Approximate Inference via Fibrations of Statistical Games

Toby St Clere Smithe

VERSES Research
&
Topos Institute

2023 August 4

Natural adaptive systems: the story so far

I'm interested in natural adaptive systems: organisms that perceive and act.

We can think of perception as a process of Bayesian inference:

- an organism has sensors, detecting data in Y;
- it maintains a belief $\beta : I \to\!\!\!\!\bullet\, X$ about the world X,
- and about how the world causes its sense data $c : X \to\!\!\!\!\bullet\, Y$.
- Given these, it updates its beliefs via $c^\sharp_\beta : Y \to\!\!\!\!\bullet\, X$.

Together, $c : X \to\!\!\!\!\bullet\, Y$ and $c^\sharp : \mathcal{C}(I, X) \to \mathcal{C}(Y, X)$ are called a **Bayesian lens** $(c, c^\sharp) : X \nrightarrow Y$. (It's a *bidirectional* process!)

- $\mathcal{C}$ is our ambient (concrete) category of 'channels' $X \to\!\!\!\!\bullet\, Y$.
- Arrows $A \to B$ are functions.
- Lens composition: $(d, d^\sharp) \circledcirc (c, c^\sharp) = (d \bullet c, c^\sharp \circ d^\sharp_c)$.

Exact vs approximate Bayesian lenses

Bayesian lenses: attach 'state-dependent' inversions $c^\sharp$ to channels c.

This gives a fibration $\pi_{\mathsf{Lens}} : \mathbf{BayesLens} \to \mathcal{C}$ by $(c, c^\sharp) \mapsto c$.

Bayes' rule (almost surely [1]) gives a section, $c \mapsto (c, c^\dagger)$.
(BUCO: "Bayesian updates compose optically" [2]).

These lenses 'exactly' satisfy Bayes' law.
But there are lots of other Bayesian lenses!

We can think of non-exact lenses as *approximate* inversions:
useful, because exact inference is computationally hard.

This entails a need to measure performance. Enter statistical games!

Overview of this talk

1 Introduction

2 Statistical games: the basics

3 Relative entropy and copy-composition

4 Loss models

5 Monoidal structures

6 Concluding remarks

Another motivation: *local* inference

Biological systems are composed of cells.

Moving information around is expensive!

(This is where most energy is spent in machine learning.)

So biology wants systems that can be optimized 'locally'.
Hence: compositional inference, plus compositional 'measurement'.

These measurements will be formalized by **loss functions**.

Statistical games: attach losses to lenses

A **statistical game** $X \to Y$ is a triple $(c, c^\sharp, L^c)$,
of which $(c, c^\sharp)$ is a Bayesian lens $X \nrightarrow Y$,
and L^c is an accordingly-typed loss function.

('Game' due to structural similarity to compositional game theory [3, 4].)

This gives a 2nd fibration $\pi_{\mathsf{Loss}} : \mathbf{StatGame} \to \mathbf{BayesLens}$.
(We'll see how loss functions compose later.)

Sections of this fibration capture some important quantities in statistics:

- relative entropy (a.k.a. Kullback-Leibler divergence);
- (log) likelihood;
- free energy (a.k.a. 'ELBO').

Let's see how this works ...

The relative entropy and its chain rule

Relative entropy measures 'divergence' between distributions:
$D_Y : \mathcal{C}(I, Y) \times \mathcal{C}(I, Y) \to \mathbb{R}_+$, with $D_Y(\alpha, \alpha') = 0 \iff \alpha = \alpha'$.

It can be indexed by channels: given parallel $c, c' : X \to\!\!\!\bullet\, Y$, define
$D_{c,c'} : X \to \mathbb{R}_+$ by $D_{c,c'}(x) := D_Y\big(c(x), c'(x)\big)$.

It satisfies a "chain rule" that we can *almost* write as

$$\begin{aligned} D_{d \bullet c, d' \bullet c'}(x) &= D_{c,c'}(x) + \big(c^* D_{d,d'}\big)(x) \\ &= D_{c,c'}(x) + \mathop{\mathbb{E}}_{y \sim c(x)} \big[D_{d,d'}(y)\big] \end{aligned}$$

... but not quite.

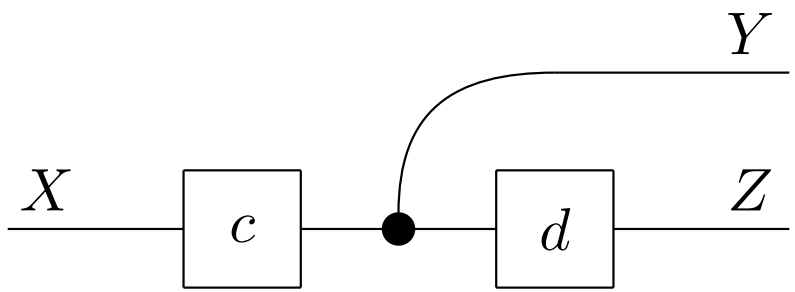
Problem: a failure of compositionality

Relative entropy is the expected difference in log probability:

$$D_Y(\alpha, \alpha') = \mathop{\mathbb{E}}_{y \sim \alpha} \big[\log p_\alpha(y) - \log p_{\alpha'}(y) \big]$$

But $p_{d \bullet c}(z|x) = \mathbb{E}_{y \sim c(x)} \big[p_d(z|y)\, p_c(y|x) \big]$.
And $\log$ doesn't commute with expectations!

