_k(h)$.

$$\begin{aligned} T_k(\mathcal{D}^*(g \circ f)(\sigma)) &= Jh_k(\sigma_{0:k}) = J(g_k \circ f_k)(\sigma_{0:k}) \\ &= Jg_k(f_k(\sigma_{0:k})) \circ Jf_k(\sigma_{0:k}) \qquad (*) \end{aligned}$$

Causal chain rule

Theorem (causal chain rule)

Suppose $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *and* $g : (\mathbb{R}^m)^\omega \to (\mathbb{R}^\ell)^\omega$ *are causal differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ *and* $f(\sigma)$*, respectively. Then* $h = g \circ f$ *is causal differentiable at* σ *and* $\mathcal{D}^* g(f(\sigma)) \circ \mathcal{D}^* f(\sigma)$*.*

Proof.

Let $f_k = T_k(f)$, $g_k = T_k(g)$, and $h_k = T_k(h)$.

$$\begin{aligned} T_k(\mathcal{D}^*(g \circ f)(\sigma)) &= Jh_k(\sigma_{0:k}) = J(g_k \circ f_k)(\sigma_{0:k}) \\ &= Jg_k(f_k(\sigma_{0:k})) \circ Jf_k(\sigma_{0:k}) \qquad (*) \\ &= Jg_k(f(\sigma)_{0:k}) \circ Jf_k(\sigma_{0:k}) \end{aligned}$$

Causal chain rule

Theorem (causal chain rule)

Suppose $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *and* $g : (\mathbb{R}^m)^\omega \to (\mathbb{R}^\ell)^\omega$ *are causal differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ *and* $f(\sigma)$*, respectively. Then* $h = g \circ f$ *is causal differentiable at* σ *and* $\mathcal{D}^* g(f(\sigma)) \circ \mathcal{D}^* f(\sigma)$*.*

Proof.

Let $f_k = T_k(f)$, $g_k = T_k(g)$, and $h_k = T_k(h)$.

$$\begin{aligned} T_k(\mathcal{D}^*(g \circ f)(\sigma)) &= Jh_k(\sigma_{0:k}) = J(g_k \circ f_k)(\sigma_{0:k}) \\ &= Jg_k(f_k(\sigma_{0:k})) \circ Jf_k(\sigma_{0:k}) \qquad (*) \\ &= Jg_k(f(\sigma)_{0:k}) \circ Jf_k(\sigma_{0:k}) \\ &= T_k(\mathcal{D}^* g(f(\sigma))) \circ T_k(\mathcal{D}^* f(\sigma)) \end{aligned}$$

Causal chain rule

Theorem (causal chain rule)

Suppose $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *and* $g : (\mathbb{R}^m)^\omega \to (\mathbb{R}^\ell)^\omega$ *are causal differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ *and* $f(\sigma)$*, respectively. Then* $h = g \circ f$ *is causal differentiable at* σ *and* $\mathcal{D}^* g(f(\sigma)) \circ \mathcal{D}^* f(\sigma)$*.*

Proof.

Let $f_k = T_k(f)$, $g_k = T_k(g)$, and $h_k = T_k(h)$.

$$\begin{aligned} T_k(\mathcal{D}^*(g \circ f)(\sigma)) &= Jh_k(\sigma_{0:k}) = J(g_k \circ f_k)(\sigma_{0:k}) \\ &= Jg_k(f_k(\sigma_{0:k})) \circ Jf_k(\sigma_{0:k}) \qquad (*) \\ &= Jg_k(f(\sigma)_{0:k}) \circ Jf_k(\sigma_{0:k}) \\ &= T_k(\mathcal{D}^* g(f(\sigma))) \circ T_k(\mathcal{D}^* f(\sigma)) \\ &= T_k(\mathcal{D}^* g(f(\sigma)) \circ \mathcal{D}^* f(\sigma)) \end{aligned}$$

Causal parallel rule

Theorem (causal parallel rule)

Suppose $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *and* $h : (\mathbb{R}^p)^\omega \to (\mathbb{R}^q)^\omega$ *are causal differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ *and* $\tau \in (\mathbb{R}^p)^\omega$*, respectively. Then* $f \| h : (\mathbb{R}^{n+p})^\omega \to (\mathbb{R}^{m+q})^\omega$ *is differentiable at* $(\sigma, \tau) \in (\mathbb{R}^{n+p})^\omega$ *and its derivative is* $\mathcal{D}^* f(\sigma) \| \mathcal{D}^* h(\tau)$.

