

Approximate Inference via Fibrations of Statistical Games

Toby St Clere Smithe

VERSES Research

Topos Institute

q@tsmithe.net

We characterize a number of well known systems of approximate inference as *loss models*: lax sections of 2-fibrations of statistical games, constructed by attaching internally-defined loss functions to Bayesian lenses. Our examples include the relative entropy, which constitutes a *strict* section, and whose chain rule is formalized by the horizontal composition of the 2-fibration. In order to capture this compositional structure, we first introduce the notion of ‘copy-composition’, alongside corresponding bicategories through which the composition of copy-discard categories factorizes. These bicategories are a variant of the **Copara** construction, and so we additionally introduce coparameterized Bayesian lenses, proving that coparameterized Bayesian updates compose optically, as in the non-coparameterized case.

1 Introduction

In previous work [1], we introduced *Bayesian lenses*, observing that the Bayesian inversion of a composite stochastic channel is (almost surely) equal to the ‘lens composite’ of the inversions of the factors; that is, *Bayesian updates compose optically* (‘BUCO’) [2]. Formalizing this statement for a given category \mathcal{C} all of whose morphisms (‘channels’) admit Bayesian inversion, we can observe that there is (almost surely) a functor $(-)^{\dagger} : \mathcal{C} \rightarrow \mathbf{BayesLens}(\mathcal{C})$ from \mathcal{C} to the category $\mathbf{BayesLens}(\mathcal{C})$ whose morphisms $(X, A) \leftrightarrow (Y, B)$ are Bayesian lenses: pairs (c, c') of a channel $X \rightarrow Y$ with a ‘state-dependent’ inverse $c' : \mathcal{C}(I, X) \rightarrow \mathcal{C}(B, A)$. Bayesian lenses constitute the morphisms of a fibration $\pi_{\mathbf{Lens}} : \mathbf{BayesLens}(\mathcal{C}) \rightarrow \mathcal{C}$, since $\mathbf{BayesLens}(\mathcal{C})$ is obtained as the Grothendieck construction of (the pointwise opposite of) an indexed category $\mathbf{Stat} : \mathcal{C}^{\text{op}} \rightarrow \mathbf{Cat}$ of ‘state-dependent channels’ (recalled in Appendix A), and the functor $(-)^{\dagger}$ is in fact a section of $\pi_{\mathbf{Lens}}$, taking $c : X \rightarrow Y$ to the lens $(c, c^{\dagger}) : (X, X) \leftrightarrow (Y, Y)$, where c^{\dagger} is the almost-surely unique Bayesian inversion of c (so that the projection $\pi_{\mathbf{Lens}}$ can simply forget the inversion, returning again the channel c).

The functor $(-)^{\dagger}$ picks out a special class of Bayesian lenses, which we may call *exact* (as they compute ‘exact’ inversions), but although the category $\mathbf{BayesLens}(\mathcal{C})$ has many other morphisms, the construction is not extravagant: by comparison, we can think of the non-exact lenses as representing *approximate* inference systems. This is particularly necessary in computational applications, because computing exact inversions is usually intractable, but this creates a new problem: choosing an approximation, and measuring its performance. In this paper, we formalize this process, by attaching *loss functions* to Bayesian lenses, thus creating another fibration, of *statistical games*. Sections of this latter fibration encode compositionally well-behaved systems of approximation that we call *loss models*.

A classic example of a loss model will be supplied by the relative entropy, which in some sense measures the ‘divergence’ between distributions: the game here is then to minimize the divergence between the approximate and exact inversions. If π and π' are two distributions on a space X , with corresponding density functions p_{π} and $p_{\pi'}$ (both with respect to a common measure), then their relative

entropy $D(\pi, \pi')$ is the real number given by $\mathbb{E}_{x \sim \pi} [\log p_\pi(x) - \log p_{\pi'}(x)]^1$. Given a pair of channels $\alpha, \alpha' : A \rightarrow B$ (again commensurately associated with densities), we can extend D to a map $D_{\alpha, \alpha'} : A \rightarrow \mathbb{R}_+$ in the natural way, writing $a \mapsto D(\alpha(a), \alpha'(a))$. We can assign such a map $D_{\alpha, \alpha'}$ to any such parallel pair of channels, and so, following the logic of composition in \mathcal{C} , we might hope for the following equation to hold for all $a : A$ and composable parallel pairs $\alpha, \alpha' : A \rightarrow B$ and $\beta, \beta' : B \rightarrow C$:

$$D_{\beta \bullet \alpha, \beta' \bullet \alpha'}(a) = \mathbb{E}_{b \sim \alpha(a)} [D_{\beta, \beta'}(b)] + D_{\alpha, \alpha'}(a)$$

The right-hand side is known as the *chain rule* for relative entropy, but, unfortunately, the equation does *not* hold in general, because the composites $\beta \bullet \alpha$ and $\beta' \bullet \alpha'$ involve an extra expectation (by the ‘Chapman-Kolmogorov’ rule for channel composition). However, we *can* satisfy an equation of this form by using ‘copy-composition’: if we write \forall_B to denote the canonical ‘copying’ comultiplication on B , and define $\beta \bullet^2 \alpha := (\text{id}_B \otimes \beta) \bullet \forall_B \bullet \alpha$, then $D_{\beta \bullet^2 \alpha, \beta' \bullet^2 \alpha'}(a)$ *does* equal the chain-rule form on the right-hand side. This result exhibits a general pattern about “copy-discard categories” [3] such as \mathcal{C} : composition \bullet can be decomposed into first copying \forall , and then discarding $\bar{\tau}$. If we don’t discard, then we retain the ‘intermediate’ variables, and this results in a functorial assignment of relative entropies to channels.

The rest of this paper is dedicated to making use of this observation to construct loss models, including (but not restricted to) the relative entropy. The first complication that we encounter is that copy-composition is not strictly unital, because composing with an identity retains an extra variable. To deal with this, in §2, we construct a *bicategory* of copy-composite channels, extending the **Copara** construction [4, §2], and build coparameterized (copy-composite) Bayesian lenses accordingly; we also prove a corresponding BUCO result. In §3, we then construct 2-fibrations of statistical games, defining loss functions internally to any copy-discard category \mathcal{C} that admits “bilinear effects”. Because we are dealing with approximate systems, the 2-dimensional structure of the construction is useful: loss models are allowed to be *lax* sections. We then characterize the relative entropy, maximum likelihood estimation, the free energy, and the ‘Laplacian’ free energy as such loss models.

Assuming \mathcal{C} is symmetric monoidal, the constructions here result in monoidal (2-)fibrations, but due to space constraints we defer the presentation of this structure (and its exemplification by the foregoing loss models) to Appendix B.

Remark 1.1. Much of this work is situated amongst monoidal fibrations of bicategories, the full theory of which is not known to the present author. Fortunately, enough structure is known for the present work to have been possible, and where things become murkier—such as in the context of monoidal indexed bicategories and their lax homomorphisms—the way largely seems clear. For this, we are grateful to Baković [5], Johnson and Yau [6], and Moeller and Vasilakopoulou [7] in particular for lighting the way; and we enthusiastically encourage the further elucidation of these structures by category theorists.

Remark 1.2. For reasons of space, detailed proofs are not included in the proceedings version of this paper; however, they are included in an appendix to the conference submission, which is available on the arXiv repository with the paper ID 2306.17009v1.

2 ‘Copy-composite’ Bayesian lenses

2.1 Copy-composition by coparameterization

In a locally small copy-discard category \mathcal{C} , every object A is equipped with a canonical comonoid structure $(\forall_A, \bar{\tau}_A)$, and so, by the comonoid laws, it is almost a triviality that the composition function

¹For details about this ‘expectation’ notation \mathbb{E} , see 3.11.

$\bullet : \mathcal{C}(B, C) \times \mathcal{C}(A, B) \rightarrow \mathcal{C}(A, C)$ factorizes as

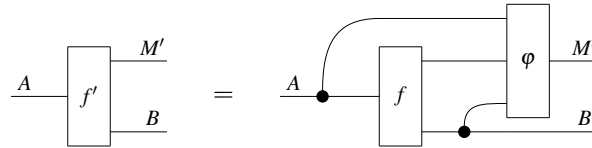
$$\begin{aligned} \mathcal{C}(B, C) \times \mathcal{C}(A, B) &\xrightarrow{(\text{id}_B \otimes -) \times \mathcal{C}(\text{id}_A, \blacktriangledown_B)} \mathcal{C}(B \otimes B, B \otimes C) \times \mathcal{C}(A, B \otimes B) \cdots \\ \cdots &\xrightarrow{\bullet} \mathcal{C}(A, B \otimes C) \xrightarrow{\mathcal{C}(\text{id}_A, \text{proj}_C)} \mathcal{C}(A, C) \end{aligned}$$

where the first factor copies the B output of the first morphism and tensors the second morphism with the identity on B , the second factor composes the latter tensor with the copies, and the third discards the extra copy of B^2 . This is, however, only *almost* trivial, since it witnesses the structure of ‘Chapman-Kolmogorov’ style composition in categories of stochastic channels such as $\mathcal{Kl}(\mathcal{D})$, the Kleisli category of the (finitary) distributions monad $\mathcal{D} : \mathbf{Set} \rightarrow \mathbf{Set}$. There, given channels $c : A \rightarrow B$ and $d : B \rightarrow C$, the composite $d \bullet c$ is formed first by constructing the ‘joint’ channel $d \bullet^2 c$ defined by $(d \bullet^2 c)(b, c|a) := d(c|b)c(b|a)$, and then discarding (marginalizing over) $b : B$, giving

$$(d \bullet c)(c|a) = \sum_{b:B} (d \bullet^2 c)(b, c|a) = \sum_{b:B} d(c|b)c(b|a).$$

Of course, the channel $d \bullet^2 c$ is not a morphism $A \rightarrow C$, but rather $A \rightarrow B \otimes C$; that is, it is *coparameterized* by B . Moreover, as noted above, \bullet^2 is not strictly unital: we need a 2-cell that discards the coparameter, and hence a bicategory, in order to recover (weak) unitality. We therefore construct a bicategory $\mathbf{Copara}_2(\mathcal{C})$ as a variant of the **Copara** construction [4, §2], in which a 1-cell $A \rightarrow B$ may be any morphism $A \rightarrow M \otimes B$ in \mathcal{C} , and where horizontal composition is precisely copy-composition.

