New Insights in Categorical Probability

Dario Stein, Radboud University Nijmegen

DutchCATs 2023
Nijmegen – 3 November 2023

Overview

Recap of the Categorical Probability Programme

Some new directions

1. Convex Analysis and Probability
2. Combining Nondeterminism and Probability
3. What are Random Variables?

Categorical Probability

Idea

Axiomatize categories of stochastic computations directly.

1. Copy-Delete categories (Cho-Jacobs)
   - unnormalized computation (failure, nondetermination, conditioning etc.)
2. Markov categories (Fritz)
   - normalized stochastic computation (sampling only)
3. high-level and graphical reasoning (no measure theory!)
4. rigorous, mechanizable and general results

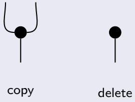
Applications

1. semantics of probabilistic programming
2. causal inference (free CD categories)
3. transfering probabilistic ideas to new domains

Definitions

CD category

A **CD category** is a symmetric monoidal category $(\mathbb{C}, \otimes, I)$ with comonoid structures



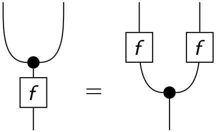
Markov category

A **Markov category** is a CD category where deletion is natural.

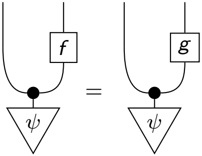
1. Axiomatizes that probability measures are normalized (integrate to 1).

Concepts in Synthetic Probability

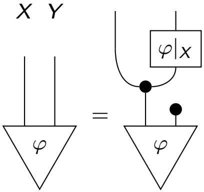
Elegant abstract definitions for probabilistic notions



f deterministic



$f = g$ ψ-almost surely



conditional distribution

Formalized: Absolute continuity, Supports, Kolmogorov extension, Kolmogorov's 0/1 law, De Finetti's theorem, Aldous-Hoover
In Progress: Conditional Expectation, Law of Large Numbers, Martingale convergence

Probabilistic Models

On a gradient (very simple – very comprehensive)

1. **FinStoch**: $p : X \to Y$ is a stochastic matrix $p \in [0,1]^{X \times Y}$, i.e.

$$\forall x, \sum_y p(y|x) = 1$$

2. **Gauss**: affine-linear maps with Gaussian noise

$$f(x) = Ax + \mathcal{N}(\mu, \Sigma)$$

   Composition $f(\mathcal{N}(\mu', \Sigma')) = \mathcal{N}(A\mu' + \mu, A\Sigma' A^T + \Sigma)$

3. measurable kernels
   - standard Borel spaces
   - compact Hausdorff spaces (continuous kernels)
   - measurable spaces
   - quasi-Borel spaces

More exotic models

A source of models

For a strong monad T on a cartesian category $\mathbb{C}$, $\mathcal{K}l(T)$ is

1. copy-delete if T is commutative (Kock)
2. Markov if T is commutative and affine ($T(1) \cong 1$)

More examples

1. partial functions, $(-) + 1 : \mathbf{Set} \to \mathbf{Set}$
2. nondeterminism, $P^+ : \mathbf{Set} \to \mathbf{Set}$
3. negative probabilities, $\mathcal{D}_\pm : \mathbf{Set} \to \mathbf{Set}$
4. fresh name generation, $N : \mathbf{Nom} \to \mathbf{Nom}$

We will discuss: Convex analysis, linear relations

Part I - Gaussians and Convexity

Convex analysis is a rich field of mathematics. But convex functions don't compose.

Definition [Rockafellar'70]

A *bifunction* $F : \mathbb{R}^m \to \mathbb{R}^n$ is a weighted relation

$$\underline{F} : \mathbb{R}^m \times \mathbb{R}^n \to \overline{\mathbb{R}}$$

where $\overline{\mathbb{R}} = ([-\infty; +\infty], \wedge, +)$ is the quantale of extended reals. A bifunction is *convex* if $\underline{F}$ is a jointly convex function. Convex bifunctions compose via infimization

$$(F; G)(x, z) = \inf_y \{F(x, y) + G(y, z)\}$$

We write $F : \mathbb{R}^m \rightharpoonup \mathbb{R}^n$ for convex and $\mathbb{R}^m \rightharpoondown \mathbb{R}^n$ for concave bifunctions (compose via supremization).

