Bayesian Brains, Cyber-categorically

Toby Smithe

6 June '23

Overview

- Intro: Bayesian brain; categories; cybernetics.

- Compositional probability and inference
  - ↳ Exact inference: Bayesian lenses
  - ↳ Approximate inference: statistical games

- Compositional dynamics
  - ↳ A category of interfaces
  - ↳ Systems on an interface

- Predictive coding and active inference
  - ↳ + societies of agents; universality of the FEP.

Free Energy and the Bayesian brain

A prominent idea in computational neuroscience is that much of what the brain does is "approximate Bayesian inference".

One method: minimize 'free energy'

(an upper bound on 'divergence'/'relative entropy')

→ Leads to influential 'predictive coding' models of visual cortex dynamics

These models have an intriguing hint of compositionality! ...

But the "free energy principle" itself is much more ambitious, and can be mathematically impenetrable ...

Why category theory?

Category theory is the mathematics of
composition, pattern, interconnection,
translation, metaphor, ...

and promises to supply a lingua franca for
science and computation (and more besides).

Each category is like a mathematical 'universe' or 'library',
and categories collect into a category of categories, Cat,
so we can translate from one universe to another
(or: connect libraries together).

Working categorically enforces discipline:
clear thinking; "carve nature at its joints."

What is a category? (briefly!)

A category $\mathbb{C}$ is a collection of 'objects' $\mathbb{C}_0$, and, for each pair of objects a, b, a set $\mathbb{C}(a, b)$ of 'morphisms' $a \to b$:

"ways of relating a to b"

such that morphisms compose associatively

$$\begin{tikzcd} & \bullet \arrow[rr, bend left, "h\circ g"] \arrow[dr, "g"] & & \bullet \\ \bullet \arrow[ur, "f"] \arrow[rr, bend right, "g\circ f"'] & & \bullet \arrow[ur, "h"'] & \end{tikzcd}$$

$$(h\circ g)\circ f = h\circ(g\circ f)$$

→ $id_b \circ f = f = f \circ id_a$ ('unitality')

and each object a has an identity morphism $id_a : a \to a$

Fundamental theorem ('Yoneda Lemma'):

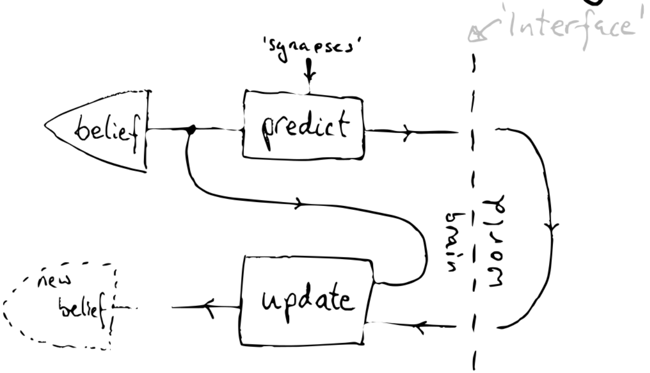
"You shall know an object by the company it keeps!"

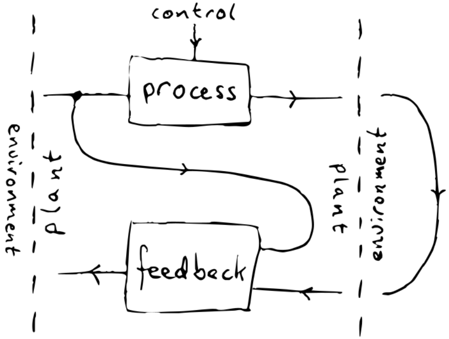
Categorical Cybernetics

The brain is an ideal test for categorical modelling.

└ A complex system made of complex systems, with 'computational' behaviour, and whose multitudes constitute societies.

It is also an exemplary cybernetic system.

