Differentiation in probabilistic coherence spaces: reconciliating differentiation and determinism?

BIRS/FMCS workshop on Tangent Categories and their Applications,
June 2021

Thomas Ehrhard
IRIF, CNRS and Université de Paris

Probabilistic Coherence Spaces, introduced by Girard and developed by Danos and E.

Model of LL with general recursion, all recursive types (pure λ-calculus).

And also of higher-order functional programs with good properties: adequacy and full abstraction (E., Pagani and Tasson).

Fact

Major feature: morphisms of the Kleisli category are "analytic functions" with ≥ 0 real coefficients. So they have derivatives.

It is a model of classical LL but not of differential LL.

Because it is not a pre-additive category: $X \oplus Y \not\simeq X \mathbin{\&} Y$.

What is a PCS?

$X = (|X|, \mathrm{P}X)$ where

- $|X|$ is a set
- and $\mathrm{P}X \subseteq (\mathbb{R}_{\geq 0})^{|X|}$: the *valuations* of X.

such that $\mathrm{P}X$ which is ($x \leq y$ if $\forall a \in |X|\ x_a \leq y_a$):

- $\downarrow$-closed
- closed under lubs of sequences $x(1) \leq x(2) \leq \cdots \in \mathrm{P}X$
- convex ($x, y \in \mathrm{P}X,\ \lambda \in [0,1] \Rightarrow \lambda x + (1-\lambda)y \in \mathrm{P}X$)
- $+$ a technical condition to avoid ∞ coefficients.

Notation: if $a \in |X|$ then $e_a \in (\mathbb{R}_{\geq 0})^{|X|}$ defined by $(e_a)_b = \delta_{a,b}$.

Linear morphisms

$\mathrm{Pcoh}(X, Y)$ is the set of all matrices $t \in (\mathbb{R}_{\geq 0})^{|X| \times |Y|}$ such that

$$\forall x \in \mathrm{P}X \quad t \cdot x = \sum_{a \in |X|, b \in |Y|} t_{a,b} x_a e_b \in \mathrm{P}Y$$

Fact

Pcoh *is a cartesian SMCC with a* $*$*-autonomous structure. It has also a resource modality comonad* $(!\_, \mathrm{der}, \mathrm{dig})$ + *Seely isomorphisms.*

$\mathrm{Pcoh}_!$ *is a CCC with least fixpoint operators* $(X \Rightarrow X) \to X$*, a model of PCF and other functional languages. All recursive types, pure* λ*-calculus etc.*

Basic constructions

- $1 = (\{*\}, [0,1])$ is the $\otimes$-unit and dualizing object;
- $|X \multimap Y| = |X| \times |Y|$, $\mathrm{P}(X \multimap Y) = \mathrm{Pcoh}(X, Y)$;
- $|X^\perp| = |X|$,
  $\mathrm{P}X^\perp = \{x' \mid \forall x \in \mathrm{P}X \ \langle x, x'\rangle = \sum_{a \in |X|} x_a x'_a \leq 1\}$;
- $X \otimes Y = (X \multimap Y^\perp)^\perp$ so that if $x \in \mathrm{P}X$ and $y \in \mathrm{P}Y$,
  $x \otimes y = (x_a y_b)_{(a,b) \in |X| \times |Y|} \in \mathrm{P}(X \otimes Y)$;
- product: $\mathrm{P}(X \mathbin{\&} Y) \simeq \mathrm{P}X \times \mathrm{P}Y$;
- coproduct, $X \oplus Y = (X^\perp \mathbin{\&} Y^\perp)^\perp$:
  $\mathrm{P}(X \oplus Y) \simeq \{(\lambda x, (1-\lambda) y) \mid x \in \mathrm{P}X,\ y \in \mathrm{P}Y \text{ and } \lambda \in [0,1]\} \subseteq \mathrm{P}(X \mathbin{\&} Y)$. Strict inclusion in general!

