Mathematical Foundations for a Compositional Account of the Bayesian Brain

Toby St Clere Smithe

Department of Experimental Psychology
St Edmund Hall
University of Oxford

2023 September 8

Introduction and motivation

- Observation of 'hierarchical' structure in predictive coding circuits ...
  - ... clear signature of compositionality!
- "Modularity of mind" (Fodor 1983)
  - and modularity in our understanding of mind!
- More broadly: ACT as tractable but rigorous approach to complex systems

Slogan: predictive coding as *dynamical semantics* for certain *statistical games*

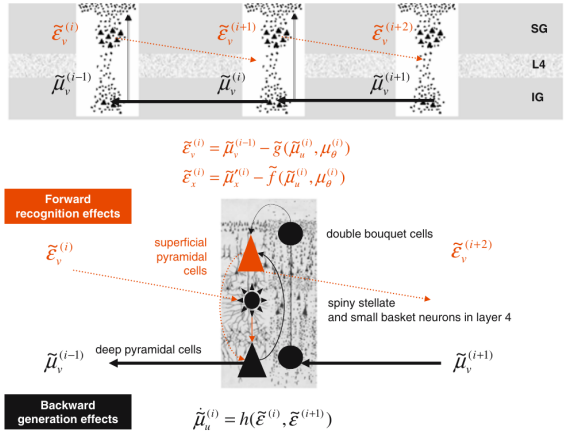


Figure: Friston and Stephan (2007, Fig.3)

Overview of main contributions

- *Bayesian lenses*, formalizing a "chain rule" for Bayesian inference.
- The corresponding result that "Bayesian updates compose optically".

- The concept of *statistical game*, by which losses are attached compositionally to lenses.
- The classification of various loss functions as sections of 2-fibrations of statistical games.
  - Relative entropy gives a strict section, capturing its chain rule.
  - Free energy is lax, inherited from log-likelihood.

- The notion of 'copy-composition', which makes the above classification work.
  - And copy-composite versions of the foregoing ...

- A new account of general open dynamical systems via polynomial coalgebra
  - plus an associated construction of dynamical lenses ('cilia').

- *Approximate inference doctrines*: dynamical semantics for statistical games
  - exemplified (laxly, non-unitally...) by predictive coding.

Plus an awful lot of exposition ... And a lot of new directions ...

Bayesian lenses: the 'chain rule' for inference

A 'hierarchical' model factorizes into one predictive process c, then another, d.

This is[1] sequential composition $d \bullet c$ in a category $\mathcal{C}$ of stochastic maps.

The task of a predictive coding system is then to invert this composite model.
That is, it pairs the predictive processes c, d with inversions $c^\sharp, d^\sharp$.

Bayes' law: inversion depends on priors as well as predictions ('likelihoods').
Hence if $c : X \to\!\!\!\!\bullet\, Y$, then $c^\sharp$ is *state-dependent*: $\mathcal{C}(I, X) \to \mathcal{C}(Y, X)$.

State-dependent channels form an indexed category $\mathrm{Stat} : \mathcal{C}^{\,\mathrm{op}} \to \mathbf{Cat}$ (Def. 4.3.2)
and *Bayesian lenses* are the morphisms of its (op)Grothendieck construction (Def. 4.3.8).

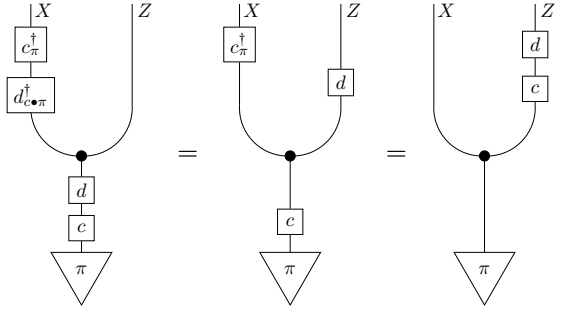
This tells us that $(d, d^\sharp) \circledcirc (c, c^\sharp) = (d \bullet c, c^\sharp \circ d^\sharp_c)$.
Alternatively: if $p_{d \bullet c}(z|x) = \sum_y p_d(z|y)\, p_c(y|x)$, then $p_{c^\sharp_\pi \bullet d^\sharp_{c \bullet \pi}}(x|z) = \sum_y p_{c^\sharp_\pi}(x|y)\, p_{d^\sharp_{c \bullet \pi}}(y|z)$.