Convex functions $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ are states $\mathbb{R}^0 \rightharpoonup \mathbb{R}^m$. Bifunctions are self-dual (hypergraph category).

Part I - Gaussians and Convexity

The indicator bifunction of $A \in \mathbb{R}^{n \times m}$ is $F_A : \mathbb{R}^m \rightharpoonup \mathbb{R}^n$

$$F_A(x, y) = \{\!|y = Ax|\!\} = \begin{cases} 0, & y = Ax \\ +\infty, & y \neq Ax \end{cases}$$

Theorem

Taking indicator bifunctions is a functor of copy-delete categories

$$F : \mathbf{Vect} \to \mathbf{CxBiFn}$$

Other subcategories of **CxBiFn**

1. linear and affine relations
2. convex relations
3. convex optimization problems

Duality

Convex analysis has a rich duality given by the **Legendre-Fenchel transform**.

Definition

The *adjoint* of a convex bifunction $F : \mathbb{R}^m \rightharpoonup \mathbb{R}^n$ is the *concave* bifunction $F^* : \mathbb{R}^n \rightharpoondown \mathbb{R}^m$ defined by

$$F^*(y^*, x^*) = \inf_{x,y} \{F(x,y) + \langle x^*, x\rangle - \langle y^*, y\rangle\}$$

Soft Theorem

Under regularity assumptions, the adjoint behaves like an idempotent functor

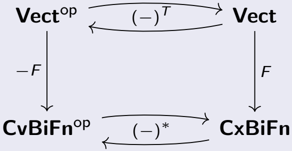
$$(-)^* : \mathbf{CxBiFn} \to \mathbf{CvBiFn}^{\mathrm{op}}$$

i.e. $(F; G)^* = G^*; F^*$, $F^{**} = F$.

Duality

Theorem

Adjoints of bifunctions generalize the matrix transpose

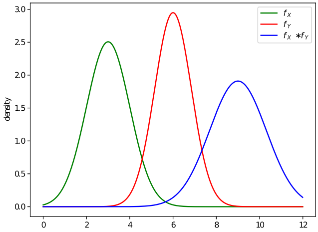


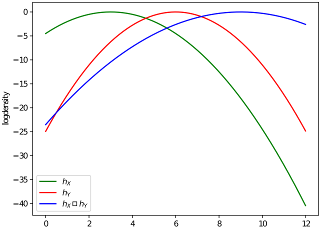
1. the same story works for linear relations (graphical linear algebra)
2. **How does probability fit into the picture?**

Part I - Gaussians and Convexity

The logpdf of a Gaussian $\mathcal{N}(\mu, \sigma^2)$ is a concave quadratic function

$$h(x) = \log f(x) = -\frac{1}{2\sigma^2}(x-\mu)^2$$





Question

Instead of computing integrals of densities, can we compute suprema of logdensities?

$$\log \int f_1(x) f_2(y-x)\mathrm{d}x \to \sup_x \{\log f_1(x) + \log f_2(y-x)\}$$

This is the "tropical limit" (aka Laplace approximation)

Question

Instead of computing integrals of densities, can we compute suprema of logdensities?

$$\log \int f_1(x) f_2(y-x)\mathrm{d}x \to \sup_x \{\log f_1(x) + \log f_2(y-x)\}$$

This is the "tropical limit" (aka Laplace approximation)

Yes (but only for Gaussians!)

Special Theorem

Logpdf is a functor of copy-delete categories **Gauss** $\to$ $\mathbf{CvBiFn}^{op}$.