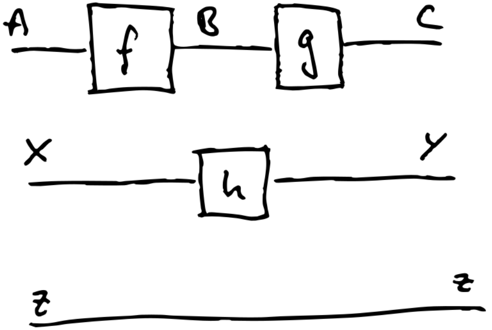




I'm interested in adaptive systems (of systems...) generally, and this is where compositional methods have most promise.

String diagrams

Not only can we use category theory to reason about the visual cortex: we can also use the visual cortex to reason about categories!

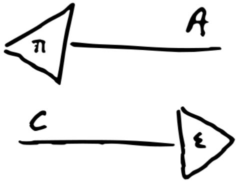


$g \circ f$

$\otimes \quad : A \otimes X \longrightarrow C \otimes Y$

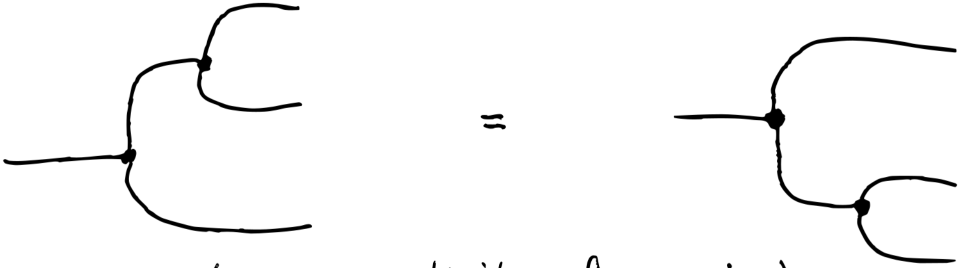
h

id_Z



$\pi : I \longrightarrow A$

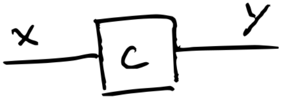
$\varepsilon : C \longrightarrow I$



(coassociativity of copying)

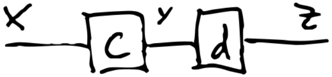
Markov Kernels

One category in which we can interpret these diagrams is that of Markov kernels (~conditional probability distributions).



depicts

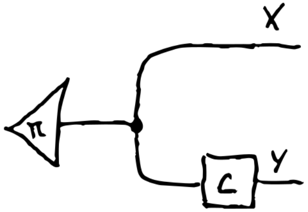
$$c : X \times \Sigma_Y \longrightarrow [0, \infty)$$
$$(x, B) \longmapsto c(B|x)$$



depicts

$$(x, C) \longmapsto \int_{y:Y} d(C|y)\, c(dy|x)$$

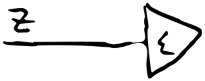
"Chapman - Kolmogorov"



depicts

$$\Sigma_X \times \Sigma_Y \longrightarrow [0, \infty)$$
$$(A, B) \longmapsto \int_{x:A} c(B|x)\pi(dx)$$