[1] or at least, appears at first to be..!

Bayesian updates compose optically

The Grothendieck construction tells us how to compose Bayesian lenses.
It *doesn't* tell us that the composite of Bayesian inversions equals the inversion of the composite.

We need to check that $(d \bullet c)^\dagger = c^\dagger \circ d^\dagger_c$. Happily, this is true, up to almost-equality (Thm. 4.3.14).



Hence † almost surely defines a section $\mathcal{C} \to \mathbf{BayesLens}$ of the Grothendieck fibration $\pi_{\mathsf{Lens}} : \mathbf{BayesLens} \to \mathcal{C}$ (Cor. 4.3.15), mapping $c \mapsto (c, c^\dagger)$.

Bayesian lens miscellany

- Strictly speaking, † is well-defined only where morphisms have inversions for all priors: a "full support" assumption (Rmk.s 4.1.20, 4.3.3).
  - Can circumvent this with support objects and dependent lenses (Braithwaite et al. 2023).
- Lawfulness (§4.3.4): Bayesian lenses are not 'lawful' lenses! (They 'mix' information.)
- $\mathbf{BayesLens}$ inherits a monoidal structure from $\mathcal{C}$ (Prop. 4.3.11), but † is only lax monoidal
  - because there may be correlations across a joint prior!
  - (We'll see more of such laxness later.)

Statistical games for approximate inference

† classifies *exact* inference — but there are many lenses outside its image.
Can we classify these *approximate* systems, too? (We'll need to, for predictive coding.)

An arbitrary lens won't be very good at inference: hence, choose one with good performance[2].
This means *measuring* that performance, or: equipping lenses with loss functions.

A *statistical game* $(c, c^\sharp, L^c) : (X, A) \to (Y, B)$ is a Bayesian lens $(c, c^\sharp) : (X, A) \nrightarrow (Y, B)$, paired with a *loss function* L^c, *i.e.* a "state-dependent effect" $B \overset{X}{\bullet\!\!\to} I$ (Def. 5.3.6).

- We require $\mathcal{C}$ to have *bilinear effects* (Def. 5.3.1).
- In $\mathbf{sfKrn}$, or $\mathbf{Vect}_{\mathbb{R}_+}$, L^c is equivalently $\mathcal{C}(1, X) \times B \to \mathbb{R}_+$ (Ex. 5.3.5).

How do these compose? What are some examples? How do they relate to predictive coding?

[2] relative to a given problem

Fibrations of statistical games: loss composition

Just as attaching inversions to channels gives a fibration, so does attaching losses to lenses.

Idea: $\sum_{X:\mathcal{C}^{\mathrm{op}}} \mathrm{Stat}(X)^{\mathrm{op}} \xrightarrow{\mathrm{Stat}(-)(=,I)} \mathbf{MonCat} \xrightarrow{\mathbf{B}} \mathbf{Bicat}$ yields an indexed bicategory (when $\mathcal{C}$ has bilinear effects).

We can lift this over $\mathbf{BayesLens}$, giving $\mathbf{StatGame} \xrightarrow{\pi_{\mathsf{Loss}}} \mathbf{BayesLens} \xrightarrow{\pi_{\mathsf{Lens}}} \mathcal{C}$.

Then $L^d \circ L^c = L^d_c + d^{\sharp *}_c L^c$. Explicitly: $(L^d \circ L^c)_\pi(z) = L^d_{c \bullet \pi}(z) + \mathbb{E}_{y \sim d^\sharp_{c \bullet \pi}(z)}\big[L^c(y)\big]$.
Well-behaved loss functions satisfy this general rule: that is,
just as † is a section of π_{Lens}, they give sections of π_{Loss}.

These *loss models* capture some well-known examples and their "chain rules":
relative entropy[3] (§5.3.3.1); log-likelihood (§5.3.3.2); free energy (§5.3.3.3) ...
... with some twists!

1. We need *copy-composition*;
2. log-likelihood (and hence free energy) is *lax*[4].

[3] Relative entropy doesn't actually need the inversions (Rmk. 5.3.23).