On the other hand, if we copy the intermediate variable ...



Solution: copy-composition

... we get a 'copy-composite' channel $d \bullet^2 c : X \to\!\!\!\!\bullet\, Y \otimes Z$
with density $p_{d \bullet^2 c}(y, z|x) = p_d(z|y)\, p_c(y|x)$.

The expectation has gone away, and so our chain rule is satisfied:

$$D_{d \bullet^2 c, d' \bullet^2 c'}(x) = D_{c,c'}(x) + \big(c^* D_{d,d'}\big)(x)$$

Challenge: adjust $\mathcal{C}$ so that composition means *copy*-composition.
Composites need to carry their intermediaries.

A bicategory of copy-composite processes

The trouble is that copy-composition is not strictly unital:
$\mathrm{id}_Y \bullet^2 c$ has type $X \to\!\!\!\!\bullet\, Y \otimes Y$, not $X \to\!\!\!\!\bullet\, Y$!

But we can think of $d \bullet^2 c$ as a Y-coparameterized morphism $X \underset{Y}{\to\!\!\!\!\bullet} Z$,
and adjust the $\mathbf{Copara}$ construction [5] accordingly.

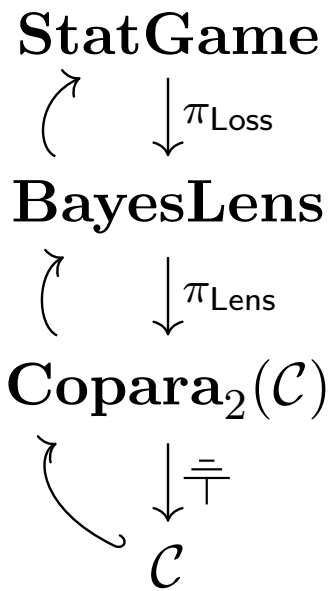
This yields a bicategory, $\mathbf{Copara}_2(\mathcal{C})$.
Horizontal composition is copy-composition.
2-cells are "changes of coparameter".

The structure morphisms deal with the book-keeping,
introducing and deleting copies where necessary.

* Still, I'm not sure this is the best way; I've since discovered others! (See bonus slide...)

Fibrations of copy-composite lenses and games

The picture starts to look like this:



- $\dagger : \mathbf{Copara}_2(\mathcal{C}) \to \mathcal{C}$ discards coparameters.
- There is a canonical (lax) inclusion $\mathcal{C} \hookrightarrow \mathbf{Copara}_2(\mathcal{C})$, yielding a section of $\dagger$.
- Sections of π_{Lens} are **inference systems**, *e.g.* $\dagger$

  (since we have a "coparameterized BUCO" result.)
- Sections of π_{Loss} are **loss models**, *e.g.* relative entropy.

So what *are* loss functions, in general? And how do they compose?

Loss functions, externally and internally

'Externally': a loss function is a map $L^c : \mathcal{C}(I, X) \times Y \to \mathbb{R}_+$, associated to a lens (c, c').

If $\mathcal{C}$ has *bilinear effects**, we can internalize this to a state-dependent effect $L^c : Y \xrightarrow{X} I$ — *i.e.*, a function $\mathcal{C}(I, X) \to \mathcal{C}(Y, I)$.

Example categories where this works:

- the Kleisli category of a free module monad;
- $\mathbf{sfKrn}$, the category of "s-finite kernels".

N.B. Each $\mathcal{C}(X, I)$ is then a monoidal category: morphisms of effects are 'differences'; monoidal product is $+$. This extends to loss functions (via "state-dependent differences").

* $\mathcal{C}(-, I)$ has a commutative monoid structure compatible with copy-composition.

Examples

Relative entropy: $\mathsf{KL}(c, c^\sharp)_\beta(y) := D_{c^\sharp_\beta, c^\dagger_\beta}(y) = D_X\big(c^\sharp_\beta(y), c^\dagger_\beta(y)\big)$.

Log likelihood: $\mathsf{MLE}(c, c^\sharp)_\beta(y) := -\log p_{c^{\dagger\!\!\!=}\bullet\beta}(y)$.

Free energy: $\mathsf{FE}(c, c^\sharp)_\beta(y) := \mathsf{KL}(c, c^\sharp)_\beta(y) + \mathsf{MLE}(c, c^\sharp)_\beta(y)$.

Alternatively, $\mathsf{FE} = \mathsf{KL} + \mathsf{MLE}$.

... But how do they compose? (Where does the chain rule enter?)

Loss models: compositional loss functions

Loss functions compose much like inversions.