Part I - Gaussians and Duality

Another special fact: The convex conjugate of a Gaussian logpdf is its cumulant generating function (cgf)

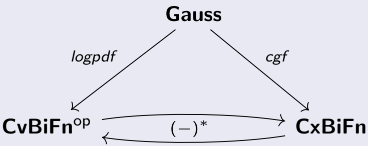
$$c_X(t) = \log \mathbb{E}[\exp(tX)]$$

e.g. for $\mathcal{N}(\mu, \sigma^2)$

$$h(x) = -\frac{1}{2\sigma^2}(x-\mu)^2 \Leftrightarrow h^*(t) = \frac{1}{2}\sigma^2 t^2 + \mu t$$

Gaussians and Duality

We have a commuting diagram of copy-delete functors



Both Gaussians and linear relations embed in bifunctions. **Combine the two?**

Definition

An *extended Gaussian distribution* on $\mathbb{R}^n$ is a pair (D, ψ) of a subspace $D \subseteq \mathbb{R}^n$ and a Gaussian distribution on the quotient $\mathbb{R}/D$.

1. D is a *nondeterministic fibre* along which we have no information
2. we can of this as a distribution on $(\mathbb{R}^n, \mathcal{E})$ with $\mathcal{E} = \text{Borel}(\mathbb{R}^n/D)$.
3. the coarse σ-algebra captures lack of information (Willems: 'open stochastic system')

Part II - Gaussians and Nondeterminism

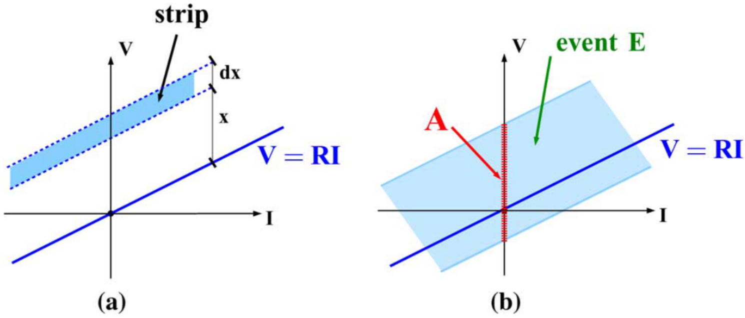


Fig. 2. Events for the noisy resistor.

Part II - Gaussians and Nondeterminism

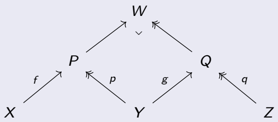
In Willems' approach, σ-algebras are part of the objects! We want to make nondeterminsm part of the morphisms:

Definition

An *extended Gaussian morphism* is a cospan of linear maps

$$X \xrightarrow{f} P \xleftarrow{p} Y$$

whose right leg is epi, together with a Gaussian distribution on P. Compose by pushout



Example

The 'uniform distribution' over $\mathbb{R}$ is represented by the cospan $0 \to 0 \xleftarrow{!} \mathbb{R}$.

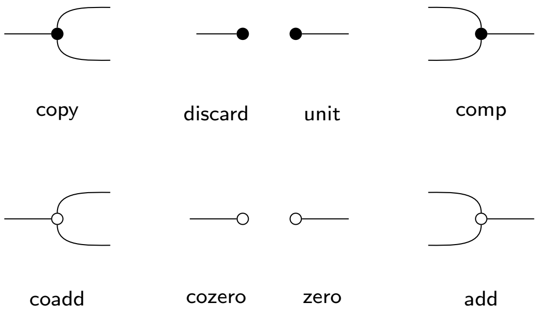
Notice the duality with partial maps

1. a partial map is a span $X \xleftarrow{m} A \xrightarrow{f} Y$ with m monic
2. a copartial map is a cospan $X \xrightarrow{f} P \xleftarrow{p} Y$ with p epic

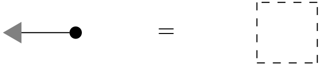
Part II - Generators and Relations

Proof technique: **presentations by generators and relations**

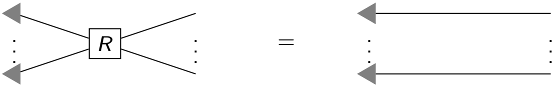
1. for linear maps and linear relations, this is Graphical Linear Algebra
2. the same generators give two hypergraph structures on **CxBiFn**



Extend affine linear algebra with a single new generator such that



(D)



(RI)

for all orthogonal matrices R.

Surprising Theorem

This presentation of **Gauss** is complete.

Next step: represent the category of partial convex quadratic functions by suitable relations.