Exponential, analytic functions

$|!X| = \mathcal{M}_{\mathrm{fin}}(|X|)$ and $\mathrm{P}!X = \{(x^\mu)_{\mu\in|X|} \mid x \in \mathrm{P}X\}^{\perp\perp}$ where $x^\mu = \prod_{a\in|X|} x_a^{\mu(a)} \in \mathbb{R}_{\geq 0}$.

Fact

$t \in \mathrm{Pcoh}(!X, Y)$ *iff for all* $x \in \mathrm{P}X$ *one has*

$$\sum_{\mu\in\mathcal{M}_{\mathrm{fin}}(|X|), b\in|Y|} t_{\mu,b} x^\mu e_b \in \mathrm{P}Y \subseteq (\mathbb{R}_{\geq 0})^{|Y|}$$

A power series with ≥ 0 real coefficients, which converges on the whole set $\mathrm{P}X$ but not outside in general.

Example

Fact

If $x : 1 \oplus 1 \vdash M : 1 \oplus 1$ then $[\![M]\!]_{x:1\oplus1}$, the semantics of M, is in $\mathrm{Pcoh}(!(1 \oplus 1), 1 \oplus 1)$, *analytic function* $\mathrm{P}(1 \oplus 1) \to \mathrm{P}(1 \oplus 1)$.

Take

$$M = \mathrm{if}(x, \mathrm{if}(x, M, \mathrm{t}), \mathrm{if}(x, \mathrm{f}, M))$$

then $f = [\![M]\!]_{x:1\oplus1}$ is the “least” function $f : \mathrm{P}(1 \oplus 1) \to \mathrm{P}(1 \oplus 1)$ such that

$$f(u) = (u_{\mathrm{t}}^2 + u_{\mathrm{f}}^2) f(u) + u_{\mathrm{t}} u_{\mathrm{f}} (e_{\mathrm{t}} + e_{\mathrm{f}})$$

$$f(u) = \begin{cases} 0 & \text{if } u_{\mathrm{t}} = 1 \text{ or } u_{\mathrm{f}} = 1 \\ \frac{u_{\mathrm{t}} u_{\mathrm{f}}}{1 - u_{\mathrm{t}}^2 - u_{\mathrm{f}}^2} (e_{\mathrm{t}} + e_{\mathrm{f}}) & \text{otherwise} \end{cases}$$

In “polar” coordinates

Any boolean $u \in \mathrm{P}(1 \oplus 1)$ can be written $u = r\theta e_{\mathrm{t}} + r(1-\theta)e_{\mathrm{f}}$ where

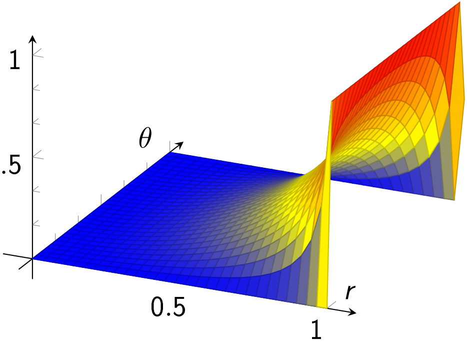
- r is the probability of convergence
- θ defines a total boolean $\theta e_{\mathrm{t}} + (1-\theta)e_{\mathrm{f}}$.

Fact

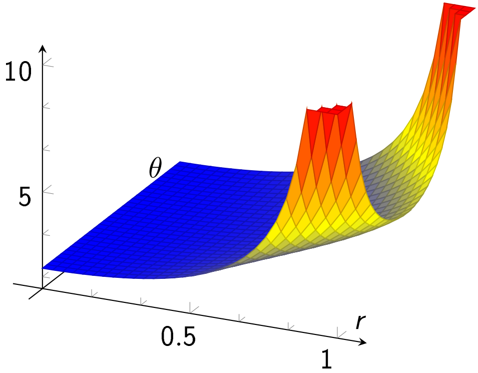
$$f(\theta e_{\mathrm{t}} + (1-\theta)e_{\mathrm{f}}) = \begin{cases} \frac{1}{2}e_{\mathrm{t}} + \frac{1}{2}e_{\mathrm{f}} & \textit{if } 0 < \theta < 1 \\ 0 & \textit{if } \theta = 0 \textit{ or } \theta = 1 \end{cases}$$