or: $\rho_c(y|x)\rho_\pi(x)$

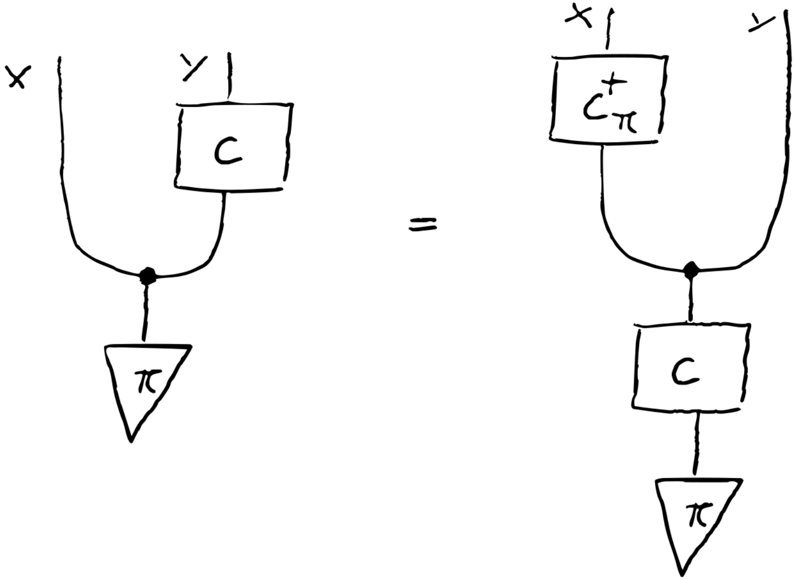


depicts

$$\varepsilon : Z \longrightarrow [0, \infty)$$

(a "fuzzy predicate")

Bayesian inversion



$c : X \mathrel{-\!\!\bullet\!\!\!\to} Y$

$c^\dagger_\pi : Y \mathrel{-\!\!\bullet\!\!\!\to} X$

$$p_c(y|x)\,p_\pi(x) \quad = \quad p_{c^\dagger_\pi}(x|y)\,p_{c\cdot\pi}(y)$$

often written $p_{c^\dagger_\pi}(x|y) \quad = \quad \dfrac{p_c(y|x)\,p_\pi(x)}{p_{c\cdot\pi}(y)}$

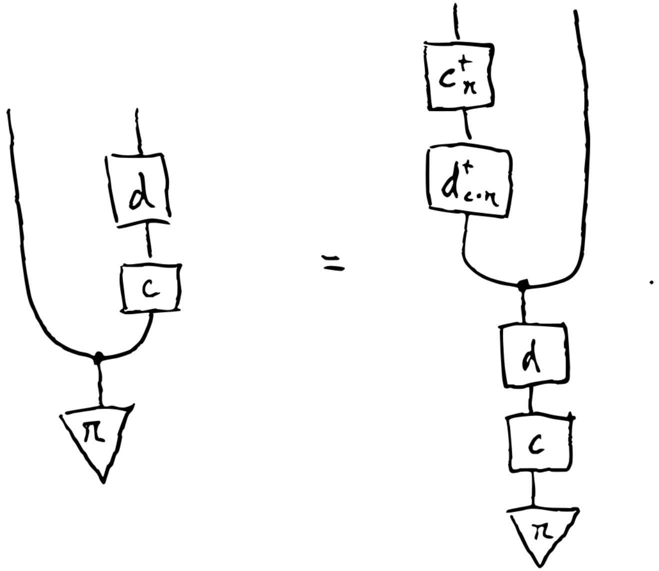
Composite Bayesian Inversion

L 'Hierarchical'

Given $X \xrightarrow{c} Y \xrightarrow{d} Z$, we may wonder what its inversion is: does it relate to $c^\dagger$ and $d^\dagger$ themselves?

(This is relevant in many contexts!)

Using two applications of Bayes' law, we can show



How can we package this up neatly?

(What is $d^\dagger_{c\cdot\pi}$?)

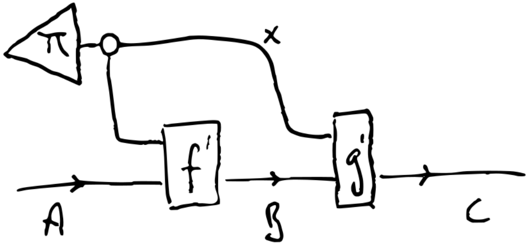
State-dependent channels

$c^+_\pi : Y \to X$ depends on $\pi : I \to X$.

Really, we have $c^+ : \mathcal{C}(I, X) \longrightarrow \mathcal{C}(Y, X)$ (in Set).

$\pi \longmapsto c^+_\pi$

For each X, we have a whole category, $\mathrm{Stat}(X)$:

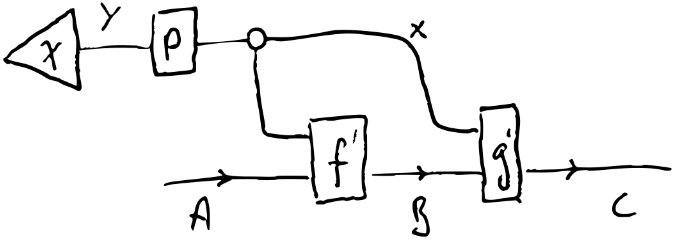


$\mathrm{Stat}(X)(A, B)$

$= \mathrm{Set}(\mathcal{C}(I, X), \mathcal{C}(A, B))$

We can 're-index' by pre-composition.