Theorem 2.1. Let $(\mathcal{C}, \otimes, I)$ be a copy-discard category. Then there is a bicategory $\mathbf{Copara}_2(\mathcal{C})$ as follows. Its 0-cells are the objects of \mathcal{C} . A 1-cell $f : A \rightarrow B$ is a morphism $f : A \rightarrow M \otimes B$ in \mathcal{C} . A 2-cell $\varphi : f \Rightarrow f'$, with $f : A \xrightarrow{M} B$ and $f' : A \xrightarrow{M'} B$, is a morphism $\varphi : A \otimes M \otimes B \rightarrow M'$ of \mathcal{C} , satisfying the *change of coparameter* axiom:

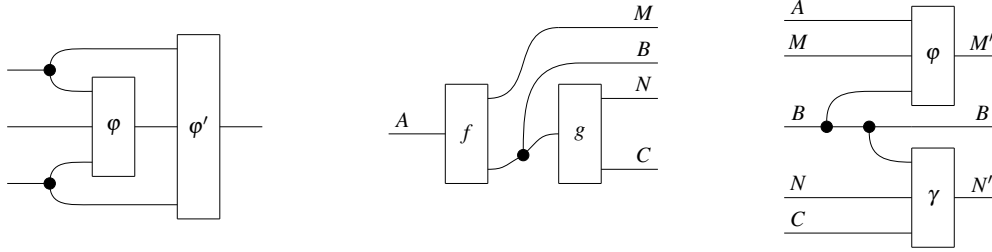


The identity 2-cell $\text{id}_f : f \Rightarrow f$ on $f : A \xrightarrow{M} B$ is given by the projection morphism $\text{proj}_M : A \otimes M \otimes B \rightarrow M$ obtained by discarding A and B , as in footnote 2. The identity 1-cell id_A on A is given by the inverse of the left unitor of the monoidal structure on \mathcal{C} , i.e. $\text{id}_A := \lambda_A^{-1} : A \xrightarrow{I} A$, with coparameter thus given by the unit object I .

Given 2-cells $\varphi : f \Rightarrow f'$ and $\varphi' : f' \Rightarrow f''$, their vertical composite $\varphi' \circ \varphi : f \Rightarrow f''$ is given by the string diagram on the left below. Given 1-cells $f : A \xrightarrow{M} B$ then $g : B \xrightarrow{N} C$, the horizontal composite $g \circ f : A \xrightarrow{(M \otimes B) \otimes N} C$ is given by the middle string diagram below. Given 2-cells $\varphi : f \Rightarrow f'$ and $\gamma : g \Rightarrow g'$ between 1-cells $f, f' : A \xrightarrow{M} B$ and $g, g' : B \xrightarrow{N} C$, their horizontal composite $\gamma \circ \varphi : (g \circ f) \Rightarrow (g' \circ f')$ is

² We define $\text{proj}_C := B \otimes C \xrightarrow{\text{id}_B \otimes \text{id}_C} I \otimes C \xrightarrow{\lambda_C} C$, using the comonoid counit and the left unitor of \mathcal{C} 's monoidal structure.

defined by the string diagram on the right below.



Remark 2.2. When \mathcal{C} is symmetric monoidal, $\mathbf{Copara}_2(\mathcal{C})$ inherits a monoidal structure, elaborated in Proposition B.1.

Remark 2.3. In order to capture the bidirectionality of Bayesian inversion we will need to consider left- and right-handed versions of the \mathbf{Copara}_2 construction. These are formally dual, and when \mathcal{C} is symmetric monoidal (as in most examples) they are isomorphic. Nonetheless, it makes formalization easier if we explicitly distinguish $\mathbf{Copara}_2^l(\mathcal{C})$, in which the coparameter is placed on the left of the codomain (as above), from $\mathbf{Copara}_2^r(\mathcal{C})$, in which it is placed on the right. Aside from the swapping of this handedness, the rest of the construction is the same.

We end this section with three easy (and ambidextrous) propositions relating \mathcal{C} and $\mathbf{Copara}_2(\mathcal{C})$.

Proposition 2.4. There is an identity-on-objects lax embedding $\iota : \mathcal{C} \hookrightarrow \mathbf{Copara}_2(\mathcal{C})$, mapping $f : X \rightarrow Y$ to $f : X \xrightarrow{I} Y$ (using the unitor of the monoidal structure on \mathcal{C}). The laxator $\iota(g) \circ \iota(f) \rightarrow \iota(g \circ f)$ discards the coparameter obtained from copy-composition.

Proposition 2.5. There is a ‘discarding’ functor $(-)^{\dagger} : \mathbf{Copara}_2(\mathcal{C}) \rightarrow \mathcal{C}$, which takes any coparameterized morphism and discards the coparameter.

Proposition 2.6. ι is a section of $(-)^{\dagger}$. That is, $\text{id}_{\mathcal{C}} = \mathcal{C} \xrightarrow{\iota} \mathbf{Copara}_2(\mathcal{C}) \xrightarrow{(-)^{\dagger}} \mathcal{C}$.

2.2 Coparameterized Bayesian lenses

In order to define (bi)categories of statistical games, coherently with loss functions like the relative entropy that compose by copy-composition, we first need to define coparameterized (copy-composite) Bayesian lenses. Analogously to non-coparameterized Bayesian lenses, these will be obtained by applying a Grothendieck construction to an indexed bicategory [5, Def. 3.5] of state-dependent channels.

Definition 2.7. We define the indexed bicategory $\text{Stat}_2 : \mathbf{Copara}_2^l(\mathcal{C})^{\text{cop}} \rightarrow \mathbf{Bicat}$ fibrewise as follows.

- (i) The 0-cells $\text{Stat}_2(X)_0$ of each fibre $\text{Stat}_2(X)$ are the objects \mathcal{C}_0 of \mathcal{C} .
- (ii) For each pair of 0-cells A, B , the hom-category $\text{Stat}_2(X)(A, B)$ is defined to be the functor category $\mathbf{Cat}(\text{disc } \mathcal{C}(I, X), \mathbf{Copara}_2^r(\mathcal{C})(A, B))$, where disc denotes the functor taking a set to the associated discrete category.
- (iii) For each 0-cell A , the identity functor $\text{id}_A : \mathbf{1} \rightarrow \text{Stat}_2(X)(A, A)$ is the constant functor on the identity on A in $\mathbf{Copara}_2^r(\mathcal{C})$; i.e. $\text{disc } \mathcal{C}(I, X) \xrightarrow{!} \mathbf{1} \xrightarrow{\text{id}_A} \mathbf{Copara}_2^r(\mathcal{C})(A, A)$.

(iv) For each triple A, B, C of 0-cells, the horizontal composition functor $\circ_{A,B,C}$ is defined by

$$\begin{aligned} & \circ_{A,B,C} : \text{Stat}_2(X)(B, C) \times \text{Stat}_2(X)(A, B) \cdots \\ \cdots & \xrightarrow{=} \mathbf{Cat}(\text{disc } \mathcal{C}(I, X), \mathbf{Copara}_2^r(\mathcal{C})(B, C)) \times \mathbf{Cat}(\text{disc } \mathcal{C}(I, X), \mathbf{Copara}_2^r(\mathcal{C})(A, B)) \cdots \\ & \cdots \xrightarrow{\times} \mathbf{Cat}(\text{disc } \mathcal{C}(I, X)^2, \mathbf{Copara}_2^r(\mathcal{C})(B, C) \times \mathbf{Copara}_2^r(\mathcal{C})(A, B)) \cdots \\ & \cdots \xrightarrow{\mathbf{Cat}(\blacktriangledown, \circ)} \mathbf{Cat}(\text{disc } \mathcal{C}(I, X), \mathbf{Copara}_2^r(\mathcal{C})(A, C)) \cdots \\ & \cdots \xrightarrow{=} \text{Stat}_2(X)(A, C) \end{aligned}$$

where $\mathbf{Cat}(\blacktriangledown, \circ)$ indicates pre-composition with the universal (Cartesian) copying functor in $(\mathbf{Cat}, \times, \mathbf{1})$ and post-composition with the horizontal composition functor of $\mathbf{Copara}_2^r(\mathcal{C})$.

For each pair of 0-cells X, Y in $\mathbf{Copara}^l(\mathcal{C})$, we define the reindexing pseudofunctor $\text{Stat}_{2,X,Y} : \mathbf{Copara}^l(\mathcal{C})(X, Y)^{\text{op}} \rightarrow \mathbf{Bicat}(\text{Stat}_2(Y), \text{Stat}_2(X))$ as follows.

- (a) For each 1-cell f in $\mathbf{Copara}^l(\mathcal{C})(X, Y)$, we obtain a pseudofunctor $\text{Stat}_2(f) : \text{Stat}_2(Y) \rightarrow \text{Stat}_2(X)$ which acts as the identity on 0-cells.
- (b) For each pair of 0-cells A, B in $\text{Stat}_2(Y)$, the functor $\text{Stat}_2(f)_{A,B}$ is defined as the precomposition functor $\mathbf{Cat}(\text{disc } \mathcal{C}(I, f^*), \mathbf{Copara}_2^r(\mathcal{C})(A, B))$, where $(-)^*$ is the discarding functor $\mathbf{Copara}_2^l(\mathcal{C}) \rightarrow \mathcal{C}$ of Proposition 2.5.
- (c) For each 2-cell $\varphi : f \Rightarrow f'$ in $\mathbf{Copara}^l(\mathcal{C})(X, Y)$, the pseudonatural transformation $\text{Stat}_2(\varphi) : \text{Stat}_2(f') \Rightarrow \text{Stat}_2(f)$ is defined on 0-cells $A : \text{Stat}_2(Y)$ by the discrete natural transformation with components $\text{Stat}_2(\varphi)_A := \text{id}_A$, and on 1-cells $c : \text{Stat}_2(Y)(A, B)$ by the substitution natural transformation with constituent 2-cells $\text{Stat}_2(\varphi)_c : \text{Stat}_2(f)(c) \Rightarrow \text{Stat}_2(f')(c)$ in $\text{Stat}_2(X)$ which acts by replacing $\mathbf{Cat}(\text{disc } \mathcal{C}(I, f^*), \mathbf{Copara}_2^r(\mathcal{C})(A, B))$ by $\mathbf{Cat}(\text{disc } \mathcal{C}(I, f'^*), \mathbf{Copara}_2^r(\mathcal{C})(A, B))$; and which we might alternatively denote by $\mathbf{Cat}(\text{disc } \mathcal{C}(I, \varphi^*), \mathbf{Copara}_2^r(\mathcal{C})(A, B))$.

Notation 2.8. We will write $f : A \xrightarrow[X]{M} B$ to denote a state-dependent coparameterized channel f with coparameter M and state-dependence on X .

