(Towards) A Compositional Account of Active Inference and Predictive Processing

Toby Smithe

Topos / Oxford

Cog Sci '22

27 July

Background

- My background in computational neuroscience / philosophy:
  * how do brains learn the structure of environment?
  * how / what do neural circuits 'compute'?
  * how does the world in my head relate to the one "out there"?
  * how should we make sense of a fragmented + underspecified literature?

- 'Free energy' framework seemingly closest we have come to a "unified theory" of intelligent systems, and it seems to have some neural grounding, too!

↳ But it's mathematically quite confusing / unclear, despite having clear signs of compositionality!

... Let's do something about it!

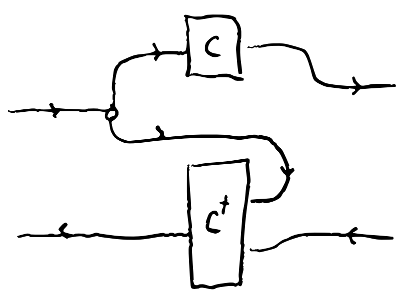
Overview

- Markov categories + Bayes' rule, abstractly
- Bidirectional systems: Bayesian lenses + statistical games
- Open dynamical systems
- Approximate inference doctrines:
  dynamical semantics for inference

Two basic ideas

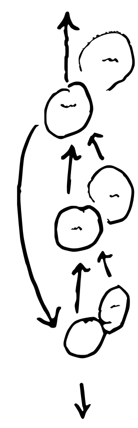
1) What is "active inference", anyway?

2) Can 'hierarchical' neural circuits solve hierarchical inference problems?



Smithe (2020)

$\longmapsto$



Bastos et al (2012)

Categorical probability: distribution monad

We describe a functor $D : \mathrm{Set} \longrightarrow \mathrm{Set}$.

On objects (sets),

$$DX := \{\text{finitely-supported distributions } X \longrightarrow [0,1]\}.$$

On morphisms (functions), $f : X \to Y$,

$$\begin{aligned} Df : DX &\longrightarrow DY \\ \pi &\longmapsto f_*\pi : Y \longrightarrow [0,1] \\ &\qquad\qquad y \longmapsto \sum_{x : f^{-1}(y)} \pi(x) \end{aligned}$$

Monad multiplication

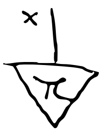
$$\begin{aligned} \mu_X : DDX &\longrightarrow DX \\ \alpha &\longmapsto \Big(x \mapsto \sum_{\pi : DX} \pi(x)\, \alpha(\pi) \Big) \end{aligned}$$

Monad unit

$$\begin{aligned} \eta_X : X &\longrightarrow DX \\ x &\longmapsto \delta_x : x' \longmapsto \begin{cases} 1 & \Leftrightarrow x = x', \\ 0 & \text{else}. \end{cases} \end{aligned}$$

The Markov category $Kl(\mathcal{D})$

(A Markov category is a copy-delete category with natural deleting.)

	Abstractly	Concretely in $Kl(\mathcal{D})$
 Objects	'spaces' $X, Y \ldots$	sets $X, Y, \ldots$
Morphisms	'stochastic channels' $X \to Y$	(finitary) conditional probability dists $c : X \longrightarrow \mathcal{D}Y : x \mapsto p_c(- \mid x)$
 'States'	channels $I \to X$	probability dists $X \longrightarrow [0,1]$