Convergence probability of f:

$$g(r,\theta) = f(r\theta e_{\mathrm{t}} + r(1-\theta)e_{\mathrm{f}})_{\mathrm{t}} + f(r\theta e_{\mathrm{t}} + r(1-\theta)e_{\mathrm{f}})_{\mathrm{f}}$$



$\min(10, r\frac{\partial g}{\partial r}/g) = \min(10, 2/(1 - r^2(1 - 2\theta - 2\theta^2)))$:



Fact

$r\frac{\partial g}{\partial r}/g$ = expectation of the number of uses of x, conditioned by termination, makes sense for all r (for $r < 1$, $u = (r\theta, r(1-\theta))$ is a partial probabilistic boolean and $g(r,\theta) = f(u)_{\mathrm{t}} + f(u)_{\mathrm{f}} < 1$).

The problem

So computing derivatives in $\mathrm{Pcoh}_!$ makes sense:

- related to execution time
- could be used also for learning (“gradient” method).

But it is problematic because derivatives:

- require sums which are not barycentric combinations (Leibniz) and hence are not freely available in the model
- seem to induce morphisms which are not in the model.

For instance if $n \in \mathbb{N}$, then $f(x) = x^n$ is in $\mathrm{Pcoh}(!1, 1)$:

- all coefficients in $\mathbb{R}_{\geq 0}$
- and maps $\mathrm{P}1 = [0, 1]$ to $\mathrm{P}1$.

However $f'(x) = nx^{n-1}$ is not in $\mathrm{Pcoh}(!1, 1)$ as soon as $n > 1$.

Fact

But, if $x, u \in \mathrm{P}1$ satisfy $x + u \in \mathrm{P}1$, then $f'(x)u \leq 1$.

This is a general phenomenon if one takes care of computing the derivative “locally”.

The local derivative

More generally, given $x \in \mathsf{P}X$ we can define a local PCS X_x as follows:

$$|X_x| = \{a \in |X| \mid \exists \varepsilon > 0 \ x + \varepsilon e_a \in \mathsf{P}X\}$$
$$\mathsf{P}(X_x) = \{u \in (\mathbb{R}_{\geq 0})^{|X_x|} \mid x + u \in \mathsf{P}X\}$$

The "local PCS at x". Observe that

$$\mathsf{P}(X_x) \simeq \{u \in \mathsf{P}X \mid x + u \in \mathsf{P}X\}$$

as convex posets.

Then given $f \in \mathrm{Pcoh}(!X, Y) = \mathrm{Pcoh}_!(X, Y)$, we can define

$$f'(x) \in \mathrm{Pcoh}(X_x, Y_{f(x)})$$

by

$$f'(x) \cdot u = \text{the "}u\text{-linear part" of } f(x+u)$$

that is

$$f'(x) \cdot u = \sum_{\substack{(a,b) \in |X_x| \times |Y_{f(x)}| \\ \mu \in \mathcal{M}_{\mathrm{fin}}(|X|)}} (\mu(a)+1) t_{\mu,b} x^{\mu} u_a e_b$$

We have

$$f(x) \leq f(x) + f'(x) \cdot u \leq f(x+u) \in \mathrm{P}Y$$

and hence $\forall u \in \mathrm{P}X_x \; f'(x) \cdot u \in \mathrm{P}Y_{f(x)}$.