[4] This makes use of the bicategorical structure of $\mathbf{StatGame}$ (*i.e.*, differences of losses) (§5.3.2).

Copy-composition

For parallel channels $c, c' : X \nrightarrow Y$, the relative entropy $D_{c,c'} : X \to \mathbb{R}_+$ is defined by

$$x \mapsto \mathop{\mathbb{E}}_{y \sim c(x)} \left[\log p_c(y|x) - \log p_{c'}(y|x) \right].$$

It *almost* satisfies the chain rule

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

Except marginalization doesn't commute with $\log$, and $p_{d \bullet c}(z|x) = \sum_y p_d(z|y)\, p_c(y|x)$.
We need composition $\bullet$ to yield the joint distribution $p_d(z|y)\, p_c(y|x)$: *copy-composition*.

§5.2 constructs a bicategory $\mathbf{Copara}_2(\mathcal{C})$ and recapitulates the Bayesian lens story.[5]
On this base, the aforementioned examples do yield sections (with relative entropy strict).

[5] I now see neater constructions: *cf.* bonus slides...

Monoidal miscellany

Worth pointing out that there are many monoidal structures around (§5.4).

This means we can model "inference systems in parallel",
and measure their performance in parallel (Def. 5.4.7, Rmk. 5.4.8, Prop. 5.4.9).
But: correlations in the prior $\implies$ loss models generally *lax* monoidal.

$+$ also gives a monoidal structure on loss models (Prop. 5.3.18): "sum the losses".
We can use this to obtain loss models compositionally: *e.g.* FE $=$ KL $+$ MLE (Def. 5.3.26).

N.B. The category theory of some of these structures is not fully worked out.
I didn't examine all the coherence conditions for *monoidal indexed bicategory*!
But everything should be well-behaved nonetheless.

Open dynamical systems via generalized polynomial coalgebra (1)

At this point we turn to dynamics.
I wanted two things, not available with existing compositional constructions:

1. Somewhere I could define "categories with dynamic morphisms";
2. with dynamics that may be both continuous-time and stochastic (*e.g.* Markov processes).

For (1), it made sense to work with coalgebra — and particularly polynomial coalgebra.
For (2), I needed some tricks.

A 'closed' dynamical system with time $\mathbb{T}$ is an action of $\mathbb{T}$ on some state space, $\mathbf{B}\mathbb{T} \to \mathbf{Cat}$.
These correspond to $\triangleleft$-comonoid homomorphisms[6] $Sy^S \to y^{\mathbb{T}}$ in $\mathbf{Poly}$.

Then: an *open* system with time $\mathbb{T}$ and interface p is a morphism $\beta : Sy^S \to [\mathbb{T}y, p]$
... that yields a $\triangleleft$-comonoid homomorphism for any 'inputs' (Def. 6.2.1).

[6] The $\triangleleft$-comonoid homomorphism condition enforces the $\mathbb{T}$-action laws.

Open dynamical systems via generalized polynomial coalgebra (2)

The preceding definition induces an opindexed category $\mathbf{Coalg}^{\mathbb{T}} : \mathbf{Poly} \to \mathbf{Cat}$ (Prop. 6.2.10)[7].

- It reduces to the 'usual' coalgebras when $\mathbb{T} = \mathbb{N}$ (Rmk. 6.2.7);
- it captures closed systems with the trivial interface y (Prop. 6.2.4);
- it has a monoidal structure, for systems in parallel (Prop. 6.2.10);

and it can be instantiated in non-deterministic settings, as follows (§6.2.2).

A monad M on $\mathbf{Set}$ induces a comonad $\bar{M}$ on $\mathbf{Poly}$.
$\mathrm{co}\mathcal{K}\ell(\bar{M})$ has morphisms with 'backward' M-effects (Rmk. 6.2.19, Def.s 6.2.14, 6.2.15).
This yields dynamical systems with M-effectful update maps.
In particular, a probability monad gives open Markov processes.

We can also do open random dynamical systems (§6.2.3), but less neatly.

[7] Reindexing is by post-composition with $\mathbf{Poly}$ morphisms.

Cilia

For the dynamical semantics, I sought categories of "dynamical lenses".
Cilia are dynamical systems that control lenses.