Given $(c, c^\sharp, L^c) : X \to Y$ and $(d, d^\sharp, L^d) : Y \to Z$, $L^d \circ L^c = L^d_c + {d^\sharp_c}^* L^c$.

More explicitly, $(L^d \circ L^c)_\beta(z) = L^d_{c^\dagger \bullet \beta}(z) + \mathbb{E}_{y \sim d^\sharp_{c^\dagger \bullet \beta}(z)}\big[L^c(y)\big]$.

A loss model is an assignment of loss functions to lenses
that is compatible with this chain rule.

Example: $\mathsf{KL}\big((d, d^\sharp) \circledcirc (c, c^\sharp)\big) = \mathsf{KL}(d, d^\sharp)_c + {d^\sharp_c}^* \mathsf{KL}(c, c^\sharp)$.
(This follows from our earlier chain rule.)

However ...

Laxness

Unlike KL, not all loss models are *strict* sections of π_{Loss}!

With morphisms between loss functions, we can consider *lax* loss models: the laxators measure the difference from $L^d \circ L^c$ to $L^{d \oplus c}$.

Hence, for a given $\mathcal{C}$, we have a category of (lax) loss models $\mathsf{Loss}(\mathcal{C})$; morphisms of loss models are 'icons'.

Examples: MLE and thus FE are lax loss models.

But this is not the only kind of laxness around! ...

Monoidal structures

There are a few monoidal structures here.

1. $\mathcal{C}$, $\mathbf{Copara}_2(\mathcal{C})$, and $\mathbf{BayesLens}$ are all standardly monoidal; and hence so is $\mathbf{StatGame}$.
2. We also have the sum of loss functions $L^c + L^{c'} : Y \overset{X}{\to} I$.
3. And this induces a monoidal structure on $\mathsf{Loss}(\mathcal{C})$.

   (We saw this when defining FE as KL + MLE.)
4. Typically there are also monoidal structures on loss models themselves!
   - Hence we have $\mathsf{MonLoss}(\mathcal{C}) \hookrightarrow \mathsf{Loss}(\mathcal{C})$.
   - These structures are generally lax (due to correlations in the priors).
   - Thus each of KL, MLE, and FE are lax monoidal.
   - (But KL on parallel channels is strict monoidal; see [6, Remark 5.4.11] ...)

Review of the talk

Statistical games formalize 'local' approximate inference, yielding a fibration over Bayesian lenses.

Using copy-composition, the relative entropy chain rule is witnessed by a (strict) section of this fibration.

Other important statistical quantities give lax sections.

And all of these admit various monoidal structures.

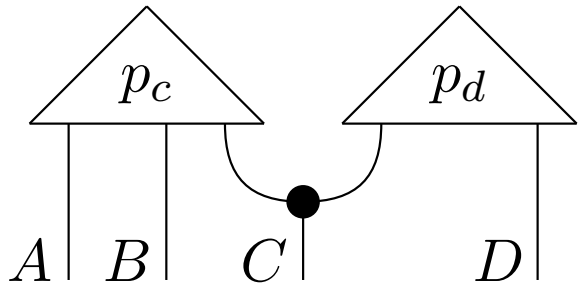
(However, the bicategory $\mathbf{Copara}_2(\mathcal{C})$ seems inessential ...)

Bonus: another route to copy-composition

I can see at least two alternatives to $\mathbf{Copara}_2(\mathcal{C})$. Here's one.

Instead of channels c as horizontal 1-cells, use their densities p_c.
Construct a decorated cospan double category accordingly:

e.g. $A \xrightarrow{p_c} C \xrightarrow{p_d} D \quad :=$



0-cells are sets of objects in $\mathcal{C}$.
Horizontal 1-cells are cospans decorated by densities.
Vertical 1-cells are 'reindexings'; 2-cells are compatible channels.

Note that this has copy-composition built in, without hacks.
It's also more faithful to "factor graphs".

References

1. Braithwaite, D., Hedges, J., and St Clere Smithe, T.: The Compositional Structure of Bayesian Inference. (2023). arXiv: 2305.06112 [cs, math]. http://arxiv.org/abs/2305.06112 (visited on 05/11/2023). preprint

2. St Clere Smithe, T.: Bayesian Updates Compose Optically. (2020). arXiv: 2006.01631 [math.CT]. preprint

3. Ghani, N., Hedges, J., Winschel, V., and Zahn, P.: Compositional Game Theory. Proceedings of Logic in Computer Science (LiCS) 2018 (2016). arXiv: 1603.04641 [cs.GT]

4. Capucci, M.: Diegetic Representation of Feedback in Open Games. (2022). arXiv: 2206.12338 [cs.GT]. preprint

5. Capucci, M., Gavranović, B., Hedges, J., and Rischel, E.F.: Towards Foundations of Categorical Cybernetics. (2021). arXiv: 2105.06332 [math.CT]. preprint

6. St Clere Smithe, T.: Mathematical Foundations for a Compositional Account of the Bayesian Brain. (2022). arXiv: 2212.12538 [cs, math, q-bio, stat]. http://arxiv.org/abs/2212.12538 (visited on 01/09/2023). preprint