Proof.

$$T_k(\mathcal{D}^*(f \| h)(\sigma, \tau)) = J(T_k(f \| h))(\sigma_{0:k}, \tau_{0:k})$$

Causal parallel rule

Theorem (causal parallel rule)

Suppose $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *and* $h : (\mathbb{R}^p)^\omega \to (\mathbb{R}^q)^\omega$ *are causal differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ *and* $\tau \in (\mathbb{R}^p)^\omega$*, respectively. Then* $f \| h : (\mathbb{R}^{n+p})^\omega \to (\mathbb{R}^{m+q})^\omega$ *is differentiable at* $(\sigma, \tau) \in (\mathbb{R}^{n+p})^\omega$ *and its derivative is* $\mathcal{D}^* f(\sigma) \| \mathcal{D}^* h(\tau)$.

Proof.

$$\begin{aligned} T_k(\mathcal{D}^*(f\|h)(\sigma,\tau)) &= J(T_k(f\|h))(\sigma_{0:k}, \tau_{0:k}) \\ &= J(T_k(f)\|T_k(h))(\sigma_{0:k}, \tau_{0:k}) \end{aligned}$$

Causal parallel rule

Theorem (causal parallel rule)

Suppose $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *and* $h : (\mathbb{R}^p)^\omega \to (\mathbb{R}^q)^\omega$ *are causal differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ *and* $\tau \in (\mathbb{R}^p)^\omega$*, respectively. Then* $f \| h : (\mathbb{R}^{n+p})^\omega \to (\mathbb{R}^{m+q})^\omega$ *is differentiable at* $(\sigma, \tau) \in (\mathbb{R}^{n+p})^\omega$ *and its derivative is* $\mathcal{D}^* f(\sigma) \| \mathcal{D}^* h(\tau)$.

Proof.

$$\begin{aligned} T_k(\mathcal{D}^*(f\|h)(\sigma,\tau)) &= J(T_k(f\|h))(\sigma_{0:k},\tau_{0:k}) \\ &= J(T_k(f)\|T_k(h))(\sigma_{0:k},\tau_{0:k}) \\ &= J(T_k(f))(\sigma_{0:k})\|J(T_k(h))(\tau_{0:k}) \qquad (*) \end{aligned}$$

□

Causal parallel rule

Theorem (causal parallel rule)

Suppose $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *and* $h : (\mathbb{R}^p)^\omega \to (\mathbb{R}^q)^\omega$ *are causal differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ *and* $\tau \in (\mathbb{R}^p)^\omega$*, respectively. Then* $f \| h : (\mathbb{R}^{n+p})^\omega \to (\mathbb{R}^{m+q})^\omega$ *is differentiable at* $(\sigma, \tau) \in (\mathbb{R}^{n+p})^\omega$ *and its derivative is* $\mathcal{D}^* f(\sigma) \| \mathcal{D}^* h(\tau)$.

Proof.

$$
\begin{aligned}
T_k(\mathcal{D}^*(f\|h)(\sigma,\tau)) &= J(T_k(f\|h))(\sigma_{0:k},\tau_{0:k}) \\
&= J(T_k(f)\|T_k(h))(\sigma_{0:k},\tau_{0:k}) \\
&= J(T_k(f))(\sigma_{0:k})\|J(T_k(h))(\tau_{0:k}) \qquad (*) \\
&= T_k(\mathcal{D}^* f(\sigma))\|T_k(\mathcal{D}^* h(\tau))
\end{aligned}
$$

□

Causal parallel rule

Theorem (causal parallel rule)

