Characterizing relative entropy on standard Borel spaces categorically

Nicolas Gagné and Prakash Panangaden
School of Computer Science
McGill University

Universidade do Minho
Braga
26th May 2017

Motivation

- Exciting new developments in understanding Bayesian inversion: Danos, Garnier, Dahlqvist, Clerc.
- Much more sophisticated understanding of categorical probability on Borel spaces.
- Theoretical underpinnings of learning.
- Categorical characterization of relative entropy for distributions on finite sets: Baez, Fritz, Leinster.
- Entropy plays a crucial role in rate of convergence of learning processes.
- First step: extend categorical characterization of relative entropy to a more general class of spaces: standard Borel spaces.

Summary

- Background: standard Borel spaces, Giry monad, $\mathbf{SRel}$, disintegration
- Categorical setting
- Relative entropy as a functor
- Uniqueness

Polish spaces and standard Borel spaces

- Basic definitions of measure theory do not mention topology but everything works best when the σ-algebra comes from the topology of a metric space.
- A Polish space is the topological space underlying a complete separable metric space.
- Start with a metric space as above and forget the metric but remember the topology.
- Note, a space like $(0, 1)$ is Polish even though it is not complete in its "usual" metric. It can be given a complete metric and is homeomorphic to $(0, \infty)$.
- A standard Borel space: take a Polish space, forget the topology but remember the Borel sets.

Categories

- $\mathbf{Pol}$: Objects Polish spaces, morphisms are *continuous functions*.
- $\mathbf{StBor}$: Objects standard Borel spaces, morphisms are *measurable functions*.
- Obvious forgetful functor $\mathbb{U} : \mathbf{Pol} \longrightarrow \mathbf{StBor}$ is not full.

The Giry monad on $\mathbf{Mes}$ I

- Her name is actually Gìry; but I will just write Giry.
- Actually proposed by Lawvere in 1964 in an unpublished manuscript.
- $\mathbf{Mes}$: Objects are sets equipped with σ-algebra (X, σ), morphisms are *measurable* functions.
- $\Gamma : \mathbf{Mes} \longrightarrow \mathbf{Mes}$ $\gamma((X, \Sigma)) = \{p | p : \Sigma \longrightarrow [0, 1]\}$; here p is a probability measure.
- For $A \in \Sigma$ define $ev_A : \Gamma(X) \longrightarrow [0, 1]$ by $ev_A(p) = p(A)$.
- Give $\Gamma(X)$ the smallest σ-algebra that makes all the ev_A measurable.
- $f : (X, \Sigma) \longrightarrow (Y, \Lambda)$ maps to $\Gamma(f) : \Gamma(X) \longrightarrow \Gamma(Y)$ by $\Gamma(f)(p)(B \in \Lambda) = p(f^{-1}(B))$.

The Giry monad on $\mathbf{Mes}$ II

- δ_x is the Dirac measure at x or point mass: $\delta_x(A) = 1$ if $x \in A$ and 0 if $x \notin A$.
- $\eta_X : X \longrightarrow \Gamma(X)$ is given by $\eta_X(x) = \delta_x$.
- $\mu_X : \Gamma^2(X) \longrightarrow \Gamma(X)$ is given by

$$\mu_X(\Omega)(A) = \int_{\Gamma(X)} ev_A d\Omega.$$

- This gives the "average" measure of A using Ω to do the weighting.
- Equations have to be checked; they all follow from the monotone convergence theorem.

The Giry monad on **Pol**

- $\mathcal{G} : \mathbf{Pol} \longrightarrow \mathbf{Pol}$. $\mathcal{G}(X)$ is the set of Borel probability measures on X. We need to make it into a topological space.
- The *weak* topology on $\mathcal{G}(X)$: given by an explicit base of open sets. Basic open neighbourhood of p:

  $$B_{f_1,\ldots,f_n;\varepsilon_1,\ldots,\varepsilon_n} := \{q :\in \mathcal{G}(X) : |\int f_i dp - \int f_i dq| < \varepsilon_i, \quad i = 1, \ldots n\}$$

  where f_i are bounded continuous functions and ε_i are positive real numbers.
- $p_n \Rightarrow p$ if for any bounded continuous function f, $\int f dp_n \longrightarrow \int f dp$. Weak convergence.
- The integrals are what one can "see" about a measure.
- Same η, μ as for Γ. One has to check that they are continuous functions now.