In 1-category theory, lenses are morphisms in the fibrewise opposite of a fibration [8]. Analogously, our bicategorical Bayesian lenses are obtained as 1-cells in the bicategorical Grothendieck construction [5, §6] of (the pointwise opposite of) the indexed bicategory Stat_2 .

Definition 2.9. Fix a copy-discard category $(\mathcal{C}, \otimes, I)$. We define the bicategory of coparameterized Bayesian lenses in \mathcal{C} , denoted $\mathbf{BayesLens}_2(\mathcal{C})$ or simply $\mathbf{BayesLens}_2$, to be the bicategorical Grothendieck construction of the pointwise opposite of the corresponding indexed bicategory Stat_2 , with the following data:

- (i) A 0-cell in $\mathbf{BayesLens}_2$ is a pair (X, A) of an object X in $\mathbf{Copara}_2^l(\mathcal{C})$ and an object A in $\text{Stat}_2(X)$; equivalently, a 0-cell in $\mathbf{BayesLens}_2$ is a pair of objects in \mathcal{C} .
- (ii) The hom-category $\mathbf{BayesLens}_2((X, A), (Y, B))$ is the product category $\mathbf{Copara}_2^l(\mathcal{C})(X, Y) \times \text{Stat}_2(X)(B, A)$.
- (iii) The identity on (X, A) is given by the pair $(\text{id}_X, \text{id}_A)$.

(iv) For each triple of 0-cells $(X, A), (Y, B), (Z, C)$, the horizontal composition functor is given by

$$\begin{aligned}
 & \mathbf{BayesLens}_2((Y, B), (Z, C)) \times \mathbf{BayesLens}_2((X, A), (Y, B)) \\
 = & \mathbf{Copara}_2^l(\mathcal{C})(Y, Z) \times \mathbf{Stat}_2(Y)(C, B) \times \mathbf{Copara}_2^l(\mathcal{C})(X, Y) \times \mathbf{Stat}_2(X)(B, A) \\
 \xrightarrow{\sim} & \sum_{g: \mathbf{Copara}_2^l(\mathcal{C})(Y, Z)} \sum_{f: \mathbf{Copara}_2^l(\mathcal{C})(X, Y)} \mathbf{Stat}_2(Y)(C, B) \times \mathbf{Stat}_2(X)(B, A) \\
 \xrightarrow{\sum_g \sum_f \mathbf{Stat}_2(f)_{C, B} \times \text{id}} & \sum_{g: \mathbf{Copara}_2^l(\mathcal{C})(Y, Z)} \sum_{f: \mathbf{Copara}_2^l(\mathcal{C})(X, Y)} \mathbf{Stat}_2(X)(C, B) \times \mathbf{Stat}_2(X)(B, A) \\
 \xrightarrow{\sum_{\circ} \mathbf{Copara}_2^l(\mathcal{C}) \circ \mathbf{Stat}_2(X)} & \sum_{g \circ f: \mathbf{Copara}_2^l(\mathcal{C})(X, Z)} \mathbf{Stat}_2(X)(C, A) \\
 \xrightarrow{\sim} & \mathbf{BayesLens}_2((X, A), (Z, C))
 \end{aligned}$$

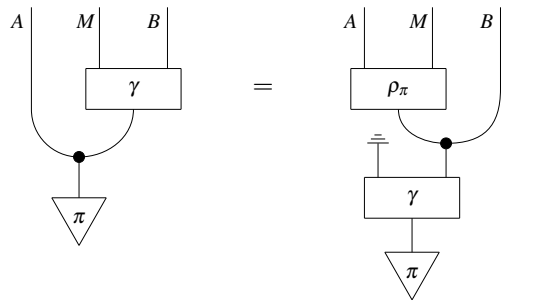
where the functor in the penultimate line amounts to the product of the horizontal composition functors on $\mathbf{Copara}_2^l(\mathcal{C})$ and $\mathbf{Stat}_2(X)$.

Remark 2.10. When \mathcal{C} is symmetric monoidal, \mathbf{Stat}_2 acquires the structure of a monoidal indexed bicategory (Definition B.2 and Theorem B.3), and hence $\mathbf{BayesLens}_2$ becomes a monoidal bicategory (Corollary B.4).

2.3 Coparameterized Bayesian updates compose optically

So that our generalized Bayesian lenses are worthy of the name, we should also confirm that Bayesian inversions compose according to the lens pattern (‘optically’) also in the coparameterized setting. Such confirmation is the subject of the present section, and so we first introduce a new “coparameterized Bayes’ rule”.

Definition 2.11. We say that a coparameterized channel $\gamma : A \rightarrow M \otimes B$ admits Bayesian inversion if there exists a dually coparameterized channel $\rho_\pi : B \rightarrow A \otimes M$ satisfying the graphical equation (with string diagrams read from bottom to top)



In this case, we say that ρ_π is the Bayesian inversion of γ with respect to π .

With this definition, we can supply the desired result that “coparameterized Bayesian updates compose optically”.

Theorem 2.12. Suppose $(\gamma, \gamma^\dagger) : (A, A) \rightarrow (B, B)$ and $(\delta, \delta^\dagger) : (B, B) \rightarrow (C, C)$ are coparameterized Bayesian lenses in $\mathbf{BayesLens}_2$. Suppose also that $\pi : I \rightarrow A$ is a state on A in the underlying category of channels \mathcal{C} , such that γ_π^\dagger is a Bayesian inversion of γ with respect to π , and such that $\delta_{\gamma_\pi^\dagger}^\dagger$ is a Bayesian

inversion of δ with respect to $(\gamma\pi)^\dagger$; where the notation $(-)^{\dagger}$ represents discarding coparameters. Then $\gamma_\pi^\dagger \bullet \delta_{\gamma\pi}^\dagger$ is a Bayesian inversion of $\delta \bullet \gamma$ with respect to π . (Here \bullet denotes copy-composition.) Moreover, if $(\delta \bullet \gamma)_\pi^\dagger$ is any Bayesian inversion of $\delta \bullet \gamma$ with respect to π , then $\gamma_\pi^\dagger \bullet \delta_{\gamma\pi}^\dagger$ is $(\delta\gamma\pi)^\dagger$ -almost-surely equal to $(\delta \bullet \gamma)_\pi^\dagger$: that is, $(\delta \bullet \gamma)_\pi^\dagger \stackrel{(\delta\gamma\pi)^\dagger}{\sim} \gamma_\pi^\dagger \bullet \delta_{\gamma\pi}^\dagger$.

In order to satisfy this coparameterized Bayes' rule, a Bayesian lens must be of 'simple' type.

Definition 2.13. We say that a coparameterized Bayesian lens (c, c') is *simple* if its domain and codomain are 'diagonal' (duplicate pairs of objects) and if the coparameter of c is equal to the coparameter of c' . In this case, we can write the type of (c, c') as $(X, X) \xrightarrow[M]{\rightarrow} (Y, Y)$ or simply $X \xrightarrow[M]{\rightarrow} Y$.

3 Statistical games for local approximate inference

3.1 Losses for lenses

Statistical games are obtained by attaching to Bayesian lenses *loss functions*, representing 'local' quantifications of the performance of approximate inference systems. Because this performance depends on the system's context (*i.e.*, the prior $\pi : I \rightarrow X$ and the observed data $b : B$), a loss function at its most concrete will be a function $\mathcal{C}(I, X) \times B \rightarrow \mathbb{R}_+$. To internalize this type in \mathcal{C} , we may recall that, when \mathcal{C} is the category **sfKrn** of s-finite kernels or the Kleisli category $\mathcal{Kl}(\mathcal{D}_{\leq 1})$ of the subdistribution monad, a density function $p_c : X \times Y \rightarrow [0, 1]$ for a channel $c : X \rightarrow Y$ corresponds to an *effect* (or *costate*) $X \otimes Y \rightarrow I$. In this way, we can see a loss function as a kind of *state-dependent effect* $B \xrightarrow{X} I$.

Loss functions will compose by sum, and so we need to ask for the effects in \mathcal{C} to form a monoid. Moreover, we need this monoid to be 'bilinear' with respect to channels, so that Stat-reindexing (*cf.* Definition A.1) preserves sums. These conditions are formalized in the following definition.

Definition 3.1. Suppose $(\mathcal{C}, \otimes, I)$ is a copy-discard category. We say that \mathcal{C} has *bilinear effects* if the following conditions are satisfied:

- (i) *effect monoid*: there is a natural transformation $+$: $\mathcal{C}(-, I) \times \mathcal{C}(-, I) \Rightarrow \mathcal{C}(- \otimes -, I)$ making $\sum_{A:\mathcal{C}} \mathcal{C}(A, I)$ into a commutative monoid with unit $0 : I \rightarrow I$;
- (ii) *bilinearity*: $(g + g') \bullet \forall \bullet f = g \bullet f + g' \bullet f$ for all effects g, g' and morphisms f such that $(g + g') \bullet \forall \bullet f$ exists.

A trivial example of a category with bilinear effects is supplied by any Cartesian category, such as **Set**. If M is any monoid in **Set**, then a less trivial example is supplied by the Kleisli category of the corresponding free module monad; bilinearity follows from the module structure. A related non-example is $\mathcal{Kl}(\mathcal{D}_{\leq 1})$: the failure here is that the effects only form a *partial* monoid³. More generally, the category **sfKrn** of s-finite kernels [10] has bilinear effects (owing to the linearity of integration), and we will assume this as our ambient \mathcal{C} for the examples below.

Given such a category \mathcal{C} with bilinear effects, we can lift the natural transformation $+$, and hence the

³Indeed, an *effect algebra* is a kind of partial monoid [9, §2], but we do not need the extra complication here.

bilinear effect structure, to the fibres of $\text{Stat}_{\mathcal{C}}$, using the universal property of the product of categories:

$$\begin{aligned}
+_X : \text{Stat}(X)(-, I) \times \text{Stat}(X)(=, I) &= \mathbf{Set}(\mathcal{C}(I, X), \mathcal{C}(-, I)) \times \mathbf{Set}(\mathcal{C}(I, X), \mathcal{C}(=, I)) \\
&\xrightarrow{(\cdot, \cdot)} \mathbf{Set}(\mathcal{C}(I, X), \mathcal{C}(-, I) \times \mathcal{C}(=, I)) \\
&\xrightarrow{\mathbf{Set}(\mathcal{C}(I, X), +)} \mathbf{Set}(\mathcal{C}(I, X), \mathcal{C}(- \otimes =, I)) \\
&\xrightarrow{=} \text{Stat}(X)(- \otimes =, I)
\end{aligned}$$

Here, (\cdot, \cdot) denotes the pairing operation obtained from the universal property. In this way, each $\text{Stat}(X)$ has bilinear effects. Note that this lifting is (strictly) compatible with the reindexing of Stat , so that $+_{(-)}$ defines an indexed natural transformation. This means in particular that *reindexing distributes over sums*: given state-dependent effects $g, g' : B \rightarrow I$ and a channel $c : X \rightarrow Y$, we have $(g +_Y g')_c = g_c +_X g'_c$. We will thus generally omit the subscript from the lifted sum operation, and just write $+$.