Short story[8]: given a category $\mathcal{D}$ enriched in $\mathbf{Poly}$, base change along $\mathbf{Coalg}^{\mathbb{T}} : \mathbf{Poly} \to \mathbf{Cat}$ yields a bicategory $\widetilde{\mathcal{D}}$ with "dynamical $\mathcal{D}$-morphisms" as 1-cells.

Longer story (§6.3): construct a $\mathbf{Poly}$-category of Bayesian lenses, with hom $[\![-,=]\!]$.
Then the (monoidal) bicategory $\mathbf{Cilia}^{\mathbb{T}}$ has hom categories $\mathbf{Coalg}^{\mathbb{T}}\big([\![-,=]\!]\big)$ (Def. 6.3.8).

We also have a 'differential' version — sketched in §6.3.2.

[8] I only noticed this account after submission...

Approximate inference doctrines: putting the pieces together

Basic idea of Def. 7.3.1: (monoidal) functors $\mathbf{P}\mathcal{C} \to \mathbf{Cilia}$ that factorize as $\mathbf{P}\mathcal{C} \to \mathbf{PBayesLens} \to \mathbf{PStatGame} \to \mathbf{Cilia}$, with the first two being sections.
Here, $\mathbf{P}$ denotes ('external') parameterization, which is itself functorial — *cf.* §3.2.2 and §7.2.

As we've seen, there's quite some laxness about ...
and each step may only be defined on the image of the preceding ...
and in this formulation, predictive coding isn't unital[9].
But with these caveats, we do get (most of) a functorial semantics!

[9] as the posterior is (instantaneously) determined by the parameter/dynamical state, not the input observation

Predictive coding as approximate inference doctrine

Predictive coding can be understood as
"gradient descent on the Laplacian free energy, with respect to the posterior mean".
Thus the *Laplace doctrine* (§7.3.1) is defined on $\mathcal{C} = \mathbf{FdGauss}$ (§7.1).
(N.B. Nonlinear Gaussian channels are not closed under composition, but they are under copy-composition.)

Schematically: $\mathbf{FdGauss} \xrightarrow{\ell} \mathbf{PBayesLens} \xrightarrow{\text{PLFE}} \mathbf{PStatGame} \xrightarrow{\nabla} \mathbf{DiffCilia} \xrightarrow{\int} \mathbf{Cilia}$.
(ℓ is the non-unital bit. ∇ computes gradient descent with respect to the parameterization. $\int$ is time-integration.)
Functoriality enforces a "mean field approximation" in ℓ. (Prop. 7.3.7.)

Hebb-Laplace doctrine (§7.3.2): make a particular choice of parameterization in $\mathbf{PFdGauss}$, *i.e.* for each $c : X \xrightarrow{\Theta_Y} Y$, $\mu_c(x) = \theta\, h(x)$, with θ a square matrix on Y and h differentiable.
(Idea is θ represents synaptic weights, h the neural activation function.)
Then, proceed as before — but now we also learn the synaptic parameters (a 'Hebb' rule).

Some next steps for compositional predictive coding

On the syntactic side — better accounts of copy-composite models (see bonus slides):

- factor graphs (via decorated cospans), and their directed cousins (right adjoints?);
- "stochastic sections" (with links to quantitative type theory and 'nested' systems).

On the semantic side — I think a better perspective is from information geometry.

- Precision-weighted prediction errors are tangent vectors to Gaussians[10]!
- This yields a precise connection to 'differential' learners
  - and constructions like dynamic categories (Shapiro and Spivak 2022)
  - and hence more generally structured P.C. architectures (Salvatori et al. 2023).
- Hopefully: a 'geometric' understanding of statistical games, perhaps via (Vigneaux 2021)?

[10] with known covariance

Categorical systems theory

More broadly, many links to categorical cybernetics / systems theory to explore.