Relating $\mathcal{G}$ and Γ

Let B be the forgetful functor from $\mathbf{Pol}$ to $\mathbf{StBor}$. It is clearly faithful. There is a natural transformation $\theta : \mathrm{B} \circ \mathcal{G} \longrightarrow \Gamma \circ \mathrm{B}$.

$$\mathcal{G} \circlearrowright \mathbf{Pol} \xrightarrow{\mathrm{B}} \mathbf{StBor} \circlearrowleft \Gamma$$

The proof is not obvious; it follows from Theorem 17.24 of *Classical Descriptive Set Theory* by Kechris.

The Kleisli category of a monad

- For a monad $T : \mathcal{C} \longrightarrow \mathcal{C}$, we define a new category $\mathcal{C}_T$ with the same objects as $\mathcal{C}$.
- A morphism $f : A \longrightarrow B$ in $\mathcal{C}_T$ is a morphism $f : A \longrightarrow TB$ in $\mathcal{C}$.
- Compose $A \xrightarrow{f} B \xrightarrow{g} C$ in $\mathcal{C}_T$ by composing

$$A \xrightarrow{f} TB \xrightarrow{Tg} T^2C \xrightarrow{\mu_C} TC \text{ in } \mathcal{C}.$$

- One can think of this as a category of "free" algebras.

The Kleisli category of Γ

- Morphisms $h : (X, \Sigma) \longrightarrow (Y, \Lambda)$ are measurable functions $h : X \longrightarrow \Gamma(Y)$.
- But $\Gamma(Y)$ is $\Lambda \longrightarrow [0, 1]$ (with some conditions).
- So $h : X \times \Lambda \longrightarrow [0, 1]$ with $h(\cdot, B)$ a measurable function and $h(x, \cdot)$ a measure. Markov kernels.
- Composition $X \xrightarrow{h} Y \xrightarrow{k} Z$ is, in terms of kernels

$$(k \circ h)(x, C \subset Z) = \int_Y k(y, C) dh(x, \cdot).$$

- Probabilistic relations, composing by integration.
- Infinite-dimensional matrix multiplication.
- This is what Lawvere defined in 1964: probabilistic mappings.

Some notation

- We write $\tilde{\circ}$ for Kleisli composition.
- $(k \,\tilde{\circ}\, h)(x, C \subset Z) = \int_Y k(y, C) dh(x, \cdot)$.
- A measure on (X, Σ) can be viewed as a Kleisli arrow from the one-point space $\mathbf{1} = \{\star\}$ to (X, Σ).
- If $s : Y \longrightarrow \Gamma(X)$ and $q : \mathbf{1} \longrightarrow \Gamma(Y)$ we have

$$(s \,\tilde{\circ}\, q)(\star)(A \in \Sigma) = \int_Y s(y, A) dq.$$

- This is just a measure on (X, Σ).
- I am going to write X instead of (X, Σ) henceforth, unless it is really necessary to emphasize the Σ.

Radon-Nikodym theorem

- If we have two measures p, q on a measurable space X we say q is absolutely continuous with respect to p, if for any measurable set A, $p(A) = 0$ implies that $q(A) = 0$. Notation: $q \ll p$.

Given a measurable space (X, Σ) if a (σ-)finite measure q is absolutely continuous with respect to a (σ-)finite measure p on (X, Σ), then there is a measurable function $f : X \to [0, \infty)$, such that for any measurable subset $A \subset X$, $q(A) = \int_A f \, dp$.

- The function f is unique up to a p-null set and is called the *Radon-Nikodym derivative*, denoted by $\frac{dq}{dp}$.