Suppose $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *and* $h : (\mathbb{R}^p)^\omega \to (\mathbb{R}^q)^\omega$ *are causal differentiable at* $\sigma \in (\mathbb{R}^n)^\omega$ *and* $\tau \in (\mathbb{R}^p)^\omega$*, respectively. Then* $f \| h : (\mathbb{R}^{n+p})^\omega \to (\mathbb{R}^{m+q})^\omega$ *is differentiable at* $(\sigma, \tau) \in (\mathbb{R}^{n+p})^\omega$ *and its derivative is* $\mathcal{D}^* f(\sigma) \| \mathcal{D}^* h(\tau)$.

Proof.

$$
\begin{aligned}
T_k(\mathcal{D}^*(f\|h)(\sigma,\tau)) &= J(T_k(f\|h))(\sigma_{0:k},\tau_{0:k}) \\
&= J(T_k(f)\|T_k(h))(\sigma_{0:k},\tau_{0:k}) \\
&= J(T_k(f))(\sigma_{0:k})\|J(T_k(h))(\tau_{0:k}) \qquad (*) \\
&= T_k(\mathcal{D}^* f(\sigma))\|T_k(\mathcal{D}^* h(\tau)) \\
&= T_k(\mathcal{D}^* f(\sigma)\|\mathcal{D}^* h(\tau))
\end{aligned}
$$

□

Causal linear rule

Theorem (causal linear rule)

If $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ is a linear causal function, it is differentiable at every $\sigma \in (\mathbb{R}^n)^\omega$ and its derivative is $\mathcal{D}^* f(\sigma) = f$.

Proof.

f is linear causal if and only if $T_k(f)$ is linear for all $k \in \mathbb{N}$. Linear functions between finite vector spaces are always their own derivatives, so f is its own derivative. □

Causal linear rule

Theorem (causal linear rule)

If $f : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ is a linear causal function, it is differentiable at every $\sigma \in (\mathbb{R}^n)^\omega$ and its derivative is $\mathcal{D}^* f(\sigma) = f$.

Proof.

f is linear causal if and only if $T_k(f)$ is linear for all $k \in \mathbb{N}$. Linear functions between finite vector spaces are always their own derivatives, so f is its own derivative. □

We can now derive many other standard rules using these three rules.

Causal sum rule

Definition (sum of causal maps)

The *sum* of $f, g : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ is $f + g \triangleq + \circ (f \| g) \circ \mathbf{dup}$.

Causal sum rule

Definition (sum of causal maps)

The *sum* of $f, g : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ is $f + g \triangleq + \circ (f \| g) \circ \mathbf{dup}$.

Theorem (causal sum rule)

If f and g as above are both differentiable at σ, so is $f + g$ and its derivative is $\mathcal{D}^* f(\sigma) + \mathcal{D}^* g(\sigma)$.

Proof.

$$\begin{aligned}
\mathcal{D}^*(f+g)(\sigma) &= \mathcal{D}^*(+ \circ (f \| g) \circ \mathbf{dup})(\sigma) \\
&= \mathcal{D}^*(+)((f \| g \circ \mathbf{dup})(\sigma)) \circ \mathcal{D}^*(f \| g \circ \mathbf{dup})(\sigma) \\
&= + \circ \mathcal{D}^*(f \| g \circ \mathbf{dup})(\sigma) \\
&= + \circ \mathcal{D}^*(f \| g)(\mathbf{dup}(\sigma)) \circ \mathcal{D}^*(\mathbf{dup})(\sigma) \\
&= + \circ \mathcal{D}^*(f \| g)(\sigma, \sigma) \circ \mathbf{dup} \\
&= + \circ (\mathcal{D}^* f(\sigma) \| \mathcal{D}^* g(\sigma)) \circ \mathbf{dup} = \mathcal{D}^* f(\sigma) + \mathcal{D}^* g(\sigma)
\end{aligned}$$

□

Causal product rule

Definition (product of causal maps)

The *product* of $f, g : \mathbb{R}^\omega \to \mathbb{R}^\omega$ is $f \times g \triangleq \times \circ (f \| g) \circ \mathtt{dup}$.