The functor of summable pairs

This suggests a kind of tangent category structure on Pcoh based on a very simple functor

$$\mathbb{S} : \mathsf{Pcoh} \to \mathsf{Pcoh}$$

given by

$$\begin{aligned} |\mathbb{S}X| &= \{0,1\} \times |X| \\ \mathsf{P}(\mathbb{S}X) &= \{s \in (\mathbb{R}_{\geq 0})^{|\mathbb{S}X|} \mid (s_{0,a} + s_{1,a})_{a \in |X|} \in \mathsf{P}X\} \end{aligned}$$

In other words $\mathsf{P}(\mathbb{S}X) = \{(x, u) \in \mathsf{P}X^2 \mid x + u \in \mathsf{P}X\}$.

And if $t \in \mathrm{Pcoh}(X, Y)$ then $\mathbb{S}t \in \mathrm{Pcoh}(\mathbb{S}X, \mathbb{S}Y)$ is defined by

$$(\mathbb{S}t)_{(i,a),(j,b)} = \begin{cases} t_{a,b} & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

In other words $\mathbb{S}t \cdot (x, u) = (t \cdot x, t \cdot u)$.

This definition makes sense since $t \cdot x + t \cdot u = t \cdot (x + u) \in \mathrm{P}Y$ by linearity.

Remark: no differentiation involved!

Remark: Of course if $x, u \in \mathsf{P}X$ and $\lambda \in [0,1]$ we have $(\lambda x, (1-\lambda)u) \in \mathsf{P}(\mathbb{S}X)$.

But an element of $\mathbb{S}X$ is not necessarily of that shape: if $X = X_1 \mathbin{\&} X_2$ and $x(i) \in \mathsf{P}X_i$ for $i = 1, 2$, we have $(x(1), 0), (0, x(2)) \in \mathsf{P}(X_1 \mathbin{\&} X_2)$ and

$$((x(1), 0), (0, x(2))) \in \mathsf{P}(\mathbb{S}(X_1 \mathbin{\&} X_2))$$

Fact

The functor $\mathbb{S}$ preserves all existing limits because it has a left adjoint $\mathbb{S}^\perp : \mathsf{Pcoh} \to \mathsf{Pcoh}$ defined by $\mathbb{S}^\perp X = \mathbb{S}(X^\perp)^\perp$ and similarly on morphisms.

$\mathbb{S}^\perp(X) = \{(x, x) \mid x \in \mathsf{P}X\}^{\perp\perp}$. *Notice that $\mathbb{S}X \not\simeq \mathbb{S}^\perp X$ in general.*

$\mathbb{S}$ has a lot of structures, in particular:

- $\iota_i : X \to \mathbb{S}X$ for $i = 0, 1$: $\iota_0 \cdot x = (x, 0)$ and $\iota_1 \cdot u = (0, u)$
- $\pi_i : \mathbb{S}X \to X$ for $i = 0, 1$: $\pi_0 \cdot (x, u) = x$ and $\pi_1 \cdot (x, u) = u$.

with $\pi_i \, \iota_j = \delta_{i,j} \, \mathsf{Id}$ (there are 0-morphisms).

$$X \oplus X \xrightarrow{[\iota_0, \iota_1]} \mathbb{S}X \xrightarrow{\langle \pi_0, \pi_1 \rangle} X \,\&\, X$$

NB: $\langle \pi_0, \pi_1 \rangle$ mono. We also have $\mathsf{s} : \mathbb{S}X \to X$ ($\mathsf{s} \cdot (x, u) = x + u$):

$$\begin{tikzcd}
X \arrow[r, "\iota_i"] \arrow[dr, "\mathsf{Id}"'] & \mathbb{S}X \arrow[d, "\mathsf{s}"] \\
 & X
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathbb{S}^2X \arrow[r, "\mathbb{S}\mathsf{s}_X"] \arrow[d, "\mathsf{s}_{\mathbb{S}X}"'] & \mathbb{S}X \arrow[d, "\mathsf{s}_X"] \\
\mathbb{S}X \arrow[r, "\mathsf{s}_X"] & X
\end{tikzcd}$$