We are now ready to construct the bicategory of statistical games.

Definition 3.2. Suppose $(\mathcal{C}, \otimes, I)$ has bilinear effects, and let $\mathbf{BayesLens}_2$ denote the corresponding bicategory of (copy-composite) Bayesian lenses. We will write $\mathbf{SGame}_{\mathcal{C}}$ to denote the following bicategory of (copy-composite) statistical games in \mathcal{C} :

- The 0-cells are the 0-cells (X, A) of $\mathbf{BayesLens}_2$;
- the 1-cells, called *statistical games*, $(X, A) \rightarrow (Y, B)$ are pairs (c, L^c) of a 1-cell $c : (X, A) \rightarrow (Y, B)$ in $\mathbf{BayesLens}_2$ and a loss $L^c : B \rightarrow I$ in $\text{Stat}(X)(B, I)$;
- given 1-cells $(c, L^c), (c', L^{c'}) : (X, A) \rightarrow (Y, B)$, the 2-cells $(c, L^c) \Rightarrow (c', L^{c'})$ are pairs (α, K^α) of a 2-cell $\alpha : c \Rightarrow c'$ in $\mathbf{BayesLens}_2$ and a loss $K^\alpha : B \rightarrow I$ such that $L^c = L^{c'} + K^\alpha$;
- the identity 2-cell on (c, L^c) is $(\text{id}_c, 0)$;
- given 2-cells $(\alpha, K^\alpha) : (c, L^c) \Rightarrow (c', L^{c'})$ and $(\alpha', K^{\alpha'}) : (c', L^{c'}) \Rightarrow (c'', L^{c''})$, their vertical composite is $(\alpha' \circ \alpha, K^{\alpha'} + K^\alpha)$, where \circ here denotes vertical composition in $\mathbf{BayesLens}_2$;
- given 1-cells $(c, L^c) : (X, A) \rightarrow (Y, B)$ and $(d, L^d) : (Y, B) \rightarrow (Z, C)$, their horizontal composite is $(c \circ d, L_c^d + L^c \circ \bar{d}_c)$; and
 - given 2-cells $(\alpha, K^\alpha) : (c, L^c) \Rightarrow (c', L^{c'})$ and $(\beta, K^\beta) : (d, L^d) \Rightarrow (d', L^{d'})$, their horizontal composite is $(\beta \circ \alpha, K_c^\beta + K^\alpha \circ \bar{d}_c)$, where \circ here denotes horizontal composition in $\mathbf{BayesLens}_2$.

Theorem 3.3. Definition 3.2 generates a well-defined bicategory.

The proof of this result is that $\mathbf{SGame}_{\mathcal{C}}$ is obtained via a pair of bicategorical Grothendieck constructions [5]: first to obtain Bayesian lenses; and then to attach the loss functions. The proof depends on the intermediate result that our effect monoids can be ‘upgraded’ to monoidal categories; we then use the delooping of this structure to associate (state-dependent) losses to (state-dependent) channels, after discarding the coparameters of the latter.

Lemma 3.4. Suppose $(\mathcal{C}, \otimes, I)$ has bilinear effects. Then, for each object B , $\mathcal{C}(B, I)$ has the structure of a symmetric monoidal category. The objects of $\mathcal{C}(B, I)$ are its elements, the effects. If g, g' are two effects, then a morphism $\kappa : g \rightarrow g'$ is an effect such that $g = g' + \kappa$; the identity morphism for each effect id_g is then the constant 0 effect. Likewise, the tensor of two effects is their sum, and the corresponding unit is the constant 0. Precomposition by any morphism $c : A \rightarrow B$ preserves the monoidal category structure, making the presheaf $\mathcal{C}(-, I)$ into a fibrewise-monoidal indexed category $\mathcal{C}^{\text{op}} \rightarrow \mathbf{MonCat}$.

As already indicated, this structure lifts to the fibres of Stat .

Corollary 3.5. For each object X in a category with bilinear effects, and for each object B , $\text{Stat}(X)(B, I)$ inherits the symmetric monoidal structure of $\mathcal{C}(B, I)$; note that morphisms of state-dependent effects are likewise state-dependent, and that tensoring (summing) state-dependent effects involves copying the parameterizing state. Moreover, $\text{Stat}(-)(=, I)$ is a fibrewise-monoidal indexed category $\sum_{X:\mathcal{C}^{\text{op}}} \text{Stat}(X)^{\text{op}} \rightarrow \mathbf{MonCat}$.

3.2 Local inference models

In the context of approximate inference, one often does not have a single statistical model to evaluate, but a whole family of them. In particularly nice situations, this family is actually a subcategory \mathcal{D} of \mathcal{C} , with the family of statistical models being all those that can be composed in \mathcal{D} . The problem of approximate inference can then be formalized as follows. Since both $\mathbf{BayesLens}_2$ and $\mathbf{SGame}_{\mathcal{C}}$ were obtained by bicategorical Grothendieck constructions, we have a pair of 2-fibrations $\mathbf{SGame}_{\mathcal{C}} \xrightarrow{\pi_{\text{Loss}}} \mathbf{BayesLens}_2 \xrightarrow{\pi_{\text{Lens}}} \mathbf{Copara}_2^l(\mathcal{C})$. Each of π_{Loss} , π_{Lens} , and the discarding functor $(-)^{\ddagger}$ can be restricted to the subcategory \mathcal{D} . The inclusion $\mathcal{D} \hookrightarrow \mathbf{Copara}_2^l(\mathcal{D})$ is a section of this restriction of $(-)^{\ddagger}$; the assignment of inversions to channels in \mathcal{D} then corresponds to a 2-section of the 2-fibration π_{Lens} (restricted to \mathcal{D}); and the subsequent assignment of losses is a further 2-section of π_{Loss} . This situation is depicted in the following diagram of bicategories:

$$\begin{array}{ccc}
 \mathbf{SGame}_{\mathcal{D}} & \hookrightarrow & \mathbf{SGame}_{\mathcal{C}} \\
 \downarrow \pi_{\text{Loss}|_{\mathcal{D}}} & & \downarrow \pi_{\text{Loss}} \\
 \mathbf{BayesLens}_2|_{\mathcal{D}} & \hookrightarrow & \mathbf{BayesLens}_2 \\
 \downarrow \pi_{\text{Lens}|_{\mathcal{D}}} & & \downarrow \pi_{\text{Lens}} \\
 \mathbf{Copara}_2^l(\mathcal{D}) & \hookrightarrow & \mathbf{Copara}_2^l(\mathcal{C}) \\
 \downarrow \ddagger|_{\mathcal{D}} & & \downarrow \ddagger \\
 \mathcal{D} & \hookrightarrow & \mathcal{C}
 \end{array}$$

This motivates the following definitions of *inference system* and *loss model*, although, for the sake of our examples, we will explicitly allow the loss-assignment to be lax. Before giving these new definitions, we recall the notion of *essential image* of a functor.

Definition 3.6 ([11]). Suppose $F : \mathcal{C} \rightarrow \mathcal{D}$ is an n -functor (a possibly weak homomorphism of weak n -categories). The *image* of F is the smallest sub- n -category of \mathcal{D} that contains $F(\alpha)$ for all k -cells α in \mathcal{C} , along with any $(k+1)$ -cells relating images of composites and composites of images, for all $0 \leq k \leq n$. We say that a sub- n -category \mathcal{D} is *replete* if, for any k -cells α in \mathcal{D} and β in \mathcal{C} (with $0 \leq k < n$) such that $f : \alpha \Rightarrow \beta$ is a $(k+1)$ -isomorphism in \mathcal{C} , then f is also a $(k+1)$ -isomorphism in \mathcal{D} . The *essential image* of F , denoted $\text{im}(F)$, is then the smallest replete sub- n -category of \mathcal{D} containing the image of F .

Definition 3.7. Suppose $(\mathcal{C}, \otimes, I)$ is a copy-delete category. An *inference system* in \mathcal{C} is a pair (\mathcal{D}, ℓ) of a subcategory $\mathcal{D} \hookrightarrow \mathcal{C}$ along with a section $\ell : \text{im}(\iota) \rightarrow \mathbf{BayesLens}_2|_{\mathcal{D}}$ of $\pi_{\text{Lens}}|_{\mathcal{D}}$, where $\text{im}(\iota)$ is the essential image of the canonical lax inclusion $\iota : \mathcal{D} \hookrightarrow \mathbf{Copara}_2^l(\mathcal{D})$.

Definition 3.8. Suppose $(\mathcal{C}, \otimes, I)$ has bilinear effects and \mathcal{B} is a subcategory of $\mathbf{BayesLens}_2$. A *loss model* for \mathcal{B} is a lax section L of the restriction $\pi_{\text{Loss}}|_{\mathcal{B}}$ of π_{Loss} to \mathcal{B} . We say that L is a *strict* loss model if it is in fact a strict 2-functor, and a *strong* loss model if it is in fact a pseudofunctor.

Remark 3.9. We may often be interested in loss models for which \mathcal{B} is in fact the essential image of an inference system, but we do not stipulate this requirement in the definition as it is not necessary for the following development.

Since lax functors themselves collect into categories, and using the monoidality of $+$, we have the following easy proposition that will prove useful below.

Proposition 3.10. Loss models for \mathcal{B} constitute the objects of a symmetric monoidal category $(\text{Loss}(\mathcal{B}), +, 0)$. The morphisms of $\text{Loss}(\mathcal{B})$ are icons (identity component oplax transformations [6, §4.6]) between the corresponding lax functors, and they compose accordingly. The monoidal structure is given by sums of losses.

3.3 Examples

Each of our examples involves taking expectations of log-densities, and so to make sense of them it first helps to understand what we mean by “taking expectations”.