Composition average over intermediate states

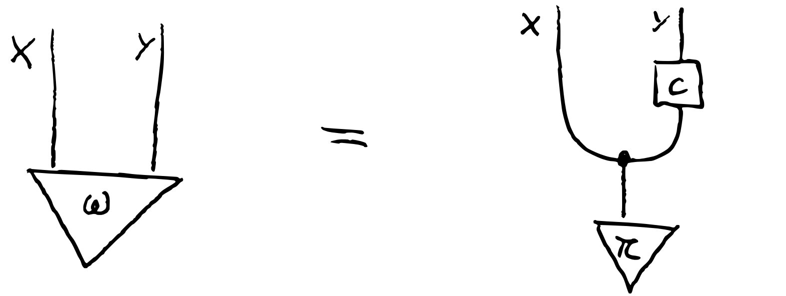
given $c : X \to Y$ and $d : Y \to Z$,

$d \circ c : X \longrightarrow \mathcal{D}Z$ is given by

$$x \longmapsto \sum_{y:Y} p_d(- \mid y) \cdot p_c(y \mid x)$$

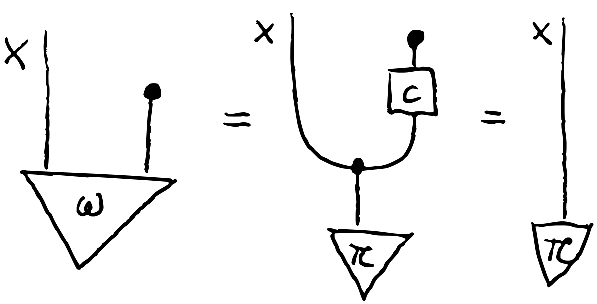
(Kleisli composition: $d \circ c = X \xrightarrow{c} \mathcal{D}Y \xrightarrow{\mathcal{D}d} \mathcal{D}\mathcal{D}Z \xrightarrow{\mu_Z} \mathcal{D}Z$)

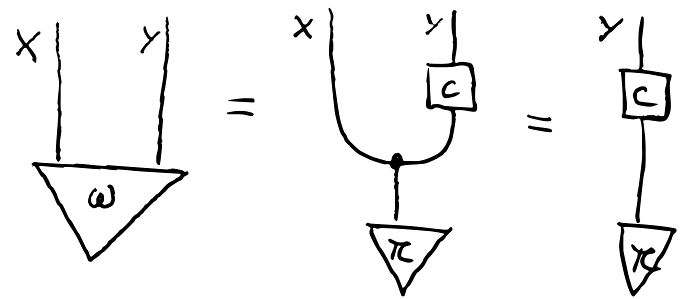
Joint states and 'generative models'



$$p_\omega(x,y) = p_c(y|x)\cdot p_\pi(x)$$

With two marginals given by discarding:

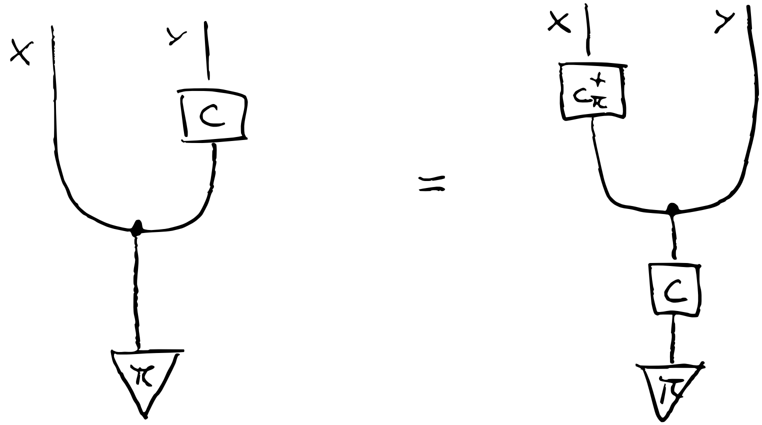




$$\sum_y p_\omega(x,y) = \sum_y p_c(y|x)\,p_\pi(x)$$
$$= p_\pi(x)$$

$$\sum_x p_\omega(x,y) = \sum_x p_c(y|x)\,p_\pi(x)$$
$$= p_{c\circ\pi}(y)$$

Bayes' rule, categorically



$$p_c(y|x)\, p_\pi(x) = p_{c^+_\pi}(x|y)\, p_{c\circ\pi}(y)$$

(when we can divide.)

$$\iff \frac{p_c(y|x)\, p_\pi(x)}{p_{c\circ\pi}(y)} = p_{c^+_\pi}(x|y)$$

Note that the inverse channel $c^+_\pi : Y \rightsquigarrow X$ depends on the prior π!