Causal product rule

Definition (product of causal maps)

The *product* of $f, g : \mathbb{R}^\omega \to \mathbb{R}^\omega$ is $f \times g \triangleq \times \circ (f \| g) \circ \mathtt{dup}$.

Theorem (causal product rule)

If f and g as above are both differentiable at σ, so is $f \times g$ and its derivative is $\mathcal{D}^* f(\sigma)(\Delta\sigma) \times g(\sigma) + f(\sigma) \times \mathcal{D}^* g(\sigma)(\Delta\sigma)$.

Proof.

Similar to the sum rule, using the derivative of Cauchy product we found in the first section. □

Stream inverse function

The *stream inverse* is the first partial causal function we will consider. This operation is defined on $\sigma \in \mathbb{R}^\omega$ such that $\sigma_0 \neq 0$ with the unbounded-order recurrence relation

$$[\sigma^{-1}]_k = \begin{cases} \frac{1}{\sigma_0} & \text{if } k = 0 \\ -\frac{1}{\sigma_0} \cdot \sum_{i=0}^{k-1} \left(\sigma_{n-i} \cdot [\sigma^{-1}]_i\right) & \text{if } k > 0 \end{cases}$$

Jan Rutten showed that $\sigma \times \sigma^{-1} = [1] \triangleq (1, 0, 0, 0, \ldots)$ for all σ satisfying $\sigma_0 \neq 0$ (2005). We can use this fact and the product rule to find the derivative of stream inverse using *implicit differentiation*.

Causal reciprocal rule

Theorem (causal reciprocal rule)

Stream inverse is differentiable everywhere it is defined, and its derivative is $\mathcal{D}^*(\cdot)^{-1}(\sigma)(\Delta\sigma) = [-1] \times \sigma^{-1} \times \sigma^{-1} \times \Delta\sigma$.

Proof.

Since $\sigma \times \sigma^{-1} = [1]$, their derivatives must also be equal.

$$\begin{aligned}[0] = \mathcal{D}^*[1] &= \mathcal{D}^*(\sigma \times \sigma^{-1})(\Delta\sigma) \\ &= \sigma \times (\mathcal{D}^*(\cdot)^{-1})(\sigma)(\Delta\sigma) + \Delta\sigma \times (\sigma^{-1})\end{aligned}$$

using the causal product rule. □

Causal reciprocal rule

Theorem (causal reciprocal rule)

Stream inverse is differentiable everywhere it is defined, and its derivative is $\mathcal{D}^*(\cdot)^{-1}(\sigma)(\Delta\sigma) = [-1] \times \sigma^{-1} \times \sigma^{-1} \times \Delta\sigma$.

Proof.

Since $\sigma \times \sigma^{-1} = [1]$, their derivatives must also be equal.

$$\begin{aligned}[0] = \mathcal{D}^*[1] &= \mathcal{D}^*(\sigma \times \sigma^{-1})(\Delta\sigma) \\ &= \sigma \times (\mathcal{D}^*(\cdot)^{-1})(\sigma)(\Delta\sigma) + \Delta\sigma \times (\sigma^{-1})\end{aligned}$$

using the causal product rule. □

Similarly, there is a causal quotient rule much like the ordinary quotient rule.

Causal functions defined by recurrence

Knowing a fact about σ^{-1} helped us find its derivative. But what if we don't have a nice algebraic property for a causal function, only a defining recurrence?

Causal functions defined by recurrence

Knowing a fact about σ^{-1} helped us find its derivative. But what if we don't have a nice algebraic property for a causal function, only a defining recurrence?

If $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ and $i \in \mathbb{R}^m$, then the causal function $\mathtt{rec}_i(g) : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ is defined by the recurrence relation

$$[\mathtt{rec}_i(g)(\sigma)]_k = \begin{cases} g(\sigma_k, i) & \text{if } k = 0 \\ g(\sigma_k, \mathtt{rec}_i(g)(\sigma)_{k-1}) & \text{if } k > 0 \end{cases}$$

Causal functions defined by recurrence

Knowing a fact about σ^{-1} helped us find its derivative. But what if we don't have a nice algebraic property for a causal function, only a defining recurrence?