Notation 3.11 (Expectations). Written as a function, a density p on X has the type $X \rightarrow \mathbb{R}_+$; written as an effect, the type is $X \multimap I$. Given a measure or distribution π on X (equivalently, a state $\pi : I \multimap X$), we can compute the expectation of p under π as the composite $p \bullet \pi$. We write the resulting quantity as $\mathbb{E}_\pi[p]$, or more explicitly as $\mathbb{E}_{x \sim \pi}[p(x)]$. We can think of this expectation as representing the ‘validity’ (or truth value) of the ‘predicate’ p given the state π [12].

3.3.1 Relative entropy and Bayesian inference

For our first example, we return to the subject with which we opened this paper: the compositional structure of the relative entropy. We begin by giving a precise definition.

Definition 3.12. Suppose α, β are both measures on X , with α absolutely continuous with respect to β . Then the *relative entropy* or *Kullback-Leibler divergence* from α to β is the quantity $D_{KL}(\alpha, \beta) := \mathbb{E}_\alpha \left[\log \frac{\alpha}{\beta} \right]$, where $\frac{\alpha}{\beta}$ is the Radon-Nikodym derivative of α with respect to β .

Remark 3.13. When α and β admit density functions p_α and p_β with respect to the same base measure dx , then $D_{KL}(\alpha, \beta)$ can equally be computed as $\mathbb{E}_{x \sim \alpha} [\log p_\alpha(x) - \log p_\beta(x)]$. It is in this form that we will adopt henceforth.

Proposition 3.14. Let \mathcal{B} be a subcategory of simple lenses in **BayesLens**₂, all of whose channels admit density functions with respect to a common measure and whose forward channels admit Bayesian inversion (and whose forward and backward coparameters coincide), and with only structural 2-cells. Then the relative entropy defines a strict loss model $\text{KL} : \mathcal{B} \rightarrow \mathbf{SGame}$. Given a lens $(c, c') : (X, X) \multimap (Y, Y)$, KL assigns the loss function $\text{KL}(c, c') : Y \overset{X}{\multimap} I$ defined, for $\pi : I \multimap X$ and $y : Y$, by the relative entropy $\text{KL}(c, c')_\pi(y) := D_{KL}(c'_\pi(y), c_\pi^\dagger(y))$, where c^\dagger is the exact inversion of c .

Successfully playing a relative entropy game entails minimizing the divergence from the approximate to the exact posterior. This divergence is minimized when the two coincide, and so KL represents a form of approximate Bayesian inference.

3.3.2 Maximum likelihood estimation

A statistical system may be more interested in predicting observations than updating beliefs. This is captured by the process of *maximum likelihood estimation*.

Definition 3.15. Let $(c, c') : (X, X) \multimap (Y, Y)$ be a simple lens whose forward channel c admits a density function p_c . Then its *log-likelihood* is the loss function defined by $\text{MLE}(c, c')_\pi(y) := -\log p_{c^\dagger \bullet \pi}(y)$.

Proposition 3.16. Let \mathcal{B} be a subcategory of lenses in $\mathbf{BayesLens}_2$ all of which admit density functions with respect to a common measure, and with only structural 2-cells. Then the assignment $(c, c') \mapsto \text{MLE}(c, c')$ defines a lax loss model $\text{MLE} : \mathcal{B} \rightarrow \mathbf{SGame}$.

Successfully playing a maximum likelihood game involves maximizing the log-likelihood that the system assigns to its observations $y : Y$. This process amounts to choosing a channel c that assigns high likelihood to likely observations, and thus encodes a valid model of the data distribution.

3.3.3 Autoencoders via the free energy

Many adaptive systems neither just infer nor just predict: they do both, building a model of their observations that they also invert to update their beliefs. In machine learning, such systems are known as *autoencoders*, as they ‘encode’ (infer) and ‘decode’ (predict), ‘autoassociatively’ [13]. In a Bayesian context, they are known as *variational autoencoders* [14], and their loss function is the *free energy* [15].

Definition 3.17. The *free energy* loss model is the sum of the relative entropy and the likelihood loss models: $\text{FE} := \text{KL} + \text{MLE}$. Given a simple lens $(c, c') : (X, X) \leftrightarrow (Y, Y)$ admitting Bayesian inversion and with densities, FE assigns the loss function

$$\begin{aligned} \text{FE}(c, c')_{\pi}(y) &= (\text{KL} + \text{MLE})(c, c')_{\pi}(y) \\ &= D_{KL}(c'_{\pi}(y), c_{\pi}^{\dagger}(y)) - \log p_{c^{\ddagger} \bullet \pi}(y) \end{aligned}$$

Remark 3.18. Beyond its autoencoding impetus, another important property of the free energy is its improved computational tractability compared to either the relative entropy or the likelihood loss. This property is a consequence of the following fact: although obtained as the sum of terms which both depend on an expensive marginalization⁴, the free energy itself does not. This can be seen by expanding the definitions of the relative entropy and of c_{π}^{\dagger} and rearranging terms:

$$\begin{aligned} \text{FE}(c, c')_{\pi}(y) &= D_{KL}(c'_{\pi}(y), c_{\pi}^{\dagger}(y)) - \log p_{c^{\ddagger} \bullet \pi}(y) \\ &= \mathbb{E}_{(x,m) \sim c'_{\pi}(y)} [\log p_{c'_{\pi}}(x, m|y) - \log p_{c_{\pi}^{\dagger}}(x, m|y)] - \log p_{c^{\ddagger} \bullet \pi}(y) \\ &= \mathbb{E}_{(x,m) \sim c'_{\pi}(y)} [\log p_{c'_{\pi}}(x, m|y) - \log p_{c_{\pi}^{\dagger}}(x, m|y) - \log p_{c^{\ddagger} \bullet \pi}(y)] \\ &= \mathbb{E}_{(x,m) \sim c'_{\pi}(y)} \left[\log p_{c'_{\pi}}(x, m|y) - \log \frac{p_c(m, y|x) p_{\pi}(x)}{p_{c^{\ddagger} \bullet \pi}(y)} - \log p_{c^{\ddagger} \bullet \pi}(y) \right] \\ &= \mathbb{E}_{(x,m) \sim c'_{\pi}(y)} [\log p_{c'_{\pi}}(x, m|y) - \log p_c(m, y|x) - \log p_{\pi}(x)] \\ &= D_{KL}(c'_{\pi}(y), \pi \otimes \mathbf{1}) - \mathbb{E}_{(x,m) \sim c'_{\pi}(y)} [\log p_c(m, y|x)] \end{aligned}$$

Here, $\mathbf{1}$ denotes the measure with density 1 everywhere. Note that when the coparameter is trivial, $\text{FE}(c, c')_{\pi}(y)$ reduces to

$$D_{KL}(c'_{\pi}(y), \pi) - \mathbb{E}_{x \sim c'_{\pi}(y)} [\log p_c(y|x)].$$

Remark 3.19. The name *free energy* is due to an analogy with the Helmholtz free energy in thermodynamics, as we can write it as the difference between an (expected) energy and an entropy

⁴Evaluating the pushforward $c^{\ddagger} \bullet \pi$ involves marginalizing over the intermediate variable; and evaluating $c_{\pi}^{\dagger}(y)$ also involves evaluating $c^{\ddagger} \bullet \pi$.

term:

$$\begin{aligned} \text{FE}(c, c')_{\pi}(y) &= \mathbb{E}_{(x,m) \sim c'_{\pi}(y)} \left[-\log p_c(m, y|x) - \log p_{\pi}(x) \right] - S_{X \otimes M} [c'_{\pi}(y)] \\ &= \mathbb{E}_{(x,m) \sim c'_{\pi}(y)} \left[E_{(c,\pi)}(x, m, y) \right] - S_{X \otimes M} [c'_{\pi}(y)] = U - TS \end{aligned}$$

where we call $E_{(c,\pi)} : X \otimes M \otimes Y \xrightarrow{X} I$ the *energy*, and where $S_{X \otimes M} : I \xrightarrow{X \otimes M} I$ is the Shannon entropy. The last equality makes the thermodynamic analogy: U here is the *internal energy* of the system; $T = 1$ is the *temperature*; and S is again the entropy.

3.3.4 The Laplace approximation

Although optimizing the free energy does not necessitate access to exact inversions, it is still necessary to compute an expectation under the approximate inversion, and unless one chooses wisely⁵, this might also be difficult. One such wise choice established in the computational neuroscience literature is the Laplace approximation [17], in which one assumes Gaussian channels and posteriors with small variance. Under these conditions, the expectations can be approximated away.

Definition 3.20. We will say that a channel c is *Gaussian* if $c(x)$ is a Gaussian measure for every x in its domain. We will denote the mean and variance of $c(x)$ by $\mu_c(x)$ and $\Sigma_c(x)$ respectively.

Proposition 3.21 (Laplace approximation). Let the ambient category of channels \mathcal{C} be restricted to that generated by Gaussian channels between finite-dimensional Cartesian spaces, and let \mathcal{B} denote the corresponding restriction of **BayesLens**₂. Suppose $(\gamma, \rho) : (X, X) \leftrightarrow (Y, Y)$ is such a lens, for which, for all $y : Y$ and Gaussian priors $\pi : I \rightarrow X$, the eigenvalues of $\Sigma_{\rho_{\pi}}(y)$ are small. Then the free energy $\text{FE}(\gamma, \rho)_{\pi}(y)$ can be approximated by the *Laplacian free energy*

$$\text{FE}(\gamma, \rho)_{\pi}(y) \approx \text{LFE}(\gamma, \rho)_{\pi}(y) \tag{1}$$

$$:= E_{(\gamma, \pi)}(\mu_{\rho_{\pi}}(y), y) - S_{X \otimes M}[\rho_{\pi}(y)] \tag{2}$$

$$= -\log p_{\gamma}(\mu_{\rho_{\pi}}(y), y) - \log p_{\pi}(\mu_{\rho_{\pi}}(y)|_X) - S_{X \otimes M}[\rho_{\pi}(y)]$$

where we have written the argument of the density p_{γ} in ‘function’ style; where $(-)_X$ denotes the projection onto X ; and where $S_{X \otimes M}[\rho_{\pi}(y)] = \mathbb{E}_{(x,m) \sim \rho_{\pi}(y)} [-\log p_{\rho_{\pi}}(x, m|y)]$ is the Shannon entropy of $\rho_{\pi}(y)$. The approximation is valid when $\Sigma_{\rho_{\pi}}$ satisfies

$$\Sigma_{\rho_{\pi}}(y) = \left(\partial_{(x,m)}^2 E_{(\gamma, \pi)} \right) (\mu_{\rho_{\pi}}(y), y)^{-1}. \tag{3}$$

We call $E_{(\gamma, \pi)}$ the *Laplacian energy*.