Given $\rho : Y \to X$,



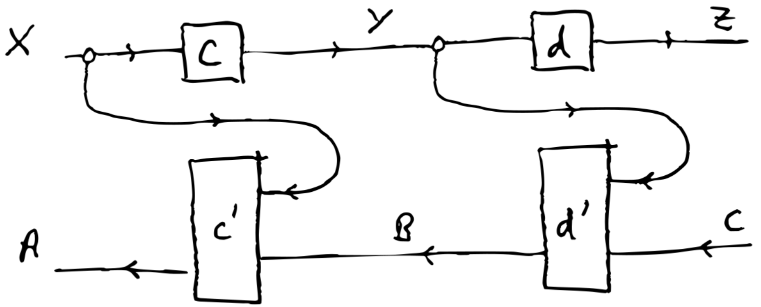
i.e., $c^+_\pi \overset{\rho^*}{\longmapsto} c^+_{\rho \cdot \pi}$

Bayesian Lenses

A Bayesian lens is a pairing of a kernel with a (state-dependent) 'inversion'.

These form a category, obtained as the "Grothendieck construction" on Stat.

Composition of Bayesian lenses $(X, A) \xrightarrow{(c, c')} (Y, B) \xrightarrow{(d, d')} (Z, C)$ is thus:



$$(d, d') \circ (c, c')$$

$$= (d \bullet c, \; c'_{(-)} \circ d'_{c \bullet (-)})$$

"Bayesian updates compose optically"

And in fact this is a common cybernetic pattern!

Sections of a Fibration

Bayesian lenses constitute a 'fibration' π

$$\begin{tikzcd}
(c, c') \arrow[d, mapsto] & \mathbf{BayesLens} \arrow[d, "\pi"'] & (c, c^{+}) \\
c & \mathcal{C} \arrow[u, bend right=60, "\sigma"'] & c \arrow[u, mapsto]
\end{tikzcd} .$$

of which exact inversions constitute a 'section'.

└ Functoriality formalizes BUCO: $\sigma(d) \circ \sigma(c) = \sigma(d \circ c)$.

What about the other lenses?

└ These represent approximate inversions.

Sections encode well behaved inference systems.

How good are my predictions?

An arbitrary inference system might not be very good.

Yet evolution has produced some that are quite good!

Can we measure the difference?

└ How do brains improve their predictions?

Note: 1) Neural computation is 'local'.

2) Inferential performance is context-sensitive, possibly depending on prior beliefs and on actual observations.

Example: relative entropy (Kullback-Leibler divergence).

Relative entropy as a local loss function

The relative entropy measures a 'divergence' $D(\alpha, \beta)$ between distributions $\alpha, \beta : I \mathrel{-\!\!\bullet\!\!\to} X$.

Minimizing the divergence $D(c'_\pi(y), c^\dagger_\pi(y))$ makes an approximate posterior $c'_\pi(y)$ as close as possible to the Bayesian posterior $c^\dagger_\pi(y)$.

↳ Observe the 'context-sensitivity': dependence on π and y. Hence we have a family of 'local' loss functions: state-dependent effects $D^{(c,c')} : \underline{\underline{Y \overset{X}{\mathrel{-\!\!\bullet\!\!\to}} I}}$

$$\begin{aligned} \mathcal{C}(I, X) &\longrightarrow \{Y \longrightarrow [0, \infty)\} \\ \pi &\longmapsto y \longmapsto D(c'_\pi(y), c^\dagger_\pi(y)). \end{aligned}$$

Notably, the relative entropy satisfies a chain rule:

$$D^{(d,d')\circ(c,c')}_\pi(z) = \mathop{\mathbb{E}}_{y \sim d'_{c\cdot\pi}(z)} \left[D^{(c,c')}_\pi(y) \right] + D^{(d,d')}_{c\cdot\pi}(z)$$

Statistical Games generalize this pattern.