Bayesian lenses

(w.r.t. π!)

Notice how the channel $c : X \to \mathcal{D}Y$ and its inverse $c^\dagger_\pi : Y \to \mathcal{D}X$ point in opposite directions.

Letting π vary, we obtain a parameterized inversion $c^\dagger : \mathcal{D}X \times Y \to \mathcal{D}X$.

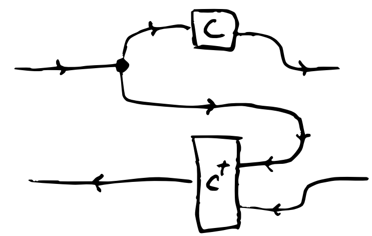
A pair of a 'forwards' morphism

$$c : X \to \mathcal{D}Y$$

and a parameterized 'backwards' morphism

$$c^\dagger : \mathcal{D}X \times Y \to \mathcal{D}X$$

is called a (simple, Bayesian) lens, with type $(X, X) \leftrightarrow (Y, Y)$.

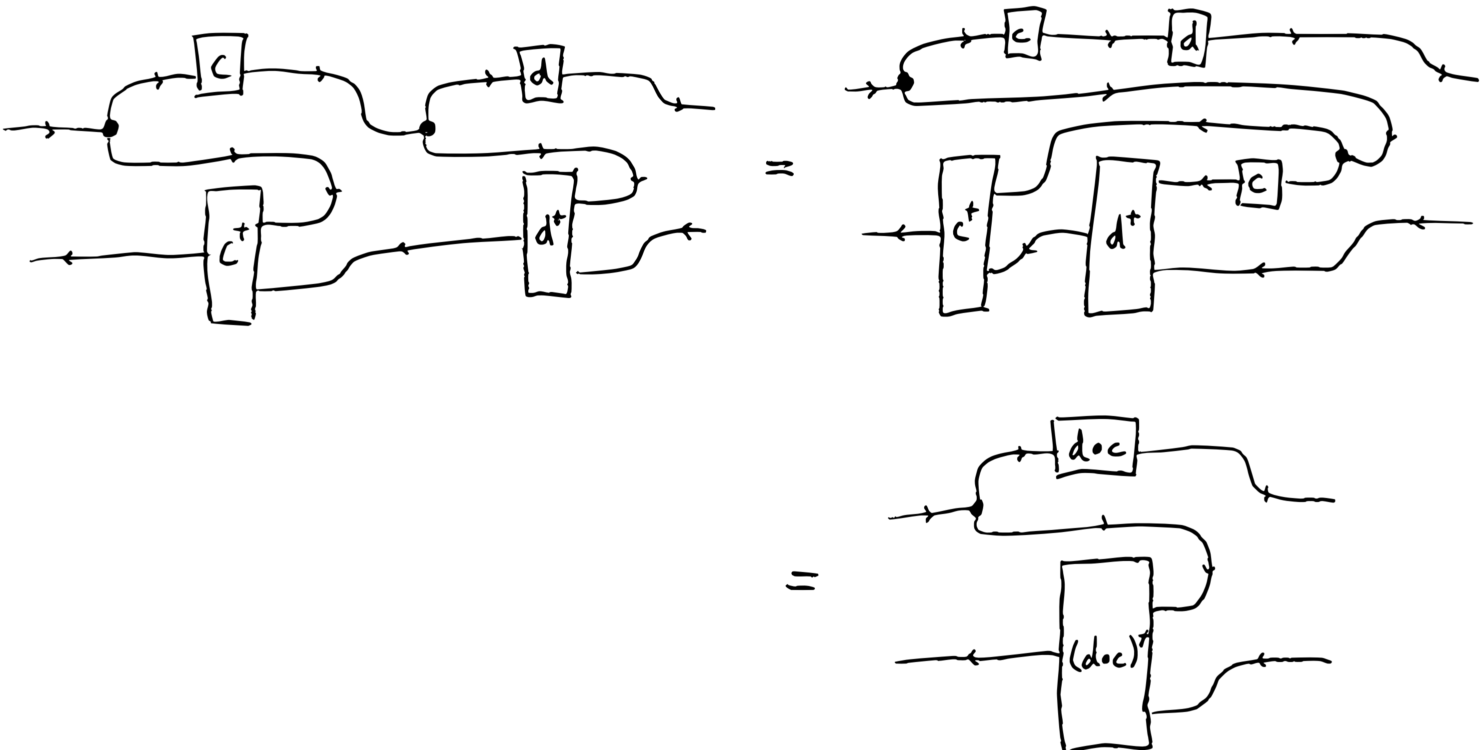


(We've already seen an indication of the prevalence of this pattern!)

* Bayesian lenses formalize abstract statistical inference systems.

Bayesian updates compose optically

Bayesian lenses compose like so:



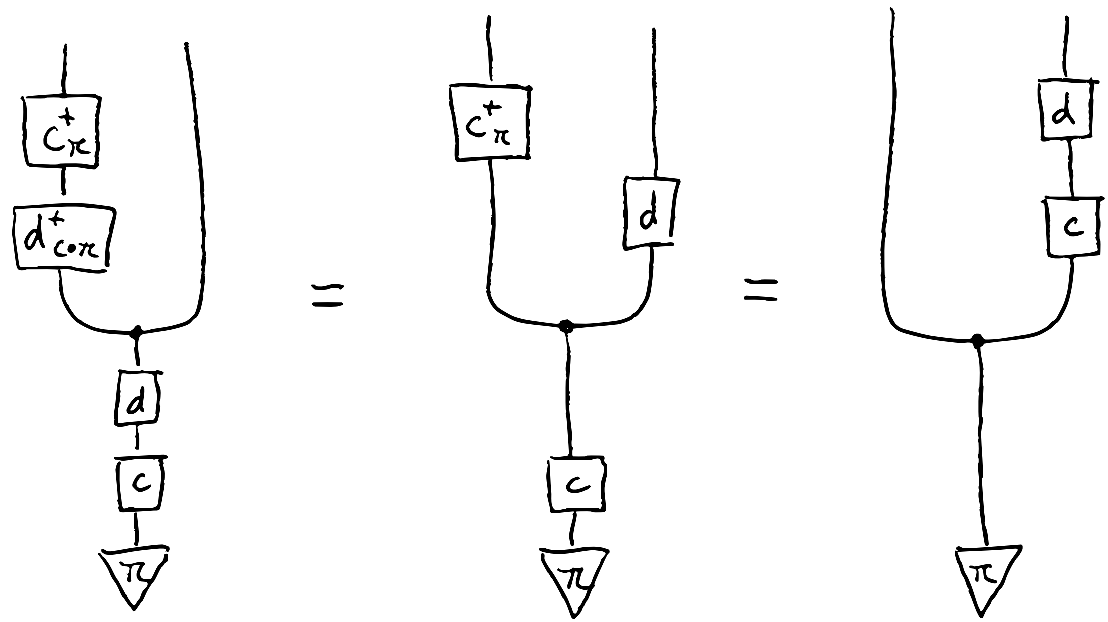
And so we might wonder,

does Bayes' rule imply $(d \circ c)^\dagger_\pi = c^\dagger_\pi \bullet d^\dagger_{c \circ \pi}$?

Bayesian updates compose optically

The answer is "yes"! That is, $(d \circ c)^\dagger_\pi = c^\dagger_\pi \circ d^\dagger_{c\circ\pi}$.