Summability

$f_0, f_1 \in \mathrm{Pcoh}(Y, X)$ are summable if there is $g \in \mathrm{Pcoh}(Y, \mathbb{S}X)$ such that

$$\begin{tikzcd}
Y \arrow[r, "g"] \arrow[dr, "\langle f_0,f_1\rangle"'] & \mathbb{S}X \arrow[r, "\mathsf{s}"] \arrow[d, "\langle\pi_0,\pi_1\rangle"] & X \\
 & X \mathbin{\&} X &
\end{tikzcd}$$

and then we have $f_0 + f_1 = \mathsf{s}\, g$. Well defined because $\langle \pi_0, \pi_1 \rangle$ is a mono.

Differential structure

Goal: differentiate morphisms in $\mathsf{Pcoh}_!$, that is linear morphisms $!X \multimap Y$

$\rightsquigarrow$ as in DiLL, differentiation is considered as a structure of the exponential.

A distributive law

$$\partial_X \in \mathsf{Pcoh}(!\mathbb{S}X, \mathbb{S}!X)$$

so that

$$\begin{tikzcd}
!\mathbb{S}X \arrow[r, "\partial_X"] \arrow[dr, "\mathsf{der}_{\mathbb{S}X}"'] & \mathbb{S}!X \arrow[d, "\mathbb{S}\mathsf{der}_X"] \\
 & \mathbb{S}X
\end{tikzcd}$$

$$\begin{tikzcd}
!\mathbb{S}X \arrow[rr, "\partial_X"] \arrow[d, "\mathsf{dig}_{\mathbb{S}X}"'] & & \mathbb{S}!X \arrow[d, "\mathbb{S}\mathsf{dig}_X"] \\
!!\mathbb{S}X \arrow[r, "!\partial X"] & !\mathbb{S}!X \arrow[r, "\partial_{!X}"] & \mathbb{S}!!X
\end{tikzcd}$$

Notice that $|!\mathbb{S}X| = \mathcal{M}_{\mathrm{fin}}(|X|)^2$ and $|\mathbb{S}!X| = \{0,1\} \times \mathcal{M}_{\mathrm{fin}}(|X|)$.

$\partial_X : !\mathbb{S}X \to \mathbb{S}!X$ is given by

$$(\partial_X)_{(\lambda,\rho),(i,\mu)} = \begin{cases} 1 & \text{if } i = 0,\ \lambda = \mu \text{ and } \rho = [\,] \\ \mu(a) & \text{if } i = 1,\ \rho = [a] \text{ and } \mu = \lambda + [a] \\ 0 & \text{otherwise.} \end{cases}$$

∂ induces a functor $\mathrm{D} : \mathbf{Pcoh}_! \to \mathbf{Pcoh}_!$ as usual by:
given $f : !X \to Y$, we set $\mathrm{D}f = (\mathbb{S}f)\,\partial_X : !\mathbb{S}X \to \mathbb{S}Y$.

Fact

If we consider f as a function $\mathrm{P}X \to \mathrm{P}Y$ and $\mathrm{D}f$ as a function $\mathrm{P}\mathbb{S}X \to \mathrm{P}\mathbb{S}Y$ then we have

$$\mathrm{D}f(x,u) = (f(x), f'(x) \cdot u)$$

Functoriality of D is chain rule.

Differentiation of a constant function

Remember that each $!X$ has a commutative comonoid structure $\mathsf{w}_X : !X \to 1$, $\mathsf{contr}_X : !X \to !X \otimes !X$.