Disintegration

Disintegration

Let (X, Σ, p) and (Y, Λ, q) be two standard Borel spaces with probability measures, where q is $q := p \circ f^{-1}$ and f is measurable $f : X \to Y$. Then, there exists a family of probability measures $\{p_y\}_{y \in Y}$ on X such that

(i) the function $y \mapsto p_y(B)$ is measurable for each $B \subset X$;

(ii) the fiber $f^{-1}(y)$ has p_y-measure 1: for q-almost all $y \in Y$;

(iii) for every Borel-measurable function $h : X \to [0, \infty]$,

$$\int_X h \, dp = \int_Y \int_{f^{-1}(y)} h \, dp_y dq.$$

Kernels from disintegration

- Write $p_y(\cdot) : \Lambda \longrightarrow [0,1]$ as $p(y,\cdot)$
- then view $p : X \times \Lambda \longrightarrow [0,1]$.
- Such a p is measurable in its first argument and a measure in its second. It is exactly a kernel.
- We will write p_y or $p(y,\cdot)$ or $p(y)$ as we find convenient.

Overview

- We follow Baez and Fritz's approach first with finite sets.
- We generalize to standard Borel spaces.
- We use the Baez-Fritz result as a major building block.
- Our work does not diminish or replace their work.

Coherence

- (X, σ, p), (Y, Λ, q) standard Borel spaces equipped with probability measures.
- A *pair* (f, s) with $f : X \longrightarrow Y$ and $s : Y \longrightarrow \Gamma(X)$ is said to be **coherent** if
- (i) f is measure preserving, *i.e.* $q = p \circ f^{-1}$, and
- (ii) $s(y)(f^{-1}(y)) = 1$. [Support condition]
- (iii) If, in addition, $p \ll s \,\tilde{\circ}\, q$, we say that (f, s) is *absolutely coherent*.

The category **FinStat**

- **Objects** : Pairs (X, p) where X is a finite set and p a probability measure on X.
- **Morphisms** : $\mathrm{Hom}(X, Y)$ are all coherent pairs (f, s), $f : X \to Y$ and $s : Y \to \Gamma(X)$.
- We compose arrows $(f, s) : (X, p) \to (Y, q)$ and $(g, t) : (Y, q) \to (Z, m)$ as follows: $(g, t) \circ (f, s) := (g \circ f, s \,\tilde{\circ}_{fin}\, t)$ where $\tilde{\circ}_{fin}$ is defined as

$$(s \,\tilde{\circ}_{fin}\, t)_z(x) = \sum_{y \in Y} t_z(y) s_y(x).$$

What it means

- We think of X as the space we are investigating and Y as the space of observations. f is the observation map and s then describes what we think the distribution over X is given our observation.
- We say that a hypothesis s is *optimal* if $p = s \,\tilde{\circ}_{\mathtt{fin}}\, q$, or equivalenty, if s is a disintegration of p along f .
- We denote by **FP** the subcategory of $\mathbf{FinStat}$ consisting of the same objects, but with only those morphisms where the hypothesis is optimal.

The category SbStat

- **Objects** : Pairs (X, p) where X is a standard Borel space and p a probability measure on the Borel subsets of X.
- **Morphisms** : $\mathrm{Hom}(X, Y)$ are all coherent pairs (f, s), $f : X \to Y$ and $s : Y \to \Gamma(X)$.
- We compose arrows $(f, s) : (X, p) \to (Y, q)$ and $(g, t) : (Y, q) \to (Z, m)$ as follows: $(g, t) \circ (f, s) := (g \circ f, s \,\tilde{\circ}\, t)$.

A graphical notation

$$
\begin{tikzcd}
(X,p) \arrow[r, bend right, "f"'] & (Y,q) \arrow[l, squiggly, bend right, "s"'] \arrow[r, bend right, "g"'] & (Z,m) \arrow[l, squiggly, bend right, "t"'] \arrow[r, Rightarrow, "\text{Composition}"'] & (X,p) \arrow[r, bend right, "g \circ f"'] & (Z,m) \arrow[l, squiggly, bend right, "s \,\tilde{\circ}\, t"']
\end{tikzcd}
$$

Basic facts

Given coherent pairs the composition is coherent. If, in addition, they are absolutely coherent, the composition is absolutely coherent.