If $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ and $i \in \mathbb{R}^m$, then the causal function $\mathtt{rec}_i(g) : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ is defined by the recurrence relation

$$[\mathtt{rec}_i(g)(\sigma)]_k = \begin{cases} g(\sigma_k, i) & \text{if } k = 0 \\ g(\sigma_k, \mathtt{rec}_i(g)(\sigma)_{k-1}) & \text{if } k > 0 \end{cases}$$

Example: The unary running product function $\prod : \mathbb{R}^\omega \to \mathbb{R}^\omega$ can be defined by a recurrence relation:

$$\prod(\sigma) = \tau \Leftrightarrow \begin{cases} \tau_0 = \sigma_0 \cdot 1 \\ \tau_{k+1} = \sigma_{k+1} \cdot \tau_k \end{cases}$$

Here g is multiplication of reals and $i = 1$.

The recurrence rule

Theorem (causal recurrence rule)

Let $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ *be differentiable and* $i \in \mathbb{R}^m$. *Then* $\mathtt{rec}_i(g) : (\mathbb{R}^n)^\omega \to (\mathbb{R}^m)^\omega$ *is causal differentiable and its derivative satisfies the following recurrence:*

$$\mathcal{D}^*\mathtt{rec}_i(g)(\sigma)(\Delta\sigma) = \Delta\tau \Leftrightarrow \begin{cases} \tau_0 = g(\sigma_0, i) \\ \tau_{k+1} = g(\sigma_{k+1}, \tau_k) \\ \Delta\tau_0 = Jg(\sigma_0, i)(\Delta\sigma_0, 0_{\mathbb{R}^m}) \\ \Delta\tau_{k+1} = Jg(\sigma_{k+1}, \tau_k)(\Delta\sigma_{k+1}, \Delta\tau_k) \end{cases}$$

Recurrence rule example

Let's find the derivative of the running product function:

$$\prod(\sigma) = \tau \Leftrightarrow \begin{cases} \tau_0 = \sigma_0 \cdot 1 \\ \tau_{k+1} = \sigma_{k+1} \cdot \tau_k \end{cases}$$

Since $J(\cdot)(x, y)(\Delta x, \Delta y) = \Delta x \cdot y + x \cdot \Delta y$, the recurrence rule tells us $\mathcal{D}^* \prod(\sigma)(\Delta\sigma) = \Delta\tau$ if and only if $\Delta\tau$ satisfies

$$\begin{cases} \tau_0 = g(\sigma_0, 1) \\ \tau_{k+1} = g(\sigma_{k+1}, \tau_k) \\ \Delta\tau_0 = Jg(\sigma_0, 1)(\Delta\sigma_0, 0) \\ \Delta\tau_{k+1} = \\ Jg(\sigma_{k+1}, \tau_k)(\Delta\sigma_{k+1}, \Delta\tau_k) \end{cases} = \begin{cases} \tau_0 = \sigma_0 \\ \tau_{k+1} = \sigma_{k+1} \cdot \tau_k \\ \Delta\tau_0 = \Delta\sigma_0 \\ \Delta\tau_{k+1} = \Delta\sigma_{k+1} \cdot \tau_k + \sigma_{k+1} \cdot \Delta\tau_k \end{cases}$$

Outline

Elman networks

An Elman network is a simple kind of recurrent neural network, introduced by Elman in 1990. It looks like:

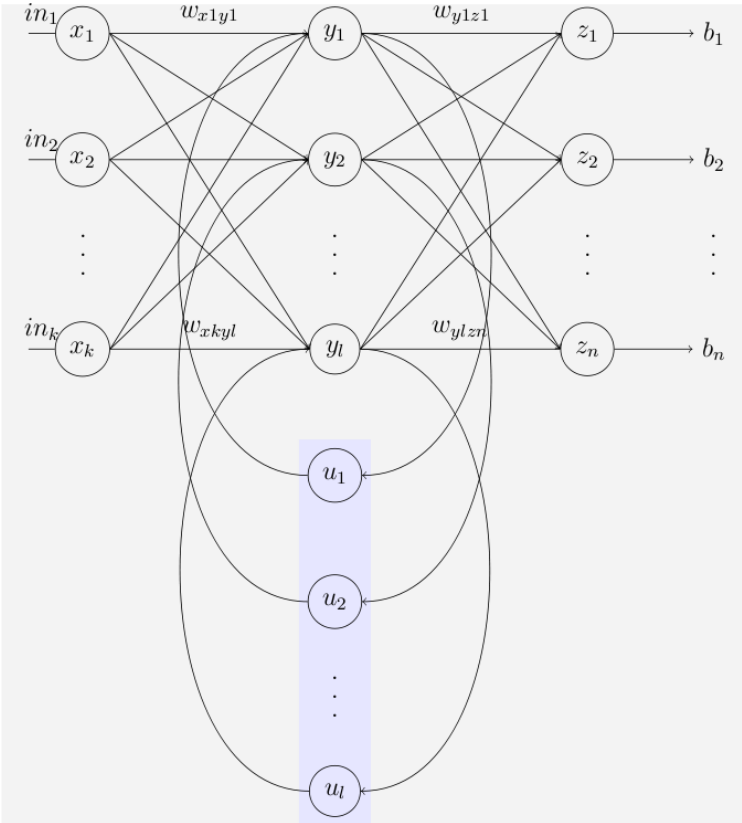


Figure: from Wikipedia

$$\begin{cases} \vec{y}_0 = \phi_1(W\vec{x}_0 + \vec{b}_1) \\ \vec{y}_{k+1} = \phi_1(W\vec{x}_{k+1} + U\vec{y}_k + \vec{b}_1) \\ \vec{z}_0 = \phi_2(V\vec{y}_0 + \vec{b}_2) \\ \vec{z}_{k+1} = \phi_2(V\vec{y}_{k+1} + \vec{b}_2) \end{cases}$$

We choose the weight matrices W, U, V and bias vectors b_1, b_2 to drive the network to some desired behavior. Usually the "activation functions" ϕ_1, ϕ_2 are fixed.

Training an Elman network, I

Take all vectors to be length 1, and fix activation functions to be sigmoids: $\phi_i(x) = \phi(x) := \frac{1}{1+e^{-x}}$. Then the Elman network structure becomes

$$E(x) = z \Leftrightarrow \begin{cases} y_0 = \phi(wx_0 + b_1) \\ y_{k+1} = \phi(wx_{k+1} + uy_k + b_1) \\ z_0 = \phi(vy_0 + b_2) \\ z_{k+1} = \phi(vy_{k+1} + b_2) \end{cases}$$

Training an Elman network, I

Take all vectors to be length 1, and fix activation functions to be sigmoids: $\phi_i(x) = \phi(x) := \frac{1}{1+e^{-x}}$. Then the Elman network structure becomes

$$E(x) = z \Leftrightarrow \begin{cases} y_0 = \phi(wx_0 + b_1) \\ y_{k+1} = \phi(wx_{k+1} + uy_k + b_1) \\ z_0 = \phi(vy_0 + b_2) \\ z_{k+1} = \phi(vy_{k+1} + b_2) \end{cases}$$

Let's imagine we are training our network, and currently $w = u = v = 1$, $b_1 = 0.1$ and $b_2 = -0.1$. Let's also imagine our trained network should satisfy $\hat{E}(1, 1, 1, 1 \ldots) = (0.60, 0.63, 0.64, 0.64, \ldots)$. Currently, $E(1, 1, 1, 1 \ldots) \approx (0.65707, 0.68226, 0.68503, 0.68533, \ldots)$. How should we adjust w?