Remark 3.22. The usual form of the Laplace model in the literature omits the coparameters. It is of course easy to recover the non-coparameterized form by taking $M = 1$.

Proposition 3.23. Let \mathcal{B} be a subcategory of **BayesLens**₂ of Gaussian lenses whose backward channels have small variance. Then LFE defines a lax loss model $\mathcal{B} \rightarrow \mathbf{SGame}$.

Effectively, this proposition says that, under the stated conditions, the free energy and the Laplacian free energy coincide. Consequently, successfully playing a Laplacian free energy game has the same autoencoding effect.

⁵In machine learning, optimizing variational autoencoders using stochastic gradient descent typically involves a “reparameterization trick” [16, §2.5].

4 Future work

This paper only scratches the surface of the structure of statistical games. One avenue for further investigation is the link between this structure and the similar structure of diegetic open (economic) games [18], a recent reformulation of compositional game theory [19]. Notably, the composition rule for loss functions appears closely related to the Bellman equation, suggesting that algorithms for approximate inference (such as expectation-maximization) and reinforcement learning (such as backward induction) are more than superficially similar.

Another avenue for further investigation concerns mathematical neatness. First, we seek an abstract characterization of copy-composition and **Copara**₂; it has been suggested to us⁶ that the computation by compilers of “static single-assignment form” [20] by compilers may have a similar structure. Second, the explicit constraint defining simple coparameterized Bayesian lenses is inelegant; we expect that using dependent optics [21–23] may help to encode this constraint in the type signature, at the cost of higher-powered mathematical machinery. Finally, we seek further examples of loss models, and more abstract (and hopefully universal) characterizations of those we already have; for example, it is known that the Shannon entropy has a topological origin [24] via a “nonlinear derivation” [25], and we expect that we can follow this connection further.

5 References

- [1] Toby St Clere Smithe. “Cyber Kittens, or Some First Steps Towards Categorical Cybernetics”. In: *Electronic Proceedings in Theoretical Computer Science*. Applied Category Theory 2020. Vol. 333. 02/08/2021, pp. 108–124. DOI: 10.4204/EPTCS.333.8. (Visited on 09/29/2023).
- [2] Toby St Clere Smithe. *Bayesian Updates Compose Optically*. 05/31/2020. arXiv: 2006.01631 [math.CT]. preprint.
- [3] Kenta Cho and Bart Jacobs. “Disintegration and Bayesian Inversion via String Diagrams”. In: *Math. Struct. Comp. Sci.* 29 (2019) 938-971 (08/29/2017). DOI: 10.1017/S0960129518000488.
- [4] Matteo Capucci et al. “Towards Foundations of Categorical Cybernetics”. In: *Electronic Proceedings in Theoretical Computer Science*. Applied Category Theory 2021. Vol. 372. 11/03/2022, pp. 235–248. DOI: 10.4204/EPTCS.372.17. (Visited on 09/29/2023).
- [5] Igor Baković. “Fibrations of Bicategories”. In: (2010). URL: <https://www2.irb.hr/korisnici/ibakovic/groth2fib.pdf>.
- [6] Niles Johnson and Donald Yau. *2-Dimensional Categories*. 06/17/2020. arXiv: 2002.06055 [math]. URL: <http://arxiv.org/abs/2002.06055> (visited on 11/03/2022). preprint.
- [7] Joe Moeller and Christina Vasilakopoulou. “Monoidal Grothendieck Construction”. In: *Theory and Applications of Categories* 35.31 (2020), pp. 1159–1207. arXiv: 1809.00727v2 [math.CT].

⁶This suggestion is due to Owen Lynch.

- [8] David I. Spivak. *Generalized Lens Categories via Functors $C \text{ Op} \rightarrow \text{Cat}$* . 08/06/2019. arXiv: 1908.02202 [math.CT]. preprint.
- [9] Bart Jacobs and Bram Westerbaan. “An Effect-Theoretic Account of Lebesgue Integration”. In: *Electronic Notes in Theoretical Computer Science* 319 (12/2015), pp. 239–253. DOI: 10.1016/j.entcs.2015.12.015.
- [10] Matthijs Vákár and Luke Ong. *On S-Finite Measures and Kernels*. 10/03/2018. arXiv: 1810.01837 [math.PR]. preprint.
- [11] nLab authors. *Essential Image*. 05/2023. URL: <https://ncatlab.org/nlab/show/essential+image>.
- [12] Bart Jacobs and Fabio Zanasi. “The Logical Essentials of Bayesian Reasoning”. In: *Foundations of Probabilistic Programming*. Ed. by Gilles Barthe, Joost-Pieter Katoen, and Alexandra Silva. 1st ed. Cambridge University Press, 12/03/2020, pp. 295–332. ISBN: 978-1-108-77075-0 978-1-108-48851-8. DOI: 10.1017/9781108770750.010. (Visited on 09/29/2023).
- [13] Mark A. Kramer. “Nonlinear Principal Component Analysis Using Autoassociative Neural Networks”. In: *AIChE Journal* 37.2 (02/1991), pp. 233–243. ISSN: 0001-1541, 1547-5905. DOI: 10.1002/aic.690370209. (Visited on 05/02/2023).
- [14] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 12/20/2013. arXiv: 1312.6114 [stat.ML]. preprint.
- [15] Peter Dayan et al. “The Helmholtz Machine”. In: *Neural Computation* 7.5 (09/1995), pp. 889–904. DOI: 10.1162/neco.1995.7.5.889. URL: <http://dx.doi.org/10.1162/neco.1995.7.5.889>.
- [16] Diederik P. Kingma. “Variational Inference & Deep Learning. A New Synthesis”. PhD thesis. University of Amsterdam, 2017. URL: <https://hdl.handle.net/11245.1/8e55e07f-e4be-458f-a929-2f9bc2d169e8>.
- [17] K. Friston et al. “Variational Free Energy and the Laplace Approximation”. In: *Neuroimage* 34.1 (01/2007), pp. 220–234. DOI: 10.1016/j.neuroimage.2006.08.035. pmid: 17055746.
- [18] Matteo Capucci. “Diegetic Representation of Feedback in Open Games”. In: *Electronic Proceedings in Theoretical Computer Science*. Applied Category Theory 2022. Vol. 380. 08/07/2023, pp. 145–158. DOI: 10.4204/EPTCS.380.9. (Visited on 09/29/2023).
- [19] Neil Ghani et al. “Compositional Game Theory”. In: *Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science*. LICS ’18: 33rd Annual ACM/IEEE Symposium on Logic in Computer Science. Oxford United Kingdom: ACM, 07/09/2018, pp. 472–481. ISBN: 978-1-4503-5583-4. DOI: 10.1145/3209108.3209165. (Visited on 09/29/2023).

- [20] Richard A. Kelsey. “A Correspondence between Continuation Passing Style and Static Single Assignment Form”. In: *Papers from the 1995 ACM SIGPLAN Workshop on Intermediate Representations*. POPL95: 22nd ACM Symposium on Principles of Programming Languages. San Francisco California USA: ACM, 03/1995, pp. 13–22. ISBN: 978-0-89791-754-4. DOI: 10.1145/202529.202532. (Visited on 05/02/2023).
- [21] Dylan Braithwaite et al. *Fibre Optics*. 12/21/2021. arXiv: 2112.11145 [math.CT]. preprint.
- [22] Pietro Vertechi. “Dependent Optics”. In: *Electronic Proceedings in Theoretical Computer Science*. Applied Category Theory 2022. Vol. 380. 08/07/2023, pp. 128–144. DOI: 10.4204/EPTCS.380.8. (Visited on 09/29/2023).
- [23] Matteo Capucci. *Seeing Double through Dependent Optics*. 04/22/2022. arXiv: 2204.10708 [math.CT]. preprint.
- [24] Tai-Danae Bradley. “Entropy as a Topological Operad Derivation”. In: *Entropy* 23.9 (09/09/2021), p. 1195. ISSN: 1099-4300. DOI: 10.3390/e23091195. URL: <https://www.mdpi.com/1099-4300/23/9/1195> (visited on 05/02/2023).
- [25] Tom Leinster. *Entropy and Diversity: The Axiomatic Approach*. 1st ed. Cambridge University Press, 04/30/2021. ISBN: 978-1-108-96355-8 978-1-108-83270-0 978-1-108-96557-6. DOI: 10.1017/9781108963558. (Visited on 09/29/2023).

A State-dependent channels

In this section, we review the indexed category $\text{Stat} : \mathcal{C}^{\text{op}} \rightarrow \mathbf{Cat}$ of state-dependent channels in \mathcal{C} , from which Bayesian lenses are obtained. We can think of Stat as a decategorified, non-coparameterized, version of Stat_2 , in which the hom-sets $\text{Stat}(X)(A, B)$ of each fibre are given by $\mathbf{Set}(\mathcal{C}(I, X), \mathcal{C}(A, B))$. Reindexing is again by pre-composition, although simplified as there are now no coparameters to discard.

Definition A.1. Let $(\mathcal{C}, \otimes, I)$ be a monoidal category. Define the \mathcal{C} -state-indexed category $\text{Stat} : \mathcal{C}^{\text{op}} \rightarrow \mathbf{Cat}$ as follows.

$\text{Stat} : \mathcal{C}^{\text{op}} \rightarrow \mathbf{Cat}$

$$X \mapsto \text{Stat}(X) := \left(\begin{array}{lll} \text{Stat}(X)_0 & := & \mathcal{C}_0 \\ \text{Stat}(X)(A, B) & := & \mathbf{Set}(\mathcal{C}(I, X), \mathcal{C}(A, B)) \\ \text{id}_A : \text{Stat}(X)(A, A) & := & \left\{ \begin{array}{l} \text{id}_A : \mathcal{C}(I, X) \rightarrow \mathcal{C}(A, A) \\ \rho \mapsto \text{id}_A \end{array} \right. \end{array} \right) \quad (4)$$

$$f : \mathcal{C}(Y, X) \mapsto \left(\begin{array}{lll} \text{Stat}(f) : \text{Stat}(X) & \rightarrow & \text{Stat}(Y) \\ \text{Stat}(X)_0 & = & \text{Stat}(Y)_0 \\ \mathbf{Set}(\mathcal{C}(I, X), \mathcal{C}(A, B)) & \rightarrow & \mathbf{Set}(\mathcal{C}(I, Y), \mathcal{C}(A, B)) \\ \alpha & \mapsto & f^* \alpha : (\sigma : \mathcal{C}(I, Y)) \mapsto (\alpha(f \bullet \sigma) : \mathcal{C}(A, B)) \end{array} \right)$$

Composition in each fibre $\text{Stat}(X)$ is as in \mathcal{C} . Explicitly, indicating morphisms $\mathcal{C}(I, X) \rightarrow \mathcal{C}(A, B)$ in

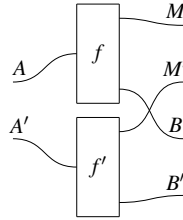
$\text{Stat}(X)$ by $A \overset{X}{\rightarrow} B$, and given $\alpha : A \overset{X}{\rightarrow} B$ and $\beta : B \overset{X}{\rightarrow} C$, their composite is $\beta \circ \alpha : A \overset{X}{\rightarrow} C := \rho \mapsto \beta(\rho) \bullet \alpha(\rho)$, where here we indicate composition in \mathcal{C} by \bullet and composition in the fibres $\text{Stat}(X)$ by \circ . Given $f : Y \rightarrow X$ in \mathcal{C} , the induced functor $\text{Stat}(f) : \text{Stat}(X) \rightarrow \text{Stat}(Y)$ acts by pre-composition.