Derivative of a constant function is 0:

$$\begin{tikzcd}
!\mathbb{S}X \arrow[r, "\partial_X"] \arrow[dr, "0"'] & \mathbb{S}!X \arrow[d, "\mathbb{S}\mathsf{w}_X"] \\
 & \mathbb{S}1
\end{tikzcd}$$

Leibniz law

Derivative of $f(x, x)$ is the sum of two partial derivatives:

$$\begin{tikzcd}
!\mathbb{S}X \arrow[rr, "\partial_X"] \arrow[d, "\mathsf{contr}_{\mathbb{S}X}"'] & & \mathbb{S}!X \arrow[d, "\mathbb{S}\mathsf{contr}_X"] \\
!\mathbb{S}X \otimes !\mathbb{S}X \arrow[r, "\partial_X \otimes \partial_X"] & \mathbb{S}!X \otimes \mathbb{S}!X \arrow[r, "\mathsf{L}_{!X,!X}"] & \mathbb{S}(!X \otimes !X)
\end{tikzcd}$$

What is L?

Fact

The two morphisms $\pi_0 \otimes \pi_1, \pi_1 \otimes \pi_0 : \mathbb{S}X \otimes \mathbb{S}Y \to X \otimes Y$ *are summable.*

We have used the lax monoidality $\mathsf{L}_{X,Y} : \mathbb{S}X \otimes \mathbb{S}Y \to \mathbb{S}(X \otimes Y)$ which satisfies

$$\begin{tikzcd}
\mathbb{S}X \otimes \mathbb{S}Y \arrow[r, "\mathsf{L}_{X,Y}"] \arrow[dr, "\pi_0 \otimes \pi_0"'] & \mathbb{S}(X \otimes Y) \arrow[d, "\pi_0"] \\
& X \otimes Y
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathbb{S}X \otimes \mathbb{S}Y \arrow[r, "\mathsf{L}_{X,Y}"] \arrow[dr, "\pi_0 \otimes \pi_1 + \pi_1 \otimes \pi_0"'] & \mathbb{S}(X \otimes Y) \arrow[d, "\pi_1"] \\
& X \otimes Y
\end{tikzcd}$$

The strength of $\mathbb{S}$ and partial derivatives

There is a kind of tensorial strength:

$$\rho_{X,Y} : X \otimes \mathbb{S}Y \xrightarrow{\iota_0 \otimes \mathbb{S}Y} \mathbb{S}X \otimes \mathbb{S}Y \xrightarrow{\mathsf{L}_{X,Y}} \mathbb{S}(X \otimes Y)$$

which gives the second partial derivative of $f : !X \otimes !Y \to Z$

$$!X \otimes !\mathbb{S}Y \xrightarrow{!X \otimes \partial_Y} !X \otimes \mathbb{S}!Y \xrightarrow{\rho_{!X,!Y}} \mathbb{S}(!X \otimes !Y) \xrightarrow{\mathbb{S}f} \mathbb{S}Z$$

Concluding remarks

Clearly very close to tangent categories.

And more specifically to *Tangent Categories from the Coalgebras of Differential Categories* by Cockett, Lemay and Lucyshyn-Wright (distributive law). Main differences:

- our objects have an “algebraic structure” $\mathsf{s} : \mathbb{S}X \to X$
- though the category is not (left) additive
- positive outcome: the category can have fixpoint operators.

In an additive category with fixpoints, the function $x \mapsto x + u$ must have a fixpoint which requires ∞ or idempotent coefficients.

Other similarity: partial traces for the GoI, partial sums (Haghverdi, Scott).

What kind of derivatives do we compute?

We consider programs typically as *analytic probability sub-distribution transformers*. We compute their Jacobians as such.

A priori it is not like having a ground type of real numbers and computing derivatives wrt. parameters in this type as in usual differential programming languages.

Thanks to this model, we will be able to internalize differentiation in a differential-probabilistic LL or λ-calculus, with fixpoints, recursive types etc.

Possible extensions

To positive cones and stable functions (E., Pagani, Tasson) which are a model of probabilistic computations compatible with continuous data-types (real line etc), conservative extension of Pcoh.

More surprisingly, similar structures are available in stable domain semantics, typically coherence spaces and hopefully also dI-domains.

Big difference wrt. standard DiLL: this “local” DiLL seems compatible with determinism. The $\mathbb{S}$ functor accounts for compatibility (boundedness in domains).

Possible connections with incremental programming?