- Regarding coalgebra and dynamics:
  - Does $\mathbf{Coalg}^{\mathbb{T}}(p)$ form a topos? Is it related to the topos of $\mathbf{Int}$-sheaves (Schultz et al. 2020)?
  - Do RDSs relate to Markov processes via "randomness pushback" (Fritz 2019, Def. 11.19)?
- How best to formulate *dynamical* inference?
- And then: *active* inference ...
  - What is the relationship between planning, backward induction, and RL?
  - What is the relationship between active inference and open economic games?
  - Or between backward induction and message-passing/expectation-maximization?
- *Spatial* systems theory: interfaces typically have geometry ...
  - Hence: multi-agent systems? Consensus, cohomology and corporations?
  - There's a sheaffy relationship between message-passing and diffusion (Peltre 2021) ...
- Universality of the free energy principle:
  - by expressing active inference systems-theoretically, we should be able to formalize its claimed universality as an adjunction of systems theories.

And many further directions sketched in Chapter 8!

References I

Fodor, Jerry A (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT press.

Friston, Karl J. and Klaas E. Stephan (Sept. 2007). "Free-Energy and the Brain". In: *Synthese. An International Journal for Epistemology, Methodology and Philosophy of Science* 159.3, pp. 417–458. DOI: 10.1007/s11229-007-9237-y. URL: http://dx.doi.org/10.1007/s11229-007-9237-y.

Braithwaite, Dylan, Jules Hedges, and Toby St Clere Smithe (May 10, 2023). *The Compositional Structure of Bayesian Inference*. arXiv: 2305.06112 [cs, math]. URL: http://arxiv.org/abs/2305.06112 (visited on 05/11/2023). preprint.

Shapiro, Brandon and David I. Spivak (May 8, 2022). "Dynamic Categories, Dynamic Operads: From Deep Learning to Prediction Markets". arXiv: 2205.03906 [math.CT].

Salvatori, Tommaso et al. (Aug. 15, 2023). *Brain-Inspired Computational Intelligence via Predictive Coding*. arXiv: 2308.07870 [cs]. URL: http://arxiv.org/abs/2308.07870 (visited on 08/24/2023). preprint.

Vigneaux, Juan Pablo (Nov. 8, 2021). *Information Structures and Their Cohomology*. arXiv: 1709.07807 [cs, math]. URL: http://arxiv.org/abs/1709.07807 (visited on 04/06/2023). preprint.

References II

Schultz, Patrick, David I Spivak, and Christina Vasilakopoulou (2020). "Dynamical Systems and Sheaves". In: *Applied Categorical Structures* 28, pp. 1–57. DOI: 10.1007/s10485-019-09565-x. arXiv: 1609.08086 [math.CT].

Fritz, Tobias (Aug. 19, 2019). "A Synthetic Approach to Markov Kernels, Conditional Independence and Theorems on Sufficient Statistics". In: *Advances in Mathematics* 370.107239. DOI: 10.1016/j.aim.2020.107239. arXiv: 1908.07021v3 [math.ST].

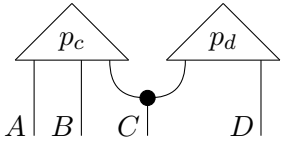
Peltre, Olivier (2021). *Belief Propagation as Diffusion*. Vol. 12829. DOI: 10.1007/978-3-030-80209-7. arXiv: 2107.12230 [math-ph]. URL: http://arxiv.org/abs/2107.12230 (visited on 08/09/2023).

Factor graphs: another route to copy-composition

Here's one alternative to $\mathbf{Copara}_2(\mathcal{C})$.

Instead of channels c as horizontal 1-cells, use their densities p_c — or, generally, costates.
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 (on the product of the objects at the apex).
Vertical 1-cells are (indexed) comonoid homomorphisms in $\mathcal{C}$; 2-cells are more general.

- This formalizes *factor graphs* (undirected graphical models).
- It has copy-composition built in, without hacks.
- There is a lax embedding of $\mathcal{C}$.

Alternatively: stochastic sections

Another alternative: observe that copy-composition is composition of 'graphs' $X \to\!\!\!\!\bullet\, X \otimes Y$.
These are "stochastic sections" of the projection $X \times Y \to X$.

There is a bifibration $\mathcal{E}$ over suitable $\mathcal{C}$ whose objects are such (deterministic) bundles in $\mathcal{C}$, and whose morphisms are (non-deterministic) morphisms in $\mathcal{C}$ between them.