The category of non-coparameterized Bayesian lenses is then obtained as the (1-categorical) Grothendieck construction of the pointwise opposite of Stat , following Spivak [8].

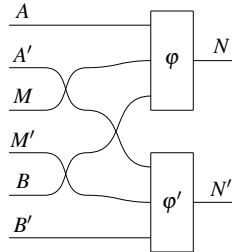
B Monoidal statistical games

In this section, we exhibit the monoidal structures on copy-composite Bayesian lenses, statistical games, and loss models, as well as demonstrating that each of our loss models is accordingly monoidal. We begin by expressing the monoidal structure on $\mathbf{Copara}_2(\mathcal{C})$.

Proposition B.1. If the monoidal structure on \mathcal{C} is symmetric, then $\mathbf{Copara}_2(\mathcal{C})$ inherits a monoidal structure (\otimes, I) , with the same unit object I as in \mathcal{C} . On 1-cells $f : A \xrightarrow{M} B$ and $f' : A' \xrightarrow{M'} B'$, the tensor $f \otimes f' : A \otimes A' \xrightarrow{M \otimes M'} B \otimes B'$ is defined by



On 2-cells $\varphi : f \Rightarrow g$ and $\varphi' : f' \Rightarrow g'$, the tensor $\varphi \otimes \varphi' : (f \otimes f') \Rightarrow (g \otimes g')$ is given by the string diagram



Next, we define the notion of monoidal indexed bicategory.

Definition B.2. Suppose $(\mathcal{B}, \otimes, I)$ is a monoidal bicategory. We will say that $F : \mathcal{B}^{\text{coop}} \rightarrow \mathbf{Bicat}$ is a *monoidal indexed bicategory* when it is equipped with the structure of a weak monoid object in the 3-category of indexed bicategories, indexed pseudofunctors, indexed pseudonatural transformations, and indexed modifications.

More explicitly, we will take F to be a monoidal indexed bicategory when it is equipped with

- (i) an indexed pseudofunctor $\mu : F(-) \times F(=) \rightarrow F(- \otimes =)$ called the *multiplication*, i.e.,
 - (a) a family of pseudofunctors $\mu_{X,Y} : FX \times FY \rightarrow F(X \otimes Y)$, along with
 - (b) for any 1-cells $f : X \rightarrow X'$ and $g : Y \rightarrow Y'$ in \mathcal{B} , a pseudonatural isomorphism $\mu_{f,g} : \mu_{X',Y'} \circ (Ff \times Fg) \Rightarrow F(f \otimes g) \circ \mu_{X,Y}$;
- (ii) a pseudofunctor $\eta : \mathbf{1} \rightarrow FI$ called the *unit*;

as well as three indexed pseudonatural isomorphisms — an associator, a left unitor, and a right unitor — which satisfy weak analogues of the coherence conditions for a monoidal indexed category [7, §3.2], up to invertible indexed modifications.

Using this notion, we can establish that Stat_2 is monoidal. (So as to demonstrate the structure, we do not omit the proof sketch.)

Theorem B.3. Stat_2 is a monoidal indexed bicategory.

Proof sketch. The multiplication μ is given first by the family of pseudofunctors $\mu_{X,Y} : \text{Stat}_2(X) \times \text{Stat}_2(Y) \rightarrow \text{Stat}_2(X \otimes Y)$ which are defined on objects simply by tensor

$$\mu_{X,Y}(A, B) = A \otimes B$$

since the objects do not vary between the fibres of Stat_2 , and on hom categories by the functors

$$\begin{aligned} & \text{Stat}_2(X)(A, B) \times \text{Stat}_2(Y)(A', B') \\ &= \mathbf{Cat}(\text{disc } \mathcal{C}(I, X), \mathbf{Copara}_2^r(\mathcal{C})(A, B)) \times \mathbf{Cat}(\text{disc } \mathcal{C}(I, Y), \mathbf{Copara}_2^r(\mathcal{C})(A', B')) \\ &\cong \mathbf{Cat}(\text{disc } \mathcal{C}(I, X) \times \text{disc } \mathcal{C}(I, Y), \mathbf{Copara}_2^r(\mathcal{C})(A, B) \times \mathbf{Copara}_2^r(\mathcal{C})(A', B')) \\ &\xrightarrow{\mathbf{Cat}(\text{disc } \mathcal{C}(I, \text{proj}_X) \times \text{disc } \mathcal{C}(I, \text{proj}_Y), \otimes)} \mathbf{Cat}(\text{disc } \mathcal{C}(I, X \otimes Y)^2, \mathbf{Copara}_2^r(\mathcal{C})(A \otimes A', B \otimes B')) \\ &\xrightarrow{\mathbf{Cat}(\multimap, \text{id})} \mathbf{Cat}(\text{disc } \mathcal{C}(I, X \otimes Y), \mathbf{Copara}_2^r(\mathcal{C})(A \otimes A', B \otimes B')) \\ &= \text{Stat}_2(X \otimes Y)(A \otimes A', B \otimes B'). \end{aligned}$$

where $\mathbf{Cat}(\multimap, \text{id})$ indicates pre-composition with the universal (Cartesian) copying functor. For all $f : X \rightarrow X'$ and $g : Y \rightarrow Y'$ in $\mathbf{Copara}_2^r(\mathcal{C})$, the pseudonatural isomorphisms

$$\mu_{f,g} : \mu_{X',Y'} \circ (\text{Stat}_2(f) \times \text{Stat}_2(g)) \Rightarrow \text{Stat}_2(f \otimes g) \circ \mu_{X,Y}$$

are obtained from the universal property of the product \times of categories. The unit $\eta : \mathbf{1} \rightarrow \text{Stat}_2(I)$ is the pseudofunctor mapping the unique object of $\mathbf{1}$ to the monoidal unit I . Associativity and unitality of this monoidal structure follow from the functoriality of the construction, given the monoidal structures on \mathcal{C} and \mathbf{Cat} . \square

Just as the monoidal Grothendieck construction induces a monoidal structure on categories of lenses for monoidal pseudofunctors [7], we obtain a monoidal structure on the bicategory of copy-composite bayesian lenses.

Corollary B.4. The bicategory of copy-composite Bayesian lenses $\mathbf{BayesLens}_2$ is a monoidal bicategory. The monoidal unit is the object (I, I) . The tensor \otimes is given on 0-cells by $(X, A) \otimes (X', A') := (X \otimes X', A \otimes A')$, and on hom-categories by

$$\begin{aligned} & \mathbf{BayesLens}_2((X, A), (Y, B)) \times \mathbf{BayesLens}_2((X', A'), (Y', B')) \\ &= \mathbf{Copara}_2^l(\mathcal{C})(X, Y) \times \text{Stat}_2(X)(B, A) \times \mathbf{Copara}_2^l(\mathcal{C})(X', Y') \times \text{Stat}_2(X')(B', A') \\ &\xrightarrow{\sim} \mathbf{Copara}_2^l(\mathcal{C})(X, Y) \times \mathbf{Copara}_2^l(\mathcal{C})(X', Y') \times \text{Stat}_2(X)(B, A) \times \text{Stat}_2(X')(B', A') \\ &\xrightarrow{\otimes \times \mu_{X,X'}^{\text{op}}} \mathbf{Copara}_2^l(\mathcal{C})(X \otimes X', Y \otimes Y') \times \text{Stat}_2(X \otimes X')(B \otimes B', A \otimes A') \\ &= \mathbf{BayesLens}_2((X, A) \otimes (X', A'), (Y, B) \otimes (Y', B')). \end{aligned}$$

And similarly, we obtain a monoidal structure on statistical games.

Proposition B.5. The bicategory of copy-composite statistical games **SGame** is a monoidal bicategory. The monoidal unit is the object (I, I) . The tensor \otimes is given on 0-cells as for the tensor of Bayesian lenses, and on hom-categories by

$$\begin{aligned}
& \mathbf{SGame}((X, A), (Y, B)) \times \mathbf{SGame}((X', A'), (Y', B')) \\
&= \mathbf{BayesLens}_2((X, A), (Y, B)) \times \mathbf{Stat}(X)(B, I) \\
&\quad \times \mathbf{BayesLens}_2((X', A'), (Y', B')) \times \mathbf{Stat}(X')(B', I) \\
&\xrightarrow{\sim} \mathbf{BayesLens}_2((X, A), (Y, B)) \times \mathbf{BayesLens}_2((X', A'), (Y', B')) \\
&\quad \times \mathbf{Stat}(X)(B, I) \times \mathbf{Stat}(X')(B', I) \\
&\xrightarrow{\otimes \times \mu_{X, X'}} \mathbf{BayesLens}_2((X, A) \otimes (X', A'), (Y, B) \otimes (Y', B')) \times \mathbf{Stat}(X \otimes X')(B \otimes B', I \otimes I) \\
&\xrightarrow{\sim} \mathbf{SGame}((X, A) \otimes (X', A'), (Y, B) \otimes (Y', B'))
\end{aligned}$$

where here μ indicates the multiplication of the monoidal structure on \mathbf{Stat} [**Smithe2022Mathematical**].

We give natural definitions of monoidal inference system and monoidal loss model, which we elaborate below.

Definition B.6. A (lax) monoidal inference system is an inference system (\mathcal{D}, ℓ) for which ℓ is a lax monoidal pseudofunctor. A (lax) monoidal loss model is a loss model L which is a lax monoidal lax functor.