Part III - Random Variables and Local State

We have talked about distributions/channels, but **what are random variables?**

1. can be meaningfully compared for equality $X = Y$ (almost surely)
2. are *cartesian*, i.e. can be copied, discarded, a random variable of type $X \times Y$ is a pair of random variables (X, Y)

Traditional answer: Measurable function $X : (\Omega, p) \to V$ where (Ω, p) is a **sample space**.

1. where does the sample space come from?
2. how to make dependence on Ω explicit?
3. in practice, the sample space gets modified or extended on the fly?

Part III - Random Variables, a Computer Science view

Alex Simpson's answer: random variables are like pointers to a heap

- we consider sheaves over sample spaces
- have a sheaf of random variables $\mathrm{RV}(X)$,

$$\mathrm{RV}(X)(\Omega, p) = \{f : (\Omega, p) \to X \text{ measurable } \}/a.s.$$

- functorial action = extension of sample spaces

Desirable results:

1. Boolean topos where everything is equivariant under change of sample spaces
2. $\mathrm{RV}(X \times Y) \cong \mathrm{RV}(X) \times \mathrm{RV}(Y)$
3. Conditional expectation $\mathbb{E} : \mathrm{RV}(\mathbb{R}) \times \mathrm{RV}(X) \to \mathrm{RV}(\mathbb{R})$
4. Allocation of random variables via a local state monad

$$(MF)(\Omega) = \int^{\Omega'} \mathrm{hom}(\Omega', \Omega) \times F(\Omega')$$

Part III - Random Variables, a Computer Science view

Question

How much of Alex' story works in a general Markov category $\mathbb{C}$? What is required for random variables to form a sheaf? If $\mathbb{C} = \mathcal{K}l(T)$, how are M and T related?

Part III - Random Variables, a Computer Science view

Question

How much of Alex' story works in a general Markov category $\mathbb{C}$? What is required for random variables to form a sheaf? If $\mathbb{C} = \mathcal{K}l(T)$, how are M and T related?

A *sample space* in $\mathbb{C}$ is a pair (Ω, p) of an object Ω and a distribution $p : I \to \Omega$. A morphism $(X, p) \to (Y, q)$ is an equivalence class of morphisms $[f]_p : X \to Y$ which are p-almost surely deterministic and measure-preserving ($f \circ p = q$).

We always have a separated presheaf

$$RV(X) : \mathbf{SamSp}(\mathbb{C})^{\mathrm{op}} \to \mathbf{Set}, \mathrm{RV}(X)(\Omega, p) = \{[f]_p : \Omega \to X \text{ } p\text{-a.s. det.}\}$$

When is this a sheaf?

Part III - Gaussian Random Variables and Nominal Sets

Testing this framework for simple Markov categories already gives very interesting results.

Example

Let $\mathbb{C} = \mathbf{Gauss}$. Then sample spaces are of the form $\mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n \times n}$ positive semidefinite. Morphisms are measure-preserving affine-linear maps.

1. We have $\mathcal{N}(\mu, \Sigma) \cong \mathcal{N}(0, I_k)$ where $k = \text{rank}(\Sigma)$.
2. it remains to classify the measure-preserving maps $\mathcal{N}(0, I_k) \to \mathcal{N}(0, I_n)$. Those are given by the co-isometries $A \in \mathbb{R}^{n \times k}$ with $AA^T = I_n$.
3. Gaussian random variables can be treated in the topos of sheaves $\mathbf{Iso} \to \mathbf{Set}$

The analogy with the Schanuel topos (nominal sets) are striking! Those consists of sheaves $\mathbf{Inj} \to \mathbf{Set}$, motivated by similar symmetry considerations.

Take home message

TL;DR

1. convex bifunctions can be seen as an exotic form of probability
2. for Gaussians, the probabilistic and convex perspective are equivalent, giving a functor of CD categories
$$\mathbf{Gauss} \to \mathbf{CxBiFn}$$
3. co-partiality helps combining probability and nondeterminism
4. The universal property of Gaussians is their rotation invariance
5. Random variables work much like local state
6. useful in implementations `https://github.com/damast93/GaussianInfer`

Thank you!