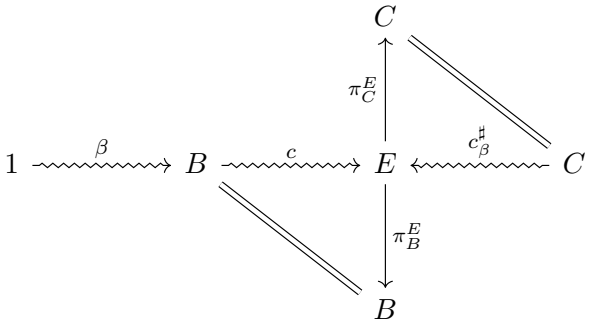
Consider the bicategory of spans of such bundles.
We can decorate each span with sections of the left leg.
Assuming $\mathcal{E}$ satisfies Beck-Chevalley: push-pull yields a strong composition of decorations.

- This gives a decorated span bicategory of "hierarchical models".
- Again, we get copy-composition automatically.

And we can construct a version of Bayesian lenses here, too.
Exact lenses make a 'butterfly' diagram commute (see next slide).

Butterfly diagram for exact Bayesian lenses

Here $(c, c^\sharp)$ is a Bayesian lens $B \nrightarrow C$ in the "stochastic section" sense, and β is a prior.



Variation on the semantics for predictive coding (1)

If we don't mind doing away with some of the structure of approx. inf. doctrines, we can define a $\mathbf{Poly}$-category with objects given by $\mathbb{N}$ and hom-polynomials

$$[\![l, m]\!] := \sum_{i \geqslant -l} \mathcal{G}(l+i)\, \Gamma(l, m)\, y^{\mathbb{R}^m\, \mathrm{H}(i+m)}$$

- posterior (output): $\mathcal{G}(n)$ is the space of Gaussians on $\mathbb{R}^n$;
- likelihood (output): $\Gamma(l, m)$ is Gaussian stochastic sections from $\mathbb{R}^l$ to $\mathbb{R}^m$;
- prediction error (input): $\mathrm{H}(n)$ is $\mathbb{R}^n \to \mathbb{R}^n$; and
- observation (input): $\mathbb{R}^m$.

This yields a dynamic bicategory $\mathbf{PC}$ (by the earlier recipe), and predictive coding gives a strong functor $\Gamma \to \mathbf{PC}$.

Variation on the semantics for predictive coding (2)

If we assume Gaussians with fixed covariance, then the points of $\mathsf{T}\mathcal{G}(n)$ are pairs $(\mu_\gamma, \eta_\gamma)$, where $\mu_\gamma : \mathbb{R}^n$ is the mean determining a Gaussian $\gamma : \mathcal{G}(n)$, and $\eta_\gamma : \mathbb{R}^n \to \mathbb{R}^n := y \mapsto \Sigma_\gamma^{-1}(y - \mu_\gamma)$ is a precision-weighted prediction error ($=$ "$\partial_y E_\gamma$"). Hence $\mathsf{T}\mathcal{G}(n) \hookrightarrow \mathbb{R}^n \, \mathrm{H}(n)$!

We can use this to cast predictive coding in the manner of Shapiro and Spivak (2022): Let $t := \sum_{x:\mathbb{R}} y^{\mathsf{T}_x \mathcal{G}(1)}$. Then $t^{\otimes m} \cong \sum_{x:\mathbb{R}^m} y^{\prod_{i:[m]} \mathsf{T}_{x^i} \mathcal{G}(1)}$.
A morphism $\varphi : t^{\otimes m} \to t^{\otimes n}$ is a pair $(\varphi_1, \varphi^\sharp)$:

- $\varphi_1 : \mathbb{R}^m \to \mathbb{R}^n$ predicts the next layer's mean;
- $\varphi^\sharp : \sum_{x:\mathbb{R}^m} \prod_{j:[n]} \mathsf{T}_{\varphi_1(x)^j} \mathcal{G}(1) \to \sum_{x:\mathbb{R}^m} \prod_{i:[m]} \mathsf{T}_{x^i} \mathcal{G}(1)$
  passes back the posterior means and prediction errors.

Then: can give a coalgebra on $[t^{\otimes m}, t^{\otimes n}]$ which updates by gradient descent on the energy.
But note that this enforces a very strict mean field approximation!
(The latter so that "horizontal composition" is strong not lax...)