A statistical game $(c, c'; L^c) : (X, A) \to (Y, B)$ pairs a lens (c, c') with a loss function $L^c : B \xrightarrow{X} I$.

(sometimes 'laxly')

They compose according to the previous pattern:

$$(L^d \circ L^c)_\pi(z) := L^d_{c\bullet\pi}(z) + L^c_\pi \bullet d'_{c\bullet\pi}(z)$$

Again we have a fibration:

$$\begin{array}{ccc} \mathrm{SGame} & \longrightarrow & \mathrm{BayesLens} \\ (c, c', L^c) & \longmapsto & (c, c') \end{array}$$

A loss model is a (possibly lax) section of this fibration.

Examples: relative entropy,

$\text{lax} \left\{ \begin{array}{l} \text{maximum likelihood,} \\ \text{free energy,} \\ \text{`Laplacian' free energy,} \\ \ldots \end{array} \right.$

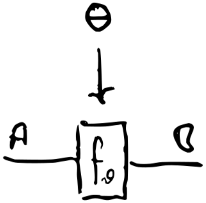
= KL + MLE

Approximates free energy.
Underlies predictive coding!

Learning as parameter-change

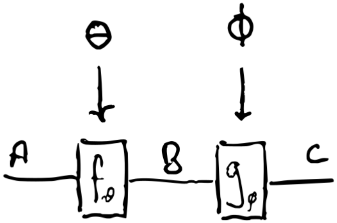
A common approach to modelling learning is to parameterize the process being learnt.

The parameter represents a "choice" of process:



$$\Theta \xrightarrow{f} \mathcal{C}(A, B) \quad \text{in } \mathbf{Set}$$

The story is much as for state-dependence:



$$\begin{tikzcd}
\Phi \times \Theta \arrow[d, "g \times f"] \\
\mathcal{C}(B, C) \times \mathcal{C}(A, B) \arrow[d, "\circ"] \\
\mathcal{C}(A, C)
\end{tikzcd}$$

Learning amounts to improving the 'choice'.

We can parameterize statistical games, and optimize the loss functions with respect to the parameters.

This is how predictive coding models are obtained!

Recap

- Bayesian lenses pair 'predictive' channels with (state-dependent) inversions.

- Exact inversions constitute a 'section' (witnessing BUCO); other lenses are approximate.

- To measure inferential performance, we can pair Bayesian lenses with loss functions → statistical games.

- The chain rule of the relative entropy makes it a loss model (a section of this 2nd fibration).

  The free energy is another loss model.

- To make space for learning, we can introduce parameters.

... The only missing piece now is a way to optimize them!

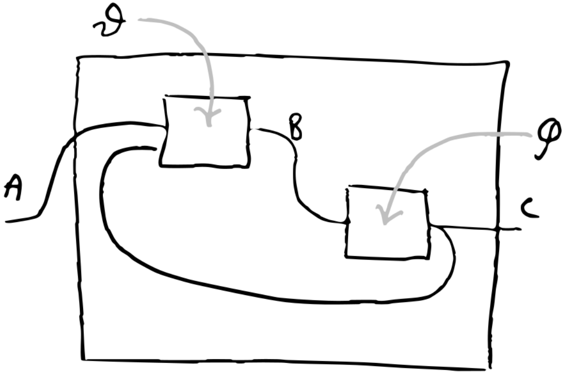
↳ Dynamics ...

Compositional Dynamics

The 'box-filling' idea of parameterized processes is very general:

we can also fill boxes with dynamical systems, treating the boxes as their interfaces.

↳ 'Wiring diagrams' are the morphisms of a category Int, whose objects are interfaces:



$[A \otimes C, B] \otimes [B, C] \longrightarrow [A, C]$

For each interface $[X, Y]$, we have a category $\mathrm{Dyn}([X, Y])$ of dynamical systems on that interface (+ homomorphisms).