Proof



This means that Bayesian inversions compose according to the "lens rule", and seems to justify abstractly some of the bidirectional hierarchical structure of cortical circuits.

Statistical Games for Approximate Inference

The previous results assume 'exact' inversions, but these are often expensive to compute.

↳ So 'real' systems approximate!

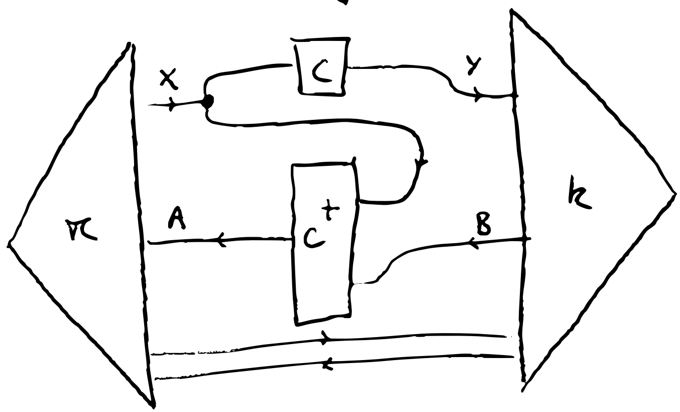
But what makes for a good approximation?

↳ It depends on the context: how the lens interacts w/ the world.
That is, it depends on the prior we choose, and the data we observe.

So we will build a category of statistical games by pairing each (possibly approximate) lens with a contextual fitness function.

Contexts for Bayesian Lenses

A lens is an open system.
A context is everything required to make it 'closed':



$(c, c^{+}) : \binom{X}{A} \to \binom{Y}{B}$

$(\pi, k) : \mathrm{Ctx}(c, c^{+})$

$$\mathrm{Ctx}(c, c^{+}) = \int^{\binom{M}{N}} \mathrm{BayesLens}\left(\binom{1}{1}, \binom{M}{N} \otimes \binom{X}{A}\right) \times \mathrm{BayesLens}\left(\binom{M}{N} \otimes \binom{Y}{B}, \binom{1}{1}\right)$$

When $\binom{M}{N} = \binom{1}{1}$, π obtains the type of a prior $\pi : 1 \to \mathcal{D}X$,
and k becomes a 'continuation' $k : \mathcal{D}Y \to \mathcal{D}B$.

Statistical Games

We define a category SGame of statistical games.

Its objects are pairs (X, A) of sets.

Its morphisms $(X, A) \longrightarrow (Y, B)$
pairs of a lens $(c, c') : (X, A) \to\!\!\!\!\to (Y, B)$ and
a loss function $\phi : Ctx(c, c') \longrightarrow \mathbb{R}$.

(Explain this!)

Composition of statistical games is
lens composition, paired with the sum of 'local' fitnesses.

Identities are given by identity lenses,
and '0' loss functions.

Examples of statistical games

Maximum likelihood estimation $(1,1) \to (X,X)$

Such a lens is determined by a distribution π on X.
The 'prior' part of the context is trivial.
So the MLE loss function is $\phi(k) = \mathbb{E}_{x \sim k(\pi)} \left[p_\pi(x) \right]$.

Bayesian inference $(X,X) \longrightarrow (Y,Y)$

Suppose the lens is $(c,c') : (X,X) \to (Y,Y)$.
The loss function is $\mathbb{E}_{y \sim \langle \pi | c | k \rangle} \left[D_{KL} \left(c'_\pi(y) \,\|\, c^{\dagger}_\pi(y) \right) \right]$.

Autoencoder / free energy $(X,X) \longrightarrow (Y,Y)$

Loss function $\mathbb{E}_{y \sim \langle \pi | c | k \rangle} \left[D_{KL} \left(c'_\pi(y) \,\|\, \pi \right) - \mathbb{E}_{x \sim c'_\pi(y)} \left[\log p_c(y|x) \right] \right]$.