Lawvere's amazing category $[0, \infty]$

- **Objects** : One single object: $\bullet$.
- **Morphisms** : For each element $r \in [0, \infty]$, one arrow $r : \bullet \to \bullet$.
- Arrow composition is defined as addition in $[0, \infty]$.
- This is a remarkable category with monoidal closed structure and many other interesting properties.

Entropy

- Given a probability distribution p on a finite set X, the **entropy** of p is

$$H(p) = -\sum_{x \in X} p(x) \ln p(x).$$

- In computer science we usually use $\log_2$ to count bits; it only changes an overall multiplicative factor.
- If we are transmitting information about the outcome of a process that produces a result in X with distribution p, the *optimal code* will use $H(p)$ *expected* number of nits (bits).
- We always assume $0 \cdot \infty = 0$.

Relative entropy

- If we have two distributions p, q the relative entropy between them is $KL(p, q) = -\sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$.
- Often called the *Kullback-Leibler divergence*.
- It is always positive (Jensen).
- It is not symmetric and does not satisfy the triangle inequality.
- If you design your optimal code thinking that the correct distribution is q when in fact it is p, $KL(p, q)$ measures how many extra nits (bits) you will need.

Bayesian inference

- Let $X = \{1, 2, \ldots, n\}$ be a finite set of outcomes.
- $\mathcal{G}(X)$ is the simplex $\Delta^{(n-1)}$. We want to estimate an unknown distribution p over X by taking samples and updating out prior beliefs.
- The prior belief is a distribution over $\mathcal{G}(X)$ *i.e.* an element of $\mathcal{G}^2(X)$; say μ.
- After observing N samples we want to update μ. We use Bayes' theorem.
- We denote by q the empirical distribution obtained by sampling.

Bayesian updating

Bayes

$$\mu(p|q) = \frac{\overbrace{\mu(p)}^{\text{Prior}} \cdot \overbrace{\mu(q|p)}^{\text{Likelihood}}}{\underbrace{\mu(q)}_{\text{Normalizing}}}.$$

The role of relative entropy

The crucial quantity is the likelihood. How does it grow with N?

Likelihood growth

$$\mu(q|p) \approx e^{-N \cdot RE(q,p)}.$$

The relative entropy controls the rate of convergence of the learning process.

Relative entropy on $\mathbf{FinStat}$

- The functor RE is from $\mathbf{FinStat}$ to $[0, \infty]$.
- **Objects** : Maps every object (X, p) to $\bullet$.
- **Morphisms** : Maps a morphism $(f, s) : (X, p) \to (Y, q)$ to $S_{\mathtt{fin}}(p, s \,\tilde{\circ}_{\mathtt{fin}}\, q)$,
- where

$$S_{\mathtt{fin}}(p, s \,\tilde{\circ}_{\mathtt{fin}}\, q) := \sum_{x \in X} p(x) \ln\left(\frac{p(x)}{(s \,\tilde{\circ}_{\mathtt{fin}}\, q)(x)}\right).$$

Relative entropy on **SbStat**

- Again the target is $[0, \infty]$.
- **Objects** : Maps every object (X, p) to $\bullet$.
- **Morphisms** : Maps every absolutely coherent morphism to $(f, s) : (X, p) \to (Y, q)$ to $S(p, s \,\tilde{\circ}\, q)$, where

$$S(p, s \,\tilde{\circ}\, q) := \int_X \log\left(\frac{dp}{d(s \,\tilde{\circ}\, q)}\right) dp,$$

where $\frac{dp}{d(s \,\tilde{\circ}\, q)}$ is the Radon-Nikodym derivative and otherwise maps to ∞.

Prop: RE is indeed a functor

$$\mathrm{RE}\left(\begin{tikzcd} (X,p) \arrow[r, bend right, "f"'] & (Y,q) \arrow[l, bend right, squiggly, "s"'] \arrow[r, bend right, "g"'] & (Z,m) \arrow[l, bend right, squiggly, "t"'] \end{tikzcd}\right) =$$

$$\mathrm{RE}\left(\begin{tikzcd} (X,p) \arrow[r, bend right, "f"'] & (Y,q) \arrow[l, bend right, squiggly, "s"'] \end{tikzcd}\right) + \mathrm{RE}\left(\begin{tikzcd} (Y,q) \arrow[r, bend right, "g"'] & (Z,m) \arrow[l, bend right, squiggly, "t"'] \end{tikzcd}\right).$$

Quite long, with some lemmas and calculations and tedious case analyses.