∟ So $\vartheta \in \mathrm{Dyn}([A \otimes C, B])$ and $\varphi \in \mathrm{Dyn}([B, C])$.

Dyn is a (monoidal) indexed category: re-indexing is re-wiring!

Cilia

An interface specifies the 'input' and 'output' types of a dynamical system.

We can define an interface type so that the resulting systems 'control' Bayesian lenses — I call these cilia.

$$\left[\!\left[\begin{matrix} X & Y \\ A, & B \end{matrix} \right]\!\right] := \left[\mathcal{C}(I, X) \times B, \ \mathrm{BayesLens}((X, A), (Y, B)) \right]$$

Then we obtain a bicategory by

$$\mathrm{Cilia}((X, A), (Y, B)) := \mathbf{Dyn}\left(\left[\!\left[\begin{matrix} X & Y \\ A, & B \end{matrix} \right]\!\right]\right).$$

A cilium $(X, A) \xrightarrow{\beta} (Y, B)$ has four parts:

(i) A state space S;

(ii) an S-parameterized forwards output kernel $\beta^o_1 : X \xrightarrow{S} Y$; 'predict'

(iii) an S-parameterized inversion $\beta^o_2 : B \xrightarrow{S, X} A$; 'infer'

(iv) an update kernel $\beta^u : B \xrightarrow{S, X} S$. 'learn'

A general recipe for predictive coding

(= local approximate inference)

category of kernels

fibrations

$$\begin{tikzcd}
\mathbf{Para}(\mathcal{C}) \arrow[r, bend right] & \mathbf{Para}(\mathbf{BayesLens}) \arrow[l] \arrow[r, bend right] & \mathbf{Para}(\mathbf{StatGame}) \arrow[l] \arrow[r, bend right] & \mathbf{Cilia}
\end{tikzcd}$$

inference system

loss model

dynamical semantics (optimization)

"Approximate inference doctrines"

What about active inference?

We can use the same 'categorical systems theory' idea:

an agent is embodied (= has an interface / a 'box'),

and its brain is within it (= a choice of box-filler).

(In the literature, this interface is sometimes known as its Markov blanket.)

More formally, for each interface p,

there is a category Agent(p) of agents with that interface.

↳ i.e. 'lenses onto the interface', $X \rightarrow\!\!\!\!\rightarrow \Sigma p$.

Given a 'wiring diagram' $w : p \to q$, we can connect agents accordingly:

$\text{Agent}(w) : \text{Agent}(p) \longrightarrow \text{Agent}(q)$

↳ Think 'getting in the car' or 'corporation'....

NB We can also do all this in dynamical contexts.

Societies of Agents

At this point, we may start to consider how agents themselves compose.

This reveals some important subtleties.

1) 'Wiring' agents into agents requires them to predict "how the environment causes their sense data".

(vs simply the sense data)

This leads to a mathematics of theory of mind.

2) Bayesian models — and this resulting theory of mind — are inherently subjective.

This means that two agents may disagree about how they are wired together!

There are various tools to deal with this:

- cohomology to measure disagreement;
- 'diffusion' to smooth it out;
- or "impose a 'manager'"!

(The choice depends on the situation at hand...)

Universality and the FEP

The free energy principle asserts that all dynamical systems that "maintain their interface" (their 'Markov blanket') can be understood as performing approximate Bayesian inference (~ playing free energy games).

This is a grand claim, and it has not been conclusively established.

Compositional active inference is a framework in which it may be possible to do so:

$$\begin{tikzcd}
\underset{\text{(biology)}}{\text{Agent}} \arrow[rr, bend left, "\text{'body'}"] \arrow[rr, phantom, "\perp"] \arrow[dr] & {} \arrow[d, phantom, "?"] & \underset{\text{(physics)}}{\text{Dyn}} \arrow[ll, bend left, "\text{'mind'}"] \arrow[dl] \\
& \text{Int} &
\end{tikzcd}$$

Fin