(See arXiv:2109.04461 for more details + examples.)

Lenses for dynamical systems

For a much more general story, see arXiv:2206.03868.

We will use a different category of lenses to define the dynamical systems which will animate our statistical games.

We will call these lenses <u>Markov lenses</u>, and their category MLens.

Its objects will again be pairs of sets.

Its morphisms $(X, A) \longrightarrow (Y, B)$ will be pairs of functions

$$\beta_1 : X \longrightarrow Y \quad \text{and} \quad \beta^\# : X \times B \longrightarrow DA.$$

Composition is lens-like, but we don't need to worry about that!

A discrete-time stochastic dynamical system ϑ with outputs in O and inputs in I and state space S is a Markov lens $(\vartheta^o, \vartheta^u) : (S, S) \longrightarrow (O, I)$; i.e.

$$\vartheta^o : S \longrightarrow O \quad \text{and} \quad \vartheta^u : S \times I \longrightarrow DS.$$

(These things form a category "over (O, I)" with morphisms being maps of state spaces compatible with dynamics; but we won't need this!)

Hierarchical bidirectional stochastic dynamical systems

By choosing the input and output types carefully, we can define a category Hier with the right structure for dynamical approximate inference, and able to capture the proposed structure of predictive coding circuits.

The objects of Hier are (again!) pairs of sets.

The morphisms $(X, A) \longrightarrow (Y, B)$ are
stochastic dynamical systems outputting Bayesian lenses $(X, A) \to (Y, B)$ and inputting elements of $\mathcal{D}X \times B$ (i.e. forwards / backwards inputs).

So if β is such a morphism, it has a state space S and three maps:

$\beta^o_1 : S \times X \longrightarrow \mathcal{D}Y$ a 'forwards output channel',

$\beta^o_2 : S \times \mathcal{D}X \times B \longrightarrow \mathcal{D}A$ a 'backwards output channel', and

$\beta^u : S \times \mathcal{D}X \times B \longrightarrow \mathcal{D}S$ a 'stochastic update channel'.

Composition is by taking the product of the state spaces, composing the output lenses, and "feeding back inputs".

Composing Hier systems (worked example)

(if there's time...)

Approximate Inference Doctrines

An approximate inference doctrine is a functor $\mathcal{Kl}(\mathcal{D}) \longrightarrow \mathcal{Hier}$.

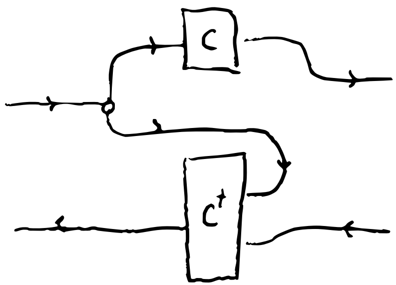
∟ Such a functor takes composite channels
(i.e. hierarchical generative models)
and returns composite dynamical systems that 'invert them'.
(Ideally we want the functor to factor as $\mathcal{Kl}(\mathcal{D}) \to \mathbf{SGame} \to \mathcal{Hier}$).

It turns out that the predictive coding neural circuits
that I opened with arise in this way,
where the first factor picks out statistical games
"under the Laplace approximation",
and the second performs stochastic gradient descent
on the resulting loss functions,
and this explains their latent compositional structure.

Details in a forthcoming paper...

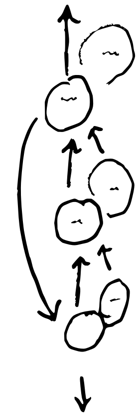
I expect many other approximate inference schemes to fit this pattern.

So we really do have something like:



Smithe (2020)

$\longmapsto$



Bastos et al (2012)

Tentative first steps towards a categorical account of cybernetic biological neural systems...

but there are many missing pieces...
(low-level details, action, planning, "world models")

and we're still rather far from cognition!