Localization of relative entropy

- Given an arrow $(f, s) : (X, p) \to (Y, q)$ in $\mathbf{StBor}$ and a point $y \in Y$, we denote by $(f, s)_y$, the morphism (f, s) restricted to the pair of standard Borel spaces $f^{-1}(y)$ and $\{y\}$.
- Equivalently,

$$(f, s)_y := (\, f|_{f^{-1}(y)}\, , s_y) : (f^{-1}(y), p_y) \longrightarrow (\{y\}, \delta_y),$$

  where δ_y is the unique probability measure on $\{y\}$.
- $(f, s)_y$ is the *local relative entropy* of (f, s) at y.
-

$$RE((f, s)_y) = \begin{cases} \int_{f^{-1}(y)} \log \left(\frac{dp_y}{d(s \,\tilde{\circ}\, q)_y} \right) \, dp_y & \text{if } p_y \ll (s \,\tilde{\circ}\, q)_y \\ \infty & \text{otherwise.} \end{cases}$$

Convexity

Definition

A functor F from $\mathbf{SbStat}$ to $[0, \infty]$ is *convex linear* if for every arrow $(f, s) : (X, p) \to (Y, q)$, we have

$$F((f, s)) = \int_Y F\left((f, s)_y\right) \, dq.$$

Theorem

RE is convex linear, i.e., for every arrow $(f, s) : (X, p) \to (Y, q)$, we have

$$RE((f, s)) = \int_Y RE\left((f, s)_y\right) \, dq.$$

Lower-semicontinuity

Definition

A functor F from $\mathbf{SbStat}$ to $[0, \infty]$ is *lower semi-continuous* if for every arrow $(f, s) : (X, p) \to (\{y\}, \delta_y)$ there is an admissible topology on X such that whenever $p_n \Rightarrow p$ and $s_n \Rightarrow s$, then

$$F\left(\begin{tikzcd}(X,p) \arrow[rr, bend right, "f"'] & & (\{y\},\delta_y) \arrow[ll, bend right, squiggly, "s"']\end{tikzcd}\right) \leqslant \liminf_{n\to\infty} F\left(\begin{tikzcd}(X,p_n) \arrow[rr, bend right, "f"'] & & (\{y\},\delta_y) \arrow[ll, bend right, squiggly, "s_n"']\end{tikzcd}\right).$$

Note the awkwardness of dealing with topological and measure-theoretic issues.

Theorem

RE is indeed lower-semicontinuous.

Follows from well-known results in information theory [Posner].

Baez and Fritz's theorem

Theorem

Suppose that a functor

$$F : \mathbf{FinStat} \to [0, \infty]$$

is lower semicontinuous, convex linear and vanishes on $\mathbf{FP}$. Then for some $0 \leqslant c \leqslant \infty$ we have $F(f, s) = cRE_{fin}(f, s)$ for all morphisms (f, s) in $\mathbf{FinStat}$.

Our theorem

Theorem

Suppose that a functor

$$F : \mathbf{SbStat} \to [0, \infty]$$

is lower semicontinuous, convex linear and vanishes on $\mathbf{FP}$. Then for some $0 \leqslant c \leqslant \infty$ we have $F(f, s) = cRE(f, s)$ for all morphisms.

Proof ideas: Our RE restricts to $\mathbf{FinStat}$ with the properties required for the Baez-Fritz theorem. We exploit density of finitely-supported measures, some tricky carving up of sets (mimicking known ideas), weak convergence and lower-semicontinuity.

Conclusions

- Functorial characterization of relative entropy on standard Borel spaces.
- Hope to link up with the theory of Bayesian inversion on such spaces.
- Ultimately hope to connect with learning.