Training an Elman network, II

With the current parameters and using the fact that $\hat{x}_k \cong 1$,

$$\begin{cases} y_0 = \phi(w\hat{x}_0 + b_1) \\ y_{k+1} = \phi(w\hat{x}_{k+1} + uy_k + b_1) \\ z_0 = \phi(vy_0 + b_2) \\ z_{k+1} = \phi(vy_{k+1} + b_2) \end{cases} = \begin{cases} y_0 = \phi(1 + 0.1) \\ y_{k+1} = \phi(1 + y_k + 0.1) \\ z_0 = \phi(y_0 - 0.1) \\ z_{k+1} = \phi(y_{k+1} - 0.1) \end{cases}$$

Training an Elman network, II

With the current parameters and using the fact that $\hat{x}_k \cong 1$,

$$\begin{cases} y_0 = \phi(w\hat{x}_0 + b_1) \\ y_{k+1} = \phi(w\hat{x}_{k+1} + uy_k + b_1) \\ z_0 = \phi(vy_0 + b_2) \\ z_{k+1} = \phi(vy_{k+1} + b_2) \end{cases} = \begin{cases} y_0 = \phi(1 + 0.1) \\ y_{k+1} = \phi(1 + y_k + 0.1) \\ z_0 = \phi(y_0 - 0.1) \\ z_{k+1} = \phi(y_{k+1} - 0.1) \end{cases}$$

Using causal derivatives, we get a recurrence:

$$\begin{cases} y_{k+1} = \phi(y_k + 1.1) & \text{sf } y_0 = \phi(1.1) \\ z_{k+1} = \phi(y_{k+1} - 0.1) & \text{sf } z_0 = \phi(y_0 - 0.1) \\ \Delta y_{k+1} = \phi'(y_k + 1.1) \cdot (\Delta w_{k+1} + \Delta y_k) & \text{sf } \Delta y_0 = \phi'(1.1) \cdot \Delta w_0 \\ \Delta z_{k+1} = \phi'(y_{k+1} - 0.1) \cdot \Delta y_{k+1} & \text{sf } \Delta z_0 = \phi'(y_0 - 0.1) \cdot \Delta y_0 \end{cases}$$

Training an Elman network, II

With the current parameters and using the fact that $\hat{x}_k \cong 1$,

$$\begin{cases} y_0 = \phi(w\hat{x}_0 + b_1) \\ y_{k+1} = \phi(w\hat{x}_{k+1} + uy_k + b_1) \\ z_0 = \phi(vy_0 + b_2) \\ z_{k+1} = \phi(vy_{k+1} + b_2) \end{cases} = \begin{cases} y_0 = \phi(1 + 0.1) \\ y_{k+1} = \phi(1 + y_k + 0.1) \\ z_0 = \phi(y_0 - 0.1) \\ z_{k+1} = \phi(y_{k+1} - 0.1) \end{cases}$$

Using causal derivatives, we get a recurrence:

$$\begin{cases} y_{k+1} = \phi(y_k + 1.1) & \text{sf } y_0 = \phi(1.1) \\ z_{k+1} = \phi(y_{k+1} - 0.1) & \text{sf } z_0 = \phi(y_0 - 0.1) \\ \Delta y_{k+1} = \phi'(y_k + 1.1) \cdot (\Delta w_{k+1} + \Delta y_k) & \text{sf } \Delta y_0 = \phi'(1.1) \cdot \Delta w_0 \\ \Delta z_{k+1} = \phi'(y_{k+1} - 0.1) \cdot \Delta y_{k+1} & \text{sf } \Delta z_0 = \phi'(y_0 - 0.1) \cdot \Delta y_0 \end{cases}$$

(Check out the demo.)

Recap & future work

Today we talked about:

- a definition for the derivative of a causal function
- rules for the causal differential calculus
- a unique rule for causal calculus—the recurrence rule
- training Elman networks with causal derivatives

Recap & future work

Today we talked about:

- a definition for the derivative of a causal function
- rules for the causal differential calculus
- a unique rule for causal calculus—the recurrence rule
- training Elman networks with causal derivatives

Future work might include:

- formalizing this as a Cartesian differential restriction category
- causal automatic differentiation
- causal integral calculus
- generalizing from sequences to other infinite data shapes

Thanks!