Remark B.7. We say ‘lax’ whenever a morphism (of any structure) only weakly preserves a monoidal operation such as composition of any order; this includes as a special case lax monoidal functors (since a monoidal category is a one-object bicategory). In this respect, we differ from [7, §2.2], who use ‘weak’ in the latter case; we prefer to maintain consistency. Following [6, Def. 4.2.1], we will continue to say *lax* when the witness to laxness maps composites of images to images of composites (and *oplax* when the witness maps images of composites to composites of images).

These conventions mean that a loss model $L : \mathcal{B} \rightarrow \mathbf{SGame}$ is lax monoidal when it is equipped with strong transformations

$$\begin{array}{ccc}
\mathcal{B} \times \mathcal{B} & \xrightarrow{L \times L} & \mathbf{SGame} \times \mathbf{SGame} \\
\otimes_{\mathcal{B}} \downarrow & \swarrow \lambda & \downarrow \otimes_{\mathbf{G}} \\
\mathcal{B} & \xrightarrow{L} & \mathbf{SGame}
\end{array}
\quad \text{and} \quad
\begin{array}{ccc}
\mathbf{1} & \xrightarrow{(I, I)} & \mathbf{1} \\
(I, I) \downarrow & \swarrow \lambda_0 & \downarrow \\
\mathcal{B} & \xrightarrow{L} & \mathbf{SGame}
\end{array}$$

where $\otimes_{\mathcal{B}}$ and $\otimes_{\mathbf{G}}$ denote the monoidal products on $\mathcal{B} \hookrightarrow \mathbf{BayesLens}_2$ and \mathbf{SGame} respectively, and when λ and λ_0 are themselves equipped with invertible modifications satisfying coherence axioms, as in Moeller and Vasilakopoulou [7, §2.2].

Note that, because L must be a (lax) section of the 2-fibration $\pi_{\text{Loss}}|_{\mathcal{B}} : \mathbf{SGame}|_{\mathcal{B}} \rightarrow \mathcal{B}$, the unitor λ_0 is forced to be trivial, picking out the identity on the monoidal unit (I, I) . Likewise, the laxator $\lambda : L(-) \otimes L(=) \Rightarrow L(- \otimes =)$ must have 1-cell components which are identities:

$$L(X, A) \otimes L(X', A) = (X, A) \otimes (X', A) = (X \otimes X', A \otimes A) = L((X, A) \otimes L(X', A))$$

The interesting structure is therefore entirely in the 2-cells. We follow the convention of [6, Def. 4.2.1] that a strong transformation is a lax transformation with invertible 2-cell components. Supposing

that $(c, c') : (X, A) \rightarrow (Y, B)$ and $(d, d') : (X', A') \rightarrow (Y', B')$ are 1-cells in \mathcal{B} , the corresponding 2-cell component of λ has the form $\lambda_{c,d} : L((c, c') \otimes (d, d')) \Rightarrow L(c, c') \otimes L(d, d')$, hence filling the following square in **SGame**:

$$\begin{array}{ccc}
 (X, A) \otimes (X', A') & \xrightarrow{L(c, c') \otimes L(d, d')} & (Y, B) \otimes (Y', B') \\
 \parallel & \nearrow \lambda_{c,d} & \parallel \\
 (X, A) \otimes (X', A') & \xrightarrow{L((c, c') \otimes (d, d'))} & (Y, B) \otimes (Y', B')
 \end{array}$$

Intuitively, these 2-cells witness the failure of the tensor $L(c, c') \otimes L(d, d')$ of the parts to account for correlations that may be evident to the “whole system” $L((c, c') \otimes (d, d'))$.

Just as we have monoidal lax functors, we can have monoidal lax transformations; again, see [7, §2.2].

Proposition B.8. Monoidal loss models and monoidal icons form a subcategory $\text{MonLoss}(\mathcal{B})$ of $\text{Loss}(\mathcal{B})$, and the symmetric monoidal structure $(+, 0)$ on the latter restricts to the former.

B.1 Examples

In this section, we present the monoidal structure on the loss models considered above. Because loss models L are (lax) sections, following Remark B.7, this monoidal structure is given in each case by a lax natural family of 2-cells $\lambda_{c,d} : L((c, c') \otimes (d, d')) \Rightarrow L(c, c') \otimes L(d, d')$, for each pair of lenses $(c, c') : (X, A) \rightarrow (Y, B)$ and $(d, d') : (X', A') \rightarrow (Y', B')$. Such a 2-cell $\lambda_{c,d}$ is itself given by a loss function of type $B \otimes B' \xrightarrow{X \otimes X'} I$ satisfying the equation $L((c, c') \otimes (d, d')) = L(c, c') \otimes L(d, d') + \lambda_{c,d}$. Following [6, Eq. 4.2.3], lax naturality requires that λ satisfy the following equation of 2-cells, where K denotes the laxator (with respect to horizontal composition \diamond) with components $K(e, c) : Le \diamond Lc \Rightarrow L(e \circ c)$:

$$\begin{array}{ccccc}
 & & (Y, B) \otimes (Y', B') & & \\
 & \nearrow^{L(c \otimes d)} & \downarrow K(e \otimes f, c \otimes d) & \searrow^{L(e \otimes f)} & \\
 (X, A) \otimes (X', A') & \xrightarrow{\quad} & L((e \circ c) \otimes (f \circ d)) & \xrightarrow{\quad} & (Z, C) \otimes (Z', C') \\
 & \searrow_{L(e \circ c) \otimes L(f \circ d)} & \downarrow \lambda(e \circ c, f \circ d) & & \\
 & & = & & \\
 & \nearrow^{L(c \otimes d)} & (Y, B) \otimes (Y', B') & \searrow^{L(e \otimes f)} & \\
 & \downarrow \lambda(c, d) & \parallel & \downarrow \lambda(e, f) & \\
 (X, A) \otimes (X', A') & \xrightarrow{Lc \otimes Ld} & (Y, B) \otimes (Y', B') & \xrightarrow{Le \otimes Lf} & (Z, C) \otimes (Z', C') \\
 & \searrow_{L(e, c) \otimes K(f, d)} & \downarrow K(e, c) \otimes K(f, d) & & \\
 & & L(e \circ c) \otimes L(f \circ d) & &
 \end{array}$$

Since vertical composition in **SGame** is given on losses by $+$, we can write this equation as

$$\begin{aligned} & \lambda(e \circledast c, f \circledast d) + K(e \otimes f, c \otimes d) \\ &= \lambda(e, f) \diamond \lambda(c, d) + K(e, c) \otimes K(f, d) \\ &= \lambda(e, f)_{c \otimes d} + \lambda(c, d) \circ (e' \otimes f')_{c \otimes d} + K(e, c) \otimes K(f, d). \end{aligned} \quad (5)$$

In each of the examples below, therefore, we establish the definition of the laxator λ and check that it satisfies equation 5.

We will often use the notation $(-)_X$ to denote projection onto a factor X of a monoidal product.

B.1.1 Relative entropy

Proposition B.9. The loss model KL of Proposition 3.14 is lax monoidal. Supposing that $(c, c') : (X, X) \mapsto (Y, Y)$ and $(d, d') : (X', X') \mapsto (Y', Y')$ are lenses in \mathcal{B} , the corresponding component $\lambda^{\text{KL}}(c, d)$ of the laxator is given, for $\omega : I \mapsto X \otimes X'$ and $(y, y') : Y \otimes Y'$, by

$$\lambda^{\text{KL}}(c, d)_{\omega}(y, y') := \mathbb{E}_{\substack{(x, x', m, m') \sim \\ (c'_{\omega_X} \otimes d'_{\omega_{X'}})(y, y')}} \left[\log \frac{P_{\omega_X \otimes \omega_{X'}}(x, x')}{P_{\omega}(x, x')} \right] + \log \frac{P_{(c \otimes d)^{\dagger} \bullet \omega}(y, y')}{P_{(c \otimes d)^{\dagger} \bullet (\omega_X \otimes \omega_{X'})}(y, y')}.$$

B.1.2 Maximum likelihood estimation

Proposition B.10. The loss model MLE of Proposition 3.16 is lax monoidal. Supposing that $(c, c') : (X, X) \mapsto (Y, Y)$ and $(d, d') : (X', X') \mapsto (Y', Y')$ are lenses in \mathcal{B} , the corresponding component $\lambda^{\text{MLE}}(c, d)$ of the laxator is given, for $\omega : I \mapsto X \otimes X'$ and $(y, y') : Y \otimes Y'$, by

$$\lambda^{\text{MLE}}(c, d)_{\omega}(y, y') := \log \frac{P_{(c \otimes d)^{\dagger} \bullet (\omega_X \otimes \omega_{X'})}(y, y')}{P_{(c \otimes d)^{\dagger} \bullet \omega}(y, y')}.$$

B.1.3 Free energy

Corollary B.11. The loss model FE of Definition 3.17 is lax monoidal. Supposing that $(c, c') : (X, X) \mapsto (Y, Y)$ and $(d, d') : (X', X') \mapsto (Y', Y')$ are lenses in \mathcal{B} , the corresponding component $\lambda^{\text{FE}}(c, d)$ of the laxator is given, for $\omega : I \mapsto X \otimes X'$ and $(y, y') : Y \otimes Y'$, by

$$\lambda^{\text{FE}}(c, d)_{\omega}(y, y') := \mathbb{E}_{(x, x') \sim (c'_{\omega_X} \otimes d'_{\omega_{X'}})(y, y')} \left[\log \frac{P_{\omega_X \otimes \omega_{X'}}(x, x')}{P_{\omega}(x, x')} \right].$$

B.1.4 Laplacian free energy

Proposition B.12. The loss model LFE of Propositions 3.21 and 3.23 is lax monoidal. Supposing that $(c, c') : (X, X) \mapsto (Y, Y)$ and $(d, d') : (X', X') \mapsto (Y', Y')$ are lenses in \mathcal{B} , the corresponding component $\lambda^{\text{LFE}}(c, d)$ of the laxator is given, for $\omega : I \mapsto X \otimes X'$ and $(y, y') : Y \otimes Y'$, by

$$\lambda^{\text{LFE}}(c, d)_{\omega}(y, y') := \log \frac{P_{\omega_X \otimes \omega_{X'}}(\mu_{(c \otimes d)'_{\omega}}(y, y')_{XX'})}{P_{\omega}(\mu_{(c \otimes d)'_{\omega}}(y, y')_{XX'})}$$

where $\mu_{(c \otimes d)'_{\omega}}(y, y')_{XX'}$ is the $(X \otimes X')$ -mean of the Gaussian distribution $(c'_{\omega_X} \otimes d'_{\omega_{X'}})(y, y')$.